# Harm and Responsibility in Hate Speech

**Robert Mark Simpson**




A Thesis Submitted for the Degree of
Doctor of Philosophy in Philosophy
at the University of Oxford

Hilary 2013

# Abstract


## Harm and Responsibility in Hate Speech

Robert Mark Simpson, Somerville College

Submitted for the degree of DPhil, Hilary 2013

The legal restriction of hate speech – i.e. speech that expresses contempt for people on the basis of their ethnicity, religion, or sexuality – is now commonplace in liberal legal systems outside the United States. This thesis takes up the question of whether restrictions on hate speech are generally justifiable. I begin by explaining why liberals should not dismiss anti-hate speech law from the outset as an intolerable violation of free speech. My analysis of the case for anti-hate speech law is thereafter framed by two main concerns. Firstly, I stress that if we are to impose legal restrictions on hate speech, we must establish not just that there are harmful outcomes associated with hate speech, but that those who engage in hate speech are *responsible* for those outcomes. Secondly, I argue that restrictions on hate speech should be assessed in two distinct classes. Inquiries into the justificatory bases of anti-hate speech law are typically conducted as if informative generalisations can be made about how the law should respond to anything that is properly called hate speech. Against this approach, I argue that while the liberal state can and should impose restrictions on directly harmful hate speech (in which hate speech is used to threaten, harass, and incite violence), restrictions on *indirectly* harmful hate speech – in which hate speech (allegedly) contributes to identity-based social hierarchies and their concomitant harms – are not justifiable. The problem with restrictions on indirectly harmful hate speech is not the structure of the liability-ascription framework under which they operate. Rather, I argue, the problem is epistemic: we cannot confidently judge that hate-speakers are in fact responsible for contributing, more than trivially, to the harmful patterns of identity-based inequality and disadvantage in light of which restrictions on indirectly harmful hate speech may be defended in principle.


Thesis length: 74,803 words (including footnotes)

# Acknowledgements

I am very grateful to Leslie Green, under whose supervision this thesis was researched and written, and whose comments, criticisms, and corrections helped me to develop my views around this subject. I am also grateful for the comments I received from others in relation to my work on this subject. My thanks in particular go to Hamish Bridges, Roger Crisp, Hannah Field, John Gardner, Rae Langton, Katherine Simpson, and James Weinstein. Special gratitude is owed to Katherine who – as well as providing comments and corrections on the written work – was a steadfast source of cheer and encouragement throughout my research.

My doctoral candidacy at the University of Oxford was made possible by the financial support of Somerville College and the Oxford Faculty of Philosophy. For support and advice at Somerville, thanks to Paula Boddington, Lesley Brown, Hilary Greaves, Julie Hage, Alice Prochaska, and Lee Walters. For support and advice in the Philosophy Faculty, thanks to Harvey Brown, Iris Geens, Catriona Hopton, Paul Lodge, and Scott Sturgeon.

# Table of Contents

**1. Hate speech, equality, and identity ..... 11**

1.1 Liberal divides in free speech theory ..... 11

1.2 Anti-hate speech law ..... 14

1.3 Liberalism, freedom, and equality ..... 20

1.4 The priority of harm and responsibility ..... 24

1.5 Inequality, oppression, and harm ..... 27

1.6 Religion, ethnicity, and sexuality ..... 29

**2. Free speech and speech acts ..... 35**

2.1 Free speech arguments ..... 35

2.2 Protected species and the meaning of 'speech' ..... 39

2.3 The First Amendment, free speech, and hate speech ..... 43

2.4 Should hate speech be a protected species? ..... 47

2.5 Viewpoint neutrality and the speech-conduct distinction ..... 51

2.6 Situating pornography ..... 58

**3. Harm and responsibility ..... 63**

3.1 Direct and indirect harms ..... 63

3.2 The harm question ..... 70

3.3 Autonomy, interests, and the harm principle ..... 75

3.4 Causation and responsibility ..... 79

3.5 Proximate causation: striking a balance ..... 82

3.6 Cognitive mediation and liability-ascription ..... 86

3.7 Activity and passivity in offence ..... 90

**4. Directly harmful hate speech ..... 95**

4.1 Responsibility-ascription in verbal assault ..... 96

4.2 What's different in hate speech? ..... 100

4.3 Mechanisms of harm in verbally-assaultive hate speech ..... 104

4.4 Responsibility-ascription in incitement ..... 108

4.5 What's different in hate speech? ..... 113

4.6 Intention and expectation ..... 118

4.7 The question of redundancy ..... 123

**5. Indirectly harmful hate speech ..... 127**

5.1 The etiology of social hierarchy ..... 128

5.2 Recognition and two senses of dignity ..... 132

5.3 Status equality and assurance ..... 138

5.4 Legitimation and authority in hate speech ..... 145

5.5 Authority in illocutionary action ..... 152

5.6 Social hierarchy and communicative impairment ..... 156

5.7 Score-keeping in an identity-oppressive language game ..... 160

5.8 Empirical data, linguistic phenomena, and indirect harm ..... 166

5.9 On setting a high bar ..... 171

**6. Conclusion ..... 177**

6.1 Summary: harm and responsibility ..... 177

6.2 Eroding social hierarchy and spreading the burden ..... 180

**Bibliography ..... 185**

# 1

# Hate speech, equality, and identity

## 1.1 Liberal divides in free speech theory

Free speech is one of the things that characterises and unites Western liberal states, or so we normally suppose. But this broad generalisation is over-simple nowadays, if it was ever otherwise. Several contemporary legal scholars – most notably Owen Fiss, in *Liberalism Divided* (1996) – have observed a major schism in free speech theory and practice since the late 1970s. In the United States, they tell us, free speech has taken on a libertarian guise, with the judiciary rejecting regulatory policy at every significant turn. Simultaneously, in most other liberal states, they see expressive liberty increasingly being constrained by, and forced to compromise with, socially egalitarian ideals. Hence we can no longer casually speak of free speech as a hallmark of Western liberalism; free speech is a partitioned axiological terrain.

Just as significant as the partition itself, however, is the fact that on either side, the evolving trajectories of free speech doctrine seem to bring other liberal values into jeopardy. In the US, the now-prevalent libertarian approach to free speech vigilantly opposes attempts by the state to constrain or dictate public opinion. But proponents of this approach seem indifferent to the corresponding ways in which, in an unregulated expressive milieu, private actors can use wealth and media-influence to sway and sometimes subvert democratic self-government. The Supreme Court's 2009 *Citizens United* ruling,[1] with its far-reaching deregulation of electoral campaign finance law, was a major victory for libertarian thinking in this arena – though it remains to be seen just how, and to what degree, the integrity of US democracy will be affected by this and other similarly-oriented developments.

The prevailing free speech paradigm outside the US is also seen by some as a threat to democratic integrity, albeit a threat from the opposite direction. Whereas, in the US, the state

[1] *Citizens United v. Federal Election Commission* [2010] 558 US 310 S. Ct. 876.

is reluctant to regulate public expression, even if this means allowing sinister or misleading voices to drown out others and distort public discourse, other liberal states are – so their critics allege – all too *eager* to interfere with public discourse, by penalising or otherwise regulating various forms of expression. The eagerness is most apparent, and most widely condemned, when it comes to the legal restriction of *hate speech*, i.e. speech that expresses contempt for people on the basis of their ethnicity, religion, or sexuality. No-one doubts that anti-hate speech laws are well-intentioned. Typically, they exist among a suite of policies aimed at redressing systemic inequality and identity-based social hierarchy. On any broadly egalitarian form of liberalism, these are legitimate (indeed, urgent) aims, which the state can and should use its legislative power to pursue. So what are the objections to anti-hate speech law? Firstly, some critics simply see restrictions on hate speech as an outright desertion of any robust commitment to free speech in the first place. Secondly, some critics express doubts about whether hate speech actually effects the ills that are attributed to it. Why, they ask, should we think hate speech is really responsible for identity-based social hierarchy? Why suppose that restricting hate speech will do anything to remedy discrimination and redress inequality? Without satisfactory answers to these objections, the widespread legal restriction of hate speech in Western liberal states today looks, at best, like a well-meaning but wrongheaded policy misadventure, and at worst, like an illiberal programme of censorship.

However, I believe adequate replies to these objections are available, and in this study I will present a partial defence of anti-hate speech law as a justifiable element in a liberal legal system. I say a *partial* defence for two reasons. Firstly, like many who write in defence of anti-hate speech law, I have reservations about the content and administration of the anti-hate speech statutes that are actually in effect in liberal societies at present. My aim is to defend the in-principle justifiability of a certain kind of legal restriction on hate speech, not to defend the in-practice justified-ness of anti-hate speech law in any given regime. The second reason why my defence will only be partial is because a particular form of hate speech law – perhaps the most recognisable form – is one that I oppose. In what follows I will argue that we need to reconfigure the adjudicatory framework in which anti-hate speech law is assessed. Evaluative analyses of anti-hate speech law in the legal, political, and philosophical literature are generally conducted as if all such laws are of a piece: as if there are useful generalisations to be made about how the law should view and respond to anything that is properly called hate speech. I

will argue that that is a mistake. Hate speech of a certain form – what I call *directly* harmful hate speech – can and should be subject to legal restrictions. However, legal restrictions on hate speech which is only *indirectly* harmful should be rejected. In this study I will explain why this is the case. In addition to this structuring distinction between direct and indirect harm, my second point of emphasis in this study is the notion of responsibility. Advocates of anti-hate speech law frequently engage in a kind of attributive brain-storming activity, in which they posit associations between hate speech and all kinds of social and individual ills, while postponing or ignoring the question of whether hate speech is ultimately responsible for those ills. Throughout this study I will stress that hate speech is only justifiably liable to restriction where we can confidently judge that it is *responsible* for harming its targets. It is primarily in virtue of this principle that I favour the two-tiered approach to anti-hate speech law.

ଔ

My discussion proceeds in the following order. In chapter 2 I begin by explaining why liberals should not dismiss legal restrictions on hate speech from the outset as an intolerable violation of free speech. In chapter 3 I spell out my distinction between (i) directly harmful hate speech, in which there is a particular kind of non-redundant causal relationship between the speech act and the harmful outcome, and (ii) indirectly harmful hate speech, in which hate speech makes a (redundant) causal contribution to a harmful outcome. I also give a detailed account of the notions of harm and responsibility that operate within this framework of analysis. In chapter 4 I focus on legal restrictions punishing and deterring directly harmful hate speech. I argue that such restrictions are justifiable in principle, and moreover, that given how hate speech interacts with identity-prejudice in background social conditions, such restrictions should apply to a more extensive range of expressive acts than what we observe with generic restrictions on other kinds of directly harmful speech (e.g. threats and harassment). In chapter 5, by contrast, I argue that legal restrictions on *indirectly* harmful hate speech are not justifiable. This is not because indirect harmfulness is an insufficient basis for liability-ascription in general. Rather, I reject restrictions on indirectly harmful hate speech for epistemic reasons: we cannot judge with any confidence that hate speech *does* contribute (more than trivially) to the social-hierarchy-based harms with which it is routinely associated.

In the remainder of this introductory chapter I will (i) provide a brief overview of anti-hate speech legislation and its administration today; (ii) explain why discussion of the justifiability of anti-hate speech law should focus on the concepts of harm and responsibility, rather than freedom and equality; and (iii) address some of the problems that arise in the discussion of ethnicity, religion, and sexuality in relation to hate speech.

## 1.2 Anti-hate speech law

In the United Kingdom, under Part 3, Section 18 of the Public Order Act 1986, it is an offence (carrying a maximum sentence of seven years imprisonment) to express racial hatred which (i) has the intent to stir up racial hatred, or (ii) is likely to stir up racial hatred. With the addition of Part 3a in 2006, the Act was extended to encompass incitement to religious hatred; the Act was amended again in 2008 to encompass incitement to hatred on the ground of sexual orientation.[2] These restrictions on hate speech are not dormant or merely for show. To take just a few examples of recent hate speech prosecutions in the UK:

- In March 2012, twenty-one year old student Liam Stacey was sentenced to eight weeks imprisonment for drunkenly posting a series of racist messages on the social networking service Twitter, in which he repeatedly vilified an injured professional footballer, Fabrice Muamba.[3]

- In May 2012, an unemployed office worker, Jacqueline Woodhouse, was sentenced to 21 weeks imprisonment after she shouted racial abuse at several passengers on a London underground train.[4]

---

[2] Under Part 4a of the Public Order Act, introduced in 1994, it is also an offence (carrying a maximum sentence of six months imprisonment) to use threatening, abusive, or insulting words to cause a person harassment, alarm, or distress. For useful overviews and discussions of UK anti-hate speech law and related statutes, see Bindman 1992; Oyediran 1992; Hare 2009; Weinstein 2009a; D. Williams 2009.

[3] 'Student jailed for Racist Fabrice Muamba Tweets', *The Guardian*, 27 March 2012.

[4] 'Racist Tube Rant Woman Jacqueline Woodhouse Jailed', *BBC News Online*, 29 May 2012, http://www.bbc.co.uk/news/uk-england-london-18251807.

- In September 2009, a British diplomat, Rowan Laxton, was convicted of racial harassment and fined £865 after he was heard shouting "f***ing Israelis, f***ing Jews" in a gymnasium.[5]

There are plenty of similar cases to be found elsewhere. For instance:

- In Canada, in March 2006, two white supremacists were fined C$13,000 by the Canadian Human Rights Tribunal for publishing articles on their website in which they vilified several different non-Caucasian groups.[6]

- In France, in June 2008, the retired actress Brigitte Bardot was convicted of inciting racial hatred (for a fifth time) over a letter published on her website in which she said that Muslims were destroying France.[7]

- In Austria, in January 2009, a conservative politician, Susanne Winter, was fined €24,000 and dealt a suspended three month prison sentence, as a result of several of her public statements about Islam, including her statement that the prophet Mohammed was a paedophile.[8]

- In Australia in December 2004, members of the Evangelical group Catch the Fire Ministries were fined under Victoria's Racial and Religious Tolerance Act over anti-Muslim leaflets distributed at a public meeting.[9]

There is considerable variation between jurisdictions with regards to the legal instruments through which these sorts of restrictions and penalties are meted out. Some states have anti-hate speech legislation that is directly modelled on anti-discrimination conventions in

---

[5] 'Foreign office diplomat found guilty of racial harassment', *The Guardian*, 24 September 2009.

[6] 'White supremacists fined for spreading hate on the internet', *CBC News Online*, 10 March 2006, http://www.cbc.ca/canada/story/2006/03/10/supremacists060310.html.

[7] 'Bardot fined over racial hatred', *BBC News Online*, 3 June 2008, http://news.bbc.co.uk/1/hi/entertainment/7434193.stm.

[8] 'Austrian far-right legislator convicted of anti-Muslim incitement', *Earth Times*, 22 January 2009, http://www.earthtimes.org/articles/show/251964,austrian-far-right-legislator-convicted-of-anti-muslim-incitement.html.

[9] 'Historic win in religious hatred case', *The Age*, 18 December 2004.

international law.[10] Under the Canadian Criminal code, for instance: "everyone who, by communicating statements, other than in private conversation, wilfully promotes hatred against any identifiable group [distinguished by colour, race, religion, ethnic origin, or sexual orientation] is guilty of... an indictable offence" (see Sumner 2009: 205, 12-19). In France, by contrast, hate speech is restricted via a patchwork of administrative statutes (e.g. regulating media content, or the activities of political parties), and criminal laws against provocation, apologism for criminal violence, abuse, harassment, and defamation (see Mbongo 2009: 222-29). In Australia, there are a variety of mediating bodies (including the Australian Human Rights Commission and various state tribunals) and higher courts (including state Supreme Courts and the Australian Federal Court) responding to complaints under the auspices of the Racial Discrimination Act 1975, according to which it is unlawful to offend, insult, or humiliate others on grounds of nationality or ethnicity (see Gelber 2007).

Anti-hate speech law is a well-established part of a number of stable and mature liberal legal systems outside the US; it is not just a prospective legislative development, or as some of its critics say (e.g. Kunzru 2005; Haworth 2007), an expedient, short-term reaction to imminent political pressures. At the same time, though, anti-hate speech policy is evidently heterogenous between different states. In some cases the tools used to penalise hate speech belong to the criminal law, in some cases they belong to private law. Certain jurisdictions respond to hate speech with specially crafted legislation, while others employ broader-scope legislative instruments to police it. We can, however, identify at least two common strands in anti-hate speech law, which represent the rejection of two central tenets of modern US free speech jurisprudence. The first of these is the idea that the value of free speech justifies "weighing interests with 'a thumb on the scales' in favour of speech" (Brison 1998a: 319). Hate speech in the US is understood as belonging to *public discourse*, i.e. the special category of expressive conduct that receives maximally stringent protection under the First Amendment. Consequently, US courts impose stringent and heavily-codified burdens of justification on legislatures which purport to restrict hate speech. By contrast, other liberal polities entrust

[10] Restrictions on hate speech are mandated by the *International Covenant on Civil and Political Rights* (article 20.2), and the *International Convention on the Elimination of All Forms of Racial Discrimination* (article 4a). Studies examining the treatment of hate speech in international law include Gordon 1992; Grimm 2009a; Hare 2009b; Miles 2011.

their legislatures to weigh-up and trade-off the demands of public discourse and social equality as and when conflict arises. Legal restrictions on hate speech outside the US are obviously still constrained by routine legislative checks and balances, but they are not subjected to anything like the onerous Constitutional tests that stand in the way of such restrictions within US legislatures. The second tenet of US free speech doctrine, eschewed elsewhere, is its narrow conception of 'viewpoint neutrality'. First Amendment jurisprudence disallows the restriction of speech where the reasoning behind the restriction adverts to the wrongness or impropriety of the viewpoint that stands to be restricted (see Eberle 1994: 1142 ff.; Weinstein 2009b: 84-88). Thus, for instance, it flatly disallows legislation that penalises a White Supremacist belief-system. By contrast, other liberal states do not see free speech as a safeguard of equal standing or protection for all viewpoints before the law, and they legislate accordingly.

In much of the legal literature on hate speech, the justifiability of anti-hate speech law is treated not as an abstract question of principle, but as a question of whether restrictions on hate speech should be permitted in some specific jurisdiction, given the parameters and institutional constraints that apply in that local context (whether due to constitutional law, or issuing from the interpretation of a common law tradition). In the US, where it is generally accepted that broadly-framed prohibitions on hate speech violate the First Amendment, work of this sort often addresses the question of whether it should be permissible for universities to prohibit hate speech within their institutional confines, via so-called 'campus speech codes'.[11] In other liberal states where anti-hate speech law is already in effect, the literature runs in a number of different directions. Unsurprisingly, there are some authors who argue that states like Canada and the UK, which profess a robust commitment to free speech, would more satisfactorily live up to that commitment by mimicing the US approach to hate speech (see

[11] See Strossen 1990; Battaglia 1991; Weinstein 1991; Cox 1995; Golding 2000; Gould 2001; Chiang 2007; Curtis 2009. Defences of campus speech codes have often been advanced alongside more general critiques of US free speech theory – see Brownstein 1991; Delgado 1993; Matsuda 1993; Lawrence 1993, 1995; Stefancic and Delgado 1993, Delgado and Stefancic 1994, 2004, 2009; and Crenshaw 1995 – and consequently this debate often engages with broader tensions in US constitutional law, between the First Amendment's free speech provisions, and the Fourteenth Amendment's equal protection guarantees. In addition to the large body of work examining the status of hate speech in US educational institutions, there is also a literature on the treatment of hate speech in the US media; see for instance Schauer 2005; Bodney 2009; Rowbottom 2009.

Sumner 2009 and Weinstein 2009a respectively). On the other hand, there are many authors who are broadly in favour of anti-hate speech laws, and who aim to show that such laws are consistent with free speech commitments in-principle, even if, in practice, the statutes we have at present are ambiguous, over-inclusive, or unsatisfactorily administered.[12]

These sorts of studies, which assume the perspective of the judge or legislator in a specific jurisdiction, account for much of the evaluative literature on anti-hate speech law. My discussion here, however, is meant to be something quite different, namely, a philosophical inquiry into the justificatory foundations of anti-hate speech law. Rather than asking whether legal restrictions on hate speech are permissible under the parameters that obtain in a certain jurisdiction at a certain time, a philosophical inquiry asks whether legal restrictions on hate speech are permissible or required *period*, irrespective of what the local law of the land happens to be. The fact that Q's constitution forbids restrictions on hate speech is beside the point in a philosophical inquiry. If we identify good reasons for restricting hate speech, this entails that there is a *pro tanto* case for amending Q's constitution. In a similar and complementary way, a philosophical inquiry precedes historical or sociological studies in anti-hate speech law. When we are determining how to best pursue our legitimate legislative aims, of course we should take account of relevant data concerning the effects of hate speech policies here or there, past or present. However, a philosophical inquiry aims to answer the prior question of what sort of legislative aims are legitimate, with respect to hate speech, in the first place.[13]

---

[12] There are numerous contributions in the literature that examine the workings of anti-hate speech law and assume the law-maker's perspective in specific jurisdictions, e.g. Australia: Eastman 1992; Eggerking 1992; Stone and Evans 2006; Gelber 2007; Canada: Manwaring 1992; Magnet 1994; Weinstein 1994; Mahoney 2009; Mullender 2007; France: Mbongo 2009; Hungary: Molnar 2009; and the UK: Bindman 1992; Oyediran 1992; D. Williams 2009. There are also a number of pieces that compare anti-hate speech law across jurisdictions, e.g. Lasson 1984; Mahoney 1995; Douglas-Scott 1999; Boyle 2001; Rosenfeld 2002; Buss 2005; A. Stone 2007; Heinze 2009b.

[13] Discussions of anti-hate speech law which try to draw lessons from the historical treatment of hate speech include Richards 1994a; Tsesis 2000; Gilreath 2009; Post 2009. Another kind of historical induction argument (e.g. in Feinberg 1985b; Curtis 1995; Weinstein and Hare 2009) suggests that because restrictions on speech which once seemed urgent now seem overbearing or groundless, we should be sceptical about the seriousness of the ills which might be thought to justify present-day speech restrictions.

But where do these judgements about legitimacy come from? Talk of a 'philosophical inquiry' into justificatory 'foundations' may be read as promising a discussion of perennial questions about goodness, value, and obligation. But it would be extremely audacious to try to evaluate hate speech policy via this approach. There are many possible starting points for our thinking about value and obligation – concepts like happiness, flourishing, virtue, reason, ownership, preferences, and interests – and it is hard enough to treat any of these concepts individually, and defend the fundamentality of one over the others, let alone to determine the legal institutions that should operate given a particular view of value and obligation, and say how a specific policy issue would be addressed under those institutions. So for practical purposes at least, the foundations that we're looking for cannot be sought at quite *that* depth. It is true, granted, that one possible starting point for thinking about the bases of value and obligation – namely, the ideal of equal moral worth, as something that all people are endowed with – might have a more immediate bearing on our surface-level concerns regarding anti-hate speech law.[14] Insofar as hate speech conspicuously flouts the ideal of equal moral worth, one who imputes foundational importance to this ideal might want to claim that hate speech is liable to be met with whatever opposition we can muster against it (legislative or otherwise). But this is implausible. In treating equal moral worth as the basis of political morality, one still cannot reach any conclusions about anti-hate speech law until one spells out (i) what this ideal collectively demands of us, (ii) the structure of the institutions that it would underwrite, and

---

[14] Although we associate the ideal of equal moral worth with Enlightenment works such as Locke's *Letter Concerning Toleration* (1689) and *Second Treatise on Civil Government* (1690), it need not be regarded as a parochial Enlightenment 'doctrine' (as in, e.g. Weinstein 2009a: 27), since it need not incorporate 17th Century prejudices about the meanings of 'personhood' or 'equality'. When taken as a putative basis of value and obligation in general, the ideal of equal moral worth functions as a locus of interpretation, around which different substantive conceptions of political morality may be constructed (see Nagel 1979: 112; Kymlicka 2002b: 44-45). Under this programmatic approach, endorsements of identity-prejudice today cannot be understood as moral claims as such. Under that guise, as Williams says (2005c: 7), they are either unintelligible or just grossly ignorant. Having said that, as a putative foundation of value or obligation in general, the normative thesis, that 'people ought to be treated as having equal moral worth' should be distinguished from the corollary descriptive thesis that 'people as a matter of fact possess equal moral worth'. Attempts to defend the latter claim, as Williams says (1973b), chart a difficult course between (i) blatant falsehoods, and (ii) trivial observations that cannot sustain the justificatory weight ascribed to them.

(iii) the rights and duties that these institutions would ascribe. In short, whatever our views about the fundamental bases of obligation and value, we cannot ascertain the justifiability of anti-hate speech law through an inquiry aimed at that foundational level.

## 1.3 Liberalism, freedom, and equality

Already, then, we need to turn our gaze back towards extant legal regimes. Though we are not interested in the institutional parameters that obtain in specific jurisdictions, with all their local contingencies, we are interested in the practices and institutions that operate across the regimes where anti-hate speech law is in effect, and the generalisations that can be made on the basis of these shared features. Suffice to say, then, that we are interested in the justifiability of legally restricting hate speech within a *liberal* system of political morality. The foundations that we are trying to uncover, in order to stand back from the contingent realities of extant regimes and say how things *should* be, reside in the values embodied in liberalism. Although the methodology is not always characterised as such, this is the approach that is taken in most philosophically-inclined work on hate speech. Certain values will be invoked, described, and then shown to require or forbid the restriction of hate speech. Values invoked in this way may include freedom of speech, liberty *per se*, autonomy, equality, respect, dignity, democracy, and tolerance. To the question of why one should care about what any of these values require, the candid author may identify her favoured values as belonging to the liberal ethos under which the anti-hate speech laws that exist in the world today generally operate. The less candid author may invoke one or more of these values without any explicit justification, assuming that her readers, as members of liberal societies, will take their demands seriously in any case. One way or the other, the approach is one which treats certain values as given *by* the liberal order to which its policy recommendations are being addressed, and which fills out the content of its recommendations by elaborating the content of those values.

This is, in essence, the approach I take in this study. However, the values that I home in on within this approach are different from those that are used – mistakenly, I believe – to frame much philosophical work on hate speech. The literature is full of entries that portray the debate over anti-hate speech law as a conflict between, on one hand, values concerning the protection of individual liberty against the state's encroachments, and on the other, values

founded in an opposition to identity-based oppression and social hierarchy. In short, the issue of hate speech policy is treated as a quandary over whether to permit hate speech in the name of freedom, or to restrict hate speech in the name of equality.[15] Through the question of whether to restrict hate speech, Owen Fiss says, liberals are "being asked to choose between their defining values" (1995: 287). Or, as the authors of an early and influential volume on hate speech say, the fight over anti-hate speech law is "a fight about the substantive content that we will give to the ideals of freedom and equality" (Lawrence et al. 1993: 15).

Despite its agreeable simplicity, this way of framing the debate is the source of much confusion, arising on both sides of the postulated dichotomy. On one side, liberals have a firm, standing preference for liberty. But the argument that a certain policy should be rejected because it restricts liberty is uninformative. Human liberty is curtailed, by design, in any and every legal restriction on conduct. In all sorts of cases that is seen as uncontroversially justifiable. Granted, the worry may be that *expressive* liberty in particular is restricted in anti-hate speech law. In chapter 2 I will address the claim that anti-hate speech law intolerably impinges upon free speech. My point for now is that an assertion of the priority of liberty *per se* provides no substantial guidance about whether we should restrict hate speech; the whole question hinges on what manner of ills or adversities hate speech produces.

At least then, so one might think, the freedom-versus-equality framework says something about the other side of the ledger: because hate speech is in some sense *anti*-equality, its restriction is required for the sake of equality. But attempts to elaborate this line of argument are neither persuasive nor cogent. Firstly, even if we accept some simple, binary description of how equality and freedom correspond with the competing positions on hate speech policy, there is no arbitrating formula in liberalism which explains when and why the demands of one of these values should give way to the demands of the other, at least, not when we are treating the values as nebulous desiderata. If the question of hate speech policy is viewed as a contest of this kind, then endorsing either side – favouring anti-hate speech law for the sake of equality, or opposing it in the name of liberty – seems like an arbitrary selection. The second problem with this framing of the issue is that a concern for equality can also underwrite an *opposition* to anti-hate speech laws. Where liberals will say that equality is

[15] Studies that frame things in these terms include Massaro 1990; Boyle 1992; Walker 1994; Delgado and Stefancic 1994, 1996; Powell 1995; Mahoney 1996; Boyle 2001; Tate 2008.

violated when hateful vilification is protected by law, others will say equality is violated, in the most important sense, when the state selectively interferes with certain individuals voicing their opinions and expressing their convictions. Neither judgement misconstrues the semantic content of 'equality'; rather, the two views encode different understandings about the moral economy of equality: what it requires, to whom it applies, and how its transactions are governed.[16] In view of this duality, one might suggest that equality is an 'essentially contested' concept.[17] Notice, however, that these divergent views about what equality requires derive from a tension at the very core of liberal theorising. Liberals want individuals to be able to govern their own lives as much as possible. On one hand, this requires a delegation of power to the state, since judicious state intervention is (in a wide range of cases) the best way to reliably prevent people from making incursions against other people's freedom, whether through sporadic acts of violence or through ongoing relationships of domination. However, having vested interventionist power in the state, liberals are equally concerned to protect people's freedom against aggressive or gratuitous encroachments from the state itself. The tensions that arise in negotiating this balance are expressed in free speech policy generally, and

[16] This recalls the dialogue between Dworkin (1981b), Cohen (1989), and others, about the 'currency' of egalitarian justice. For Dworkin, equality resides in individual rights that enable participation in collective governance; "it would be a serious misunderstanding of citizen equality to suppose that allowing... damaging political opinions free circulation offends the equality in question" (2000: 366). Unsurprisingly, there are contrasting facets in the concept of freedom here too. Several authors argue that anti-hate speech laws are justified for the sake of the freedom of hate speech's targets (e.g. Michelman 1995; Mahoney 1996; others defend a parallel view about anti-pornography law, e.g. Easton 1995; Hornsby 1995a). Those who adopt a freedom-versus-equality framework should also entertain the possibility of these two values being made wholly compatible through well-crafted social policy (see Dworkin 1978; 2001). Although as Williams notes (2005b), this can have a sinister complexion, for it requires us to say that people whose conduct is restricted for the sake of others' equality have not had their *freedom* curtailed, and that where it seems this way, we (and they) have just misunderstood what freedom truly consists in.

[17] The idea (from Gallie 1962, and elaborated (e.g.) in Gray 1983 and Connolly 1993) is that when interlocutors with deeply divergent views seek to orient their discussion around a complex evaluative concept (such as, in this case, 'equality'), the application of the concept is bound to be contested at points of substantive disagreement, but at the same time, likely to be agreed upon over a significant range of cases, such that the interlocutors can (rightly) take themselves to be discussing the same concept, while nevertheless disagreeing irreconcilably about its evaluative ramifications.

in hate speech policy specifically (see Peters 2005: 14-22). In relation to anti-hate speech law, the tension is this: that if the state restricts speech which erodes *social* equality, this threatens equality *in rights* (of expression) held against the state. Concern for equality *per se* can therefore draw us in either direction with respect to the restriction of hate speech.

Now, one might accept that there is flexibility in how our moral vocabulary maps on to different views about hate speech policy, while still maintaining that hate speech does real damage to its targets, and in a manner that leaves them worse off than groups and individuals who are not victimised via hate speech. Is it not enough, then, to more clearly specify the kinds of damage and disadvantage that one has in mind in advancing the claim that hate speech should be restricted for the sake of equality? This is a step in the right direction, but it indicates, quite rightly, that the case for restricting hate speech does not finally hinge on the demands of equality, any more than the case for restricting racially-motivated violence does. Instead, the case for legally restricting hate speech hinges on the claim that hate speech *harms* people. Now, if this is right – if hate speech really does harm people – then it will often also be true that those harmed by hate speech are victims of identity-based inequality at large. And the harm done by hate speech may consist in its perpetuating or reinforcing the patterns of unequal treatment that its targets experience. Nevertheless, in stating the case for restrictions on hate speech, we should put the emphasis on the (alleged) harm, and not the inequality with which that harm is or may be enmeshed. Our question is not just what society looks like with or without hate speech in it; our question is whether hate speech is in some ascertainable way *to blame* for the ills which give rise to identity-based hierarchy, discrimination, harassment, etc., and whether we can therefore use the law to deter or redress hate speech. Talk of equality and inequality may be used to describe – in a general manner, fit for certain purposes – what is at stake in permitting or restricting hate speech. But this vocabulary is unsuited to articulating the causal connections between speakers, hate-speech-acts, and the adverse outcomes produced by such acts, and these are the connections that we need to examine when judging whether the state may hold individuals answerable for engaging in hate speech.

## 1.4 The priority of harm and responsibility

Two questions are of key importance, then, in determining whether legal restrictions on hate speech are justifiable in the liberal legal regimes where they operate. The first is *the harm question*: are there harmful outcomes that can plausibly be associated with hate speech? Or to put it another way: among the various *adverse* outcomes associated with hate speech, which of them (if any) are genuinely *harmful*? The prevention of harmful action is among the law's primary purposes, on a liberal view, insofar as this protects individuals from other people's liberty-impairing incursions. If this is right – if the law *should* aim primarily at preventing harmful acts – then legal restrictions on hate speech are only justifiable where the negative effects of hate speech are properly harmful. It's true, of course, that some conservative theorists see the criminal law as a tool for prohibiting immoral acts regardless of whether they harm individuals, and advocates of this moralistic conception of law may endorse restrictions on hate speech merely on account of its abject moral character. I follow most legal theorists who discuss hate speech in rejecting that view from the outset. If hate speech were harmless, its legal restriction would not be justifiable, as far as my analysis here goes.

The second key question, in determining whether restrictions on hate speech are justifiable, is *the responsibility question*: with regards to the harmful outcomes that are or can be associated with hate speech, for which of those outcomes is hate speech actually *responsible*, such that it is fair and fitting for hate-speakers (i.e. those who perform acts of hate speech) to be made liable for those harms? The overarching aim of harm-prevention does not justify ascribing blame and liability to anyone whose punishment may somehow be useful to this end. We must ensure, firstly, that the forms conduct which we are proscribing are in fact the (or *a*) source of the harms whose prevention we are aiming at, and secondly, that those punished through the enforcement of our proscriptions are, in all appropriate respects, culpable for engaging in the proscribed activities. To punish those who are innocent (in either of the above senses), is to run roughshod over the same ideals that underwrite the harm principle itself, i.e. people being able to live freely according to their own values and purposes.

The responsibility question is raised less often in the hate speech literature than the harm question, and this is most likely because it appears to contain its own answer. When we ask whether hate-speakers are responsible for the harms of hate speech, the terms of the question itself reply in the affirmative. But ascriptions of responsibility are complicated in this

arena by the character of the causal processes through which hate speech affects its targets. Many of the harms attributed to or associated with hate speech are outcomes which it seems like hate speech contributes to without being strictly necessary for their occurrence. Nearly all the harms associated with hate speech are such that where hate speech *does* contribute to them, the causal processes linking the speech to the harm involve intermediating mental states (intentions, beliefs, feelings, etc.) on the part of responsible agents besides the speaker. When such complexities are observed in the literature, they are usually seen as complicating factors pertaining to the question of whether hate speech is harmful in the first place. However, we do better to treat the harm and responsibility questions separately.[18] After all, it is one thing to claim that (some of) the negative outcomes attributed to hate speech are properly harmful, rather than merely adverse; it is another thing to claim that among the numerous intersecting forms of conduct and patterns of social organisation that give rise to those outcomes, hate speech is the activity (or one of the activities) whose proponents should ultimately be made answerable for this. With regards to the harm question, then, our aims are to characterise the adverse outcomes associated with hate speech, and to distinguish the genuinely harmful outcomes from sub-harmful negative experiences, such as mere offence, annoyance, or disappointed preferences. With regards to the responsibility question, our aim is to determine whether, and under what circumstances, the contributions hate speech makes to its associated harms are sufficient, both in their extent and character, to render hate-speakers legally answerable for those harms, and concomitantly liable to punishment.[19]

ଓ

My emphasis on the priority of harm and responsibility here reflects a particular stance on 'the distributional problem' of law, i.e. the question of *who* may be made liable to

[18] The only evaluative analysis of anti-hate speech law that I know of that stresses responsibility-ascription is from Evan Simpson (2006), building on ideas that emerge in Raz's (1994a) work on free speech and identity.

[19] The suggestion here, to be clear, is that for a legal penalty to be justifiably imposed on the hate speaker, A, he must be responsible for harm in two senses: (i) it must be the case that his act of hate speaking is actually causally responsible for the harm that is attributed to it, and in virtue of which it is proscribed, and (ii) it must be the case that A culpably performed the act of hate speech and thus breached that proscription. Although both dimensions of responsibility are important, my discussion mostly focuses on the former.

punishment in our legal institutions. As H.L.A. Hart (1959) famously argued, this is a question that we can and should distinguish from inquiries concerning the general justifying aims of punitive legal institutions, e.g. whether they should be primarily retributive (i.e. according just deserts to those who engage in harmful wrongdoing, see von Hirsch 1985) or consequentialist (i.e. minimising the incidence of harmful wrongdoing, see Gibbs 1975). I will leave those questions unanswered here. My aim is to determine whether hate speech should be liable to punishment *as* harmful conduct, rather than to give an account of the normative framework under which punishment for harmful conduct in general is morally defensible. It is worth observing, however, that for several different candidate theses about the general justifying aim of law, their plausibility in relation to anti-hate speech law seems entirely dependent on it having already been established that hate speech *is* responsible for harming people. Anti-discrimination law is sometimes defended by reference to the aim of making restitution for systemic, identity-based injustices that have occurred in the past and continue into the present (see Fredman 2002: 20ff.). Can anti-hate speech law be defended via the same route? Perhaps, but the credibility of this defence ultimately hinges on whether hate speech is responsible for inflicting harms that cumulatively produce identity-based injustice. Similarly, those who cite *civic* goods (e.g. the fostering of harmony and social unity, see Kymlicka 1996) among the general justifying aims of law, can only plausibly invoke that rationale in relation to anti-hate speech law if hate speech actually does effect harmful outcomes which imperil civic goods. To take one more example: an *expressivist* view – on which legal restrictions are partly justified by their role in publicly commending certain values, ideals, or commitments – has sometimes been cited to justify anti-hate speech laws (Banton 1992; Mahoney 1995; Sunstein 1996; Post 2009). The problem in this case is that there is considerable ambiguity in what it is exactly that anti-hate speech laws are supposed to commend,[20] and so unless we are already confident that hate-speakers are responsible for harming their targets, it seems incautious to make them liable to legal punishment, and thereby 'stamp them in some degree or dimension as

[20] Anti-hate speech law may commend non-prejudicial attitudes in general, or the avoidance of prejudice and hostility in speech specifically (Sunstein 1996); it may purport to inform people that others oppose identity prejudice (Banton 1992: 353); it may purport to stabilise support for anti-prejudice norms to which people already subscribe (Post 2009: 129); or it may have a reflexive purpose, of commending and instilling positive regard for abiding with restrictions on hate speech (Gardner 2007b).

outlaws',[21] just because their speech contravenes norms related to identity-prejudice and respect.

## 1.5 Inequality, oppression, and harm

I have argued that it is a mistake to appeal to egalitarian concerns to try to justify anti-hate speech law. If hate speech *is* legitimately liable to restriction, this is because hate speech is responsible for harming its targets. Nevertheless, as I said in §1.3, it is reasonable to suppose that many of the harms associated with hate speech will in certain respects be causally 'entangled' with identity-based inequality. When the hate-speaker, A, expresses contempt for others on the basis of their ethnicity, say, it is a mistake to think of him as merely voicing an opinion. Why? Because the attitudes conveyed in his hate speech are, we may suppose, partly cultivated by the identity-prejudicial elements of A's social milieu. In my analysis of harm and responsibility I certainly don't want to erase or obscure individual responsibility: above all I am interested in whether A's speech act harms B. But judgements about how responsibility operates in this area should be guided by an apprehension of the deep, identity-based social inequalities that vest power into hate speech, and which are – so some authors say, at least – perpetuated and reinforced by it in turn. To understand harm and responsibility in hate speech, we must understand the identity-oppression to which it is related.[22]

On the other hand, we should resist any attempt to efface the distinction between hate speech and identity-oppression at large. Social subordination obviously consists in much more than just being the target of verbal hostility. A conflation of hate speech and oppression *per se* is reminiscent of a certain type of 'identity-politics' – one which insists on respect and recognition as crucial elements of a just political order, but which fails to distinguish between different modes of disrespect and non-recognition, and thus portrays all hostility towards a

[21] I borrow the expression from Dworkin (1986: 1).

[22] We might be tempted to say there is an *ideological* affinity between hate speech and identity oppression, but the term 'ideology' is potentially misleading here, being ambiguous between a descriptive sense, used to refer to something like a 'worldview', and a more critical and pejorative sense, used to describe a kind of false consciousness engendered by repressive power structures (see Geuss 1981: 9ff.).

certain group as an injustice of the highest order.[23] Against this kind of identity-politics, Iris Young (1997) argues that identity-oriented approaches to normative political theory can – and frequently do – adhere to much the same individualistic conception of justice as liberal theory at large. What is distinctive in identity-politics, according to Young, is a focus on the problems that difference-blind institutions face in the pursuit of justice, thus understood, given the way that identity markers pervasively impact upon people's welfare and prospects. Where my discussion in this study draws connections between hate speech and identity-oppression, this is the concern by which I am motivated. We should be mindful of the relation between hate speech and identity-oppression, not out of a parochial interest in the welfare of any particular group, but simply because identity-oppression is itself a systemically harmful form of social arrangement, and we need to see if and how hate speech interacts with it.

There is some common ground between this approach and a liberal multiculturalist conception of justice, of the kind that emerged from liberal-communitarian debates of the 1980s, and according to which cultural identity is an instrumentally valuable thing, insofar as people's autonomy partly depends on them having an identity-community in whose formative environs they can acquire and revise a conception of good (see Kymlicka 1995: 82-83). On this view, as Joseph Raz says, we care about specific cultures and communities because of the goods they make possible for individuals (1994b: 178). I would reject the suggestion made by a handful of authors (e.g. Arkes 1988; Massey 1992), however, that anti-hate speech law as a whole is justifiable via a liberal multiculturalist rationale. It is one thing to enjoin positive aid (e.g. subsidised cultural resources for immigrant communities), and another to defend restrictive measures aimed at combating hostility towards social groups. There are doubts, moreover, about liberal multiculturalism's institutional feasibility (Taylor 1992b; Walzer 1992) and its conceptual coherence (Waldron 1991; Fish 1997; Marcus 1998; O'Neill 1999). Nothing is gained by saddling the case for anti-hate speech law with these problems.

[23] A related manoeuvre, criticised by Waldron (1991), is to esteem 'belonging' as a key element of political justice, without specifying the level at which belonging matters (family? village? state?), thus setting up a nebulous category through which any parochial demands may be represented, on the political stage, as demands of justice.

## 1.6 Religion, ethnicity, and sexuality

Evaluative analyses of anti-hate speech law sometimes veer into discussion of other forms of controversial speech, such as incitement to terrorism,[24] radical sedition,[25] blasphemy,[26] and Holocaust denial.[27] Hate speech can surely be examined separately from these other forms of speech, and so far as there is a different case to be made for restriction in relation to each kind of speech, it seems prudent to confine my attention to hate speech for the purposes of this study.[28] One type of controversial speech which is not as easily separable from hate speech, however, is religious critique. Two well-known modern cases in which the distinction between hate speech and religious critique has come under pressure are the *Satanic Verses* controversy,[29] and the Danish cartoons scandal.[30] The profound offence taken by some Muslims in these affairs, along with confusion about the purpose of anti-hate speech law, may make it tempting for the casual observer to read 'religious hate speech' as a euphemism for speech that offends religious believers, and to perceive the purpose of anti-hate speech law as preventing such offence.[31] There would obviously be little to say for anti-hate speech law if this were the case, and this is true whether we are (i) religious believers whose faith commends proselytic criticism of other faiths (McNamara 2007), or (ii) Millian liberals who see religious critique as

---

[24] See Barendt 2009; Choudhury 2009.

[25] See Leets and Bowers 1999.

[26] See Brown 2008; Hare 2009a.

[27] See Stein 1986; Douglas 1998; D. Fraser 2009; Grimm 2009b; Weil 2009; Whine 2009; Altman 2012.

[28] Kenneth Lasson says that Holocaust denial "is a particularly pernicious form of hate speech" (1997: 86). It is unclear, though, why he makes this identification. Surely we can just say that Holocaust denial is hate speech when it is used to express contempt for Jewish people. Arguments for restricting Holocaust denial needn't hinge on whether it can be classified as hate speech.

[29] See Parekh 1990a, 1990b; Modood 1993; Slaughter 1993; Waldron 1993b; O'Neill 1999.

[30] See Fish 2006; Price 2006; McNaron 2008; Cram 2009; Klausen 2009; Reichman 2009.

[31] This view is encouraged, for example, by European Court of Human Rights rulings which explain restrictions on anti-religious hate speech as attempts to uphold the targets' religious freedom (see Leigh 2011). However, those who suspect anti-hate speech law of being a tool for religious groups to shirk criticism should note that religious speech is itself targeted by such laws, insofar as they restrict speech that expresses contempt for people based on their sexual preference (see Evans 2009; Leigh 2009).

an integral element in a vital public sphere, or (iii) people for whom a critical response to religion represents 'an outlook in its own right' (see Taylor 1989: 119). In any case, misunderstandings about the character and purpose of anti-hate speech law can be cleared-up; the larger problem is how we distinguish religious critique and anti-religious hate speech in the first place. Those who support anti-hate speech law will accept that we must have "freedom to advocate and debate ideas". Their concern is that we must not attack "the rights of others or their status as... members of the community" (Heyman 2008: 182). The worry, however, is that this dichotomy is untenable. Anti-hate speech law may not *purport* to restrict religious critique, but for some "it is hard to see how a line can be drawn between... hate speech targeted at a group" and the "abuse of religious beliefs and practices" (Barendt 2011: 49). Why? Because for certain kinds of religious believers, our identities are inextricably bound up in our belief-systems, such that the individual is necessarily assailed in the vilification of her belief-system. As the novelist Philip Pullman says, "to criticise the religion of someone who makes that religion the primary marker of their identity will be, specifically, to criticize them" (2005: 110). This idea is implicit, as well, in Peter Jones's suggestion that since respecting people involves respecting them *qua* believers, "we have a reason, *ceteris paribus*, not to subject [people's] most cherished beliefs to vilification and ridicule" (2011: 86-87).

The conception of identity embodied in these sorts of claims descends in part from the work of communitarians like Alasdair MacIntyre (1981) and Michael Sandel (1982) who – against Rawls (1971) and others who say we must be free to select and revise any and all of our views about the good life – argue that people's religious commitments are *constitutive* of their conceptions of the good. On one hand, it seems reasonable to grant that people do not select their ends from an un-situated position of choice, and that our choices may be presided over by our constitutive ends. It is quite another thing, though, to take seriously the notion that people's worldviews ought to be immune from criticism in light of this thesis about how people's worldviews originate. This is just the kind of obdurate, divisive identity-politics that Young (and many others) oppose. In a pluralistic, multi-faith society, we can only escape all criticism of our deeply-held convictions by forcefully quashing the worldviews of others, or by removing ourselves from the company of others altogether. To regard any and every critique of one's religion as an attack on oneself requires a blindness to the diverse character of the societies we inhabit, and to what it would take to evade such critiques.

It is a well-worn rejoinder to these kinds of liberal precepts that they enforce a partisan dogma of their own, one which is hostile to the religious worldviews they seek to domesticate. It would be too large a detour to address this worry here, except to say that such complaints do not attach in any distinctive way to our distinction between religious critique and religious hate speech. Someone who insists that universal esteem be given to her religious beliefs is not just cavilling over a particular legislative programme; she is disputing the whole liberal view that rejects authoritatively mandated conceptions of the good.[32] In this study I assume the cogency of that aspect of liberalism, and therefore, that the most we can aim to secure is esteem for all *persons*, rather than esteem for all *beliefs*. In this we have the germ of a distinction between religious critique (i.e. the derision of commitments, practices, and beliefs) and religious hate speech (i.e. the derision of the people behind these things).

It remains to be shown how this distinction will be illuminating in practice, especially given that religious critique and hate speech are often intermingled. From the outset, we will just have to reconcile ourselves to the fact that we cannot distinguish hate speech and religious critique in a way that will neatly and uncontroversially classify every utterance as being either one or the other. Anti-religious tirades motivated by bigotry may articulate cogent criticisms of religious belief and practice (Blackford 2012: 189). Good-faith critiques of religion may shade into derision of religious believers (Asad 2009: 55-56), and even where the critic expressly condemns identity-prejudice, a persuasive religious critique may still subtly contribute to an identity-prejudicial social environment (Post 2007: 82-84).[33] However, we cannot allow religious critique to be classified as hate speech just because of how others behave in response

---

[32] One might wonder why exactly we should even locate religious adherence among the categories of identity addressed in anti-hate speech law. Why not tell religious adherents to simply opt out their beliefs if the attacks are too much to bear? The evident problem with this voluntarist 'solution' is that it ignores the social and psychological character of religious commitment. The question for people of religious faith, as Gardner (1998: 172) says, is not whether their faith can change, nor whether it is a product of past choices, but whether they can change their faith now merely by choice. Religious commitments are an important and enduring part of the lives of people who hold them, and this makes religion an appropriate marker of identity for the purposes of anti-hate speech law.

[33] In some such cases – namely, "when the acts of... some small number of members of a group are taken to be the defining actions and beliefs of the group itself" (Butler 2009: 134) – a critique may devolve into mere racism.

to it. In practice, distinguishing hate speech from religious critique means looking to the words that are used and to the inferrable attitudes of the speaker, not to the innumerable ramifying consequences of every utterance.[34] Given that caveat, the distinction that we're trying to pin down is surely not as elusive as some have suggested. There are clear enough differences between a condemnation in epistemic terms (idiot, fool, imbecile, ignoramus), and one that attributes evil or sub-human characteristics to the target (vermin, scum, ape, pig, dog, devil). Similarly, there are clear enough differences between contexts that invite dialogic communicative aims (debates, scholarly arguments, editorial writings, artistic spaces), and contexts that tend to function as a vehicle for slander and abuse (propaganda leaflets, vandalism, face-to-face confrontations, and online 'trolling'). It may be difficult to draw bright lines across the various hazy penumbras here, as it often is, but the underlying contours of the sought-after classificatory schema are not especially mysterious or opaque.

Another difficulty with the way identity figures in discussions of anti-hate speech law – one that is less commonly discussed[35] – is what we might call 'the reification problem'. One condition for the perpetuation of identity-based social hierarchy is a widespread acceptance that the relevant identity-categories are in some sense real or genuine. The simplest example here is 'race'. The logic of a race-based hierarchy will be profoundly undermined if people no longer accept that there are significant biological differences between individuals which track the folk concept of 'race'. The problem for us is that in discussing the evils of racism, we can carry out our discussion as if 'race' *is* a genuine marker of identity and difference, and thus in our efforts to combat identity-based oppression, accidentally bolster one of its pre-

---

[34] This is denied, to bizarre effect, in Monica Mookherjee's (2007) argument for the regulation of speech which is critical of the injustices carried out in minority religious communities. Mookherjee begins by contending, plausibly enough, that this sort of criticism can erode the ethos of tolerance that is necessary for other members of minority communities to speak out effectively about their experiences of injustice in those communities. But Mookherjee then wants to say that the Sikh woman who speaks out publicly about coercive gender practices in her community is partly responsible, if it subsequently transpires that male members of the Sikh community become even more domineering in an attempt to stop other women speaking up. Mookherjee's argument for the regulation of religious critique lays responsibility for these sorts of injustices in religious communities at entirely the wrong feet.

[35] One notable exception is in Judith Butler's *Excitable Speech* (1997).

conditions.[36] How else might we approach things? Sally Haslanger (2000) argues that we should only use the social-identity concepts that advance legitimate political aims (such as overcoming oppression) for the people to whom the concepts apply.[37] Hence, so one might argue, we should try to find a way to talk about hate speech and its restriction (or permission), which does not implicitly accept the racist hate-speaker's racial categorisations, and thereby tacitly support one part of the view that he seeks to promulgate. As I said, 'race' is a relatively easy case, with its feeble foundations and shameful history. Instead of 'race' concepts we may invoke 'ethnicity' concepts, which at least sometimes track genuine differences between persons with respect to things like language, culture, and custom. In a parallel way, we can eschew any rigid taxonomy that contrasts heterosexual normality with homosexual pathology, and use words like 'queer', 'gay', and 'lesbian' to refer to differences between persons with respect to choice of sexual or romantic partners. But there are still problems. Firstly, ethnicity concepts do not seem sufficient to diagnose many of the behaviours and phenomena that are usually described, in informal discourse, as 'racist'. In Western societies people who are non-Caucasian, but who do not identify with any culturally- or linguistically-based ethnic identity, can of course still experience discrimination or prejudice based on a spurious folk conception of race. We might not want to describe hate speech in this vein as 'racially-based', but it is surely a misnomer to describe it as ethnicity-based. Secondly, even where social-identity

---

[36] This phenomenon is insightfully addressed in the work of Kwame Anthony Appiah (see for instance 1989: 55ff; 2005: 189-92).

[37] For instance, Haslanger (2003, 2005) suggests that the concept 'woman' should be understood as 'someone who is systematically subordinated due to her female sex'. Haslanger's claim is not that this is how we understand the concept 'woman' currently (it is not our *operative* concept), nor that this is the view of the concept we would have after careful reflective equilibrium (it is not our *manifest* concept), rather, she argues, it is our *target* concept: it is the understanding of the concept 'woman' that best serves our legitimate anti-oppression purposes in deploying that concept. Against this suggestion, Jennifer Saul (2006a) argues that if our social role concepts are negotiable, it may be a bad idea to embed claims about subordination into those concepts, since the new vocabulary we arrive at, when deployed in the untidy cut and thrust of political activity, may be just as likely to entrench subordination as to destabilise it. In more recent work on the meaning of social role words, however, Saul (2012) gravitates towards Haslanger's approach, insofar as she accepts that the ongoing process of revising and negotiating the meanings of words (like 'woman', or 'marriage') can and should be influenced by our legitimate (e.g. anti-oppression) political purposes.

categories based on sexuality or ethnicity *do* suffice to describe different forms of hate speech, our use of them seems to betray something of the same pernicious taxonomising instinct to which Haslanger objects. In discussing hate speech, why should we continue verbally classifying human beings into identity-based tribes, given that such classifications are an element of mindset that underpins hate speech, and which is (arguably) perpetuated by it in turn?

Such concerns may encourage reforms in how we talk about human difference and identity, and it is possible within this that we would seek to eschew ethnic and sexual categories entirely. However, it is compatible with this prospect, and with Haslanger's critical approach to identity concepts, to continue employing these categories in our discussion of hate speech and its legal restriction. This is not a case of *allowing* hate speech to set the agenda for how identity is understood and verbally encoded, so much as a case of acknowledging that at present, for people maligned in hate speech, hate speech *already* sets that agenda. In hate speech it is a real and important marker of human difference for someone to be black, Arab, lesbian, or queer, and consequently, wherever hate speech circulates, these things may well become important markers of difference. As Lynne Tirrell (1998: 309) says, words "have the power to make something *seem* real, and sometimes seeming is the first step toward *being*". If we want to describe this process in hate speech and use the law to redress its harms, we will inevitably attend to the hate-speaker's perspective: the fact that *she* imputes negative traits to people identified, in *her* eyes, by their ethnicity or sexuality. We do not become normatively invested in the hate-speaker's social taxonomy, though, just because we are trying to redress harm done by her speech. We can always see it as a further question, which identity categories would serve our legitimate purposes in a future that is less oppressive than our present.

# ☙ 2 ❧

# Free speech and speech acts

## 2.1 Free speech arguments

Suppose we have reached the conclusion that hate speech *is* responsible for harming its targets. Would a provisional case for restricting hate speech, based on this finding, be overridden by the fact that what stands to be restricted is a form of speech? It is a banal observation, but one worth repeating, that nowhere is speech absolutely free. No-one thinks criminal conspiracy, perjury, extortionate threats of violence, or even nuisances should be beyond the law's reach, just because such acts can be verbally constituted. Legal restrictions on false advertising and child pornography are widely accepted as proper and necessary, even though most of what they restrict is speech, in the sense of 'speech' that applies in free speech discourse (which uncontroversially encompasses images, audiovisual material, electronic communication, and other non-enunciatory modes of expression). Every social and private institution imposes some limits on what its members, participants, and clients can and cannot say, and in a broad range of cases the law (criminal, administrative, contract, tort, and others) stands behind these restrictions. There is no easy shortcut, therefore, from hate speech being 'speech' to it being immune to restriction. First we need to discern whether 'hate speech' is the sort of thing that free speech principles are meant to protect; and if it is, we need to see whether the protections extended to it are, or should be, so stringent as to immunise hate speech against legal restriction, irrespective of the harm it may be found to inflict.

So why should speech be free? And should everything that counts as 'speech' receive protection under the auspices of 'free speech' in the same way, and to the same degree? I will confine my discussion to *restrictions* on speech, thus setting aside questions about whether the state has a duty to refrain from making positive contributions to public discussion (e.g. via public information campaigns, or funding for the arts, or scholarly research, see Post 1996).

And since our interest is in *legal* restrictions on hate speech, I will also confine my discussion to free speech as a constraint on what law may restrict – i.e. a putative principle of legislation, as opposed to a wider obligation we may have, *qua* private individuals, to not excessively inhibit or interfere with one another's expression (see Kagan 1992).

Some arguments for free speech have a deontological shape. They identify features of human individuals related to our communicative capacities, e.g. our rationality, or reflectivity, or our capacity for critical thinking and independent moral judgement, and they argue that given these characteristics, a respect for persons gives rise to negative duties for speech to not be silenced, nor coercively deterred, nor significantly interfered with via other means. These arguments suggest that legally enshrined rights to free speech are somehow founded on rights grounded in human capacities. Some arguments focus on our capacity simply to disclose our ideas and attitudes to others. Among sentient creatures, humans alone have the ability to verbally articulate our inner lives, and government, so the argument goes, must allow space for this ability express itself (see Redish 1982: 625; Peonidis 1998: 3-5; Baker 2009: 142ff.).[1] Other arguments focus on how speech is specially implicated in our relationships and our capacity to create and participate in cooperative communities (see Hornsby and Langton 1998: 37). Probably the most well-known arguments in this area, however, are those that appeal to the demands generated by our ability to think for ourselves and make up our own minds about things. In his 1972 article 'A Theory of Freedom of Expression', Thomas Scanlon argues that a government which seeks to regulate public expression, for fear that false or dangerous ideas will attain a wider currency, thereby fails to treat people as intellectually responsible agents, and therefore relinquishes its claim to legitimate governance.[2] Ronald Dworkin (1996: 200)

---

[1] According to Jonathan Gilmore, this value is not just relevant to our social relations, because self-disclosure also helps us to understand our own attitudes. Our interest in expressing our thoughts, he says, "lies not just in the results that may follow from their articulation... but also in the very role such expression plays in developing and discovering the content of those thoughts" (2011: 518).

[2] In response to criticism from Robert Amdur (1980) and others, Scanlon retreated from this argument in his later work, arguing that the government's duty of non-interference in speech is based on a consequentialist interest in our intellectual independence as a good to be promoted, rather than as the basis of an obligation-conferring right (see Scanlon 1979). Fish (1994b: 14-15) argues that the Scanlon's retreat to consequentialism typifies all *prima facie* deontological reasoning in free speech arguments; duties that are posited as categorically binding in fact turn out to be justified, if at all, by the benefits their general observance would promote.

makes a similar claim: "Government insults its citizens, and denies their moral responsibility" he says, "when it decrees that they cannot be trusted to hear opinions that might persuade them to dangerous or offensive convictions" (see also Richards 1986: 166-74).

There are other arguments for free speech that are more consequentialist in character, and these come in both a positive variety, appealing to goods that are impaired by restrictions on speech, and a negative variety, appealing to evils that are caused or threatened by such restrictions. The most well-known positive arguments, traceable back to Mill, or even to John Milton's *Areopagitica* (1644), have an epistemic bent, claiming that the cause of truth and human knowledge is held back when free speech is curtailed (for modern accounts, see Chevigny 1980; Scoccia 1987; Canavan 1999; Blocher 2008; Talisse 2008). Often these arguments are cashed-out using a familiar economic analogy: free speech allows the marketplace of ideas to promote true beliefs, thus avoiding the perils of government-mandated dogma and orthodoxy (analogous to the perils of a managed economy in liberal market theory, see Ingber 1984: 6-16).[3] In addition to these epistemic rationales, there are arguments, typically associated with the work of Hans Kelsen (1945) and Alexander Meiklejohn (1948; 1961), that advert to the positive effects of free speech on democracy. Free speech, so these arguments say, imparts vital benefits to the processes of criticism, conversation, transmission of information, and the dissection of information in argument, on which the health of a democratic system depends. For democracy to flourish, in short, these

[3] Such arguments have long been criticised for imagining society as 'a debating club devoted to the pursuit of truth' (Kendall 1960: 977). For justificatory purposes, the free market analogy is as much a liability as an asset, given the comparison it invites between volatile, deregulated economic markets and *laissez-faire* conditions for speech (for a theoretical analysis, see Baker 1978: 974-81; for a more policy-oriented analysis, see the Report of the Great Britain Committee on Obscenity and Film Censorship 1979: 55ff.). Even if we accept the economic analogy, the claim that a marketplace of ideas produces truth assumes that consumers in the market want truth. As Goldman and Cox (1996) argue, though, if the consumers want anodyne entertainment more than truth, that is what the marketplace will deliver to them. Goldman has defended various anti-marketplace-of-ideas proposals, suggesting, for instance, that just as courts engage in epistemic paternalism – making decisions about which information should be available and withheld from jurors – so too might epistemically paternalistic policies be justified in other institutions, under certain circumstance, in order to promote veritistic outcomes (Goldman 1991).

forms of speech must remain unfettered.[4] One particular aspect of democratic flourishing that is sometimes emphasised, within this broader rationale, is the public presentation – in speech, custom, and symbolic expression – of different worldviews and ways of life. Raz (1994a) argues that such presentations should be protected from restriction since they are the primary vehicle through which different ways of life can be advertised to others as viable possibilities, and through which lifestyles can be propagated between generations.[5] This emphasis speaks to a wider sense of democratic flourishing connected with 'self-determination', i.e. how a people establish a way of life together over time, as opposed to the narrower and shorter-term aspects of democracy to do with election cycles and managing day-to-day policy issues.

Negative consequentialist arguments for free speech, naturally, advert to negative outcomes that are purportedly threatened by restrictions on speech. Given the inevitability of erroneous judgements in law, some speech that is restricted on account of its supposed harmfulness will in fact turn out not to be harmful. The prospect of such error is not unique to restrictions on speech, but it is identified as a distinct problem in this arena, because the wrongful silencing of speech is regarded as a disproportionate travesty of justice, similarly to the way that a mistaken conviction is seen as disproportionately worse than a mistaken acquittal (see Schauer 1982: 136-9). Errors aside, some authors argue that even moderate and judicious restrictions on speech unacceptably increase the likelihood of tyrannical repression at later times, a pessimism that is based on either (i) observation of such cases in the historical record (Weinstein and Hare 2009: 5), or (ii) an eagerness to pre-empt the worst manifestations of human intolerance (Blasi 1985: 449-50), or (iii) a judgement that the legitimate reasons for protecting speech will seem remote or counterintuitive to those involved in the administration of the law: prosecutors, judges, and jurors (Schauer 1985: 376-77). Finally, even if we set aside the prospects of erroneous and tyrannical silencing, the restriction of speech is commonly

[4] There is a connection between the Scanlonian legitimacy argument and the democracy-based argument for free speech, for if the state's legitimacy depends not just on its respect for people's autonomy *per se*, but its respect for people's capacity to govern themselves via democratic procedures specifically (see J. Cohen 2006), then the state's legitimacy hinges on its instituting the kind of flourishing democratic system to which, so the democracy argument says, restrictions on free speech do particular damage.

[5] Moreover, as Leslie Green (1991) suggests, this sort of rationale may in certain cases entail that free speech principles disallow not only restrictions on speech *per se*, but also institutions that restrict people's choice of language in the public sphere.

alleged to exert an invidious 'chilling effect' on the communicative environment, such that people with good reason to speak, and who face no legal barrier to their doing so, are nevertheless deterred from speaking, for fear of running afoul of the established legal restrictions on speech, whose precise contents and purpose the speakers do not know (see Schauer 1978).

## 2.2 Protected species and the meaning of 'speech'

All these arguments may be taken to underwrite an individualised political right to free speech. With respect to the consequentialist arguments, this right can be seen as an institutional tool for promoting the goods advanced by speech, or for forestalling the evils begotten by its silencing (see Raz 1994a). As for the deontological arguments, the natural rights and duties they postulate can just be translated into the political arena. But how would an individualised, political right to free speech function as a constraint on what the law restricts, if respect for the right itself is not enforced by a prohibition on its infringement as such? The right to free speech exerts its influence, in theory, as a consideration that 'channels' the kinds of reasons and arbitrates among the kinds of justifications that governments may act upon in legislative undertakings that have a restrictive or otherwise adverse effect upon speech (Pildes 1998: 761). To be clear, then, infringements against 'free speech' are *acts of government*. They are acts carried out via the implementation of policies backed by state power (or via the immediate deployment of state power), and which are *pro tanto* wrongful, insofar as they ignore or intrude upon people's rightful sphere of expressive liberty, warranted on the sort of deontological or consequentialist grounds cited above. Whatever rights-respecting duty of non-interference the state has in relation to speech will be translated into the legal domain, where it matters for our purposes, in a manner such that certain kinds of legislative restrictions on speech are rendered illegitimate or invalid *qua* legislation (Lyons 1970: 49-54).

The question remains, though, of what should qualify as 'speech' for the purposes of a legislative principle of free speech, like the one I have sketched. Schematically, the principle may be stated as follows: if a law, L, would function (by design or in practice) to restrict conduct that belongs to the class 'speech', then L infringes against the right to free speech, and must meet a more stringent standard of justification than a law which does not infringe

that right. But the class 'speech' cannot plausibly be identified with the everyday concept of 'speech'. The latter includes many trivialities – idle blather, pleasantries exchanged with the shop assistant, utterances aloud to oneself – that bear no evident relationship to the values that animate the arguments sketched above. The sense of 'speech' that is invoked in those arguments (and in the rights and principles they underwrite) is both narrower *and* wider than the everyday sense: narrower because it bears no evident connection to trivial verbal actions, and wider because it does bear some connection to certain non-verbal behaviours, such as the creation and exhibition of visual art, the display of symbols for expressive purposes, and the broadcasting of images, music, etc. We cannot adduce arguments dealing in this technical conception of 'speech' to defend protections for all behaviours which belong to the everyday conception of speech. Indeed, all we get 'for free', from the term 'speech' in this discussion, is a general indication that our concern is with behaviour that is expressive or communicative in some sense. Within that loose bound, what counts as 'speech' for the purposes of 'free speech' will simply be the class of expressive or communicative behaviours whose protection would be justified by the best arguments in favour of protection for some class of expressive or communicative behaviour (Braddon-Mitchell and West 2004). 'Speech' will thus be a technical term that is used to refer to the entire "heterogeneous set of non-violent, non-coercive expressive activities" around which these arguments revolve (Baker 1989: 47).

ଔ

But again, we may ask, what kinds of behaviours *do* the arguments revolve around? If we treat the standard free speech arguments as a unified and mutually complementary suite of justifications, they seem to locate the value of speech, primarily, within the individualistic concerns of classical liberal thought. Firstly, these arguments ascribe significant worth to morally and intellectually independent individuals expressing and publicising their personal views, whatever they may be. Secondly, rather than ascribing any overriding worth to group associations, these arguments see the value in social relations as contingent on them being entered into by informed, voluntary, and cooperative individuals, and serving the purposes of those individuals in turn. Thirdly, these arguments suggest that groups imposing their view upon others, and thereby repressing individuality, pose an especially grave threat to human flourishing. Deontological free speech arguments articulate these kinds of judgements by

mapping out a relationship between the capacities of human individuals and the expressive rights that are owed to individuals in light of those capacities. Positive consequentialist arguments advert to various goods that underpin healthy, individual-respecting social relations – democratic integrity, the dissemination of truth and knowledge, the expression of different ways of life – and identify free speech as an essential tool for promoting such goods. Negative consequentialist arguments advert to bad outcomes that arise in degraded social relations – ignorance, fear, repression, coercive orthodoxy, hollow dogmatism – and identify free speech as a bulwark holding these things at bay.

Given this understanding of how arguments for free speech hang together, references to the special value of 'speech' *per se* are somewhat misleading. The goods and values that animate these arguments are evidently vested to a greater extent in certain *species* of expression or communication than in others. These values seem to be particularly vested, for instance, in: (i) expression that contributes to a debate or dialogue about what is good and desire-worthy in life – literature, art, philosophy, religious proselytisation, and the display of symbols that convey people's views in these respects; (ii) expression that contributes to the spread of information and knowledge – journalism, scientific research, educational speech, and public information campaigns; and (iii) expression that questions or combats the development of oppressive or hegemonic social relations – criticism, protest, satire, iconoclastic art, sedition, and dissent. One might specify other forms of expression besides these. Schematically, the suggestion is that some species of expressive and communicative behaviour require special protection, since they specially embody the goods and values adverted to in the justificatory underpinnings of free speech.[6]

---

[6] One might ask whether, so far as these typical free speech arguments have an individualistic bent, they are vulnerable to the familiar communitarian charge of assuming an implausibly atomistic social ontology, or in a similar vein, of excessively prioritising individual liberties over obligations of civic unity or belonging. Charles Taylor (1985a: 198) argues that if we can only develop moral consciences within a political society, then freedom of conscience cannot trump our obligations to society; "in the event of conflict" he says, "we should have to acknowledge that we [are] legitimately pulled both ways… for in undermining such a society we should be making the activity defended by the rights assertion impossible of realization". One way to defend individualistic free speech theory from charges of pernicious individualism would be to point out that the 'impossibility of realisation problem' cuts both ways: we cannot attain moral independence if the civic bonds that make political society possible fall apart, but nor can we attain it if we relinquish robust freedom

What about the expressive behaviours that do not belong on that protected species list? Although the conception of speech's value that underwrites free speech arguments does not ascribe a special status to all and every species of expressive or communicative conduct, it does suggest the need for a constraint upon the type of rationale that can be used to justify restrictions on any conduct within this genus, irrespective of the particular species that stands to be restricted. Given the high price free speech arguments place on individuality, and the dangers they associate with official dogma, social hegemony, and tyrannies of the majority, they suggest that expression of whatever kind should not be subject to restriction simply on account of the content it conveys. Ideas and opinions that are false, outré, taboo, out of favour, or simply absurd, cannot be restricted for those reasons.

The picture I am sketching is one on which we have a unified 'principle of free speech', grounded in multiple sources of value (see Perry 1983). One may, instead, see 'free speech' as an umbrella term for an array of norms, principles, and ideals, which are all related in some way to values residing in communicative and expressive conduct, and which largely agree with one another, even though they occasionally pull in different directions (see Schauer 1982; Alexander and Horton 1984). In defence of the former view of things, one might argue that the singular importance of free speech as a principle arises precisely *because* it governs an area of human behaviour that implicates a multifarious complex of values (see Schauer 1984: 1303-05). In any case, with the referent of 'speech' thus elaborated, 'free speech' as a legislative principle can be spelled-out as a two-tiered set of constraints. The more stringent tier specifies various protected species of expressive behaviour, and requires laws which would prohibit one or more of these protected species to answer to a more demanding justificatory standard. We may restrict religious proselytisation, or iconoclastic art, or the dissemination of scientific research (for instance), only if the restriction of such expression is strictly necessary for some legitimate governmental aim (e.g. preventing harm). Why the high threshold? Because these forms of speech specially embody the individualistic values adverted to in the

---

of conscience from the outset. If the conditions under which such liberties become widely possible are therefore highly elusive, this in turn seems to provide all the more reason to guard those liberties jealously where they have been formally enshrined.

justificatory underpinnings of free speech.[7] The less stringent tier applies to all expressive behaviours, i.e. all behaviours that convey some kind of content: ideas, opinions, propositions, etc., and it stipulates that wherever such behaviours stand to be restricted, the content they convey *per se* cannot be invoked as the justificatory basis for their restriction. Why not? Because the individualistic values in the underpinnings of free speech are jeopardised whenever official dogmas are allowed dictate what people say.

## 2.3 The First Amendment, free speech, and hate speech

The two elements of a legislative free speech principle that I identify above are operative, but instituted via a notably different structure, in our most sophisticated operative theory of free speech as a legislative principle, namely, the jurisprudential framework around the First Amendment to the US Constitution.[8] Where I say the values behind free speech should lead us to identify protected *species* of expression, First Amendment jurisprudence identifies a protected *domain* of expression, namely, 'public discourse', which it defines in terms of a certain subject matter, i.e. "matters of public concern", being addressed in certain contexts, i.e. "settings dedicated or essential to public discussion such as books, magazines, films, the Internet, or in public fora such as the speaker's corner of the park" (Weinstein 2009b: 83).

---

[7] In contrast to this 'protected species' approach, some authors begin by classifying all expression as 'speech' initially, before trying to identify a criterion or test that excludes specific forms of expression from the initially-protected classification. Greenawalt excludes utterances which "actually change the social world in which we live" from the protected domain (1989a: 58); Maitra and McGowan (2007; 2010) exclude 'significantly obligation-enacting utterances'; Solum (1989) invokes Habermas's distinction (1984: 295ff.) between *communicative* action (pursuit of illocutionary aims alone) and *strategic* action (pursuit of perlocutionary aims), classifying only the former as 'speech' for the purposes of 'free speech'. Such approaches assume that there are general features which explain our judgements about which forms of expressive conduct need special protection as 'speech'. Even if that is right, though, it's not clear what we gain from contriving criterial tests to sort the cases, when we can just ostensively specify the sorting of cases we have reason to favour (and against which we would be calibrating our criteria in any case).

[8] I say the framework *surrounding* the First Amendment is our most sophisticated theory; the text of the free speech clause itself is simple and terse: "Congress shall make no law... abridging the freedom of speech, or of the press...".

Within this protected realm, the reasons that may be invoked to justify restrictions on expression are confined to those concerning the time, place, or manner of the expression at issue. Restrictions on public protests *may* be permitted if their purpose is to prevent disruptions to business, leisure, and other goings-on. But such restrictions are strenuously opposed if their purpose is to prevent certain issues being subject to public discussion, and all the more so if they violate 'viewpoint-neutrality', i.e. if their aim is to suppress the point of view being espoused in the demonstration. Such restrictions are subject to a judicial review standard known as 'strict scrutiny' which requires, on pain of unconstitutionality, that the restriction in question not only aims narrowly at a legitimate purpose (e.g. public order), and is demonstrably likely to be effective to this end, but also that the government cannot effectively attain or promote the relevant purpose *via any other means*. Restrictions on public discourse are not flatly prohibited, but the standards of justification they must meet, strictly interpreted, will almost never be met. Expression that does not belong to the specially esteemed domain of public discourse receives far less stringent protection. Indeed, First Amendment jurisprudence specifies various forms of expression as being content-based *exceptions* to its neutrality doctrines, including incitement to criminal conduct, true threats, commercial advertising, child pornography, and face-to-face insults or provocation (i.e. 'fighting words') (see Weinstein 2009b: 82-83). Expression of these kinds may in practice be restricted much the same as any other behaviour, as far as First Amendment requirements are concerned.

There is a dismaying irony in the way this framework functions in US Constitutional theory and practice. First Amendment doctrine, as James Weinstein rightly observes, "is characterised by hard-edged, determinative rules", in contrast to other jurisdictions, which "tend to employ soft, flexible standards that balance the free speech interest... against the government's interest in suppressing the speech" (2009b: 90). Weinstein grants that there are likely to be positive and negative aspects to both kinds of approaches. But the upside of the US's hard-edged approach, he says, is that it does not "allow law enforcement officials and judges to smuggle their own ideology into the analysis" (2009b: 90). This is a very sanguine view. Given the inflexibility of the US approach, the categorisation of a given form of expression – whether it is elevated to the hallowed realm of public discourse, or consigned to one of the lowly categories of expression exempted from special protection – will largely seal the fate of any legislation under which that expression might be restricted. Standards of strict

scrutiny are nearly un-satisfiable, so that even legitimately-purposed restrictions on expression are bound to fail before them, while exempted categories are ostensibly *un*protected, so that legitimately-purposed restrictions on such expression face no real First Amendment-based hurdles. Consequently, when lawmakers ask, for instance, whether corporate-funded election propaganda should be deemed 'public discourse' or 'commercial expression', they are asking, by proxy, whether such expression ought to be seen as a candidate for legal restriction (e.g. on accounts of its corrosive effects on democracy), or whether it should, instead, be treated as off-limits to government interference for all practical purposes. The judge's options are inflexibly spelled-out once the classificatory process is settled. But that simply means that the influence of ideology, partisan bias, and *realpolitik* will enter at an earlier stage of the diagnostic procedure, i.e. in decisions about which behaviours belong to which categories of expression, and thus where the stringent demands of content- and viewpoint-neutrality will be imposed upon expression. All of this is old news.[9] It is pertinent in this context, nevertheless, because it

[9] See for instance Schauer (1989, 2004) and Fish (1994d). For discussion of how the political battles around this 'classification problem' have been waged across party lines in US politics, see Hentoff (1992). Fish's critique extends past the point about the politicised classification of speech for legal purposes, since he ultimately denies that there is any free speech *principle*; for Fish, any such 'principle' inevitably (like all principles, on his view) is a mere pretext invoked for its utility in furthering the partisan values that bestow its content. A similarly relativistic mood is evinced in Judith Butler's writings about censorship (1998), in which she discerns (and endorses) in Foucault the thesis that because censorship is ubiquitous and inescapable (e.g. we are censored by our language, by our socio-cultural location, by our imaginative limits, etc.), there is no distinct species of wrong that is picked out by what we normally and informally call 'censorship'. Schauer also flirts with this thought when he says that to call something censorship is to offer "a conclusion masquerading as an analytic device" (1998: 160). Such views about free speech epitomise the deconstructive approaches that Miranda Fricker criticises on account of the way they, having rejected the traditional view which treats rationality in speech as something *independent* of power relations, subsequently *reduce* reason to power (1998: 160). Joshua Cohen (1993) calls such views 'free speech nihilism'. But it would be more apt, given the politics of people like Fish and Butler, to call these views 'ironic pragmatist' views of free speech, in the sense of that term associated with Richard Rorty (1989) (this seems particularly apt in Fish 1994a, for example). These approaches deny the purported impartiality of the moral concepts we invoke to criticise societies in which, for instance, blasphemy or apostasy are severely punished (see Post 1998). However, the authors largely share the convictions that motivate such criticisms. What they assert is the cultural contingency of those convictions and of the putative principles of free speech which give expression to them.

reminds us that despite the sophisticated structure of US free speech jurisprudence, its decidedly negative stance on the restriction of hate speech is not a result of its free speech policy framework alone, but of its history and politics filtered through that framework. Landmark judicial decisions, in particular *R.A.V. v. City of St. Paul*,[10] have created a situation in which restrictions on expressions of identity-based contempt are just accepted in US free speech jurisprudence as involving an illegitimate breach of viewpoint neutrality, either because such expression belongs to 'public discourse', or because it is a form of regular expression to which viewpoint neutrality still applies.

Might this be due to systemic identity-prejudice in the US legal system? That suggestion has been raised repeatedly in critical legal analysis, and I will not evaluate it afresh here.[11] Any legal framework in this area is liable to become a political battleground, and so even if the protected status of hate speech in the US *is* due to systemic identity-prejudice in the US legal system, that is not a decisive argument against the First Amendment framework *per se*. As I said above, moreover, the US system does put into effect the two components of a legislative free speech principle that should be in effect, given the individualistic values that underwrite free speech, namely: (i) the ranking of certain species of expression as especially valuable and in need of especially stringent protection, and (ii) the disqualification of a certain kind of repressive anti-individualistic rationale for any restriction on expression. Having said all that, however, I think that the way these components of a free speech principle are expressed in the formal structure of First Amendment jurisprudence encourages us to ask the wrong questions about hate speech, or to ask the right questions in the wrong way. I will come to the issues of viewpoint neutrality in §2.5. Given the axiological foundations on which free speech principles are based, we're right to ask whether a given restriction on speech is viewpoint neutral. But the First Amendment model misconstrues viewpoint neutrality, and its

[10] *R. A. V. v. City of St. Paul* [1992] 505 US 377 112 S. Ct. 2538. The case involved a minor convicted under a Minneapolis city ordinance for burning a cross on a black family's lawn. On appeal the Minnesota Supreme Court upheld the ruling and the relevant statute, which it took to apply to a particular sub-class of 'fighting words'. Subsequently, however, the statute was struck down in the US Supreme Court because, so Justice Scalia argued, it singled-out the racist viewpoint of some fighting words for restriction, and thereby constituted an unjustifiable attempt by the state to impose a form of political orthodoxy (see Kagan 1992; Matsuda and Lawrence 1993).

[11] See for instance Delgado 1993; Lawrence 1993; 1995; Delgado and Yun 1995b.

misconstrual (misleadingly) makes hate speech restrictions seem like blatant violations of this principle, which they need not be. As for how we cordon-off the especially valuable forms of expression, it is difficult to see how the category of 'public discourse', elaborated by reference to 'expression on matters of public concern', could ever be sufficiently narrow and unambiguous to serve its purpose. These are terminological banners that, depending on which way the political wind is blowing, may be used count harassment and slurs as 'public discourse', while denying that military whistle-blowing is 'expression on a matter of public concern'. Instead of asking whether hate speech is entitled to special protection as an instance of public discourse, then, I think we do better to simply ask whether we have reason to treat it as a protected species of expression, among the ranks of protest, journalism, satire, and so on.

## 2.4 Should hate speech be a protected species?

Some argue that hate speech should be a protected species because this is a necessary condition for the legitimacy of laws which promote inclusion and equality for ethnic, religious, and sexual groups. The idea is that legislation to these ends is illegitimate if there has not been the opportunity for those governed by the legislation to engage in a free and open debate about such laws and what they purport to restrict, and that hate speech is a part of free and open debate on these matters. Ronald Dworkin espouses arguments of this kind, both in relation to hate speech in general (2009), and in relation to parallel concerns about the censorship of religious critique in certain quarters, following the Danish cartoons controversy (2006). To restrict hate speech, Dworkin says, would be "to intervene too soon in the process through which collective opinion is formed" on legislation aimed at promoting equality and social inclusion, and to thereby "spoil the democratic justification we have for insisting that everyone obey these laws, even those who hate and resent them" (2009: viii).[12] Adapting Dworkin's own language, Jeremy Waldron aptly characterises this as an upstream-downstream

[12] This line of argument may derive from a broader view of free speech as the key vehicle for democratic self-determination, the idea being that if identity-based inequality and oppression are to be overcome, the impetus to do so must arise from a joint communicative effort to wrestle with the relevant questions, of which hate speech may be a part. "If the unjust cultural construction of racism and sexism is to be remedied, it must be on the terms and in the authentic voice of the moral powers of free people who understand their ultimate

rationale: "if you interfere coercively upstream", by restricting hate speech, "then you undermine political legitimacy downstream" for laws restricting the kind of identity-prejudice that hate speech allegedly promotes (Waldron 2012: 178). Waldron also has a swift and compelling rebuttal to this rationale: "the racist doesn't *need* to use the sort of vicious hate propaganda that the law punishes in order to express his opposition to laws about discrimination and so on" (2012: 183). Perhaps if there were no way for people to express objections to hate-crime laws, anti-discrimination laws, and the like, other than by expressing contempt for the identity groups (ethnic, religious, or sexual) whose maltreatment such laws proscribe, the legitimacy of those laws would be undermined, for reasons like those Dworkin states. But there are numerous other ways in which such objections can be articulated and registered in public discourse, e.g. on grounds of gratuitous expense, structural inefficiency, personal disadvantage, or even on the basis of one's opposition to abstract egalitarian ideals as such (Waldron 2012: 189). With those alternative possibilities on the table, Dworkin's argument seems to devolve into a patently excessive demand: that unless there are stringent protections for every possible means through which people might express opposition to some law, L – however inarticulately that opposition is expressed, and however damaging that expression may be in other respects – L is illegitimate. Of course we still need a sufficient reason to restrict hate speech; legislation which restricts hate speech precisely in order to quash opposition to an ambitious legislator's egalitarian initiatives would still be illegitimate. The point is that if we do have sufficient harm-based reasons for restricting hate speech, the case for anti-hate speech law grounded in those reasons is not undermined by concern for the standing of other laws that oppose manifestations of identity-prejudice.

---

and inalienable moral responsibility reasonably to define the moral meaning of their lives and struggles against a culture so traditionally hostile to such claims. Liberalism is the doctrine that requires both such strenuous moral freedoms and such demanding moral responsibilities" (Richards 1994b: 114).

Another argument for the protected status of hate speech is Lee Bollinger's argument in *The Tolerant Society* (1986), that by specially protecting blatant expressions of identity-based contempt, the state promotes a social ethos of toleration in the face of profound disagreement.[13] An appeal to toleration here evokes thoughts of Locke, and perhaps there is something of the spirit of the *Letter Concerning Toleration* in Bollinger's account. Locke argues that trying to coerce people in their beliefs is a mistake, because beliefs are not produced by the will, which is all that coercion can act upon.[14] Therefore, at least as far as the hate-speakers themselves go, tolerating intolerance may be a better way to promote a tolerant ethos than trying to legally compel a tolerant ethos. But Bollinger's argument is not about what fails to promote tolerance; it rests on a claim about what will *succeed* in this aim, and not only among hate-speakers, but in all of society. Even if anti-hate speech law is unhelpful with regards to promoting a tolerant ethos, it doesn't follow that special protection for hate speech *is* helpful. Whether that is the case depends on how the meaning of the state's inaction on hate speech is interpreted by the public, and how those interpretations influence people's own views in turn. If special protections for hate speech are interpreted (as in most critical race theory) as more evidence of the state's complicity in identity-based subordination, then the 'tolerance' enacted by those protections will not be a harbinger of wider tolerance in the community. Even the description of the policy here may affect perceptions of meaning. Where the state is taken to be 'permitting' intolerance, rather than 'tolerating' intolerance, then its stance with respect to matters of tolerance, can be described, quite naturally, as a *pro-intolerance* stance. In any case, however the 'meaning' of protections for hate speech are interpreted, there is too much speculation involved in Bollinger's account – about how such interpretations will favourably 'rub off' on people – to warrant treating hate speech as a protected species of expression. If hate speech is responsible for inflicting harm, then in order to override our *pro tanto* case for

[13] Bollinger's account is the result of an attempt to make sense of 'the Skokie affair' in which First Amendment protections were extended to an anti-Semitic neo-Nazi parade in the suburb of Skokie, Illinois.

[14] As Waldron observes, however, Locke's reasoning proceeds rather too swiftly, for even if it is the case that beliefs are unresponsive to the will, they may still be indirectly guided by behaviours which are produced by the will, and which are therefore susceptible to coercion. What Locke's argument overlooks, Waldron says, is that trying to coerce someone's beliefs is wrongful, irrespective of its efficaciousness (see 1993a: 113-14).

anti-hate speech law, we need more than a bare conjecture that the protection of hate speech, despite its harms, will positively influence the political temper of society at large.[15]

ଓ

Instead of casting around for reasons to put a protective ring around hate speech in particular, we may do better to ask again which kinds of expression ought to be protected and why, and then see whether the answer we arrive at can plausibly encompass hate speech. In §2.2 I enumerated various types of expression that should be entitled to a protected species status, given the individualistic conception of speech's value that underwrites the standard roster of free speech arguments. I mentioned (i) expression that contributes to social dialogue about the good life (literature, art, philosophy, and religious proselytisation), (ii) expression that spreads information (journalism, research, educational speech, and public information work), and (iii) expression that combats the development of oppressive or hegemonic social relations (criticism, protest, dissent, satire, and iconoclasm). Now, there is a way of viewing contemporary liberal societies that makes hate speech seem like an instance of, or an appropriate adjunct to, categories (i) and (iii). If we perceive liberal societies as overbearingly insistent upon progressive mores and political correctness, and inherently hostile to conservative views on country, faith, and sexual morality, then hate speech begins to resemble the kind of dissident rabble-rousing or moral unorthodoxy that free speech principles ought to specially protect. Conversely, though, we may understand modern liberal societies as engaged in a reformatory endeavour that has, against much resistance, begun to make modest inroads to alleviate the exclusion and disadvantage experienced by individuals who in some way run afoul of historically dominant views on country, faith, and sexual morality. From that perspective, hate speech looks like just another tired manifestation of a familiar, conventional

[15] There is also a false equivalence in Bollinger's claims about tolerating the intolerant. As Leslie Green says, "toleration is called for only in the face of a judgement that the tolerated conduct... is somehow wrong or deficient" (2008: 284). Bollinger wants the state to tolerate hate speakers, thereby (supposedly) encouraging hate speakers to tolerate those they despise in turn. But it is apt to describe non-interference with hate speech as toleration, only because hate speech is morally wrong. If we call the ethos that is being encouraged in hate speakers 'tolerance', we are implying that hate speakers have a legitimate reason for feeling contempt or resentment towards their targets, reasons they should abstain from acting on (in the name of tolerance).

worldview, rather than a bold, individual stand against orthodoxy.[16] Perspectives alone are not enough to secure protected species status for hate speech. Those who want hate speech to be a protected species of expression on account of its bold unorthodoxy would still have to explain why hate speech is a specially important tool for those who want to dissent to progressive or egalitarian ideals. (After all, if the ideals are really the problem, why not criticise them forthrightly, instead of taking the indirect route by sounding-off against a social group whose standing is implicated in those ideals?) My suggestion, in any case, is that we will only see hate speech as a candidate for protected species status if we have a reactionary view about how contemporary opposition to hate speech functions in our society more generally. The candid *tone* of much hate speech may violate certain modern norms of civility, which penalise frankness and reward weaselly dissembling in public expression, but the identity-political *content* of hate speech goes as much with the grain of our social climate as against it.

## 2.5 Viewpoint neutrality and the speech-conduct distinction

Even if hate speech is not a protected species, though, this doesn't mean that restrictions against it will meet the requirements of viewpoint neutrality. Viewpoint neutrality is not a superfluous artefact that we inherit from a particular jurisdiction's legal system. It derives from the individualistic values that animate free speech arguments at large, specifically, those that condemn governments forcing people to subscribe to approved dogmas and orthodoxies. Now, the legitimacy of anti-hate speech law, as I argued in chapter 1, primarily depends on whether hate-speakers are responsible for harming people. 'Viewpoint' is not and need not be invoked at that juncture. Nicholas Wolfson is wrong, then, when he says that the argument for

[16] Consider the following from Kevin Boyle (2001: 493): "hate speech was once mainstream speech. It was central to European culture. There were no "hate groups" espousing racism and white superiority when it was in fact the official ideology or mainstream idea. Today's racists wear our castoffs, and we have a responsibility for what is done with those castoffs... The prejudice and hatred against Jews came to its apotheosis in the Holocaust but these crimes came out of centuries of prejudice built into mainstream Christianity – Catholic and Protestant. The purveyors of hate against blacks, Jews, and other groups seek to protest the abandonment of these prejudices... They reject democratic ideas such as the equality of all citizens, and their hatred is directed at the beneficiaries of those struggles, such as the black population. Hate speech... seeks to restore theories and ideas that were defeated by democratic struggle".

anti-hate speech laws boils down to the claim that "once society reaches a consensus, as in the area of racist or sexist speech, it is the duty of government to sustain true belief and wipe out false opinion" (1997: 30).[17] If anti-hate speech law is justified, it is because of a duty to prevent harmful wrongdoing, not a duty to impose truth. At the same time, however, if the harm- and responsibility-based justificatory conditions for anti-hate speech law are fulfilled, then such laws will naturally classify the conduct that stands to be restricted by reference to its viewpoint. The question is whether viewpoint neutrality ought to prohibit restrictions on expression that have *this* structure: not that they fundamentally purport to silence a particular viewpoint, but that they make viewpoint the classificatory basis upon which individual expressive acts become liable to punishment under laws justified in other terms.

Some free speech theorists defend a stringent interpretation of viewpoint neutrality, according to which not only the aim of viewpoint suppression, but even the unintended *side-effect* of such suppression (even in the pursuit a legitimate government interest) is enough to render a law invalid (see G. R. Stone 1987: 81ff.). Anti-hate speech law in any guise would undoubtedly be invalid on that interpretation of what viewpoint neutrality requires. However, the stringency of this interpretation is not necessitated by anything in the individualistic conception of speech's value which underwrites viewpoint neutrality. Governments are subject to legislative constraints relating to speech. They *must not* impinge upon people's self-governing status by enforcing orthodoxy through the law. They *must* privilege forms of expression involved in dialogue about the good life, furthering epistemic goods, and resisting hegemony. But the unintended suppression of a viewpoint in the pursuit of a legitimate aim does not violate either of these precepts in itself, and the principle of viewpoint neutrality should be elaborated in a way that reflects this.[18] In the case of anti-hate speech law, as long as

---

[17] Eric Barendt makes a similarly misleading claim when he says the arguments used to justify the restriction of hate speech "can also be adopted to justify banning any speech encouraging the formation of beliefs and attitudes the government dislikes" (2005: 171-72). This concern only holds if one supposes that it is the 'objectionable-ness' of hate speech that is used to justify its restriction, rather than its harmfulness.

[18] As Scanlon says, the paradigmatically wrongful kind of content-based restriction on speech involves an authority forcing people to adopt a view on an issue about which they have good reason to decide for themselves; "not every law that, literally, restricts expression on the ground of its content is threatening in this way" (2004: 1483).

the ultimate justificatory reason for restricting the speech is the harm it inflicts, rather than the view it conveys, viewpoint neutrality should be taken to be satisfied.

The question, though, is what exactly we mean by 'reason for restricting'. Waldron is dubious about attempts to portray anti-hate speech law as compatible with viewpoint neutrality. We can always claim that hate speech is restricted "not because of its content *per se*, but either because of the way it is expressed or because of the likely effect of what is said or published upon society's maintenance of the basic dignity of the members of the targeted group". But instead of such hair-splitting, he says we should grant that hate speech is restricted "because of the effect *of its content* in helping undermine the assurance that members of vulnerable groups are supposed to be able to draw from the public affirmation of their dignity" (Waldron 2012: 151).[19] Waldron is right, on one hand, to emphasise the entanglement of harm and content in hate speech. If hate speech *does* inflict harm on its targets, the process through which this occurs will depend (in many cases, at least) upon the alliance or alignment between the identity-oppressive content of hate speech and the vulnerable status of those against whom it is directed. The content of hate speech is a crucial part of the explanation of what makes it harmful. But underlying purposes matter all the same. If we care about the individualistic values from which free speech principles arise, then we will care about the aims of the legal restrictions under which speech stands to be restricted. Suppose we take a specific act of expression that conveys, say, anti-gay sentiment, φ, and which is liable to punishment under a general prohibition on hate speech, L. When we ask why φ is being punished, the first part of the answer will be: 'because φ satisfies certain content-based criteria, such that it constitutes an instance of hate speech, which is liable to punishment under L'. But then the next question will be: 'why *should* hate speech be liable to restriction under L?' and at this point it matters a great deal whether our answer is (i) 'because hate speech inflicts harm on people'; or (ii) 'because identity-based contempt is morally objectionable'; or (iii) 'because identity-based contempt is inconsistent with the government's stance on questions of identity and human worth'. The reasons for restricting hate speech in (ii) and (iii) may be plausible as far as they go, but they are paternalistic incursions against individual self-government, of a kind inimical to the moral framework upon which free speech is founded. It is not splitting hairs too finely, then, to observe that appropriately formulated anti-hate speech law invokes reason

[19] I will discuss the dignity-based conception of harm that Waldron invokes in these passages in §5.3.

(i), which does not appeal to 'wrongness of viewpoint', rather than reasons like (ii) and (iii), which do, and to defend its legitimacy to proponents of free speech on that basis. The point at which viewpoints *cannot* be invoked in the giving of reasons and justifications in this arena, on pain of breaching viewpoint neutrality, is in explaining the underlying rationale for laws like L in the first place. The fact that the viewpoint expressed in a speech act, φ, is cited to explain why φ is liable to restriction under L, does not mean L itself violates viewpoint neutrality.

ଔ

My claims, then, are that hate speech should not be a protected species of expression, and that its restriction under a harm-based rationale does not violate viewpoint neutrality. If there a residual unease in this position, I suggest it is due to misleading conclusions encouraged by the juxtaposition of (i) our (entirely standard) content-based definition of hate speech,[20] and (ii) the speech-conduct distinction. The latter is a simple construct, implicitly invoked in much free speech theory, which identifies the scope of 'speech' (i.e. what is meant to be free) by contrasting it with 'conduct' or 'action' *simpliciter* (see Emerson 1963; Stevens 1993; Wertheimer 1994; Sunstein 1995). The distinction goes back at least to Mill in *On Liberty* (1859). Mill argues for unfettered expression of opinion, but he acknowledges that

> ... even opinions lose their immunity when the circumstances in which they are expressed are such as to constitute their expression a positive instigation to some mischievous act... Acts, of whatever kind, which without justifiable cause do harm to

[20] Compare with the following definitions: "a generic term that has come to embrace the use of speech attacks based on race, ethnicity, religion and sexual orientation or preference" (Smolla 1992: 152); "symbolism, linguistic or otherwise, that expresses intense antipathy towards some group or toward an individual on the basis of membership in some group" (Corlett and Francescotti 2002: 1083); speech that employs the language of hatred, degradation, and persecution to advocate the superiority of one group of human beings over another (see Matsuda's contribution, in Borovoy et al. 1989: 361-62). Some authors prefer to define hate speech in way that stresses its typical uses (see Brison 1998a: 314-15; Bakircioglu 2008: 4-5), or its typical affective consequences (see Nielsen 2002, 2004). But these are, in most cases, elaborations of a content-based definition, which try to give the reader a grasp of what hate speech *is* by adverting to the sorts of acts or effects for which it is typically used.

> others may be, and in more important cases absolutely require to be, controlled... by the active interference of mankind. (Mill 1859: 119)

The suggestion is that expression only merits protection as 'speech' up to the point at which its character in the context of performance changes from being mere opining to a (potentially harmful) act, like a provocation or an incitement. Free speech by definition requires "protection for speech that goes beyond limitations on government interference with other activities" (Greenawalt 1989b: 120); the idea here is that verbal behaviour itself becomes 'other activity' in some circumstances. Charles Collier speaks of the "obviously experienced differences between speech and conduct" (2001: 233) and, whatever is the case phenomenologically, Collier is surely right at least so far as we are inclined to draw a linguistic contrast along these lines. Phrases like "I was only *saying*", or "nothing's happening, *it's just talk*", come easily.[21] Notwithstanding the distinction's appealing simplicity, though, it carries misleading connotative baggage. Where 'action' and 'conduct' sound like things that impact on the world, 'speech' sounds inert; where 'action' and 'conduct' elicit thoughts of objects, change, and materiality, 'speech' suggests a domain of immaterial thoughts and ideas.[22] The

---

[21] The appeal of the speech-conduct distinction is particularly evident when we see writers trying to distinguish 'speech' from '*real* action', so as to downplay the significance of verbal acts. A recent (and *perfect*) example of this can be found in Richard Dawkins' widely-reported comments, written in response to a young writer's allegation that she had been sexually harassed in an elevator at an atheist convention. After being criticised for downplaying the seriousness of such harassment by comparing it to extreme misogynistic abuse in Middle Eastern states, Dawkins argued, in his defence, that the offender "didn't physically touch her, didn't bar her way out of the elevator, didn't even use foul language at her. He spoke some words to her. Just words. She no doubt replied with words. That was that. Words. Only words, and apparently quite polite words at that... Muslim women suffer physically from misogyny, their lives are substantially damaged by religiously inspired misogyny. Not just words, real deeds". See 'Dawkins Gets into Comments War with Feminists', *The Atlantic Wire*, 6 Jul 2011, http://www.theatlantic-wire.com/national/2011/07/richard-dawkins-draws-feminist-wrath-over-sexual-harassment-comments-39637/.

[22] Catharine MacKinnon sees the inertness of speech as a crucial assumption in responses to her critique of pornography; her opponents misconstrue her claims about pornography's harm, she says, because they insist that "a form of communication cannot, as such, do anything bad except offend" (1994: 8). Fish similarly objects to the assumption in US free speech theory "that mental activities, even when they emerge into

problem, when we come to think about hate speech – which is (i) defined in terms of the ideas it conveys, and (ii) identified (in its very name) as a type of speech – is that these connotations shape our perceptions of what we are discussing, and of what stands to be restricted through the operations of anti-hate speech law. Under the speech-conduct distinction's influence, it is easy think of hate speech as nothing more than its content, and to imagine anti-hate speech law, therefore, as a straightforward violation of viewpoint neutrality, something inherently aimed at suppressing disapproved ideas.

All of this is misleading, as I say. But how can we map the conceptual terrain around speech and 'non-speech' in order to shake off the implications that undermine the case for anti-hate speech law? We can begin by noting the elementary insight behind the work of J. L. Austin, John Searle, and other early speech act theorists. Speech is not something that contrasts with conduct. Speech is a kind of conduct: a voluntary or intentional 'doing' that affects the world. In *How to Do Things with Words*, Austin's initial idea is that the significance of some utterances lies in the acts they constitute. For example, when uttered in the right conditions, the sentences "I now pronounce you husband and wife", or "I dub thee Sir Michael" constitute the acts of marrying the couple and knighting Jagger respectively. What Austin finds, in trying to specify criteria which can categorise all utterances as either 'performatives' like these, or as mere 'constatives', is that all utterances have both 'act' and 'statement' aspects. He is thus compelled to develop a verbal typology, "based not on *types* of sentences, but on *aspects* of a sentence's signification... between the thought that it may express and the action that it may be designed to perform or assist in performing" (Gorman 1999: 109-10).[23] Austin's typology specifies three forms of speech act. When A says to B "it's raining", she conveys meaningful content about the weather, but she also performs the act of warning B about the rain, and simultaneously brings about further effects, e.g. B may be surprised, or persuaded to take his umbrella out. Austin classifies these three dimensions of

---

speech, remain safely quarantined in the cortex and do not spill over into the real world, where they can inflict harm" (1994c: 125-26).

[23] Austin's provisional distinction between *performatives* and *constatives* resembles the speech-conduct distinction, and there are some who favour this distinction over the more complex taxonomy Austin ultimately favours (e.g. Sadock 1994 suggests classifying speech acts as constative or performative depending on which aspect is 'dominant').

verbal conduct as: (i) *locutionary* acts, which consist in conveying meaningful thoughts/sentences; (ii) *illocutionary* acts, which are constituted by meaningful utterances, or performed in utterances, e.g. marrying, naming, warning; and (iii) *perlocutionary* acts, i.e. the bringing about of certain effects through such utterances (1962: 109ff.).[24]

Hate speech can be characterised more precisely, and without misleading connotations, under this schema. The defining feature of hate speech is its locutionary content: it expresses contempt for people on account of their ethnicity, religion, or sexuality. However, such locutions can be put to harmful illocutionary uses (e.g. to harass, threaten, incite, provoke, or bully), and/or used to bring about harmful perlocutionary effects (distress, humiliation, anxiety, and so on). If hate speech is liable to restriction, it is in view of the harms effected in its illocutionary and perlocutionary dimensions, not due to its locutionary content as such. This framework also makes sense of the language some theorists use to try to subvert the speech-conduct distinction in discussions of hate speech. Charles Lawrence, for instance, says hate speech is "both 100 percent speech and 100 percent conduct" (1993: 62). I take this to mean that although hate speech of its essence conveys ideas about identity and human worth, and thereby embodies the principal characteristic of 'speech', it simultaneously has a harmful impact on people's lives, and thereby exhibits a central characteristic of 'conduct'. The speech act theoretic framework can acknowledge hate speech's illocutionary potency,

[24] The locutionary contents and illocutionary uses of speech acts are independent according to the account that I have sketched here. We may need to know what speech says to understand its use, but besides that epistemic link, there is no relationship between the two. Against this view, Alston (1994) says we should analyse locutionary contents in terms of their illocutionary potential, such that facts about use *determine* facts about content. How would that proposal bear on our concerns here? If we define hate speech in terms of its content, we are saying that all instances of hate speech are alike with respect to meaning. However, we do not need to espouse any specific theory of meaning to say this. Meanings may reside in the speaker's communicative aims (Grice 1957), or her expressive aims (Davis 1992), or, as Alston argues, in the uses to which the speech may be put, whether communicative or expressive. All we commit to, in defining hate speech in terms of its semantic content, is that judgements about meaning are ultimately truth-apt. Moreover, the rationale for restricting hate speech remains intact even if Alston is right about illocutionary use being prior to locutionary meaning; treating illocutionary use as conceptually prior to locutionary meaning consorts comfortably with a rationale for restricting speech acts that sees harmful illocutionary use as the basis for restriction.

while allowing it to be defined by its content.[25] A further benefit of the framework is its recognition that all (meaningful) expression has a locutionary aspect, and hence that any restriction on expression – whether aimed at curbing threats, criminal conspiracy, false advertising, the publication of private information, or whatever – involves a suppression of semantic tokens. The case for anti-hate speech law need not and should not say that the locutionary content definitive of hate speech must be suppressed, full stop. It should say, rather, that such locutionary content, expressed publicly, may constitute harmful illocutionary action, or have harmful perlocutionary effects (or a tendency to these effects), and that it should be liable to legal restriction for the purpose of preventing those things. In this rationale there is no need to appeal to the dubious notion that hate speech is 'really' conduct, rather than speech. The idea, as in Lawrence's remark above, is that hate speech is both conduct *and* speech. But it is the aspects pertaining to the former dimension which ground the case for its restriction.

## 2.6 Situating pornography

So far I have set aside questions about how pornography fits into my view on anti-hate speech law. One question is whether expressions of contempt on the basis of people's sex or gender should be liable to restriction under the auspices of anti-hate speech law, or extrapolating further, whether (and how) we might set general parameters on the sorts of groups that may be protected against identity-based vilification under the auspices of anti-hate speech law. I have opted to focus on ethnicity-, religion-, and sexuality-based hate speech merely in order to adhere to the typical approaches in the literature.[26] Having said that, there are complexities

[25] Having stressed the illocutionary dimensions of verbal conduct, some authors (e.g. Langton 1993; Hornsby 1994; 1995b) argue that free speech principles should incorporate protection of the conditions necessary to engage in certain kinds of illocutionary action specifically. I discuss this proposal in §5.6.

[26] One objection to singling out ethnicity, religion, and sexuality in anti-hate speech law is that this discriminates against other groups targeted in hate speech. According to Eric Heinze (2006, 2009a), anti-hate speech laws that are only addressed to attacks on these bases violate the kind of anti-discrimination principle which would, so he maintains, provide the best justification for the existence of such laws in the first place. One who defends anti-hate speech law, Heinze says, must produce sample hate speech legislation that would

that arise in discussing pornography specifically, which may require careful treatment of their own, and this is a further reason for cordoning-off pornography and hate speech against women in this study.[27] In any case, as a conclusion to my discussion of free speech, it is worth spelling out the structural resemblance between the case for anti-hate speech law and the contemporary feminist case for restrictions on pornography, insofar as the two lines of argument can (or should) call on a similar reply to the claim that they impose intolerable limits on free speech. Some authors see more extensive affinities in the two arguments than just what they say about free speech (see Futch 1995; La Caze 2004), although others argue that critiques of misogynistic speech and racist speech are in fact in tension with one another (see Crenshaw 1993). Here, in any case, I will just consider the free speech parallels.[28]

In the 1970s a major shift in the moral critique of pornography began. Under the older *obscenity* paradigm (operative, for example, in Crawford 1968, Richards 1974, R. Dworkin 1981a, Skipper 1993, and Kristol 1994; and dissected in MacKinnon 1984 and Koppelman 2005), pornography's wrongness was primarily understood in terms of its capacity to corrupt the moral character of its consumers, or to upset the sensibilities of people unwillingly

---

avoid such discriminatory results, and this, he believes, cannot be done (2006: 564). This seems over-demanding. If our aim is to redress the harm done by hate speech, surely we can tackle some harm first, whether by principle of triage, or by virtue of the need to start somewhere. In any case, my analysis of hate speech based on ethnicity/religion/sexuality, can, without major changes in the framework of discussion, be programmatically extended, to encompass 'speech that expresses contempt for people on the basis of any marker of social identity, *including, but not limited to*, ethnicity, religion, or sexuality'.

[27] For instance, as Saul (2006b) argues, any general claim about pornography's illocutionary use must take account of the fact that pornography consists in recordings, manufactured for repeated use, in unspecified contexts. Bianchi (2008) seeks to defend the Langton/MacKinnon-style critique of pornography against Saul's claim that only utterances in fixed contexts can be speech acts, by suggesting the context of pornography's interpretation as the illocution-determining context of pornography-based illocutions. That response leads to the peculiar position, though, that the character of many pornography-based illocutions is not determined by the pornographer.

[28] Presumably there are some instances of pornography that uncontroversially qualify as hate speech, insofar as some audiovisual material that is produced for the purposes of sexual arousal also contains expressions of identity-based contempt towards a particular sexuality or ethnic/religious group. Those expressions of contempt may be integrated into the purportedly arousing elements on the pornographic materials, but they need not be.

exposed to it or brought into its vicinity. Pornography's harm, under this paradigm, was to the moral purity or mental equanimity of its audience. Under the newer *inequality* paradigm, disseminated and developed by Catharine MacKinnon and Andrea Dworkin,[29] pornography's moral wrongness is instead understood primarily as a matter of the harm it does to women – the women involved in its production, for one thing, but also women at large, so far as pornography "contributes significantly to the continuing subordinate position of women" (Langton 1990: 333). Precisely how pornography makes its contribution to subordination is a matter of ongoing inquiry. The most common suggestion, initially from MacKinnon, but elaborated by others since, is that pornography *silences* women, which is to say (roughly) that it creates an environment in which women's speech capacities are impaired in specific respects. (I discuss silencing further §5.6, in relation to parallel contentions about the effects of hate speech.) More generally, the inequality paradigm posits a significant causal (or constitutive) relationship between the production and consumption of pornography, on one hand, and the disadvantages and harms inflicted upon women on account of their sex, on the other. Arguments for legal restrictions on pornography, within this paradigm, primarily consist in attempts to vindicate claims about pornography's role in inequality.

The feminist anti-pornography movement was less successful than some other late 20th century feminist initiatives opposing sexual inequality.[30] This was at least partly due to the insistence of many free speech advocates (e.g. Parent 1990; Rodgerson and Wilson 1991; Strossen 1993; R. Dworkin 1995) that anti-pornography laws would fall afoul of free speech principles, and notwithstanding the patient efforts of other authors to show that such laws were in fact compatible with established free speech thinking (e.g. Sunstein 1986). By the time the movement started attracting criticism from pro-pornography feminists (e.g. Carter 2000; Royalle 2000) this division seemed to substantially defuse the movement as a political force.[31]

---

[29] The inequality paradigm has since been developed by several authors, including Vadas 1987; Langton 1990; 1993; 1998; Hornsby 1994, 1995a; 1995b; McGowan 2003; 2005, and Maitra 2009; 2012.

[30] For example, the anti-sexual harassment movement; see MacKinnon 1979; Brownmiller 2000.

[31] One thing evident in those critiques was that anti-pornography feminists, in critiquing women's objectification, had little or no sense of the possibility of positive forms of objectification, of the sort adverted to, e.g. by Nussbaum (1995) and Green (2000).

Even if pornography as 'free speech' was not thoroughly vindicated as a theoretical position, the widespread legal toleration of pornography continues today in practice.

What do the arguments for anti-pornography law and anti-hate speech law have in common? The resemblance lies in the combination of a content-based definition of the expressive conduct that stands to be restricted, and a harm-prevention rationale for the restriction. As with hate speech, pornography can be categorised by reference to the negative view it conveys about a certain group. One may question the anti-pornography feminists' claim that pornography necessarily conveys a view of women as second-class people. But presumably a decent portion of pornography conveys that view, and whatever the case, there are clearly some uses of pornography (e.g. its circulation in a workplace) that are almost certain to convey a demeaning view of women given the communicative context. However, similarly to anti-hate speech law, the putative justification for anti-pornography law – under the inequality paradigm – doesn't hinge on the judgement that pornography's content is morally wrong and should therefore be suppressed. The claim, rather, is that the content of pornography, when put to certain expressive uses, constitutes harmful illocutionary action or has harmful perlocutionary effects, and that it should be liable to restriction for the purpose of preventing or redressing these harms.[32] In both the pornography case and the hate speech case, the claim is not that what stands to be restricted is *really* conduct, and not speech. Rather, the claim is that these things are simultaneously conduct and speech, but that it is the aspects associated with the former dimension (how these things can harm people), not the latter (the locutionary contents they convey), that underwrite the case for legal intervention.

ଔ

For any given legal system, there are a variety of objections that may be raised against legal restrictions on hate speech. Such laws may be condemned as an exorbitant use of government resources. They may be deemed difficult to police or prosecute. They may be

[32] Williams (2005a) says that locating anti-pornography arguments within anti-hate speech arguments requires proponents of the former to abandon their claims about pornography's harmfulness. On my view the contrary is true: the two arguments *unite* around claims about the harmful uses of the speech in question.

judged more likely to stir up identity-based hostility than to defuse it.[33] In the literature such objections are most commonly pressed, as part of a two-stage argumentative strategy, by authors who identify themselves as defenders of free speech. At the first stage, the author argues that restrictions on hate speech are an inappropriate and ineffective tool for trying to prevent or redress the harms and inequalities associated with hate speech. At the second stage, the author argues that even if restrictions on hate speech *were* appropriate and effective tools for preventing harm and combating inequality, they remain unjustifiable infringements of free speech. Objections put forward at the first stage must be assessed case-by-case, and I have said nothing here that stands as a general reply to concerns about the ineffectiveness or inefficiency of anti-hate speech law.[34] My discussion can be taken as a response to the arguments free speech defenders advance in the second stage of their critique of anti-hate speech law. On a duly elaborated understanding of free speech as a legislative principle, the fact that anti-hate speech laws restrict verbal conduct is not a good in-principle reason to reject such laws, much less a decisive reason. Legal restrictions on hate speech do not apply to a 'protected species' of speech, and properly formulated, they do not violate viewpoint neutrality. The harm and responsibility questions should still be central, therefore, to any inquiry into the in-principle justifiability of anti-hate speech law.

---

[33] To cite just a few examples of these kinds of objections from the literature, we find free speech advocates arguing (i) that anti-hate speech law is/would be simply ineffective in its aims (Coliver 1992; Lively 1994), (ii) that anti-hate speech law would lead to suppression of artistic work and other forms of potentially subversive expression (Adler 1996; Edgar 2009), and (iii) that anti-hate speech law is in fact likely to be used as a further tool of oppression against vulnerable and marginal groups (Strossen 1995).

[34] As Ruth Wedgwood (1988: 330) says "these are pragmatic arguments that depend upon a particular time, a particular place, and a particular history – and they are not arguments in-principle about the power of the state".

# ☙ 3 ❧

# Harm and responsibility

## 3.1 Direct and indirect harms

We can make a start on the harm and responsibility questions by mapping out different causal routes through which hate speech might inflict harm. Schematically, the simplest case is one in which the hate-speech-act is directed towards someone targeted in the locutionary content of the speech, and used to perform the illocutionary act of harassing or threatening the target, or to achieve the perlocutionary effect of making the target feel afraid, upset, or humiliated. Imagine, for instance, someone saying "f**k off home, rag-head", to a Sikh man on the street late at night. Call this type of case – in which the audience for the hate-speech-act is also the person targeted in the speech, and the person liable to be harmed by it – a 'hate-speech-based verbal assault'. We can contrast this with a case in which the audience for the hate-speech-act differs from the person/s targeted in or liable to be harmed by that expression; e.g. a case in which anti-Sikh hate speech is used to provoke or encourage non-Sikhs to commit acts of violence against Sikhs. Call such cases 'hate-speech-based incitements'. Verbal assaults and incitements may occur in parallel, e.g. in a case where anti-Sikh hate speech is expressed to a mixed group, including Sikhs (who are harassed) and non-Sikhs (who are incited to violence). More complex scenarios may also arise, e.g. a case of anti-Sikh hate speech provoking people to attack others whom they wrongly believe to be Sikh, or provoking Sikhs into reacting violently against the hate-speaker, or against some associated third party.

The differences between these cases involve different causal pathways connecting the hate-speaker, the hate-speech-act, the audience, and the people targeted in – and/or who stand to be harmed by – the hate-speech-act. At the same time, the cases are all alike in one significant respect, in that the hate-speech-acts have something like a counterfactual causal connection to the harmful outcome in each instance. In verbal assaults, incitements, and acts

of expression combining the two, the relevant harms cannot and do not eventuate if the relevant speech acts are not carried out. This does not straightforwardly entail that hate-speakers are liable for harm in all such cases. As we will see, liability-ascriptions hinge (or should hinge) on judgements about how all the parties implicated in a harm-inflicting causal sequence contribute to the harm that eventuates. Setting aside liability for now, though, the point is that in verbal assaults and incitements, speech is necessary for harm.

How could it be otherwise? How could hate-speech-acts inflict harm in cases in which there *wasn't* causal necessity linking speech and harm? First, a red herring: hate-speech-acts may overdetermine harmful outcomes. Suppose two people verbally harass the same victim, in such a way that either one's speech, absent the other's, would inflict the sort of harms (involving distress, anxiety, etc.) associated with harassment. In such a scenario, as in 'two bullet' cases of overdetermination, it does not seem to be true of either person's speech that had it not been performed, the harm would not have occurred. Therefore, given a bare counterfactual conception of causation, neither individual's harassing speech caused the harm. First things first, then, we need a more finely calibrated analysis of causation. The crucial common feature of the scenarios described thus far (i.e. harassment, threats, incitement) is not that the hate-speech-act is strictly necessary for the harm to occur, but that in each case it is a necessary element in a certain kind of causal sequence, one that is sufficient – but not uniquely sufficient, and hence not *necessarily* necessary – to bring about the harmful outcome. This qualified conception of causal non-redundancy can be spelled-out in a number of different ways;[1] in any case, the question that I was trying to pose above was how hate speech can be

[1] Counterfactual analyses of causation have been defended in the philosophical literature on causation (e.g. see Mackie 1974), but they are famously ill-equipped for characterising cases of causal overdetermination and/or joint determination. One prominent alternative account of causation invokes the so-called NESS condition: $\varphi_1$ is a cause of x iff $\varphi_1$ is a Necessary Element of a Sufficient Set of conditions for bringing about x under the circumstances (see Hart and Honoré 1985; R. W. Wright 1985; Honoré 1995). A NESS analysis of causation preserves the definitive notion that causes are non-redundant with respect to their effects, but unlike a bare counterfactual analysis, it embeds this modal relationship in a disjunctive structure; $\varphi_1$ is a necessary part of a set of conditions, $S_1$, but $S_1$ is merely sufficient for bringing about x, and another set of conditions $S_2$ (involving other actions, $\varphi_2$, $\varphi_3$, etc.) may also be sufficient to bring about x, consistently with $\varphi_1$ having caused x via $S_1$. As a rather different alternative, some have sought to devise scalar accounts of

said to inflict harm under circumstances in which even *this* kind of non-redundant causal connection between hate-speech-act and harm does not obtain.

The short answer is: through a non-necessary causal contribution. But I want to approach that answer indirectly. Some authors talk about hate speech harming society *per se*. Anthony Cortese, for instance, says hate speech harms "not only the victim and the performer of individual racist acts but also society as a whole" (2006: 50). One way to read this claim is as follows: 'the very existence of hate speech in a society makes that society morally worse, and thus lessens its moral calibre'. On that interpretation, Cortese's claim has some independent plausibility, but it also represents a patently illiberal rationale for restricting hate speech. For liberals, the law does not aim at society's general moral edification; rather, it aims to prevent harmful incursions that impair people's ability to live according to their own values and purposes. So shouldn't we just ignore those who claim that hate speech harms society? No, because their claim can be interpreted in another way, on which it does say something relevant to a liberal conception of the law's purpose. So far as the welfare of individuals is profoundly dependent on the character of the social environment they inhabit, talk of hate speech harming society can be treated as a claim to the effect that hate speech damages the social environment in such a way that harm ultimately redounds to individuals within that environment. Joshua Cohen adverts to this kind of harm in a discussion about the negative consequences of speech more generally. Speech, he says, "may help to constitute a degraded, sickening, embarrassing, humiliating, obtrusively moralistic... or demeaning environment", and although in such cases we cannot "trace particular harmful or injurious consequences to particular acts of expression that help to constitute the unfavourable environment", we may still judge that certain acts of expression do contribute to the degradation of the social environment, with harmful results in turn (Cohen 1993: 231).[2] In *Social Philosophy* (1973), Joel

---

causation which quantify the degree to which acts or conditions contribute to outcomes (e.g. Moore 2009). A NESS-type analysis would suffice for our purposes here, in any case.

[2] Some authors in the hate speech literature argue that liberal legal systems are ill-suited to recognising harms brought about via this environmentally-mediated causal route. Delgado (1996: 82) says that "focus on the imminence of the harm impairs the court's understanding of the totality of the harm relationship... the law requires [victims] to tell a stylized story that might or might not correspond to [their] injury". MacKinnon (1984: 337-38) presses this point in her critique of pornography: "Words work in the province of attitudes... pornography must cause harm like negligence causes car accidents or its effects are not cognizable as harm.

Feinberg sets out a framework for distinguishing these kinds of environmentally-mediated harms from harms inflicted via a more direct route. After describing a 'private harm principle', which permits restrictions on acts that directly harm individuals, Feinberg contrasts this with a 'public harm principle', which permits restrictions on acts when such restrictions are "necessary to prevent impairment of institutional practices and regulatory systems that are in the public interest" (1973: 25). Feinberg identifies tax evasion and contempt of court as paradigmatic forms of the 'publicly harmful' conduct liable to restriction under the auspices of a public harm principle. Isolated instances of these behaviours are not (typically) harmful in any familiar counterfactual sense, but they still inflict harm, Feinberg says, "insofar as they weaken public institutions in whose health we all have a stake" (1973: 25-26).

My suggestion, then, is that casual remarks about the damage hate speech does to society can be interpreted as adverting to public harms in Feinberg's sense. My preferred terminology for discussing such harms, in relation to hate speech, rests on the simple contrast I introduced in §1.1, between *directly* harmful hate speech and *indirectly* harmful hate speech (or correspondingly, between the direct and indirect harms *of* hate speech). The former are cases like verbal assaults and incitements, in which hate-speech-acts play a causally non-redundant role in bringing about discrete harms to specifiable individuals. Indirectly harmful hate speech, by contrast, involves a graded causal relation between hate-speech-act and harmful outcome. This kind of causal relationship is naturally described using words like *contribution* (e.g. to a harmful outcome) or *impairment* (e.g. of some institution, in a manner such that harm redounds to individuals).[3] Contributions and impairments are not all-or-nothing ways of bringing about outcomes. Clinton's convention speech can contribute to Obama's re-election without causing it. Ryan's gaffe may impair Romney's campaign, without causing Romney's defeat. When we

---

The trouble with this individuated, atomistic, linear... conception of injury is that the way pornography targets and defines women for abuse and discrimination does not work like this".

[3] Sumner (2009: 208-09) offers a distinction along similar lines, and we also find something like this distinction in Clay Calvert's analysis of hate speech. Calvert contrasts a *transmission* model of communication with James Carey's (1989) *ritual* model, where the latter sees communication as a process of constructing shared meanings. Calvert distinguishes harms associated with hate speech under the transmission model, i.e. "direct, immediate, and overt behavioural changes, physical responses, and mental anguish", from those related to the ritual model, i.e. "cumulative harm that accrues with repeated use of racist epithets" leading to "the perpetuation and reinforcement of discriminatory attitudes and behaviours" (1997: 6).

say that Clinton or Ryan contributed to the election outcome, we are saying that although they did have some influence on the outcome, their influences were ultimately non-decisive or redundant with respect to the outcome that is salient to our causal inquiry.[4]

How can hate-speech-acts exemplify this kind of contributive causal relationship, with respect to harms that may be inflicted on their targets? The most common contention we find in the literature is that hate speech contributes to identity-based social hierarchy, i.e. the patterns and practices of disadvantage and mistreatment experienced by people in oppressed or culturally marginalised social groups. Hate speech does not single-handedly cause cultural divisions based on sexuality, or the subordination of ethnic groups, or discrimination against religious groups. However, so many authors claim, it does contribute to these systemic states of affairs in discernible and significant ways. In the words of David Dyzenhaus (1992: 536), hate speech operates in this arena as "a pernicious prop of inequality".

Just what it would take to substantiate such contentions is a difficult question in its own right. Suffice to say, given that responsibility for harm is necessary for any justifiable restriction on hate speech, and given that there are differences in how responsibility for causing harm and responsibility for contributing are ascribed, the considerations that would need to be cited to justify restrictions on indirectly harmful hate speech are different to those needed to justify restrictions on directly harmful hate speech. For this reason we do better, as I said in §1.1, to treat anti-hate speech law as an overarching genus with two sub-species – (i) restrictions on *directly* harmful hate speech, and (ii) restrictions on *indirectly* harmful hate speech

---

[4] In chapter 4 of *On Liberty*, Mill posits a distinction that roughly corresponds with the distinction I have outlined here. When an individual "is led to violate a distinct and assignable obligation to any other person or persons" Mill says, "the case is taken out of the self-regarding class [of actions] and becomes amenable to moral disapprobation... and might be justly punished" (1859: 148). A paragraph later he contrasts this kind of harmful conduct with the kind of harmful conduct in which "a person disables himself, by conduct purely self-regarding, from the performance of some definite duty incumbent upon him to the public" (Ibid: 149). Mill discerns a liability-grounding harm in both kinds of cases, but where the former (like directly harmful hate speech) involves a specific incursion against a specific individual, the latter (like indirectly harmful hate speech) involves a dereliction of duty towards public goods in whose health many people have a rightful stake. The law, according to Mill, may legitimately intervene to prevent both kinds of harm: "Whenever, in short, there is a definite damage, or a definite risk of damage, either to an individual or to the public, the case is taken out of the province of liberty and placed in that of morality or law" (Ibid: 149).

– with the legitimacy of each resting on different justificatory grounds.[5] When hate-speech-acts inflict discrete harms on identifiable victims, as in cases of verbal assault or incitement, there are no in-principle reasons to oppose their legal restriction, once free-speech-based concerns have been addressed. In such cases, the justificatory questions are about (i) how we ascertain the nature or severity of the harms in question; (ii) when and why, given the conduct of other parties, the hate-speaker's causal role in the harmful outcome is such as to render the speaker liable for the harm; and (iii) whether we ought to treat such acts as a distinct class of offence, or instead merely as an instance of some broader class of offence, such as verbal harassment.[6] Attempts to justify restrictions on *indirectly* harmful hate speech will call upon an altogether different set of answers to the harm and responsibility questions. Where indirect harms are at issue, the hate-speaker's liability does not depend on the identification of a non-redundant causal relationship between his hate-speech-acts and some discrete harm to an identifiable person (or people). Rather, the speaker's liability depends on him having performed a type of speech act whose contribution to the existence of a harmful system or

[5] The distinction is motivated by pragmatic concerns about how responsibility for harm can be ascribed to individual speakers. It may, however, distort the phenomenology of being targeted in hate speech. A well-worn claim in the critical race theory literature on hate speech (e.g. Jones 1980; Williams 1987; Bell 1995) is that victims of hate speech feel the immediate harm of hate speech to be exacerbated by wider conditions identity-oppression and prejudice, simultaneously with hate speech perpetuating those conditions.

[6] I will take up (i)-(iii) in the remainder of this chapter. One further question which I will not discuss, but which does need to be addressed at some point in the legislative details of restrictions on directly harmful hate speech, is whether there should be any distinctive forms of excuses or exculpatory considerations which should override a *pro tanto* liability-ascription to the speaker for harms resulting from directly harmful hate-speech-acts. One might argue, for example, that given the characteristic social dynamics involved in 'tit-for-tat' verbal conflicts, liability for harms resulting from *retaliatory* verbal assaults need not, or perhaps in general should not, be ascribed to the retaliatory speaker.

state of affairs has been independently established, and, moreover, has been judged significant enough that people who perform acts which token that type may be justifiably held accountable for the harms arising from the whole pernicious system.

ଔ

Can these justificatory burdens be satisfied? As I said in chapter 1, my view is that whereas restrictions on directly harmful hate speech are justifiable, and should limit a relatively wide range of hate-speech-based expressions, restrictions on indirectly harmful hate speech are not justifiable, because we cannot establish that hate speech does significantly contribute to the harms associated with it via indirect causal routes. I will spell out the reasons that I take to support these claims in the following chapters, focusing on directly harmful hate speech in chapter 4, and indirectly harmful hate speech in chapter 5. In the rest of this chapter I address two more general sets of questions about harm and responsibility in hate speech. In §§3.2-3.3 I consider how we should distinguish, among the adverse effects associated with hate speech, between harms proper and *merely* adverse experiences.[7] In §§3.4-3.6 I ask under what circumstances responsibility for harmful outcomes should render the hate-speaker legally liable for harm, given that (nearly) all of the harms at issue are such that when hate speech *does* contribute to them, the causal sequences linking speech to harm involve intermediating cognitive processes on the part of responsible agents other than the hate-speaker. In §3.7 I

[7] When I say an outcome is 'associated with' hate speech, I just mean that some causal or contributive relationship is alleged (rightly or wrongly) to obtain between the two. This expression makes it easier to distinguish the harm question (is outcome x, associated with hate speech, *harmful*?) from the responsibility question (is hate speech *responsible* for outcome x?). Note that I am treating harms or harmful outcomes as conceptually prior to (i) *acts* of harming and (ii) the *property* of an act/state of affairs being harmful. If we treat acts of harming or the property of harmfulness as conceptually fundamental, then we cannot separate the harm and responsibility questions. This worry, more than anything, explains my disagreement with Matthew Hanser's claim that "suffering harm is a matter of undergoing an event whose status as the undergoing of a harm does not derive from the badness of any associated *state* at all" (2008: 422). That view forces us to bundle together any diagnosis of a harm being done with a responsibility-ascription for that harm.

take a brief tangent to show how a focus on responsibility-ascription, rather than outcome alone, may resolve some difficulties in liberal theorising about offence.

## 3.2 The harm question

Which among the adverse effects of hate speech constitute harms? One may be tempted by the 'sticks and stones' view, i.e. that *none* of the negative outcomes associated with hate speech are harms proper, because speech is ostensibly inert and unable to do any harm on its own.[8] In §2.5 I described how free speech theorists commonly invoke a contrast between speech (which is immaterial and inert) and action or conduct (which are concrete and effectual). The sticks and stones view descends from a similar perspective. If there is a kernel of plausibility in it, it lies in the idea that until expression is met with an interpretation and response, it is only a noise or marking. Audience capacities have to be engaged in some form before speech inflicts harm, and in that limited sense, speech by itself is harmless. Or so one might argue. The first problem with this view is that it is false *even if* we grant that speech acts are not genuinely

[8] The children's adage that these words come from has been cited by US Supreme Court Justice Scalia as an encapsulation of the First Amendment's core thesis: "Sticks and stones will break my bones, but words can never hurt me... that's the First Amendment" (see *Schenck, v. Pro-Choice Network of Western New York* [1996] No. 95-1065 S. Ct LW 608239). To be fair, these remarks were Scalia's summary of the view being defended by the petitioners in *Schenck*, rather than his own view of the First Amendment. In the hate speech literature, the 'sticks and stones' view often takes the form of a complaint about how a focus on hate speech distracts from the *real* work of combating injustice; as in Marjorie Heins' remark that "too much work remains in the battle against the... institutional racism entrenched in our society to squander resources suing loudmouths" (1983: 592), or Henry Louis Gates' claim that efforts to curb racist speech merely pay lip service to civil rights, and ignore the real economic aspects of inequality (1993: 43); or in Martha Minow's statement that "if the goal is to eradicate the dehumanization of the hated group, attention would more fruitfully turn to... the surrounding social traditions and practices that have made and continue to make that group subject to dehumanization." (2000: 1274). Sometimes the view operates more subtly. For instance, in Cass Sunstein's discussion of anti-hate speech laws, he says such laws "could not plausibly be focused on consequences", since "the stakes are relatively low" (1996: 2023). This claim – that 'the stakes are relatively low' in hate speech – does not come after a reply to those who argue that the harm stakes in hate speech are actually very *high*; it is just stated as a commonsense observation. The prevalence of the sticks and stones view is what allows Sunstein (and others) to see this assertion as a matter of assertible commonsense.

harmful if their injurious effects depend on other people's reactive engagement. Person A can scream abuse in B's ear, or spraypaint messages on B's front window. In either case the harm is brought about whether or not B's agential capacities are exercised to interpret or make sense of the expression. The second problem is that the proposed criterion for an expressive act being genuinely harmful (i.e. that no-one's agential capacities should have to be engaged to interpret its meaning) conflates two distinct assessments. To say an act is harmful is to say something about its consequences. If a speech act, φ, is causally to blame for a harmful outcome, x, then φ can be deemed harmful irrespective of where and how people's agential capacities are brought to bear in the causal sequence leading to x. Questions about whose agential capacities were engaged at what point become pertinent in ascriptions of blame and liability, not in ascriptions of harmfulness *per se*. It is true that, given the cognitively-mediated process through which (nearly) all speech gets its purchase on the world, there are (nearly) always further questions to be asked about where responsibility for the harmful consequences of speech rightly belongs. But this is no reason to deny speech's harmfulness.

One may concede that speech is harmful in some sense, while claiming that the harm it inflicts is an order of magnitude beneath full-blooded 'material' harm.[9] However, this claim fares no better than the sticks and stones view. Firstly, it faces the same counterexamples (e.g. the ear example) of speech that does inflict immediate material harm. Secondly, it is vulnerable to counterexamples in which speech inflicts material harm by causing a third party to inflict material harm. Suppose B hits C, at A's verbal command, and that B would not have done so

[9] Frederick Schauer (1993) calls this 'the lesser harm hypothesis'; Cohen (1993) calls it 'low-cost minimalism'. This view has been used to defend free speech principles in the first place. Michael Bayles (1986: 54) argues that liberty in speech is justified in part because it is "less likely to interfere with the exercise of other liberties than is... liberty of action". The claim that speech is relatively harmless can arguably be identified in *On Liberty* as well. Mill classifies expression of opinion as a 'self-regarding' action, and argues that it should be immune from restriction on that account. Given that, for Mill, (i) harm to others is a proper reason for government interference with people's conduct, and (ii) self-regarding actions (like speech) are meant to be immune from government interference, the implication seems to be that self-regarding actions are harmless, and that as a member of the class 'self-regarding actions', expression of opinion is harmless. Against this reading, Daniel Jacobson (2000) argues that Mill sees self-regarding actions – including speech – as immune from restriction solely due to their ultimately beneficial effects on individual welfare. On this reading, speech is protected not because it is harmless, but for its independent utility and *despite* its harm-causing capacity.

were it not for A's command. In this case, A's speech is clearly causally necessary for the harm done to C. Again, there are further questions about how responsibility-ascription should go in such cases. But the initial causal diagnosis – that A's speech caused a material harm to C – is independent from the answers to those questions. The third problem with the 'lesser harm' thesis is that when we consider the adverse non-material consequences of speech – that is to say, the mental or emotional hurts that it may bring about – it seems implausible that all such consequences are, as a general rule, an order of magnitude beneath material harms.[10] Even if mental harm is harder to objectively detect or quantify than material harm, those are epistemic discrepancies, not differences in the gravity or severity of the adverse experiences in question. And moreover, such epistemic discrepancies tend to be overstated. It is not clear, for instance, that the harms involved in physical injuries are more able than those involved in mental anguish to be quantified in a manner that does not call on the subjective judgements of the affected individuals (see Brison 1998b: 44-46). Charles Ogletree (1996: 503) stresses this point in relation to hate speech specifically: the mere fact that a hurt is mental rather than physical does not make it less genuine or severe. I believe Schauer (1993) is right, then, to say that the only way to establish that the psychological hurts inflicted by speech are sub-harmful, or that they are mere ersatz harms, is to write this judgement into one's definition of 'harm'.

ଔ

The question of how we should characterise the adverse effects of hate speech cannot, then, be brushed aside by appeal to theses about the relative inefficaciousness of speech. We need to enumerate the different types of adverse effects for individuals that may be plausibly causally associated with hate speech, and ask which of those constitute harms, given some general guidelines for distinguishing harm from mere offence, dislike, discomfort, or disappointed preferences. On the first (enumerative) point, two types of adverse outcomes associated with hate speech have already been mentioned. First, hate speech may negatively affect people's minds, thoughts, and feelings. This is the kind of effect we associate with speech used to harass or threaten people. Second, hate speech may cause physical injury, to

[10] To be clear, my claim is not that mental events as such are immaterial (e.g. occurring in some sort of disembodied soul), but rather, that the *damage* involved in mental pain is not material damage.

people's bodies or property. This is the kind of outcome we typically associate with speech used to provoke or incite violence. A third kind of adverse outcome associated with hate speech is what I will call a 'status injury'. Hate speech may lower people's standing, rob them of their dignity, subordinate them, or in some other respect damage their social position or status relative to others. Status injuries may not be neatly distinguishable from adverse mental or physical effects; that is to say, they may partly consist in the experience of mental and physical adversities of particular kinds and in particular patterns (see J. R. Cohen 2007). Nevertheless, in at least some instances, status harms are not entirely reducible to mental or physical harms. Suppose, for example, that all of person A's neighbours despise her, but that this dislike is manifested in actions which, taken individually, do not inflict any mental or physical harm upon A (e.g. suppose others only interact with A if they are forced to, and that they are distant and cold when they do). If B causes A to be in this position, where A had previously enjoyed warm relations with others, then B seemingly harms A's status, even though no other harms as such have been inflicted on A by her neighbours.

So there we have three types of negative effects associated with hate speech (which may of course overlap or coexist in particular cases). Our question is: in virtue of what do specific instances of these types constitute harms proper, rather than *merely adverse* outcomes – like feelings of offence, discomfort, dislike, and disappointed preferences?

It is unlikely that there will be anything distinctive to say about the adverse *physical* effects associated with hate speech. At some point on the continuum from fleeting soreness to a permanent and disabling injury, the fitting characterisation of an adverse physical effect changes from a mere hurt to a fully-fledged harm. Wherever and however that threshold is identified for legal purposes in general, it can be applied in cases of hate speech as well. When hate speech is causally implicated in outcomes that fall on the 'harm' side of the threshold, all the important and difficult questions are about how liability ought to be ascribed.

Status injuries are more complicated. Certain types of ups and downs in one's social standing are part of the normal, inevitable hurly-burly of human relationships. What makes a status injury harmful, when it is harmful, is at least partly a matter of its invidious origins, e.g. the use of slander, threats, lies, or the cultivation of false consciousness and a self-loathing psychology, in order to originate or perpetuate the status hierarchy in question. Note, however, that contrastive ways of thinking about such injuries makes little sense where the

injuries descend from a widespread identity-based hierarchy. When we say that φ contributes to the subordination of gay people, we are not claiming that gay people's social standing was in good shape before φ, but not afterwards. The contention, rather, is that φ is implicated in the processes through which the subordination of gay people is replicated over time. But then, what makes the status injury in question harmful, if it is harmful, cannot be understood simply in terms of gay people being 'worse off'. It will have to be something about the inherently bad character of occupying a subordinate position in which the harm of a status injury resides: something about the severity of its impact upon one's life chances, or on one's day-to-day lived experience. To characterise a certain status adversity as a status *harm*, then, is to render a judgement both about its origins (that they are sinister, and not merely accidental), and about its severity: that it is a pervasive or enduring source of disadvantage, rather than an ephemeral downward 'blip' in the continual waverings of one's relative social esteem.

Adverse mental or emotional effects occupy a similarly complicated position. In a free society we have no strict duty to be kind or friendly to one another. We cannot expect our neighbours to behave in ways that we find agreeable, nor are we obliged to conduct our own affairs in ways they find agreeable. It is almost inevitable, therefore, that we will experience *some* annoyance, dislike, and offence in interactions with others. Moreover, where we are in competition with others, or have other-regarding preferences, our preferences will sometimes, inevitably, be disappointed. However, at some point the character of our unhappy interactions with others changes from mere social friction into something more malicious, like bullying, harassment, or threats. The adverse mental effects in these cases cannot be counted as harms if they are only very mild (there is no harm in the bad feeling elicited by a single rude remark, for instance). But severity of feeling alone is not sufficient for something's being a mental harm. Some adverse mental effects arising from competition and social friction may be more severe than those brought about by wilful hostility. I may be mildly upset by being threatened, but devastated about being spurned by a lover, or bested by a rival to some coveted prize. Mental harms, therefore, should be identified as those adverse mental effects which not only cross a threshold of severity, but which also originate in particular types of unhappy social interactions, involving targeted manifestations of hostility or ill-will.

It seems almost certain that some of the emotional hurts brought about by hate speech, as well as some of the status injuries related to ethnicity-, religion-, and sexuality-based

oppression and social hierarchy (so far as the latter *are* causally connectible with hate speech), will fall on the 'harm' side of the thresholds that I have sketched, rather than on the 'mere adversity' side. Someone who is humiliated and cowed via relentless harassment using cruel epithets and identity-based slander would not (so we would naturally expect) suffer only a mild twinge of irritation and dislike. The feelings generated by such an experience – given both their severity and their malicious origins – can be appropriately characterised as psychological or mental harms. Similarly, someone who experiences pervasive discrimination and loathing from others, on no grounds other than her social identity, is not just 'having a bad run' in the routine ups and downs of dynamic social relations. The invidious origins of such an experience (insofar as it stems from entrenched, identity-based social hierarchies), and its potentially ruinous impact upon the target's prospects and everyday life, are such that this sort of experience can be appropriately characterised as a status harm.

## 3.3 Autonomy, interests, and the harm principle

One might wonder whether we can find a more principled characterisation of the crucial distinctions between harm and mere adversity. The best approach to fleshing out that distinction, I believe, is to consider the purpose that it is supposed to serve, so far as it is put to use in the harm principle. For liberals, as I said in §1.4, the harm principle is a device used to try to forge a liberty-optimal balance both in our relations with governments and with one another. It permits the legal restriction of harmful acts, in order to protect people against hostile incursions from others, but it insists that *only* harmful acts can be restricted, and thus aims to protect people against governments attempting to impose substantive values, ideals, or purposes on them. Liberals often contrast the aim of preventing acts that harm others with other – by their lights, illegitimate – aims that the law might have. In the most famous passage of *On Liberty* (1859: 68), Mill contrasts the harm principle with a paternalistic rationale for legal restrictions. For Mill, it is legitimate for the state to restrict a person's actions to prevent harm to *others*, but not, contra the paternalist, in order to prevent harm to the actor *herself*. In Hart's seminal defence of the harm principle in *Law, Liberty, and Morality* (1963), the harm-prevention rationale is contrasted with illegitimate moralistic rationales. Hart's liberal state legislates to prevent people harming others, not to punish or deter harmless immoralities. In Feinberg's

multi-volume work on the moral limits of the criminal law (1985a; 1985b; 1988), his defence of the harm principle is advanced alongside a sustained critique of paternalistic and moralistic rationales for legal restrictions on conduct.[11] The thematic thread that unifies these defences of the harm principle, and accounts for the contrasts with paternalism and moralism, is the individualistic notion that – to use a phrase from Kymlicka (2002b: 216) – people must be allowed to live their lives 'from the inside', according to their own values and purposes. Given this ideal, we want to ensure protection against all liberty-affecting incursions, whether from governments or from our neighbours. The purpose of the harm principle, and of its essential distinction between harms proper and sub-harmful adversity, is to track that ideal across its dual points of focus. We can define 'harms', then, as negative outcomes which are sufficiently detrimental to people's liberty that if the state *were* to coercively intervene for the sake of their prevention, it would still be acting in the interests of people's liberty overall. Of course the state's intervention itself constitutes an incursion against the liberty of certain individuals, namely, those whose conduct is being prevented or punished. But so long as it is carried out under the auspices of a rights-respecting legal system, the intervention is justifiable in the name of liberty, so liberals will characteristically say, since such interventions sustain a social environment in which there is a liberty-optimal balance for individuals overall.

Correlatively, if we classify an outcome as a sub-harmful type of adversity – if we say that something is *merely* annoying, or morally offensive, or that someone's preferences are *merely* disappointed – what we are saying, in effect, is that the outcome in question, unwelcome though it may be, is *not* sufficiently detrimental to the victim's liberty that the state, if it were to coercively intervene, would be acting in the service of people's liberty overall, insofar as the intervention itself would again involve an incursion against someone's liberty.[12]

---

[11] However imperfectly it is practiced in contemporary regimes, it is a nearly-universal *theoretical* tenet of liberalism that harmless immoralities should not be liable to legal restriction (see Feinberg 1988, 1994). Those who defend legal restrictions on harmless immorality may argue either that they are essential for our ongoing moral education over the course of our lives (Stephen 1874), or that moralistic laws serve to secure a society's survival (Devlin 1968), or that we all become moralists in extreme cases anyway, and that there is no principled way to differentiate the more extreme cases from everyday cases (Kristol 1994).

[12] Some liberals see the legitimacy of harm-prevention and the illegitimacy of moralism as jointly descended from a norm of moral neutrality, on which the state should not use the law to pursue substantive moral aims, and should instead only aspire to the (morally neutral) aim of harm-prevention. The harm principle can also

Given this characterisation, the concept of harm need not correspond with any singular, generic, or readily quantifiable feature of human experience. The concept is first and foremost a theoretical device, employed for a constructive moral purpose, in the kind of liberal regime that purports to guard against threats to people's liberty from two directions simultaneously: vertical (i.e. government-harming-citizen) and horizontal (i.e. neighbour-harming-neighbour).[13] There are different ways to translate this into a criterial definition of harm. Feinberg's definition emphasises interests. Harms do not include all hurts and damages, on his view, only setbacks to people's interests. And not any old setbacks to interest (those that result from fair competition or everyday conflict are excluded, for instance); "only setbacks to interests that are wrongs, and wrongs that are setbacks to interest, are to count as harms" (Feinberg 1985a: 36). If we dig deeper into Feinberg's account, we find that he defines wrongs as violations of the pre-legal moral rights that protect a sphere of individual liberty within which people can pursue their everyday activities and their broader projects and values. Thus, for Feinberg, harms ultimately involve negative treatment of both rights and (rights-

---

be defended, however, under a (morally non-neutral) liberal perfectionist framework. Even if, as perfectionists in general claim, there are no principled limits to the state's pursuit of moral aims under the law, those perfectionists who prize liberty especially can still argue that criminal prohibitions are so injurious to liberty, both in their first-order effects, and in the relationship of domination they express (and the social meanings they convey accordingly), that they are unacceptable as means for pursuing any purpose less urgent than the prevention of liberty-impairing harms themselves (see Raz 1986: 412-20).

[13] Ben Bradley (2012) argues that we should banish the concept of harm from serious moral discourse. His dissatisfaction with the concept owes to his judgement that it cannot be characterised in any way that is, amongst other things, axiologically neutral (i.e. does not presuppose any substantive theory of well-being), and not morally theory-laden (i.e. such that the judgement that an act φ is harmful does not automatically entail further moral judgements about φ, e.g. that φ is all-things-considered morally wrongful or vicious). Bradley may well be right to claim that the concept of harm cannot be defined in a way that satisfies these criteria. He is surprisingly confident, however, that other widely-employed moral concepts like utility and rights are readily characterisable in ways that satisfy these criteria. Unlike Bradley, I suspect that, given careful scrutiny, all of the concepts invoked in our moral discourse will turn out to be things that we read into the social world in view of our constructive moral purposes, rather than objectively discernible features of life and society that we uncover, and to which our moral judgements respond. But unlike Bradley again, I do not believe this renders our moral discourse 'unserious'. At any rate, Bradley's critique of the concept of harm rests on more controversial meta-ethical foundations than he acknowledges.

protected) interests. Some authors, like Thomas Nagel (1991: 139ff.), define harm in a way that simply identifies harm with certain kinds of rights-violations. Others look to supplant the harm principle with a direct appeal to the autonomy-based considerations that underwrite liberal conceptions of the harm principle. For instance, Arthur Ripstein (2006: 229) defends a 'sovereignty principle', under which "the only legitimate restrictions on conduct are those that secure the mutual independence of free persons from each other". The advantage of this approach is that it seems to make better sense of our prohibition of certain non-harmful actions – in particular, harmless trespasses – whose prohibition is widely accepted despite their harmlessness.[14] Conversely, the difficulty with this approach is that it is unclear what, if anything, could set limits on the aim of 'securing people's mutual independence', since even very minor incursions could be restricted for that purpose.[15] Each approach has certain benefits and costs. If we are prepared to compromise on parsimony, it may be that the best way to translate the purposes of the harm principle into a formal definition of harm is to draw on a combination of the elements I have mentioned here. For instance, we might say that an action is harmful only if it inflicts a *rights-violating* setback to someone's *sovereignty-essential* interests. The formulation would be unwieldy, but it conveys a fairly simple notion: that among our many interests, special protection is needed for those interests whose impairment would fundamentally threaten our capacity for autonomous self-government.

What we are unlikely to get, in any case, even from the most painstakingly engineered definition of harm, is a definition that will easily and uncontroversially resolve the difficult borderline cases. But it would be a mistake to expect that from our definitions in the first

---

[14] On Ripstein's view, non-harmful trespasses may be prohibited because they violate people's sovereignty over their property. In his discussion of structurally similar cases, Gardner argues that such acts may be prohibited under the auspices of a harm principle, not because they themselves are harmful but because criminalisation reduces their occurrence, and thereby shores up goods whose erosion in the absence of criminalisation would be harmful (Gardner 2007d: 30-31). Gardner's understanding of the harm involved in harmless trespass, in these respects, resembles Feinberg's account of public harm (see §3.1).

[15] As Bird (2007) argues, a sovereignty principle would be too weak if it only licensed intervention in cases of extreme sovereignty violation (e.g. unlawful imprisonment); but too strong if it licensed restrictions on any action that impaired others' sovereignty in any respect. Individual sovereignty may be an appropriate underlying aim or purpose of a system of legal rights and obligations, but it is not always a useful source of guidance when it comes to articulating the limiting conditions for legal restrictions on conduct.

place. For any case lying near the threshold between mental harm and mere annoyance, or between status harm and mere social friction, some sort of discretionary criterion will need to be called upon to arbitrate; e.g. what someone could *reasonably* be expected to endure, or what would not *excessively* impair someone's capacity for self-government. Such discretionary criteria – there to be interpreted by judges, juries, etc. – are a routine and ubiquitous part of the law's administration. The important thing for our purposes here is that, having identified the role that the concept of harm is supposed to fulfil, and surveyed different ways that the concept might be elaborated in that connection (e.g. in terms of interests, rights, or sovereignty), we have found no reason to favour a position that denies the significance of the adverse effects hate speech has on people. Of course, not *every* bad consequence associated with hate speech will qualify as a harm. But there is nothing trumped-up about the suggestion that many of them will. Certain adverse consequences of hate speech involve severe setbacks to individuals, arising from hostile incursions on the part of the hate-speaker. That is most apparent where the adverse consequences of hate speech involve material harms, but it is no less the case where the adverse consequences are mental/psychological or status-based.

## 3.4 Causation and responsibility

So, the adverse consequences of hate speech include physical and mental harms, and status harms can also be (provisionally) counted among these consequences, so far as (i) hate speech allegedly contributes to pervasive, identity-based social hierarchies, and (ii) many of the negative effects that arise out of those hierarchies can be construed as status harms. This brings us back to the responsibility question. Once these causal relations have been identified, under what circumstances are hate-speakers responsible for the harms in question, not just causally, but in the sense of being answerable or liable for them as well? As we shift focus to responsibility, we will have to negotiate ambiguities in the various senses of the key term 'responsible' and its cognates. Following Hart, we can distinguish four senses of responsibility. In some cases, to say person A was responsible with respect to state of affairs, x, is just to say that A caused x or contributed to x; e.g. we might say that the vandal was responsible for the broken letterbox, or that the baker was responsible for the dead mouse in the bread. Call this *causal responsibility* (Hart 2008c: 214-15). In other cases, to say A was responsible with respect

to x, is to say that so far as A had a hand in x's coming about, A's behaviour expressed various agential capacities, e.g. (i) to understand the normative demands to which she was subject vis-à-vis x, (ii) to deliberate in light of her understanding, and (iii) to control her x-related actions as guided by her deliberations. This is the sense of responsibility invoked when we say, for instance, that the child was not responsible for setting off the alarm, or that the sleepwalker was not responsible for unlocking the door.[16] Call this *capacity-responsibility* (Ibid: 227-30). In other cases, to say A is responsible with respect to some action, φ, is to say that A is obliged to φ in view of wrongdoing on her part, or in view of some misfortune for which she was accountable; e.g. we might say the thief is responsible for compensating her victim; or that the oil company is responsible for repairing the damaged ecosystem. Call this *liability-responsibility* (Ibid: 215-22).[17] Finally, in some cases to say A is responsible with respect to φ is to say that A has a duty to φ because of a social role she occupies or a commitment she has undertaken; e.g. we might say that the surgeon is responsible for seeking consent, or that the insurer is responsible for processing claims punctually. Call this *role-responsibility* (Ibid: 212-14).

Philosophical questions about responsibility in law can be more easily articulated once these different senses of responsibility have been distinguished. First, there are questions about the normative structure underlying the law's imposition of liability-responsibility, e.g. are such impositions supposed to promote good consequences or exact retribution? Second, there are metaphysical questions about the grounds of liability-responsibility, e.g. when and why is the recipient of punishment identifiable with the wrongdoer? And third, there are questions about how judgements involving liability-responsibility depend on judgements about other kinds of responsibility: (i) when does causal responsibility merit an ascription of liability-responsibility? (ii) When does failure to fulfil one's role-responsibilities justify an imposition of liability-responsibility? And (iii) what makes someone responsible *capacity*-wise

[16] In such cases, we are not denying that the child/sleepwalker was causally 'to blame' for the outcome; we are saying they are not *evaluatively* to blame for the outcome, because their causal role in bringing about that outcome was not agential in an appropriately blame-grounding manner.

[17] Liability-responsibility is similar to what Dworkin (2000: 287) calls *consequential* responsibility: A is consequentially responsible with respect to φ, if A is obliged to bear the burdensome consequences constituted by or related to φ. Ascriptions of causal and liability-responsibility may be run together in informal discourse. When someone says "BP is responsible for the oil spill", she may mean *both* that BP caused the spill and that they are obliged to repair it (see Oshana 1997: 81).

for her conduct, and when, precisely, does an ascription of liability-responsibility depend on this?[18] While a fully developed account of responsibility in law would say something about all these questions, here our concern is just with the circumstances under which an ascription of *causal* responsibility to the hate-speaker, for harms associated with his speech, is sufficient, in degree and kind, to justify an ascription of *liability*-responsibility to the hate-speaker. It is a familiar notion that any imposition of liability-responsibility for harmful conduct should be based on a cause-to-liability relationship. Most of us, as Judith Thomson says, share the conviction "that causation is necessary for liability: that the one or ones who caused the harm are the one or ones who must pay for it" (Thomson 1986b: 60). Call this, following Shelly Kagan (1986), an internalist view of liability. One way to defend internalism about liability is to consider the way in which its falsehood would undermine any putative general moral justification for imposing liability-responsibility on people. The law may purport to minimise harmful wrongdoing, or to exact retribution for it, or to distribute the social costs associated with it. But given any of these purposes (or any combination of them), allowing the 'causally innocent' to be punished confounds the aim of the law; for this does not afford just desserts

[18] Note the resemblance between this question and the perennial question in moral philosophy of whether praise and blame can be reconciled with a deterministic view of human action. The notion that we cannot praise or blame someone for his actions unless he exercises control in performing them traces back to Aristotle's *Nicomachean Ethics* (1110a-1111b4). Determinism threatens praise and blame, in short, because it suggests that our actions stem from things outside our control. Liability-ascriptions in law are informed to some degree by judgements about the limits of voluntary control, insofar as all legal systems recognise some individuals as lacking capacity-responsibility, and being immune from liability-ascription for that reason (for contemporary discussion see, e.g., Honoré 1995; 1998; Pettit 2001; M. Smith 2001; Gardner 2007c; Hart 2008a). However, the more radical concerns in moral philosophy – that human action may be determined such that we can never really be responsible for our actions – have little currency in legal contexts. Why should that be the case? One reason is that the law's imposition of liability-responsibility is more evidently compatible with determinism. According to Tony Honoré, "to judge people according to their general capacity neither rules out nor requires determinism. General capacities can be measured by how people generally perform when they try to execute a given type of action... Whether their choices and performances are determined by pre-existing factors has no obvious bearing on the fairness of judging them on this basis" (1999b: 38-39). Another reason may be that metaphysical objections to liability-ascription are self-defeating. If those who are punished should instead be held blameless, on metaphysical grounds, it presumably follows that those enacting the punishment should also be immune to blame, on the same metaphysical grounds.

to harmful wrongdoers, it does not fairly distribute the costs of harmful wrongdoing, and it is unlikely (in the long run) to deter harmful wrongdoing. These things speak in favour of the core internalist intuition Thomson identifies: that causation is *necessary* for liability. Our concerns extend further, though. What we want to identify are the circumstances in which causation is *sufficient* for liability: given that some causal connection obtains between a certain hate-speech-act and a certain harmful outcome (and assuming that the hate-speaker is capacity-wise responsible for her actions), when is it fair and fitting to make the hate-speaker answerable – to *hold* her responsible – for the harms associated with her speech?

Attributions of causation in law are commonly treated as having two elements. First, we can ask whether φ was a *cause-in-fact* of x. In this we are asking whether x's occurrence causally depended upon φ in any way at all. There are various ways to analyse causation-in-fact (see note 1 in §3.1), but these are of little consequence for our purposes here. If we grant that some of the adverse effects associated with hate speech are harms proper, then we are granting that some hate-speech-acts are part of the causal ancestry – that is to say, are causes-in-fact – of harms to persons. Our interest is in the second element of legal causation, namely, *proximate* causation. An action, φ, is adjudged a proximate cause of outcome x, when it contributes to x in such a way as to warrant an ascription of blame or liability to the person who performed φ, for the harmful consequences involved in x. As an example, suppose that person A insults person B, and that B, driving away in an agitated state, then knocks the rear-view mirror off person C's car. In this case A's speech act is quite clearly part of the causal ancestry of the damage to C's car; that is to say, A's speech is one of the causes-in-fact of the damage to C's car. At the same time, though, A's speech – even if ill-mannered, hurtful, gratuitous, etc. – is not where accountability for C's broken mirror ought to redound. The judgement that A is not an appropriate liability-bearer for the damage to C's car is expressed in the verdict that A's speech was not the *proximate* cause of that damage.

## 3.5 Proximate causation: striking a balance

Judgements about proximate causation require us to discriminate among the causes-in-fact of an outcome to identify the factors that are most salient in an explanation of how that outcome came about. One way to think about proximate causation, then, is to ask how causes-in-fact of

x may *lack* explanatorily salience with respect to x. Probably the simplest case, exemplified in the rear-view mirror example above, is when the putative cause's place in the events leading to x is temporally or sequentially remote from x. Relatively 'nearby' causes are, *ceteris paribus*, more salient in the explanation for an outcome than relatively 'distant' causes. One particularly notable species of remoteness – addressed in the legal doctrine of *novus actus interveniens*, or colloquially, 'breaking the chain' – is the case in which the sequence connecting the putative cause to the outcome is interrupted along the way by the intervening, voluntary action of another agent. A's $\varphi_1$-ing and B's $\varphi_2$-ing may both be causes-in-fact of x, but if B's $\varphi_2$-ing occurs after A's $\varphi_1$-ing, then, provided both actions are performed independently and voluntarily, B's $\varphi_2$-ing is more salient in an explanation of how x came about than A's $\varphi_1$-ing, and hence it 'screens off' A's $\varphi_1$-ing from being ascribed proximate causal responsibility in relation to x. B's $\varphi_2$-ing, as we might say, 'sets a new chain of causation going', independently of what has come before (G. Williams 1989: 392). The general heuristic is that wherever the causes-in-fact of x include multiple, non-simultaneous agential actions, we assume (defeasibly) that the temporally posterior action constitutes the proximate cause of x.

Another ground for limiting an ascription of proximate causation is evident in cases in which a certain condition, c, is, along with A's $\varphi$-ing, one of the causes-in-fact of x, but where c, because of its abnormality in the circumstances, has greater explanatory power with respect to x than A's $\varphi$-ing. Like intervening acts, abnormal conditions derive their explanatory power from being salient difference-makers in the world. As Tony Honoré (1999a: 6) says "conditions that are abnormal... have a prior claim to be selected as causes of an unexplained event". What they have in common with cases involving 'breaking the chain', he says "is that they are, or can be regarded as, interventions in the world. They bring about changes in the existing or expected state of affairs" (Ibid: 6). So far as judgements about proximate causation involve discrimination and selection, Honoré's claim is that if the causes-in-fact of x include abnormal conditions in the circumstances in which x came about, this is a *pro tanto* reason *against* ascribing proximate causal responsibility for x to nearby agential acts. The crucial question, however, is how we judge one of the cause-in-fact conditions of an outcome to be abnormal. In some instances our judgements about abnormality will just be statistical. The statistical unusualness of B's weak heart, for instance, may be sufficient to prevent A's verbal mockery of B from being deemed the proximate cause of B's heart attack. The heart attack

may not have occurred when it did but for A's mockery, but nor would it have occurred but for B's physiological ailment, and among the two causes-in-fact, the physiological defect is more out of the ordinary, and accordingly has more explanatory power. On the other hand, for cases in which the conditions among the causes-in-fact of x include the dispositions and attitudes of one or more agents involved in the causal chain of events leading up to x, our judgements about the normality of the relevant conditions may not be wholly descriptive – they may be partly evaluative as well. For instance, there will be cases in which an outcome x would not have transpired were it not for a certain individual's apathy, cowardice, laziness, meanness, gullibility, or oversensitivity. When we judge that one of these conditions was abnormal as a cause-in-fact of the outcome, and attribute explanatory power to it for that reason, it may be that the basis of our judgement is not the condition's statistical unusualness (we may not have any information about that matter, and even if we do, we may not be interested in it), but instead that the condition was abnormal evaluatively speaking – that regardless of whether individuals in these sorts of cases typically betray the dispositions that the individual in this case betrayed, they generally ought not to do so.

If we make judgements about normality in this way, we may worry that our ascriptions of causal responsibility abandon any claim to 'metaphysical purity'. To say that condition c bears proximate causal responsibility for x, on the approach that I am suggesting, would seemingly be to institute, in the guise of a judgement about causation, a policy specifying the sorts of dispositions people ought to have, and making people accountable for shortcomings in this respect. There is something correct in this complaint, but also something misleading. It is true that we institute substantive policy if we make our judgements about proximate causation hinge partly on evaluative judgements, and that this renders our judgements about causation 'metaphysically impure'. What is misleading is the suggestion that, as far as the law goes, things could ever really be otherwise. From the outset, a legal inquiry into the cause of any harmful outcome, x, seeks an explanation for x that is tailored to a restricted and evaluatively-oriented purpose. The law is interested, firstly, in the way that individuals (rather than, for example, microphysical processes, or family dynasties, or whole societies) contribute to adverse outcomes, and secondly, in when and how these contributions can be taken as grounds for an imposition of liability-responsibility, with a view to redressing (whether under a retributivist, or consequentialist, or distributive framework) those adverse

outcomes. This is, as Hart (1948) observed, evinced in the fact that we use terms that imply an element of decision – like 'ascribe', 'assign', or 'attribute' – to describe what is performed when verdicts are rendered, in legal contexts, about the causal bases of adverse outcomes, rather than terms like 'find out' or 'discover', which imply the mere observation of something already there. We might say that ascriptions of causal responsibility in law always proceed with one eye on evaluative questions about who has behaved well and poorly, and who should be liable to bear the costs that come with redressing harm accordingly.[19] When we judge that A's suggestibility or oversensitivity is abnormal, and that it has greater explanatory power with respect to x than B's remark to A, we are matching our ascription of x's proximate cause with our judgement that A, rather than B, is the more appropriate party to bear the burdens that come with x being redressed. This does not mean that judgements about causation in law are entirely based on evaluative judgements. In order to qualify as a proximate cause of x, an act or condition must, in the first place, be a cause-in-fact of x. And in selecting among causes-in-fact under the proximate cause criterion, judgements about 'breaking the chain', for instance, do not issue from an evaluative verdict. So person A may behave well and still be ascribed proximate causal responsibility for x, while B may behave appallingly without being ascribed proximate causal responsibility for x. Our aim, in selecting proximate causes, is to strike a balance. On one hand we want to do justice to a 'pure metaphysical' conception of what it means for an agent to bring about an outcome, as it applies to particular cases. On the other hand, where there is more than one metaphysically plausible candidate for an ascription of

[19] One may still worry about the use of causal language in the characterisation of judgements about who bears liability for harm. Jane Stapleton, for instance, argues that "the traditional mixing of factual and certain evaluative issues under the legal category of "causation"... is the root of doctrinal obscurities in the area" (2001: 183-84). When we judge that A's φ-ing, which numbers among the causes-in-fact of a harm, x, is sufficient to ground a liability-ascription to A, we should not, on Stapleton's view, describe A's φ-ing as the 'proximate cause' of x. Instead, we should candidly admit that A's φ-ing fares badly against some (extra-causal) evaluative criterion, and that this is what grounds the liability-ascription to A. Perhaps the reason why many theorists want to retain a causal characterisation for these judgements is to express some sort of commitment to harm-causation being the essential criterion for criminal liability. Where judgements about causation rest on further judgements about salience relative to the explananda, the choice to cast primarily evaluative judgements in causal terminology reflects the fact that, when the law explains harmful outcomes in order to found liability-ascriptions, harm-causation is its principal concern.

proximate causation, as is usually the case, we want to do justice to individuals. In redressing harm, we want the liabilities in question to fall upon those who have fallen short of a standard of conduct that we ought reasonably to be able to expect from each other.

## 3.6 Cognitive mediation and liability-ascription

What is the upshot of all this when it comes to assigning causal- and liability-responsibility for harms associated with hate speech? Suppose we judge that an act of hate speech, φ, is a cause-in-fact of some harmful outcome, x, in a simple counterfactual sense: but for φ, x would not have occurred. If φ is a capacity-wise responsible action, then, as an initial default judgement, it is liable to be adjudged the proximate cause of x. The question, then, is whether there is any defeater for this responsibility ascription: any reason to think that some other action, event, or preceding state of affairs would be a more fitting candidate for being judged the proximate cause of x. What complicates this question is the fact that the causal processes connecting acts of hate speech to harmful outcomes involve (in most cases) some form of mediating cognition on the part of the audience. As we observed in §3.2, when hate speech is used to harass, abuse, or threaten people, the psychological harms that may ensue, if they do indeed ensue, are not brought forth automatically upon the expressive debut of the words in question. The causal sequence connecting the speech to the harmful outcome (nearly) always includes some kind of intellectual, emotional, or doxastic response on the part of the audience.[20] Likewise, when hate speech is used to incite violence, the causal sequence connecting speech to harm includes various kinds of intellectual, emotional, or doxastic responses from the audience. Similarly again, when hate speech functions as an attack on the status of its targets, the causal connections that relate the speech acts to the harm include, among their innumerable mediating steps, various kinds of cognitive processing on the part of the third-party audience members exposed to the hate speech, or 'second-party' audience members targeted in the speech. In all cases, the harmful effects of hate speech depend on it eliciting a particular type of intellectual or emotional response from those to whom it is expressed.

[20] The 'nearly' here is meant to allow for cases like an individual (i) screaming in someone's ear, or (ii) creating a gratuitous nuisance, or (iii) vandalising another individual's property.

The fact that hate speech inflicts harm by acting on people's minds complicates our ascriptions of liability-responsibility in hate speech, because people are not, as a matter of course, completely passive and incapable of playing any agential role in their interactions with speakers, and in the further consequences of those interactions. We may sometimes be tempted to conceive of people's reactions to speech in mechanistic terms. We can imagine the content that is expressed in the speech act entering the audience member's consciousness via perception, acting upon her mind, which then reacts in turn (depending on its state at the moment of impact), in a reliable 'input-output' fashion, to determine the individual's response. If a tidied-up version of that were an acceptable characterisation of how speech affects people's minds, then we would not have any qualms about automatically ascribing proximate causal responsibility to the hate-speaker for any harm occurring in the wake of his speech. But the characterisation is patently untenable. In the normal run of cases, it is simply false that when others assert that p, they *make* us believe p, or *cause* us to φ in response to p. Rather, we are, and we regard ourselves as, the originators of our beliefs, attitudes, and actions. And this is rightly how the law regards us as well. Any attempt to make sense of holding people responsible for the outcomes of their actions must invest itself in the notion that people are – again, in the normal run of cases – centres of agency. If we are to be answerable for the consequences of any actions at all, then we cannot be viewed as mere flotsam and jetsam being swept around in a sea of events. The law cannot deny that people have an agential capacity to originate changes in the world, and yet retain its constitutive commitments to (i) imposing liability-responsibility upon people, and (ii) regarding its doing so, in-principle, as reasonable and just.[21] Among other things, this is what makes 'breaking the chain' more than

[21] The idea that a system of liability-ascription relies on, and in turn affirms, human agency, is at the heart of Honoré's work on responsibility in law. Honoré is especially concerned with how we justify strict-liability, in which liability is imposed not for fault, but (e.g.) for failure to meet an objective standard of care, or for risky activity which results in material damages (see 1999a: 8-9). Honoré thinks it is important that we are seen as the authors of our actions, since we would otherwise be unable to make sense of praising or rewarding success. It is a matter of taking the bad with the good, he says, to accept discredit for some bad but faultless outcomes of one's actions: "to bear the risk of bad luck is inseparable from being a choosing person" (1999b: 40). In a similar spirit, Gardner argues that the fundamental sense of responsibility in law is "the ability to explain oneself, to give an intelligible account of oneself, to answer for oneself, as a rational being" (2003: 161). We *hold* people responsible in law, Gardner says, as a way of investing in the idea that people *are*

just an arbitrary rule in judgements about proximate causes. Where there are two capacity-wise responsible agents, A and B, both performing blameworthy acts that are causally necessary precursors to x, and where B's contribution is enacted after A's, we cannot judge A's contribution to be the proximate cause of x without either demeaning B, by viewing her behaviour as a mechanistic response to input, or being inconsistent, by treating some agential acts, but not others, as genuine difference-makers in the world.

ଔ

The idea that I am trying to defend here, in short, is that people normally bear some measure of causal responsibility for the way that the expressive conduct to which they are exposed affects them, and hence, that they bear some liability-responsibility for what occurs in consequence.[22] This is most plausible, I expect, when it comes to incitement-type cases. The

---

agentially responsible – which is something to be prized, despite the risk of costly liability it carries (Ibid: 169-70). Both authors seek to reconcile two desiderata. On one hand we want to ensure that when someone is punished it is due to actions attributable to her, and not to events beyond her control. On the other hand, we want it to turn out that *some* actions *are ultimately* attributable to persons. The tension arises because the more we explore the role of circumstantial luck in liability-ascription, the more we see circumstances producing decisions that initially seem attributable to persons, so that people come to look less like actors *in* the world, and more like things acted on *by* the world (in moral psychology, this motivates 'situationist' critiques of character, e.g. Harman 1999; Doris 2002). The Honoré/Gardner response to this, roughly, is to say that if we expect unfavourable liabilities to be outweighed in the long-run by favourable liabilities, and by the inherent good of being counted as responsible, the second desideratum should override the first.

[22] The claim is not that people straightforwardly choose their beliefs, attitudes, desires, etc. Williams (1973a) and others have persuasively argued that our attitudes are in all important respects unchosen. But while there is an obvious sense in which that is true, there are also various compatibilist ways to identify people as having some accountability for their mental lives (see Nozick 1981, chapter 4 on the general conditions required for compatibilist accounts of liability). For instance, one can argue that mental involuntarism proves too much, and thus reduces to a banal fatalism (Stocker 1982). One can follow Mark Leon (2002) in arguing that we have a measure of indirect control over (and thus responsibility for) our mental lives, in our ability to engage in reflective thoughts that have a salutary effect on our attitudes. Alternatively, one may concede that we are involuntary in our mental lives, but deny that accountability here depends on voluntariness. Pamela Hieronymi (2008) argues that A can be responsible for involuntary attitudes if those attitudes embody what A would reflectively affirm as her 'take on the world'. Angela Smith (2005) argues that A is responsible for

person who performs the final harmful act is more liable for its effects than the speaker who provokes or encourages him. The suggestion is less plausible – and sometimes less palatable – in cases of psychological harm due to verbal assault. If we say the target of verbal abuse is responsible for the way another person's abuse affects her, we run the risk of victim-blaming. The suggestion is harder to grasp, one way or the other, in cases of indirectly-inflicted status harms. On one hand, it seems *prima facie* plausible that hate-speakers should not be answerable for harms which arise from the identity-prejudicial workings of society at large, given that 'society at large' is made up of individuals who, if they are influenced by hate speech, should remain answerable for how they are influenced. On the other hand, if the way hate speech works is more like conditioning people into identity-prejudicial attitudes, than persuading them, then perhaps the hate-speaker *is* a legitimate liability-bearer for the eventuating harm; although it is difficult to say what becomes of this prospect if everyone is generally ignorant about the conditioning processes to which hate speech (allegedly) contributes.

The different kinds of cases will throw up different problems. So far as there is a general answer to the question of when, in relation to hate speech, causal responsibility for harm is sufficient to ground liability-responsibility, the answer will have to be programmatic. The hate-speaker can be judged liable for the harm that occurs in consequence of his speech, if the performance of the relevant hate-speech-acts, under the circumstances (both local and general), could reasonably have been expected to have harmful consequences for others, given normal or reasonable reactions to those speech acts on the part of the speaker's audience. So why is it necessary to stress the idea that audiences are responsible? Because if, as I have claimed, liability-ascriptions for the harms of hate speech must be determined by judgements about how the hate-speaker's audience can be expected to react, *those* judgements have to be rendered in light of a conception of audiences as responsible actors, rather than mere passive patients, in their responses to hate speech. On this account, we are all entitled to expect that the people exposed to our speech can and should exercise some intellectual independence and mental resilience in the way they respond to it. The same is true for the hate-speaker. The fact

---

her attitudes if, in holding such attitudes, A is an appropriate target for certain kinds of demands, e.g. that she justify, modify, or abandon those attitudes. Views of the latter variety, in which responsibility is untethered from voluntariness, trail in the wake of P. F. Strawson's (1962) argument that being responsible primarily means being the proper target of other people's reactive attitudes (e.g. their gratitude or resentment).

that she is speaking in an identity-prejudicial idiom does not give her words supernatural powers to control other people. The hate-speaker becomes liable for the harmful outcomes that occur in consequence of her speech, however, when those outcomes could have been reasonably anticipated, even while allowing for a normal degree of intellectual independence and mental resilience being manifested in the reactions of her audience.

In cases where we are confident in judging that the speaker *is* liable for the harmful consequences of her speech, e.g. a case in which someone repeatedly makes explicit threats to another's safety, this confidence stems from the related judgement that anyone who faced such treatment could be expected to suffer mental harms (involving fear, anxiety, etc.) even while manifesting a reasonable degree of emotional resilience. Conversely, in cases where we are confident in saying that a speaker is *not* liable for the harmful effects of her speech, e.g. when A calls B "a stupid yuppie" or "a dumb Yank", this confidence derives from the related judgement that if B felt more than superficially wounded by such epithets, then B would be reacting hyper-sensitively, and as such, would be responsible for bearing the cost of managing those feelings himself.[23] The lesson, in relation to liability-ascriptions in anti-hate speech law, is that when hate speech is found to be among the causes-in-fact of harmful outcomes, we cannot arrive at the further conclusions – that hate speech is the proximate cause of those outcomes, and that the hate-speaker is legally liable for those outcomes – without first attending to the position and perspective of the individuals exposed to the relevant speech acts. Obviously the hate-speaker plays *some* part: that is entailed by his speech being among the causes-in-fact of the harm. However, it is only in considering what can be expected of audiences that a proper characterisation of the hate-speaker's role is evident.

## 3.7 Activity and passivity in offence

Before examining the details of responsibility-ascription in directly and indirectly harmful hate speech, I want to show how the account of audience responsibilities given here can address a tension in liberal theorising about the concept of offence. As we saw in §3.2, offence is one of several kinds of merely adverse outcomes (including irritation and thwarted preferences) in

[23] The example comes from Corlett and Francescotti (2002: 1092-93).

contrast with which harm is naturally defined. Where harm-prevention is the central justifying purpose for restrictions on conduct in a liberal legal system, offence is regarded as something milder, which – if it is a basis for restrictions on conduct at all – operates within a narrower framework and with further caveats.[24] A number of reasons might be given for thinking of offence in this way. Some authors emphasise the inevitability of offence in a diverse, liberty-respecting society. "Occasional psychic injury is an unavoidable side effect of vigorous public discourse in an open society" says Weinstein (1994: 206); "the essence of negative liberty is freedom to offend" says Dworkin (1995: 117). Others observe that offense can be a positive thing, either for the offended person, who benefits from her attitudes being put to the test (Waldron 1987), or for all of us, since people's being offended is "a welcome sign that the oxygen of debate and dissent still flows through society" (Geddis 2004: 874).[25] Others observe that, so far as it is a subjective matter, offence is often unreasonably taken (see Bayles 1973; von Hirsch 1985; 2000 Shoemaker 2000; Simester and von Hirsch 2002). Someone whose feelings of offence stem from a baseless prejudice – e.g. a dislike of garish outfits – has no real objection to the offending party's conduct (von Hirsch 2000: 86), and all the more so if the offensive thing embodies weighty interests held by the 'offending' party, as in cases where people are offended by, say, the sight of a mixed-ethnicity couple (Ellis 1984: 23).

The best-developed attempt to defend offence as a basis for legal restrictions on conduct, in Feinberg's *Offence to Others* (1985b), acknowledges the subjectivity and inevitability of offence, but still sees offence-causing behaviour as in-principle liable to legal restriction, on the same grounds as things (like loud noises or noxious odours) which create a nuisance for others.[26] For Feinberg, offensive acts may not be harmful, but they are burdensome and distracting: the sort of thing we find ourselves having to cope with, rather than just being able to ignore (1985b: 9). If we can restrict nuisances to secure conditions that are conducive to people effectively pursuing their goals and day-to-day activities, Feinberg argues, then we can

[24] Some authors deride anti-hate speech laws as attempts to protect minority groups from offence, under the nebulous pretext of 'cultural respect', e.g. Richards 1988; Alexander 1994; Kamm 2007; McNaron 2008.

[25] If offence *is* valuable in these ways, we may also endorse deliberate efforts to offend (see Vandeveer 1979).

[26] Feinberg is the notable outlier in defending the legitimacy of offence-based restrictions in a liberal regime, but his view does receive occasional support in the literature, e.g. Cohen-Almagor 1993; Peard 2004.

restrict offence-causing activities for the same reason – subject, of course, to some limiting constraints which may override the case for restriction in any given instance.[27]

A problem with Feinberg's view is that it seems to accommodate indiscriminately the sensitivities of people who take offence unreasonably. Although he ultimately agrees that we should not use the law to protect the feelings of people horrified by garish clothing – or more seriously, by the sight of a gay couple in public – his route to this conclusion is disconcertingly timid. On Feinberg's view, the offence taken by bigoted and oversensitive people is always a *pro tanto* reason for restricting garish clothes and gay people kissing. The law should remain uninvolved, he thinks, only because other considerations (e.g. the legitimate interests of the people whose activities would be curtailed, or the fact that no malice or insult is intended in their actions) override the case for restriction. But why treat offence stemming from bigotry or oversensitivity even as a *pro tanto* reason for restricting people's conduct? Judith Thomson takes this line of objection even further. While some psychological malaises are relatively visceral (e.g. nausea or dizziness), others – those she calls 'belief-mediated distresses' – only occur due to the subject's beliefs or attitudes. Thomson argues that we do not have *any* claim against others that they not cause us feelings of this sort (1990: 253-57).[28] Her reasoning is similar to the rationale we see in theorising about offence. People may feel distress over all sorts of strange things, but if we legislate to protect people against such feelings, everyone's lives will become constrained by the delicate or overbearing sensibilities of the few. So the tension is this: on one hand, as Feinberg emphasises, offence – *gratuitous* offence, at least – is

---

[27] Feinberg motivates his 'offence-as-nuisance' analysis with a series of stories about offensive behaviour on a public bus. The reader is asked to imagine that while travelling to an appointment, nearby passengers drool, belch, and fart (story 5), consume vomit or faeces (stories 7-8), mutilate a corpse (story 10), masturbate (story 16), or fellate a dog (story 23) (1985b: 10-13). On Feinberg's view, such acts are ultimately harmless, but they are liable to restriction all the same because they "commandeer one's attention from the outside, forcing one to relinquish control of one's inner state, and drop what one was doing in order to cope" (1985b: 22).

[28] Thomson makes an exception, though, for cases in which such feelings are instrumental in some further harm. I do not have a claim against you causing me affront *per se*, but I *do* have a claim against you doing so while I am up a ladder, to try to make me fall off. A view related to Thomson's is that where harmful effects are due to the harmed party's belief that the offending action is wrong, they cease to legitimately ground legal restrictions. C. L. Ten (1980: 56-61) identifies this idea as a component of Mill's harm principle, although his reading on this point is disputed by Honderich (1982) and Gordon (1984).

something we have a legitimate interest in minimising, given its nuisance-like character. Whether we think of offence as a type of low-key harm or as something different to harm is not the issue. What matters is that there are *pro tanto* good reasons for liberals to try to minimise these kinds of nuisances.[29] On the other hand, some people 'cry offence' and suffer indignation all too easily. If, like Feinberg, we treat all offence as a *pro tanto* legitimate concern in the law, then the calculations on which policy decisions are based seemingly cannot but be influenced to some extent by these pernicious tendencies – and this is surely not what we want from a legal system that purports to operate in the service of liberty.

I think the tension can be resolved if we emphasise responsibility for offence. Two separate questions are conflated in the dispute between Feinberg and his critics. There is a question about the gravity of offence: is it 'bad' enough to be the law's concern? But there is also a question about the origins of offence: when is the person whose conduct instigates offence actually *responsible* (causally or liability-wise) for that offence? Thomson's claims about belief-mediated distress are more pertinent to the latter question; her view, in essence, is that A is not answerable for what occurs due to B's unusual sensibilities. If that is how we approach things, we don't have to say that offence only 'counts' if it is reasonably taken. We can follow Feinberg in saying that all feelings of offence may, in principle, admit of legal remedy. We accommodate the concerns of Thomson and others, however, by requiring that no restrictions be levied on a person whose conduct instigates feelings of offence, unless that person is, in addition to being the instigator of those feelings, also responsible for them in the wider sense. This does not require us to say that people have a duty to not be offended by others' actions (contra Barrow 2005), or that people with delicate sensibilities choose to have them in any sense. The claim is about how offence-related liabilities ought to be distributed: those who find themselves with unusual sensibilities are liable to deal with the unwelcome

[29] Can offence be treated as a type of harm? Thomson (1986a) argues that the anti-nuisance rationale that grounds Feinberg's case for offence-based restrictions can be subsumed under the harm principle. Feinberg resists that approach, but his stated reason – that offensive nuisances do no lasting damage – is peculiar, since, by his own account, harm consists in rights-violating setback to interest, rather than enduring damage. In Feinberg's defence, one might suggest that it is essential to the concept of harm that one can be harmed without realising or caring that this is the case (see Hanser 2008), and that neither condition holds in relation to offence. In recent work, Thomson (2011) follows Hanser in treating harm as something isolatable from subjective elements, and in this she provides support for Feinberg's view against her own earlier criticisms.

consequences that follow from this. Attending to audiences' responsibilities in our thinking about the legal status of offence thus has similar ramifications to an emphasis on audience-responsibility in relation to the psychological harms associated with hate speech. It means that when we are specifying the conditions under which the speaker/offender, A, is liable to having his conduct restricted, we cannot construe the audience as passive patients, while thinking of A as the lone agential actor. So far as they underwrite liability-ascriptions, judgements about causal responsibility for offence-causing conduct in general, as with judgements about causal responsibility for the harms of hate speech in particular, should be guided by a recognition of the agential participation of both actor and audience.

# ☙ 4 ❧

# Directly harmful hate speech

In this chapter I elaborate the case for restrictions on directly harmful hate speech, that is to say, hate speech which is causally non-redundant in harmful outcomes inflicted upon specific individuals. The contrast, to reiterate from §3.1, is with indirectly harmful hate speech, in which hate speech merely contributes to a harmful outcome. Restrictions specifically aimed at directly harmful hate speech, as I understand them, have a narrower scope than general prohibitions on hate speech. They establish avenues through which hate-speakers can be made liable, case-by-case, for the harmful effects of their speech. In principle, such restrictions may be instituted either under criminal law or civil law. Criminal prohibitions are appropriate as far as the negative outcomes go (after all, the aim is harm-prevention, rather than merely settling disputes). However, it may be that there are certain practical difficulties that come with criminalisation in this arena (in policing, for example), and I take my claims about responsibility-ascription in directly harmful hate to apply similarly with respect to the prospect of treating directly harmful hate speech as a basis for civil legal action.

As I argued in my discussion of free speech in chapter 2, legal restrictions on hate speech must, on pain of illegitimacy, operate – both in *de jure* purpose and *de facto* practice – as harm-prevention tools, rather than tools for suppressing identity-prejudicial beliefs. Where this condition is satisfied, restrictions on directly harmful hate speech are surely justifiable in principle. The rationale is simple. States can justifiably restrict harmful conduct; directly harmful hate speech is harmful conduct; therefore, states can justifiably restrict it. But we have to proceed cautiously from this syllogism. For one thing, as I observed in §§3.2-3.3, hate speech will often have effects that are *sub*-harmful (e.g. merely offensive), and our laws will have to be formulated so that speakers cannot be punished in view of those consequences (especially since the punishments themselves will be harmful). There should be no legal intervention here without harm. Secondly, and this is the consideration that I stressed in §§3.4-3.6, even when there *is* harm, it is (nearly) always a further question whether the hate-

speaker should be deemed answerable for it, given the agential participation of others in the sequence of events through which the harm transpires. In §3.6 I proposed a rule of thumb for ascribing liability to hate-speakers for the harm causally associated with their speech. Speakers can be adjudged liable for harm, I said, only if the performance of the relevant speech acts under the circumstances could have reasonably been expected to have harmful consequences for others, given normal or reasonable reactions to those speech acts on the part of the audience. The question remains, however, as to how this liability-ascription principle should operate in practice. In what follows here, I argue that liability for the direct harms of hate speech should redound to hate-speakers in a relatively broad range of cases – broader, at least, than what we would have if we treated hate speech the same way we typically treat other directly harmful kinds of speech. In short, this is because of the way the identity-prejudicial content conveyed in hate speech interacts with the identity-prejudicial attitudes, practices, and conventions that exist in the context where hate speech is spoken and received. Because of this interaction, even some relatively isolated expressions of hostility in hate speech can be expected to harm the people against whom they are directed. In §§4.1–4.3 I show how this is the case in relation to hate-speech-based verbal assaults. In §§4.4–4.6 I show how this it the case for hate-speech-based incitements. In §4.7 argue in favour of treating directly-harmful forms of hate speech as distinct offences governed by dedicated statutes, rather than treating them as instances of more general types of harmful verbal conduct, which are only liable to restriction under the auspices of more broadly-formulated statutes.

## 4.1 Responsibility-ascription in verbal assault

As I said in §3.1, hate-speech-based verbal assaults are cases in which hate speech is (i) addressed to people targeted in the locutionary content of the speech (i.e. against whom contempt is expressed), and (ii) used to perform the illocutionary act of harassing or threatening those people, or to bring about perlocutionary effects like making them feel anxious, distressed, or humiliated. To be clear, the harms here all involve negative mental states. I am not too concerned with how we classify sub-species of verbal assault, beyond an elementary distinction between threat-type cases, in which the negative mental effects relate to an expectation of future harm being inflicted, and harassment-type cases, in which the

negative effects do not relate to expectations of future harm.[1] Obviously verbal assaults of either kind need not only be carried out using hate speech: people can threaten or harass without engaging in hate speech. Our interest is in responsibility-ascription in the hate-speech-based cases in particular, but instead of homing in on those cases immediately, it will be useful to begin by sketching guidelines for responsibility-ascription in verbal assaults generally. We can then consider how those guidelines might be adapted or applied in hate speech-based verbal assaults specifically, given their distinctive features.

So, let's consider some (non-hate-speech-based) verbal assault scenarios in which the speaker seems clearly morally culpable – and, *ceteris paribus*, legally liable – for the harm that results. Consider (i) a stalker who constantly calls his victim to leave sinister messages, or (ii) an office sociopath, who cruelly insults her co-worker on a daily basis. In the language of our responsibility-ascription principle from §3.6, this conduct patently would be expected to have harmful consequences for the targeted individuals, given normal and reasonable reactions to the speech. The question is: so far as that diagnosis seems obvious in these cases, what features of the cases account for this? The first important consideration in such cases is that they involve the targeting of a specific individual. Occasional disputes and incidental hostilities with the people we interact with are inevitable, and most of us – being resigned to this – will endure such encounters with only minor distress. If someone yells at me in traffic, or if am rudely addressed by a work colleague, I can attribute this to someone having a bad day, or to adverse circumstances bringing out the worst in people. By contrast, when I face verbal hostility directed at me in a deliberate and calculated way, rather than spontaneously and indiscriminately, I will reasonably feel anxious about this. I will wonder what, if anything, I can do to avoid the abuse. The victims of the stalker and sociopath in our examples may reasonably infer that, having been victimised once, they are likely to be victimised again, and this may lead to an anticipatory distress. Compounding these effects, in both of these cases, is the repetition and persistence in the use of verbal assaults. This confirms the targets' expectation that their regular activities can and will be subjected to hostile interruptions. The

[1] Kinney (1994) presents an empirically-informed taxonomy of different types of verbal aggression. Note that verbal assault is not supposed to be thought of as a sub-species of assault *simpliciter*. Assault is standardly defined as an act causing apprehension of harmful contact between assailant and target. Harassment-type instances of verbal assault need not generate any such apprehension.

momentary distresses are elicited by the aggression conveyed in the speech acts themselves, but they also lead to abiding distress, grounded in an anticipation of subsequent episodes of momentary distress, occurring unpredictably but consistently.

These are harassment-type cases; what about threat-type cases? We should note, firstly, that threatening speech is not always credible. If I doubt that the speaker really intends harm, or has the ability to inflict the harm he intends, then his speech will be more like a harassment-type case of verbal assault: it is liable to be harmful if and when, through targeting and repetition, it creates an abiding feeling of dread or anxious anticipation – not that the threat itself will be carried out, but that I will continue to be berated and belittled with the (idly) threatening language. When threatening speech *is* credible, however, it can exact its toll even in one-off uses. In such cases, the speaker is most clearly liable for harming his target when the threatened action promises to be severely harmful, and when anticipation of that harm is forced onto the target. Suppose the ringleader of a group of thugs on a train says to someone travelling alone: "I'm going to kick your ass". Being outnumbered, the target knows he may be badly injured if attacked, and being unable to absent himself from the altercation, he cannot help but anticipate the imminent possibility of that injury. The mental harm will still be inflicted, in such a scenario, even if the threatening speech turns out to be idle. Similarly to harassment-type cases, there is a momentarily negative state of mind, which consists in fear, humiliation, and so on, as well as a more enduring setback to the victim's equanimity. The victim's future experiences in travelling alone are liable to be coloured by the memory of fear and humiliation and the anticipation of a similar situation recurring. And the threatening speaker is responsible for all this, insofar as, given the context of his speech, he forces the anticipation that leads to these feelings for the target in the initial situation.

These claims about when speakers are responsible for harm in verbal assault express a view of social intercourse that I will call 'the simple view'. The simple view has at its core a normative claim: that we are all entitled to pursue our everyday business without being subjected to malicious interference from others. This is not an alternative to the harm principle, as much as a (partial) elaboration of it. Malicious interference need not involve a tactile interaction with the target of one's malice; it can also mean simply inserting oneself into the target's conscious awareness in a way that makes one's malice evident. Granted, we are all entitled to *feel* malice towards others. But we are not concomitantly entitled to confront people

with our malice in a way that will make them anxious about their safety, status, or well-being. This normative claim is founded on the same considerations that underwrite the harm principle generally: malicious incursions, even if they are merely verbal, are liable to generate anxiety and distress of a sufficiently severe kind to impair individual self-government, the protection of which is the central pillar of liberal political morality. Where the first part of the simple view says we are entitled to be free from malicious incursions, the second part admits that some degree of low-key malice will inevitably arise in a society which manages to secure conditions that are conducive to autonomous living. Free people feel upset, get into quarrels, vent their frustrations, and speak their minds. According to the simple view, these behaviours are ordinary and permissible, and mature individuals have to negotiate interactions involving these behaviours with whatever aplomb they are able to muster. There may be some who find routine low-key hostility completely unbearable. However, *ceteris paribus*, it is the responsibility of those people to manage the psychological travails that arise for them in consequence of this, and not the responsibility of everyone else to 'walk on eggshells' throughout all corners of social intercourse.

The third part of the simple view is a set of claims about what it takes for someone's behaviour to cross the threshold from being merely a manifestation of routine social friction, to being a malicious incursion, such that the aggressor, rather than the target, is responsible for the mental disturbance that ensues. In harassment-type cases, repetition and targeting are characteristic of (and *pro tanto* reasons to judge a speech act as being) malicious interference rather than routine hostility. In threat-type cases, any *credible* threat may be judged as falling on the malicious interference side of the distinction, and the more so to the extent that (i) the threatened harm is severe, and (ii) the apprehension of the threatened harm is unavoidable in the context. In both kinds of cases, the crossing of the threshold is a matter of acts being performed which conspicuously manifest the actor's malicious disposition towards the target, either implicitly, or explicitly as in the case of an overt threat. The simple view doesn't provide a fine-grained algorithm for sorting cases across this distinction. What it provides are discretionarily-applied criteria for sorting cases, grounded in a general account of how social intercourse works.[2] The account is one that is meant to consort with my claims about

[2] Here I am trying to do something similar to what Feinberg does in *Offence to Others*, in his discussion of the mediating maxims for determining the circumstances under which offence warrants legal intervention.

responsibility-ascription in §§3.4–3.6. Audiences are not imagined as inactive patients in the face of the speech they encounter. Their agential capacities as listeners are recognised, and they are expected to show resilience in the face of routine social friction accordingly. The speaker becomes answerable for the mental adversities that his audience experiences, only when his verbal conduct manifests malicious intentions, the very manifestation of which constitutes, *ipso facto*, an unacceptable incursion against the audience member's rightful sphere of self-government.

## 4.2 What's different in hate speech?

How might this simple view of social intercourse guide our responsibility-ascriptions for harm in hate-speech-based verbal assaults specifically? Let's begin, again, by considering some cases. Below are four examples of hate-speech-based verbal assault drawn from the literature.[3]

> A. An African-American worker found himself repeatedly subjected to racist speech when he came to work. A noose was hanging one day in his work area. "KKK" references were directed at him, as well as other... typical racist slurs... His employers discouraged him from calling the police, attributing the incidents to "horseplay"... The employee was subjected to repeated hate messages including "Get out nigger, you ain't wanted here", "Nigger Ben, KKK", [and]

---

"When standards conflict and none apply in a pre-emptive way" he says, "there will be no automatic mathematical way of coming to a clearly correct decision. The theorist can identify the factors that must be considered and compared, but, in the end, there is no substitute for judgement" (1985b: 45-46).

[3] There are a number of contributions to the literature which describe first-hand experiences of hate-speech-based verbal assault, some drawing on the author's own experiences, others drawing on second-hand accounts and courtroom testimony. Notable examples include Downs 1985; Love 1990; Delgado 1993; F. M. Lawrence 1993; Barnes 1995; Lawrence 1995; Lederer 1995; Mahoney 1995, 1996; Leets 2002; Delgado and Stefancic 2004; Eastman 2007; Bakircioglu 2008; and Gilreath 2009. Of course first-hand testimony and anecdotes cannot be taken as automatically veridical evidence for legal purposes. However, as is typical in the literature, the aim here is just to formulate a picture of the kinds of harm-inflicting causal processes in which hate speech is implicated.

"KKK for you, Ben"... Repeated threats over a nine-year period resulted in a diagnosis of hypertension and depression. (Matsuda 1993: 2327-28)[4]

B. Michael told a story of being called "faggot" by a man on a subway... He found himself in a state of semishock, nauseous, dizzy, unable to muster the witty, sarcastic, articulate rejoinder he was accustomed to making. He was instantly aware of the recent spate of gay bashing in San Francisco and that many of these incidents had escalated from verbal encounters... he realized that any response was inadequate to counter the hundreds of years of societal defamation that one word – "faggot" – carried with it. (C. R. Lawrence 1993: 70)[5]

C. [The victim] was a twenty-two year old Mexican national who had been residing legally in the United States... She was the director of the Senior Citizens Program in the Village of Central, New Mexico. The defendant was a white member of the Board of Trustees of Central that governed the village and supervised the Senior Citizens Program. During a public meeting of the Village Trustees, the defendant learned... that the plaintiff was a Mexican, not a Spanish American... [He] asked the plaintiff whether she had applied for citizenship and whether she had registered to vote in the United States. When she declined to answer these questions, he became outraged and said: "I feel... that anybody who works for the City of Central should be a citizen and especially if she's over five American citizens. I think one of them should be the director." The Village Trustees then went into a closed executive session. The trustees asked the plaintiff and others to leave the meeting room. The plaintiff sat outside the room and heard the defendant shout: "She's a Mexican. You see her? She's a Mexican. That's not fair. There's no way. Why should she be employed? She's a Mexican." (Love 1990: 134)[6]

---

[4] The example is from *Citchen v. Firestone Steel Products Co.* [1984] Nos. 12190-EM; 15389-EM Michigan Civil Rights Commission 1984. The victim was eventually awarded damages of US$1.5 million.

[5] This example was related to the author by one of his law school students, after the author had asked his class to describe their first-hand experiences of witnessing or being subjected to hate speech.

[6] The example is from *Dominguez v. Stone* [1981] NM 211 638 P.2d 423 Ct App 1981, in which a New Mexico appellate court reversed an earlier trial court's judgement in favour of the defendant (see Acosta 1989: 447-48; Love 1990: 134-35).

D. Phillip and Renee Smith, a mixed-race couple, agreed to purchase a house. While they were visiting the premises, their future next-door neighbour, a man named Talley, built a 4-foot tall cross in his own yard and set it on fire. The neighbour complained to bystanders that "having niggers next door" would ruin his property value. (Nockleby 1994: 677).[7]

Case A has all the features of malicious incursion as understood in the simple view of social intercourse. The verbal assaults were persistent ('repeated threats over a nine year period'), clearly targeted the victim (many of them were addressed to him by name), and while they weren't as overtly threatening as they might have been, references to the KKK (for instance) powerfully suggest the thought of violent action. Even a very resilient person subjected to verbally assaultive speech in this pattern could be expected to suffer anxiety and distress. In such a case, then, those who performed the hate-speech-based verbal assaults would rightly be legally accountable, in principle, for the harm suffered by the victim.

It's not clear that the same can be said about the other cases, however, so far as our judgements are guided by the simple view. Firstly, in each case the assaultive speech occurred in a one-off encounter. Secondly, to the extent that the speech in these cases might be received as threatening, the threats were veiled behind ambiguous epithets and symbols (ambiguous not in their hostility *per se*, but in whether that hostility contained violent intentions). Thirdly, in none of the cases did the speakers target the victims in a way that betrayed malice towards the victims *qua* individuals. In each case the malice was directed toward a social group; the victims were, as we might say, 'caught in the crossfire' of the speakers' malice. Indeed, unlike case A, in which the verbal assaults occurred as part of an ongoing relationship between speakers and victim, there is little to suggest that the speakers in cases B, C, and D even knew their victims. These verbal assaults were manifestations of *indiscriminate* prejudice, whose associated feelings of aggression were expressed toward specific individuals ostensibly due to accidents of social circumstance.

---

[7] The example is from *State v. Talley* [1993] 122 WA 2d 192 858 P.2d 217 S. Ct 1993, a criminal harassment case under the Washington Criminal Code, under which criminal harassment is defined as words or conduct performed with the intent to intimidate and which places the target 'in reasonable fear of harm to his person or property'. Despite the fact that the code specifies cross-burning as a particular form that such harassment may take, the court in this case dismissed all charges against Talley.

In cases like B, C, and D, then, the simple view suggests that if the targets suffered mental harm, they themselves were responsible for this, since the verbal assaults were neither repeated, nor narrowly targeted, nor overtly threatening. And yet it only takes a little empathic effort to see that the victims of these verbal assaults may have reasonably reacted to their experiences with the sort of distress that the simple view associates with malicious interference. The simple view imagines social intercourse playing out in a setting in which we can all safely presume that others do *not* have malicious intentions towards us, until they demonstrate otherwise, either by overtly communicating such intentions in their speech, or by executing a pattern of behaviour that transparently betrays such intentions. However, in societies where identity-prejudicial attitudes govern some people's actions, and osmotically influence the actions of many more, that is clearly not something that everyone can safely presume. The simple view is *over*-simple, then, in at least two ways. Firstly, it does not recognise that a reasonable judgement, about whether A's speech betrays intentional malice towards B, is partly determined by the prior likelihood of a randomly-chosen person harbouring malicious intentions towards B. If people in B's social group are held in contempt by 50% of the population, then it is rational for B to impute malevolent intentions to A, after A makes an isolated hostile remark to B, in a way that is not rational for person C to do in a parallel scenario, where C's social group is hated by only 1% of the population. More generally, it is rational to pay attention to existing patterns of malevolence in the wider community, when trying to ascertain where routine social friction ends and malicious incursion begins. Secondly, the simple view assumes that if no individual's speech, in isolation, betrays malicious intentions towards person B, then any verbal aggression B faces must be attributed to routine social friction. The simple view thus ignores the way in which patterns of conduct constituted by multiple people's independent verbal treatment of B may reasonably be taken as evidence (by B or others) that some of B's neighbours harbour malevolent intentions towards her. If B is met with verbal hostility in 50% of her one-off interactions with strangers, and if she sees that others face verbal hostility in only 1% of their one-off interactions, it would surely be irrational for B *not* to infer that others in her community harbour ill-will towards her. More generally, the point is that it is reasonable to attend to patterns of aggression in the behaviour of many people, not just patterns in the behaviour of individuals, when judging where routine social friction becomes malicious incursion.

The simple view of social intercourse needs to be replaced with a more sophisticated view, one which recognises that the reasonableness of how person A reacts to verbal assaults is partly determined by background facts about the identity-based hostility that A experiences and observes in her relations with others. Where identity-prejudicial attitudes and patterns of behaviour condition and suffuse people's experience of social intercourse, even isolated and indiscriminate hate-speech-based verbal assaults can be expected, given normal and reasonable reactions, to have harmful consequences for their targets.

## 4.3 Mechanisms of harm in verbally-assaultive hate speech

We can postulate a variety of specific processes or mechanisms through which these harmful effects may be achieved in verbally-assaultive hate speech. One is 'threatening implicature'.[8] Even when hate speech does not explicate the complex intention that is constitutive of a true threat,[9] it can convey an implicated suggestion of threat which, in light of the target's relative vulnerability, serves to generate the same kind of negative feelings that characteristically accompany true threats (see Tsesis 2004: 390ff.; Wiener and Richter 2008).[10] In linguistic pragmatics, implicatures are normally defined as ideas or propositions that are conveyed in speech indirectly, via the communication of some other idea or proposition (Davis 2010: §§1-2). Suppose, for example, that A asks whether B's blind date was good-looking, and B replies "well, he had a nice personality". Here the content of B's statement is distinct from, and

[8] Richard Delgado says that because the harms of hate speech are "embedded and ordinary" in our society, they are "virtually invisible" (1998: 878). The remark concerns the mere *existence* of harm, but it applies similarly to judgements about responsibility for harm. Where people in certain social groups routinely endure hate-speech-based verbal assaults, it may be difficult – once the cumulative negative effect of those assaults becomes visible as an individualised harm – for us to recognise the speaker's contribution to that harm.

[9] In true threats: (i) speakers deliberately convey intentions to harm the target, and (ii) reasonable listeners would interpret the speech as conveying such intentions, see Greenawalt 1989a: 90-94; Knechtle 2005: 573.

[10] John Nockleby stresses this point in his discussion of the cross-burning in case D. "Whatever the message is" he says, "it cannot be understood apart from the history of burning crosses in our culture. The burning cross – especially in front of a black family's house – has been used to intimidate blacks for more than a century... [it] has often been followed by a lynching. Why should a black or interracial family risk physical harm just to live in a neighbourhood of their choosing?" (1994: 678-79).

indeed doesn't even entail, the meaning that B purports to convey, namely, that the date was unattractive. Nevertheless, given the conversational context of B's speech, this meaning is successfully conveyed. The standard account of implicature, due to H. P. Grice, distinguishes two forms. *Conventional* implicatures use either the semantic content of what is uttered,[11] or established regularities of linguistic use and expectation,[12] to convey ideas and propositions without explicating them. *Conversational* implicatures, by contrast, trade on our typical expectations about linguistic cooperativeness in order to communicate meanings indirectly. In the above example B is asked whether her date was attractive. If B is a cooperative conversationalist – if she aims to be relevant, sincere, and informative – then B will say that her date was good looking if he indeed was. By saying that her date had a nice personality, B implicates that he was unattractive, since this being the case is the simplest way to interpret her apparent non-sequitur such that what she is conveying constitutes a relevant, informative contribution to the conversation as it stands (Grice 1989: 26-31).

How might this phenomenon work in hate-speech-based verbal assaults? Both types of implicature may be in effect here. In a case of conventional implicature, when A says to a stranger, B, on public transport "go back to your own country, you f***ing Arab terrorist", A obviously does not, in that phrase, issue an explicit threat of violence. But a threat may be implicated via the utterance nevertheless. According to our conventions of social expectation pertaining to an aggressive demand using this sort of language, there are very few reasons – basically, either: (i) for the sake of an over-the-top joke, or (ii) in order to intimidate – for which a person might say such a thing. Thus, if the context renders implausible interpretations on which the expression does *not* aim at intimidation, it is reasonable for B, whatever A's actual intentions might have been, to interpret A's expression as containing an (implicated) threat. An example of conversational implicature in hate-speech-based verbal assault is a

---

[11] E.g. if A utters S1: 'B is very tidy for a man', she implicates S2: 'men are untidy'; S1 would be a non-sequitur but for S2's truth, and so by uttering the S1, the content of S2 is also conveyed (Grice 1989: 25).

[12] E.g. if A says 'B is meeting a woman tonight' she implicates that B is meeting someone other than his mother or sister, even though this is neither logically entailed nor necessarily conversationally implicated by what A says. The relevant information is implicated, rather, by virtue of a communicative convention such that "meeting a woman" is generally understood to mean "rendezvousing with a female adult for a romantic liaison" instead of "meeting any female adult" (Grice 1989: 37).

scenario like case D, in which A says B should not hold a job because B is Mexican. The only way to read A's remark "but B is a *Mexican*", on which A is being informative and relevant in a discussion of B's fitness for the job, is if we suppose that B's being Mexican entails, by A's lights, that B has the negative traits associated with derogatory Mexican stereotypes.[13] Given the content of the statement in the discursive context, and assuming (as we do) conversational cooperativity on A's behalf, the imputation of these negative traits to B is conversationally implicated, which is to say, is conveyed without explicit statement.[14]

A second, related, process through which hate-speech-based verbal assaults may inflict harm, despite being isolated and not overtly threatening, is by calling on the affective force of identity-based epithets and slurs. Waldron says slurs intimate "the desirability of returning to a time when members of a racial minority were kept in their place by terrorizing threats" (2012: 118). I take it that this generalises to slurs based on all forms of identity-prejudice. When someone addresses a stranger on public transport as "faggot" (as in case B), he does not explicitly threaten violence. And yet violent intentions can still be conveyed, because, again, according to conventions of use and expectation, there is a narrow range of reasons for which one person might address a stranger as "faggot" publicly and face-to-face, and the context rules out the non-threatening interpretations. The main concern in recent philosophical work on slurs (e.g. Hornsby 2001; Hom 2008; Croom 2011) is to give an account of their semantic character, the primary question being whether or not their derogatory utility is due to their semantics. Williamson (2009) argues that slurs have the same referential properties as their non-abusive counterparts, but that they conventionally implicate other (false and abusive) sentences, beyond their (descriptive) categorisations.[15] Anderson and

---

[13] In discussing this case, Jean Love (1990) suggests that in the local cultural context of the utterance (in the US state of New Mexico) 'Mexican' in fact *conventionally* implicates these derogatory meanings.

[14] Again, in cases like these, it may be true that the hate speaker who implicates a threat does not *intend* to issue a threat. But that fact by itself does not entail that a threat is not issued. As Saul (2002) explains, what is essential to a Gricean conception of implicature is not content being *intentionally* transmitted to the audience (without being explicated), but rather, content being *made available* to the audience (without being explicated). I will discuss intention and liability in directly harmful hate speech more generally in §4.6.

[15] Here my interest is in epithets so far as they operate in directly harmful hate speech, but it is plausible to suppose that if hate speech has an indirectly harmful effect on its targets, epithets will play a part in this as well. Adam Croom suggests that prejudicial claims implicated in epithets are likely to permeate the cultural

Lepore (forthcoming) argue that the derogatory valence of slurs derives simply from the fact that they are socially illicit. In either case, what is distinctive about slurs for our purposes is their 'efficiency'. Without saying anything overtly threatening, the wily homophobe can use linguistic niceties, or seize opportunities that open up in the conversational space, to convey a threatening message to his target. But what can the simple and guileless homophobe do? *His* recourse is to say "what are you looking at, faggot?", for in so doing he can convey similar menace, while likewise immunising himself against the charge of issuing an over threat ("what do you mean?" he says, "I call my friends 'faggot' all the time"). Given the rich web of abusive associations that can be called forth with the 'appropriate' use of a slur, even unsophisticated language users can do something sophisticated in their speech: they can convey menacing suggestions – "your kind are disgusting; you could use a brutal thrashing; maybe I'll see to that for you now" – without explicating the content of those menacing suggestions.

Another mechanism through which isolated hate speech may nevertheless directly inflict harm is through what we might call 'cumulative harassment'. On a simple view of social intercourse we recognise that it is normal and reasonable for A to feel greatly distressed if she is insulted by her sociopathic co-worker twice daily for a month. On a sophisticated view of social intercourse, we should recognise that it is likewise normal and reasonable for B to feel greatly distressed if she is vilified with hate speech by 60 different, randomly-encountered strangers, twice daily over a month. No individual speaker's speech by itself constitutes harassment, but each speaker participates in the cumulative harassment of the target. Now, our programmatic liability-ascription principle in §3.6 says hate-speakers can be adjudged liable for the harm that occurs in consequence of their speech, if their performance of the hate-speech-acts under the circumstances could reasonably have been expected to have harmful consequences for others, given normal or reasonable reactions to those speech acts on the part of the speaker's audience. It may only rarely be the case that hate-speakers satisfy a 'reasonable expectation' criterion, so as to be liable for harm inflicted via this cumulative mechanism. Hate-speakers do not omnisciently observe the harassment their targets experience. However, what we *can* say, at least, is that if speaker A has reason to believe that

---

understandings of the society in which epithets are in circulation, and he suggests this may be manifested "not only in a superficial layer of disconnected racist acts towards other people, but rather in the adoption of a way of facing others and organizing our thoughts about them that is morally bankrupt" (2008: 39).

target B is experiencing cumulative harassment, A's knowingly joining in that cumulative harassment, even with just an isolated contribution, is more like a malicious incursion than a manifestation of routine social friction. In conditions of pervasive identity-prejudice, awareness of cumulative harassment, which may in principle ground liability-ascriptions of this sort, is attainable in at least some cases.[16]

One thing that becomes evident, in the description of these different 'mechanisms of harm' in verbally-assaultive speech, is that established social-hierarchies and identity-prejudices play an integral role in how acts of harassments and threats are carried out using hate speech. But identifying this integration does not absolve hate-speakers of responsibility for their part in the process. The speakers are still the agents whose conduct conveys the threatening or harassing meanings. Pervasive, identity-prejudicial social conditions make it possible for them to do this covertly (e.g. via implicature), but the hate-speech-act is the decisive event, and the hate-speaker, by extension, is the decisive actor, in the causal analysis of how the harms involved in someone being harassed or threatened come about in such cases.

## 4.4 Responsibility-ascription in incitement

Whereas verbal assaults result in mental harm, hate-speech-based incitements will typically result in physical harm. Of course hate speech is not the only type of expression through which violent (or otherwise harmful) acts can be incited. As with my analysis of verbal assault, I will begin here by discussing liability-ascription for incitement generally, using this as a basis for discussing liability-ascriptions in hate-speech-based incitement specifically. I will assume throughout that those who carry out the final harmful acts in incitement-type cases are liable,

[16] In such cases one might wonder whether 'society at large' can be deemed responsible for the harm. Isolated hate-speech-based verbal assaults may be experienced as harmful just because they emblematise, for their targets, the pain of living with oppression. One problem with this is that it cannot account for the unvarying emphasis placed on the *content* of hate speech by those who postulate a harm in such cases. No-one suggests that people in oppressed groups are unusually vulnerable to verbal aggression *per se*. Critiques here are uniformly focused on speech that supports or affirms identity-prejudice. An explanation of the potency of hate-speech-as-verbal-assault, it therefore appears, cannot advert to the positional vulnerability of hate speech's targets alone.

*ceteris paribus*, for the consequent harms. A may have incited B's harming of C, but if B is capacity-wise responsible and in-principle liable for her actions, she remains (defeasibly) liable for the harm to C.[17] My interest is not in liability-ascription for actors in the 'B-position', it is whether, and on what bases, speakers in the 'A-position' are also liable.[18]

Robert Amdur says it is intuitively obvious that if a speaker "persuades or incites violence", she "bears some responsibility for the resultant harm" (1980: 295). I would grant that there is an obvious, intuitive judgement in this vicinity, but Amdur obscures it by mentioning persuasion and incitement side-by-side, as if equivalent. Once we acknowledge audiences' agential capacities, incitement is something that will need to be characterised in contrast to persuasion. Incitement functions by "getting the subject worked up or agitated rather than by offering a convincing argument" (Sumner 2009: 215). The contrast may not always be stark in practice. There is presumably a continuum of processes, between 'pure persuasion' and 'pure incitement', through which we influence others' actions.[19] Nevertheless, the distinction is a morally significant one. Where B's act, φ, occurs as a result of B judging, by her own lights, that there are good reasons to φ, the person who previously persuaded B towards this judgement, A, cannot not be made liable for B performing φ, without B being

---

[17] So we can still say with Feinberg that "the 'reasonable person' in a democracy must... have enough self-control to refrain from violent responses to odious words and doctrines" (1975: 388), and that people who can exert such self-control but who fail to do so in a given case are, *ceteris paribus*, liable for the harmful consequences of this failure. The listener may be deemed liable for the harmful outcomes of her actions where we judge (i) that she is normally able to exercise her critical faculties to make a reasonable, controlled response to the speaker's speech, and (ii) that she is culpable for her failure to exercise those faculties under the circumstances.

[18] Note that I am rejecting from the outset any 'bucket' theory of responsibility, on which there is a fixed quantum of liability, α, associated with a given harm, x, to be allotted between A and B, such that if A is solely liable for x, A incurs exactly α units of liability, or, if A and B are equally liable for x, they both incur 0.5α units (I take the term 'bucket theory' from Nozick's critique of this view, 1974: 129-30). Also, I will be assuming that where speakers in the A-position are liable in incitement-type cases, what they are liable *for* is the eventual harm done to C, rather than (as in Gardner's analysis of complicity, 2007a) for their contribution to B's violent act.

[19] This is stressed in Scoccia's (1996) work on the relationship between violent pornography and sexual violence.

thought of as a pliable rube. Consequently, absent any prior, discernible reason to regard B in this light, B should be adjudged wholly and exclusively liable for φ in such a case. Persuasive influence is a ubiquitous part of our social relations. An unpreparedness to ascribe exclusive liability in a case like this will generalise to an unwillingness to ever ascribe exclusive liability. In free speech theory, this lends intuitive support to 'the persuasion principle', according to which states cannot justify restrictions on speech "by invoking harmful consequences that are caused by the persuasiveness of the speech" (Strauss 1991: 334; see also Sadurski 1996; Alexander 2000).[20] Certain dangers are created in treating persuasiveness as above the law,[21] but these dangers are inherent to a liberal society. Where people are living according to their own aims, values, and intentions, it seems inevitable that at some point some individuals will be persuaded towards having maleficent intentions. For liberals, the law's aim is to prevent people from acting on those intentions, not to impose conformity and upright objectives in the discursive space from which such intentions emerge.

What is distinctive of incitement, in contrast to persuasion, is that in incitement the causal connection between A's speech and the harmful consequences of B's action are such as to circumvent B's agential capacities. Unlike the person who has been persuaded, who admits (or would admit, or should admit) ownership of the reasons that compel her action, the

[20] Wollheim (1973) discerns a similar principle in *On Liberty*, arguing that Mill wants to exempt from coercive interference acts which either have no effect on others, or which, if they do have an effect, do so only through their potential to alter people's beliefs. Some contemporary authors draw a further connection between (i) the illegitimacy of interfering with acts of persuasion, and (ii) self-determination as a condition of democratic legitimacy generally. According to Weinstein (2009a: 28), for instance, "it is a core democratic precept that government may not restrict speech on the ground that it might cause the audience to adopt some unwise or even disastrous social policy. To do so would be contrary to the precept that the people, collectively, have a right to govern and, individually, have a right to participate freely and equally in this collective decision making".

[21] On this approach, as David Barnum (2006: 276) says, we are "precluded from punishing a broad spectrum of messages that are both intended and highly likely to bring about various kinds of unlawful action, including horrific violence". Kathleen Mahoney expresses concerns in a similar vein about protecting (potentially) persuasive hate propaganda. In response to the idea that "adequate protection for vulnerable and marginalized victims of hate propaganda will come from the audiences' rational capacity to disregard hateful messages", Mahoney insists that "neither history nor the present state of the world" lends support to this notion (2009: 338).

incited person, having been agitated, worked up, or provoked into violent action, acts under the governance of reasons which do not belong to her in the fullest sense. Qua incited person, she does not reflectively affirm the reasons that imminently compelled the incited conduct; those reasons only had a purchase on her behaviour due to some agency-obstructing distortion in her decision-making conditions, whether involving mob psychology, or emotional pique, or her being forced to make a 'snap decision'.

The question is: when are speakers liable for harmful outcomes that result from such interactions? Consider Mill's well-known corn dealer example (1859: 119). A mob has gathered at the corn dealer's property. Circumstances do not conduce to anyone making temperate judgements about the corn dealer's alleged wrongdoing, or about what should be done in response. Anger is in the air. Each person's aggression fuels, and is fuelled by, the aggression of his neighbour. No system is in place to manage the discussion or to govern a sober decision-making process. The escalatory dynamics of violence, as we might say, are present and ready to be triggered. Now, what is wrong with someone then stating, to the gathered mob, that all corn dealers are thieving scoundrels? The answer is: this expression fails a 'clear and present danger test', because it is carried out despite the significant, foreseeable likelihood that it will incite the gathered mob to harmful violence.[22] For more general purposes, I suggest we think of foreseeability as having two components. Firstly, there is the bare fact of volatility: it is apprehensible, in such a case, that people who would certainly *not* act violently under normal conditions *may* act violently now, given the acrimonious temper and the possible effects of mob psychology. Secondly, there is the fact that this volatility is directed: it is apprehensibly likely that the formless acrimony that is present will be turned against certain people who can be (subjectively) regarded as its proper object.[23]

---

[22] For Mill, these circumstances deprive the speech of its presumptive entitlement to protection *as* 'speech'. The opinion that corn dealers are thieves may not be suppressed, but airing that opinion in certain conditions constitutes "a positive instigation", for Mill, as opposed to a 'mere' expression of opinion (1859: 119).

[23] In the UK, incitement offences have been replaced with more broadly-framed statutes restricting conduct which assists or encourages a crime. Under this framework, the foreseeability criteria for liability-ascription invoked in traditional incitement law are replaced with requirements either that the defendant intentionally cooperated in the crime, or assisted/encouraged the proponent in the belief that a crime would be committed (See Ashworth 2009: 458-61). This reconfiguration of incitement law fails, I think, to appreciate the distinctive nature of incited harm. Incitements are not, as the revised legislative framework in the UK

Our responsibility-ascription principle from §3.6 says that speakers are responsible for the harm that their speech brings about if their speech could reasonably be expected to result in harmful consequences for its targets, given normal or reasonable reactions from the audience. Provisionally, then, if we are going to treat the clear and present danger test as the key criterion for liability-ascription in harms resulting from incitement, we will have to finesse our understanding of what a 'normal and reasonable' reaction involves in such cases. After all, a *reasonable* person would not react to provocative speech with violent action. What we are requiring the speaker to anticipate, for the purposes of liability-ascription, is the possibility – given observable and directed volatility – that otherwise non-violent people will perform violent actions. If speech can reasonably be expected to result in harm, in these respects, then the speaker will be liable in principle if such harm *does* eventuate.

My claim in §3.6 was that liability-ascriptions should be guided by a conception of audiences as responsible agents, who are capable of exercising self-control and mental resilience in response to other people's speech. Here, as in the conception of social intercourse that I sketched in §4.1, I am trying to outline the boundaries of this capacity. Just as people's resilience to verbal assault can become enfeebled due to unfavourable conditions, so too there can be destabilising forces that impair people's capacity to sober-mindedly withstand the escalatory dynamics of incited violence. When an individual speaks in public, she can rightly and reasonably expect her audience to take responsibility for the way they react to what she says. However, at the same time, it is incumbent upon the speaker to recognise that the conditions which enable her audience to take such responsibility sometimes break down, and that when it is apprehensible that this has occurred, she is answerable for the consequences of her injecting provocative sentiments into a volatile environment, through her speech. The speaker can incur liability for inciting the harms that her audience enacts, where it

---

imagines them to be, fully-fledged collusive crimes, tantamount to supplying weapons or strategic advice to someone planning a violent act. But neither are incitements merely reckless forms of public speech which increase the risk of an unimagined future harm (and as such, conduct of a sort that is best regulated via generic public order offences). Incitements occupy an intermediate space between increasing the risk of harm and colluding in the harmful act itself. Within that intermediate space, the appropriate liability-ascription condition for the inciting party is his recklessness in the face of foreseeable risk of harm, rather than (as in the UK nowadays), his fully-fledged intention or conscious expectation of a harmful outcome eventuating. I will say more about intentionality and culpability in incitement in §4.6.

is reasonably foreseeable that her speech will trigger the kind of (directed, volatile) escalatory dynamics that are characteristically involved in speech-incited violence.[24]

## 4.5 What's different in hate speech?

Should we have a different view of how these escalatory dynamics may be anticipated, and of how liability should be incurred for the resultant harm, when the putatively inciting speech is hate speech? There are a number of authors in the literature who advert to the inciting effects of hate speech as a justificatory basis for anti-hate speech law,[25] although the connection between hate speech and violence is often described in ways that obscure the distinction between persuasion and incitement. We are told that hate speech 'sets the stage' for identity-based mistreatment (Delgado and Yun 1994: 1813), 'paves the way' for aggression (Tsesis 2002), is 'a fertilizer' for violence (Weinman 2006), and 'lays the foundation' for murder (Mahoney 1996: 792). These claims capture the idea of hate speech having a pro-violence influence, but they do not acknowledge the different modes of influence that hate speech might exert. Of course we can allow that different modes of influence may operate in different cases. However, our interest is confined to cases in which (i) hate speech triggers the escalatory, agency-circumventing dynamics that are characteristic of incitement *as opposed to*

[24] Our interest in liability-ascription in particular explains why we cannot simply take an expected-utility approach to anti-incitement law (as sketched in McCloskey 1970; Glass 1978). Where responsibility for harm is accepted as the principal condition for someone being liable to punishment under law, it is illegitimate to treat the restriction of inciting speech as if it were merely a minor proximate disutility, whose infliction is warranted to prevent the small chance of a larger disutility (due to incited violence ) eventuating later.

[25] See for instance Knechtle 2005; Barendt 2009; Tsesis 2009. Note that discussions of incitement in the hate speech literature sometimes focus on incitement to *hatred*, rather than incitement to violence. This is partly due to the fact that in certain jurisdictions, including the United Kingdom, 'incitement to hatred' is an expression that appears in the definition of the legal offence under which certain forms of hate speech are liable to restriction (S. D. Williams 2009). Incitement to hatred is peculiar as legal offences go, insofar as the outcome whose incitement is prohibited under such an offence (i.e. the outcome of people feeling hatred for a certain group), is not *itself* an offence, in the UK or anywhere else. Here, in any case, I will just focus on incitement to *violence*.

persuasion, and (ii) the hate-speaker is liable for this, in that there is a significant and foreseeable likelihood of these escalatory dynamics being triggered by his speech.

Let's firstly consider an example of hate-speech-incited violence which appears to satisfy both of these conditions, namely, the Rwandan massacre. Over a period of several weeks in 1994, upwards of 500,000 members of Rwanda's Tutsi minority were killed by members of the Hutu majority, following months of inflammatory anti-Tutsi hate messages being broadcast through media outlets such as the newspaper *Kangura*, and the radio station Television Libre des Mille Collines. Based on reports on events in Rwanda leading up to the massacre, it seems clear, firstly, that the simple fact of volatility was apprehensible for those who engaged in the media-proliferated hate speech at issue. Journalistic records show that in the lead-up to the main period during which the massacre took place, *ad hoc* acts of violent aggression against Tutsis, and the mobilisation of armed militia groups throughout many parts of the country, had been widely-discussed in the Rwandan media (see Prunier 1995; Malvern 2004; Mamdani 2004; G. S. Gordon 2004). It seems clear that it was known, throughout Rwanda in 1994, that once-temperate rivalries between the two groups had devolved into an open and volatile conflict. It seems clear, moreover, that this volatility would have been apprehensibly likely to be directed against specific targets. The Hutu militias were not indiscriminately angry mobs, seeking any old outlet against which to vent their anger. Their aggression was specifically against, and known to be directed specifically against, members of the Tutsi minority. Those who engaged in hate speech in this environment could have and should have foreseen that violent acts were likely to be triggered by their speech. Granted, it is highly artificial to describe the hate-speakers' perspective in these terms. Many of the individuals implicated in hate broadcasts during and prior to the Rwandan massacre were later convicted of inciting genocide.[26] Given the verdicts in these cases, it is misleading to say that

[26] See Benesch 2007: 491-98. To be convicted of incitement to genocide under Article 3 of the UN Convention on Genocide, the accused must be found to have had a specific intent to cause genocide, and that intention must have been directly manifested in public acts of expression; see the *Convention on the Prevention and Punishment of the Crime of Genocide*, December 9, 1948, 102 Stat. 3045, 78 U.N.T.S. 277. Although acts of incitement to genocide, thus defined, need not incorporate hate speech, it is natural to expect that they often will, as in the Rwandan case. Schabas (2000) argues that the Convention on Genocide should allow punishment of hate speech that occurs prior to the other escalatory dimensions of violence being in place; Lynne Tirrell (2012) goes even further, arguing that such hate speech should be viewed not merely as

Hutu hate-speakers *could* have foreseen their speech triggering murderous violence against Tutsis, since, in many cases, the speakers did *in fact* foresee this outcome, and in their hate broadcasts enacted an intention to bring it about. When we say the violence could have been foreseen, we are speaking in anticipation of the hate-speaker's attempt to evade liability, on grounds of ignorance, for his role in instigating violence. Whether or not the Hutu hate-speaker *deliberately* intended to incite violence in his speech, the claim is that he should be liable for the harm that his speech instigated, since that instigation was foreseeable.

In a case like this one, in which the directed volatility (of the audience to whom the hate speech was broadcasted) was clearly apprehensible on all sides, the fact that the inciting speech was hate speech does not seem to be essential to an explanation of how the violence came about and how liability ought to be ascribed. As with Mill's corn dealer example, it was evident that the escalatory dynamics of violence were liable to be triggered by speech, and it is primarily in view of this – that is, the 'trigger conditions' being *apprehensible* – that liability for harm redounds to the hate-speaker. There are other kinds of cases, however, in which the trigger conditions for the escalatory dynamics of violence are less evident, but in which it still seems right to judge that hate speech is responsible for triggering violence, in such a way that hate-speakers could be liable for the harm that occurs in consequence. Consider the case of the 2005 Cronulla riots in Australia. The riots occurred shortly following an incident in which several men of Lebanese descent had assaulted two lifeguards during an argument on a South Sydney beach. An informal gathering aimed at 'reclaiming the beach' was subsequently arranged by white youths in the local community, and in the days leading-up to the gathering, some members of a white supremacist group in Sydney began circulating inflammatory text messages about the gathering, one of which was reprinted in *The Daily Telegraph* (Australia's then highest-circulating daily newspaper). "This Sunday every Aussie in the Shire get down to North Cronulla to help support Leb and wog bashing day... Bring your mates and let's show them that this is our beach and they are never welcome... let's kill these boys" (quoted in Poynting 2006: 86-87). A mob of approximately 5,000 white men gathered on Cronulla beach the following Sunday and, over the course of several hours, confronted and violently assaulted several dozen men of Middle-Eastern appearance. Most of the individuals participating in the

---

an antecedent to the events of genocide, but instead as a component among the temporally extended events that jointly constitute genocide.

violence had no prior affiliation with any white supremacist groups, and were just local young men responding to 'reclaim the beach' messages circulated via social networking, text messages, and in the mainstream media (Poynting 2006).

What kind of harmful outcomes should we suppose were reasonably foreseeable for those who approvingly circulated messages, on social networking sites, like "bring your mates and let's show them that this is our beach and they are never welcome"? Was there so much anger and volatility in the community that people's ability to resist the escalatory dynamics of incited violence had already broken down? By the time a mob of 5,000 gathered on the beach, that threshold was presumably crossed. But earlier in the week when the messages were first sent? Surely not. The circulation of these messages quite evidently contributed to the gathering of the mob in the first place. However, so the speakers who circulated the messages might say in their own defence, few would have expected 5,000 people to show up, and in so doing, to create the conditions in which frustrations and simmering resentments would blossom into mindless acts of violence against innocent strangers. We can certainly say that it was reckless and foolish – and for those who were hoping or expecting that violence *would* eventuate, hateful and malicious – for these messages to be expressed and circulated. But there is a sense in which, when we set aside the racial dynamics from our consideration, many of the instigators in the Cronulla case probably could not have reasonably anticipated their speech acts setting off a chain of events in which people were violently attacked. The dangers were not clear and present, they were conjectural and relatively remote.

Or so it will appear, as I say, if we bracket the considerations of underlying identity-prejudice from our assessment. On the other hand, if we take those patterns of prejudice and hostility into account, the dangers associated with circulating a message about "Leb and wog bashing day" seem rather less remote, even when no mob has yet assembled in the presence of people who may be picked out as targets by such epithets. When we're trying to understand how the crude, half-baked provocations of anonymous social media users managed to whip up a violent mob of 5,000, we will naturally conclude that identity-prejudice had some role in events. The mob of (we may assume) otherwise (mostly) non-violent individuals was able to be gathered in the first place, and incited to violence subsequently, partly because identity-based prejudice against people of Middle-Eastern ethnicities suffused the beliefs and attitudes of the South Sydney community at large. Some of those who spread these messages hoped or

expected to take advantage of precisely that fact. But what of those people who just took themselves to be 'stirring the pot' or venting frustrations? Can they be made liable for their harm-inciting speech, carried out with a failure to foresee the escalatory social dynamics that became evident in hindsight? Well, insofar as failures in foresight provide a basis for liability in incitement-type cases generally, they can provide a basis for liability in hate-speech-based incitements as well. In the corn dealer case, our judgement that the harmful outcome was reasonably foreseeable derives from the fact that the directed acrimony (towards the corn dealer) was already evident: the speaker incurs liability for spreading a message that stoked the (already evident) acrimony into full-blown violence. In conditions of pervasive identity-prejudice, the same sort of foresight can ground liability for harm related to hate-speech-based incitements. To say that the harm is reasonably foreseeable is to say that the directed acrimony (e.g. against people of Middle Eastern descent in Cronulla) was already evident (among certain people in the speakers' audience). Speakers who spread contemptuous messages calling for violent action in accordance with directed acrimony may therefore be liable for the harm that ensues, if and when the already evident acrimony is fanned into violence.

In her discussion of the clear and present danger test, Lyrissa Lidsky suggests that the test's appropriateness, as a liability-ascription tool, depends on the social conditions in which it operates. It does its job when there is "a particular form of public discourse and particular types of citizens" providing "built-in safeguards against mass violence". But in the absence of the safeguard of an engaged and rational citizenry who have the ability to deliberate about policy and refute those who call for violence, Lidsky sees the clear and present danger test as dangerously *laissez faire* (2002: 1021-22). David Anderson defends a similar view: as a regulatory response to inciting speech, he says, the clear and present danger test is "a rule that bears no relation to the kind of risks that the activity actually creates" (2002: 1006-07). What both authors recognise is that social relations sometimes devolve to the point where some kind of 'state of emergency' must take effect. Liability for inciting harm is incurred when the speaker increases the volatility in an already-destabilised set of social relations, and when the likely harmful effects of his doing so were foreseeable. What is different in cases of hate speech is that, because hate-speakers espouse attitudes of identity-prejudice that already hold sway in certain parts of the wider community, the further destabilisation of volatile social relations is never too remote or too conjectural a possibility.

## 4.6 Intention and expectation

In both hate-speech-based incitements and hate-speech-based verbal assaults, there is a question about whether liability for harm should be contingent on the hate-speaker having a conscious intention to harm. Should there be a *mens rea* requirement in legal restrictions on directly harmful hate speech, or is the mere *actus reus* of a hate-speech-based incitement or hate-speech-based verbal assault sufficient to render the speaker liable to punishment for the harm that occurs in consequence of those acts? Throughout my discussion in this chapter I have assumed that the hate-speaker is a capacity-wise responsible agent in the general sense (which is to say that he has a general ability to deliberate, control his actions, etc.), and that the hate-speaker is therefore, in principle, morally culpable for his actions.[27] Why does moral culpability matter in this context? In short, due to the basic precept that A's being morally culpable relative to x is a necessary condition for A's being made legally liable relative to x (see Duff 1990: 102-03). Note, however, that A's being capacity-wise responsible relative to x is not always regarded as sufficient for A's being morally culpable relative to x. In our informal practices of blaming and punishing, at least, moral culpability for x is standardly thought to require some degree of intentionality vis-à-vis x. With respect to directly harmful hate speech, the relevant element of intentionality will at least sometimes be absent in cases where all other culpability conditions are satisfied. For instance, the hate-speaker may view his speech merely as a form of 'letting off steam', or he may seem himself engaging in harmless joking or banter. The speaker may have no idea that his speech is liable to cause his targets grave anxiety and distress, or to incite acts of violence against them. In such cases, so one may argue, where the harms resulting from hate speech are accidental relative to the speaker's intentions, the hate-speaker should not ultimately be adjudged liable for the eventuating harm.

One may wonder how often we can really accept that hate-speakers *are* naive or ingenuous in the ways I have described here. Waldron writes of hate-speakers as if they are, characteristically, executing a deliberate strategy. He talks of hate speech being "calculated to

[27] It is a vast and difficult problem to specify precisely which conditions should qualify as impairments to an individual's capacity-responsibility, and to determine when and how, exactly, such impairments preclude liability-ascription (see note 18 in §3.4). Having said that, these problems do not arise in any distinctive way for anti-hate speech law. Any general account of capacity-responsibility and its relation to liability-responsibility in law can be invoked at the necessary junctures in relation to anti-hate speech law.

undermine the dignity of other people" (2012: 60), and he insinuates that hate-speakers are typically aware that their speech impairs the social status of their targets, and that they intend this outcome in speaking (2012: 74).[28] While these are empirical questions to some extent, I do not find Waldron's view persuasive, outside of cases in which we can identify a wider pattern of conduct that betrays the speaker's contempt for those he impugns in his speech, such that his speech can be understood as an articulation of those more broadly-manifested attitudes. When it comes to people who perform isolated hate-speech-acts, but who do not belong to hate organisations or ethno-nationalist political parties, and who have no prior track-record of harassment or physical aggression against the people targeted in their speech, hate speech is more plausibly attributed to ingenuous or mindless spite, than to a conscious and deliberately menacing purpose on the speaker's behalf. Do these isolated instances of hate speech express or manifest the existence of systemic identity-prejudice towards certain identity-groups? In many cases, as I have already stated, they certainly do. My claim is that those who sub-consciously internalise these forms of systemic identity-prejudice, and who then externalise what they have internalised, cannot plausibly be regarded as executors of a deliberate intention to harm their targets: not unless their hate-speech-acts are performed as part of a larger pattern of behaviour that manifests that kind of deliberate intention.

It still remains as a separate question, however, whether there *should* be a *mens rea* requirement in restrictions on directly harmful hate speech, and I suggest that there should not be any such requirement. So far as attitudes or intentions are a necessary component of moral culpability in this arena, we should follow Hart (2008b) in downplaying the essentiality of an intention to harm as such. In-principle moral culpability, in relation to hate speech, should just be contingent on the hate-speaker (i) possessing the normal mental and physical capacities for doing what the law requires, and (ii) having a fair opportunity to act under those capacities.[29] If an agent whose conduct satisfies (i) and (ii) infringes legal restrictions on

---

[28] A similar view is suggested in Richard Delgado and David Yun's claim that hate speech is "an instrument of majoritarian identity politics" and "a central weapon in the struggle by the empowered to maintain their position in the face of formerly subjugated groups clamouring for change" (1995b: 1298-99).

[29] Against Hart, Mackie (1977) argues that a causal connection between the agent's desires and her performance of an *actus reus* is the key determinant of criminal liability. At the same time, though, Mackie wants to allow liability-ascriptions in cases in which the *actus reus* is performed with only a vague awareness

potentially directly harmful hate speech, and harm does indeed eventuate, the attitudinal faults in view of which he is morally culpable for this harm are faults pertaining to his failure to govern his conduct in adherence to legitimate legal restrictions, despite having the capacity to do so. To borrow some phrases from Antony Duff, where hate speech is performed by people whose lives "exhibit the continuing structures of thought, attitude, and motivation", those people express a form of "inappropriate attitude towards the law and the values it protects" (Duff 1993: 380). It is the expression of such attitudes, rather than consciously malicious intention or egregious recklessness, that makes such hate-speakers legally answerable, in-principle, for the harms resulting from their speech.

In §3.6 I said, regarding the relationship between causal responsibility and liability, that hate-speakers can be adjudged liable for the harm that results from their speech if their performance of the relevant hate-speech-acts, under the circumstances, could have reasonably been expected to have harmful consequences for others, given responsible reactions to those speech acts on the part of the speaker's audience. It should now be clear that my reference to 'reasonable expectations' in this formulation was not intended to 'subjectivise' responsibility-ascriptions in this arena. Given many hate-speakers' narrowly situated points of view – what *they* do and do not know, what *they* have and have not experienced, where *their* empathic capacities do and do not extend – it may be the case that they would not normally expect their hate speech to inflict harm on those against whom it is directed. But we should not treat ingenuousness or obliviousness of these kinds as exculpating considerations, because this amplifies the burdens faced by people targeted in identity-oppressive speech. Not only must they deal with the harmful consequences of being subjected to such speech, they must also educate hostile interlocutors about those harms before those interlocutors become culpable for effecting those harms through their verbal aggression (Goodall 2007: 111-13). Reasonable expectations, in this arena, are the expectations that would occur to someone who evinces a genuine concern for the welfare of all others in society, and who is not grossly ignorant or

---

that harm may eventuate. He suggests that in such cases, the actor (obliquely) intends her negligent attitude (Mackie 1977: 184). But it is surely better in such cases to say that the actor is liable for the *actus reus* despite her lack of maleficent intention, rather than making liability depend on 'obliquely intended negligence', for while the policy ramifications are largely equivalent, the former approach does not require us to grossly distort the concept of intention.

obtuse about the ways in which hate speech can inflict harm within an identity-oppressive social milieu.[30] To be ignorant or obtuse is certainly not an offence by itself. But nor is it a general barrier to one's being answerable for the harm inflicted via one's hate speech. Where the adverse consequences of hate speech *are* genuinely harmful, and where the hate-speech-act can be identified as the proximate cause of the harm, a general capacity-responsibility is all that is required, with respect of the hate-speaker's attitudes or intentions, to render the hate-speaker an appropriate liability-bearer for those harms.

It is important to stress, once again, that audiences are not being imagined as passive bystanders in the processes through which the harms in question are brought about. The hate-speaker who uses hate speech to verbally assault his target is not being told to imagine his target as a mentally fragile child, who will fall to pieces at a single bullying remark. The hate-speaker who addresses a group of third parties (whether face-to-face, or via communication technology) is not being told to imagine his audience as a mob of mindless automatons, ready to run riot at the merest encouragement. What the hate-speaker is being told to recognise, first, is the general fact that prior events and conditions can impair people's agential capacities to be resilient in the face of verbal hostility, or to resist the escalatory dynamics of verbally-

---

[30] Reasonableness in conduct is a quality related to one's cooperativity with others. As Ripstein says "the familiar common-law idea of the reasonable person gives expression to [the] idea of a fair balance between liberty and security. The reasonable man has long been a central character in the common law, taking appropriate precautions against accidentally injuring others, making only allowable mistakes, and maintaining an appropriate level of self-control when provoked. The reasonable person is neither the typical nor the average person. Nor is the reasonable person to be confused with the rational person, who acts effectively in pursuit of his or her ends. Instead, the reasonable person needs to be understood as the expression of an idea of fair terms of social cooperation" (Ripstein 1999: 7). The contrasting conception of rationality and reasonability that Ripstein invokes here has its earliest clear articulation in Sibley (1953), and is important in the contractualist approaches of Scanlon (1982) and Rawls (1993). My uses of 'reasonable' above and below can be taken as correlative the with the reasonable person device, as Ripstein describes it. To be reasonable is to conduct oneself in a way that manifests an intention to abide by fair terms of social cooperation. The notion of a reasonable *expectation* could be elaborated in several ways, given this general characterisation of reasonableness; it might be (i) what one would expect to occur if others are reasonable, or (ii) what one could expect to occur if one is being reasonable in one's expectation itself (i.e. forming an expectation such that one is, in forming that expectation, acting under fair terms of social cooperation), or (iii) an expectation that is reasonable in both of these other senses, (i) and (ii).

incited violence, and second, that the vulnerable and disadvantaged position of minority ethnic, religious, and sexual groups in our society is such that those kinds of prior events and conditions are more likely to obtain in relation to hate speech. These conditions can make hate speech a more potently harmful action than it may appear to the critically disengaged. When such conditions obtain, and when it is apprehensible that that is the case, the individual who engages in hate speech anyway is liable in-principle for the harm that eventuates. Where we have instituted statutes that formalise these liability-ascription judgements, the would-be hate-speaker may not even be required to engage his empathic imagination in order to anticipate the potential harmful consequences of his speech. Our statutes will issue guidelines for when hate speech is likely to be directly harmful. So far as the hate-speaker's intentions are concerned, all that is required is that he possesses a general capacity to adhere to the law's demands, and a fair opportunity to exercise that capacity. If the hate-speaker engages in hate speech while that condition is satisfied, and if harm to his targets eventuates – whether via harassment, threats, or incited violence – then the law can properly hold the hate-speaker accountable for that harm.

ଔ

The problem we began with was how we ought to ascribe liability for the mental harms resulting from verbally-assaultive hate speech, and for the physical harms resulting from hate-speech-based incitements. If we bracket off the hate speech, and think about 'regular' verbal assaults and incitements, then the key conditions for ascribing responsibility seem to be repetition, targeting, and overtness (in the case of verbal assaults) and apprehensible, directed volatility (in the case of incitements); the speaker is presumptively liable for harm when these conditions are satisfied, and the audience (if anyone) is liable for harm in their absence. However, when we consider the interaction between identity-oppressive speech and identity-oppressive social forces, those rules of thumb seem to excuse and exonerate hate-speakers too quickly. In verbal assaults, they lead us to imagine victims of hate speech as being like the thin-skinned yuppie (from §3.6), because we do not take into account the demoralising effects of the mutually-reinforcing relationship between hate speech and an identity-prejudicial social milieu, and how those effects are experienced by the person targeted in a hate-speech-based verbal assault. In incitements, our rules of thumb too readily

excuse hate-speakers who inflame dormant identity-based hostilities in the community. The solution is not to dispense with these rules of thumb, and render judgements about responsibility based entirely on particularised, case-by-case appraisals of who has behaved well or poorly under the circumstances. Judgements of this sort will inevitably be invoked at certain points in the law's ascriptions of responsibility for harm. However, we can still make progress in augmenting and extending our initial guiding principles to make them more suitable for liability-ascription in directly harmful hate speech.

## 4.7 The question of redundancy

The final question I will take up in this chapter is whether our legal systems should view hate-speech-based verbal assaults and incitements as offences that are distinct from other forms of verbal assault. After all, where there are already legal remedies in place for things like verbal harassment, threats, and incitement, won't these be sufficient for curbing or redressing *hate speech* that is used to threaten, harass, and incite? If so, then further, dedicated restrictions on hate-speech-based verbal assaults would be gratuitous or redundant, since all the hate-speech-based offences in question would be governed under the broader offence. As Wojciech Sadurski says, "there is no need for a new criminal category to describe incitement to racially motivated violence", because "the offence is already handled by general anti-incitement rules" (1999: 205); and the parallel argument can obviously be advanced in reference to threats and harassment. The suggestion here, more generally, is that hate speech policy should take an 'umbrella approach', on which hate-speech-based offences are liable to restriction as instances of other verbal offences (described in more general terms), and subject to the same constraints that govern the restriction of any other instances of those offences.

Against this approach, L.W. Sumner argues that there are various independent reasons to favour dedicated anti-hate speech statutes: "for the symbolic importance to minorities of acknowledging their particular vulnerability, or for publicly communicating the message that racism and homophobia have no place in a liberal social order, or for singling out certain uses of speech as hate crimes for the purposes of penalty enhancement". At the same time, Sumner says, these are matters "which a liberal society may be left to decide on pragmatic grounds" (Sumner 2009: 219). In my discussion of liability-ascription in directly

harmful hate speech, we have seen an additional set of considerations that tell in favour of treating hate-speech-based verbal assaults and incitements as separate offences. Given any general-purpose criteria for ascribing legal liability for harassing or threatening speech, the isolated, non-targeted, and non-overt instance of hate speech is unlikely to be recognised as a matter of concern for the law. An isolated instance of hate speech that is used to intimidate someone will generally fall short of constituting a threat proper, even though, given the social context, its effects would be comparable to those of a true threat. We can draw a useful comparison here with sexual harassment jurisprudence. One may wonder why sexual harassment should be treated as a distinct phenomenon and offence, given that it is, in one evident sense, merely a specific form of 'harassment' as a larger category of offence. Even if there is some symbolic or bureaucratic value in treating sexual harassment as a separate offence, an important part of the reason for our doing so is that the sexually-oriented form of the phenomenon of harassment operates in a distinctive manner (or a variety of manners, some of which are distinctive to the particular sub-category), and that its operations – being different from other forms of harassment, and primarily affecting a group whose perspectives and concerns have not played a major role in shaping legal norms around harassment *per se* – are such that it is not always recognisable *as* a form of harassment. In other words, although sexual harassment *is* simply another form of harassment in one respect, it becomes visible in the legal system when it is treated separately from the overarching category of harassment, and its harms are therefore more likely to be effectively redressed if it is recognised as a distinct type of offence.[31] In a similar vein, directly harmful forms of hate speech, even if they are assimilable in some respects to broader categories of harmful speech – threats, harassment, and incitement – are likely to be more visible in a given legal system when treated as distinct classes of offence, and more effectively redressed under that system as a result. Given the distinct mode of operation we see in hate-speech-based verbal assaults and incitements, a separate set of legislative instruments for preventing or redressing the related harms seems appropriate and justifiable. Of course those devices should be subject to the basic demands of institutional parsimony. My claim is not that there ought to be a rapid proliferation anti-hate speech law focussing on harassment, threats, and incitement. The claim is simply that general-

[31] This sketch of a rationale for a 'separate-offence' approach to sexual harassment is broadly in agreement with the approach MacKinnon defends in *Sexual Harassment of Working Women* (1979).

purpose legal restrictions on these forms of speech are ill-suited, given their broad-ranging remit, to taking account of the distinctive and often inconspicuous ways in which directly harmful forms of hate speech do their work, and hence that specially-established categories of legal offence, judiciously framed and administered, will be better able to identify and remedy those cases in which hate-speakers are causally responsible for inflicting harm on their targets through verbal assaults and incitements.

ଔ

We have been examining two types of cases in which hate speech is directly causally responsible for harming people: (i) verbal assaults, in which the hate-speaker uses his speech to harass or threaten his targets, and thus inflict adverse mental states like anxiety, distress, and humiliation; and (ii) incitements to violence, in which the hate-speaker provokes third parties into violent action against his targets. When we classify a harmful outcome associated with these kinds of cases as a *direct* harm of hate speech, we are saying two things: firstly, that the hate-speaker's speech was among the causes-in-fact of the harm; and secondly, that among the *set* of causes-in-fact of the harm, the hate-speaker's speech was one of, or the, decisive difference-maker vis-à-vis the harm, such that it is reasonable and appropriate to make the hate-speaker liable for the harm, i.e. to make him bear the costs, in the form of punishment, penalty, or some other coercive restraint on conduct, that come with redressing or remedying the harm, or preventing further instances of it. I have sought to identify the bases on which speakers may be judged directly responsible for harmful outcomes associated with hate speech, in both verbal assaults and incitements to violence, and in so doing to illuminate certain complexities in the harm-causing operations of hate speech. In both types of cases the identity-prejudicial milieu in which hate speech is delivered is causally implicated in the process through which hate speech inflicts its harm. Systemic practices of identity-prejudice, across lines of difference in ethnicity, religion, and sexuality, create a communicative environment in which, in order to convey identity-prejudicial locutionary contents, and thus generate harmful perlocutionary effects in the audience, the hate-speaker does not need to issue overtly aggressive or provocative statements. By exploiting linguistic implicatures and suggestive symbolism, the hate-speaker can covertly endorse identity-oppression to those who

are victimised by it, or encourage identity-oppression to third parties with whom the speaker seeks to align his purposes.

Part of my aim in this analysis is just to contribute to an explanation of the operations of a certain kind of hostile verbal conduct. But I have suggested on the basis of this analysis an argument for treating hate-speech-based verbal assaults and incitements to violence as distinct legal offences. Having recognised that identity-prejudicial locutions can be put to directly harmful illocutionary uses, one might still prefer to rely on broadly-framed legal restrictions on harmful illocution – harassment, threats, and incitements to violence – to regulate the instances involving hate speech. Against this approach, I have argued that principles for attributing responsibility for harm in cases of verbal incitement, harassment, and threats are liable to misattribute responsibility in a significant range of cases, because of the covert processes through which hate speech inflicts harm. These principles quite rightly direct our attention to the agential participation of the hate-speaker's *audience* in harm-causing processes, whether they, the audience, are (i) being attacked, or (ii) being provoked into attacking others. But given this emphasis, the principles are unlikely to recognise the hate-speaker's causal contribution to the harms in question. They are unlikely to recognise *what* the hate-speaker is able to convey to others, and *how* he does so, when his speech inherits its harm-causing power from a pervasive, identity-prejudicial social system.

# ☙ 5 ❧

## Indirectly harmful hate speech

Restrictions on directly harmful hate speech are supposed to penalise only hate-speech-acts that are, as an ascertainable matter of fact, responsible for harming people. They do not prohibit hate speech as such. Granted, it *is* part of their aim to deter people from performing potentially harmful hate-speech-acts. However, on a proper understanding of free speech (see chapter 2), this sort of 'chilling effect' does not make such laws illegitimate; firstly, because hate speech isn't a protected species of speech requiring stringent protection, and secondly, because the restrictions in question aim at a legitimate purpose, i.e. harm-prevention, rather than the illegitimate goal of viewpoint suppression.

Where restrictions on directly harmful hate speech would merely chill or discourage hate speech, restrictions on indirectly harmful hate speech would make all hate speech liable to restriction. The idea behind this sweeping approach, as I explained in §3.1, is that all hate speech, even that which isn't directly harmful, contributes to an identity-prejudicial social climate that harms members of vulnerable groups. In what follows I investigate this kind of case for anti-hate speech law. As I said in §3.1, the ascription of liability to the hate-speaker for indirect harm is not based on his hate-speech-acts being judged causally non-redundant in discrete harms done to a specific person. Rather, the liability-ascription is based on a judgement about the effects of hate speech as a widely-practiced form of expressive conduct, namely: that it contributes, in a non-trivial way, to the perpetuation of harmful, identity-based social hierarchies. If the hate-speaker is liable under this approach, it is for harmful outcomes ensuing cumulatively from the *form* of expression in which he engages. However, I will argue that the case for general restrictions on hate speech under this indirect harm rationale fails. My objection is not to the structure of the liability-ascription as I have described it. Contributive or participatory liability is a plausible account of how liability is incurred in cases of 'social harm' like those Feinberg cites (involving tax evasion and contempt of court, see §3.1), and of how liability can be incurred for *literal* environmental harms (e.g. related to pollution). Instead,

the problem is epistemic: we cannot confidently judge that hate speech does contribute, more than trivially, to the social-hierarchy-based harms associated with it. There are many attempts in the literature to spell out the contributive connections between hate speech and identity-based social hierarchy, but I argue that these are inadequate as putative justifications for anti-hate speech law, in short, because they only identify causal possibilities. Either they describe processes through which hate speech *could* contribute to social hierarchy and its harms, or they offer a description of social hierarchy on which it is easier to imagine how such contributions might occur. They do not enable us to arrive at firm conclusions regarding the indirect harmfulness of hate speech, as would be required to justify its general restriction.

After discussing how we should think about the causal bases of social hierarchy in §5.1, in §§5.2–5.3 I examine accounts from Steven Heyman and Jeremy Waldron, which characterise the indirect harms of hate speech in terms of infringements against people's dignity. In §§5.4–5.5 I examine several ways in which one might characterise hate speech as something that authoritatively legitimates social hierarchy, and which thereby contributes to social-hierarchy-based harm. In §§5.6–5.7 I examine accounts which seek to characterise the indirect harms of identity-prejudicial expression as impairments to people's communicative capacities. In §5.8 I discuss the prospect of appealing to findings in social and behavioural psychology to substantiate claims about the indirect harmfulness of hate speech. Finally, in §5.9 I discuss standards of evidence and justification: if the analyses that I examine cannot substantiate claims about the indirect harmfulness of hate speech, what would it take to substantiate such claims so as to justify general prohibitions on hate speech?

## 5.1 The etiology of social hierarchy

According to Katharine Gelber (2002: 87), racist hate-speech-acts "reproduce and reinforce inequality on the grounds of race"; Laura Beth Nielsen says that so far as "prejudice is about relative group position... hate speech provides a clear example of one of the ways in which such social hierarchies are constructed and reinforced on a day-to-day basis".[1] Such remarks

[1] Nielsen understands prejudice as being essentially a pro-social hierarchy attitude. Prejudice thus conceived need not be consciously malicious. One may regard identity-based hierarchies as simply *natural*, and act this

seem *prima facie* plausible. Modern liberal societies may be formally egalitarian, but they contain complex, structural disadvantages for people in marginal social groups. Hate speech is sometimes directly responsible for mental and physical harms to its targets (as we saw in chapter 4), and even when it isn't, its sentiments evidently affirm the identity-prejudicial aspects of our social arrangements. It would be implausible, then, simply to assert that hate speech bears *no* relation at all to identity-based social hierarchy. But there is an unwarranted assuredness in Gelber's and Nielsen's remarks as well. It is difficult enough to perspicuously describe how social hierarchies work, let alone what causally underwrites their workings. We needn't demand a simple explanation; as several authors say, there are likely to be numerous causal factors at work here, including various verbal factors (Lederer and Delgado 1995: 5; Savage and Liht 2009: 488-89). However, given the complexity of the system, the contributory role of hate speech cannot be taken as obvious from the outset.

So how might we ascertain hate speech's role in this arena? It is crucial not to think of social hierarchy as a static outcome, for which we are trying to identify preceding links in a causal chain, as in our discussion of the direct harms of hate speech. Rather than being the final node in a causal sequence, identity-based social hierarchies are dynamic outcomes *constituted* by complex causal networks of actions, practices, and institutions; they are states of affairs that supervene on the wide-scale operations of these networks.[2] We can think of hate

---

out in ways that are generally non-confrontational (this is the conception of prejudice we find, for example, in the work of people like Frye 1983; Young 1990; and Bobo 1999).

[2] The idea that hate speech *constitutes* a harm for its targets appears, for example, in Maitra 2012: 98-99, and Waldron 2012: 166-67. Something like the account I have sketched here is, I think, the most promising way to interpret talk of constitutive harm in this context. But alternative accounts have been given. Maitra sees constitutive harm as what occurs when a harmful illocutionary act is felicitously performed. I will discuss illocutionary conceptions of hate speech's indirect harm in §5.5. Another account of constitutive speech harm, due to Schauer (2000), likens hate speech to spitting on someone. The harm consists not in what the sputum does, but in how the act changes the target's relational properties, making him someone who has been spat upon. (See also Delgado and Stefancic's description of face-to-face hate speech as something akin to a slap in the face, 2009: 369.) This example may enrich our understanding of the harm associated with certain hate-speech-based verbal assaults. But with respect to the indirect, social-hierarchy-based harms of hate speech, this is not a useful elaboration of constitutive-harm-talk. Even if social hierarchies are partly a matter of relations similar to those in the spitting case, they surely involve more (and further-reaching) dimensions of inequality and disadvantage than such relations alone.

speech as an element in the networks, but there are many other elements besides. There are activities and institutions related to the distribution of material goods and resources. There are activities that involve physical violence, either institutionalised in the military, legal system, and police force, or unstructured on the street and in the home. There are activities involving cooperation and coordination in government bodies and other public institutions. And there are activities that involve the communication of ideas outside of, or adjunct to, these other areas: in the media, the arts, and in everyday conversation. Identity-based social hierarchies exist when, as a net outcome of the networks comprised by all these activities, certain markers of identity systematically correlate with inferior positions over multiple dimensions of social activity (see Haslanger 2004; Fricker 2007: 27ff). The question is how all the actions in these networks generate this state of affairs, and how hate speech functions in this.

Before we try to answer that question, we may ask what it means exactly for a practice to have a role in a dynamic outcome constituted by a causal network. One reasonable approach is to ask whether the practice's elimination would have a reformatory effect on the character of the system, all else being equal. In our case: if hate speech was eliminated, all else being equal, would society be less identity-hierarchical in its operations?[3] If not, then we cannot say hate speech contributes to identity-oppressive social hierarchies. The indirect harm rationale for anti-hate speech law accepts that single instances of hate speech may be redundant with respect to the relevant harms. But the practice of hate speech *at large* cannot be likewise redundant. It is insufficient, at least, to just observe that hate speaking occurs alongside social hierarchy, and has some kind of affinity with social hierarchy. After all, the co-existence and affinity of the outcome and the putative contributing factor may be explained by a causal relation running in the opposite direction. An ascription of contributory liability, under the indirect harm rationale, at a minimum requires us to diagnose the processes through

[3] Such questions might be formulated probabilistically, e.g. we might consider the likelihood of a specified reduction in some quantifiable aspect of identity-oppression if hate speech were eliminated. The correlative rationale would be that hate speech is liable to restriction so far as it increases the likelihood of certain negative outcomes associated with identity-based oppression (such an approach is suggested, for instance, in Schauer 1987, 2000; Nalezinski 1996; and Brown 2008).

which the allegedly contributory activity gets some causal purchase in the larger causal network through which the outcome is effected or constituted.

If this is how we think about causal contribution in complex networks, it is relatively easy to see how non-verbal factors make their contribution to identity-based hierarchies. *De jure* discrimination in family law is a clear route through which gay people face identity-based exclusion. *De facto* discrimination in education and employment are clear routes through which ethnic and religious minorities experience identity-based disadvantage. The question is: why suppose that hate speech, as a form of verbal practice, makes a contribution that is somehow comparable to these things? [4] Perhaps we suspect that language has a special power in shaping social forms, as in MacKinnon's claim that "authoritatively *saying* someone is inferior is largely how structures of status... are demarcated and actualized" (1994: 21), or in Onder Bakircioglu's assertion that hate speech plays "a major role in the construction of social reality [including in] the demonization of minority groups" (2008: 5), or in Shannon Gilreath's claim that "words create the hierarchies and people fill them" (2009: 604).[5] However, claims about the constructive power of speech and language *per se* are not enough here, because even if they can be given a credible elaboration, they still leave us to wonder why hate speech *specifically*, among the many forms of language that nurture identity-prejudice, is the one that is ultimately responsible for the relevant harms. Hate speech is the vulgar and conspicuous face of identity-prejudice, but identity-prejudice may be manifested – less overtly, and in a more reputable

[4] Nancy Fraser presses a similar point in her discussion of the different aspects of gender injustice, namely, (i) those that are primarily economic/distributive, and (ii) those pertaining to identity, recognition, and representation. The two kinds of injustices, she says, "are so complexly intertwined that neither can be redressed entirely independently of the other" (2007: 34).

[5] The constructivist ideas that underlie these claims are a recurrent concern in Charles Taylor's work on social structure and its relation to language. Taylor is interested in how accounts of human nature significantly draw upon understandings of language (1985b: 216), and in language's role in "making the discriminations which are foundational to human concerns, and hence opening us to these concerns" (1985c: 263), and, similarly, its role in "making possible new purposes, new levels of behaviour, [and] new meanings" (1992a: 248). Constructivist views of language also have a notable influence in some major modern critical works in the social sciences, like Said's *Orientalism* (1978) and Foucault's *History of Sexuality* (1979), which explore how value-laden classifications, about ethnicity and sexuality respectively, are represented as neutral descriptions of natural difference.

guise – in the media, the arts, academia, the church, and in certain sectors of serious political discourse.[6] To justify anti-hate speech law under an indirect harm rationale, we need to see not only that language shapes society, but that the brash vulgarities of hate speech in particular have this kind of constructive power.

## 5.2 Recognition and two senses of dignity

One thought might be that even though hate speech is brash and inarticulate, it still succeeds in the very simple aim of publicising a view of its targets as inferior or contemptible, and that this by itself does harm to people's dignity or esteem. The concept of dignity is often invoked in attempts to characterise the harms of hate speech.[7] And naturally so, given that we're considering status-based harms, and given that dignity is a moral concept uniquely attendant to considerations of status, rank, and hierarchy (see Walzer 1983: 258; Nussbaum 2000a: 123-24). Of course our concern is not with dignity in its aristocratic guise (the imagined thing that distinguishes gentlemen from commoners), but dignity in something like a Kantian sense: the inherent, shared store of worth in virtue of which all people have equal standing and are treated as equals in society. Talking about dignity discourages from thinking of the hate-speaker and his target as opponents in a private conflict, and encourages us to see the targets

---

[6] According to Teun van Dijk (1995: 24), responsibility for identity-based hierarchy ultimately lies with these elite forms of discourse, which, he says "preformulate, though often in more 'moderate' terms, the kind of modern racism that will then be taken up... by large segments of the general population".

[7] For instance: "the most stringent demands of... respect for the dignity of persons... pull strongly in favour of formally uniform hate speech rules" (Wright 2006: 569); "a society that permits toleration to run roughshod over human dignity deceives itself into thinking that undeserved abusive expression is worth the cost of denying many persons what is basic and decent" (Corlett and Francescotti 2002: 1097); "a racial insult is always a dignitary affront: a direct violation of the victim's right to be treated respectfully" (Delgado and Stefancic 1997: 8); see also David Kretzmer's defence of anti-hate speech law on the grounds that "certain interests, such as dignity, ought to be protected" (1987: 465). We also see appeals to dignity in hate speech legislation, e.g. in Germany, where "the Federal Constitutional Court has ruled that the Basic Law establishes 'an objective ordering of values' with some rights being more important than others" so that, "freedom of speech, press, assembly, and association are... inferior to the government's interest in securing and protecting human dignity" (Krotoszynski 2004: 1562-63).

of hate speech as people vying to secure esteem in a wider social ecosystem, and to see this aspiration as what is maligned in hate speech.[8] That is the upside of this terminology. The downside is that it is difficult to be clear about what an infringement of a person's dignity involves, given the ambiguity in our vocabulary for describing actions or practices that stand in a negative relationship to people's dignity. Dignity may be 'infringed', 'invaded', 'assaulted', 'trampled-upon' – or it may be 'respected', 'recognised', 'protected', and 'upheld'. When A insults B in a public forum, there is one obvious sense in which A's speech 'goes against' B's dignity, insofar as A publicises (to B and others) her own refusal to accord B the respect she is owed by virtue of her status as a fellow member of society. But how do we conceive of the antagonism between A's insult and B's dignity, beyond this observation? Inspecting the content of A's insult by itself cannot settle things, because there is a further question whether the insult will actually diminish B's standing in the eyes of B's peers and neighbours (or in B's own eyes). And if B's social esteem is unaffected, then it is misleading or overwrought to say that B's dignity has been 'assaulted' or 'trampled-upon' in A's speech. More broadly, the ambiguity is between the two senses in which someone's dignity can be acted against, one sense involving the actor's attitudes, the other involving social outcomes. Words like 'infringe', 'trample-upon', and 'assault' (on one hand) and 'respect' and 'uphold' (on the other) are ambiguous between the two senses. Some authors see it as uncontroversial that hate speech infringes its targets' dignity, and this seems rightly uncontroversial on the first sense. But the case for general legal restrictions on hate speech depends on us establishing that hate speech infringes against people's dignity in the second sense, by diminishing people's status in the wider social ecosystem, as opposed to it merely being a symptom of the speaker's identity-prejudicial attitudes.

[8] Some authors (e.g. Carmi 2007) oppose talk of dignity in discussions of hate speech precisely *because* it opens up a thematic connection between hate speech and identity-based hierarchy, and thus (arguably) lends some greater sense of moral urgency to the case in favour of anti-hate speech law. But this is evidently question-begging: if there are *prima facie* good reasons to characterise the indirect harms of hate speech in terms of dignity, we should see how that characterisation goes, and how the case for anti-hate speech law fares accordingly.

Steven Heyman's attempt to characterise the harm of hate speech as an infringement of dignity is unsuccessful, largely because it obscures this distinction.[9] On Heyman's view, A's dignity is upheld if, and to the extent that, A's personhood is recognised and respected by others – and this recognition is in turn, on his view, a necessary condition for A's political rights in general obtaining (166-167). An emphasis on recognition in moral and political theory is not at all eccentric. In the German Idealist tradition, Hegel, most notably, emphasises recognition – roughly, the apprehension of self-government in others, and the redirecting of one's own intentions and desires in light of this apprehension – as the foundation of ethical understanding (see R. Williams 1997). Among modern authors, Jürgen Habermas (1987: 41ff.) espouses a Hegelian ideal of recognition, as an alternative to the prevalent ethical frameworks of modernity,[10] while Axel Honneth (1995: 131ff.) argues that we should conceive of social conflicts generally as battles over the recognition of different ways of life. For others, the concept of recognition provides a locus around which we can seek to reconcile an individualistic view of rights with the need for respecting diversity that arises in culturally heterogeneous societies.[11] Heyman's account is situated within this family of ideas, but what is distinctive about it is that it calls for a 'right to recognition' to be enshrined via stringent legal obligations for individuals. On his view recognition requires not just that A is formally enfranchised in the eyes of the state, but that A's status, in virtue of which A receives that formal recognition, is also seen and honoured, informally, by everyone else in A's community. Heyman understands the harm of hate speech as a diminution of the target's dignity, enacted through the violation of this duty of recognition. The hate-speaker, in publicising his

[9] Parenthetical page numbers refer to Heyman 2009; this article elaborates views sketched in Heyman 2008.

[10] The counterpoint, so Habermas says, lies in the fact that where the ethical frameworks of modernity favour introspective attention to one's own desires and purposes as the starting point for ethical thinking, Hegelian approaches constantly direct our ethical attention toward the desires and purposes of others.

[11] Consider the following from Charles Taylor: "The thesis is that our identity is partly shaped by recognition or its absence... and so a person or group of people can suffer real damage, real distortion, if the people or society around them mirror back to them a ... contemptible picture of themselves" (1992b: 25). Note, however, that the recognition Heyman says people are owed is different to the recognition whose deprivation inflicts harm, on the view Taylor describes here. As Steven Rockefeller says, "from the liberal democratic point of view, a person has a right to claim equal recognition first and foremost on the basis of his or her universal human identity and potential, and not primarily on the basis of an ethnic identity" (1992: 88).

contempt for a particular social group, signals to others his non-recognition of the individuals who identify with that group, and thereby diminishes their social standing.

One way to understand this view is as part of a move away from institutional preoccupations in liberal egalitarian thinking. Will Kymlicka (2002a: 102-03) reminds us that "until the 1960s, the state itself was still engaged in explicit discrimination against women, homosexuals, racial minorities, and so on", and that many expected the decline of official state discrimination to give rise to an egalitarian milieu for historically subordinated groups. With these expectations having been disappointed in the post-civil rights era, one might think that governments must now promote equality in a way that goes beyond institutional egalitarianism, using their coercive power to influence civil society away from identity-prejudicial views. Heyman's right to recognition may be seen as a tool for mobilising such influence. His view also has affinities with the move among some liberals away from conceiving of citizenship as a set of status-based entitlements, towards seeing citizenship as encompassing both rights and duties (Kymlicka and Norman 1994). In this case, the individual's right to free speech might be understood as correlative with a duty to recognise the personhood of others.[12] But one could endorse these trends while rejecting the obligations Heyman imposes. Most liberals will accept that 'hate speech does not respect others' equality or dignity', but they see the crucial consideration that is 'at stake in evaluating the content of the legal order' as the respect or lack of respect that is demonstrated by *the state*, rather than by lone individuals (Baker 2009: 143). Heyman seemingly wants to deny the normative significance of that distinction. He sees person A's status as sensitive to, and able to be diminished by, the verbal mistreatment A receives in hate speech, even when that speech does not take place within the context of a government institution, and even when it does not represent the values of, and is not made on behalf of, the community at large.

---

[12] This appears to be what Jonathan Quong has in mind in his work on the rights of unreasonable citizens. "Unreasonable people have all the normal rights of citizenship" Quong argues, but "they cannot exercise these rights in the pursuit of unreasonable objectives". A white supremacist cannot be denied a general right to free speech because he denies the freedom and equality of persons, but "he can be prevented from exercising those rights when his aims are explicitly unreasonable – indeed they cease to be rights when he exercises them in this way" (2004: 332). (A similar line of argument appears in Waldron 1989: 518 ff.)

Now, it is clear that the hate-speaker's dereliction of his 'duty of recognition' is an infringement of dignity in the first sense, involving the speaker's attitudes. But why does Heyman think this can be equated with an infringement of dignity in the second sense, involving social outcomes? His idea, in short, is that person A's dignity is not only determined by her *de jure* legal status, but also partly by her *de facto* civic status – how she is viewed and the treatment she receives from all others in her community. The question, then, is why we should suppose that the way A is viewed and spoken of by the hate-speaker is of great enough significance to relegate A to an inferior status, given that she is not viewed or treated by most others, nor by the state itself, in this light. Consider the position of mixed-ethnicity couples in modern liberal societies. Today such couples have the same rights as others, both individually and with regards to the status of their relationship. Of course this was nowhere the case 200 years ago, and there remains a small minority in modern liberal societies who would prefer the legal and cultural norms of the early-19th century to remain in effect in this area. Now, imagine an advocate of 'traditional values' in this area, Z, who is so zealous in his views that he treats a particular mixed-ethnicity couple, A/B, with whom he interacts (in their mutual workplace, say) the way he believes they deserve to be treated. Thus, Z regularly condemns A/B's relationship, and tries within his sphere of influence to spoil and disrupt their pleasant and untroubled participation in society. If Z keeps at this long enough, he may be guilty of harassing A/B, and liable for harm through that connection. Whether or not this is the case, though, it seems clear that the imagined social order that underwrites Z's spiteful treatment of A/B does not, and without a seismic change in society could not, determine A/B's civic status (or the status of any other mixed-ethnicity couple). When Z treats A/B as second-class citizens, there is a sense in which, in addition to his patent moral failings, he is *factually mistaken*. The equal status of mixed-ethnicity couples in modern liberal societies is enshrined both in anti-discrimination legislation, and in informal attitudes and customs which largely (if imperfectly) affirm the ideals that underpin such legislation. The judgement that A/B's status is *non*-subordinate, in the contemporary context, is based as much upon the latter – there being support for the relevant norms of anti-discrimination – as it is on the fact that those norms carry the special imprimatur of being legally enforceable in certain domains.[13]

[13] This distinction is overlooked in Andrew Altman's arguments in defence of US campus speech codes. According to Altman, hate speech constitutes an illocutionary act of "treating someone as a moral

In rejecting the idea that formal legal apparatuses are the sole determinants of a person's civic status, an advocate for Heyman's account may be tempted to say that there are no abstract facts of the matter about what any person's social status amounts to, only facts about how an individual is treated by others. But this cannot be what Heyman thinks, because if there were no facts about people's statuses, there would be no intelligible rationale for singling out violations of people's putative right to recognition for condemnation. The view has to be that there *are* facts about people's status, and that these facts are determined by the way people are informally treated by others, in such a way that someone's being treated as a second-class citizen by being the target of hate speech means that, *ipso facto*, they are turned into a second-class citizen. Once the view is elaborated in this way, it is evident that it assigns too much weight to the attitudes and actions of those who are typically socially marginal individuals. When a political and legal regime is egalitarian only 'on paper', and is pervaded in reality by identity-prejudice and discriminatory practices, then it may indeed be the case that people in vulnerable groups have an inferior *de facto* status, notwithstanding their *de jure* equality. However, if a group's equal standing is affirmed both by the law and by widely-accepted social norms (as they now are in the mixed-ethnicity couple's situation), then merely being spoken of as socially inferior by an isolated bigot does not by itself confer a subordinate status on members of that group.

Heyman's view, if it is qualified in certain ways, contains an important insight. We will see the damaging effects of hate speech in a different light if we understand people's dignity as something constituted by the recognition people receive in *all* their social interactions, not just the recognition they receive in their *de jure* institutional status (this point is also stressed by Demaske 2004). However, in thinking about people's status from this perspective, what we encounter in modern liberal societies is a complex mixture of social forces, which includes some manifestations of identity-based prejudice, but which also includes many forces that reject and counteract those identity-prejudicial elements; including a (largely) functional and (largely) well-supported legal system, which purports to identify, mitigate, and redress identity-

---

subordinate" (1993: 310), and on this basis he argues that hate speech should be restricted in educational environments. However, Altman fails to acknowledge any difference between something being a 'speech act of subordination' (i.e. something that *actually* subordinates), and a speech act that merely treats someone *as though* they are subordinate.

prejudicial practices. Nothing in Heyman's account explains why, if we're appraising such a tangle of social forces, we should think hate-speakers decide the social status of people in vulnerable groups. Where *de jure* and *de facto* social equality both largely obtain, the hate-speaker's dissident voice by itself is surely insufficient to diminish people's status, and where *de jure* equality co-exists with forms of *de facto* inequality, we have no reason, apart from the mere fact of their 'alignment', to see hate speech as responsible for this larger *de facto* inequality. We can grant, again, that the hate-speaker assaults his targets' dignity in the individualistic sense that he personally fails to accord them the respect and recognition they are owed; but whether or not they are accorded these things more generally, by the wider community, is not, as far as anything Heyman shows us, determined by the conduct of hate-speakers.

## 5.3 Status equality and assurance

The two senses of dignitarian infringement that I have identified are distinguished by a contrast between the hate-speech-target's status as it actually stands in the wider social arena (which is the crucial consideration from a legal perspective), and the target's status merely in the eyes of the hate-speaker. Perhaps, though, this distinction overlooks the importance of how hate speech affects its targets' perceptions of their own status. Waldron's conjecture is that hate speech infringes the dignity of its targets through this avenue: not by degrading their equal status as it is enshrined in law or manifested in egalitarian attitudes and practices, but by undermining the targets' *assurance* as to the security of their dignitarian status.[14]

Waldron thinks of hate speech as a type of group defamation. Given this initial thought, we may imagine anti-hate speech laws modelled on the tort of libel, aimed at

[14] Here I focus on the account Waldron puts forward in 'Dignity and Defamation' (2010). The key arguments have appeared in several guises. They first emerged, embryonically, in 'Free Speech and the Menace of Hysteria' (2008). The 2010 article that I will be focusing on is a published version of the 2009 Oliver Wendell Holmes lectures, delivered at Harvard Law School partly in response to the reception the 2008 article received. The arguments have since reappeared, with minor amendments and additional discussion of surrounding issues, in *The Harm in Hate Speech* (2012). Parenthetical page references here refer to the 2010 article, with one noted exception.

restricting speech which impugns the reputations of individuals indirectly, *by* impugning the reputation of groups they identify with.[15] Whatever the merits of that approach may be, however, Waldron's account of anti-hate speech law is modelled on *criminal* libel statutes. Criminal libel has largely been cast aside in modern legal systems, but the onetime purpose of this offence, on Waldron's view, provides an instructive model for our thinking about hate speech law. Where civil libel allows individuals to protect their reputations against attack, criminal libel statutes were a way for states foster the orderly social conditions necessary for the maintenance of people's reputations generally. How do libellous statements corrode public order? On one hand, libellers may provoke outrage and a desire for revenge.[16] But Waldron draws our attention to a more subterranean relationship. Criminal libel statutes aim to protect a crucial precondition for orderly social relations: a widely-shared understanding and acceptance among the populace of the status-respect that we all owe each other as members of society.[17] Libellous statements jeopardise the acceptance of our mutual standing *qua* citizens, on which public order relies in turn (1604-05). Waldron's suggestion, then, is that anti-hate speech laws can be seen as operating in the same way as criminal libel statutes once did, as a way of protecting these dignitarian foundations of public order.[18] Anti-hate speech

---

[15] Something along these lines is espoused in Lasson (1985); Jones (1998) endorses anti-hate speech law on the basis of harm to group reputation as such, rather than harms done to individuals via group reputation.

[16] For an argument to this effect, in relation to what we would nowadays call hate speech, see Hughes 1966.

[17] There are traces here of Waldron's thoughts on Locke and the theological foundations of moral equality. In the *Letter Concerning Toleration* (1689), as Waldron explains (2002: 226-28), it is the atheist's inability to internalise the concept of human moral equality that necessitates the intolerance he receives, since, for Locke, this concept is essential to underwriting political doctrines concerning property, political representation, and the common good. Waldron obviously rejects Locke's conclusions about the atheist, but he is sympathetic, at certain points, to the anxiety manifested in the argument that leads Locke there. When Waldron says "equality cannot do its work unless it is accepted among those whom it consecrates as equals" (2002: 243), he simultaneously recalls Locke's doubts about what would remain of a political morality cut loose of theological foundations, and anticipates his own claim that anti-hate speech law is needed to secure the basic, dignitarian constituents of an egalitarian social order.

[18] The notion that anti-hate speech law somehow protects public order has occasionally been mooted by other authors (e.g. Tsesis 2009). However, Waldron articulates this relationship more perspicuously than the

laws, he says, aim at "the protection of a certain sort of precious public good: a visible assurance offered by society to all of its members that they will not be subject to abuse, defamation, humiliation, discrimination, and violence on grounds of race, ethnicity, religion, gender, and in some cases sexual orientation" (1599).

The concept of 'assurance' is central in Waldron's account. In a well-ordered society the state is committed to equality and justice, and structures its institutions accordingly. But these commitments, for Waldron, do not by themselves *make* a society well-ordered; a further condition is that the conspicuous features of the sensory environment must support (or at least, not undermine) the state's commitments to justice and equality.[19] Why? Because people need to be able to feel *secure* in these commitments obtaining. The assurances in question are mostly implicit and unspoken, and for this reason, Waldron says, they are easily susceptible to being undermined, particularly for members of vulnerable groups (1634). Even if state Q's labour laws are governed by egalitarian ideals, formally *and* effectively, it still may only take one person's public proclamation that 'terrorist immigrants are stealing our jobs' to undermine, for many of Q's workers, the confidence that they are fully-entitled member of Q's workforce. Given the fragility of assurance, Waldron argues that "the prime responsibility... that falls upon the ordinary citizen is to refrain from doing anything to undermine it or to make the furnishing of this assurance more laborious or more difficult", and on his view, this is the obligation that hate speech laws enforce (1630).

The reference to 'the well-ordered society' and the idea of assurance as an element of social well-ordered-ness, are both things Waldron borrows from John Rawls (e.g. Rawls 1996). Waldron does not attribute a position on anti-hate speech laws to Rawls. Rather, he sees himself as adopting conceptual tools and normative themes from Rawls' theory to formulate an argument in defence of such laws. Now, there is a question about whether the concept of assurance at work in Waldron's argument bears more than a nominal relation to Rawls's

---

others who have adverted to it, his key claims being that public order only obtains when certain public goods are secured, and that hate speech threatens public order by jeopardising one such good.

[19] Waldron is especially concerned with *visual* manifestations of identity-prejudice. His work thus intersects with the small subset of the hate speech literature that focuses on non-verbal representations of identity-prejudice, e.g. Delgado and Stefancic 1992a; Whillock 1995; Levinson 1998; Zick 2004; and Newman and Rackow 2011.

claims about assurance. For Waldron, assurance is about "conveying to people a sense of security in their enjoyment of their most fundamental rights" (1626). For Rawls, by contrast, the need for assurance comes from a game-theoretic 'assurance problem', to do with how we coordinate expectations based on fair-minded rules of cooperation.[20] In any case, so far as Waldron's argument is concerned with how the bare visibility of hate speech undermines assurance (as a felt sense of security), it sidesteps many of the empirical uncertainties that arise in claims about hate speech's indirect harmfulness. Waldron is not treating hate speech as the causal root of identity-based social inequality. Rather, he is just observing that in societies in which hate speech freely circulates – on the internet, on talk radio, in books and pamphlets – it becomes a fixed feature of the social environment. The observation is straightforward, but Waldron's insights are about why it matters from a liberal perspective: (i) because providing assurance is an important part of securing people's dignitarian status, and (ii) because the assurance of vulnerable groups is a delicate undertaking, which is jeopardised when the social environment is 'disfigured by hate', even if only in isolated pockets.

This seems plausible in its basic outline. Presumably part of the reason we think status matters is because we see that social hierarchies are not only unfair, but also painful or humiliating for subordinated parties. An appeal to the good of assurance may be partly motivated, then, by the insight that the distress of feeling oneself to be subordinate might be

[20] For a discussion of assurance problems in political theory see Runge 1984. In Rawls, roughly, the idea is that if reasonable people are to comply with institutions that limit their liberty, they need to be assured that others will also comply, so that in complying one is not making oneself vulnerable to exploitation and free-riding. Certain social institutions like the criminal law may then be defended in view of their ability to provide such assurances. How does Waldron's view of assurance depart from this? One difference is that in game-theoretic assurance problems, as in Rawls, everyone stands to gain equally from assurance, since everyone depends on the institutions that assurance supports; indeed, this is the unifying feature of contractualist assurance games (Skyrms 1996). In Waldron's account, on the other hand, the assurances cut across antagonisms within the community, protecting one group against the hostile incursions of another. Another difference is that for Rawls, although the provision of assurance is intended to foster a subjective sense of security for subjects in their basic rights, the moral import of that secure feeling derives from its tendency to promote institutional compliance, and the import of *that* derives in turn from the fact that compliance is necessary to sustain just institutions. In Waldron's account, by contrast, providing a subjective sense of security seems to be viewed a moral end in its own right. That is not inherently implausible, but it does seem different to how Rawls understands the rationale for assurance.

similarly severe to the distress of actual subordination. It is still not clear, however, that hate speech is to blame for this problem. There are two routes Waldron takes in explaining why a concern for assurance requires restrictions on hate speech, and both routes run into difficulty. At certain points, Waldron appears to be saying that anti-hate speech law is an effective means for protecting the dignity-respecting elements of social order, and for improving things in this respect in the future. That reading is encouraged by a passage like this:

> ... even if a well-ordered society could dispense with laws prohibiting group defamation, it would be a mistake to infer from this that the societies we know must be prepared to dispense with those laws, as a necessary way of becoming well-ordered. Societies do not become well-ordered by magic. The expressive and disciplinary work of law may be necessary as an ingredient in the change of heart within its racist citizens... [If] the good to be secured is a public good, a general and diffuse assurance to all the inhabitants of a society concerning the most basic elements of justice, then it is natural to think that the law would be involved... (Waldron 2012: 81)[21]

On one hand, I think Waldron is right in challenging the Millian faith that identity-prejudice will wither on its own in the marketplace of ideas, and suggesting that the law may have to get involved for progress to be made.[22] But on the other hand, the argument from assurance is robbed of its initial appeal when it is elaborated in this way, as an appeal to the potential long-

---

[21] Here I quote from the more recent version of Waldron's argument, since this version uses more circumspect language than the corresponding (otherwise similar) passage in the 2010 article (at 1623).

[22] The view Waldron rejects is epitomised in remarks like the following from Richard Posner (1986: 31-32), who says "the more remote the danger, the more time there is for the marketplace in ideas to prevent the danger from materializing by correcting any falsities in the speech"; and from Thomas Nagel (1999: 41), who insists that anti-hate speech law is "extremely unhealthy, with its suggestion that the opposite, right-thinking view is a dogma that cannot survive challenge, and cannot be justified on ordinary rational and evidential grounds. The status of blacks and women *can only be damaged* by this kind of protection" (my emphasis). Some theorists drawn to the marketplace model are more sensitive to its limitations, taking note of social research which shows that very subtle institutional changes can significantly influence the character of public discourse – determining, for instance, whether it generates thoughtful debate or encourages sectarianism and polarisation (see Sunstein 2000, 2003).

term benefits of formally deterring the most overt expressions of identity-prejudice. Here we move into the realm of consequentialist speculation: anti-hate speech law *may* influence society away from identity-prejudice over time, but it may not. Whether that is the case depends on how important a role hate speech plays in the complex causal networks out of which identity-based hierarchies arise and are sustained. Hate speech might have an integral role in these causal networks, or it might have a trivial, incidental role. That is the crucial question, and it is not settled by the observation that assurance is an element of dignitarian status, and by our noting the possibility that restrictions on hate speech may be a good way to promote assurance in the long run. Under the first interpretation, these modest conclusions are all that Waldron's argument leads us to.

At other points, however, Waldron's argument from assurance seems to be appealing to the immediate harm of undermining the hate-speech-targets' felt security. Even if person A is a fully-entitled member of society by *de jure* right, the insecurity that comes with knowing she is despised by some of her neighbours may leave her humiliated, cowed, and unable to live the way that this status should make it possible for her to live. This reading of Waldron's argument is encouraged by passages like this:

> We are talking about a display [in hate speech] that matters practically to individuals. It matters to them in their reliance on the principles of justice in the ordinary course of their lives, and in the security with which they enjoy that reliance. In a well-ordered society... People know that they can reasonably count on not being discriminated against, humiliated, or terrorized. They feel secure in the basic rights that justice defines; they can face social interactions without the elemental risks that interaction would involve if one could not count on others to act justly. (1627)

In one respect, this version of the argument is on surer footing. There is no need for speculation here about how the toleration or restriction of hate speech may affect society in the long run. The claim is that hate speech deprives its targets, here and now, of something they are owed, namely: a felt sense of security in their dignitarian status, which is partly constitutive of their having that status in the fullest and most complete sense. The problem with this elaboration of Waldron's argument is that it paints a picture on which hate speech is not the actual source of the threat to people's assurance. What has gone awry, when assurance fails, is that people cannot reasonably *count on* not being humiliated, terrorised, or

discriminated against. If hate speech was not backed by genuine prejudice or hostile intentions – if it was merely superficial, loud-mouthed abuse – then it seems doubtful that people's sense of assurance about their status would be seriously threatened. Hate speech gets its potency, in real-life cases, when and where it *does* reflect deeper, sinister currents of identity-prejudicial malice. But then, it is surely the underlying enmities which prevent people from being able to count on not being discriminated against, humiliated, or terrorised. Granted, hate speech can be involved in the process through which these underlying states of affairs become known about. However, for people in vulnerable, marginal, and historically-oppressed groups, the legal restriction of hate speech will not alleviate the assurance-eroding knowledge that there are others in their wider political community who feel contempt or hostility towards them. That knowledge can be conveyed and circulated in all sorts of indirect, coded ways. And of course, it can be – and where anti-hate speech laws are in effect, already is – powerfully conveyed in people's preparedness to express their identity-based contempt *even while faced* with the threat or reality of prosecution.[23]

It is useful to reflect on how we might relate to someone who lived in a society in which she was subject to identity-based prejudice, but in which she failed to realise that this was the case. One thought might be that someone ought to tell this person what she has not yet apprehended. Perhaps this will help her to keep on her guard against those who harbour maleficent intentions towards her.[24] Perhaps there is a sense in which facing up to painful

---

[23] Some argue that restrictions on hate speech will inadvertently energise its proponents. Michael Curtis, for instance, says that silencing bigots is "more likely to temporarily repress problems than to resolve them", because bigots will "simply find cleverer and more appealing ways to package the message" (2009: 494). Although I accept that there is something important underlying that concern (which I will return to in §5.9), my point here is not that silencing hate speech makes hate speech more potently harmful. My point is that *if* the harm of hate speech consists in its making identity-prejudice widely known, prohibiting hate speech will not prevent *that* harm, because hate speech performed in the face of prohibition exhibits identity-prejudice just as plainly, if not more so, than hate speech performed under legal tolerance.

[24] This recalls an argument in the hate speech literature (e.g. in R. G. Wright 1988: 24-25), that hate speech should be permitted because it functions as a 'bellwether' indicating the prevalence of prejudice in the wider community (I borrow the term 'bellwether' from Delgado and Stefancic 1995: 484; Delgado 2006: xi). As I say above, however, it seems unlikely that this benefit, whatever it amounts to, would be threatened by anti-

truths about how one is seen by others is itself partly constitutive of having one's dignity in the richest sense.[25] In any case, even if we do *not* ultimately think it is a good idea to relay the information that would disabuse this person of the assurance they feel, there is a question about how we should understand the loss that the person experiences when, having been confronted with hate speech, they come to see their status in the community as tenuous or threatened. If, when the hate-speaker's malevolent attitudes are brought out into the open, it is the fact of the malevolence itself that is ultimately responsible for the erosion of people's dignitarian assurance, then a dispiriting conclusion is awaiting us. When a group of people is held in contempt by others in their community, there is no way to secure the dignity of those belonging to the group in the fullest sense, short of eradicating the contempt itself. What we want is a social milieu in which people know they will not be discriminated against or terrorised on account of their identity, because, as a matter of fact, these thing will not happen. Granted, this aspiration brings us back to the first interpretation of Waldron's argument. Anti-hate speech law is one of the measures that we could employ to try to advance our societies along this trajectory. My response on that front was that we need more evidence that the restriction of hate speech would in fact be a fitting and effective measure to this end. As far as the second strand of the argument goes, I believe Waldron misidentifies the source of the problem. If hate speech undermines dignitarian assurances, it does so indirectly, by manifesting attitudes in view of which people may entirely rationally come to doubt the security of their status as fully-accepted and fully-entitled members of society.

## 5.4 Legitimation and authority in hate speech

Attempts to ascribe a dignity-based harm to all hate speech, like Heyman's and Waldron's accounts, work by presenting an original characterisation of social structure, and of how people's status within that structure can be harmed. Heyman suggests, implausibly, that being publicly derogated in hate speech *ipso facto* harms a person's status. Waldron fares better, identifying a credible and overlooked aspect of civic status – namely, assurance about the

---

hate speech law. If legally tolerated hate speech helps people to notice and ascertain the dangers of prejudicial enclaves in their society, hate speech performed under threat of punishment surely does the same.

[25] This prospect is explored Dan-Cohen 1999: 402ff.

*security* of one's civic status – and a related form of status harm. However, he implausibly sees hate speech, rather than the malice that compels it, and which endures even under anti-hate speech law, as the fundamental source of the status harm in question.

Unlike Heyman and Waldron, most authors who purport to identify a far-reaching, indirect causal connection between hate speech and identity-based inequality do so without offering a novel characterisation of social structure and/or the bases of dignity and status. Instead, most who undertake this task try to show how hate-speakers, despite their apparent powerlessness, nevertheless exert a potent influence in favour of identity-prejudicial social arrangements. In the words of Bhikhu Parekh (2006: 314), the idea is that identity-based social hierarchy is "legitimised by a wider moral climate which is built up and sustained by, amongst other things, gratuitously disparaging and offensive remarks... collectively contributing to the dehumanisation or demonisation of the relevant groups". To say hate speech legitimises social hierarchy is not to envisage a sinister conspiracy whose covert machinations allow hate-speakers to influence society in a manner akin to the influence of heads of state or high court judges. The claim is just that, whatever else it does, hate speech as a form of expressive practice makes identity-based hierarchy seem, to many people, natural and normal. What this calls for, though, is an explanation of how this effect can be achieved by low-status speech like hate speech: speech that is largely marginalised in the mainstream news and entertainment media, which is almost entirely absent in expert discourses in academia and professional life, and whose proponents are, by and large (by conventional social markers), marginal and low-status members of society. There are many disreputable folks espousing disreputable views in the world. We don't think that Ayn Rand's acolytes are responsible, with their many online screeds, for the erosion of social justice under late capitalism. Why suppose, then, that hate speech has any real society-shaping force? Perhaps, one might conjecture, there is a uniquely collaborative type of communicative dynamic in hate speech, where speakers perform the task of articulating an identity-prejudicial worldview, and audiences perform the task of rallying behind and broadcasting that view, so that identity-prejudice is normalised on all sides. This may be roughly how things work in reactionary nationalist political groups, with leaders and followers interactively facilitating each other's prejudice.[26] But if this is how hate speech in

[26] This kind of dynamic might explain why some authors perceive a particularly grave threat to social equality in hate speech that is circulated in online communities; see for instance Breckheimer 2001, Leets 2001.

general achieves its legitimating effects, it is unclear why hate-speakers should be specially liable. Under an indirect harm rationale, if one purports to restrict hate speech because it legitimates inequality, one's judgement must be that hate-speakers exercise some power or authority *over* the beliefs and attitudes of their audiences or society at large. To substantiate claims about hate speech's legitimating operations, then, we need an explanation of how hate-speakers attain such authority. And it is not enough simply to say something like that the hate-speaker 'speaks with the force of an oppressive structure behind him' (Schwartzman 2002: 431), or that there is a "language-culture cycle that keeps oppressive categories intact" (Tirrell 1998: 288). We need a specification of the hate-speaker's role in events that makes sense of our singling-out the hate-speaker for special disapprobation and liability.

Perhaps the hate-speaker has authority not because he occupies any stable *position* of authority, but because the views he advances are in some sense uncontested. Gilreath says that social hierarchies are created by "speech that says, quite authoritatively because it encounters no official resistance, that a group is second-class or no-class" (2009: 604). Matsuda says that "government protection of the right of the Klan to... spread a racist message promotes the role of the Klan as a legitimizer of racism" (1993: 49). Sumner says that for a government to allow hate speech "would inevitably be to confer upon [the hate speech] a certain degree of legitimacy" (1994: 172-73).[27] Douglas Wells suggests a slightly different and perhaps stronger version of the 'authority-as-non-contestation' thesis; he claims that racist hate speech "is particularly devastating, because many in the targeted group are likely to fear that society tolerates the message of the hate speech, because a sizable portion of the white population at some level condones, if not endorses, the message" (1995: 321). In this case the thought appears to be that the hate-speaker's authority is not the delegated formal authority of the state, but rather the delegated informal authority of society at large, which supports the hate-speaker's message tacitly, through a collective failure to contest it.

The first thing to say here is that even if we grant that non-contestation authorises, it is just false that in modern liberal societies which do not restrict hate speech, like the US, the

[27] Although in Sumner's recent work (2004, 2009) he renounces his earlier support for anti-hate speech law.

message of hate speech goes uncontested.[28] Identity-prejudice is contested by governments, in their ongoing efforts to reform social institutions along egalitarian lines, and by private individuals and organisations. Countless civil-society institutions, including churches, schools, universities, community groups, and sporting organisations, promote civility and oppose prejudice, not merely by taking a formal stance to this effect, but by modelling that stance in their institutional activities. Remarks like those in the previous paragraph simply equate some *forms* of non-contestation – e.g. an absence of 'official' resistance (Gilreath), legal toleration (Sumner, Matsuda), or tacit social consent (Wells) – with non-contestation *per se*. Moreover, at least in Matsuda's and Sumner's claims, there is evident confusion about the difference between tolerating and promoting a state of affairs. Tolerant policies rest on judgements about how countervailing reasons should be balanced. In short, to tolerate x is to permit x while recognising *pro tanto* reasons against doing so (Green 2008). Where hate speech is tolerated, its toleration rests on the judgement that the reasons against its restriction, such as they are, are weightier than the reasons for its restriction. Tolerators of hate speech can agree with advocates of anti-hate speech law about the reasons in favour of restriction, while disagreeing about the appropriate policy all things considered. To say the tolerators promote hate speech in such cases is just to efface any distinction between toleration and promotion.

If there is a kernel of truth in the claims about non-contestation, it is that where anti-hate speech law is not in place, the state foregoes an opportunity to *ensure* the contestation of hate speech. Other kinds of (informal) social opposition to hate speech may be little help in those situations and segments of society where hate speech is readily accepted and circulated. These are the segments Ishani Maitra (2012) is thinking of when she claims the hate-speaker's authority comes from its sentiments being licensed by people who witness hate speech (as third parties) and do nothing to respond. Having equated authority with tacit licensing in this way, Maitra says it is a further question, in hate speech or in any other case, as to the range of circumstances over which the authority extends (2012: 102). It is true that we can think of authority that way if we want – separating the questions of (i) whether it exists and (ii) how far it extends – but if we take that approach, everything hangs on the second question, what

[28] Of course, we might reject the view that non-contestation authorises in the first place, since this view fails to recognise the modally robust character of authoritative power (the idea being, roughly, that it is characteristic of authoritative commands that they enact norms *irrespective* of whether they are contested).

Maitra calls the 'jurisdictional aspect' of authority. Conjectures about hate speech legitimating social hierarchy, and thus indirectly harming its targets, receive no support if it turns that the hate-speaker's putative 'authority' is confined to the imminent context in which his claims are licensed or supported by the inaction of others. To say that hate-speakers have authority in *that* sense is just to say that in certain places and at certain times, hate-speakers speak with local support (or without local opposition). It remains implausible, in that case, to say that hate speech legitimates social hierarchy by virtue of it being *generally* uncontested.

ଔ

Consider another possibility: perhaps hate speech functions as if it were authoritative, not because it is backed by any social, intellectual, or moral authority, but because it suffuses our discussions of human identity and difference in a way that conditions the attitudes of impressionable individuals, in particular, children and adolescents. Whatever wider credibility accrues (or fails to accrue) to the notion that language constructs the social world, this notion has an enduring influence in developmental psychology. Piaget's *The Construction of Reality* (1954) defines each stage of cognitive development as an extension in the child's capacity to construct a world-image through sympathetic identification with other people's points of view. It is easy to imagine how patterns of identification and non-identification laid down in Piagetian developmental stages might prefigure prejudices later in life (see Cortese 2006: 144-46). Some authors stress the negative impact of hate speech expressly targeted at children, for example, its tendency to foster a demeaning or belittling self-conception (see Delgado 1989: 2437; Sunstein 1991: 31-32).[29] Equally important for our purposes, however, is the influence that hate speech may have on children in privileged demographic groups, who are (so the claim goes) conditioned into identity-prejudice before they possess any resources for critical appraisal and resistance.[30] While the child is acquiring the rudiments of her self-image and her

[29] Delgado and Stefancic also talk of how hate speech "begins to shape and determine us, who we are, what we see, how we select, reject, interpret, and order subsequent reality" (1992a: 221); and they observe that "from as young as three, children make value judgements about their racial identities based on the views they hear expressed by others" (2004: 94-95).

[30] The worry has been around for some time: "I am unable to understand why there is so little support... for governmental action designed to... prevent the indoctrination of children which produces adults who seek

conception of society, the hate-speaker – by stirring up powerful anxieties about identity, community, and security – lays down an identity-prejudicial template in her attitudes. And he can do this even while his speech has a negligible influence and lowly reputation in society more generally.[31] If nothing else, thinking about the impressionability of children casts doubt on certain crude liberal arguments in favour of free speech. For Dworkin (1985: 339), part of the reason to protect speech is because "human development must be self-development or else its value is compromised from the start". Thus, he argues, the state cannot try to make us good or wise by silencing bad ideas. But young children have a very meagre capacity for *self*-development at best. There is no control centre from which children reflectively govern their first-order responses to ideas and opinions. By shielding young minds from hate speech, one might argue, we do not stymie their development. Rather, we aid it, by warding off the spectre of a distorting, identity-based contempt for others in one's society.[32]

Some of the concerns here can be addressed straightforwardly. If there is evidence that children targeted in hate speech suffer direct psychological harm, this is a *pro tanto* good

---

relief in persecution. Do we believe that impregnation with hate ought to be tolerated so that all of us may be enlightened and ennobled by the debate it excites?" (Hyneman 1962: 849).

[31] This conception of human cognition is evident in political theory that emphasizes the dialogical nature of human identity. "We become full human agents, capable of understanding ourselves, and hence defining our identity" says Taylor, "through our acquisition of rich human languages of expression... People do not acquire the languages needed for self-definition on their own. Rather, we are introduced to them through interaction with others who matter to us" (1992b: 32).

[32] This line of argument regularly crops up in feminist critiques of pornography. Langton (1998) suggests that pornography's authority ultimately resides in its captivating influence on boys who, in their sexual naïveté, come to regard pornographic representations of women's sexuality as veridical and action-guiding. The thought also occurs in Langton's work in feminist epistemology, where she critiques the general notion that knowledge is shaped by the world, arguing instead that men's putative knowledge about women (e.g. drawn from cultural stereotypes) is itself something that shapes women into being how they are known (by men) to be (2009b: 280). Bringing the two strands together, Langton claims that the knowledge pornographers create *about* women can be a kind of harm inflicted *against* women (2009c). In these works Langton is exploring constructivist aspects of MacKinnon's critique of pornography (e.g. in 1987a; 1994), in particular, MacKinnon's claim that although women are constructed by pornography, the construction is false. Pornography, for MacKinnon, misrepresents what women are actually like, even as it partially succeeds in transforming society to more closely resemble its misrepresentations.

reason to have especially stringent measures restricting hate speech directed against children. But other problems raised by thinking about children's exposure to hate speech run deeper. Can we treat the family as a social institution for the purposes of preventing children's exposure to hate speech's debasing influence? Are parents who see an identity-prejudicial belief-system as integral to their way of life entitled to pass that on to their children? And if we grant this entitlement provisionally, where is it overridden by individual imperatives (e.g. the need for children to be free to decide for themselves whether or not to accept an identity-prejudicial belief-system), or by corporate imperatives (e.g. the need for societies, if they are to attain ongoing stability, to oppose destructive enmities within their borders).

At their root, questions about when the interests of society override the parent's child-regarding preferences descend from larger problems about how we understand social responsibility in the relations between individuals in families. And of course there is more riding on this than just hate speech policy. If we accept that we cannot make any real progress on these big-picture problems here, and bracket them off accordingly, what we are left with is a question about how, when people *do* come to see identity-based hierarchies as natural or normal, their prior, incidental exposure to hate speech as children contributes this outcome. Thus construed, however, the question of hate-speakers' authority relative to children just leads us back to the problematic outlined in §5.1, where we are trying to ascertain how the myriad practices in the causal networks that constitutively give rise to social hierarchy all work together, and why hate speech might be regarded as a component whose elimination would destabilise or reconfigure the identity-prejudicial character of that causal network. Identity-based hierarchies can be socially normalised, but this is not as simple as people being exposed to hate speech as children, so that they later accept, as adults, a system where some of hate speech's values are manifested in social arrangements. What we are initiated into as children, rather, are complex ways of life, with a range of activities and practices which jointly constitute identity-based social hierarchies.[33] Verbal patterns and practices are part of this. But our

[33] Amartya Sen (2006) argues that essentialist ideas about ethnic and religious identity are perpetuated as much by well-meaning bureaucratic initiatives (such as government aid in the proliferation of religious schools), as they are by sinister or corrupting communicative practices. On Sen's view, bureaucratic initiatives reinforce misguided ideas that cohere with the hate speaker's perspective, for instance, his crude understanding of religious identity as monolithic and alien to mainstream society.

impressionability as children is obviously not confined to what is conveyed and habituated in speech, while being absent in our experiences of family, friendship, romance, education, work, finance, law, and all the other parts of human life through which identity-prejudice is sustained. Thus, without any independent reason to see hate-speakers as authoritative, observations about children's impressionability do not support the claim that hate speech plays a significant role in legitimating inequality and identity-based hierarchies.

## 5.5 Authority in illocutionary action

Feminist critiques of pornography suggest a different approach to understanding the authority of hate-speakers. Andrea Dworkin and Catharine MacKinnon famously claim that pornography is – not just causes, *is* – the subordination of women in pictures and words (1988: 138-39). While some hastily dismiss this as rhetorical bluster (e.g. Parent 1990: 209; Strossen 1995: 121), Rae Langton shows how this view of pornography can be interpreted, via Austin's speech act theory (see §2.5), as a claim about the illocutionary acts that pornographic speech is used to perform. In addition to the locutionary content of pornography (what it *says* about women and sex), and its perlocutionary effects (its capacity to arouse, titillate, etc.), pornographic expression is employed to *do* those things that MacKinnon and Dworkin ascribe to it, specifically, to subordinate women (Langton 1993: 297).[34]

Authority in illocutionary action is not identical with authority *per se*. Speech acts typically operate via conventions: their successful performance requires "an accepted conventional procedure having a certain conventional effect, the procedure to include the uttering of certain words by certain persons in certain circumstances" (Austin 1962: 26).[35]

---

[34] In relation to hate speech, the notion underwriting this kind of claim – namely, that harm is done in the speech act itself, regardless of eventual consequences – is implicit in Wright's talk of hate speech's harm consisting in its "elemental wrongness" (1988: 6), and in Delgado and Stefancic's claim that identity-prejudicial speech is a "*per se* harm", which "occurs irrespective of what happens later" (1992b: 1048).

[35] Among early proponents of speech act theory, Searle departs from Austin's position slightly, for although he agrees that in general there must be "some conventions or other... in order that one can perform illocutionary acts" (1969: 38), Searle also maintains that in a narrow range of cases, successfully communicated expressive intentions make illocutionary acts possible without the use of convention.

Standardly construed, a convention is a regularity in conduct that is sustained because people conform to it, motivated (partly) by the conformity of others (see Lewis 1975: 5-6). Authority in illocutionary action, then, means being the right person in the right circumstances to say something that will trigger a generally accepted regularity of conduct.[36] Authority in an *exercitive* act like subordination – that is, an act which changes what is permissible or prohibited for others – means being the right person in the right situation to say something that will then enact conventionally accepted permissions or prohibitions. Notice, however, that with regards to authority *per se*, as opposed to authority in illocution, no such serendipity is needed to enact prohibitions. The master can enact prohibitions for the slave no matter the circumstances. The crucial question, then, about Langton's idea of subordinating illocution, is whether there even *exist* exercitive conventions that can be triggered by expressive acts like pornography or hate speech. Identity-prejudicial expression elicits a range of responses: some rally around it, while others fight back. In small-scale, local circumstances, there may exist embryonic regularities in response to would-be subordinating speech. But these do not become general conventions unless they are accepted by all the parties to whom the nascent conventions are purported to apply, and where this would require certain groups of individuals to accede to their own subordination, such conventions are unlikely to catch on.

Granted, there are some ways in which hate-speech-acts make use of conventions. There are linguistic conventions in light of which certain phonemes constitute utterances (see Strawson 1964: 442), and there are semantic conventions in light of which, say, certain epithets convey hostile content. What we would need to see, however, for there to be subordinating illocutions, would be conventions through which speech triggers wide-scale patterns of (subordinating) conduct. Illocutionary action may be possible without convention if and when all that illocutionary success requires is *uptake*, i.e. the speaker's expressive

[36] To see how conventions make illocutionary action possible, consider the illocutions performed by a judge, of sustaining or overruling objections in a courtroom. There are conventions regarding *what* must be said in order to perform these act, *when* these things are to be said relative to other utterances, the conditions under which a ruling one way or the other is appropriate, and how all the parties must behave depending on the judge's ruling. There are also conventions governing how people become judges, how judges behave in the courtroom, and what qualifies judges for presiding over cases. The acceptance of these regularities of conduct on the part of parties engaged in the legal system is what empowers the relevant illocutionary acts.

intention being understood by the audience.[37] But much more is needed for the 'success' of a subordinating illocution: no subordination occurs if no-one is in fact relegated to an inferior status. In order for speech acts to subordinate, what needs to be secured is not just uptake, but *compliance* with the permissions and prohibitions conveyed in the putatively subordinating illocution. Securing compliance with norms that are not already mobilised in generally-accepted conventions, however, will not just be a matter of saying the right words in the right place and time. To secure others' compliance with one's intended prohibitions is something that can only be achieved by speakers with some independent social power.

Thinking about the legitimation of inequality from a speech act theory perspective ultimately returns us, therefore, to questions about the *per se* authority of the pornographer or hate-speaker. If there were conventional illocutionary acts of subordination, then hate-speakers might succeed in subordinating (or legitimating the subordination of) their targets, merely by deploying their hate speech appropriately. But there surely aren't any speech-to-conduct conventions of the kind that would make the hate-speaker's task so easy. The hate-speaker may *want* to elicit from his audiences a structured response that enacts his identity-prejudicial purposes. But in order to elicit that sort of response, he needs his audience's compliance, and securing this requires more than mere illocutionary authority. Langton recognises this in her account of pornography's subordinating force. The cases she uses to illustrate subordinating illocutions (e.g. the passing of apartheid legislation) are cases in which speech evidently *is* backed by authority, and she uses these cases precisely to suggest that pornographers do possess considerable social power after all, not over a wide spectrum, but in the domain of discourse on sex and women's sexuality. The suggestion may be controversial,[38]

---

[37] Both Strawson and Davidson observe that speech acts like *warning* or *objecting* can be performed without the speaker having to trigger speech-to-conduct conventions. Strawson adverts to scenarios in which contextual factors betray an utterance's illocutionary force, e.g. "A is a black belt in karate" clearly has the force of a warning when said to someone expressing an intention to fight A (1964: 444). Davidson, generalising, says that "there is no known, agreed upon... convention for making assertions [or] giving orders, asking questions, or making promises. These are all things we do, often successfully, and our success depends in part on our having made public our intention to do them. But it was not thanks to a convention that we succeeded" (1984: 7).

[38] Green criticises Langton's ascription of authority to pornographers on a number of grounds, arguing (i) that pornography presents conflicting 'prescriptions', which cannot be followed as one would follow

but there is at least a *prima facie* case for pornographers having an ascendant influence in the formation of cultural understandings of what is natural and desirable in sex. It is hard to see how any case along similar lines can be made about hate speech. If the claim about pornography's authority has some plausibility, it comes from (i) the fact that many people willingly consume pornographic 'speech', and (ii) the fact that although pornography's 'claims' about sex and sexuality are contested, they may still, arguably, reach and influence more minds than the competing claims do.[39] These considerations, rather than the speech act theoretic framework, are doing the heavy lifting in Langton's claims about pornography's subordinating force, and – while there *is* a parallel in pornography and hate speech both being disreputable, low-status expression – the considerations that lend credibility to Langton's claims about the authority of pornographers are not paralleled in relation to hate speech.

---

authoritative prescriptions, (ii) that authority in matters of sexuality cannot be equated with authority with respect to women's status *per se*, and (iii) that pornographers compete with other putative sources of authority (e.g. the state, the family, the church) in the norms they prescribe (1998: 295-97). In reply, Langton maintains that even if pornography lacks the efficacy it would need to be authoritative *per se*, it may still be authoritative locally; "the possibility of escape does not alter the fact that such a woman, where she is, in that context, is subordinated" (1998: 273). Green's reply is that it is implausible, given the competition around the norms prescribed in pornography, for pornographers to be deemed responsible even for these local states of affairs. Pornography he says (like hate speech), is tolerated, low-status speech, which competes with other views, including that of a formally equitable legal system which prohibits identity-based discrimination (1998: 305-06). The law's failure to silence endorsements of identity-prejudice does not authorise identity-prejudice where identity-prejudice is already in effect locally.

[39] This suggests that Langton is not really imputing special power to pornographic *speech*, but rather adverting to the *bona fide* social power that purveyors of pornography have, and which we (and often they) generally fail to recognise. "What creates the power of words and slogans" Pierre Bourdieu says, "is the belief in the legitimacy of words and of those who utter them", and "words alone cannot create this belief" (1991: 170). My claim is that Langton's analysis will ultimately have to affirm that sort of view, despite initial appearances to the contrary. In her more recent attempts to spell-out the character of the pornographer's power, Langton (2009c) invokes a concept she calls 'maker's knowledge'; pornographers make it true that women *are* sexually submissive (for instance) by influentially representing women *as* sexually submissive. Saul (2011) argues that this analysis imputes an unrealistic degree of self-understanding to pornographers, most of whom deny any responsibility for shaping sexual norms. What pornographers really possess, according to Saul, is better called 'perpetuator's ignorance': they help to maintain sexual hierarchies, while seeing themselves as reflecting, rather than shaping, the desires and attitudes which those hierarchies rely upon.

## 5.6 Social hierarchy and communicative impairment

Another way to give an account of hate speech's contribution to social hierarchy is through claims about the way that hate speech, as a communicative practice, inhibit its targets communicative capacities in turn. Speaker authority is not as much of a problem in this sort of account, because the channels via which one person's speech can hinder or constrain another person's speech are in a certain sense 'built into' the interactive nature of communicative conduct. We can easily see how A's speech may negatively affect B's speech: the question is whether and how this might be characterised as a harm.[40] One suggestion that crops up in work on hate speech and pornography is that speech which impairs speech should be restricted for the *sake* of free speech, because such speech abuses, rather than embodies, the institution of free speech (West 2012). On this line of thought, the state's choice is not between restraint and non-restraint, it is "between two forms of restraint: one carried out by private groups operating outside the law, and another... carried out by legal authorities under the constraints of a formal statute" (Arkes 1974: 284-85; see also Buchanan 1979: 558-59). In feminist work, this is often motivated by concern for the harm women suffer when their speech (e.g. in the refusal of a sexual advance) is rendered ineffective, the argument being that if expression (like pornography) enables this, it should be liable to restriction in order to protect women's speech. As a general approach to the issue of speech-impairing-speech, however, this introduces more problems than it solves. Firstly, it drastically reconfigures the principle it invokes. It is true that "the guarantee of *equal* liberty of expression" is central to liberal understandings of free speech (Karst 1975: 68), but the guarantee is not of a level communicative playing field, secured through coercive intervention. Free speech is a constraint on the methods the state may employ in pursuit of its legitimate purposes. Its role isn't to inspire initiatives that will promote an egalitarian communicative environment,[41] but

[40] What we then say about authority can go one of two ways. We can say that impairing communicative capacities requires *no* authority, or (like Wieland 2007: 439) that it requires a kind of authority that is 'very easy to come by'.

[41] This understanding is suggested in Kathleen Mahoney's statement that minority groups "look at freedom of expression in terms of the liberal ideal and ask, does this make us free? Do we have the same freedom of speech that the dominant elite have?" (Borovoy et al. 1989: 345); and in David Gillborn's claim that "genuinely *free* speech is an impossibility in a context where "common sense" (what is rational and irrational)

rather to invalidate laws which aim to restrict speech, or which excessively impair speech in pursuit of some other aim. Besides, secondly, there is no need to characterise communicative impairments as infringements of free speech. If A severs B's tongue, we do not condemn A for infringing B's free speech; we condemn A grievously *harming* B. Part of what makes the act harmful is that it impairs communicative interests that are of essential importance to B's autonomy and well-being. But in our legal response to the act, the fact of the harm is all we need to make sense of B's punishment. Similarly, if hate speech harms its targets by impairing their communicative capacities, we can simply advert to the harm to justify the intervention. Framing the issue as a matter of free speech is at best redundant.[42]

So how, then, *should* the harms in question be characterised? One might make the claim about illocutionary disablement without bringing in any of the free speech baggage. The claim would be that hate speech contributes to a communicative environment in which its targets are harmed through being rendered unable to perform certain kinds of illocutionary acts – e.g. things like testifying, reporting, protesting – which they have a rightful interest in performing. The illocutionarily disabled can speak and write, but their speech doesn't do what it is intended to do. As Langton says, the person in this situation is "free to make sounds, say words, but he is not free... to do the illocutionary things he might wish to do" (2009a: 72).[43] Claims about illocutionary disablement may be adjoined to a 'capability theory' approach to understanding what kind of illocutionary abilities we have rightful interest in. Many of the capabilities Martha Nussbaum identifies as requisite for an individual's autonomy have

---

is determined by, and for, White people" (2009: 553). Owen Fiss is the most prominent advocate of this revisionist approach to free speech. The state, he says, is sometimes entitled to "stop the general advocacy of an idea when that idea has the effect of interfering with the speech rights of others"; it thereby acts "not as a censor, but rather as a parliamentarian, requiring some to shut up so others can speak" (1991: 2059; see also 1986: 1412, 1995: 287).

[42] There are a number of other arguments against the idea of free speech as something that encompasses freedom of illocution. Alexander Bird (2002) argues that this alienates the content of a free speech principle from the arguments characteristically used in its defence. Daniel Jacobson (2001) argues that freedom of illocution *per se* would mean freedom for a great deal of misconduct that we rightly prohibit.

[43] A variant on this sort of account is Maitra's (2009) analysis of *communicative disablement*, which conceives of communication in Gricean terms as the conveying of an intention to elicit a specific response in the recipient.

communicative dimensions.[44] Katharine Gelber (2010: 319) spells out the implications of this in relation to hate speech; there are grounds for its restriction, she says, where it impairs "the development of individual capabilities in ways that prevent individuals from becoming capable of constructing and implementing their own conception of the good life". These harms would be indirect, in that they would arise out of a diversely-constituted communicative environment to which individual acts of hate speech contribute. It is not a matter of hate-speakers defaming specific individuals, and diminishing their communicative capacities in that manner. Rather, the harm comes from hate speech's contribution to creating and sustaining a communicative environment in which certain people have inferior capacities, because of their membership in or association with subordinate identity-based social groups.

Alternatively, one might claim not that hate speech interferes with the illocutionary potential of its targets' speech when they do speak, but rather that it just makes them less likely to speak in the first place (Tsesis 2009: 499). This may be, in certain quarters, a result of hate speech 'crowding out' the speech of its targets; although as Frank Michelman (1992) observes, in most cases it is inapt to imagine speakers in a zero-sum contest for finite resources. More plausibly, the claim would be that hate speech undermines the conditions enabling participation in deliberative social activities for people in targeted groups (Fiss 1995: 287). As Mortimer Sellers puts it, the idea is that hate speech "disrupts efforts to cooperatively address common needs through a functional public discourse" (2004: 22).[45]

---

[44] For example, an individual's capacity for things like (i) practical reason, (ii) social affiliation, (iii) imagination and thought, and (iv) political engagement, may all be impaired if her ability to speak, and the effectiveness of the speech she engages in, have been undermined by the speech of others impugning her credibility or character (see Nussbaum 2000b: 231-33; 2003: 41-42).

[45] Feminist critics say similar things about pornography creating "a social climate that makes women reluctant to speak at all", such that "they literally do not utter or pen a string of words" (West 2003: 298). A similar point can be pressed from the opposite direction, as in Kim Watterson's claim that "once the dynamics of hate speech are revealed", campus speech codes can be seen "as a mechanism for opening the university forum to all voices" (1991: 974), or David Brink's claim that "hate speech regulations aim to establish the kind of culture in which vigorous discussion... can be both possible and profitable" (2001: 154). Theorists who hold deliberative views of justice and democratic legitimacy may see communicative impairments as violations of a criterial condition for a just social order (e.g. see Habermas 1996; Bohman 1997; Cohen 1997, 2006). However, even those who do not see deliberative inclusiveness as a condition for justice may regard it

A third way hate speech might impair communicative capacities is through what Miranda Fricker calls *testimonial injustice*. Fricker's idea is not that people are disabled outright in their testimonial capacities. Whatever obstacles a speaker faces, she can still, if she is not literally silenced, perform verbal acts which satisfy the felicity conditions of assertion or attestation. Testimonial injustice is what occurs, rather, when the speaker's assertions are dismissed or discounted because she "receives a credibility deficit owing to identity prejudice in the hearer" (Fricker 2007: 28).[46] Something like this notion of testimonial injustice seems to be in play in things critical race theorists say about hate speech, e.g. Lawrence's claim that hate speech "distorts the marketplace of ideas" (1993: 78), or Delgado and Stefancic's claim that hate speech "warps the dialogic community" (2009: 368). Testimonial injustices have both epistemic and moral aspects. If A prejudicially discounts B's credibility, failing to recognise B's assertion that p as good *pro tanto* evidence for p, then A does poorly *qua* epistemic agent. But B's credibility deficit in the eyes of A is also, Fricker explains, a form of moral infringement against B. For one thing, certain harms may ensue for B as a consequence of her credibility being discounted (e.g. the asylum-seeker may find that her testimony to immigration officials explaining her circumstances is discounted, with ruinous consequences, for the primary reason that the story is told by *her*, an asylum-seeker, see K. Jones 2002). There is also, however, a kind of injustice intrinsic to identity-prejudicial credibility deficit. Being a fully-recognised person, Fricker says, partly involves being seen by others as a knowing subject.[47] If hate

---

as instrumentally important in securing justice (see for example Post 1990a; Estlund 1997; Nagel 1999; Rawls 2001). There are some, like Robert Post, who see hate speech as having deleterious effects on public discourse – indeed, effects which render public discourse unfit as a tool for self-determination for people in marginal groups (1990b: 312) – but who reject anti-hate speech law, on the grounds that infringements of "formal conditions of freedom in order to recuperate democratic values... frustrate the very *raison d'être* of the democratic enterprise" (1993: 677; see also Post 1995: 309 ff.).

[46] Testimonial injustice is one form of what Fricker calls *epistemic* injustice. The other form of epistemic injustice on her account is *hermeneutical* injustice, i.e. the wrong someone experiences when her conceptual resources for interpreting the world are alienated from or prejudiced against her perspective, such that there are important aspects of her experience which she lacks the interpretative tools to grasp or describe.

[47] Fricker invokes a Kantian conception of personhood to explain the moral significance of being counted as a knower: "In testimonial injustice, one person undermines another's status as a subject of knowledge; in Kant's conception of immorality, one person undermines another's status as a rational agent... they are both

speech creates a communicative environment in which people in vulnerable groups have their statuses as knowers undermined, then it contributes to testimonials injustices even without further bad consequences ensuing.

For any of these forms of communicative impairment, there is room for disagreement about whether the impairments in question can be characterised as harms, in the sense of being setbacks to the victim's autonomy-essential interests, and about the extent to which the impairments actually obtain in contemporary, formally egalitarian, liberal societies.[48] However, granting that these *are* genuine and non-hypothetical harms, we are faced with the challenge of establishing that it is verbal conduct generally, and hate speech in particular, which is shaping the communicative environment through which these sorts of communicative impairments arise. Non-verbal factors can surely be formative forces in a communicative environment, as much as verbal factors. Economic inequalities, for instance, might profoundly influence people's participation in communicative practices, while affecting people's illocutionary effectiveness or testimonial credibility. The shift from considering social inequality generally, to considering inequality in communication specifically, may obviate the need for an account of hate-speakers' authority, but we still need an explanation of how hate speech gets a grip on the communicative environment.

## 5.7 Score-keeping in an identity-oppressive language game

Might it be the case, then, that hate speech subordinates its targets *within* communicative interactions? David Lewis's account of conversational kinematics provides a useful framework for investigating this possibility. Lewis devises a concept he calls 'conversational score', which is, roughly, the informal register of how a conversation C stands at time t, and of which

---

instances of the undermining of a dimension of a person's rationality, where that rational capacity is conceived of as essential to human value" (2007: 136).

[48] To take one example, Daniel Jacobson (1995) argues that women are not illocutionarily disabled in sexual refusal in modern liberal societies, because women are almost universally understood – in the legal system and in wider society – as being fully able to verbally refuse sex. When a particular woman's verbal sexual refusal fails, he maintains, this is not a symptom of a general state of women's illocutionary disablement. Rather, it is just an instance of egregious wrongdoing on the part of the person ignoring the refusal.

contributions would be appropriate in C at t accordingly. Conversational score is determined by what has transpired in C up to, and especially immediately prior to, t. However, Lewis observes, conversational scores – in contrast with scores in other score-governed activities – have a distinctive plasticity; they automatically update so that *any* conversational contributions the participants accept are accommodated, and retrospectively *made* felicitous, in the conversational score, whether or not they were felicitous in the first place (1979: 346-47). As Mitchell Green puts it, "speakers uttering indicative sentences have those sentences entered into the conversational record unless there is a demurral from an addressee", and thereafter "the content of those sentences may be used as fodder for future inference as well as be presupposed by their speech acts" (2000: 466). So if, in a conversation about religion, A makes an obscure remark about cheese, the score will immediately change to accommodate A's remark, making remarks about cheese felicitous conversational contributions thereafter – that is, unless A's interlocutor blocks this, e.g. by saying "what's this about cheese? We're talking religion"). In a similar way, the presuppositions in a statement like "she's very able for a woman" will, if unchallenged, become part of the common ground for a conversation, so that other statements consonant with those presuppositions become apt conversational contributions.[49] Anyone engaged in a conversation has the ability, given this plasticity, to make certain things conversationally licit or illicit for her interlocutor, and one might see this as a mechanism through which hate speech has a shaping influence, at least in the communicative dimensions of identity-based social hierarchy, however the larger causal story about social hierarchy goes. Like all verbal conduct, hate speech determines "a space of admissible illocutions that is partially ordered in terms of preferability for conversational purposes" (M. S. Green 1999: 81). Where hate speech is allowed to pervade public discourse, its targets can expect to find their communicative aims – assertorical or otherwise – shut off or disadvantageously prefigured, both by the presuppositions, and by the overtly prejudicial calumnies, that hate-speakers enter into the conversational score.

[49] On the typical view of presupposition (due to Stalnaker 1973), presuppositions are intentional relationships between persons and propositions (such that A's presuppositions are the propositions whose truth she takes for granted in a conversation or inquiry) as opposed to semantic relationships between propositions. In Stalnaker's recent work on presupposition (2002), he explores its other-regarding character, i.e. the fact that in presupposing p, one assumes that one's interlocutors likewise presuppose p.

Some feminist philosophers formulate accounts of women's communicative impairment using this Lewisian framework. For instance, Langton and West use it to elaborate feminist claims about the silencing effects of pornography.[50] Mary Kate McGowan constructs a more general-purpose analysis of conversational score-keeping and oppressive speech (in 2003; 2004; 2009; 2012), focusing on what she calls 'conversational exercitives': speech acts which enact permissibility facts in conversations, by altering, in accordance with a Lewisian conception of conversational score-keeping, which conversational moves are subsequently licit and illicit.[51] Such speech acts are 'exercitive' in that they enact 'permissibility facts', and 'conversational' so far as the conversation is the domain in which they do this enacting (2004: 95). In a wide-scale, identity-prejudicial system of social arrangements, McGowan says, conversational exercitives make certain specific courses of action apt or permissible, and thereby bring the force of identity-prejudice to bear in specific contexts (2009: 392). For example, when A makes a sexist remark in the workplace, he makes it conversationally appropriate in what follows to *speak* degradingly of women, and in so doing, he "makes it acceptable, in this immediate environment and at this time, to degrade women" (2009: 400). The key innovation resides in McGowan's fine-grained articulation of the relations between systems and structures of oppression, and the points at which these things affect their subjects. Part of what it means for B to be a victim of *structural* oppression, is that A cannot, by himself, act so as to cause or prevent B's oppression. However, A can still be a culpable participant in the realisation of the patterns of conduct through which B's oppression *is* effected, and McGowan's account explains how identity-prejudicial speech may constitute such participation. In a conversation such as the following one:

---

[50] "Presuppositions established by pornography as a component of an on-going conversational score", they say, prevent "women from making the moves they intend to make" (Langton and West 1999: 318).

[51] McGowan sees her account as supplementary to feminist critiques of pornography, and Langton has endorsed McGowan's approach to these issues, noting in particular that McGowan's framework is well-equipped to explain how non-authoritative speech nevertheless has far-reaching social consequences (Langton 2012: 137-38).

> John: So, Steve, how'd it go last night?
>
> Steve: I banged the bitch.
>
> John: [smiling] She got a sistuh?

McGowan wants to say that

> Steve's utterance... makes women second-class citizens (locally and for the time being)... it is akin to a sign reading: 'It is hereby permissible, in this local environment and at this time, to treat women as second class citizens.' (McGowan 2009: 399-400)

On McGowan's account, the structural oppression of women is a rule-governed activity, in the sense that it is a complex, coordinated system of social interaction, which ranks people and imposes norms for how people are to be treated in view of their ranking, with associated expectations and informal penalties (2009: 397). And like other rule-governed activities, McGowan says, women's oppression is governed by two types of rules (i) general rules, or *g-rules*, which "govern all instances of the rule-governed activity in question and... are not enacted by the performing of any particular such activity", and (ii) specific rules, or *s-rules*, which are "enacted by the performing of the very rule-governed activity over which they preside", and which are limited in scope and duration (2009: 396). G-rules set parameters for how the activity proceeds in general, and s-rules determine, within those parameters, what happens (at any given) 'here' and 'now'. The g-rules of identity-oppression (of women or any group) prioritise and subordinate the interests and prerogatives of one group relative to one another. McGowan sees Steve's comment as enacting s-rules, in the local context, to the effect that women may be treated as inferiors 'here and now'. The comment can only enact s-rules because it is consonant with the g-rules of the gender system that Steve and his co-workers all live under, but what it does is bring the latent force of that system to bear locally. The oppressive force comes from something Steve is not responsible for – women's oppression *per se* – but Steve *is* responsible for activating the oppressive force in the local context.[52]

[52] Katharine Gelber suggests a Habermasian analysis of hate speech's harm, which bears some resemblance to McGowan's work on conversational exercitives. Hate speech impairs its targets' communicative capacities, Gelber says, when it "raises 'truth' claims of an objective world characterised by inequality, and where the hate speaker is in a position of power relative to the hearer" (2002: 117-18). The Habermasian ideas here

One difficulty with this account is the 'the reversibility problem': if permissibility facts can be enacted very easily, by anyone participating in a conversation, why can't the individual who stands to be oppressed by a conversational exercitive counter with a conversational exercitive of her own? Communicative impairments brought about in identity-prejudicial speech, given this account, can seemingly be remedied via the free speech theorist's favourite panacea: speaking back/counter-speech (espoused, for instance, in Richards 1988; 1999). McGowan's reply is that there is an asymmetry in the plasticity of conversations involving identity-prejudicial speech. Asymmetries in conversational plasticity were recognised by Lewis in his presentation of the score-keeping framework. When A says 'I know we're out of milk', and B replies 'but what if you're being deceived by a Cartesian demon?', B alters the conversational score to increase standards of stringency for the application of 'know'. In cases like this, Lewis says, "the boundary readily shifts outward if what is said requires it, but does not so readily shift inward if what is said requires that" (1979: 355).[53] McGowan discerns a comparable asymmetry in the way that identity-prejudicial utterances impact upon the communicative environment. "Setting the record straight in response to a sexist remark" she says, is "akin to trying to unring a bell... there is something complicated and covert going on that is difficult to pinpoint and hence undo" (McGowan 2009: 403).

Some *pro tanto* support for this claim may be found in psychological studies examining people's responses to identity-based epithets,[54] and the phenomena in question do not even

---

relate to the notion of "validity claims": claims proffered for acceptance in communicative conduct, whose recognition by speaker and listener is a constitutive condition of the ideal speech situation (Habermas 1979: 1-3). Validity claims consist in (i) a truth claim about how the world is, (ii) an implicit normative claim as to the propriety of saying things like (i); and (iii) an implicit claim as to the speaker's sincerity and seriousness (Habermas 1984: 38-42). Within this framework of analysis, identity-prejudicial speech may be seen as (i) assuming others' assent for identity-prejudicial claims which impugn the speaker's targets, (ii) assuming others' assent to the notion that it is acceptable to make such claims, and (iii) indicating to all one's sincere investment in the views expressed in (i) and the norms implied in (ii).

[53] Lewis (1996) also discusses such cases in his contextualist analysis of the conditions of knowledge.

[54] Although an asymmetry in the activation and deactivation of negative schemata has not been demonstrated in any study that I know of, there is evidence to suggest that negative schemata in accordance with typical identity-prejudicial attitudes can be activated (that is to say, made to influence people's judgements) with extraordinarily *little* effort. Single uses of derogatory slurs or allusions to negative stereotypes can lead

seem all that unusual if we understand them as consequences of associational thinking. Consider an example: a well-known blogger reports that a pop singer, LG, is transsexual. The report is quickly discredited, so that we all return to our earlier evidentiary state of having no good reason to suppose that LG is transsexual. And yet the association remains. It remains, in part, because it is subtly perpetuated by its discrediting. The statements "LG is really a woman", or "reports that LG is transsexual have been discredited" reinforce the notion that, on some level, LG and transsexuality have some connection. 'No smoke without fire', so we may think, even when the 'smoke' is something with no evidentiary force by our own lights. The association remains, even while it is seen as having spurious or sinister origins. In relation to McGowan's asymmetry claim, then, the idea is that there are schemas of identity-prejudicial associations which can easily be brought to bear on our social interactions through speech, but which cannot easily be removed again.[55] On a standard prejudicial schema for sexual identity, female-ness is associated with weakness, sexual passivity, domesticity, etc. A sexist remark brings forth these associations, and exercitively creates presuppositions and expectations that accord with them, and it does so easily, (i) because the associations are a familiar distillation of the structurally-prejudicial social arrangements into which we're all already acculturated, and (ii) because conversational scores *in general* are easily altered, and thus readily reshaped by the introduction of *these* associations and the exercitive enactments that come with them. Subsequently, however, the associations are immune to attempts to refute or displace them, because conversational moves which try to discredit them tend to accidentally reinforce them. Given the nature of associative thinking, A's rebuttal to the idea of women's natural domesticity still lends a small measure of credibility to that idea. McGowan's asymmetry claim can be seen, then, as a result of a more general asymmetry in creating and eliminating associations. It is easier to shift conversational scores along sexist trajectories because sexist

---

subjects to downgrade their assessments of, or expectations towards, members of targeted groups, and this sort of effect has been demonstrated both for third parties (Greenberg and Pyszczynski 1985), and for members of the targeted groups themselves (Steele and Aronson 1995). It seems improbable, although it remains to be demonstrated empirically, that identity-prejudicial schemata can be 'de-activated' (so to speak) as effortlessly as they are, so the evidence seems to indicate, activated.

[55] A schema, here, is some sort of psychological or doxastic construct which "contains in a schematic or abbreviated form someone's concept about... a group of people or events", including the group's "main characteristics, from the perceiver's point of view" (Valian 1998: 103).

presuppositions and expectations are governed by invidious associations, and associations in general are relatively easy to initiate through verbal activity, and relatively difficult to eliminate via verbal activity, if indeed they can be eliminated in that manner at all.[56]

This is in keeping with the picture McGowan paints. She says that responding to a sexist remark is like trying to 'unring a bell'. I think we can put a finer point on the insight contained in this simile by reflecting on the character of associative thinking. Along the way, we find that there are similarities after all between Lewis's examples of asymmetric conversational pliability, involving stringency, and the cases that interest McGowan. In both areas, the asymmetric pliability of conversation seems to be due to the asymmetry between what it takes to make something salient in speech, and what it takes to make the same thing subsequently unsalient. In conversation generally, ideas can be made salient in an instant, through a participant's speech, but cannot be made unsalient in any similarly straightforward manner. The salience of certain considerations will – in accordance with Lewisian rules of conversational kinematics – alter how it is permissible, apt, or felicitous for participants to behave in a given conversation. In Lewis's pet cases, the result is (for instance) that non-stringent usages of certain words will be rendered 'out of play'. In McGowan's cases, the result is the local mobilisation of systemic identity-prejudices.

## 5.8 Empirical data, linguistic phenomena, and indirect harm

McGowan's central example involves sexist speech in the workplace, but her analysis can extend to explain the role speech plays more generally in various forms of identity-based hierarchy. With respect to our concerns, the argument would be that all hate speech,

---

[56] These claims about associative schemas suggest an intersection between McGowan's work and the emerging philosophical literature on implicit bias (e.g. Blum 2004; Kelly and Roedder 2008), so far as associative schemas may be an important part of the processes through which implicitly-biased judgements are rendered. I have said little here about the content of identity-oppressive schemas, and how they vary from one social identity group to another. Such substantive questions are the focus of much feminist theory, critical race theory, queer theory, and other critical discourses. If a view like McGowan's is elaborated in terms of associative schemas, it may be used to unify an account of the semantics and pragmatics of identity-prejudicial speech.

irrespective of its direct harmfulness, is liable to prohibition under an indirect harm rationale, because all hate speech functions as a mediating conduit through which identity-based social hierarchies can be brought to bear in specific contexts.

One question, even if we grant McGowan's analysis of the linguistic phenomena, is whether we have seen enough to conclude that these phenomena *produce* identity-based oppression, or whether they are generic linguistic phenomena, which are active in identity-oppression just because they are ubiquitous in verbal intercourse. Consider an example (adapted from Richard 2008): the Philosophy Department at Mantown University hires two women in 2013, and the Vice-Chancellor of Mantown later says to a colleague: "this year Philosophy hired *three broads*". Despite the numerical error, Richard says the remark should not be judged false, because an assertion of its falsity assents to its contemptuous representation of women. We can say the remark is erroneous, but its erroneousness is not the truth-apt erroneousness of falsehood (2008: 7). Now, the linguistic phenomenon exemplified in the remark – a representation of Xs as scorn-worthy via the exhibition of scorn *for* Xs – is not unique to remarks about women. The Vice-Chancellor of Rotund University may (erroneously) tell her colleagues that "the History Department hired three bean-poles". In this case it is doubtful that identity-oppression is occurring, because tall and slim people are (generally) not oppressed for their body shape. But then, so far as the linguistic phenomenon that facilitated the oppressive speech in the first case occurs in the second case as well, it seems doubtful that the linguistic phenomenon in any way *explains* the identity-oppressive operations of the first Vice-Chancellor's remarks, any more than his grammatical competence explains that putative fact. If we want to explain the identity-oppression in a case like this, looking to the linguistic phenomena will not get us very far. A parallel point applies to McGowan's account. Conversational exercitives and the asymmetric pliability of conversations may be collocated with identity-oppression, and may sometimes facilitate it. We will do well, for these reasons, to understand how the phenomena function. However, absent any reason to think that these phenomena are confined to the use of identity-prejudicial language against members of subordinated social groups, these observations do not by themselves vindicate a language-oriented understanding of how identity-based oppression is created or sustained. Granted, we can say speech is 'involved' in identity-based oppression, but this postpones the key question, which is whether speech's 'involvement' consists merely in collocation with

oppression, or in the facilitation of oppression, or whether its involvement is somehow more causally integral in enacting oppression.

A further problem with using McGowan's approach to defend general prohibitions on hate speech is that it seems like only hate speech carried out in a conversational *mode* will do the oppressive work McGowan describes. Obviously there is much verbal conduct generally (and hate speech specifically) that is not in the conversational mode. There are books, speeches, lectures, editorials, and creative forms of speech (poems, plays, novels), which are addressed to audiences in an ostensibly 'one-way' fashion. This matters, because the conversational pliability that Lewis identifies isn't a *sui generis* quirk, rather, it is a consequence of conversation's cooperative character. Part of what makes a verbal exchange a conversation, as opposed to 'dysfunctional quarrelling' or 'free-associative yammering', is the participants' joint aim to be talking about the same things to similar communicative ends (e.g. solving a problem or exchanging information). This tacit cooperation is what confers the presumptive status of 'common ground' on unchallenged presuppositions. There is no evident reason, however, for hate speech which *is not* carried out in the conversational mode to be seen as presumptively changing our 'common-ground', such that it can exercitively enact permissibility facts for people targeted in hate speech. One may want to claim that all publicly performed speech becomes part of some nebulous entity called '*the* social conversation'. But if this wider 'social conversation' is a conversation at all, it is an ersatz one at best; it lacks the characteristics of 'regular' conversation that facilitate the phenomena McGowan describes.[57] What McGowan's account explains, if it is successful, is how hate speech in specific conversational contexts (e.g. in the co-workers' banter), can bring the force of identity-prejudice to bear in those contexts. It doesn't substantiate the more ambitious suggestion underlying an indirect harm rationale for anti-hate speech law, that *all* hate speech, irrespective of the context of its performance, contributes to social hierarchy.

As I suggested above, in the Vice-Chancellor example, further evidence is needed before we can judge McGowan's account successful even in substantiating its more modest

[57] As Andrew Kenyon (2010) says, a great deal of free speech discourse is framed by a conversational notion of public speech, and this conception has a more general distorting potential, since it implicitly suggests a public in which there is cooperation and mutual comprehension, and thus discourages a sober apprehension of the depth of some of the disagreements dividing our societies.

conclusions. In the rule-governed activity of identity-oppression, McGowan says, the verbal moves and the non-verbal moves are all mutually-reinforcing enactments of the activity: "since speech is one way to... differentially treat people in virtue of a person's membership in a socially marked group" she says, "speech is certainly sometimes a move (in the rule-governed activity) of oppression" (2009: 397). Now, while it's true that speech is one way to differentially treat people, differential forms of treatment are not all created equal. If A invites B but not C to join a basketball team, that is a way of differentially treating C; if A's non-invitation owes to the fact that C is a woman and that the team plays in a men's sporting league, then that is a way of differentially treating C in virtue of her membership in a 'socially-marked group'. It does not necessarily follow, however, that A has enacted sexual oppression. Whether that is the case will depend on whether men's sporting leagues contribute to the causal networks constituting sexual hierarchy, and the view that they do requires an argument. McGowan describes a process through which, in the normal course of conversational kinematics, hate speech can enact identity-prejudicial permissibility facts for people in the local conversational context. It remains to be seen, however, whether such permissibility facts actually alter the way people subsequently behave. The hate-speaker might exercitively make it permissible to degrade Xs at a particular time and place, but this does not show that acts of degradation against Xs, where they occur, are elicited by such permissions. The critical race theory literature is full of anecdotal accounts of hate speech's enactment of oppressive norms, but anecdotes only reveal the facts about specific scenarios; they are insufficient by themselves to substantiate a general thesis about hate speech's social function.[58] There is work in social psychology on the impact of identity-oppressive speech specifically, but what little findings there are do not appear to speak directly to our concerns.[59] Some authors express doubt about

[58] See for instance Delgado 1989; Griffin 1994; Bell 1995.

[59] Some studies examine how perception of hate speech differs across lines of social difference. There are qualitative studies which find members of disadvantaged social groups more likely to feel personally attacked in public verbal confrontations (Nielsen 2002, 2004), and studies in social psychology which find subjects' judgements about the offensiveness of hate speech to vary depending on gender, race, sexual preference, and to be sensitive to several framing effects (Cowan and Hodge 1996; Cowan and Mettrick 2002; Cowan et al. 2002; Cowan and Khatchadourian 2003; Cowan et al. 2005). There are other studies that explore correlations between identity-prejudicial speech and markers of structural inequality (e.g. suicide rates, rates of naturalisation and inter-marriage, segregation), and seek to ascertain the extent to which these correlations are

whether it would even be possible to obtain sufficiently fine-grained data on the relationship between hate speech and social inequality. Eric Heinze, for instance, says that general restrictions on hate speech cannot find support in empirical studies, since such restrictions are based on the view that hate speech is detrimental to its targets "in ways which are often subtle and pernicious, and therefore not amenable to precise empirical observation" (2009a: 278; see also Weinstein 1999: 127-35; Sumner 2004: 158-59; 2009: 210). I am less sure than Heinze about the unattainability of useful data here. For instance, it seems like a manifest form of structural, identity-based inequality when Caucasians are rewarded and non-Caucasians are penalised (in pecuniary ways or otherwise) for being assertive in the workplace. If people's prior exposure to identity-prejudicial representations in speech (e.g. representations of white people being more polite, and non-whites being more combative) affects people's subsequent relative appraisals of whites and non-whites behaving assertively, then evidence of that could surely be obtained through an appropriately-designed study. If appraisals of character along axes of assertiveness and civility affect patterns of reward and penalty in typical work environments, then evidence of *that* can surely be found as well. Such data would establish that verbal conduct is continuous with paradigmatic 'moves' in the rule-governed activity of identity-prejudice – where by 'continuous with', I mean that the verbal practices causally interact with the other practices in a non-epiphenomenal way.

My point above, however, was that even if McGowan's account of what is going on in identity-prejudicial speech (or something like it) were evidentially vindicated, through this sort of data, what we would have would not be a justification for making all hate speech liable to restriction under an indirect harm rationale. Rather, we would have evidence of previously undiagnosed mechanisms through which identifiable hate-speech-acts are causally responsible

---

sensitive to the character of (e.g. the degree of complexity in) the identity-prejudicial language in question (Mullen and Johnson 1995; Mullen and Rice 2003; Mullen 2004; Mullen and Smyth 2004; Leader et al. 2009). In many cases, though, these studies have an unhelpfully etiological view of the causal processes they investigate. For instance, Mullen and Rice (2003) ask whether hate speech, rather than contributing to social exclusion, is an after-the-fact consequence of attempts to rationalise exclusion, and they see their arguments against the latter as supporting the conclusion that hate speech *does* contribute to exclusion. The form of the argument suggests that they haven't considered either (i) an independent common cause that explains a correlation between hate speech and exclusion, or (ii) hate speech being part of complex causal networks which cumulatively constitute identity-based exclusion.

for inflicting harms on specific individuals in specific contexts. That is to say, we would have evidence of processes through which hate speech *directly* harms its targets. The infliction of these direct harms would eventuate from causal interplay between the proximate hate-speech-act and underlying patterns of identity-based inequality; but in this respect, they would not be unlike those discussed in chapter 4, related to harassment, threats, and incitement.

Speaking more generally, when we shift our focus to the question of how hate speech may inflict communicative impairments on its targets, and how this may occur within the communicative interaction, it becomes easier to formulate an account, like McGowan's, on which responsibility for the relevant harm can be plausibly ascribed to the hate-speaker. In such accounts, the hate-speaker's being responsible doesn't rest on hate-speakers being authoritative speakers, nor does it rest on there being baroque causal mechanisms in virtue of which hate speech turns out to be the root of all the evils of identity-based social hierarchy. Rather, when the aim is to show how hate speech as one form of verbal practice might impair other forms of verbal practice (for its targets), the character of the verbal practices themselves can be called upon to provide explanations of how this might transpire. What we will uncover in a successful account, however, is going to be a more fine-grained description of a heretofore unrecognised, verbally-enacted mode of harassment or discrimination: the kind of act for which the hate-speaker's responsibility lies simply in the fact that if he hadn't engaged in hate speech, certain specifiable harms would not have eventuated. We will not establish the bolder conclusion – that *all* hate speech contributes to social hierarchy – which is needed to justify hate speech's general prohibition under an indirect harm rationale.

## 5.9 On setting a high bar

I have argued against a number of attempts to show that hate speech significantly contributes to identity-based hierarchy. I do not think we can say with certainty that hate speech *doesn't* significantly contribute to the identity-prejudicial climate through which harmful consequences redound to the people targeted in hate speech. My claim is that in order to make hate-speakers liable to punishment on this basis, we need to firmly establish – whether by uncovering new empirical data, or by offering a persuasive re-description of things we already recognise – that a causal contribution of the relevant kind actually occurs.

None of the accounts that we have seen succeed to this end. Heyman's account implausibly claims that being targeted in hate speech *ipso facto* harms a person's status. Waldron identifies a plausible form of dignity-based harm that people subordinated in social hierarchies may suffer, but he misidentifies hate speech as its causal source. Those who say hate speech legitimates identity-based hierarchy due to non-contestation exaggerate their key premise – that hate speech is uncontested by the state and/or society – beyond credibility. The prospect of hate speech having a profound conditioning influence on children is *prima facie* plausible, but we do not have any reason to single out hate speech as an important factor in acculturating young minds into identity-prejudice and the social arrangements influenced by it. Hate speech may, theoretically, be used in subordinating illocutions, like those Langton identifies, but that depends on hate-speakers having the social power to secure compliance with the subordinating norms they purport to enact, and this seems not to be true of hate-speakers typically. Hate speech may create a communicative climate in which its targets' communicative capacities are disabled, but once again, we need some reason to think that such a climate isn't simply produced by material or institutional factors, in order to hold hate-speakers liable for the harm it does. McGowan's account of conversational exercitivity shows one way in which hate speech might exert its communication-impairing effects on its targets, but what the account amounts to, ultimately, is in fact a description of a different form of direct harm for which hate speech might be responsible, such that evidence unearthed in support of McGowan's account would, if anything, justify expanding the scope of anti-hate speech law under a direct harm rationale, rather than implementing general prohibitions on hate speech under an indirect harm rationale.

ଔ

Perhaps I have set an unduly arduous task for advocates of restrictions on indirectly harmful hate speech. Isn't it enough to assert the *prima facie* plausible claim that all hate speech makes *some* contribution to the social-environment-based harms at issue? If an act contributes to a harmful outcome as a mere 'drop in the bucket', isn't that contribution sufficient for liability?[60] Waldron objects to the way that, in discussions of hate speech, it is often supposed

[60] For discussion of moral obligations related to 'drop in the bucket' contributions, see Parfit 1986, chapter 3.

that we're entitled "to raise the bar very high indeed in our assessment of the harms that racist speech might do", while ignoring evidence of harm "unless it is established beyond a scintilla of doubt" (2012: 148). Am I not doing this here: setting the bar for what it takes to establish claims about hate speech's indirect harmfulness unreasonably high? I am not.

Firstly, note that the reformation of identity-based social hierarchies is recognised as an urgent moral imperative on all sides of the hate speech literature. If hate speech *doesn't* contribute, more than trivially, to the perpetuation of social hierarchies, its prohibition might not only be ineffective in reforming social hierarchies, it might positively impede that aim: (i) by diverting attention away from other legal and extra-legal measures that may be used to combat social hierarchy; or (ii) by sparking attention, and perhaps even sympathy in some quarters, for the perspective of hate-speakers who have been punished for their speech. The risks are conjectural, but not far-fetched. At the very least, if these are genuine sources of risk in the pursuit of an urgent moral imperative, then it behoves us to be very confident of the veracity of any complicated causal claims that are invoked to justify general prohibitions on hate speech – more confident, I have argued, than we can be at present.[61]

Secondly, we need to be mindful of the unconscious 'classism' that may lead us to favour general prohibitions on hate speech under an indirect harm rationale, even in the absence of a clear understanding of how hate speech in fact contributes to social hierarchy. When we are asking how all the economic, institutional, and discursive practices in our society work together to perpetuate social hierarchy, the frank answer must be that it is very difficult to ascertain where and how the driving causal forces operate. Hate speech is the vulgar and conspicuous facade of identity-prejudice. It is an expressive practice that in our societies seems to belong mainly to the poor, the uneducated, and to the fanatically unhinged. Drastic changes to the distribution of resources in the financial or education systems may have an extremely profound impact on social hierarchy. Ending the global war on drugs might do so

[61] Stefancic and Delgado (1993: 744) are critical of those who claim, without evidence, that anti-hate speech law "turns hatemongers into martyrs". I admit that the risks are conjectural, and that adverting to conjectural negative consequences is sometimes used as a cynical way of defending an unjust status quo. But their principal concern, I take it, is that we should not allow merely conjectural risks to override the case for prohibiting hate speech where a harm-prevention rationale for its restriction is already established. My point is that because the harm-prevention rationale for prohibiting hate speech is *not* firmly established, we should be wary about opening up (even merely conjectural) risks in the project of reforming social hierarchy.

as well. Reforming cultural apartheids in representations of ethnicity, sexuality, and religion in other verbal practices – for instance, in news and entertainment, or in academia – might also have a far-reaching effect on patterns of marginalisation and exclusion. All of these things would be extraordinarily difficult reforms. Restricting hate speech is much easier. But part of what makes it easy is the relatively lowly social standing of most who engage in hate speech, compared to most of us whose relationship to social hierarchy is one of quietly enjoying its benefits, rather than crudely and inarticulately rallying behind it. If we want to ensure that hidden currents of classism are not at play, again, we had better be confident that legal prohibitions on hate speech are not just a way of finding convenient scapegoats for harmful structures in which we are all complicit, even if only as idle beneficiaries.[62]

The final thing to say here is that attributing liabilities for 'drop in the bucket' contributions still ultimately requires us to judge, with some confidence, that the relevant

[62] Larry Alexander says "the movement to ban hate speech... is motivated primarily by hatred of those with bigoted attitudes and a desire to exercise power over them" (1996: 98). That seems unwarranted, whether for reasons of over-generalisation or gross exaggeration. The putatively 'classist' attitudes that may encourage support for anti-hate speech law, as I have sketched them here, principally descend from an eagerness to attribute blame and liability somewhere, rather than from a structured antipathy for the people toward whom (it turns out) such blame can be most easily attributed. To whatever extent a form of 'scapegoating' is in play here, there are several ways to characterise its wrongfulness. Talk of scapegoating might be used to gesture toward a general concern about how the law stigmatises people as 'outlaws', the worry being that this status should not be conferred on anyone unless we are very confident of his criminality, not only in the sense (i) that he has performed the relevant maleficent act, φ, but also in the sense (ii) that φ does in fact stand in the right kind of relation to the bad outcome, x, in light of which φ is understood as a criminal act (rather than a *merely* maleficent act) in the first place. Alternatively, the wrong here may be understood as pertaining to the punishment of expressive conduct in particular. Unlike other (putatively) indirectly harmful forms of conduct (e.g. river pollution), speech bears a special dignitary relationship to the speaker; it is an act that is somehow emblematic of the actor's agency, so far as it embodies (part of) his view of the world. Consequently, so one might argue, when we are trying to redress bad outcomes with inscrutably complex causal origins, it is particularly pernicious to pin the blame on those who have *spoken* in a way that seems potentially complicit in the outcomes, while withholding liability from others who have performed (non-verbal) behaviours that seem equally likely to be complicit in the outcomes.

liability-incurring acts are in fact 'drops' in the relevant bucket.[63] We can grant that person A should be liable for contributing a 'drop in the bucket' to pollution when she deposits her used batteries into the river. But should B likewise be liable if she drops a banana skin in the river? Given the accretive nature of pollution-based harm, neither act will be counterfactually responsible for inflicting any discrete harm on any identifiable individual. However, we cannot judge that B's act even *indirectly* contributes to pollution-based harm, until we have identified a process through which banana skins in rivers, like batteries in rivers, make *some* incremental negative difference to the salutary functioning of the system whose degradation would be harmful. We can allow that hate speech used to harass, threaten, and incite violence functions as a drop in the bucket, but that is something which we can redress through restrictions on *directly* harmful hate speech. The task is to explain why this might also be true of hate speech which is *not* used to harass, threaten, or incite. And there is a dilemma that arises when we try to provide an explanation to this end. Unlike in the river pollution analogy, hate speech's negative effects are not entirely dilute and accretive. Whereas negative incremental polluting effects are typically brought about in such a way that no potential victim is (indirectly) more affected than any other, the negative incremental effects of hate speech will always be achieved through hate speech having a negative impact on a particular person or people's attitudes, in some sense or other. Where those negative impacts are mild and sub-harmful, the hate-speech-acts are, for the reasons discussed in chapter 4, rightly regarded as a part of normal social intercourse. On the other hand, where those impacts can be shown to function as the originating events in specifiable harms inflicted on specifiable individuals – as would be the case, so I have suggested, in a fully evidentially-vindicated version of McGowan's account – there the hate-speech-acts in question should just be conceived of as directly harmful. Unlike with the banana skin in the river, satisfactorily identifying the mechanism through which hate speech makes an incremental contribution to social hierarchy will involve identifying a process whereby specific acts of expression figure (causally non-redundantly) in

[63] Christopher Kutz suggests accountability in participatory liability might be based on "an expanded notion of individual participation, to include participation in a culture or a way of life" (2000: 167). Even if that is right, however, my point here is that further work is required to establish that hate speech which *is not* directly harmful *can* be legitimately construed as part of the kind of harmful 'way of life', in which one's participation might make one (indirectly) liable for harm.

specific negative outcomes. Indirect, contributory harm may be a defensible way to ground liability-ascriptions in principle, but given how hate speech gets its causal purchase on the world, it seems unlikely that any tenable account of indirectly harmful hate speech is in the offing.

# ☙ 6 ❧

# Conclusion

## 6.1 Summary: harm and responsibility

In the second half of the 20th century free speech commonly functioned as a totemic political value in liberal societies. Whether in the civil rights movement, or in the liberalisation of the entertainment media, or in the gradual ascent of progressive cultural and sexual mores, 'free speech' was a banner behind which advocates of a reformist social agenda could unite against repression and conservatism. It seems unlikely at present that free speech will continue to occupy such a position in 21st century liberal politics. In the US, libertarian conservatives have helped transform 'free speech' into an ally in the electoral system's ongoing corporatisation. Granted, capital has an influence on electoral politics in all liberal societies. What is distinctive about the US situation, as it currently stands, is that an exalted liberal value has been invoked – successfully – to rationalise and enable this influence. Free speech has not been twisted and repurposed in the same manner, certainly not the same extent, at least, in other contemporary liberal states. As I said in §1.1, though, there are anxieties about the future and the integrity of free speech on both sides of this contemporary divide in liberal politics. Outside the US, one may worry that well-meaning egalitarian strains of liberalism have transformed 'free speech' into a hollow aspiration, one that sits idly alongside the repressive policing of public opinion. If our hope is that free speech will remain a salutary value in liberal politics, the axiological conflicts on the both sides of this rift will have to be examined and resolved.

The legal restriction of hate speech is now commonplace in liberal legal systems outside the US. To the question of whether that is generally justifiable, the answer I have given suggests a two-tiered approach: the legal restriction of hate speech is justifiable, so long as it is confined to legislation which aims narrowly to penalise (and thus, to deter) hate-speech-acts which play a non-redundant role in inflicting identifiable harms on specifiable

individuals. When hate speech is used to verbally assault its targets, or to incite third parties to act violently against the people targeted in hate speech, then the hate-speaker is, in principle, liable to punishment in view of the harm that eventuates, if harm does indeed eventuate. Moreover, I have argued that the power of hate speech to inflict harm in these kinds of direct harm cases is enhanced and extended by the way that the identity-prejudicial content of hate speech interacts with identity-prejudicial features of the social environment in which hate speech is delivered. Consequently, so I have argued, there are *pro tanto* good reasons for treating hate-speech-based verbal assaults and hate-speech-based incitements as distinct offences, liable to restriction under purpose-built statutory instruments, rather than treating them merely as specific instances of more general varieties of verbal wrongdoing, such that they are liable to legal penalty (if at all) under more broadly-framed restrictions on verbal wrongdoing (such as, for instance, statutes criminalising certain forms of threats). Restrictions of this kind will be incompatible with a commitment to free speech, only if (i) hate speech is the kind of speech that is entitled to a 'protected species' status under properly-elaborated free speech principles, or (ii) the aim of the restrictions is to penalise the expression of identity-prejudicial opinion, rather than to prevent harm. I have argued that (i) is not the case, and that in appropriately-formulated restrictions on hate speech, (ii) will not hold either.

On the other hand, where restrictions on directly harmful hate speech are justifiable in principle, I have argued that hate speech which is not used to threaten or harass its targets, or to incite violence against its targets, should not be subject to legal punishment. A number of authors have recognised that the most promising case in favour restricting such hate speech would be to show that it is responsible for harming its targets after all, and have formulated accounts which aim to establish just that: that *all* hate speech *is* indirectly harmful. Even when hate speech does not inflict the kind of direct harms associated with threats, harassment, and incitement, these authors claim, it contributes to an identity-prejudicial social environment, under whose operations the targets of hate speech – distinguished by ethnicity, religion, or sexuality – experience a range of status-harms. I have explained why such accounts rest on a liability-ascription framework that is tenable in principle. But at the same time I have argued, against these accounts, that the complexity of the processes through which status-harms of the relevant kind are inflicted is such that we cannot confidently judge that hate speech does in fact contribute, more than trivially, to the environmental conditions and concomitant harms

to which these indirect-harm analyses draw our attention. Granted, it may still be the case, for all that I have argued in chapter 5, that all hate speech is indirectly harmful. However, all we can currently assert is that possibility, and as a mere possibility this is an insufficient basis for imposing punitive legal liability upon those who engage in hate speech.

Among outspoken proponents of free speech, one common line of response to the question of anti-hate speech law's justifiability is to discount or downplay the negative effects that are attributed to hate speech. David Goldberger, for instance, says he doubts the existence of any "generalized, automatic harm" wherever hate speech occurs, "there is a pain in those circumstances" he grants, but he sees it as a relatively mild discomfort; "an effect which we should not criminalize" (Borovoy et al. 1989: 366).[1] Larry Alexander says that "while hate speech can be harmful, its actual harmfulness has been greatly exaggerated by those who wish to ban it" (1996: 98). Ronald Dworkin insists that many of the claims in the literature about the harm of hate speech are 'inflated', and that some are outright 'absurd' (2009: vi). It is true that, in some parts of the literature, the harmfulness of hate speech is described in very emotive terms, and this is probably what inspires these sorts of complaints from Goldberger, Alexander, Dworkin, and others. However, in light of my discussion in this study, I suggest that a key component in the conflict between defenders of anti-hate speech law, and liberals who see the case for anti-hate speech law as inflated, is a disagreement that only comes into view once we stress the distinction between harm and responsibility. Parties on both sides of this conflict can agree that people in marginal social groups experience grave and enduring forms of disadvantage and exclusion on account of their social identity. In arguments about the legal restriction of hate speech, the genuine nexus of controversy is not (or should not be) whether there *are* genuine harms at stake in this arena, but whether the provenance and character of the harms that exist are such that hate-speakers can legitimately be regarded as responsible, answerable, liable, or accountable for them. In this study I have attempted to show that once we set aside the scenarios in which a direct causal connection between hate speech and personal harm can be discerned, the harms that we are left with – those that descend from multifariously-constituted, identity-based social hierarchy – are ones for which

[1] Goldberger is notable in this context for having previously acted as lead counsel in the American Civil Liberties Union's defence of the Skokie March. The ACLU is a frequent focus of criticism in work that opposes US free speech orthodoxy (see for instance Delgado 1983; MacKinnon 1987b).

hate-speakers cannot be adjudged liable. It is consistent with this conclusion, nevertheless, to criticise those who endorse a similar view about the justifiability of anti-hate speech law, only because they downplay the seriousness of the harms that have been associated with hate speech in the literature. The reason we should reject any general prohibition on hate speech is not that the adversities in this arena are minor, insubstantial, and overblown. Nor is it that a commitment to free speech peremptorily renders such prohibitions illegitimate. General prohibitions on hate speech are to be rejected, primarily, because we cannot provide a perspicuous and confident diagnosis about how identity-based social hierarchies are sustained and perpetuated. It may be the case, as several authors maintain, that hate speech does have an integral role in these processes. But that remains a conjecture, as far as any evidence or analyses we find in the literature go. It is not the kind of firm conclusion by whose lights we can use the coercive power of the law to designate hate-speakers – the vulgar and inarticulate face of identity-prejudice – as liability-bearers for the harms in question.

## 6.2 Eroding social hierarchy and spreading the burden

On the view that I have defended, the illegitimacy of general prohibitions of hate speech is largely a consequence of the epistemic limitations we face in ascertaining *how* identity-based social hierarchies are created and sustained, and relatedly, *who* can appropriately be held liable for the harms effected by those hierarchies. One important question to ask, in light of this view, is how the burdens which arise from those epistemic limitations can be and should be distributed. Systemic, socially-mediated harms are inherently difficult to trace to their 'source'. We may only very rarely be in a position where we can confidently identify a certain type of conduct as one that (i) contributes to a systemic socially-mediated harm, despite the fact that (ii) individual acts which token that type of conduct do not inflict harm on any specifiable individuals. If it is only legitimate to make people liable for harm when these judgements can be confidently rendered – which is what I have argued here – then we may find ourselves, inadvertently, adding to the burdens of those who already suffer under the weight of identity-based social hierarchies. Because the disciplinary and reformatory force of the law cannot be recruited to the cause of eroding social hierarchy under these circumstances, it is up to those

who are harmed by social hierarchy to promote reformatory efforts themselves, or instead, if they cannot do so, to stoically endure the harms that social hierarchy effects.[2]

Proponents of free speech orthodoxy have a ready – though all too often facile – response to these concerns. Assuming (for argument's sake) that hate speech *does* indirectly contribute to the harms of social hierarchy, but granting that legal restrictions are not justifiable because this contribution cannot be adequately demonstrated, the way forward, they will say, is to combat the pernicious effects of hate speech by engaging in 'counter-speech', or to use the common expression, by 'speaking back'. Can something helpful be made of this suggestion? As a putative remedy for the systemic, socially-mediated harms of hate speech, it is easy for 'speaking back' to be invoked as a formless panacea, in a way that ignores the difficulties and burdens involved in counter-speech itself. In many cases where speaking back is proposed as a response to hate speech, the suggestion just seems to be that the targets of hate speech should verbally defend themselves against hate-speakers, or level abuse at hate-speakers in turn. This fails to recognise, firstly, that identity-based social hierarchies impair people's communicative capacities, rendering such counter speech relatively ineffective, and secondly, that this 'remedy' continues to lay the burden of reforming social hierarchy entirely at the feet of those already disadvantaged by it (see Delgado and Yun 1995a: 299; Nielsen 2012). We would do somewhat better to say that it is the responsibility of *everyone* to engage in counter-speech against hate-speakers, not just of those who are personally victimised through hate speech. This proposal at least takes some account of how the burdens are distributed. However, there is also an unwarranted optimism in this putative remedy to the problems of hate speech. Those who aspire to a utopian pluralism may express a hope that "habits of empathy and altruism will be transported from the private to the public sphere primarily through free choice, not through legal coercion" (Neuborne 1992: 376). On its face, however, this seems to be a case of policy-making as wishful thinking. Surely we must do more than just gesturing towards the vague hope that societies will be able, through a spontaneous collective enlightenment, to do away with the attitudes underlying hate speech.

---

[2] Here I agree with Schauer again, that "the immediate fact that the cost of a constitutional right is being borne disproportionately by victims of its exercise ought at least to occasion more thought, especially in the First Amendment area, than it has to date" (1992: 1357).

A few authors in the hate speech literature have taken up this concern, outlining the rudiments of a more substantial version of 'the counter-speech remedy' to hate speech. Katharine Gelber argues that "individual targets of hate-speech-acts ought to be provided with the appropriate institutional, educational, and material support to empower them to respond to, and to seek to contradict, the discrimination enacted in hate-speech utterances" (Gelber 2012: 54; see also Gelber 2002). In a similar vein, Maleiha Malik says that

> cultural policy that encourages the participation of minorities in 'free speech' should be supported as a key way of addressing hate speech in the public sphere. Supply side investment that increases capacity within minority groups to respond to 'hate speech' may be preferable to the use of the criminal law. This could be an alternative to incitement to hatred legislation which often causes a conflict between equality and freedom of speech. (Malik 2011: 39)

Precisely what would be involved in the kind of 'supply side investment' and 'material support' that Gelber and Malik refer to is something to be spelled out in detail elsewhere, and not necessarily in an idiom native to legal and political philosophy. If what is needed to reform social hierarchy is a far-reaching array of educative initiatives, then it is reasonable to suppose that education theorists and developmental psychologists will have more of worth to say than political or legal philosophers about the shape that those initiatives take.

Whatever initiatives might emerge from this approach, it is important that we recognise the prospect of some middle ground between the alternatives of (i) prohibiting hate speech outright, and (ii) waiting patiently for anti-prejudicial, egalitarian views to conquer all in the marketplace of ideas. Where people are suffering greatly at the hands of their hate-filled neighbours, the liberal state has a responsibility to use its position of institutional authority and influence to intervene. For a certain class of harms associated with hate speech, I have argued that the most obvious tool for intervention – the general legal prohibition of hate speech – would not be justifiable, given what we can ascertain about where and with whom responsibility for that class of harms ultimately resides. However, having deprived the state of that particular means for engineering reform, our next port of call should be to conceive of and encourage the development of alternative means for engineering reform.

In a discussion of the challenges liberal states face in trying to engineer social reforms, Thomas Nagel says that human individuals "do not change most effectively and *en masse*

through personal conversion, but through the development of practices which form their sense of themselves and make it natural for them to be guided by different priorities and values, different requirements and inhibitions" (Nagel 1991: 60). The criminal law is one very important institution through which the liberal state can guide the reformation of individuals' behaviour, without relying on the futile hope of widespread 'personal conversion'. When the harms associated with hate speech are such that responsibility for them can be clearly and convincingly ascribed to hate-speakers – that is to say, when hate speech is *directly* harmful – the liberal state can and should use legal penalties to engender values and inhibitions that will deter the verbal behaviours which bring about those harms. However, when the harms in question are such that responsibility for them cannot be ascribed to hate-speakers, and where legal restrictions must be eschewed accordingly, the state's response cannot just be to hope that counter-speech will bring about personal conversions that cumulatively erode harmful social hierarchies. The erosion of social hierarchy is a goal that liberal societies must aim collectively to pursue through the organs of the state. However, the methods employed in the pursuit should not lay blame and liability for social hierarchy at the feet of hate-speakers. Hate-speakers may resist, obstruct, and resent the reforms in question in a way that most of us do not. Nevertheless, we cannot confidently judge that their speech makes a significant contribution to the harmful social systems that we are trying to eradicate.

# Bibliography


Acosta, X. E. (1989). 'The Tort of "Outrageous Conduct" in New Mexico: Intentional Infliction of Emotional Harm without Physical Injury', *New Mexico Law Review* 19 (2): 425-49.

Adler, Amy (1996). 'What's Left?: Hate Speech, Pornography, and the Problem for Artistic Expression', *California Law Review* 84 (6): 1499-572.

Alexander, Larry (1994). 'Harm, Offense, and Morality', *Canadian Journal of Law and Jurisprudence* 7 (2): 199-216.

_____ (1996). 'Banning Hate Speech and the Sticks and Stones Defense', *Constitutional Commentary* 13 (1): 71-100.

_____ (2000). 'Incitement and Freedom of Speech', in David Kretzmer and Francine Kershman Hazan (Eds.), *Freedom of Speech and Incitement against Democracy*, London: Kluwer Law International.

Alexander, Lawrence, and Paul Horton (1984). 'The Impossibility of a Free Speech Principle', *Northwestern University Law Review* 78 (5): 1319-57.

Alston, William P. (1994). 'Illocutionary Acts and Linguistic Meaning', in Savas L. Tsohatzidis (Ed.), *Foundations of Speech Act Theory: Philosophical and Linguistic Perspectives*, London: Routledge.

Altman, Andrew (1993). 'Liberalism and Campus Hate Speech: A Philosophical Examination', *Ethics* 103 (2): 302-17.

_____ (2012). 'Freedom of Expression and Human Rights Law: The Case of Holocaust Denial', in Ishani Maitra and Mary Kate McGowan (Eds.), *Speech and Harm: Controversies over Free Speech*, Oxford: Oxford University Press.

Amdur, Robert (1980). 'Scanlon on Freedom of Expression', *Philosophy & Public Affairs* 9 (3): 287-300.

Anderson, David A. (2002). 'Incitement and Tort Law', *Wake Forest Law Review* 37: 957-1008.

Anderson, Luvell, and Ernie Lepore (forthcoming). 'Slurring Words'.

Appiah, Kwame Anthony (1989). 'The Conservation of 'Race'', *Black American Literature Forum* 23 (1): 37-60.

_____ (2005). *The Ethics of Identity*, Princeton: Princeton University Press.

Aristotle (1985). *The Nicomachean Ethics* (Terence Irwin, Trans.), Indianapolis: Hackett.

Arkes, Hadley (1974). 'Civility and the Restriction of Speech: Rediscovering the Defamation of Groups', *Supreme Court Review* 1974: 281-335.

_____ (1988). 'A Communitarian Defense of Group Libel Laws', *Harvard Law Review* 101 (3): 682-701.

Asad, Talal (2009). 'Free Speech, Blasphemy, and Secular Criticism', in *Is Critique Secular? Blasphemy, Injury, and Free Speech*, Berkeley: The Townsend Centre for the Humanities.

Ashworth, Andrew (2009). *Principles of Criminal Law* (sixth edition), Oxford: Oxford University Press.

Austin, J. L. (1975). *How to Do Things with Words* (second edition), Cambridge Massachusetts: Harvard University Press (originally published 1962).

Baker, C. Edwin (1978). 'Scope of the First Amendment Freedom of Speech', *UCLA Law Review* 25 (5): 964-1040.

_____ (1989). *Human Liberty and Freedom of Speech*, New York: Oxford University Press.

_____ (2009). 'Autonomy and Hate Speech', in Ivan Hare and James Weinstein (Eds.), *Extreme Speech and Democracy*, Oxford: Oxford University Press.

Bakircioglu, Onder (2008). 'Freedom of Expression and Hate Speech', *Tulsa Journal of Comparative & International Law* 16: 1-49.

Bannister, Judith (2008). 'It's Not What You Say but the Way That You Say It', *Florida State University Law Review* 36 (1): 23-40.

Banton, Michael (1992). 'The Declaratory Value of Laws against Racial Incitement', in Sandra Coliver (Ed.), *Striking a Balance: Hate Speech, Freedom of Expression, and Non-Discrimination*, London: Article 19, London and Human Rights Centre.

Barendt, Eric (2005). *Freedom of Speech* (second edition), Oxford: Oxford University Press.

_____ (2009). 'Incitement to, and Glorification of, Terrorism', in Ivan Hare and James Weinstein (Eds.), *Extreme Speech and Democracy*, Oxford: Oxford University Press.

_____ (2011). 'Religious Hatred Laws: Protecting Groups or Belief?', *Res Publica* 17 (1): 41-53.

Barnes, Robin D. (1995). 'Everyday Racism and the Campaign of Hatred', in Laura J. Lederer and Richard Delgado (Eds.), *The Price We Pay: The Case against Racist Speech, Hate Propaganda, and Pornography*, New York: Hill and Wang.

Barnum, David G. (2006). 'Indirect Incitment and Freedom of Speech in Anglo-American Law', *European Human Rights Law Review* 3: 258-80.

Barrow, Robin (2005). 'On the Duty of Not Taking Offence', *Journal of Moral Education* 34 (3): 265-75.

Battaglia, Jack M. (1991). 'Regulation of Hate Speech by Educational Institutions: A Proposed Policy', *Santa Clara Law Review* 31 (2): 345-92.

Bayles, Michael (1973). 'Offensive Conduct and the Law', in Norman S. Care and Thomas K. Trelogan (Eds.), *Issues in Law and Morality*, Cleveland: Western Reserve University.

_____ (1986). 'Mid-Level Principles and Justification', in J. Roland Pennock and John W. Chapman (Eds.), *Nomos XXVIII: Justification*, New York: New York University Press.

Bell, Derrick A. (1995). 'Who's Afraid of Critical Race Theory?', *University of Illinois Law Review* 1995 (4): 893-910.

Benesch, Susan (2007). 'Vile Crime or Inalienable Right: Defining Incitement to Genocide', *Virginia Journal of International Law* 48 (3): 485-528.

Bianchi, Claudia (2008). 'Indexicals, Speech Acts and Pornography', *Analysis* 68 (4): 310-16.

Bindman, Geoffrey (1992). 'Incitement to Racial Hatred in the United Kingdom: Have We Got the Law We Need?', in Sandra Coliver (Ed.), *Striking a Balance: Hate Speech, Freedom of Expression, and Non-Discrimination*, London: Article 19, London and Human Rights Centre.

Bird, Alexander (2002). 'Illocutionary Silencing', *Pacific Philosophical Quarterly* 83 (1): 1-15.

Bird, Colin (2007). 'Harm Versus Sovereignty: A Reply to Ripstein', *Philosophy & Public Affairs* 35 (2): 179-94.

Blackford, Russell (2012). *Freedom of Religion and the Secular State*, Oxford: Wiley-Blackwell.

Blasi, Vincent (1985). 'The Pathological Perspective and the First Amendment', *Columbia Law Review* 85 (4): 449-514.

Blocher, Joseph (2008). 'Institutions in the Marketplace of Ideas', *Duke Law Journal* 57 (4): 821-89.

Blum, Lawrence (2004). 'Stereotypes and Stereotyping: A Moral Analysis', *Philosophical Papers* 33 (3): 251-89.

Bobo, Lawrence D. (1999). 'Prejudice as Group Position: Microfoundations of a Sociological Approach to Racism and Race Relations', *Journal of Social Issues* 55 (3): 445-72.

Bodney, David J. (2009). 'Extreme Speech and American Press Freedoms', in Ivan Hare and James Weinstein (Eds.), *Extreme Speech and Democracy*, Oxford: Oxford University Press.

Bohman, James (1997). 'Deliberative Democracy and Effective Social Freedom: Capabilities, Resources, and Opportunities', in James Bohman and William Rehg (Eds.), *Deliberative Democracy: Essays on Reason and Politics*, Cambridge Massachusetts: MIT Press.

Bollinger, Lee C. (1986). *The Tolerant Society*, Oxford: Oxford University Press.

Borovoy, Alan, Kathleen Mahoney, Barry Brown, Jamie Cameron, David Goldberger, and Mari Matsuda (1989). 'Language as Violence Versus Freedom of Expression: Canadian and American Perspectives on Group Defamation', *Buffalo Law Review* 37 (2): 337-73.

Bourdieu, Pierre (1991). 'On Symbolic Power' (Gino Raymond and Matthew Adamson, Trans.), in John B. Thompson (Ed.), *Language and Symbolic Power*, Cambridge: Polity Press.

Boyle, Kevin (1992). 'Overview of a Dilemma: Censorship Versus Racism', in Sandra Coliver (Ed.), *Striking a Balance: Hate Speech, Freedom of Expression, and Non-Discrimination*, London: Article 19, London and Human Rights Centre.

_____ (2001). 'Hate Speech: The United States Versus the Rest of the World?', *Maine Law Review* 53 (2): 488-502.

Braddon-Mitchell, David, and Caroline West (2004). 'What Is Free Speech?', *The Journal of Political Philosophy* 12 (4): 437-60.

Bradley, Ben (2012). 'Doing Away with Harm', *Philosophy and Phenomenological Research* 85 (2): 390-412.

Breckheimer, Peter J. (2001). 'A Haven for Hate: The Foreign and Domestic Implications of Protecting Internet Hate Speech under the First Amendment', *Southern California Law Review* 75: 1493-528.

Brink, David O. (2001). 'Millian Principles, Freedom of Expression, and Hate Speech', *Legal Theory* 7 (2): 119-57.

Brison, Susan J. (1998a). 'The Autonomy Defense of Free Speech', *Ethics* 108 (2): 312-39.

_____ (1998b). 'Speech, Harm, and the Mind-Body Problem in First Amendment Jurisprudence', *Legal Theory* 4: 39-61.

Brown, Alexander (2008). 'The Racial and Religious Hatred Act 2006: A Millian Response', *Critical Review of International Social and Political Philosophy* 11 (1): 1-24.

Brownmiller, Susan (2000). *In Our Time: Memoir of a Revolution*, London: Aurum Press.

Brownstein, Alan E. (1991). 'Regulating Hate Speech at Public Universities: Are First Amendment Values Functionally Incompatible with Equal Protection Principles?', *Buffalo Law Review* 39 (1): 1-52.

Buchanan, Allen (1979). 'Autonomy and Categories of Expression: A Reply to Professor Scanlon', *University of Pittsburgh Law Review* 40: 551-61.

Buss, William (2005). 'Constitutional Words About Words: Protected Speech and "Fighting Words" under the Australian and American Constitutions', *Transnational Law and Contemporary Problems* 15 (2): 489-513.

Butler, Judith (1997). *Excitable Speech: A Politics of the Performative*, New York: Routledge.

_____ (1998). 'Ruled Out: Vocabularies of the Censor', in Robert C. Post (Ed.), *Censorship and Silencing: Practices of Cultural Regulation*, Los Angeles: Getty Research Institute for the History of Art and the Humanities.

_____ (2009). 'The Sensibility of Critique: Response to Asad and Mahmood', in *Is Critique Secular? Blasphemy, Injury, and Free Speech*, Berkeley: The Townsend Centre for the Humanities.

Calvert, Clay (1997). 'Hate Speech and Its Harms: A Communication Theory Perspective', *Journal of Communication* 47 (1): 4-19.

Canavan, Francis (1999). 'Speech That Matters', *Society* 36 (6): 11-15.

Carey, James W. (1989). *Communication as Culture: Essays on Media and Society*, New York: Routledge.

Carmi, Guy E. (2007). 'Dignity – the Enemy from Within: A Theoretical and Comparative Analysis of Human Dignity as a Free Speech Justification', *University of Pennsylvania Journal of Constitutional Law* 9 (4): 957-1001.

Carter, Angela (2000). 'Polemical Preface: Pornography in the Service of Women', in Drucilla Cornell (Ed.), *Feminism and Pornography*, Oxford: Oxford University Press.

Chevigny, Paul G. (1980). 'Philosophy of Language and Free Expression', *New York University Law Review* 55 (2): 157-94.

Chiang, Jerry C. (2007). 'Plainly Offensive Babel: An Analytical Framework for Regulating Plainly Offensive Speech in Public Schools', *Washington Law Review* 82 (2): 403-30.

Choudhury, Tufyal (2009). 'The Terrorism Act 2006: Discouraging Terrorism', in Ivan Hare and James Weinstein (Eds.), *Extreme Speech and Democracy*, Oxford: Oxford University Press.

Cohen-Almagor, Raphael (1993). 'Harm Principle, Offence Principle, and the Skokie Affair', *Political Studies* 41 (3): 453-70.

Cohen, G. A. (1989). 'On the Currency of Egalitarian Justice', *Ethics* 99 (4): 906-44.

Cohen, Jonathan R. (2007). 'Coping with Lasting Social Injustice', *Washington and Lee Journal of Civil Rights and Social Justice* 13 (2): 259-84.

Cohen, Joshua (1993). 'Freedom of Expression', *Philosophy & Public Affairs* 22 (3): 207-63.

_____ (1997). 'Procedure and Substance in Deliberative Democracy', in James Bohman and William Rehg (Eds.), *Deliberative Democracy: Essays on Reason and Politics*, Cambridge Massachusetts: MIT Press.

_____ (2006). 'Deliberation and Democratic Legitimacy', in Robert E. Goodin and Philip Pettit (Eds.), *Contemporary Political Philosophy: An Anthology* (second edition), Oxford: Blackwell.

Coliver, Sandra (1992). 'Hate Speech Laws: Do They Work?', in Sandra Coliver (Ed.), *Striking a Balance: Hate Speech, Freedom of Expression, and Non-Discrimination*, London: Article 19, London and Human Rights Centre.

Collier, Charles W. (2001). 'Hate Speech and the Mind-Body Problem: A Critique of Postmodern Censorship Theory', *Legal Theory* 7 (2): 203-34.

Connolly, William E. (1993). *The Terms of Political Discourse* (third edition), Oxford: Blackwell.

Corlett, J. Angelo, and Robert Francescotti (2002). 'Foundations of a Theory of Hate Speech', *The Wayne Law Review* 48 (3): 1071-1100.

Cortese, Anthony (2006). *Opposing Hate Speech*, Westport Connecticut: Praeger.

Cowan, Gloria, Becky Heiple, Carolyn Marquez, Desiree Khatchadourian, and Michelle McNevin (2005). 'Heterosexual's Attitudes Towards Hate Crimes and Hate Speech against Gays and Lesbians: Old-Fashioned and Modern Heterosexism', *Journal of Homosexuality* 49 (2): 67-82.

Cowan, Gloria, and Cyndi Hodge (1996). 'Judgements of Hate Speech: The Effects of Target Group, Publicness, and Behavioural Responses of the Target', *Journal of Applied Social Psychology* 26 (4): 355-74.

Cowan, Gloria, and Desiree Khatchadourian (2003). 'Empathy, Ways of Knowing, and Interdependence as Mediators of Gender Differences in Attitudes Towards Hate Speech and Freedom of Speech', *Psychology of Women Quarterly* 27 (4): 300-08.

Cowan, Gloria, and Jon Mettrick (2002). 'The Effects of Target Variables and Setting on the Perception of Hate Speech', *Journal of Applied Social Psychology* 32 (2): 277-99.

Cowan, Gloria, Miriam Resendez, Elizabeth Marshall, and Ryan Quist (2002). 'Hate Speech and Constitutional Protection: Priming Values of Equality and Freedom', *Journal of Social Issues* 58 (2): 247-63.

Cox, Phil (1995). 'The Disputation of Hate: Speech Codes, Pluralism, and Academic Freedom', *Social Theory and Practice* 21 (1): 113-44.

Cram, Ian (2009). 'The Danish Cartoons, Offensive Expression, and Democratic Legitimacy', in Ivan Hare and James Weinstein (Eds.), *Extreme Speech and Democracy*, Oxford: Oxford University Press.

Crawford, Donald W. (1968). 'Can Disputes over Censorship Be Resolved?', *Ethics* 78 (2): 93-108.

Crenshaw, Kimberle Williams (1993). 'Beyond Racism and Misogyny: Black Feminism and 2 Live Crew', *Words That Wound: Critical Race Theory, Assaultive Speech, and the First Amendment*, Boulder Colorado: Westview Press.

_____ (1995). 'Comments of an Outsider on the First Amendment', in Laura J. Lederer and Richard Delgado (Eds.), *The Price We Pay: The Case against Racist Speech, Hate Propaganda, and Pornography*, New York: Hill and Wang.

Croom, Adam M. (2008). 'Racial Epithets: What We Say and Mean by Them', *Dialogue* 51 (1): 34-45.

_____ (2011). 'Slurs', *Language Sciences* 33 (3): 343-58.

Curtis, Michael Kent (1995). 'Critics of "Free Speech" and the Uses of the Past', *Constitutional Commentary* 12: 29-65.

_____ (2009). 'Be Careful What You Wish For: Gays, Dueling High School T-Shirts, and the Perils of Suppression', *Wake Forest Law Review* 44 (2): 431-95.

Dan-Cohen, Meir (1999). 'Harmful Thoughts', *Law and Philosophy* 18 (4): 379-405.

Davidson, Donald (1984). 'Communication and Convention', *Synthese* 59 (1): 3-17.

Davis, Wayne (1992). 'Speaker Meaning', *Linguistics and Philosophy* 15 (3): 223-53.

_____ (2010). 'Implicature', in Edward N. Zalta (Ed.), *Stanford Encyclopedia of Philosophy*, http://plato.stanford.edu/archives/win2010/entries/implicature/.

Delgado, Richard (1983). 'Professor Delgado Replies', *Harvard Civil Rights-Civil Liberties Law Review* 18 (2): 593-97.

_____ (1989). 'Storytelling for Oppositionists and Others', *Michigan Law Review* 87 (8): 2411-41.

_____ (1993). 'Words That Wound: A Tort Action for Racial Insults, Epithets, and Name Calling', *Words That Wound: Critical Race Theory, Assaultive Speech, and the First Amendment*, Boulder Colorado: Westview Press.

_____ (1994). 'Rodrigo's Eighth Chronicle: Black Crime, White Fears. On the Social Construction of Threat', *Virginia Law Review* 80 (2): 503-48.

_____ (1996). 'Rodrigo's Eleventh Chronicle: Empathy and False Empathy', *California Law Review* 84 (1): 61-100.

_____ (1998). 'Are Hate-Speech Rules Constitutional Heresy? A Reply to Steven Gey', *University of Pennsylvania Law Review* 146 (3): 865-79.

_____ (2000). 'Toward a Legal Realist View of the First Amendment', *Harvard Law Review* 113 (3): 778-802.

_____ (2006). 'Foreword: Several Things We Know About Hate Speech... And Several Misconceptions', *Opposing Hate Speech*, Westport Connecticut: Praeger.

Delgado, Richard, and Jean Stefancic (1992a). 'Images of the Outsider in American Law and Culture: Can Free Expression Remedy Systemic Social Ills?', *Cornell Law Review* 77 (6): 1258-97.

_____ (1992b). 'Pornography and Harm to Women: "No Empirical Evidence?"', *Ohio State Law Journal* 53 (4): 1037-56.

_____ (1994). 'Hateful Speech, Loving Communities: Why Our Notion of "a Just Balance" Changes So Slowly', *California Law Review* 82 (4): 851-69.

_____ (1995). 'Ten Arguments against Hate Speech Regulation: How Valid?', *Northern Kentucky Law Review* 23 (3): 475-90.

_____ (1996). 'Apologize and Move On?: Finding a Remedy for Pornography, Insult, and Hate Speech', *University of Colorado Law Review* 67 (1): 93-112.

_____ (1997). *Must We Defend Nazis? Hate Speech, Pornography, and the New First Amendment*, New York: New York University Press.

_____ (2004). *Understanding Words That Wound*, Boulder Colorado: Westview.

_____ (2009). 'Four Observations About Hate Speech', *Wake Forest Law Review* 44 (2): 353-70.

Delgado, Richard, and David Yun (1994). 'The Neoconservative Case against Hate Speech Regulation – Lively, D'Souza, Gates, Carter, and the Toughlove Crowd', *Vanderbilt Law Review* 47 (6): 1807-25.

_____ (1995a). 'Pressure Valves and Bloodied Chickens: An Assessment of Four Paternalistic Arguments for Resisting Hate-Speech Regulation', in Laura J. Lederer and Richard Delgado (Eds.), *The Price We Pay: The Case against Racist Speech, Hate Propaganda, and Pornography*, New York: Hill and Wang.

_____ (1995b). '"The Speech We Hate": First Amendment Totalism, the ACLU, and the Principle of Dialogic Politics', *Arizona State Law Journal* 27 (4): 1281-300.

Demaske, Chris (2004). 'Modern Power and the First Amendment: Reassessing Hate Speech', *Communication Law and Policy* 9 (3): 273-316.

Devlin, Patrick (1968). *The Enforcement of Morals*, Oxford: Oxford University Press.

Doris, John M. (2002). *Lack of Character: Personality and Moral Behaviour*, Cambridge: Cambridge University Press.

Douglas-Scott, Sionaidh (1999). 'The Hatefulness of Protected Speech: A Comparison of the American and European Approaches', *William and Mary Bill of Rights Journal* 7 (2): 305-46.

Douglas, Lawrence (1998). 'Policing the Past: Holocaust Denial and the Law', in Robert C. Post (Ed.), *Censorship and Silencing: Practices of Cultural Regulation*, Los Angeles: Getty Research Institute for the History of Art and the Humanities.

Downs, Donald A. (1985). 'Skokie Revisited: Hate Group Speech and the First Amendment', *Notre Dame Law Review* 60 (4): 629-85.

Duff, R. A. (1990). *Intention, Agency and Criminal Liability: Philosophy of Action and the Criminal Law*, Oxford: Basil Blackwell.

_____ (1993). 'Choice, Character, and Criminal Liability', *Law and Philosophy* 12 (4): 345-83.

Dworkin, Andrea, and Catharine A. MacKinnon (1988). *Pornography & Civil Rights: A New Day for Women's Equality*, Minneapolis: Organizing Against Pornography.

Dworkin, Ronald (1978). 'Liberalism', in Stuart Hampshire (Ed.), *Public and Private Morality*, Cambridge: Cambridge University Press.

_____ (1981a). 'Is There a Right to Pornography?', *Oxford Journal of Legal Studies* 1 (2): 177-212.

_____ (1981b). 'What is Equality? Part 1: Equality of Welfare', *Philosophy & Public Affairs* 10 (2): 185-246.

_____ (1985). *A Matter of Principle*, Cambridge Massachusetts: Harvard University Press.

_____ (1986). *Law's Empire*, Cambridge Massachusetts: Belknap.

_____ (1995). 'Liberty and Pornography', in Susan Dwyer (Ed.), *The Problem of Pornography*, Belmont California: Wadsworth.

_____ (1996). 'Why Must Speech Be Free?', in *Freedom's Law: The Moral Reading of the American Constitution*, Cambridge Massachusetts: Harvard University Press.

_____ (2000). *Sovereign Virtue: The Theory and Practice of Equality*, Cambridge Massachusetts: Harvard University Press.

_____ (2001). 'Do Values Conflict? A Hedgehog's Approach', *Arizona Law Review* 43 (2): 251-59.

_____ (2006). 'The Right to Ridicule', *New York Review of Books*, March 23, 2006.

_____ (2009). 'Foreword', in Ivan Hare and James Weinstein (Eds.), *Extreme Speech and Democracy*, Oxford: Oxford University Press.

Dyzenhaus, David (1992). 'John Stuart Mill and the Harm of Pornography', *Ethics* 102 (3): 534-51.

Eastman, Kate (1992). 'Racial Vilification: The Australian Experience', in Sandra Coliver (Ed.), *Striking a Balance: Hate Speech, Freedom of Expression, and Non-Discrimination*, London: Article 19, London and Human Rights Centre.

_____ (2007). 'Problems with Evidence in Hate Speech Cases', in Katharine Gelber and Adrienne Stone (Eds.), *Hate Speech and Freedom of Speech in Australia*, Sydney: Federation Press.

Easton, Susan (1995). 'Autonomy and the Free Speech Principle', *Journal of Applied Philosophy* 12 (1): 27-39.

Eberle, Edward J. (1994). 'Hate Speech, Offensive Speech, and Public Discourse in America', *Wake Forest Law Review* 29 (4): 1135-213.

Edgar, David (2009). 'Shouting Fire – From the Nanny State to the Heckler's Veto: The New Censorship and How to Counter It', in Ivan Hare and James Weinstein (Eds.), *Extreme Speech and Democracy*, Oxford: Oxford University Press.

Eggerking, Kitty (1992). 'Australia: The Role of the Media in Perpetuating Racism', in Sandra Coliver (Ed.), *Striking a Balance: Hate Speech, Freedom of Expression, and Non-Discrimination*, London: Article 19, London and Human Rights Centre.

Ellis, Anthony (1984). 'Offense and the Liberal Conception of the Law', *Philosophy & Public Affairs* 13 (1): 3-23.

Emerson, Thomas I. (1963). 'Toward a General Theory of the First Amendment', *The Yale Law Journal* 72 (5): 877-956.

Estlund, David (1997). 'Beyond Fairness and Deliberation: The Epistemic Dimension of Democratic Authority', in James Bohman and William Rehg (Eds.), *Deliberative Democracy: Essays on Reason and Politics*, Cambridge Massachusetts: MIT Press.

Evans, Carolyn (2009). 'Religious Speech That Undermines Gender Equality', in Ivan Hare and James Weinstein (Eds.), *Extreme Speech and Democracy*, Oxford: Oxford University Press.

Feinberg, Joel (1973). *Social Philosophy*, Englewood Cliffs: Prentice-Hall.

_____ (1975). 'Freedom of Expression and Its Limits', in Joel Feinberg and Hyman Gross (Eds.), *Philosophy of Law*, Encino California: Dickenson.

_____ (1985a). *Harm to Others* (The Moral Limist of the Criminal Law, Volume 1), Oxford: Oxford University Press.

_____ (1985b). *Offence to Others* (The Moral Limist of the Criminal Law, Volume 2), Oxford: Oxford University Press.

_____ (1988). *Harmless Wrongdoing* (The Moral Limist of the Criminal Law, Volume 4), Oxford: Oxford University Press.

_____ (1994). 'Harmless Wrongdoing', in Gerald Dworkin (Ed.), *Morality, Harm, and the Law*, Boulder Colorado: Westview Press.

Fish, Stanley (1994a). 'The Common Touch, or, One Size Fits All', *There's No Such Thing as Free Speech, and It's a Good Thing Too*, Oxford: Oxford University Press.

_____ (1994b). 'Introduction: "That's Not Fair"', *There's No Such Thing as Free Speech, and It's a Good Thing Too*, Oxford: Oxford University Press.

_____ (1994c). 'Jerry Falwell's Mother, or, Where's the Harm?', *There's No Such Thing as Free Speech, and It's a Good Thing Too*, Oxford: Oxford University Press.

_____ (1994d). 'There's No Such Thing as Free Speech, and It's a Good Thing Too', *There's No Such Thing as Free Speech, and It's a Good Thing Too*, Oxford: Oxford University Press.

_____ (1997). 'Boutique Multiculturalism, or Why Liberals Are Incapable of Thinking About Hate Speech', *Critical Inquiry* 23 (Winter 1997): 378-95.

_____ (2006). 'Our Faith in Letting It All Hang Out', *New York Times*, February 12, 2006.

Fiss, Owen M. (1986). 'Free Speech and Social Structure', *Iowa Law Review* 71 (5): 1405-25.

_____ (1991). 'Freedom and Feminism', *Georgetown Law Journal* 80 (2041-2062).

_____ (1995). 'The Supreme Court and the Problem of Hate Speech', *Capital University Law Review* 24 (2): 281-91.

_____ (1996a). *The Irony of Free Speech*, Cambridge Massachusetts: Harvard University Press.

_____ (1996b). *Liberalism Divided: Freedom of Speech and the Many Uses of State Power*, Boulder Colorado: Westview Press.

Foucault, Michel (1979). *The History of Sexuality: Volume I* (Robert Hurley, Trans.), London: Allen Lane.

Fraser, David (2009). "On the Internet, Nobody Knows You're a Nazi': Some Comparative Legal Aspects of Holocaust Denial on the WWW', in Ivan Hare and James Weinstein (Eds.), *Extreme Speech and Democracy*, Oxford: Oxford University Press.

Fraser, Nancy (1990). 'Feminist Politics in the Age of Recognition: A Two-Dimensional Approach to Gender Justice', *Studies in Social Justice* 1 (1): 23-35.

Fredman, Sandra (2002). *Discrimination Law*, Oxford: Oxford University Press.

Fricker, Miranda (1998). 'Rational Authority and Social Power: Towards a Truly Social Epistemology', *Proceedings of the Aristotelian Society* 98: 159-77.

_____ (2007). *Epistemic Injustice: Power and the Ethics of Knowing*, Oxford: Oxford University Press.

Frye, Marilyn (1983). *The Politics of Reality*, Freedom California: Crossing Press.

Futch, Devona (1995). 'The Cases for Regulation of Hate Speech and Pornography: Are They the Same or Different?', in Laura J. Lederer and Richard Delgado (Eds.), *The Price We Pay: The Case against Racist Speech, Hate Propaganda, and Pornography*, New York: Hill and Wang.

Gallie, W. B. (1962). 'Essentially Contested Concepts', in Max Black (Ed.), *The Importance of Language*, Englewood Cliffs New Jersey: Prentice-Hall.

Gardner, John (1998). 'On the Ground of Her Sex(uality)', *Oxford Journal of Legal Studies* 18 (1): 167-87.

_____ (2003). 'The Mark of Responsibility', *Oxford Journal of Legal Studies* 23 (2): 157-71.

_____ (2007a). 'Complicity and Causality', *Offences and Defences: Selected Essays in the Philosophy of Criminal Law*, Oxford: Oxford University Press.

_____ (2007b). 'The Functions and Justifications of Criminal Law and Punishment', *Offences and Defences: Selected Essays in the Philosophy of Criminal Law*, Oxford: Oxford University Press.

_____ (2007c). 'The Gist of Excuses', *Offences and Defences: Selected Essays in the Philosophy of Criminal Law*, Oxford: Oxford University Press.

_____ (2007d). 'The Wrongness of Rape', *Offences and Defences: Selected Essays in the Philosophy of Criminal Law*, Oxford: Oxford University Press.

Gates, Henry Louis (1993). 'Let Them Talk: Why Civil Liberties Pose No Threat to Civil Rights', *New Republic* 37 (September 27): 43-49.

Geddis, Andrew (2004). 'Free Speech Martyrs or Unreasonable Threats to Social Peace? "Insulting" Expression and Section 5 of the Public Order Act 1986', *Public Law* 2004 (4): 853-74.

Gelber, Katharine (2002). *Speaking Back: The Free Speech Versus Hate Speech Debate*, Amsterdam: John Benjamins.

_____ (2007). 'Hate Speech and the Australian Legal and Political Landscape', in Katharine Gelber and Adrienne Stone (Eds.), *Hate Speech and Freedom of Speech in Australia*, Sydney: Federation Press.

_____ (2010). 'Freedom of Political Speech, Hate Speech and the Argument from Democracy: The Transformative Contribution of Capabilities Theory', *Contemporary Political Theory* 9 (3): 304-24.

_____ (2012). "Speaking Back': The Likely Fate of Hate Speech Policy in the United States and Australia', in Ishani Maitra and Mary Kate McGowan (Eds.), *Speech and Harm: Controversies over Free Speech*, Oxford: Oxford University Press.

Geuss, Raymond (1981). *The Idea of a Critical Theory: Habermas and the Frankfurt School*, Cambridge: Cambridge University Press.

Gibbs, Jack P. (1975). *Crime, Punishment, and Deterrence*, New York: Elsevier.

Gillborn, David (2009). 'Risk-Free Racism: Whiteness and So-Called "Free Speech"', *Wake Forest Law Review* 44 (2): 535-55.

Gilmore, Jonathan (2011). 'Expression as Realization: Speakers' Interests in Freedom of Speech', *Law and Philosophy* 30 (5): 517-39.

Gilreath, Shannon (2009). "'Tell Your Faggot Friend He Owes Me $500 for My Broken Hand": Thoughts on a Substantive Equality Theory of Free Speech', *Wake Forest Law Review* 44 (2): 557-615.

Glass, Marvin (1978). 'Anti-Racism and Unlimited Freedom of Speech: An Untenable Dualism', *Canadian Journal of Philosophy* 8 (3): 559-75.

Golding, Martin P. (2000). *Free Speech on Campus*, Oxford: Rowman & Littlefield.

Goldman, Alvin I. (1991). 'Epistemic Paternalism: Communication Control in Law and Society', *The Journal of Philosophy* 88 (3): 113-31.

Goldman, Alvin I., and James C. Cox (1996). 'Speech, Truth, and the Free Market for Ideas', *Legal Theory* 2 (1): 1-32.

Goodall, Kay (2007). 'Incitement to Religious Hatred: All Talk and No Substance?', *The Modern Law Review* 70 (1): 89-113.

Gordon, David (1984). 'Honderich on Morality Dependent Harm', *Political Studies* 32 (2): 288-95.

Gordon, Gregory S. (2004). "'A War of Media, Words, Newspapers, and Radio Stations": The ICTR Media Trial Verdict and a New Chapter in the International Law of Hate Speech', *Virginia Journal of International Law* 45 (1): 139-98.

Gordon, Paul (1992). 'Racist Violence: The Expression of Hate in Europe', in Sandra Coliver (Ed.), *Striking a Balance: Hate Speech, Freedom of Expression, and Non-Discrimination*, London: Article 19, London and Human Rights Centre.

Gorman, David (1999). 'The Use and Abuse of Speech-Act Theory in Criticism', *Poetics Today* 20 (1): 93-119.

Gould, John B. (2001). 'The Precedent That Wasn't: College Hate Speech Codes and the Two Faces of Legal Compliance', *Law & Society Review* 35 (2): 345-92.

Gray, John (1983). 'Political Power, Social Theory, and Essential Contestability', in David Miller and Larry Siedentop (Eds), *The Nature of Political Theory*, Oxford: Clarendon Press.

Great Britain Committee on Obscenity and Film Censorship, Chairman: Bernard Williams (1979). *Report of the Committee on Obscenity and Film Censorship*. London: Her Majesty's Stationery Office.

Green, Leslie (1991). 'Freedom of Expression and Choice of Language', *Law & Policy* 13 (3): 215-29.

_____ (1998). 'Pornographizing, Subordinating, and Silencing', in Robert C. Post (Ed.), *Censorship and Silencing: Practices of Cultural Regulation*, Los Angeles: Getty Research Institute for the History of Art and the Humanities.

_____ (2000). 'Pornographies', *The Journal of Political Philosophy* 8 (1): 27-52.

_____ (2008). 'On Being Tolerated', in Matthew H. Kramer, Claire Grant, Ben Colburn, and Antonty Hatzistavrou (Eds.), *The Legacy of H. L. A. Hart: Legal, Political, and Moral Philosophy*, Oxford: Oxford University Press.

Green, Mitchell S. (1999). 'Illocutions, Implicata, and What a Conversation Requires', *Pragmatics & Cognition* 7: 65-91.

_____ (2000). 'Illocutionary Force and Semantic Content', *Linguistics and Philosophy* 23 (5): 435-73.

Greenawalt, Kent (1989a). *Speech, Crime, and the Uses of Language*, Oxford: Oxford University Press.

_____ (1989b). 'Free Speech Justifications', *Columbia Law Review* 89 (1): 119–55.

Greenberg, Jeff, and Tom Pyszczynski (1985). 'The Effect of an Overheard Slur on Evaluations of the Target: How to Spread a Social Disease', *Journal of Experimental Social Psychology* 21 (1): 61-72.

Grice, H. P. (1957). 'Meaning', *The Philosophical Review* 66 (3): 377-88.

_____ (1989). 'Logic and Conversation', *Studies in the Way of Words*, Cambridge Massachusetts: Harvard University Press.

Griffin, Anthony P. (1994). 'The First Amendment and the Art of Storytelling', *Speaking of Race, Speaking of Sex: Hate Speech, Civil Rights, and Civil Liberties*, New York: New York University Press.

Grimm, Dieter (2009a). 'Freedom of Speech in a Globalized World', in Ivan Hare and James Weinstein (Eds.), *Extreme Speech and Democracy*, Oxford: Oxford University Press.

_____ (2009b). 'The Holocaust Denial Decision of the Federal Constitutional Court of Germany', in Ivan Hare and James Weinstein (Eds.), *Extreme Speech and Democracy*, Oxford: Oxford University Press.

Habermas, Jurgen (1979). *Communication and the Evolution of Society* (Thomas McCarthy, Trans.), London: Heinemann.

_____ (1984). *The Theory of Communicative Action, Volume 1: Reason and the Rationalization of Society* (Thomas McCarthy, Trans.), London: Heinemann.

_____ (1987). *The Philosophical Discourse of Modernity* (F. Lawrence, Trans.), Cambridge Massachusetts: MIT Press.

_____ (1996). *Between Facts and Norms: Contributions to a Discourse Theory of Law and Democracy* (William Rehg, Trans.), Cambrige: Polity Press.

Hanser, Matthew (2008). 'The Metaphysics of Harm', *Philosophy and Phenomenological Research* 77 (2): 421-50.

Hare, Ivan (2009a). 'Blasphemy and Incitement to Religious Hatred: Free Speech Dogma and Doctrine', in Ivan Hare and James Weinstein (Eds.), *Extreme Speech and Democracy*, Oxford: Oxford University Press.

_____ (2009b). 'Extreme Speech under International and Regional Human Rights Standards', in Ivan Hare and James Weinstein (Eds.), *Extreme Speech and Democracy*, Oxford: Oxford University Press.

Harman, Gilbert (1999). 'Moral Philosophy Meets Social Psychology: Virtue Ethics and the Fundamental Attribution Error', *Proceedings of the Aristotelian Society* 99: 315-31.

Hart, H. L. A. (1948). 'The Ascription of Responsibility and Rights', *Proceedings of the Aristotelian Society* 49: 171-94.

_____ (1959). 'Prolegomenon to the Principles of Punishment', *Proceedings of the Aristotelian Society* 60: 1-26.

_____ (1963). *Law, Liberty, and Morality*. Oxford: Oxford University Press

_____ (2008a). 'Acts of Will and Responsibility', *Punishment and Responsibility: Essays in the Philosophy of Law* (second edition), Oxford: Oxford University Press.

_____ (2008b). 'Negligence, *Mens Rea*, and Criminal Responsibility', *Punishment and Responsibility: Essays in the Philosophy of Law* (second edition), Oxford: Oxford University Press.

_____ (2008c). 'Postscript: Responsibility and Retribution', *Punishment and Responsibility: Essays in the Philosophy of Law* (second edition), Oxford: Oxford University Press.

Hart, H. L. A., and Tony Honore (1985). *Causation in the Law* (second edition), Oxford: Clarendon Press.

Haslanger, Sally (2000). 'Gender and Race: (What) Are They? (What) Do We Want Them to Be?', *Nous* 34 (1): 31-55.

_____ (2003). 'Future Genders? Future Races?', *Philosophic Exchange* 34 (1): 5-27.

_____ (2004). 'Oppressions: Racial and Other', in Michael P. Levine and Tamas Pataki (Eds.), *Racism in Mind*, Ithaca: Cornell University Press.

_____ (2005). 'What Are We Talking About? The Semantics and Politics of Social Kinds', *Hypatia* 20 (4): 10-26.

Haworth, Alan (2007). 'On Mill, Infallibility, and Freedom of Expression', *Res Publica* 13 (1): 77-100.

Heins, Marjorie (1983). 'Banning Words: A Comment on "Words That Wound"', *Harvard Civil Rights-Civil Liberties Law Review* 18 (2): 585-92.

Heinze, Eric (2006). 'Viewpoint Absolutism and Hate Speech', *The Modern Law Review* 69 (4): 543-82.

_____ (2009a). 'Cumulative Jurisprudence and Hate Speech: Sexual Orientation and Analogies to Disability, Age, and Obesity', in Ivan Hare and James Weinstein (Eds.), *Extreme Speech and Democracy*, Oxford: Oxford University Press.

_____ (2009b). 'Wild-West Cowboys Versus Cheese-Eating Surrender Monkeys: Some Problems in Comparative Approaches to Hate Speech', in Ivan Hare and James Weinstein (Eds.), *Extreme Speech and Democracy*, Oxford: Oxford University Press.

Hentoff, Nat (1992). *Free Speech for Me but Not for Thee: How the American Left and Right Relentlessly Censor Each Other*, New York: HarperCollins.

Heyman, Steven J. (2008). *Free Speech and Human Dignity*, New Haven: Yale University Press.

_____ (2009). 'Hate Speech, Public Discourse, and the First Amendment', in Ivan Hare and James Weinstein (Eds.), *Extreme Speech and Democracy*, Oxford: Oxford University Press.

Hieronymi, Pamela (2008). 'Responsibility for Believing', *Synthese* 161 (3): 357-73.

Hom, Christopher (2008). 'The Semantics of Racial Epithets', *The Journal of Philosophy* 105 (8): 416-40.

Honderich, Ted (1982). '"On Liberty" and Morality-Dependent Harms', *Political Studies* 30 (4): 504-14.

Honneth, Axel (1995). *The Struggle for Recognition: The Moral Grammar of Social Conflicts* (Joel Anderson, Trans.), Cambridge: Polity Press.

Honoré, Tony (1995). 'Necessary and Sufficient Conditions in Tort Law', in David G. Owen (Ed.), *Philosophical Foundations of Tort Law*, Oxford: Clarendon Press.

_____ (1998). 'Being Responsible and Being a Victim of Circumstance', *Proceedings of the British Academy* 97: 169-87.

_____ (1999a). 'Introduction', *Responsibility and Fault*, Oxford: Hart Publishing.

_____ (1999b). 'Responsibility and Luck: The Moral Basis of Strict Liability', *Responsibility and Fault*, Oxford: Hart Publishing.

Hornsby, Jennifer (1994). 'Illocution and Its Significance', in Savas L. Tsohatzidis (Ed.), *Foundations of Speech Act Theory: Philosophical and Linguistic Perspectives*, London: Routledge.

_____ (1995a). 'Disempowered Speech', *Philosophical Topics* 23 (2): 127-48.

_____ (1995b). 'Speech Acts and Pornography', in Susan Dwyer (Ed.), *The Problem of Pornography*, Belmont California: Wadsworth.

_____ (2001). 'Meaning and Uselessness: How to Think About Derogatory Words', *Midwest Studies in Philosophy* 25: 128-41.

Hornsby, Jennifer, and Rae Langton (1998). 'Free Speech and Illocution', *Legal Theory* 4: 21-37.

Hughes, Graham (1966). 'Prohibiting Incitement to Racial Hatred', *The University of Toronto Law Journal* 16 (2): 361-68.

Hyneman, Charles S. (1962). 'Free Speech: At What Price?', *The American Political Science Review* 61 (4): 847-52.

Ingber, Stanley (1984). 'The Marketplace of Ideas: A Legitimizing Myth', *Duke Law Journal* 1984 (1): 1-91.

Jacobson, Daniel (1995). 'Freedom of Speech Acts?', *Philosophy & Public Affairs* 24 (1): 64-79.

_____ (2000). 'Mill on Liberty, Speech, and the Free Society', *Philosophy & Public Affairs* 29 (3): 276-309.

_____ (2001). 'Speech and Action: Reply to Hornsby and Langton', *Legal Theory* 7: 179-201.

Jones, Karen (2002). 'The Politics of Credibility', in Louise M. Antony and Charlotte E. Witt (Eds.), *A Mind of One's Own: Feminist Essays on Reason and Objectivity* (second edition), Boulder Colorado: Westview Press.

Jones, Peter (2011). 'Religious Belief and Freedom of Expression: Is Offensiveness Really the Issue?', *Res Publica* 17 (1): 75-90.

Jones, Thomas David (1980). 'Article 4 of the International Convention on the Elimination of All Forms of Racial Discrimination and the First Amendment', *Howard Law Journal* 23 (3): 429-80.

_____ (1998). *Human Rights: Group Defamation, Freedom of Expression, and the Law of Nations*, The Hague: Kluwer Law International.

Kagan, Elena (1992). 'The Changing Face of First Amendment Neutrality: R.A.V. v. St. Paul, Rust v. Sullivan, and the Problem of Content-Based Underinclusion', *The Supreme Court Review* 1992: 29-77.

Kagan, Shelly (1986). 'Causation, Liability, and Internalism', *Philosophy & Public Affairs* 15 (1): 41-59.

Kamm, Oliver (2007). 'New Labour: The Tyranny of Moderation', *Index on Censorship* 36 (2): 82-86.

Karst, Kenneth L. (1975). 'Equality as a Central Principle in the First Amendment', *The University of Chicago Law Review* 43 (1): 20-68.

Kelly, Daniel, and Erica Roedder (2008). 'Racial Cognition and the Ethics of Implicit Bias', *Philosophy Compass* 3 (3): 522-40.

Kelsen, Hans (1945). *A General Theory of the Law and State* (A. Wedberg, Trans.), Cambridge Massachusetts: Harvard University Press.

Kendall, Willmore (1960). 'The "Open Society" and Its Fallacies', *The American Political Science Review* 54 (4): 972-79.

Kenyon, Andrew T. (2010). 'What Conversation? Free Speech and Defamation Law', *The Modern Law Review* 73 (5): 697-720.

Kinney, Terry A. (1994). 'An Inductively Derived Typology of Verbal Aggression and Its Association to Distress', *Human Communication Research* 21 (2): 183-222.

Klausen, Jytte (2009). *The Cartoons That Shook the World*, New Haven: Yale University Press.

Knechtle, John C. (2005). 'When to Regulate Hate Speech', *Penn State Law Review* 110 (3): 539-78.

Koppelman, Andrew (2005). 'Does Obscenity Cause Moral Harm?', *Columbia Law Review* 105 (5): 1635-79.

Kretzmer, David (1987). 'Freedom of Speech and Racism', *Cardozo Law Review* 8 (3): 445-513.

Kristol, Irving (1994). 'Pornography, Obscenity, and the Case for Censorship', in Gerald Dworkin (Ed.), *Morality, Harm, and the Law*, Boulder Colorado: Westview Press.

Krotoszynski, Ronald J. (2004). 'A Comparative Perspective on the First Amendment: Free Speech, Militant Democracy, and the Primacy of Dignity as a Preferred Constitutional Value in Germany', *Tulane Law Review* 78 (5): 1549-609.

Kunzru, Hari (2005). 'Respecting Authority, Taking Offence', in Lisa Appignanesi (Ed.), *Free Expression Is No Offence*, London: Penguin.

Kutz, Christopher (2000). *Complicity: Ethics and Law for a Collective Age*, Cambridge: Cambridge University Press.

Kymlicka, Will (1995). *Multicultural Citizenship: A Liberal Theory of Minority Rights*, Oxford: Clarendon Press.

_____ (1996). 'Social Unity in a Liberal State', *Social Philosophy and Policy* 13 (1): 105-36.

_____ (2001). *Politics in the Vernacular: Nationalism, Multiculturalism, and Citizenship*, Oxford: Oxford University Press.

_____ (2002a). 'Civil Society and Government: A Liberal-Egalitarian Perspective', in Nancy L. Rosenblum and Robert C. Post (Eds.), *Civil Society and Government*, Princeton: Princeton University Press.

_____ (2002b). *Contemporary Political Theory: An Introduction* (second edition), Oxford: Oxford University Press.

Kymlicka, Will, and Wayne Norman (1994). 'The Return of the Citizen: A Survey of Recent Work on Citizenship Theory', *Ethics* 104 (2): 352-81.

La Caze, Marguerite (2004). 'If You Say So: Feminist Philosophy and Antiracism', in Michael P. Levine and Tamas Pataki (Eds.), *Racism in Mind*, Ithaca: Cornell University Press.

Langton, Rae (1990). 'Whose Right? Ronald Dworkin, Women, and Pornographers', *Philosophy & Public Affairs* 19 (4): 311-59.

_____ (1993). 'Speech Acts and Unspeakable Acts', *Philosophy & Public Affairs* 22 (4): 293-330.

_____ (1998). 'Subordination, Silence, and Pornography's Authority', in Robert C. Post (Ed.), *Censorship and Silencing: Practices of Cultural Regulation*, Los Angeles: Getty Research Institute for the History of Art and the Humanities.

_____ (2009a). 'Dangerous Confusion? Response to Ronald Dworkin', *Sexual Solipsism: Philosophical Essays on Pornography and Objectification*, Oxford: Oxford University Press.

_____ (2009b). 'Feminism in Epistemology: Exclusion and Objectification', *Sexual Solipsism: Philosophical Essays on Pornography and Objectification*, Oxford: Oxford University Press.

_____ (2009c). 'Speaker's Freedom and Maker's Knowledge', *Sexual Solipsism: Philosophical Essays on Pornography and Objectification*, Oxford: Oxford University Press.

_____ (2012). 'Beyond Belief: Pragmatics in Hate Speech and Pornography', in Ishani Maitra and Mary Kate McGowan (Eds.), *Speech and Harm: Controversies over Free Speech*, Oxford: Oxford University Press.

Langton, Rae, and Caroline West (1999). 'Scorekeeping in a Pornographic Language Game', *Australasian Journal of Philosophy* 77 (3): 303-19.

Lasson, Kenneth (1984). 'Group Libel Versus Free Speech: When Big Brother *Should* Butt In', *Duquesne Law Review* 23 (1): 77-130.

_____ (1985). 'Racial Defamation as Free Speech: Abusing the First Amendment', *Columbia Human Rights Law Review* 17 (1): 11-55.

_____ (1997). 'Holocaust Denial and the First Amendment: The Quest for Truth in a Free Society', *George Mason Law Review* 6 (1): 35-86.

Lawrence, Charles R. (1993). 'If He Hollers Let Him Go: Regulating Racist Speech on Campus', *Words That Wound: Critical Race Theory, Assaultive Speech, and the First Amendment*, Boulder Colorado: Westview Press.

_____ (1995). 'Cross-Burning and the Sound of Silence: Anti-Subordination Theory and the First Amendment', in Laura J. Lederer and Richard Delgado (Eds.), *The Price We Pay: The Case against Racist Speech, Hate Propaganda, and Pornography*, New York: Hill and Wang.

Lawrence, Charles R., Mari J. Matsuda, Richard Delgado, and Kimberle Williams Crenshaw (1993). 'Introduction', in *Words That Wound: Critical Race Theory, Assaultive Speech, and the First Amendment*, Boulder Colorado: Westview Press.

Leader, Tirza, Brian Mullen, and Diana R. Rice (2009). 'Complexity and Valence in Ethnophaulisms and Exclusion of Ethnic out-Groups: What Puts the "Hate" into Hate Speech?', *Journal of Personality and Social Psychology* 96 (1): 170-82.

Lederer, Laura J. (1995). 'The Case of Cross-Burning: An Interview with Russ and Laura Jones', in Laura J. Lederer and Richard Delgado (Eds.), *The Price We Pay: The Case against Racist Speech, Hate Propaganda, and Pornography*, New York: Hill and Wang.

Lederer, Laura J., and Richard Delgado (1995). 'Introduction', in Laura J. Lederer and Richard Delgado (Eds.), *The Price We Pay: The Case against Racist Speech, Hate Propaganda, and Pornography*, New York: Hill and Wang.

Leets, Laura (2001). 'Responses to Internet Hate Sites: Is Speech Too Free in Cyberspace?', *Communication Law and Policy* 6: 287-317.

_____ (2002). 'Experiencing Hate Speech: Perceptions and Responses to Anti-Semitism and Antigay Speech', *Journal of Social Issues* 58 (2): 341-61.

Leets, Laura, and Peggy J. Bowers (1999). 'Loud and Angry Voices: The Insidious Influence', *Communication Monographs* 66 (4): 325-40.

Leigh, Ian (2009). 'Homophobic Speech, Equality Denial, and Religious Expression', in Ivan Hare and James Weinstein (Eds.), *Extreme Speech and Democracy*, Oxford: Oxford University Press.

_____ (2011). 'Damned If They Do, Damned If They Don't: The European Court of Human Rights and the Protection of Religion from Attack', *Res Publica* 17 (1): 55-73.

Leon, Mark (2002). 'Responsible Believers', *The Monist* 85 (3): 421-35.

Levinson, Sanford (1998). 'The Tutelary State: "Censorship", "Silencing", and the "Practices of Cultural Regulation"', in Robert C. Post (Ed.), *Censorship and Silencing: Practices of Cultural Regulation*, Los Angeles: Getty Research Institute for the History of Art and the Humanities.

Lewis, David (1975). 'Languages and Language', in Keith Gunderson (Ed.), *Language, Mind, and Knowledge*, Minneapolis: University of Minnesota Press.

_____ (1979). 'Scorekeeping in a Language Game', *Journal of Philosophical Logic* 8: 339-59.

_____ (1996). 'Elusive Knowledge', *Australasian Journal of Philosophy* 74 (4): 540-67.

Lidsky, Lyrissa Barnett (2002). 'Brandenburg and the United States' War on Incitement Abroad: Defending a Double Standard', *Wake Forest Law Review* 37: 1009-33.

Lively, Donald E. (1994). 'Racial Myopia in the Age of Digital Compression', *Speaking of Race, Speaking of Sex: Hate Speech, Civil Rights, and Civil Liberties*, New York: New York University Press.

Locke, John (1946). 'A Letter Concerning Toleration', in J. W. Gough (Ed.), *The Second Treatise of Civil Government and a Letter Concerning Toleration*, Oxford: Basil Blackwell (originally published 1689).

_____ (1946). 'The Second Treatise of Civil Government', in J. W. Gough (Ed.), *The Second Treatise of Civil Government and a Letter Concerning Toleration*, Oxford: Basil Blackwell (originally published 1690).

Love, Jean C. (1990). 'Discriminatory Speech and the Tort of Intentional Infliction of Emotional Distress', *Washington and Lee Law Review* 47: 123-59.

Lyons, David (1970). 'The Correlativity of Rights and Duties', *Nous* 4 (1): 45-55.

MacIntyre, Alasdair (1981). *After Virtue: A Study in Moral Theory* (third edition), London: Duckworth.

Mackie, J. L. (1974). *The Cement of the Universe: A Study of Causation*, Oxford: Clarendon Press.

_____ (1977). 'The Grounds of Responsibility', in P. M. S. Hacker and Joseph Raz (Eds.), *Law, Morality, and Society: Essays in Honour of H. L. A. Hart*, Oxford: Clarendon Press.

MacKinnon, Catharine A. (1979). *Sexual Harassment of Working Women: A Case of Sex Discrimination*, New Haven: Yale University Press.

_____ (1984). 'Not a Moral Issue', *Yale Law & Policy Review* 2 (2): 321-45.

_____ (1987a). 'Francis Biddle's Sister: Pornography, Civil Rights, and Speech', *Feminism Unmodified: Discourses on Life and Law*, Cambridge Massachusetts: Harvard University Press.

_____ (1987b). 'On Collaboration', *Feminism Unmodified: Discourses on Life and Law*, Cambridge Massachusetts: Harvard University Press.

_____ (1994). *Only Words*, London: HarperCollins.

Magnet, Joseph (1994). 'Hate Propaganda in Canada', in W. J. Waluchow (Ed.), *Free Expression: Essays in Law and Philosophy*, Oxford: Clarendon Press.

Mahoney, Kathleen (1995). 'Recognizing the Constitutional Significance of Harmful Speech: The Canadian View of Pornography and Hate Propaganda', in Laura J. Lederer and Richard Delgado (Eds.), *The Price We Pay: The Case against Racist Speech, Hate Propaganda, and Pornography*, New York: Hill and Wang.

_____ (1996). 'Hate Speech: Affirmation or Contradiction of Freedom of Expression?', *University of Illinois Law Review* (3): 789-808.

_____ (2009). 'Hate Speech, Equality, and the State of Canadian Law', *Wake Forest Law Review* 44 (2): 321-51.

Maitra, Ishani (2009). 'Silencing Speech', *Canadian Journal of Philosophy* 39 (2): 309-38.

_____ (2012). 'Subordinating Speech', in Ishani Maitra and Mary Kate McGowan (Eds.), *Speech and Harm: Controversies over Free Speech*, Oxford: Oxford University Press.

Maitra, Ishani, and Mary Kate McGowan (2007). 'The Limits of Free Speech: Pornography and the Question of Coverage', *Legal Theory* 13 (1): 41-68.

_____ (2010). 'On Racist Hate Speeeh and the Sccope of a Free Speech Principle', *Canadian Journal of Law and Jurisprudence* 23 (2): 343-72.

Malik, Maleiha (2011). 'Religious Freedom, Free Speech and Equality: Conflict or Cohesion?', *Res Publica* 17 (1): 21-40.

Malvern, Linda (2004). *Conspiracy to Murder: The Rwandan Genocide*, London: Verso.

Mamdani, Mahamood (2004). *When Victims Become Killers: Colonialism, Nativism, and the Genocide in Rwanda*, Princeton: Princeton University Press.

Manwaring, John (1992). 'Legal Regulation of Hate Propaganda in Canada', in Sandra Coliver (Ed.), *Striking a Balance: Hate Speech, Freedom of Expression, and Non-Discrimination*, London: Article 19, London and Human Rights Centre.

Marcus, George E. (1998). 'Censorship in the Heart of Difference: Cultural Property, Indigenous Peoples' Movements, and Challenges to Western Liberal Thought', in Robert C. Post (Ed.), *Censorship and Silencing: Practices of Cultural Regulation*, Los Angeles: Getty Research Institute for the History of Art and the Humanities.

Massaro, Toni M. (1990). 'Equality and Freedom of Expression: The Hate Speech Dilemma', *William and Mary Law Review* 32 (2): 211-65.

Massey, Calvin R. (1992). 'Hate Speech, Cultural Diversity, and the Foundational Paradigms of Free Expression', *UCLA Law Review* 40 (1): 103-98.

Matsuda, Mari J. (1993). 'Public Response to Racist Speech: Considering the Victim's Story', in *Words That Wound: Critical Race Theory, Assaultive Speech, and the First Amendment*, Boulder Colorado.

Matsuda, Mari J., and Charles R. Lawrence (1993). 'Epilogue: Burning Crosses and the R.A.V. Case', in *Words That Wound: Critical Race Theory, Assaultive Speech, and the First Amendment*, Boulder Colorado: Westview Press.

Mbongo, Pascal (2009). 'Hate Speech, Extreme Speech, and Collective Defamation in French Law', in Ivan Hare and James Weinstein (Eds.), *Extreme Speech and Democracy*, Oxford: Oxford University Press.

McCloskey, H. J. (1970). 'Liberty of Expression: Its Grounds and Limits (I)', *Inquiry* 13 (1): 219-37.

McGowan, Mary Kate (2003). 'Conversational Exercitives and the Force of Pornography', *Philosophy & Public Affairs* 31 (2): 155-89.

_____ (2004). 'Conversational Exercitives: Something Else We Do with Our Words', *Linguistics and Philosophy* 27 (1): 93-111.

_____ (2005). 'On Pornography: MacKinnon, Speech Acts, and 'False' Construction', *Hypatia* 20 (3): 22-49.

_____ (2009). 'Oppressive Speech', *Australasian Journal of Philosophy* 87 (3): 389-407.

_____ (2012). 'On 'Whites Only' Signs and Racist Hate Speech: Verbal Acts of Racial Discrimination', in Ishani Maitra and Mary Kate McGowan (Eds.), *Speech and Harm: Controversies over Free Speech*, Oxford: Oxford University Press.

McNamara, Lawrence (2007). 'Salvation and the State: Religious Vilification Laws and Religious Speech', in Katharine Gelber and Adrienne Stone (Eds.), *Hate Speech and Freedom of Speech in Australia*, Sydney: Federation Press.

McNaron, David L. (2008). 'Meeting Objections to the Danish Cartoons of Muhammed', *Florida Philosophical Review* 8 (1): 51-80.

Meiklejohn, Alexander (1948). *Free Speech and Its Relation to Self-Government*, New York: Harper.

_____ (1961). 'The First Amendment Is an Absolute', *The Supreme Court Review* 1961: 245-66.

Michelman, Frank I. (1992). 'Liberties, Fair Values, and the Constitutional Method', *The University of Chicago Law Review* 59 (1): 91-114.

_____ (1995). 'Civil Liberties, Silencing, and Subordination', in Laura J. Lederer and Richard Delgado (Eds.), *The Price We Pay: The Case against Racist Speech, Hate Propaganda, and Pornography*, New York: Hill and Wang.

Miles, J. K. (2011). 'Hatred, Hostility, and Defamation: The United Nations' Exceptions for Free Speech', *International Journal of Applied Philosophy* 25 (1): 25-32.

Mill, John Stuart (1985). *On Liberty*, London: Penguin Books (originally published 1859).

Milton, John (1999). 'Areopagitica', *Areopagitica, and Other Political Writings of John Milton*, Indianapolis: Liberty Fund (originally published 1644).

Minow, Martha (2000). 'Regulating Hatred: Whose Speech, Whose Crimes, Whose Power? An Essay for Kenneth Karst', *UCLA Law Review* 47 (5): 1253-77.

Modood, Tariq (1993). 'Kymlicka on British Muslims: A Rejoinder', *Analyse & Kritik* 15: 97-99.

Molnar, Peter (2009). 'Towards Improved Law and Policy on 'Hate Speech' – the 'Clear and Present Danger' Test in Hungary', in Ivan Hare and James Weinstein (Eds.), *Extreme Speech and Democracy*, Oxford: Oxford University Press.

Mookherjee, Monica (2007). 'Permitting Dishonour: Culture, Gender, and Freedom of Expression', *Res Publica* 13 (1): 29-52.

Moore, Michael S. (2009). *Causation and Responsibility: An Essay in Law, Morals, and Metaphysics*, Oxford: Oxford University Press.

Mullen, Brian (2004). 'Sticks and Stones Can Break My Bones, but Ethnophaulisms Can Alter the Portrayal of Immigrants to Children', *Personality and Social Pyschology Bulletin* 30 (2): 250-60.

Mullen, Brian, and Craig Johnson (1995). 'Cognitive Representation in Ethnophaulisms and Illusory Correlation in Stereotyping', *Personality and Social Pyschology Bulletin* 21 (5): 420-33.

Mullen, Brian, and Diana R. Rice (2003). 'Ethnophaulisms and Exclusion: The Behavioural Consequences of Cognitive Representation of Ethnic Immigrant Groups', *Personality and Social Pyschology Bulletin* 29 (8): 1056-67.

Mullen, Brian, and Joshua M. Smyth (2004). 'Immigrant Suicide Rates as a Function of Ethnophaulisms: Hate Speech Predicts Death', *Psychosomatic Medicine* 66 (3): 343-48.

Mullender, Richard (2007). 'Hate Speech and Pornography in Canada: A Qualified Deontological Response to a Consequentialist Argument', *Canadian Journal of Law and Jurisprudence* 20 (1): 241-55.

Nagel, Thomas (1979). 'Equality', in *Mortal Questions*, Cambridge: Cambridge University Press.

_____ (1991). *Equality and Partiality*, New York: Oxford University Press.

_____ (1999). 'Rights and Public Space', in Harold Hongju Koh and Ronald Slye (Eds.), *Deliberative Democracy and Human Rights*, New Haven: Yale University Press.

Nalezinski, Alix (1996). 'Acceptable and Unacceptable Levels of Risk: The Case of Pornography', *Social Theory and Practice* 22 (1): 83-104.

Neuborne, Burt (1992). 'Ghosts in the Attic: Idealized Pluralism, Community and Hate Speech', *Harvard Civil Rights-Civil Liberties Law Review* 27 (2): 371-406.

Newman, Christopher J., and Peter Rackow (2011). 'Undesirable Posters and Dubious Symbols: Anglo-German Legal Solutions to the Display of Right-Wing Symbolism and Propaganda', *Journal of Criminal Law* 75 (2): 142-55.

Nielsen, Laura Beth (2002). 'Subtle, Pervasive, and Harmful: Racist and Sexist Remarks in Public as Hate Speech', *Journal of Social Issues* 58 (2): 265-80.

_____ (2004). *License to Harass: Law, Hierarchy, and Offensive Public Speech*, Princeton: Princeton University Press.

_____ (2012). 'Power in Public: Reactions, Responses, and Resistance to Offensive Public Speech', in Ishani Maitra and Mary Kate McGowan (Eds.), *Speech and Harm: Controversies over Free Speech*, Oxford: Oxford University Press.

Nockleby, John T. (1994). 'Hate Speech in Context: The Case of Verbal Threats', *Buffalo Law Review* 42 (3): 653-713.

Nozick, Robert (1974). *Anarchy, State, and Utopia*, New York: Basic Books.

_____ (1981). *Philosophical Explanations*, Oxford: Clarendon Press.

Nussbaum, Martha C. (1995). 'Objectification', *Philosophy & Public Affairs* 24 (4): 249-91.

_____ (2000a). 'Aristotle, Politics, and Human Capabilities: A Response to Antony, Arneson, Charlesworth, and Mulgan', *Ethics* 111 (1): 102-40.

_____ (2000b). 'Women's Capabilities and Social Justice', *Journal of Human Development* 1 (2): 219-47.

_____ (2003). 'Capabilities as Fundamental Entitlements: Sen and Social Justice', *Feminist Economics* 9 (2): 33-59.

O'Neill, Daniel I. (1999). 'Multicultural Liberals and the Rushdie Affair: A Critique of Kymlicka, Taylor, and Walzer', *The Review of Politics* 61 (2): 219-50.

Ogletree, Charles J., Jr (1996). 'The Limits of Hate Speech: Does Race Matter?', *Gonzaga Law Review* 32 (3): 491-510.

Oshana, Marina A. L. (1997). 'Ascriptions of Responsibility', *American Philosophical Quarterly* 34 (1): 71-83.

Oyediran, Joanna (1992). 'The United Kingdom's Compliance with Article 4 of the International Convention on the Elimination of All Forms of Racial Discrimination', in Sandra Coliver (Ed.), *Striking a Balance: Hate Speech, Freedom of Expression, and Non-Discrimination*, London: Article 19, London and Human Rights Centre.

Parekh, Bhikhu. (1990a). *The Rushdie Affair and the British Press: Some Salutary Lessons*. London: Commission for Racial Equality.

_____ (1990b). 'The Rushdie Affair: Research Agenda for Political Philosophy', *Political Studies* 38 (4): 695-709.

_____ (2006). *Rethinking Multiculturalism: Cultural Diversity and Political Theory*, New York: Palgrave MacMillan.

Parent, William (1990). 'A Second Look at Pornography and the Subordination of Women', *Journal of Philosophy* 87 (4): 205-11.

Parfit, Derek (1986). *Reasons and Persons*, Oxford: Oxford University Press.

Peard, Thomas (2004). 'Regulating Racist Speech on Campus', in Christine T. Sistare (Ed.), *Civility and Its Discontents: Essays on Civic Virtue, Toleration, and Cultural Fragmentation*, Kansas: University of Kansas Press.

Peonidis, Filimon (1998). 'Freedom of Expression, Autonomy, and Defamation', *Law and Philosophy* 17 (1): 1-17.

Perry, Michael J. (1983). 'Freedom of Expression: An Essay on Theory and Doctrine', *Northwestern University Law Review* 78 (5): 1137-211.

Peters, John Durham (2005). *Courting the Abyss: Free Speech and the Liberal Tradition*, Chicago: University of Chicago Press.

Pettit, Philip (2001). 'The Capacity to Have Done Otherwise: An Agent-Centred View', in Peter Cane and John Gardner (Eds.), *Relating to Responsibility: Essays for Tony Honoré on His Eightieth Birthday*, Oxford: Hart Publishing.

Piaget, Jean (1954). *The Construction of Reality*, New York: Basic Books.

Pildes, Richard H. (1998). 'Why Rights Are Not Trumps: Social Meanings, Expressive Harms, and Constitutionalism', *Journal of Legal Studies* 27 (2): 725-63.

Posner, Richard A. (1986). 'Free Speech in an Economic Perspective', *Suffolk University Law Review* 20 (1): 1-54.

Post, Robert C. (1990a). 'The Constitutional Conception of Public Discourse: Outrageous Opinion, Democratic Deliberation, and Hustler Magazine v. Falwell', *Harvard Law Review* 103 (3): 601-86.

_____ (1990b). 'Racist Speech, Democracy, and the First Amendment', *William and Mary Law Review* 32 (2): 267-327.

_____ (1993). 'Managing Deliberation: The Quandary of Democratic Dialogue', *Ethics* 103 (4): 654-78.

_____ (1995). 'Reprise: The Racist Speech Problem', *Constitutional Domains: Democracy, Community, Management*, Cambridge Massachusetts: Harvard University Press.

_____ (1996). 'Subsidized Speech', *The Yale Law Journal* 106 (1): 151-95.

_____ (1998). 'Censorship and Silencing', in Robert C. Post (Ed.), *Censorship and Silencing: Practices of Cultural Regulation*, Los Angeles: Getty Research Institute for the History of Art and the Humanities.

_____ (2007). 'Religion and Freedom of Speech: Portraits of Muhammad', *Constellations* 14 (1): 72-90.

_____ (2009). 'Hate Speech', in Ivan Hare and James Weinstein (Eds.), *Extreme Speech and Democracy*, Oxford: Oxford University Press.

Poynting, Scott (2006). 'What Caused the Cronulla Riot?', *Race & Class* 48 (1): 85-92.

Price, Terry L. (2006). 'Feinberg's Offense Principles and the Danish Cartoons of Muhammad', *APA Newsletter* 6 (1): 6-12.

Prunier, Gérard (1995). *The Rwanda Crisis: History of a Genocide*, New York: Columbia University Press.

Pullman, Philip (2005). 'Against 'Identity'', in Lisa Appignanesi (Ed.), *Free Expression Is No Offence*, London: Penguin.

Quong, Jonathan (2004). 'The Rights of Unreasonable Citizens', *The Journal of Political Philosophy* 12 (3): 314-35.

Rawls, John (1971). *A Theory of Justice*, Cambridge Massachusetts: Belknap Press.

_____ (1993). *Political Liberalism*, New York: Columbia University Press.

_____ (2001). *Justice as Fairness: A Restatement*, Cambridge Massachusetts: Harvard University Press.

Raz, Joseph (1986). *The Morality of Freedom*, Oxford: Oxford University Press.

_____ (1994a). 'Free Expression and Personal Identification', *Ethics in the Public Domain*, Oxford: Oxford University Press.

_____ (1994b). 'Multiculturalism: A Liberal Perspective', *Ethics in the Public Domain*, Oxford: Oxford University Press.

Redish, Martin H. (1982). 'The Value of Free Speech', *University of Pennsylvania Law Review* 130 (3): 591-645.

Reichman, Amnon (2009). 'Criminalizing Religiously Offensive Satire: Free Speech, Human Dignity, and Comparitive Law', in Ivan Hare and James Weinstein (Eds.), *Extreme Speech and Democracy*, Oxford: Oxford University Press.

Richard, Mark (2008). *When Truth Gives Out*, Oxford: Oxford University Press.

Richards, David A. J. (1974). 'Free Speech and Obscenity Law: Toward a Moral Theory of the First Amendment', *University of Pennsylvania Law Review* 123 (1): 45-91.

_____ (1986). *Toleration and the Constitution*, New York: Oxford University Press.

_____ (1988). 'Toleration and Free Speech', *Philosophy & Public Affairs* 17 (4): 323-36.

_____ (1994a). 'Free Speech as Toleration', in W. J. Waluchow (Ed.), *Free Expression: Essays in Law and Philosophy*, Oxford: Clarendon Press.

_____ (1994b). 'Liberalism, Free Speech, and Justice for Minorities', in Jules L. Coleman and Allen Buchanan (Eds.), *In Harm's Way: Essays in Honor of Joel Feinberg*, Cambridge: Cambridge University Press.

_____ (1999). *Free Speech and the Politics of Identity*, Oxford: Oxford University Press.

Ripstein, Arthur (1999). *Equality, Responsibility, and the Law*, Cambridge: Cambridge University Press.

_____ (2006). 'Beyond the Harm Principle', *Philosophy & Public Affairs* 34 (3): 215-45.

Rockefeller, Steven C. (1992). 'Comment', in Amy Gutmann (Ed.), *Multiculturalism and "the Politics of Recognition"*, Princeton: Princeton University Press.

Rodgerson, Gillian, and Elizabeth Wilson (Eds.) (1991). *Pornography and Feminism: The Case against Censorship*. London: Lawrence & Wishart.

Rorty, Richard (1989). *Contingency, Irony, and Solidarity*. Cambridge: Cambridge University Press.

Rosenfeld, Michel (2002). 'Hate Speech in Constitutional Jurisprudence: A Comparitive Analysis', *Cardozo Law Review* 24 (4): 1523-67.

Rowbottom, Jacob (2009). 'Extreme Speech and the Democratic Functions of the Mass Media', in Ivan Hare and James Weinstein (Eds.), *Extreme Speech and Democracy*, Oxford: Oxford University Press.

Royalle, Candida (2000). 'Porn in the USA', in Drucilla Cornell (Ed.), *Feminism and Pornography*, Oxford: Oxford University Press.

Runge, Carlisle Ford (1984). 'Institutions and the Free Rider: The Assurance Problem in Collective Action', *The Journal of Politics* 46 (1): 154-81.

Sadock, Jerrold M. (1994). 'Toward a Gramatically Realistic Typology of Speech Acts', in Savas L. Tsohatzidis (Ed.), *Foundations of Speech Act Theory: Philosophical and Linguistic Perspectives*, London: Routledge.

Sadurski, Wojciech (1996). 'On 'Seeing Speech through an Equality Lens': A Critique of Egalitarian Arguments for Suppression of Hate Speech and Pornography', *Oxford Journal of Legal Studies* 16 (4): 713-23.

_____ (1999). *Freedom of Speech and Its Limits*, Boston: Kluwer.

Said, Edward W. (1978). *Orientalism*, London: Routledge.

Sandel, Michael J. (1982). *Liberalism and the Limits of Justice*, Cambridge: Cambridge University Press.

Saul, Jennifer (2002). 'Speaker Meaning, What Is Said, and What Is Implicated', *Nous* 36 (2): 228-48.

_____ (2006a). 'Gender and Race', *Aristotelian Society Supplementary Volume* 80 (1): 119-43.

_____ (2006b). 'Pornography, Speech Acts and Context', *Proceedings of the Aristotelian Society* 106 (1): 229-48.

_____ (2011). 'Maker's Knowledge or Perpetuator's Ignorance', *Jurisprudence* 2 (2): 403-08.

_____ (2012). 'Politically Significant Terms and Philosophy of Language', in Sharon L. Crasnow and Anita M. Superson (Eds.), *Out from the Shadows: Analytical Feminist Contributions to Traditional Philosophy*, Oxford: Oxford University Press.

Savage, Sara, and Jose Liht (2009). 'Radical Religious Speech: The Ingredients of a Binary World View', in Ivan Hare and James Weinstein (Eds.), *Extreme Speech and Democracy*, Oxford: Oxford University Press.

Scanlon, Thomas (1972). 'A Theory of Freedom of Expression', *Philosophy & Public Affairs* 1 (2): 204-26.

_____ (1979). 'Freedom of Expression and Categories of Expression', *University of Pittsburgh Law Review* 40: 519-50.

_____ (1982). 'Contractualism and Utilitarianism', in Amartya Sen and Bernard Williams (Eds.), *Utilitarianism and Beyond*, Cambridge: Cambridge University Press.

_____ (2004). 'Adjusting Rights and Balancing Values', *Fordham Law Review* 72 (5): 1477-86.

Schabas, William A. (2000). 'Hate Speech in Rwanda: The Road to Genocide', *McGill Law Journal* 46 (1): 141-72.

Schauer, Frederick (1978). 'Fear, Risk and the First Amendment: Unraveling the "Chilling Effect"', *Boston University Law Review* 58 (5): 685-732.

_____ (1982). *Free Speech: A Philosophical Enquiry*, Cambridge: Cambridge University Press.

_____ (1984). 'Must Speech Be Special?', *Northwestern University Law Review* 78 (5): 1284-306.

_____ (1985). 'Slippery Slopes', *Harvard Law Review* 99 (2): 361-83.

_____ (1987). 'Causation Theory and the Causes of Sexual Violence', *Law & Social Inquiry* 12 (4): 737-70.

_____ (1989). 'Review: Harry Kalven and the Perils of Particularism', *The University of Chicago Law Review* 56 (1): 397-414.

_____ (1992). 'Uncoupling Free Speech', *Columbia Law Review* 92 (6): 1321-57.

_____ (1993). 'The Phenomenology of Speech and Harm', *Ethics* 103 (4): 635-53.

_____ (1998). 'The Ontology of Censorship', in Robert C. Post (Ed.), *Censorship and Silencing: Practices of Cultural Regulation*, Los Angeles: Getty Research Institute for the History of Art and the Humanities.

_____ (2000). 'Speech, Behaviour, and the Interdependence of Fact and Value', in David Kretzmer and Francine Kershman Hazan (Eds.), *Freedom of Speech and Incitement against Democracy*, London: Kluwer Law International.

_____ (2004). 'The Boundaries of the First Amendment: A Preliminary Exploration of Constitutional Salience', *Harvard Law Review* 117 (6): 1765-809.

_____ (2005). 'Towards an Institutional First Amendment', *Minnesota Law Review* 89 (1256-1279).

Schwartzman, Lisa H. (2002). 'Hate Speech, Illocution, and Social Context: A Critique of Judith Butler', *Journal of Social Philosophy* 33 (3): 421-41.

Scoccia, Danny (1987). 'Autonomy, Want Satisfaction, and the Justification of Liberal Freedoms', *Canadian Journal of Philosophy* 17 (3): 583-602.

_____ (1996). 'Can Liberals Support a Ban on Violent Pornography?', *Ethics* 106 (4): 776-99.

Searle, John. R (1969). *Speech Acts: An Essay in the Philosophy of Language*, Cambridge: Cambridge University Press.

Sellers, Mortimer (2004). 'Ideals of Public Discourse', in Christine T. Sistare (Ed.), *Civility and Its Discontents: Essays on Civic Virtue, Toleration, and Cultural Fragmentation*, Kansas: University of Kansas Press.

Sen, Amartya (2006). *Identity and Violence: The Illusion of Destiny*, London: Allen Lane.

Shoemaker, David W. (2000). '"Dirty Words" and the Offense Principle', *Law and Philosophy* 19 (5): 545-84.

Sibley, W. M. (1953). 'The Rational Versus the Reasonable', *Philosophical Review* 62 (4): 554-60.

Simester, A. P., and Andrew von Hirsch (2002). 'Rethinking the Offence Principle', *Legal Theory* 8 (3): 269-95.

Simpson, Evan (2006). 'Responsibilities for Hateful Speech', *Legal Theory* 12 (2): 157-77.

Skipper, Robert (1993). 'Mill and Pornography', *Ethics* 103 (4): 726-30.

Skyrms, Brian (1996). *Evolution of the Social Contract*, Cambridge: Cambridge University Press.

Slaughter, M. M. (1993). 'The Salman Rushdie Affair: Apostasy, Honour, and Freedom of Speech', *Virginia Law Review* 79 (1): 153-204.

Smith, Angela M. (2005). 'Responsibility for Attitudes: Activity and Passivity in Mental Life', *Ethics* 115 (2): 236-71.

Smith, Michael (2001). 'Responsibility and Self-Control', in Peter Cane and John Gardner (Eds.), *Relating to Responsibility: Essays for Tony Honoré on His Eightieth Birthday*, Oxford: Hart Publishing.

Smolla, Rodney (1992). *Free Speech in an Open Society*, New York: Knopf.

Solum, Lawrence (1989). 'Freedom of Communicative Action: A Theory of the First Amendment Freedom of Speech', *Northwestern University Law Review* 83 (1): 54-135.

Stalnaker, Robert (1973). 'Presuppositions', *Journal of Philosophical Logic* 2 (4): 447-57.

_____ (2002). 'Common Ground', *Linguistics and Philosophy* 25 (5): 701-21.

Stapleton, Jane (2001). 'Unpacking "Causation"', in Peter Cane and John Gardner (Eds.), *Relating to Responsibility: Essays for Tony Honoré on His Eightieth Birthday*, Oxford: Hart Publishing.

Steele, Claude M., and Joshua Aronson (1995). 'Stereotype Threat and the Intellectual Test Performance of African Americans', *Journal of Personality and Social Psychology* 69 (5): 797-811.

Stefancic, Jean, and Richard Delgado (1993). 'A Shifting Balance: Freedom of Expression and Hate Speech Regulation', *Iowa Law Review* 78 (3): 737-50.

Stein, Eric (1986). 'History against Free Speech: The New German Law against the "Auschwitz" – and Other – "Lies"', *Michigan Law Review* 85 (2): 277-324.

Stephen, James Fitzjames (1994). 'The Doctrine of Liberty and Its Application to Morals', in Gerald Dworkin (Ed.), *Morality, Harm, and the Law*, Boulder Colorado: Westview Press (originally published as chapter four of *Liberty, Equality, Fraternity*, 1874).

Stevens, John Paul (1993). '*The* Freedom of Speech', *The Yale Law Journal* 102 (6): 1293-313.

Stocker, Michael (1982). 'Responsibility Especially for Beliefs', *Mind* 91 (363): 398-417.

Stone, Adrienne (2007). 'How to Think About the Problem of Hate Speech: Understanding a Comparative Debate', in Katharine Gelber and Adrienne Stone (Eds.), *Hate Speech and Freedom of Speech in Australia*, Sydney: Federation Press.

Stone, Adrienne, and Simon Evans (2006). 'Freedom of Speech and Insult in the High Court of Australia', *International Journal of Constitutional Law* 4 (4): 677-88.

Stone, Geoffrey R. (1987). 'Content-Neutral Restrictions', *The University of Chicago Law Review* 54 (1): 46-118.

Strauss, David A. (1991). 'Persuasion, Autonomy, and Freedom of Expression', *Columbia Law Review* 91: 334-71.

Strawson, P. F. (1962). 'Freedom and Resentment', *Proceedings of the British Academy* 48: 1-25.

_____ (1964). 'Intention and Convention in Speech Acts', *The Philosophical Review* 73 (4): 439-60.

Strossen, Nadine (1990). 'Regulating Hate Speech on Campus: A Modest Proposal?', *Duke Law Journal* 1990 (3): 484-573.

_____ (1993). 'A Feminist Critique of "the" Feminist Critique of Pornography', *Virginia Law Review* 79 (5): 1099-190.

_____ (1995). *Defending Pornography: Free Speech, Sex, and the Fight for Women's Rights*, New York: Scribner.

Sumner, L. W. (1994). 'Hate Propaganda and Charter Rights', in W. J. Waluchow (Ed.), *Free Expression: Essays in Law and Philosophy*, Oxford: Clarendon Press.

_____ (2004). *The Hateful and the Obscene: Studies in the Limits of Free Expression*, Toronto: University of Toronto Press.

_____ (2009). 'Incitement and the Regulation of Hate Speech in Canada: A Philosophical Analysis', in Ivan Hare and James Weinstein (Eds.), *Extreme Speech and Democracy*, Oxford: Oxford University Press.

Sunstein, Cass R. (1986). 'Pornography and the First Amendment', *Duke Law Journal* 1986 (4): 589-627.

_____ (1991). 'Preferences and Politics', *Philosophy & Public Affairs* 20 (1): 3-34.

_____ (1995). 'Words, Conduct, Caste', in Laura J. Lederer and Richard Delgado (Eds.), *The Price We Pay: The Case against Racist Speech, Hate Propaganda, and Pornography*, New York: Hill and Wang.

_____ (1996). 'On the Expressive Function of the Law', *University of Pennsylvania Law Review* 144 (5): 2021-53.

_____ (2000). 'Deliberative Trouble? Why Groups Go to Extremes', *The Yale Law Journal* 110 (1): 71-119.

_____ (2003). *Why Societies Need Dissent*, Cambridge Massachusetts: Harvard University Press.

Talisse, Robert B. (2008). 'Toward a Social Epistemic Comprehensive Liberalism', *Episteme: A Journal of Social Epistemology* 5 (1): 106-28.

Tate, John William (2008). 'Free Speech or Equal Respect? Liberalism's Competing Values', *Philosophy & Social Criticism* 34 (9): 987-1020.

Taylor, Charles (1985a). 'Atomism', *Philosophy and the Human Sciences: Philosophical Papers 2*, Cambridge: Cambridge University Press.

_____ (1985b). 'Language and Human Nature', *Human Agency and Language: Philosophical Papers 1*, Cambridge: Cambridge University Press.

_____ (1985c). 'Theories of Meaning', *Human Agency and Language: Philosophical Papers 1*, Cambridge: Cambridge University Press.

_____ (1989). 'The Rushdie Controversy', *Public Culture* 2 (1): 118-22.

_____ (1992a). 'Heidegger, Language, and Ecology', in Hubert Dreyfus and Harrison Hall (Eds.), *Heidegger: A Critical Reader*, Oxford: Basil Blackwell.

_____ (1992b). *Multiculturalism and "the Politics of Recognition"*, Princeton: Princeton University Press.

Ten, C. L. (1980). *Mill on Liberty*, Oxford: Clarendon Press.

Thomson, Judith Jarvis (1986a). 'Feinberg on Harm, Offense, and the Criminal Law: A Review Essay', *Philosophy & Public Affairs* 15 (4): 381-95.

_____ (1986b). 'A Note on Internalism', *Philosophy & Public Affairs* 15 (1): 60-66.

_____ (1990). *The Realm of Rights*, Cambridge Massachusetts: Harvard University Press.

_____ (2011). 'More on the Metaphysics of Harm', *Philosophy and Phenomenological Research* 82 (2): 436-58.

Tirrell, Lynne (1998). 'Aesthetic Derogation: Hate Speech, Pornography, and Aesthetic Contexts', in Jerrold Levinson (Ed.), *Aesthetics and Ethics: Essays at the Intersection*, Cambridge: Cambridge University Press.

_____ (2012). 'Genocidal Language Games', in Ishani Maitra and Mary Kate McGowan (Eds.), *Speech and Harm: Controversies over Free Speech*, Oxford: Oxford University Press.

Tsesis, Alexander (2000). 'The Empirical Shortcomings of First Amendment Jurisprudence: A Historical Perspective on the Power of Hate Speech', *Santa Clara Law Review* 40 (3): 729-86.

_____ (2002). *Destructive Messages: How Hate Speech Paves the Way for Harmful Social Movements*, New York: New York University Press.

_____ (2004). 'Regulating Intimidating Speech', *Harvard Journal on Legislation* 41 (2): 389-405.

_____ (2009). 'Dignity and Speech: The Regulation of Hate Speech in a Democracy', *Wake Forest Law Review* 44 (2): 497-532.

Turner, Ronald (1995). 'Regulating Hate Speech and the First Amendment: The Attractions of, and Objections to, an Explicit Harms-Based Analysis', *Indiana Law Review* 29 (2): 257-337.

Vadas, Melinda (1987). 'A First Look at the Pornography/Civil Rights Ordinance: Could Pornography Be the Subordination of Women?', *The Journal of Philosophy* 84 (9): 487-511.

Valian, Virginia (1998). *Why So Slow? The Advancement of Women*, Cambridge Massachusetts: MIT Press.

van Dijk, Teun A. (1995). 'Elite Discourse and the Reproduction of Racism', in Rita Kirk Whillock and David Slayden (Eds.), *Hate Speech*, London: Sage.

Vandeveer, Donald (1979). 'Coercive Restraint of Offensive Actions', *Philosophy & Public Affairs* 8 (2): 175-93.

von Hirsch, Andrew (1985). *Past or Future Crimes: Deservedness and Dangerousness in the Sentencing of Criminals*, New Brunswick: Rutgers University Press.

_____ (2000). 'The Offence Principle in Criminal Law: Affront to Sensibility or Wrongdoing?', *King's College Law Journal* 11 (1): 78-89.

Waldron, Jeremy (1987). 'Mill and the Value of Moral Distress', *Political Studies* 35 (3): 410-23.

_____ (1989). 'Rights in Conflict', *Ethics* 99 (3): 503-19.

_____ (1991). 'Minority Cultures and the Cosmopolitan Alternative', *University of Michigan Journal of Law Reform* 25 (3): 751-94.

_____ (1993a). 'Locke, Toleration, and the Rationality of Persecution', *Liberal Rights: Collected Papers 1981-1991*, Cambridge: Cambridge University Press.

_____ (1993b). 'Rushdie and Religion', *Liberal Rights: Collected Papers 1981-1991*, Cambridge: Cambridge University Press.

_____ (2002). *God, Locke, and Equality: Christian Foundations of John Locke's Political Thought*, Cambridge: Cambridge University Press.

_____ (2008). 'Free Speech and the Menace of Hysteria', *The New York Review of Books* 55 (9).

_____ (2010). 'Dignity and Defamation: The Visibility of Hate (2009 Oliver Wendell Holmes Lectures)', *Harvard Law Review* 123 (7): 1596-657.

_____ (2012). *The Harm in Hate Speech*, Cambridge Massachusetts: Harvard University Press.

Walker, Samuel (1994). *Hate Speech: The History of an American Controversy*, Lincoln: University of Nebraska Press.

Walzer, Michael (1983). *Spheres of Justice: A Defence of Pluralism and Equality*, Oxford: Basil Blackwell.

_____ (1992). 'Comment', in Amy Gutmann (Ed.), *Multiculturalism and "the Politics of Recognition"*, Princeton: Princeton University Press.

Watterson, Kim M. (1991). 'The Power of Words: The Power of Advocacy Challenging the Power of Hate Speech', *University of Pittsburgh Law Review* 52 (4): 955-88.

Wedgwood, Ruth (1988). 'Freedom of Expression and Racial Speech', *Tel Aviv University Studies in Law* 8: 325-37.

Weil, Patrick (2009). 'The Politics of Memory: Bans and Commemorations', in Ivan Hare and James Weinstein (Eds.), *Extreme Speech and Democracy*, Oxford: Oxford University Press.

Weinman, Michael (2006). State Speech vs. Hate Speech: What to Do About Words That Wound? *Essays in Philosophy, 7*(1). Retrieved from http://commons.pacificu.edu/eip/vol7/iss1/18.

Weinstein, James (1991). 'A Constitutional Roadmap to the Regulation of Campus Hate Speech', *Wayne Law Review* 38 (1): 163-248.

_____ (1994). 'An American's View of the Canadian Hate Speech Decisions', in W. J. Waluchow (Ed.), *Free Expression: Essays in Law and Philosophy*, Oxford: Clarendon Press.

_____ (1999). *Hate Speech, Pornography, and the Radical Attack on Free Speech Doctrine*, Boulder Colorado: Westview Press.

_____ (2009a). 'Extreme Speech, Public Order, and Democracy: Lessons from *the Masses*', in Ivan Hare and James Weinstein (Eds.), *Extreme Speech and Democracy*, Oxford: Oxford University Press.

_____ (2009b). 'An Overview of American Free Speech Doctrine and Its Application to Extreme Speech', in Ivan Hare and James Weinstein (Eds.), *Extreme Speech and Democracy*, Oxford: Oxford University Press.

Weinstein, James, and Ivan Hare (2009). 'General Introduction: Free Speech, Democracy, and the Suppression of Extreme Speech Past and Present', in Ivan Hare and James Weinstein (Eds.), *Extreme Speech and Democracy*, Oxford: Oxford University Press.

Wells, N. Douglas (1995). 'Whose Community? Whose Rights? Response to Professor Fiss', *Capital University Law Review* 24 (2): 319-29.

Wertheimer, Aviva O. (1994). 'The First Amendment Distinction between Conduct and Content: A Conceptual Framework for Understanding Fighting Words Jurisprudence', *Fordham Law Review* 63 (3): 793-851.

West, Caroline (2003). 'The Free Speech Argument against Pornography', *Canadian Journal of Philosophy* 33 (3): 391-422.

_____ (2012). 'Words That Silence? Freedom of Expression and Racist Hate Speech', in Ishani Maitra and Mary Kate McGowan (Eds.), *Speech and Harm: Controversies over Free Speech*, Oxford: Oxford University Press.

Whillock, David E. (1995). 'Symbolism in the Representation of Hate in Visual Discourse', in Rita Kirk Whillock and David Slayden (Eds.), *Hate Speech*, London: Sage.

Whine, Michael (2009). 'Expanding Holocaust Denial and Legislation against It', in Ivan Hare and James Weinstein (Eds.), *Extreme Speech and Democracy*, Oxford: Oxford University Press.

Wieland, Nellie (2007). 'Linguistic Authority and Convention in Speech Act Analysis of Pornography', *Australasian Journal of Philosophy* 85 (3): 435-56.

Wiener, Richard L., and Erin Richter (2008). 'Symbolic Hate: Intention to Intimidate, Political Ideology, and Group Association', *Law and Human Behaviour* 32 (6): 463-76.

Williams, Bernard (1973a). 'Deciding to Believe', *Problems of the Self: Philosophical Papers 1956-1972*, Cambridge: Cambridge University Press.

_____ (1973b). 'The Idea of Equality', *Problems of the Self: Philosophical Papers 1956-1972*, Cambridge: Cambridge University Press.

_____ (2005a). 'Censorship', in Geoffrey Hawthorn (Ed.), *In the Beginning Was the Deed: Realism and Moralism in Political Argument*, Princeton: Princeton University Press.

_____ (2005b). 'Conflicts of Liberty and Equality', in Geoffrey Hawthorn (Ed.), *In the Beginning Was the Deed: Realism and Moralism in Political Argument*, Princeton: Princeton University Press.

_____ (2005c). 'Realism and Moralism in Political Theory', in Geoffrey Hawthorn (Ed.), *In the Beginning Was the Deed: Realism and Moralism in Political Argument*, Princeton: Princeton University Press.

Williams, David (2009). 'Hate Speech in the United Kingdom: An Historical Overview', in Ivan Hare and James Weinstein (Eds.), *Extreme Speech and Democracy*, Oxford: Oxford University Press.

Williams, Glanville (1989). '*Finis* for *Novus Actus*?', *Cambridge Law Journal* 48 (3): 391-416.

Williams, Patricia (1987). 'Spirit-Murdering the Messenger: The Discourse of Fingerpointing as the Law's Response to Racism', *University of Miami Law Review* 42 (1): 127-57.

Williams, Robert R. (1997). *Hegel's Ethics of Recognition*, Berkeley: University of California Press.

Williamson, Timothy (2009). 'Reference, Inference, and the Semantics of Pejoratives', in Joseph Almog and Paolo Leonardi (Eds.), *The Philosophy of David Kaplan*, Oxford: Oxford University Press.

Wolfson, Nicholas (1991). 'Free Speech Theory and Hateful Words', *Cincinnati Law Review* 60 (1): 1-42.

_____ (1997). *Hate Speech, Sex Speech, Free Speech*, Westport Connecticut: Praeger.

Wollheim, Richard (1973). 'John Stuart Mill and the Limits of State Action', *Social Research* 40 (1): 1-30.

Wright, R. George (1988). 'Racist Speech and the First Amendment', *Mississippi College Law Review* 9 (1): 1-28.

_____ (2006). 'Dignity and Conflicts of Constitutional Values: The Case of Free Speech and Equal Protection', *San Diego Law Review* 43 (3): 527-76.

Wright, R. W. (1985). 'Causation in Tort Law', *California Law Review* 73: 1737-828.

Young, Iris Marion (1990). *Justice and the Politics of Difference*, Princeton: Princeton University Press.

_____ (1997). 'Difference as a Resource for Democratic Communication', in James Bohman and William Rehg (Eds.), *Deliberative Democracy: Essays on Reason and Politics*, Cambridge Massachusetts: MIT Press.

Zick, Timothy (2004). 'Cross Burning, Cockfighting, and Symbolic Meaning: Toward a First Amendment Ethnography', *William and Mary Law Review* 45 (5): 2261-398.