Ethical Theory and Moral Practice (2019) 22:539–553
https://doi.org/10.1007/s10677-019-10002-0

What Is the Harm of Hate Speech?

Eric Barendt[1]

Published online: 25 May 2019

Abstract
In Jeremy Waldron's book, The Harm in Hate Speech, it is not always clear whether he argues that hate speech causes harm or whether it constitutes harm. This article considers this uncertainty, concluding that the best understanding of Waldron's argument is that hate speech tends to cause harm - a weak form of the consequentialist case for its proscription. His argument is not advanced by his apparent reliance on speech-act theory.

Keywords Waldron · The harm in hate speech · Causing or constituting harm · Speech-act theory

1 Introduction

What is the harm of hate speech - the harm for which hate speech is responsible? Is this harm sufficient to justify its proscription or regulation, given the strong protection liberal democracies often give freedom of speech in their constitution? The United States is unusual in affording racist hate speech more or less absolute protection under the First Amendment, but European jurisdictions too, including the United Kingdom, ban extreme speech only in relatively narrowly circumscribed circumstances.[1] One reason for reluctance to take this step is that hate speech is hard to distinguish from political speech which is strongly protected in liberal democracies; indeed, some speech, say, by a politician at a public rally calling for a halt to immigration can be regarded as both political and hate speech.

The recent literature on hate speech is much more voluminous than that on any other area of free speech philosophy or law. One of the most notable contributions has been that of Jeremy Waldron in his book, *The Harm in Hate Speech* (2012), a revision of the Holmes lectures he gave at Harvard in 2009. The title of the book is suggestive. It can be taken as intimating that the harm for which hate speech is responsible is not so much a result or consequence of speech targeted at, say, racial groups or homosexuals, but rather

[1]In the UK racist hate speech must be threatening, abusive or insulting, and intended or likely to cause racial hatred.

✉ Eric Barendt
e.barendt@ucl.ac.uk

[1] University College London, London, UK

lies in the speech itself. In other words, the harm is not *caused* by the speech, but the speech itself *constitutes* the harm. This distinction between speech causing harm and speech constituting harm has been drawn in recent writing on hate speech (see Maitra and McGowan 2012, 4–8) and may have significant implications. If hate speech amounts to harm, it might be equivalent to a harmful act such as an assault or environmental pollution, and in that event it would not even be covered by a freedom of speech principle or constitutional provision (see further Section 5 of this article). If that is Waldron's argument, we do not need to ask what harm is caused by extreme racist speech – psychological damage to members of the targeted group, a breakdown in harmonious community relations, or the threat to law and order – nor pose hard questions whether there is persuasive evidence that hate speech really does cause harm. But we do need to enquire whether Waldron's argument is that hate speech as such constitutes harm and whether that argument is really distinguishable from the more familiar claim that hate speech is likely to produce harmful consequences.

In this article I contend that Waldron's book is not always clear on this fundamental question: does hate speech cause or constitute harm? This uncertainty is troubling. The usual causation argument requires the production of persuasive evidence about the impact of hate speech if it is to trump freedom of speech, while the alternative claim that such speech constitutes harm is open to challenge for reasons explored in Sections 4 and 5 of this article. Waldron's claim may be both that hate speech is harm-producing speech and that it also amounts to a speech-act constituting harm itself. It is however far from clear that shortcomings of the traditional case for hate speech proscription – doubts whether there is evidence of its impact strong enough to outweigh freedom of speech – are compensated by resort to the alternative (or additional) claim that it constitutes harm.

This article begins by setting out the central claims in Waldron's argument and identifying some problems with them (Section 2). This section will be relatively brief, as it is intended to provide a background to the principal thesis of this article that his argument is sometimes unclear on the nature of the harm of hate speech. Waldron's claims have already been subject to criticism (Heinze 2016; Simpson 2013; Weinstein 2017a), largely on the ground (not discussed in this article) that Waldron's justification of bans on hate speech would weaken, or even annihilate, the legitimacy of the application of anti-discrimination laws to racists by depriving them of the opportunity to challenge their validity (Dworkin 2009, vii-ix, Weinstein 2017a, 530–33). Section 3 is the most important part of this article; it examines whether Waldron is arguing that hate speech *causes* real harm, significant enough to justify its proscription, or if it is part of his case that it also *constitutes* harm in itself. The latter argument seems to be an application to hate speech of the contention that *speech* (or at least some types of speech) amount to 'performative-utterances' (Austin 1975) or 'speech-acts' (Langton 1993; McGowan 2012; Maitra 2012). The coherence and scope of this claim is discussed in Section 4 of the article, in particular how far it is right to apply it to hate speech. Although some of this discussion might appear peripheral to the main thesis of this article, it is important insofar as Waldron's argument relies on this claim on the nature of speech.

Section 5 examines the relationship of the constitutional principle of freedom of speech to the philosophical arguments discussed in Section 4. If hate speech, let alone any kind of political speech, were regarded primarily as performative or as doing performative work (see Waldron 2012, 166 at note 35) the repercussions might be considerable; such speech might even be treated as 'conduct' rather than pure 'speech' and so fall entirely outside the scope of a

free speech provision like the First Amendment. The final section (Section 6) offers a brief conclusion on how Waldron's argument might best be understood: it should interpreted as making a 'weak' consequentialist claim that hate speech has a tendency, or is likely, to bring about social or personal harm (see the analysis of 'weak' consequentialist claims in Section 3).

2 Waldron's *The Harm in Hate Speech*

Waldron helpfully summarizes his argument early in the book (Waldron 2012, 4–5). He disagrees with the view, widely held in the United States, that the targets of hate speech should learn to put up with it, as freedom of speech is more important than minimizing their feelings of anxiety or distress. Waldron's argument is that hate speech undermines the sense of assurance, to which we are all entitled, that we will not be discriminated against or subject to violence just because we are a member of a particular racial, religious, or other vulnerable group. This assurance is a public good, providing a sense of inclusiveness to which a good society is committed. A second way of looking at the harm in hate speech is from the perspective of members of the targeted groups; hate speech injures their social standing, their *dignity* (his emphasis on p 5) which they should be able to rely on if they are to live with confidence. Hate speech laws vindicate the dignity of members of the targeted groups as equal members of society, a proposition which was upheld in 1952 by a majority of the US Supreme Court in *Beauharnais v Illinois,*[2] though that decision is now widely discredited as a matter of US constitutional law.[3] It is legitimate to ban hate speech, for it puts forward as a rival to the public good of assurance a nightmare vision of intolerance and hatred (Waldron 2012, 94–96).

Waldron is careful to distinguish the protection of dignity from the prevention of mere offence. If it means anything at all, freedom of speech must entail a liberty to disseminate offensive ideas, but that is not the same as a freedom to assault the dignity of members of vulnerable groups. This distinction, as Waldron admits, is hard to draw in the context of religious expression, where it may be very difficult to determine whether, say, anti-Catholic or Muslim expression is an offensive criticism of a religious belief or practice, or amounts instead to religious hate speech (Waldron 2012, 118–26; Barendt 2011).

Waldron's argument is attractive and vigorously maintained. Nevertheless, there are some uncertainties in it. For example, it is not entirely clear what the public good of assurance means and whether it can sharply be distinguished from the feelings of anxiety and insecurity which members of vulnerable targeted groups may experience when they encounter hate speech. Waldron may have in mind the impact of hate speech in the long term – whether a society exposed to it over years will continue to reflect the public goods of tolerance and acceptance of all racial and other groups – or he may be more concerned with the immediate impact of such speech on members of these groups (Simpson 2013, 722–724). On either interpretation it seems that he is making a consequentialist argument; hate speech causes a loss of the social assurance to which we are all entitled, while it equally advances a rival vision of a society where intolerance and discriminatory practices are acceptable. There may also be problems in

[2] *Beauharnais v Illinois*, 343 US 250 (1952).

[3] In 1952 libel was generally considered to fall entirely outside the First Amendment free speech guarantee, but that view is no longer tenable after the landmark ruling in *New York Times v Sullivan* 376 US 254 (1964) holding that the libel of public officials was protected by freedom of speech unless actual malice was proved.

evaluating the impact of hate speech, given that it is inevitably expressed in a society where a variety of attitudes to racial questions are held and where the government generally discourages or even outlaws racially (and religiously and other) discriminatory practices (Simpson 2013, 724–726). Consequently it may be difficult, perhaps impossible, to decide whether the dissemination of hate speech is responsible for the loss of assurance or whether that loss should be ascribed to other social phenomena, for example, discriminatory practices.

The argument that hate speech infringes human dignity is equally problematic, though it has been made by a number of scholars as well as Waldron (Heyman 2009, 166–169; Tsesis 2009). What does it mean to claim that hate speech *injures* human dignity? Does it do more than disparage members of the targeted groups, so that their feelings are injured and they fear they will be discriminated against? Hate speech does not as such amount to a denial of their rights, say, to vote or to secure adequate housing and other social benefits – although of course the widespread dissemination of such speech may make discriminatory practices more likely and more widely accepted. While a government might injure dignity by denying members of a racial group the right to vote, it is hard to see that any similar injury is done by the hate speech of racist bigots (Simpson 2013, 710–17). Moreover, as with the assurance argument, it is hard to isolate the contribution of hate speech to any loss of dignity from that occasioned by contemporary practices of discrimination and other conduct which may be targeted at members of a vulnerable racial (or other) group.

The dignity argument, as Eric Heinze has argued, is ambivalent (Heinze 2016, 32–38, 153–57). It does not make a consequentialist claim that could easily be established, given first the uncertain meaning of *injury* to dignity or social standing, and secondly the difficulty of showing that it is hate speech, rather than associated discriminatory practices and conduct, which causes the loss of dignity. It is perhaps better, more intellectually honest in Heinze's view, to see the dignity argument as a disguised deontological claim that hate speech as such *is* harmful, or that it constitutes harm. (Heinze 2016, 125–27). On that interpretation of Waldron's argument, hate speech as such amounts to an indignity, a loss of dignity for members of the targeted groups; it is pointless to ask questions about its causal impact. Whether that is Waldron's argument will be discussed in the next section of this article.

One final problem with Waldron's case for the proscription of hate speech should be briefly discussed: which groups should be protected by hate speech bans? Waldron usually refers to the need to protect the dignity of members of vulnerable minorities, to provide them with the assurance that they are full members of society in good standing (Waldron 2012, 5, 31–33, 37, 47). Emphasis is given to the protection of racial and religious groups (see Waldron 2012, ch 5, where he discusses racial epithets and the distinction between religious hatred and religious offence). Hate speech laws should also protect gays against homophobia (Waldron 2012, 65, 116). But should other groups be protected against hate speech, for example, asylum-seekers, single mothers, disabled people, welfare claimants, or political minorities? The failure of the UK hate speech laws to protect all vulnerable groups has been criticized for allowing discrimination between different points of view (Heinze 2006, 2009).[4] On this ground the US Supreme Court invalidated a city ordinance proscribing the placing of any object on public or private property calculated to cause alarm or anger on the basis of race, colour, creed,

[4] In contrast to UK law, the German Penal Code, sect 130 penalizes insults to 'segments of the population' (Barendt 2005, 180).

religion or gender (*RAV v City of St Paul* (1992).[5] The provision banned some types of hate speech, but not those directed against homosexuals or members of a political party.

It is not clear whether Waldron would apply his argument to protect, say, welfare claimants or members of minority pressure groups if they are targeted by hate speech. In principle it ought to be applied insofar as it could plausibly be claimed that their dignity had been injured or they lacked assurance that they stood in good standing as citizens. But that claim would be very hard to substantiate, just as it would be if it were made by a member of a (white) majority racial group targeted by a (black) racist speaker.[6] There must be some basis for a claim that a group is sufficiently vulnerable to need the protection of hate speech bans. This point surely suggests that Waldron's argument is fundamentally consequentialist in its character: hate speech is harmful only when there is evidence that it weakens the reputation and standing of members of particular groups, whose position has historically been vulnerable and who (still) need the protection of the law against speech which might further damage their dignity. If there is no evidence to support such a claim by members of a particular group, the case for a ban to protect them collapses. An alternative argument that hate speech necessarily constitutes an indignity for a targeted group lacks any plausibility if that group forms a (white) majority or is socially well protected as, say, an established political party or pressure group.

3 The Nature of Harm in Waldron's Argument

It is not entirely clear whether Waldron believes that hate speech *causes* harm sufficiently damaging for the law to step in, or whether it is inherently harmful, so that it *constitutes* harm and there is no need to examine its causal impact. Much of his argument suggests he is concerned with the consequences of hate speech, and that it is how it has been interpreted by some critics (Simpson 2013, 706–7, 723–24; Weinstein 2017b (2), 769–72). Waldron argues that hate speech undermines the public good of the assurance of security 'by intimating discrimination and violence' and by 'reawakening living nightmares' of what previous societies have been like. It poses an 'environmental threat to social peace, a sort of slow-acting poison…', producing its effects over years (Waldron 2012, 4). From the perspective of members of the targeted groups, hate speech 'sets out to make the establishment and upholding of their dignity' much harder (Waldron 2012, 5). The proscription of hate speech also ensures the maintenance of social peace and civic order (Waldron, 103–104). These arguments suggest that Waldron is concerned with the harmful consequences of extreme racist and other varieties of hate speech.

A striking analogy with laws concerned to protect the environment is drawn. In reply to the argument of free speech proponents that the law should intervene only when hate speech poses a clear and present danger to public order, Waldron points out that this argument does not hold good for environmental protection: we do not ask whether pollution is caused by *my automobile* (his emphasis), but look at the impact of millions of polluting vehicles. This reasoning – a development in 'consequentialist moral philosophy' – should be applied when examining the impact of the expression of hate on the social order (Waldron 2012, 96–97). This analogy surely shows that Waldron's case for the legitimacy of hate speech regulation is consequentialist in character: society is entitled to be concerned with the cumulative impact

[5] *RAV v City of St Paul* 505 US 377 (1992).
[6] The UK ban on racist hate speech has been applied to penalize militant black spokesmen: Barendt 2005, 178.

over the long term of the expression of racist ideas, whether or not particular harmful effects can be attributed to an individual pamphlet or social media message.

Waldron does not adduce empirical arguments which might show the damaging effects of hate speech in order to substantiate his case for a ban on its dissemination. But in reply to the point that it is impossible satisfactorily to distinguish between a wound to dignity on the one hand from mere offence on the other, he argues that the law can identify types of expression that 'are likely to have an impact on the dignity of members of vulnerable minorities' (Waldron 2012, 113). He appears here, as in other parts of his argument, to be concerned with the likely impact of hate speech on its victims. This is a weak form of consequentialist argument, for it refers to the *likely* impact of hate speech, rather than resting, as would a strong consequentialist argument, on a link between the speech and particular harm established by empirical evidence.[7] (There may well be evidence to establish such a link, but Waldron does not put it forward as part of his case.) An analogy can be drawn with libel law. Under the common law a defamatory allegation has been actionable if its publication had a tendency to injure the claimant's reputation: there was no need to establish that it did have this result.[8]

But other passages in *The Harm in Hate Speech* suggest a different interpretation of the argument: Waldron's case is not only that hate speech may cause harm, but also that it constitutes harm in itself. It should not primarily be regarded as asserting pernicious propositions with damaging consequences, but as equivalent perhaps to polluting acts like the defacement of buildings or waste-tipping. If that is what it means for the harm to be constituted by the speech itself, hate speech might even be regarded as not really 'speech', but treated as conduct which falls outside the scope of a provision such as the First Amendment (see Section 5 for consideration of this argument). As will be explained shortly, Waldron does not draw this possible implication of an argument that speech constitutes harm.

The most explicit indication that for Waldron the harm of racist speech is not caused, but is *constituted,* by it comes in his reply to an argument of Ed Baker that a ban on hate speech would wrongly limit the freedom of (racist) speakers to self-disclosure and to express their own view of the world (Baker 2009, 143); racist speech is expressive of their autonomy, and should not therefore be equated with acts of arson or assault (Baker 1997, 990–91). In response Waldron writes that hate speech is not purely expressive, but that by dispelling its victims' sense of assurance its work 'is largely performative' (Waldron 2012, 166). The harms with which Waldron is concerned are not only, as Baker argues, bad consequences of hate speech, but are also *constituted* by speech. 'The harm *is* the dispelling of assurance, and the dispelling of assurance is the speech act – it is what the speaker is doing in his self-disclosure….'(Waldron 2012, 167). It is perhaps hard to distinguish this bald conclusion from the argument that hate speech has a tendency to cause anxiety and feelings of insecurity among members of vulnerable groups – the weak consequentialist argument which dispenses with the production of evidence.

This last point can be made in respect of another disagreement between Baker and Waldron. The former argues (Baker 1997, 991–93) that hate speech harms because the audience reacts in

[7] The offence of incitement to racial hatred in the UK uses this weak consequentialist argument, for since 1976 it has been committed if speech is likely to stir up racial hatred: Barendt 2005, 178.

[8] Under the UK Defamation Act 2013 a claimant has to show that the libel has caused or is likely to cause serious harm to his reputation.

a particular way – by, for example, feeling wounded or more anxious. Waldron is unpersuaded (Waldron 2012, 168–172). Hate speech *inevitably* dispels the sense of assurance to which its victims are entitled, and there is harm in requiring them simply to ignore it or shrug it off. This point can be understood as a type of weak consequentialist argument about the likely tendency of hate speech.

There are other indications of Waldron's approach to the character of the harm in hate speech. At several points (Waldron 2012, 58, 73–4, 89–92) he refers enthusiastically to the writings of Catharine MacKinnon (MacKinnon 1991, 1994). His concerns about hate speech are similar to hers about pornography: just as sexually explicit material does not really communicate an idea for a masturbator to ponder, so racist defamation 'is not just an idea contributed to a debate.' Rather hate speech 'can become a world-defining activity… (Waldron 2012,74). Mackinnon does not regard pornography as propositional speech depicting a subordinate position for women and advocating discrimination against them. It amounts itself to acts of subordination and discrimination; it is 'used as sex, is sex' (Mackinnon 1994, 12). Interestingly, while sometimes MacKinnon appears to bracket hate speech with pornography (MacKinnon 1994. 9–10, 23), she has also drawn distinctions between them. Racist speech does not use its victims as pornography does, and it provides an argument, albeit one which is false and pernicious, while pornography does not work through thought and is not an argument at all (MacKinnon 1991, 803–4, and 1994, 43). Waldron might concur with this view, though his reply to Baker suggests that he does not think hate speech should be treated as only expressing ideas.

A final point is that Waldron clearly approves the decision of the US Supreme Court in *Beauharnais v Illinois* (1952) when it held that a group libel ordinance was compatible with the First Amendment guarantee of free speech (Waldron 2012, 47–52). In the view of the Court group libel fell outside the protection of the First Amendment. Waldron argues that hate speech laws should be viewed, like the ordinance in *Beauharnais,* as protecting vulnerable groups against defamatory attacks on their reputation and social standing; they are akin to group libel laws. Although such laws are enacted to protect the dignity of racial and other groups, and to avert the long-term risks to law and order from a breakdown in community relations, a prosecution for their infringement does not have to establish that the particular publication is likely to lead to any proven harm – whether to identifiable individuals or to the community at large. That would be required on a strong form of consequentialist argument. The publication of group libel, it is assumed, amounts automatically to harm. Waldron's support of hate speech bans on this basis suggests his view is that the dissemination of racist and other extreme speech constitutes harm in itself. But it could equally be taken as supporting the weak consequentialist claim that hate speech has an inevitable tendency to cause harm.

Waldron may therefore appear equivocal about the nature of the harm of hate speech: is it caused or constituted by such speech? Even when developing his reply to Baker, he states that 'the harmful *consequences* of the speech' (my emphasis) with which he is concerned do not disappear in the face of Baker's free speech arguments (Waldron 2012, 165). Further, he resists the temptation to treat all hate speech as falling entirely outside the scope of a free speech principle. When deciding whether to proscribe it, consideration must be given to the importance of free political speech which should then be balanced against the gravity of the harm occasioned by it (Waldron 2012, 145–47). That view seems hard to reconcile with his endorsement of the decision in *Beauharnais* which did not adopt this balancing approach.

Does it matter much which view Waldron holds of the harm of, or in, hate speech? Courts and commentators may often be unsure whether a particular type of speech can be banned because it falls outside the scope of a free speech principle altogether or because the arguments for its proscription trump the case for its protection as an exercise of freedom of speech. This is so with hard-core pornography and commercial speech, particularly advertising. So it is perhaps hardly surprising if Waldron sometimes appears to argue that hate speech should be banned because it causes harm of sufficient gravity to outweigh the case for its protection as an exercise of free speech, while on other occasions he apparently considers that hate speech constitutes harm itself. But the different approaches to the categorisation of the harm might have significant implications: if hate speech can be banned because it causes harm, some evidence should surely be produced to show the link between the speech and the consequent harm, but that link need not be required if the harm is inherent in the speech itself. In that event it could be claimed that hate speech, like hard-core pornography, is not really 'speech' for the purposes of a free speech principle, though powerful arguments would need to be adduced for that counter-intuitive conclusion (see Section 5). The best interpretation of Waldron's argument is that for him hate speech is both inherently harmful and that harm has, or is likely to cause, damaging consequences. Perhaps any shortcomings in the usual consequentialist arguments are compensated by Waldron's (additional) contention that hate speech is a speech-act doing performative work. In the following section I consider the coherence of that contention.

4 Speech-Act Theory Applied to Hate Speech

When Waldron wrote that the work that hate speech does is 'largely performative' he referred (Waldron 2012, 166, note 35) to lectures given by the linguistic philosopher JL Austin at Harvard University (published as *How to do Things with Words* 1975). Hate speech does performative work insofar as it dispels the sense of assurance which a good society provides for its members that their dignity is valued. Waldron also referred to the 'speech act', what the speaker 'is doing' with hate speech (Waldron 2012, 167), terms used by Austin. So it is important to unpack Austin's arguments and examine whether they strengthen Waldron's thesis.

In his first lecture Austin argued that philosophers had misunderstood the significance of making statements, by asking whether such statements were true or false. Not all speech describes or reports states of affairs. When someone says 'I do' at a marriage ceremony, names a ship, makes a contractual promise or places a bet, they are not making descriptive statements, but making a 'performative utterance' or more simply a 'performative' (Austin 1975, 4–7). When making a performative, a person is doing something by uttering words under a conventional procedure under which the utterance is valid – a lawful marriage, an enforceable bet or contractual promise – rather than true. Many of these performatives are authoritative, for example, the verdict of a judge, referee or umpire, or the pronouncement of a legislature or other official body (Langton 1993, 304–305).

In a subsequent lecture Austin coined the term 'illocutionary act' to refer to performatives when '*in* saying something we do something' (Austin 1975, 99–100). The 'locutionary' act *of* making a statement refers to its content, sense and meaning, and should be distinguished from the 'illocutionary' force of what the speaker is doing with the statement: acts such as asking or

answering a question, providing information, advice or a warning, pronouncing a verdict or sentence, or giving a description (Austin 1975, 99).[9]

Austin's argument, as he himself admitted, bristles with difficulties and uncertainties. In particular it is doubtful whether a sharp distinction can be drawn between descriptive statements which may be true or false and performatives which in appropriate circumstances may be valid or, if the circumstances do not apply, are null and void. While the early lectures suggested the distinction was clear and significant, many passages in subsequent lectures can be read as putting this in question. For example, Austin wrote: performing a locutionary act is 'in general … also and *eo ipso* to perform an illocutionary act…' (Austin 1975, 98). What is necessary for an illocutionary act is *uptake,* indicating that the meaning and force of the locutionary act is understood by its audience (Austin 1975, 117). There seems to be no requirement of an invocation of a conventional procedure in appropriate circumstances; uptake may be sufficient, as well as necessary, for the illocutionary act (Langton 2009, 78, 80). Indeed in the last two lectures (Lectures XI and XII) Austin doubted whether there was any real distinction between performatives and descriptive statements, given that every statement has a locutionary meaning and an illocutionary force. Both locutionary and illocutionary acts were aspects of 'the total speech act', a term used by Austin in his final lecture (Austin 1975, 148).[10]

Austin did not discuss how hate speech or pornography might be categorized. He found it difficult to characterise swearing or the expression of emotion as a performative because they were not acts done in conformity with a convention (Austin 1975, 105, 133) He might therefore have had problems in deciding how to treat racial abuse and other epithets. However, these utterances might have illocutionary force as what he termed a 'Behabitive', as being connected with the communication of attitudes and social behaviour (Austin 1975, 151, 160–61).

Despite its difficulties Austin's argument has influenced the approach of feminist philosophers to the treatment of pornography. They regard its depiction of women as an illocutionary speech-act (Langton 1993, 295–97). It amounts to acts of subordination and discrimination, rather than pure speech to be assessed primarily in terms of its content. MacKinnon has taken a similar view:'[p]ornography is more actlike than thoughtlike'(MacKinnon 1987. 154, 193).

Although Waldron only referred to Austin's work in a footnote (Waldron 2012, 166, note 35) it was probably attractive to him as it advocates that we should be concerned not only with the content of hate speech – its meaning – and with its consequences, but by what it does. Hate speech constitutes a type of speech-act in which the speaker may 'resent', 'curse' or express hatred of members of the targeted groups.[11] However, it is doubtful whether hate speech could be regarded as performative if we adopt the distinction drawn by Austin in the early lectures between performative utterances and descriptions or reports of states of affairs. Unlike marriage vows, the naming of a ship, an umpire's verdict or the making of a bet, hate speech is not disseminated under a conventional procedure in particular circumstances appropriate for use of the procedure. But if any speech can be regarded as having illocutionary force whenever it receives 'uptake' – the audience appreciates its meaning and force, as, say, a threat or

[9] Austin also distinguished from locutionary and illocutionary speech-acts what he termed 'perlocutionary' acts, the consequential effects of statements, for example, the speaker persuaded me to do x or convinced me that x is right (Austin 1975. 101–3). It is not important for the purposes of this article to say more about perlocutionay speech.

[10] Austin also used the term 'speech-act' in an unscripted talk on the BBC in 1956 (Austin 1970, 245, 251).

[11] 'Resent' and 'curse; are among the 'behabitives' in Austin's taxonomy of illocutionary speech-acts (Austin 1975, 160.

warning – then hate speech may be treated as having illocutionary force. Rae Langton regards this reciprocity of understanding as a sufficient condition (alternative to that of the authority of a conventional procedure like that for regulating a marriage ceremony) for recognizing pornography as an illocutionary speech-act (Langton 2009, 78). On her interpretation of Austin's argument, hate speech can certainly be treated as having illocutionary force: in expressing hate, the speaker is not only expressing pernicious ideas, but is doing something such as threatening, abusing, or warning.

One possible problem with this broad interpretation of Austin's argument is that any expression of racial (or other) antipathy, however moderate, might be treated as doing something with words, for it could be understood as a warning or veiled threat. Political speech pointing, say, to the dangers of unrestricted immigration or the large number of sexual crimes committed by a particular racial group might on this approach be regarded as illocutionary speech-acts and bracketed with hate speech. Waldron would not draw that conclusion but his reliance on Austin's work risks that misinterpretation. Of course, in some circumstances even moderate political speech might have a significant impact on a group's sense of social assurance; but a decision to proscribe it should only be taken if there is persuasive evidence that it does have that impact and if account is taken of the value of free political speech. These factors must be considered if Waldron's argument is that hate speech causes harm, but need not if his case is also that it constitutes harm.

Another narrower approach can be taken. Some philosophers have made strenuous attempts to argue that hate speech can treated as equivalent or similar to the speech-acts of persons in authority, which are clearly performative utterances under the arguments Austin made in his early Harvard lectures. For example, there can be little doubt that if a legislature such as that in apartheid South Africa deprives members of a racial group of its entitlement to vote, the enactment has both locutionary content and meaning – this group is not allowed to vote – and illocutionary force in annulling a prior entitlement and directing election officers not to allow certain people to vote (Langton 1993, 302–303). The performatives are made by a body – the legislature - with authority to act under prescribed procedures. A sign 'Whites Only' placed on a property for sale or outside a restaurant can similarly be regarded as a performative utterance discriminating against Blacks and other non-Whites; it is made by the owner of the premises with authority over its sale or use. The notion of authority, it has been argued, can be extended so it covers the communication of a message by someone who has been given permission to take control in a certain situation (derived authority) or by someone whose control is accepted by others in this situation (licensed authority); in this way racial abuse of targeted individuals in a crowded bus should be equated with the legislative denial to members of a racial group of the right to vote (Maitra 2012, 94–120). Mary Kate McGowan argues that a demand to an African American to get off a bus and go back to Africa is equivalent to a 'Whites Only' sign, in that it covertly grants permission to discriminate against the targeted passenger in a similar way to the display of the sign (McGowan 2012, 121–147).[12]

These arguments are ingenious, but unpersuasive. A racist on a crowded bus or underground carriage surely has no derived authority to abuse another passenger just because the driver and other passengers fail to intervene. Nor can it be said that the racist is 'licensed' to

[12] This sentence oversimplifies McGowan's complex argument: while the display of the sign is a standard 'exercitive' (the term used by Austin for the granting of permissions and authorizations), McGowan argues that racist speech on the bus is a *covert* exercitive, in that it constitutes a move in the norm-governed activity of racism: McGowan 2012, 132–36. Also see McGowan 2019.

communicate his ideas in these circumstances: by whose authority is he licensed? In any case it would be difficult to apply this argument to more typical examples of hate speech when racist ideas are disseminated in pamphlets left in bars or on social media, or put over at a political rally. If it can be said that any speaker trolling on the Net is exercising derived or licenced authority, the notion of 'authority' is emptied of any real meaning. McGowan's argument applies only to hate speech which can be understood to enact racial discrimination, but not to every verbal or written manifestation of hate (McGowan 2012, 141–42). So attempts to assimilate hate speech to performative utterances, insofar as they are understood as authoritative speech made under conventional procedures, seem to fail. If, however, Austin's broader view of such utterances is accepted, then not only hate speech, but all types of political speech and other speech, may be understood as having illocutionary force and as performative. So Austin's argument does not really advance Waldron's thesis: on a narrower approach to the argument, hate speech is not really a performative, while if the broader interpretation is taken too much speech might be caught as doing performative work.

5 Speech-Acts and Freedom of Speech

Waldron's argument that hate speech does performative work should now be examined from a legal perspective, in particular to see how far it is compatible with a free speech principle. In this context the approach taken by Kent Greenawalt, a leading American constitutional theorist, to the scope of freedom of speech can usefully be compared with Austin's philosophical views. A freedom of speech principle asserts that there is something special about *speech* and that it should not be banned or regulated on grounds that would justify comparable restrictions on conduct (Schauer 1982, 7–12). Two straightforward examples of this distinction can be given. It would be wrong to ban the expression of political or social views merely because they are regarded as offensive, but a community may limit planning development which it regards as aesthetically unattractive or out of character with the area. A society could legitimately ban fly-tipping or regulate the use of diesel cars because they cause environmental harm, but could not ban all political demonstrations on the ground that they interfere with the flow of traffic or disrupt commercial life. Speech enjoys greater immunity than conduct from state regulation for a number of reasons much discussed by political and legal philosophers: the communication of ideas is essential for the discovery of truth and for social progress, and it is crucial for the maintenance of a participatory democracy, while the freedom to express and receive information and opinion is vital for individual intellectual and moral self-development (Schauer 1982, chs 2–6).

But not all types of 'speech' in the dictionary sense of the word enjoy this immunity; in constitutional or legal terms they are not covered by a provision such as the First Amendment protecting freedom of speech.[13] Examples of such speech are the exchange of marriage vows, contractual promises and bets, bribery, perjury, and incitement to criminal conduct. Nobody could claim immunity from an action for breach of contract or from criminal prosecution for soliciting murder with an argument that in making the promise or inciting murder he was exercising his right to freedom of speech. These types of speech often change legal relations by creating obligations (marriage vows, promises and bets) or are made in a special legal or other

[13] In contrast some frms of 'conduct', for example, burning a flag or draft card, may be covered by freedom of speech because they are intended to communicate an idea and are so understood: Schauer 1982, 95–101.

environment (perjury, a jury verdict, an umpire calling 'Out' in cricket or tennis). Kent Greenawalt has termed these categories of uncovered speech 'situation-altering utterances'; the use of the words changes the situation – legal or otherwise – rather than providing information about the world or revealing the speaker's beliefs or values (Greenawalt 1989, 57–63). The reasons for treating speech as special under the free speech principle simply do not apply to situation-altering utterances; a contractual promise or bet does not contribute to the discovery of truth or to the working of a participatory democracy, and there is no good free speech justification for protecting speech which incites murder or indeed any crime.

To some extent the categories of 'speech' not covered at all by the freedom of speech principle are similar to the types of performative utterance discussed by Austin in his early Harvard lectures. But the categories are much narrower than Austin's performative utterances (Greenawalt 1989, 58). This is particularly evident in relation to the treatment of performatives in Austin's later lectures: once the dichotomy of performatives and descriptive statements is abandoned, there is an enormous family of related 'speech-acts' – utterances with both locutionary content and illocutionary force – which contain an immense range of expression (Austin 1975, Lecture XII, esp. 150).

Greenawalt was concerned to elucidate the appropriate constitutional and legal approaches to the understanding of language rather than the purely philosophical questions canvassed by Austin. The practical consequences of their different approaches can be teased out in the context of hate speech; in particular it can be asked how far the freedom of speech principle applies to extreme racist and other hate speech, or whether such speech falls outside the scope of the principle as might be the case if its dissemination constitutes harm. For Greenawalt hate speech is clearly covered by a free speech principle or constitutional clause, as is political speech generally; racist speech reveals the speaker's beliefs, and may even provide information about the world, though it will generally consist of bogus claims, as in the case of Holocaust denial. In his view (Greenawalt 1989, 300–301) that was what happened in the *Beauharnais* group libel case where a racist pamphlet had made highly provocative claims about the incidence of Black crime in Chicago: it was wrong for the Supreme Court to have ruled the speech outside the scope of the First Amendment. Irrespective of its content and damaging consequences, hate speech does not amount to a situation-altering utterance. But as was explained in Section 4 of this article, Austin's later Harvard lectures may be interpreted as suggesting that hate speech does something (its illocutionary force) – warning, urging, ordering, resenting, blaming, cursing – as well as containing words with a hateful meaning (its locutionary content). Indeed, if we are primarily concerned to emphasise the illocutionary force of speech, any exercise of freedom of speech can be understood as *doing* something with words.

But it would be wrong to infer from that conclusion that the distinction drawn by courts and constitutional commentators – necessary for the sensible application of a legal free speech principle – between 'speech' and 'conduct' is eroded. First, it is worth bringing out that Austin's preoccupation was with the elucidation of what he termed in his last lecture the 'total *speech* act' (my emphasis) (Austin 1975, 148). Although the theme of his lectures is that words are used to do things as well as state propositions, their focus is on speech and language, not on conduct. Moreover, Austin was not concerned with the interpretation of the free speech clause in a constitution, but with how philosophers should understand the use of words. Constitutional interpretation involves understanding the history and culture of a particular society and its legal precedents in addition to general (philosophical) principles for appreciating language. So it is a mistake to apply Austin's approach to distinguish speech from conduct (Langton

1993, 296). The distinction between 'speech' and 'conduct' in US constitutional law is made on the basis of contemporary social practices for the communication of ideas – for example, whether flag-desecration is a well-understood means of political protest – and taking into account previous court decisions on the issue (Barendt 2005, 78–86). Philosophers have disputed whether freedom of speech is the freedom of locutionary acts, the freedom to determine the content of speech (Jacobson 1995), or the freedom of illocution, a freedom to communicate by doing various things with words (Langton 2009, 84–87). But free speech lawyers, and philosophers interested in a political free speech principle, need not be troubled with this debate; it is immaterial for their concerns whether speech is a locutionary act or an act with illocutionary force. For them it can be both.

Waldron would surely accept this conclusion. In his view hate speech has content and meaning, but it is speech which does largely performative work. But crucially hate speech remains an exercise of free (political) speech, not conduct, and so is covered by constitutional provisions such as the First Amendment. If it did not qualify as 'speech' at all, there would be no need to balance the value of the particular speech against the gravity of the harm it caused; the law does not do that with, say, perjury and incitement to murder. But, as explained in Section 3, Waldron agrees that the value of hate speech as an exercise of free speech rights must be weighed against the seriousness of the harm for which it was responsible (Waldron 2012, 147). It is not clear how the characterisation of hate speech as doing performative work adds much, if anything, to this balancing process. What is more important is that, as Greenawalt has reminded us, from the perspective of the free speech principle, hate speech like that in *Beauharnais* is an exercise of free speech rights, so that strong arguments must be adduced for its exercise to be restricted.

6 Conclusions

If it were Waldron's case that hate speech constitutes, rather than causes, harm, he would not have substantiated it. At all events the philosophical arguments made by JL Austin do not advance that case. A racist pamphlet or message does say something about the world or the speaker's beliefs. It is not a 'performative utterance' in the narrow understanding of that concept derived from the early Harvard lectures, while on its broader interpretation much political speech can be treated as doing something with words. Austin's argument does not help to distinguish hate speech constituting harm from, say, moderate discussions on racial issues.

Hate speech is a particularly nasty type of speech, but it is still speech, not conduct, and is covered by the freedom of speech principle (Section 5). Waldron clearly agrees with that conclusion when he states that its value must be balanced against the gravity of the harm for which it is responsible. As explained in Section 3 of this article, though Waldron argues that hate speech does performative work, his case is essentially consequentialist: it is legitimate to proscribe it because of its impact on the sense of social assurance to which we are all entitled.

If the right to free speech is taken seriously, strong arguments must be advanced to justify its restriction and evidence adduced to establish a link between hate speech and the harm it is alleged to cause. It would be unreasonable to expect this evidence to be provided in Waldron's book, which is concerned to put forward general arguments of political principle. The best interpretation of his argument is that it is legitimate to ban hate speech because it has harmful

tendencies to endanger social cohesion and injure the dignity of targeted groups. That is the weak form of consequentialist argument: hate speech may be banned because of a general apprehension of its effects, not because there is evidence that it really does cause substantial harm, whether to social order or its victims (Section 3). This argument leaves much to the judgment of government when it is appropriate to intervene; for that reason alone it is unattractive to advocates of the free speech principle who are suspicious of government regulation of freedom of speech (Schauer 1982, 85–6).

Acknowledgements I am grateful to David Bentley, Eric Heinze and two anonymous reviewers for their comments on earlier drafts of this article.

References

Austin JL (1970) Performative utterances. In: Urmson JO, Warnock GJ (eds) Philosophical papers. Oxford University Press, Oxford, pp 233–255

Austin, JL (1975) *How to Do Things with Words,* 2nd ed (eds JO Urmson and M Sbisà). Oxford: Oxford University Press

Baker CE (1997) Harm, liberty, and free speech. South Calif Law Rev 70:979–1020

Baker CE (2009) Autonomy and hate speech. In: Hare I, Weinstein J (eds) Extreme speech and democracy. Oxford University Press, Oxford, pp 139–157

Barendt E (2005) Freedom of speech, 2nd edn. Oxford University Press, Oxford

Barendt E (2011) Religious hatred Laws: protecting groups or belief? Res Publica 17:41–53

Beauharnais v Illinois 343 US 250 (1952)

Dworkin, RM (2009) Foreword to *Extreme Speech and Democracy* (eds I Hare and J Weinstein), v-ix. Oxford: Oxford University Press

Greenawalt K (1989) Speech, crime, and the uses of language. Oxford University Press, New York

Heinze E (2006) Viewpoint absolutism and hate speech. Mod Law Rev 69:643–682

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

Heinze E (2016) Hate speech and democratic citizenship. Oxford University Press, Oxford

Heyman SJ (2009) Hate speech, public discourse, and the first amendment. In: Hare I, Weinstein J (eds) Extreme speech and democracy. Oxford University Press, Oxford, pp 158–181

Jacobson D (1995) Freedom of speech acts? A response to Langton. Philos Public Aff 24:64–79

Langton R (1993) Speech acts and unspeakable acts. Philos Public Aff 22:293–330

Langton R (2009) Sexual Solipsism. Oxford University Press, New York

MacKinnon CA (1987) Feminism unmodified. Harvard University Press, Cambridge

MacKinnon CA (1991) Pornography as defamation and discrimination. Boston Univ. Law Rev 71:793–815

MacKinnon CA (1994) Only words. Harper Collins, London

Maitra I (2012) Subordinating speech. In: Maitra I, McGowan MK (eds) Speech and harm. Oxford University Press, Oxford, pp 94–120

Maitra I, McGowan MK (2012) Introduction and overview. In: Maitra I, McGowan MK (eds) Speech and harm. Oxford University Press, Oxford, pp 1–23

McGowan MK (2012) On "whites only" signs and racist hate speech: verbal acts of racial discrimination. In: Maitra I, McGowan MK (eds) Speech and harm. Oxford University Press, Oxford, pp 121–147

McGowan MK (2019) Just words: on speech and hidden harm. Oxford University Press, Oxford

RAV v City of St Paul 505 US 377 (1992)

Schauer F (1982) Free speech: a philosophical enquiry. Cambridge University Press, Cambridge

Simpson RM (2013) Dignity, harm, and hate speech. Law Philos 32:701–728

Tsesis A (2009) Dignity and speech: the regulation of hate speech in a democracy. Wake Forest Law Rev 44: 497–532
Waldron J (2012) The harm in hate speech. Harvard University Press, Cambridge, Mass
Weinstein J (2017a) Hate speech bans, democracy, and political legitimacy. Const Comment 32:527–583
Weinstein J (2017b) Viewpoint discrimination, hate speech, and political legitimacy: a reply. Const Commen 32(2):715–782