Note: this is an article whose final and definitive version will be published in the *Journal of Law and Biosciences*; an 'online first' version of this article is available via: http://jlb.oxfordjournals.org/content/early/2015/10/16/jlb.lsv040.full.

Dehumanization: its operations and its origins

Robert Mark Simpson

Abstract. Murrow and Murrow offer a novel account of dehumanization, by synthesizing data which suggest that where subject S has a dehumanized view of group G, S's neural mechanisms of empathy show a dampened response to the suffering of members of G, and S's judgments about the humanity of members of G are largely non-conscious. Here I examine Murrow and Murrow's suggestions about how identity-based hate speech bears responsibility for dehumanization in the first place. I identify a distinction between (i) accounts of the nature of the harm effected by identity prejudice, and (ii) accounts of how hate speech contributes to the harms of identity prejudice. I then explain why Murrow and Murrow's proposal is more aptly construed as an account of type (i), and explain why accounts of this type, even if they're plausible and evidentially well-supported, have limited implications in relation to justifications for anti-hate speech law.

1. Introduction

Gail Murrow and Richard Murrow examine the phenomenon of *dehumanization*, by which they mean the thing that occurs when a subject, S, perceives people who belong to some cultural outgroup, G, as having a lesser degree of humanity than herself and others she identifies with.[1] Their inquiry is partly motivated by the conjecture – which they credit to the psychologist Gordon Allport, and the philosopher Hannah Arendt – that dehumanization can lead to mass, ethnically-motivated human rights abuses, like those of

[1] Gail B. Murrow and Richard Murrow, *A hypothetical neurological association between dehumanization and human rights abuses*, JOURNAL OF LAW AND THE BIOSCIENCES (2015), 1-29 (doi: 10.1093/jlb/lsv015).

the Holocaust. The principal contribution of their article is to articulate a more detailed (hypothetical) account of how dehumanization alters the cognition of affected subjects, which they do by synthesizing contemporary findings in neuroscientific research on empathy and psychological research on prejudice. They contend that where subject S has a dehumanized view of group G, S's neural mechanisms of empathy exhibit a dampened response to the suffering of members of G, and S's judgments, vis-à-vis the degree of humanity of members of G, are largely *non-conscious*, i.e. not a product of S's conscious beliefs, but rather a product of S's conditioned associations connecting G to sub-human traits. If this is how dehumanization works then it's plain to see why it's so dangerous. Affected subjects feel less disquiet about certain people's suffering, and consequently they're less averse to inflicting or permitting that suffering, and these dispositions are difficult to modify, because they're borne of subjects' implicit associations and not their conscious beliefs. Dehumanization, thus characterized, is a phenomenon that societies should take urgent measures to counteract, even at some cost to other values.

In what follows I'll provisionally grant Murrow and Murrow's account of how dehumanization operates once it's in effect. What I want to examine are the ramifications they seek to trace from this for questions about the legitimacy of laws restricting *hate speech*, i.e. speech that expresses or incites hatred towards people on the basis of some aspect of their identity, e.g. their ethnicity, nationality, or religion. Murrow and Murrow don't claim to be advancing a justification for general prohibitions on hate speech; what they purport to offer, rather, is a "theory regarding how hate speech might be threatening to its targets in ways which jurists, scholars, and scientists have not perhaps previously been aware."[2] They summarize the part of their account that's relevant in connection with this concern as follows:

> If hate speech conditions… an implicitly dehumanized view of its human targets, and, if this reduces empathy for targets that motivates prosocial behavior toward them, then, in legal terms, it also reduces the sense of equality that conspecifics have toward the targets and thus deprives the latter of equal protection of their rights, which is provided not simply by positive law, but by the capacity to provoke empathy. Therefore… it may pose a true threat to such targeted groups' safety or human rights.[3]

Here I'll question whether Murrow and Murrow's account of dehumanization in fact has the implications for anti-hate speech law that they suggest. I'll begin by sketching different candidate justifications for different kinds of anti-hate speech laws, and indicating the point at which an account like Murrow and

[2] *Ibid*, 19.

[3] *Ibid*, 19.

Murrow's has the potential to make an impact on the case for anti-hate speech law. I'll then briefly outline the views advanced by a few other scholars interested in the malign effects of identity-prejudicial speech, and in so doing flesh-out a distinction that becomes salient in examining such views – one that's also crucial for an assessment of Murrow and Murrow's proposal – between (i) accounts of the nature of the harm effected by identity-prejudice, and (ii) accounts of how hate speech contributes to the harms of identity-prejudice. I'll then explain why Murrow and Murrow's proposal is properly construed as an account of type (i), and explain why accounts of this type, even if they're plausible and evidentially well-supported, have limited ramifications in relation to justifications for anti-hate speech law.

2. Two justificatory approaches to anti-hate speech law

We should denounce the views of people who express identity-prejudicial sentiments, but we shouldn't legally restrict such expression except in cases where it does a discernible, non-trivial harm to some identifiable victim/s. This premise isn't predicated upon a commitment to any strong free speech principle, of the kind that underpins American First Amendment jurisprudence, and towards which Murrow and Murrow (and many others) are sceptical. Rather, to say that we should only penalise harmful hate speech is just to accept the tenet that the law should be used to remedy harms done to others, not to censure pernicious attitudes or penalise wrongs *per se*. I'll assume that tenet henceforth, without any defence except to observe that hate speech is one area of law in which non-harm-based justifications are especially at risk of devolving into expedient scapegoating. Assuming that hate speech may only be restricted when it harms others, it's then important to distinguish between:

> (i) Putative justifications for anti-hate speech law which claim that *specific acts* of hate speech *cause* harm to others; and
> (ii) Putative justifications which claim that, *in general*, hate speech *contributes* to harm to others.

Vis-à-vis (i), someone might claim that being confronted with face-to-face hate speech can cause distress of a magnitude sufficient to qualify as psychological harm ('cause' in a counterfactual sense, i.e. *but for* the relevant acts, the harm wouldn't have occurred). If this claim is evidentially well-supported, it seems sufficient to justify, in principle, an anti-hate speech statute narrowly aimed at restricting instances of hate speech that inflict the relevant harms. By contrast, vis-à-vis (ii), someone might claim that *all* hate speech – even where it isn't counterfactually responsible for harming others – contributes to the existence of *de*

facto social hierarchies which inflict harm upon targets of group-based identity-prejudice. What does 'contribute to' mean here? Roughly, the idea is that eliminating hate speech would increase the likelihood of the *de facto* social hierarchies being eradicated, but that holding other factors constant, this intervention alone wouldn't be sufficient to eradicate the hierarchies. Claims along these lines are much harder to substantiate, but if one were well-supported, it would seem sufficient to justify – again, at least in principle – a broad-scope statute aimed at legally restricting all instances of hate speech.

Claims of the type that are necessary to underwrite the first type of putative justification for anti-hate speech law are highly plausible. Only the most extreme free speech zealot would deny that *some* hate speech can do *some* harm to *some* of its targets.[4] Granted, there's room for disagreement about how exactly we should formulate laws restricting such directly harmful hate speech. Nonetheless, the more significant question about the in-principle justifiability of anti-hate speech law is whether any putative justification for restrictions of the second type can be sustained, e.g. whether it's plausible to say that instances of hate speech that are individually harmless nevertheless contribute to some sort of socially-mediated harm, in a way that renders all hate speech legitimately liable to legal restriction. If Murrow and Murrow's account of how dehumanization operates has any impact to make on the case in favour of anti-hate speech law, those debates are the place in which its conclusions need to register.

3. Two elements in an adequate account of speech-harm

In the literature on identity-prejudicial speech and its malign effects, we observe efforts to describe both (i) the *nature* of the harm that such speech (allegedly) contributes to, and (ii) the *route via which* such speech (allegedly) makes its contribution. Where a putative justification for anti-hate speech law is purporting to show that all hate speech contributes to a socially-mediated harm, an account needs to be given in regards

[4] Having said that, I should acknowledge that what's sometimes called the 'sticks and stones' view – roughly, the view that speech is ostensibly harmless, irrespective of the purposes to which it's being put – has rather too often been taken seriously in the literature. A significant part of the contribution made in early, influential work on hate speech by critical race theorists – e.g. Mari J. Matsuda, *Public response to racist speech: considering the victim's story*, and Charles R. Lawrence, *If he hollers let him go: regulating racist speech on campus*, both in WORDS THAT WOUND: CRITICAL RACE THEORY, ASSAULTIVE SPEECH, AND THE FIRST AMENDMENT (Boulder Colorado: Westview Press, 1993) – was to provide detailed analyses of the very significant forms of psychological harm inflicted by various kinds of face-to-face hate speech. For the definitive, sustained critique of the sticks and stones view, see Susan J. Brison, *Speech, harm, and the mind-body problem in First Amendment jurisprudence*, LEGAL THEORY 4 (1998), 39-61.

to both points. A quick sketch of some prominent recent work in the legal philosophical literature on identity-prejudicial speech will illuminate the nature of this demand that I'm outlining.

So, for example, in philosophical and legal theoretic work on the *silencing* effects of sexist expression, we get (i) an account of the nature of a certain kind of harm – namely, people being rendered unable to perform important speech acts that they intend to perform, i.e. suffering *illocutionary disablement* – and (ii) an account of the route via which sexist expression makes a crucial contribution to that harm – namely, by subverting certain key audience expectations which need to be in place in order for the important speech acts in question to be reliably and successfully performed.[5]

Or, in Jeremy Waldron's work on hate speech, we get (i) an account of the nature of a certain kind of harm – namely, people's losing their sense of assurance that they are fully socially enfranchised members of their society – and (ii) an account of the route via which racist expression makes a crucial contribution to that harm – namely, by making it unavoidably salient to people in targeted outgroups that they are held in contempt by others in their society.[6]

Or, in Jason Stanley's work on propaganda, we get (i) an account of the nature of a certain kind of harm – namely, disadvantage stemming from systematic identity-based discrimination across multiple areas of social policy – and (ii) an account of the route via which a certain form of racist propaganda makes a key contribution to that harm – namely, by effectively misrepresenting systematically identity-prejudicial practices as if they were egalitarian and meritocratic.[7]

Although the proponents of these accounts hold different views about the justifiability of legally restricting the expressive practices they're examining, each of them supplies the elements that would be needed in order to formulate a case for restriction. In each account we get an analysis of the nature of a complex socially-mediated harm, *plus* an explanation of how the conditions that facilitate the harm are created and sustained, and – crucially – this explanation is one that indicates how the relevant form of expression might be effecting an identifiable, distinct, and important contribution to those conditions. If

[5] Key works include: Jennifer Hornsby, *Disempowered speech*, PHILOSOPHICAL TOPICS 23 (1995), 127-48; Rae Langton, *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, 1998); Ishani Maitra, *Silencing speech*, CANADIAN JOURNAL OF PHILOSOPHY 39 (2009), 309-38.

[6] Jeremy Waldron, *The Harm in Hate Speech* (Cambridge Massachusetts: Harvard University Press, 2012).

[7] Jason Stanley, *How Propaganda Works* (Princeton: Princeton University Press, 2015).

Murrow and Murrow's analysis of dehumanization is to be usefully employed in an argument for anti-hate speech law, it likewise needs to supply each of these elements.

4. Dehumanization: its operations and its origins

Murrow and Murrow's paper does supply the first of these elements; it presents an account of the nature of a complex, socially-mediated harm. What their discussion doesn't supply is an explanation of why we should suppose that hate speech makes a crucial contribution to that harm. The mere fact that there exist instances of hate speech which manifest the speaker's dehumanized view of a certain group doesn't suffice to bridge this argumentative gap. Obviously people sometimes say things that overtly or covertly betray a view of groups as essentially sub-human. But the question that matters for our purposes here is whether such expressions are actually responsible, by themselves or alongside other factors, for non-trivial numbers of people coming to accept a dehumanized view of others.

I don't think we're in a position to speak confidently in the affirmative. Even if we could show that the incidence of dehumanizing attitudes and dehumanizing hate speech reliably co-vary across cultural and historical circumstances, this *still* wouldn't reveal which way the causal arrows point. *Maybe* an increase in dehumanizing hate speech in public discourse is responsible for an increase in people's having dehumanized views; but then maybe dehumanizing attitudes flourish in a political milieu for other reasons, and they are a cause, rather than a consequence, of increases in dehumanizing hate speech in public discourse. Or alternatively – probably this is the safest conjecture to advance from the armchair – the causal arrows criss-cross in multiple directions simultaneously. In any case, if we have credible alternative hypotheses about what could be the crucial contributing factors in the inculcation of dehumanizing attitudes (and we clearly *do* have such hypotheses to consider, e.g. historical enmities, short-term conflicts borne of economic crises, innate tendencies towards ethnic outgroup enmity, various combinations of the above), then it's conjectural to attribute a proliferation of dehumanized attitudes to the influence of hate speech. Granted, there is work in social psychology that aims to shift questions about hate speech's effects away from the realms of intractable sociological speculation, into more

controlled experimental conditions,[8] but the all-things-considered implications of this work are, I think it's fair to say, uncertain.

The natural reply to this, for Murrow and Murrow, is to reiterate that they only ever promised an account of how hate speech threatens its targets in ways jurists and scholars haven't paid close attention to previously. But that is, I'd contend, a subtle mischaracterization of what their analysis can and does achieve. Having advanced an explanation of how dehumanization works and why it's so dangerous (by my lights, one that's plausible on both fronts) what their account entails for political practice really ought to be described schematically. Dehumanization carries significant threats to people's basic rights, which (arguably) have previously been underappreciated, and therefore the acts or practices that make a significant contribution to dehumanization – *whatever those acts may be, precisely* – endanger people's rights in a way that calls for an urgent response. Murrow and Murrow's account tells us something noteworthy about dehumanization's dangers, but rather less about its originating sources.

5. Danger, Urgency, and Uncertainty

My point in all this isn't to endorse the idea that in general hate speech *doesn't* contribute significantly to dehumanization, it's to say we don't know to what extent hate speech is involved here, and that if there *are* milieus in which it's a driving causal force, we know little about what it is within those milieus that facilitates the uptake of the dehumanizing attitudes being promulgated in hate speech.

None of this should detract from the primary implications of Murrow and Murrow's theses about how cognition is altered for subjects affected by dehumanization. Dehumanization is dangerous; it's plausible to suppose that it's some part of the explanation of mass state violence against ethnic outgroups, both in cases like Nazi Germany, where racialized justifications for violence were espoused forthrightly, and in cases like contemporary mass incarceration in the United States, where racialized violence runs amok even while overtly identity-prejudicial justifications for this state of affairs are largely confined to the political fringes. However, the gravity of such wrongs and the urgency of redress means that the standards we apply, in assessing explanations of how the wrongs occur and what redress would require, should be more exacting, not less. In this I'm echoing Henry Louis Gates' worry, that calls to

[8] See for example the products of the research program pursued by the late Brian Mullen and collaborators at the University of Syracuse, e.g. Brian Mullen and Diana R. Rice, *Ethnophaulisms and exclusion: the behavioural consequences of cognitive representation of ethnic immigrant groups*, PERSONALITY AND SOCIAL PSYCHOLOGY BULLETIN 29 (2003), 1056-67.

curb racist speech may encourage societies to downplay the deeper, less easily remediable, economic underpinnings of identity-based social hierarchies.[9] I'm also betraying a sympathy for Martha Minow's claim 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."[10] These ideas aren't new, but they bear repeating.

I should close by noting one area in which claims about hate speech's dehumanizing influence are not so conjectural, but instead empirically well-informed. International relations theorists and political philosophers trying to diagnose the causes of mass atrocity crimes, as in Rwanda in 1994, have identified intervention in dehumanizing propaganda as being, in at least some cases, a key preventative strategy.[11] The intersection between this work, with its empirical insights into the origins of dehumanized attitudes (at least in certain kinds of contexts), and Murrow and Murrow's account of how exactly dehumanization modifies the cognition of affected subjects, is likely to be a fruitful space for inquiry.

[9] Henry Louis Gates, *Let them talk: why civil liberties pose no threat to civil rights*, NEW REPUBLIC 37 (1993), 43-49, at 43.

[10] Martha Minow, *Regulating hatred: whose speech, whose crimes, whose power? An essay for Kenneth Karst*, UCLA LAW REVIEW 47 (2000), 1253-77, at 1274.

[11] See for instance: Susan Benesch, *The ghost of causation in international speech crimes cases* in Predrag Dojcinovic (Ed.), PROPAGANDA, WAR CRIMES TRIALS, AND INTERNATIONAL LAW (London: Routledge, 2012); Lynne Tirrell, *Genocidal language games* in Ishani Maitra and Mary Kate McGowan (Eds.), SPEECH AND HARM: CONTROVERSIES OVER FREE SPEECH (Oxford: Oxford University Press, 2012).