'Won't somebody please think of the children?' Hate speech, harm, and childhood

Robert Mark Simpson

Abstract: Some authors claim that hate speech plays a key role in perpetuating unjust social hierarchy. One *prima facie* plausible hypothesis about how this occurs is that hate speech has a pernicious influence on the attitudes of children. Here I argue that this hypothesis has an important part to play in the formulation of an especially robust case for general legal prohibitions on hate speech. If our account of the mechanism via which hate speech effects its harms is built around claims about hate speech's influence on children, then we will be better-placed to acquire evidence that demonstrates the processes posited in our account, and better-placed to ascribe responsibility for these harms to individuals who engage in hate speech. I briefly suggest some policy implications that come with developing an account of the harm of hate speech along these lines.

1. Introduction: hypothesizing about hierarchy

The equal moral standing of all people, regardless of their social identity, is an axiomatic commitment in liberal democratic thought. And yet all liberal democracies have unjust *de facto* social hierarchies, correlated with social identity categories like sex, race, and religion, to some extent. By '*de facto* hierarchies' I don't mean arrangements in which groups are officially disenfranchised or made into second-class citizens. I mean systematic identity-based inequalities in material resources, labor conditions, social mobility, and social influence – inequalities which, in modern liberal democracies, exist alongside *de jure* guarantees of

civic equality. The larger questions behind this paper are how to understand the causal forces involved in people being subordinated in these *de facto* hierarchies, and how the hierarchies may be changed.[1]

When progressives today engage with these questions they often pay considerable attention to communication. One example of this is criticism levelled against satirists of the *Charlie Hebdo* type, arguing that they're not legitimate critics of religion, but accomplices in the persecution of a racialized minority.[2] Another example is the considerable emphasis on 'misgendering' in transgender activism.[3] These are two instances of a broader trend, in which progressives attribute harmful power to communicative behaviors and then try to police them. In response to this trend we see many critics denouncing 'political correctness gone mad' and complaining about the 'regressive left'.[4] And even if we are sceptical about the anti-PC brigade's complaints, we should grant that they are right about at least one thing: many progressives in the 21st century *have* been treating speech as a key battle-ground in campaigns for social change.

Plenty of philosophers defend views that favor the idea that communication plays a key role in perpetuating social injustice. Consider philosophical contributions to the feminist anti-pornography literature,[5] or recent work on slurs,[6] or stereotype threat,[7] or Butlerian analyses of how people's performance of social identities reifies cultural norms.[8] I'm not saying that philosophers run as a pack in their views on how communication supports social hierarchy, just that many of them agree that communication is an important factor. And this shouldn't be surprising. Philosophical inquiry often focuses our attention on

---

[1] At various points in the paper I will use the term 'subordination' to refer to the process of people being assigned an inferior position in these *de facto* social hierarchies. The question of how these hierarchies operate, and what perpetuates them, is one that many scholars have examined. I am just addressing one narrow class of hypotheses about the role of hate speech in this. Influential examples of more general inquiries around this topic include Iris Marion Young, *Justice and the Politics of Difference* (Princeton: Princeton University Press, 1990); Nancy Fraser and Axel Honneth, *Redistribution or Recognition? A Political Philosophical Exchange* (London: Verso, 2003).

[2] E.g. Anshuman Mondal, 'Charlie Hebdo reinforces the very racism it is trying to satirise', *The Conversation*, 16th January 2016; theconversation.com/charlie-hebdo-reinforces-the-very-racism-it-is-trying-to-satirise-53263.

[3] E.g. Y. Gavriel Ansara and Peter Hegarty, 'Methodologies of misgendering: recommendations for reducing cisgenderism in psychological research', *Feminism & Psychology* 24(2) (2014): 259-70.

[4] One recent and prominent presentation of complaints along these lines comes in Greg Lukianoff and Jonathan Haidt, 'The coddling of the American mind', *The Atlantic* 316 (2015): 42-53.

[5] E.g. Rae Langton, *Sexual Solipsism* (Oxford: Oxford University Press, 2009).

[6] E.g. Renée Jorgensen Bolinger, 'The pragmatics of slurs', *Noûs* 51(3) (2017): 439-62.

[7] E.g. Jennifer Saul, 'Implicit bias, stereotype threat, and women in philosophy', in K. Hutchison and F. Jenkins (eds.), *Women in Philosophy: What Needs to Change?* (Oxford: Oxford University Press, 2013), pp. 39-60.

[8] Judith Butler, *Gender Trouble: Feminism and the Subversion of Identity* (London: Routledge, 1990) is the classic text.

the subtleties of language, and there is a temptation in this mode of inquiry to assign language a position at the explanatory center of everything. In short, plenty of philosophers are fellow travellers with progressives who propose that communication plays a major role in sustaining *de facto* social hierarchies.

Although we should treat this view seriously, we should be wary of any over-confident or hyperbolic characterization of the causal role that communication plays in *de facto* social hierarchy. The forces that underpin social hierarchies are of course enormously complex. It is hard enough to explain how the major economic elements of a social order function: trade, jobs, housing, and the organisation of business and government. Institutional practices (e.g. in policing, the courts, schools, and workplaces) further complicate the picture, and there are also private domains in which social arrangements are largely shaped by informal norms and customs. With these other factors in view, the cautious position would be to say that social hierarchies aren't sustained by communicative behaviors as such, but by an interlocking network of policies, institutions, and material conditions, which advantage some people and groups over others. But then, given that sort of picture, there is room for doubt about whether speech is making *any* distinctive and significant causal contribution to social hierarchy. After all, it may be that when people say racist things, for instance, this is largely an epiphenomenal symptom of a deeper racist social order, whose etiology lies in other material and institutional forces.[9] And in response to this it isn't enough to simply assert that speech *is* doing the work. Our account becomes speculative past the point of credibility if it suggests that words can create social realities *ex nihilo*. We need hypotheses to specify how, exactly, speech might be an operative factor in these causal systems, and then credible evidence to substantiate those hypotheses.

Among the different kinds of speech that might be seen as contributing to social hierarchy, one that is often singled-out for attention is speech that overtly expresses contempt or disdain towards people on the basis of their social group, e.g. speech that essentializes groups with negative traits ("all Muslims are terrorists"), or uses slurs or dehumanizing terms to convey a view of the group as disgusting, evil, or in some other way of lesser status. Like many others, I will narrow my inquiry to this class of communicative conduct. And I will use the term 'hate speech' to refer to it.[10] In narrowing my inquiry like this, I

---

[9] This kind of view doesn't entail that speech is completely *uninvolved* in the causal processes through which social hierarchies are sustained, just that the power of speech and communication, where they are causally involved in these processes, is parasitic upon material and institutional factors. As Pierre Bourdieu says, in this vein, "what creates the power of words… is the belief in the legitimacy of words and of those who utter them", and "words alone cannot create this belief"; 'On symbolic power', G. Raymond and M. Adamson, (Trans.), in J. B. Thompson (ed.), *Language and Symbolic Power* (Cambridge: Polity Press, 1991), p. 170.

[10] In this definition I am following James Weinstein and Ivan Hare. "In its purest form" they say, "hate speech is simply expression which articulates hatred for another individual or group, usually based on a characteristic (such as race) which is perceived to

don't mean to deny that other forms of communication, apart from overt hate speech – like everyday chatter in the home and workplace, or the stereotyping of groups in the mainstream media – might play an important causal role in sustaining social hierarchies.[11] Indeed, given that these other forms of communication are in certain ways more ubiquitous and less avoidable than hate speech, and sometimes more subtle in encoding identity-prejudicial views, they may sustain *de facto* social hierarchies in ways that overt hate speech doesn't or couldn't replicate. Nevertheless, much of the work that has been done by philosophers and legal theorists around this topic focuses on the effect of more overt forms of identity-prejudicial communication, and that is where I will focus too.

More specifically, I want to examine the hypothesis that hate speech contributes to identity-based social hierarchies by influencing children to support or accept those hierarchies. This hypothesis isn't entirely novel (see §2.1). What I am trying to do here is to build on lines of inquiry that are suggested in the work of other authors, by identifying certain merits in this kind of hypothesis that aren't fully recognized in the literature on this topic. I should acknowledge, at the outset, that imploring others to 'think of the children' can sometimes just be cheap, emotive bluster.[12] But while it is important to be mindful of this concern, we shouldn't dismiss a *prima facie* plausible hypothesis about the role of communicative factors in social hierarchy just because of its superficial similarity with moralistic rhetoric. It is reasonable to 'think of the children' in a discussion about the harm of hate speech, as long as we proceed cautiously.

---

be shared by member of the target group"; see 'General introduction: free speech, democracy, and the suppression of extreme speech past and present", in I. Hare and J. Weinstein (eds.), *Extreme Speech and Democracy* (Oxford: Oxford University Press, 2009), pp. 1-7, 4. Some authors define 'hate speech' in a way that also emphasizes the feelings that certain speech characteristically *elicits*, and not just the feelings it *expresses*; e.g. Rae Langton, 'The authority of hate speech', forthcoming in *Oxford Studies in Philosophy of Law*, Vol. 3. Alexander Brown has recently argued that, on the understanding of the term 'hate speech' that is acquiring popular currency beyond legal discourse, hatred needn't be involved in hate speech in *any* respect, either in the attitudes it expresses *or* elicits; 'What is hate speech? Part 1: the myth of hate', *Law and Philosophy* 36(4) (2017): 419-68. Nevertheless, I take it that the class of communicative acts picked out by the definition that I have given merits attention in its own right, in part because its members are the paradigmatic instances of the kinds of speech that are regulated by the kinds of laws customarily identified as 'hate speech regulations'; for an overview of these, see Ivan Hare and James Weinstein (eds.), *Extreme Speech and Democracy* (Oxford: Oxford University Press, 2009), pp. xxxix-xlvii.

[11] On everyday chatter, see Mary Kate McGowan, 'Oppressive speech', *Australasian Journal of Philosophy* 87(3) (2009): 389-407; on media speech, see Jacob Rowbottom, 'Extreme speech and the democratic functions of the mass media', in I. Hare and J. Weinstein (eds.), *Extreme Speech and Democracy* (Oxford: Oxford University Press, 2009), pp. 608-30.

[12] Indeed, such rhetoric can be used to stir moral panics; see Marjorie Heins, *Not in Front of the Children: 'Indecency', Censorship, and the Innocence of Youth* (New Brunswick: Rutgers University Press, 2007). This is lampooned in episodes of *The Simpsons* where the shallowly pious Helen Lovejoy cries "won't somebody *please* think of the children".

The merits of focusing on hate speech's influence on children don't really come into play if our question is just whether some instances of hate speech are harmful to particular individuals. The answer to that is uncontroversial. Token instances of speech that expresses contempt towards people on the basis of their social group can be used to harass, threaten, and incite violence. We don't need to be specially convinced that these instances of hate speech are harmful, or that there is an in-principle justification for legally restricting them.[13] There is room for debate about what the right regulatory approach is in this area, e.g. whether we should have customized restrictions on hate-speech-as-harassment, or rely on generic anti-harassment laws. But notwithstanding these open questions, the real controversy over hate speech – and the controversy to which my discussion in this paper is addressed – is not about whether hate speech is harmful in specific instances, for people whom it is used to directly and personally attack. The controversy, rather, is about whether *all* instances of hate speech are implicated in harming others, in a way that would give us an in-principle justification for what I will call BANS, i.e. general legal prohibitions on hate speech, which apply *irrespective* of whether the targeted speech is being used to harass, incite violence, or in any other direct way threaten or harm people.[14]

It is true that in focusing on the case for BANS we are setting a high bar for opponents of hate speech. One could argue that we have grounds for thinking hate speech makes *some* contribution to social hierarchy – one which justifies *some* form of legal response, like anti-discrimination laws that disallow hate speech in workplaces – while at the same time believing we lack the evidence we would need in order to assert that all hate speech is harmful in a manner that would justify BANS. Still, the question I intend to explore here is what it would take to satisfy that more demanding standard of justification. And this is part of what makes the focus on children relevant. If we are aiming to develop an evidentially-supported defense of the thesis that hate speech plays a significant causal role in sustaining social hierarchies – one

[13] Granted, this hasn't always been true. One contribution of critical race scholarship on hate speech has been to create a wider recognition of the threatening and harassing power of hate speech; see in particular, the seminal collection Mari J. Matsuda, Charles R. Lawrence, Richard Delgado, Kimberlé Williams Crenshaw, *Words That Wound: Critical Race Theory, Assaultive Speech, and the First Amendment* (Boulder: Westview Press, 1993).

[14] Authors who defend BANS include Anthony Cortese, *Opposing Hate Speech* (Westport Connecticut: Praeger, 2006), Alexander Tsesis, 'Dignity and speech: the regulation of hate speech in a democracy', *Wake Forest Law Review* 44(2) (2009): 497-532, and Jeremy Waldron, *The Harm in Hate Speech* (Cambridge: Harvard University Press, 2012). Representative examples of authors who criticize BANS include James Weinstein, 'Extreme speech, public order, and democracy', in I. Hare and J. Weinstein (eds.), *Extreme Speech and Democracy* (Oxford: Oxford University Press, 2009), pp. 23-61, and Eric Heinze, *Hate Speech and Democratic Citizenship* (Oxford: Oxford University Press, 2016).

with the potential to underwrite an in-principle justification for BANS – the hypothesis most likely to realize this aim is one focusing on hate speech's influence on children. Or so I will argue.

The rest of the paper is organized as follows. In §2 I survey some of the existing work on hate speech's influence on children, and I discuss three conditions that an account of hate speech's harm needs to meet in order to provide an in-principle justification for BANS. First, it should explain how all instances of hate speech make a contribution to the postulated harm. Second, it should be the kind of account for which it is possible in principle to acquire evidence that substantiates the key claims about how this contribution occurs. And third, it should explain how the person who engages in hate speech, i.e. the 'hate speaker', can justifiably be ascribed responsibility for the harm. In §3 I discuss the advantages of focusing on hate speech's influence on children when formulating an account of its harm, in a way that links up with these three conditions. I conclude in §4 by sketching some of the policy implications that may follow if we account for hate speech's harm in a way that emphasizes its influence on children.

2. What is needed in an account of hate speech's harm?

In scholarly writing that examines identity-prejudicial communication, and the case for its regulation, we find a number of passing comments on the negative impact that such communication can have on children.[15] There are only a few authors, though, who explicitly claim that hate speech contributes to social hierarchy specifically *through* its influence on children.

---

[15] In his 1962 Presidential address to the American Political Science Association, Charles S. Hyneman derided the marketplace of ideas ethos in First Amendment theory, and said he couldn't see "why there is so little support… for governmental action designed to lessen or prevent the indoctrination of children" into racist views; see 'Free speech: at what price?', *The American Political Science Review* 56(4) (1962): 847-52, 849. More recently, in arguing that pornography subordinates women, Rae Langton suggests that a key issue is whether pornography "is authoritative for… the *fifty percent of boys* who 'think it is okay for a man to rape a woman if he is sexually aroused by her'"; see 'Speech acts and unspeakable acts', *Philosophy & Public Affairs* 22(4) (1993): 293-330, 311-12, my emphases. She also discusses pornography's influence on children by examining the findings of the 2013 Report of the UK Office of the Children's Commissioner, into adolescents' views about consent; see 'Is pornography like the law?', in M. Mikkola (ed.), *Beyond Speech: Pornography and Analytic Feminist Philosophy* (Oxford: Oxford University Press, 2017), pp. 23-38. Jeremy Waldron is another influential scholar who alludes to the impact of identity-prejudicial speech on children. He opens his book on the subject with the following story. "A man out walking with his seven-year-old son and his ten-year-old daughter turns a corner on a city street in New Jersey and is confronted with a sign. It says: 'Muslims and 9/11! Don't serve them, don't speak to them, and don't let them in'. The daughter says, 'What does it mean, papa?' Her father, who is a Muslim… doesn't know what to say"; see *The Harm in Hate Speech*, p. 1. For Waldron, such episodes reveal hate speech's *raison d'être*, which is to ensure that "for the father walking with his children… there will be no knowing when they will be confronted by one of these signs"; see *ibid*: 3.

2.1 Existing work on hate speech and childhood

Delgado and Stefancic make this claim in *Understanding Words that Wound*, a text that digests the key ideas in discussions about hate speech from critical race theory. They devote a chapter to hate speech's bad influence on children, arguing that "much of the blame" for feelings of inferiority among minority groups "rests with the words and names children are exposed to while growing up", and that since children have fewer coping mechanisms than adults, they are "particularly susceptible to the wounds words can inflict".[16] But the evidence cited to support all this is not entirely convincing. For instance, when Delgado and Stefancic say that much of the blame for feelings of racial inferiority lies with the words and names children are exposed to while growing up, they cite Delgado's claim that minority children "question their competence, intelligence, and worth" primarily because "they constantly hear racist messages".[17] The supporting citations for this come from classic social scientific texts on racial prejudice from the mid-20th century, by Robert Redfield, Gordon Allport, and Mary Goodman, and the evidence in these texts just indicates that racist communicative practices are one factor among others in generating racial stigma, not something to which 'much of the blame' can be assigned.[18]

Cortese is another scholar who devotes particular attention to the effect of hate speech on children. Drawing on Piaget's developmental psychology, Cortese claims that progress in the child's cognitive development consists in the expansion of her ability to construct a 'world-image' through sympathetically identifying with other people's perspectives.[19] On this picture, in-group and out-group associations are 'hardwired' in our cognition, as a consequence of this developmental process. And Cortese's contention, then, is that hate speech impairs the processes involved in the child's 'socioemotional' development, in a way that ultimately leads to identity-prejudicial attitudes in adulthood, which in turn contributes to the perpetuation of identity-based social hierarchy.[20] Again, though, the evidence used to support these claims about the influence of hate speech is unpersuasive. The empirical studies that Cortese cites indicate that

---

[16] Richard Delgado and Jean Stefancic, *Understanding Words that Wound* (Boulder: Westview, 2004), pp. 93, 95.

[17] Richard Delgado, 'Words that wound: a tort action for racial insults, epithets, and name-calling', *Harvard Civil Rights-Civil Liberties Law Review* 17(1) (1982): 133-82, 146.

[18] For instance, the text from Allport that they cite says that "plural causation is the primary lesson we wish to teach", and that economic exploitation and social structure are important contributors to prejudice, and also that it is "a serious error to ascribe prejudice and discrimination to any single taproot"; Gordon W. Allport, *The Nature of Prejudice* (Reading Massachussetts: Addison-Wesley, 1954), p. xvi.

[19] Cortese, *Opposing Hate Speech*, 144-45.

[20] See *ibid*: 144.

prejudicial attitudes manifest in children's thoughts at alarmingly young ages. But evidence of these effects doesn't substantiate his key claim that hate speech is responsible for them. Cortese's attempt to link questions about hate speech's effects to developmental psychology adds an important element to the hate speech literature, but it doesn't vindicate the suggestion that hate speech should be legally restricted because of its pernicious influence on children.[21]

2.2 Causing and contributing to harm

As well as the limitations in the evidence they cite, the accounts discussed above are partly orthogonal to our purposes, since they are partly concerned with individual-level harmful effects of particular instances of hate speech, which, as explained in §1, will not be the focus of our inquiry here. But having said that, the authors cited above are right to develop an account of hate speech's harm in a way that is responsive to empirical research in this area. And for reasons that I will present and discuss in §3, these authors are also right to focus on hate speech's impact on children. It is worth doing a little more background theoretical work, however, to clarify exactly what is needed from an account of hate speech's harm, if it is going to be able to provide an in-principle justification for BANS. Having clarified these criteria, we can use them to identify some of the advantages of emphasizing hate speech's effect on children.

In this I will be treating the harm principle as a necessary (but not sufficient) condition on the permissible prohibition of speech. The state cannot use the coercive apparatus of the law to prohibit speech if its aim is to edify the speaker, or penalize bad speech *per se*, or send a message condemning certain speech. If the state is going to prohibit hate speech, it owes us (and the speaker) a rationale based on the ultimate aim of preventing harms to others.[22]

---

[21] In a similar vein to the above, Brown surveys a range of claims made about the harms of hate speech, including in critical race theory, and finds that in several cases authors cite empirical studies which they say show that hate speech harms its targets, when in fact the cited studies show that 'discriminatory treatment' is the relevant causal factor, and don't license the inference that hate speech in particular (as a specific form of discriminatory treatment), is responsible for the relevant harms; see *Hate Speech Law: A Philosophical Examination* (New York: Routledge, 2015), pp. 56-58. Further critical discussion can be found in Heinze (see *Hate Speech and Democratic Citizenship*, pp. 125-29) of how legal theoretic work in this area represents the findings of empirical research on the effects of hate speech.

[22] Sunstein suggests that the prohibition of hate speech might be defended in terms of its 'expressive' function, the idea being that such laws send a message condemning the attitudes of the hate speaker; Cass R. Sunstein, 'On the expressive function of the law', *University of Pennsylvania Law Review* 144(5) (1996): 2021-53. In assuming the harm principle I am ruling out any such justification for BANS.

How might one substantiate the claim, then, that all instances of hate speech are harmful in a way that suffices to justify BANS – even hate speech that isn't used to threaten, harass, or incite violence, and thus isn't directly responsible for any particular harm being done to any particular individual? The kind of claim one must defend here is one that says all instances of public hate speech make a contribution to some general state of affairs that *is* the cause of concrete harms to particular individuals. Feinberg proposes a theoretical framework to distinguish this kind of environmentally-mediated harm from harm that is directly inflicted by particular acts. Where a 'private harm principle' only allows prohibitions on directly harmful acts, he says, a 'public harm principle' would permit prohibitions on acts whose restriction is necessary "to prevent impairment of institutional practices and regulatory systems that are in the public interest".[23] Tax evasion and contempt of court are Feinberg's paradigmatic examples of acts whose harmfulness is more aptly characterized by way of this public harm principle. Even though isolated instances of tax evasion don't directly harm anyone, they are still genuinely harmful, he says, "insofar as they weaken public institutions in whose health we all have a stake".[24]

Joshua Cohen describes how environmentally-mediated harms may be effected by speech in particular. Speech, he says, "may help to constitute a degraded, sickening, embarrassing, humiliating... or demeaning environment", and when this is the case, although we cannot "trace particular harmful or injurious consequences to particular acts of expression that… constitute the unfavorable environment", we can judge that individual speech acts are contributing to the social environment's degradation, and that specifiable harms are resulting from this.[25] This is the kind of account of hate speech's harm that one needs to provide in order to defend BANS: one on which all instances of hate speech are contributing to a social environment that harms people in targeted groups, or that weakens institutions protecting their interests.

The type of environmentally-mediated harms that I am especially concerned with, as I explained in §1, are those constituted by occupying a subordinate position in an identity-based social hierarchy, e.g. those that come with systematic disadvantages in resources, labor conditions, or social opportunities. Another class of environmentally-mediated harms that one could focus on would be those fostered by a "climate of hatred" towards a group, "associated with an increased chance of acts of discrimination, violence,

[23] Joel Feinberg, *Social Philosophy* (Englewood Cliffs: Prentice-Hall, 1973), p. 25.

[24] See *ibid*: 25-26.

[25] Joshua Cohen, 'Freedom of expression', *Philosophy & Public Affairs* 22(3) (1993): 207-63, 231.

[and] damage to property".[26] Granted, the two kinds of harms may be causally interrelated, insofar as climates of hatred can be produced by *de facto* social hierarchies, and can reinforce those hierarchies in turn.[27] But there is a particular set of complexities associated with harms borne of a climate of hatred that I don't want my account to inherit – specifically, complexities in how we conceive of the relations of causation and responsibility that obtain between a speaker fueling a climate of hatred, and someone under that climate performing an act of violence or discrimination against a particular victim.[28] By contrast, the harms generated by social hierarchy *per se*, are harms for which, typically, there is *no* actor whose conduct is the proximate cause of the harm – in other words, they are harms that are necessarily conceived of as *structural* rather than agential.[29] The kind of account of hate speech's harm whose prospects I want to focus on, then, is one where the hate speaker is culpable for causally contributing to a social order in which such structural harms are effected.

2.3 Evidential support

A harm-based rationale for BANS should be backed by evidence that supports its claims about the causal processes through which hate speech contributes to this kind of social order. Even if it isn't the legal theorist's job to supply the data, she should not be indifferent to whether and how evidence may be adduced in support of her conjectures. After all, as discussed in §1, there are other credible hypotheses about the

---

[26] See Brown, *Hate Speech Law*, p. 67.

[27] E.g. see Cecilia L. Ridgeway's account of how, in a social order where one group gains a positional advantage over another, recognition of the initially accidental disparity will be transformed over time into a set of beliefs about the inferiority of the disadvantaged group; 'The emergence of status beliefs: from structural inequality to legitimizing ideology', in J. T. Jost and B. Major (eds.), *The Psychology of Legitimacy: Emerging Perspectives on Ideology, Justice, and Inter-group Relations* (Cambridge: Cambridge University Press, 2001), pp. 257-77.

[28] Specifically, as Brown says, to substantiate this kind of rationale we won't just need an evidentially-supported account of how hate speech contributes to the degraded social environment, but also evidence that the 'climate of hatred' makes proximately harmful acts (e.g. violence) against members of target groups *likely* and *imminent*. There are reasons to doubt that the climate-to-act causal pathways operate so straightforwardly, and practical difficulties in conducting studies that would demonstrate these pathways if they were in effect; Brown, *Hate Speech Law*, 68-70.

[29] In structurally harmful social hierarchies, as Iris Marion Young puts it, "in most cases it is not possible to trace which specific actions of which specific agents cause which specific parts of the structural… outcomes"; *Political Responsibility and Structural Injustice* (Kansas: The University of Kansas, 2003), p. 7. The idea isn't that the harms are ultimately done to some amorphous entity, e.g. to society *per se*, or to a social group abstractly conceived. The environmentally-mediated harms regulated by a public harm principle still redound to individuals. What's distinctive about them (and prevents their regulation under a *private* harm principle) is that there is no direct causal link between perpetrator and victim. Individual actors contribute to system-effects, and people's interests are wrongfully setback by those system-effects, but in a way such that we typically cannot attribute specific effects to specific actors.

principal causal forces behind *de facto* social hierarchy. It is easy to imagine hate speech as the culprit, because it is the conspicuous facade of identity-prejudice. It is also an expressive practice that is often the province of low-status speakers, who are less skilled than elites in finessing their expression to avoid the infringement of mainstream expressive customs. We can see that deep, structural changes in employment and social mobility would have a major impact on identity-based social hierarchies. But these reforms are hard to achieve. Restricting hate speech is easier, partly due to the limited social capital of the people BANS typically penalize. But BANS would be patently illegitimate if they were essentially an exercise in expedient scapegoating. For all these reasons, advocates of BANS should not be content with a plausible-sounding just-so story about how hate speech is playing a role in the perpetuation of social hierarchy. They should want their story to be backed up by evidence.[30]

There are ways of conceptualizing the harm that hate speech inflicts that can notionally sidestep this demand for empirical support. Some authors appeal to conceptions of harm on which hate speech doesn't cause, but rather *constitutes*, a harm to its targets. While there is a case to be made for this approach, it also has limitations, insofar as it makes the truth of claims about hate speech's harm primarily hinge on esoteric normative and social-theoretic theses, which lie outside the sphere of empirical arbitration. Complex problems of legitimacy arise if our justification for BANS appeals to an understanding of harm-infliction which rests on philosophical conjectures that many in the polity reject (or would reject, if the question arose). By contrast, if our claims about the harmfulness of hate speech are based on the results of the application of widely-accepted methods for assessing the impact of different factors on people's welfare and interests (defined in terms that are standard to mature social scientific disciplines), then we will have an especially robust justification for BANS – the kind that *anyone* should recognize as legitimate, in principle, on pain of irrationality or general scepticism. The call for evidential support, then, in debates about BANS, is at least in part about seeing whether this decisive type of justification for BANS is

[30] There is a considerable, but disciplinarily disparate, body of empirical research investigating the effects of hate speech, and it is unclear what sort of confident conclusions can be drawn from it (if any) about hate speech's distinctive role in perpetuating identity-based social hierarchies. One important recent cluster of papers on this topic comes from a collaborative project between law and political theory on the effects of hate speech and its regulation in Australia; see Katharine Gelber and Luke McNamara, 'Freedom of speech and racial vilification in Australia: 'the Bolt case' in public discourse', *Australian Journal of Political Science* 48(4) (2013): 470-84; Katharine Gelber and Luke McNamara, 'The effects of civil hate speech laws: lessons from Australia', *Law & Society Review* 49(3) (2015): 631-64; and Katharine Gelber and Luke McNamara, 'Evidencing the harms of hate speech', *Social Identities* 22(3) (2016): 324-41. The most comprehensive synthesis and analysis of published empirical research relating to the harm of hate speech, in the recent legal theoretic literature, is in Brown, *Hate Speech Law* (see in particular footnote 28).

in the offing. This need not be motivated by the prioritization of expressive liberty above all other considerations, as some authors suggest.[31]

One might argue that the burden of proof should be reversed, such that *opponents* of BANS have to provide evidence for the view that hate speech *doesn't* harm or endanger its targets.[32] This sort of precautionary approach gains *prima facie* plausibility from the historical record of cases where hate speech seems to have helped to fuel genocidal movements. If we have reason to think that hate speech can contribute to catastrophic harms, in contexts where identity-prejudice gives way to murderous atrocities, then we arguably also have reason to think it can contribute to the routine harms associated with identity-based hierarchies in relatively stable societies.[33] On the other hand, there are several difficulties with the appeal to precautionary justifications in this area. Precautionary laws don't eliminate risk as such, they just trade one set of risks (the risks of legal inaction) for another set.[34] And in a liberal legal system in particular, there will be a presumptive opposition to precautionary laws that infringe against basic civil rights, insofar as such laws themselves run the risk of allowing specious assertions about imminent dangers to erode the rule of law and usher in authoritarianism.[35] At any rate, whatever the best general defense of precautionary principles might amount to, a *preventive* rationale – that adverts to the imminent aim of redressing the empirically-demonstrated harms of hate speech – gives us a stronger case for BANS than a precautionary

---

[31] E.g. Waldron, *The Harm in Hate Speech*, 148.

[32] Brown indicates some sympathy for a precautionary approach, e.g. he says an authority may impose restrictions on certain instances of hate speech, "because having identified the possibility… that a proportion of the individuals targeted by hate speech will not participate in the formation of public opinion, and bearing in mind the conditions of uncertainty that surround these outcomes, it errs on the side of precaution"; Brown, *Hate Speech Law*, 199.

[33] For discussion of hate speech's role in genocidal social movements, see Lynne Tirrell, 'Genocidal language games', in I. Maitra and M. K. McGowan (eds.), *Speech and Harm: Controversies over Free Speech* (Oxford: Oxford University Press, 2012), pp. 174-221. For discussion of how, even in stable democratic societies, hate speech may contribute to a kind of 'slow-burn' incitement of anti-democratic movements against marginalized groups; see e.g. Tsesis, 'Dignity and speech'. Whether we can learn something about hate speech's likely impact in stable democracies, from observing its involvement in genocidal movements elsewhere, is a complex question in its own right. One of Heinze's core theses in *Hate Speech and Democratic Citizenship* is that we cannot make ready inferences across this divide.

[34] On this point see Cass R. Sunstein, *Laws of Fear: Beyond the Precautionary Principle* (Cambridge: Cambridge University Press, 2005). For a critical discussion of how precautionary rationales were employed in Western democracies to legitimize 'the war on terror', and the semi-permanent shift to a mode of governance based on 'extraordinary emergency', see Claudia Aradau and Rens Van Munster, 'Governing terrorism through risk: taking precautions, (un)knowing the future', *European Journal of International Relations* 13(1) (2007): 89-115.

[35] Just as historical cases can be cited to emphasize the risks of legal inaction, they can also be cited to identify risks associated with infringing civil rights *for the sake of* addressing first-order risks, e.g. see Geoffrey R. Stone, *Perilous Times: Free Speech in Wartime from the Sedition Act of 1798 to the War on Terrorism* (New York: W. W. Norton, 2004).

rationale. Once again, the aim here is to explore whether that kind of particularly robust justification for BANS is in the offing.

2.4 Responsibility for harm

If one is seeking to defend BANS, the story one tells about the harm done by hate speech needs to be one on which the hate speaker can be ascribed responsibility for the harm. An example will help to convey what I have in mind. Suppose someone were to argue as follows.

> Racial hierarchies, in-groups and out-groups, us and them: our penchant for social taxonomies and rankings reflects the structure of language itself. Identity-prejudice is not due to particular speech acts, but to the underlying grammars and vocabularies that frame all verbal communication. Language prefigures the discriminations that define social cognition, opening us up to some people and closing us off to others.[36]

This picture identifies speech as responsible for causally contributing to *de facto* social hierarchy, but it does so in a way which suggests that social hierarchy cannot be meaningfully combatted by trying to single out and counteract the effect of any particular speech acts. The problem with this picture is not the fact that it sees speech as contributing to harms that are structural, indirect, or environmentally-mediated. As I explained in §2.2, this is precisely the type of account of hate speech's harm that we are looking to develop. The problem here, rather, in the characterization of the mechanism through which communication is harmful, is that it doesn't enable us to discriminate between instances of communication that are contributing to harm, and instances whose effects are benign or positive. On this picture all language-users collaborate in sustaining social hierarchy, and there is little any of us can do to resist this.

In order to underwrite a credible justification for BANS, an account of hate speech's contribution to the structural harms of social hierarchy cannot have this form. If we are going to punish individual hate speakers, we need reason to think that they are responsible for making a distinctive contribution to the harms of social hierarchy, one that is different from – and more important than – the contribution made by the listener or by the citizenry at large. There is little justificatory payoff in an account that ultimately

---

[36] Although I'm not attributing this sketch of a position to anyone, the kind of linguistic constructivism that underpins it appears in some of Charles Taylor's work; e.g. see 'Theories of meaning', in *Human Agency and Language: Philosophical Papers 1* (Cambridge: Cambridge University Press, 1985), pp. 248-92, 263.

depicts hate speakers as flotsam and jetsam, drifting around in a sea of deeper forces that are the real drivers of social inequality.

We also need our account to support the notion that, where the hate speaker contributes to harmful outcomes by *influencing other people*, he can still be rightfully conceived of as responsible for contributing to the relevant harms. At the same time, however, the account we give must not have the upshot that in *any* context in which B is influenced by A into φ-ing, A can be deemed responsible for contributing to whatever ensues from B's φ-ing. If our account of responsibility for contributory harm was that inclusive – in societies like ours, which have complex, multidirectional networks of cross-cutting influences – it would too easily break down into an implausible picture, on which everyone who speaks in public ends up being partly responsible for contributing to a vast range of downstream consequences. In short, in seeking evidence of the hate speaker's contribution to a harmful social order, we need to be able to understand the speaker's responsibility for that contribution in a way such that he is at least partly accountable for actions performed by the people he influences, but without making it the case that responsibility for harms which result from influencing others comes too cheaply.

3. Hate speech and children: evidence and responsibility

I have argued that an account of hate speech's harm must explain how all instances of hate speech contribute to a harmful state of affairs, and be amenable to empirical confirmation of its claims about these effects, *and* show how hate speakers can be understood as responsible for the relevant harms. In this section I explain the advantages of focusing on children in developing an account of hate speech's harms that has the potential to satisfy these conditions.

3.1 The legitimation and normalization hypothesis

As I explained at the outset, in §1, our hypothesis about how hate speech contributes to unjust *de facto* social hierarchies, and their resultant harms for members of targeted groups, must not downplay historical and material factors. Our account becomes speculative past the point of credibility if it suggests that words can conjure up social realities *ex nihilo*. With this constraint in mind, the most credible type of hypothesis about hate speech's contribution to social hierarchy is one which posits that hate speech *legitimates* and *normalizes* social hierarchies.

Several authors appeal to something like this in their discussion of the relationship between speech and social hierarchy. In Matsuda's ground-breaking work on hate speech, she says the power of racist

groups "derives from their offering legitimation and justification for otherwise socially unacceptable emotions of hate, fear, and aggression".[37] Parekh says identity-based social hierarchy is "legitimized by a wider moral climate which is built up and sustained by… gratuitously disparaging and offensive remarks".[38] Among MacKinnon's charges against pornography, she says that it "authorizes and legitimizes" sexual abuse.[39] And Langton echoes this, arguing that although pornographers lack formal authority, they legitimate discrimination against women by representing women's subordinate social position as "ordinary and normal".[40] If we are employing this critical vocabulary in order to describe hate speech's contribution to a harmful social structure, then an initial version of our hypothesis might be stated as follows.

> The LAN (i.e. Legitimation and Normalization) Hypothesis: Hate speech causally contributes to the harms of *de facto* social hierarchies by legitimating and normalizing systematic material and institutional inequalities that track social identity categories.

What exactly does it mean to say that hate speech legitimates and normalizes material and institutional inequalities? In particular, how are social facts about what is legitimate affected by hate speech, if most hate speech comes from people who lack any formal authority to impose norms for others? One possibility, proposed by Ishani Maitra, is that the hate speaker can acquire 'licensed authority'. Like the person who takes charge in a chaotic social situation and finds that others fall into line behind his leadership, the hate speaker can assume a kind of *de facto* authority, due to his contingent situational influence rather than

---

[37] Mari J. Matsuda, 'Public response to racist speech: considering the victim's story', *Michigan Law Review* 87 (1989): 2320-81, 2378.

[38] Bhikhu Parekh, *Rethinking Multiculturalism: Cultural Diversity and Political Theory* (New York: Palgrave MacMillan, 2006), p. 314.

[39] Catharine A. MacKinnon, 'Francis Biddle's sister: pornography, civil rights, and speech', in *Feminism Unmodified: Discourses on Life and Law* (Cambridge Massachusetts: Harvard University Press, 1987), pp. 163-97, 171.

[40] Rae Langton, 'Subordination, silence, and pornography's authority', in R. C. Post (ed.), *Censorship and Silencing: Practices of Cultural Regulation* (Los Angeles: Getty Research Institute for the History of Art and the Humanities, 1998), pp. 261-84, 269. Langton's work in this area is perhaps best-known for the way that it uses Austin's speech act theory to explicate the claims that pornography subordinates and silences women. But another integral element of her work is its development of the concepts of legitimation and normalization, and the relations between them. As well as the above source, these elements are at work in her 2015 Locke Lectures on 'Accommodating Injustice' (see philosophy.ox.ac.uk/john-locke-lectures), and in other works of hers cited here, including 'Speech acts and unspeakable acts', 'Is pornography like the law?', and 'The authority of hate speech'. One aspect of Langton's development of these concepts is her emphasis on epistemic authority. The makers of pornography don't *accidentally* succeed in normalizing a picture of women as objects, and in causing people's beliefs to reflect that picture. Rather, she argues, pornography shapes the world such that makes it (partly) true that women *are* what pornography represents them as; pornographers normalize and legitimate women's subordination, by expressing the epistemic authority they have as *architects* of a patriarchal social order, and authoritatively transmitting knowledge of their design to others. This aspect of her view is the focus in 'Speaker's freedom and maker's knowledge', in *Sexual Solipsism*, pp. 289-310.

any recognized positional authority.[41] It is unclear, though, whether it would make sense to say that hate speech 'legitimates' social hierarchy, if the hate speaker's 'authority' to influence the social facts about what is legitimate is reliant upon other people's voluntarily acceding to his leadership.[42] In light of this concern, we can see why it makes sense to conceive of legitimation and normalization as complementary processes. We are all subject to powerful *de facto* social norms that enjoin us to act in accordance with whatever practices and behaviors we understand to be descriptively normal in the context where we are acting.[43] The conjecture, then, will be that when a person engages in public hate speech, even if they do not possess any formal political authority, they can represent the subordinate status of their targets as being (descriptively) normal, and in so doing they can give identity-based *de facto* social hierarchies the appearance of (normative) legitimacy.

To be clear, in stating that the LAN Hypothesis is a credible one, I am not leaping to the conclusion that it is true. It remains for the hypothesis to be assessed in light of relevant data, and as I will explain in the next section, certain challenges are likely to arise in trying to acquire evidence that demonstrates the specific process that the hypothesis describes. Still, the hypothesis has two important features to recommend it. First, it assigns a distinct role to hate speech in sustaining *de facto* social hierarchy, but – crucially – without denying the primacy of the historical-material forces that underpin racial social structures, patriarchy, and other identity-based social hierarchies. I expect that almost no-one who is seriously engaged in these debates actually believes that hate speech summons racism or heteronormativity into existence out of thin air. But it is easy to add rhetorical flourishes in describing the causal powers of com-

---

[41] Ishani Maitra, 'Subordinating speech', in I. Maitra and M. K. McGowan (eds.), *Speech and Harm: Controversies over Free Speech* (Oxford: Oxford University Press, 2012), pp. 94-120. Note that McGowan offers another kind of account of how social facts about what is legitimate can be altered by regular, low-status speakers, based on the idea that speakers can verbally enact 'conversational exercitives' which (in the first instance) alter what is proper and improper conduct within the particular conversation in which they are performed; e.g. see 'Oppressive speech'. But it is a further question whether this kind of account can be extended to explain how low-status speakers can alter legitimacy facts in a further-reaching way, which affects the whole hierarchical ordering of a society.

[42] Given this account of what is occurring, a more fitting characterization might be to say that the social hierarchy is being mutually enacted by speaker and audience together. For further discussion of these kinds of cases, see Saray Ayala and Nadya Vasilyeva, 'Responsibility for silence', forthcoming in *Journal of Social Philosophy*.

[43] See footnote 45, below.

munication and, in so doing, downplay historical-material factors. It is easy to say things like "words create the hierarchies and people fill them",[44] which are meant to highlight the harmful potential of communication, but which seemingly do this by attributing magical powers to speech. The LAN hypothesis doesn't make the influence of hate speech a rival explanatory hypothesis to one that emphasizes the historical-material bases of social hierarchy. Rather, it posits that hate speech plays a key role in cementing the conditions that historical-material forces set in place.

The second advantageous feature of the LAN Hypothesis is that it posits the operation of a phenomenon that is widely recognized and which has been empirically observed, i.e. the phenomenon of normalization. Whether the phenomenon is actually in effect in this particular context – whether hate speech does in fact normalize identity-based social hierarchies – is a further empirical question. But the phenomenon of normalization itself isn't merely some speculative or conjectural critical concept. It is an empirically observed phenomenon, of interest to researchers in a number of empirical disciplines, including psychology and sociology.[45]

Given that I defined hate speech in terms of disdain and contempt, one might note that the mere fact of someone expressing 'disdain' for a group wouldn't necessarily represent that group's subordination as either descriptively normal or normatively legitimate – at least, not for all values of 'disdain'. If we are proposing that hate speech legitimates social hierarchy, then, we need to interpret 'disdain' and 'contempt'

---

[44] Shannon Gilreath, "'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) (2009): 557-615, 604.

[45] For starters, there is a large body of research on how norm-abiding behavior can be sustained by people attending to environmental cues that indicate other people's norm-adherence. One of the seminal papers on this topic is Robert B. Cialdini, Raymond R. Reno, and Carl A. Kallgren, 'A focus theory of normative conduct: recycling the concept of norms to reduce littering in public places', *Journal of Personality and Social Psychology* 58(6) (1990): 1015-26. These findings are part of the background for contemporary philosophical work on social norms, e.g. Cristina Bicchieri, *The Grammar of Society: The Nature and Dynamics of Social Norms* (Cambridge: Cambridge University Press, 2006). There is also research in social psychology that indicates a connection between people's understanding of what is normal and legitimate, specifically with respect to social inequality; see Rui Costa-Lopes, John F. Dovidio, Cícero Roberto Pereira, and John T. Jost, 'Social psychological perspectives on the legitimation of social inequality: past, present and future', *European Journal of Social Psychology* 43(4) (2013): 229-37, 230. Recent work by Adam Bear and Joshua Knobe offers further support for the notion that things can be made to seem legitimate by being represented as normal. Their key finding is that when people are asked what they believe are the 'normal', 'ideal', and 'average' quantities of a variable (e.g. what fraction of students cheat on exams) judgements about what is *normal* deviate from judgements about what is descriptively *typical*, in a way that is influenced by judgements about what is *ideal*. Judgements about what is normal and what is ideal are not independent, then – they interact and influence each other; see 'Normality: part descriptive, part Prescriptive", *Cognition* 167(1) (2017): 25-37. Some studies suggest that children are especially disposed to form views about proper behavior based on observations of what's descriptively typical, e.g. Marco F. H. Schmidt, Hannes Rakoczy, and Michael Tomasello, 'Young children attribute normativity to novel actions without pedagogy or normative language', *Developmental Science* 14(3) (2011): 530-39.

for x as meaning something like 'seeing x as worthy of a subordinate position'.[46] All liberal democracies subscribe to some form of the doctrine that "human beings are born free and equal in dignity and rights".[47] We encode this in our legal systems in various ways, and in most liberal countries this enjoys majoritarian support. Any form of discriminatory, identity-based mistreatment should be straightforwardly recognizable as unjust, in this kind of formally egalitarian social milieu. But hate speech impairs the recognition of this, by making salient to its audience a representation of its targets as second-class beings, who are *quite rightly* assigned a subordinate social position. And all instances of public hate speech make a contribution to the salience of these derogatory group representations. In this way, so a proponent of the LAN Hypothesis would suggest, hate speech helps people to see the disadvantages faced by the target group as normal, natural, and legitimate, and it thus deters efforts at reforming the wider structural hierarchies that generate these disadvantages.

3.2 Evidential support, children, and the LAN hypothesis

As discussed in §2.3, a harm-based rationale for BANS should be backed by evidence that supports its claims about the causal processes through which hate speech contributes to the harms of social hierarchy. The natural question, then, is what evidence can be adduced in support of the LAN Hypothesis? However, for reasons that will become evident shortly, we actually need to delve into the adjacent – more theoretical – question, of what evidence *could* be unearthed and adduced in support this hypothesis, if it were in fact true. I argued that the LAN Hypothesis is credible because it is compatible with the highly plausible assumption that material and institutional factors have causal primacy in the creation and perpetuation of *de facto* social hierarchies. The LAN Hypothesis sees hate speech's role as *bolstering* those hierarchies, by shaping people's attitudes in a way that favors them. But this picture of how the causal factors work together creates difficulties if we are trying to evidentially demonstrate hate speech's effects. If hate speech's influence follows on from (and interacts with) other more fundamental causal forces, then its distinctive effects will, in the normal run of cases, be difficult to isolate and detect.

Consider the everyday bigot, A, who mostly keeps her prejudiced views to herself, but is regularly exposed to hate speech in her daily life. We want to see whether there is any evidence that this exposure

---

[46] Some authors make broadly similar points, about how the form of the 'contempt' that is conveyed in hate speech should be understood not in terms of a speaker's negative emotion toward the target group, but rather, in terms of the their speech's ascription of a diminished status to the target, e.g. Waldron, *The Harm in* Hate Speech, 34-37.

[47] Universal Declaration of Human Rights (1948), Article 1.

contributes to A's view of the racial inequalities in her society as normal and legitimate, as the LAN Hypothesis claims. But there is an important rival hypothesis in the background. A's entire life has been spent in a society ordered by innumerable forms of racial inequality. White people dominate the upper ranks in politics, business, law, academia, the arts, and the military. Among the various ways in which they are socially outranked by white people, black people generally achieve worse outcomes in education and in other proxies of intellectual ability. The complex historical and institutional forces that explain these patterns are beyond A's comprehension, and the persistence of this social order confers upon it the appearance of naturalness. By applying simplistic explanatory heuristics, A comes to believe that the best explanation for the inequalities that she observes and experiences in her society is one that attributes some kind of general inferiority in intellectual capacities to black people.

We can generalize from the uncertainty that this rival hypothesis creates. Any study that aims to gauge the influence of hate speech on adult subjects will have to examine individuals for whom this rival hypothesis would be a *prima facie* plausible explanation of how they came to regard identity-based social hierarchies as normal and legitimate. In order to control for the factors that are emphasized in this rival explanation, while gauging the influence of hate speech, we would have to screen out the conditioning influence of a whole life spent in the shadow of inequality. It may be relatively easy to devise studies to test whether behavioral *manifestations* of identity-prejudice can be activated through exposure to hate speech. But the evidencing of this effect wouldn't demonstrate that hate speech is essentially involved in the formation of identity-prejudice. When examining adult subjects who regard racial inequalities as normal and legitimate, there are many other factors besides hate speech that could plausibly be causally responsible for this, such that it is going to be hard to acquire clear evidential support for any hypothesis that purports to isolate the distinctive contribution of hate speech. The confounding factors to be controlled for are too many, and too much enmeshed with people's everyday experiences of living in societies like ours, to simply screen off. But now: consider a modified version of the LAN Hypothesis, which adverts to hate speech's influence on children in particular.

> The CLAN (i.e. Childhood Legitimation and Normalization) Hypothesis: Hate speech causally contributes to the harms of *de facto* social hierarchies by influencing children's attitudes in a way that legitimates and normalizes systematic material and institutional inequalities that track social identity categories.

Studies that could find evidence in support of the CLAN Hypothesis will be easier to devise and execute, in comparison to the LAN Hypothesis. It is verging on impossible to find adult experimental subjects who

have been insulated from the wider social conditions that might lead one to think of *de facto* social hierarchies as normal, simply due to living a society in which they *are* normal. It will be easier to find children who have been insulated from their society's overall conditions like this. Especially at pre-school ages, some children live relatively cloistered lives, in which they don't see the material and institutional elements of identity-based social inequality, or indeed, in which they don't even encounter people from other social groups. Obviously this isn't true of all children. But it is true of some children, and they may in principle become subjects for studies aiming to isolate the influence of hate speech on people's attitudes.

For example, suppose we take a four year-old, C_4, living in an ethnically homogenous community, and under appropriately controlled conditions, we expose her to examples of hate speech against a social group, G, that she and her family have no interaction with, and know little or nothing about. Suppose we then find, in follow-up tests weeks or months later, that C_4 starts manifesting a pattern of negative attitudes towards members of G, e.g. she shows less distress at the mistreatment of Gs compared to members of other groups. In trying to explain this finding, there would be no reason to wonder whether C_4's anti-G attitudes could be explained in terms of her attempts to independently interpret the patterns of power and disadvantage that she has been observing in a social system where G's are structurally subordinated. Rather, we would attribute the change in C_4's anti-G attitudes to the influence of hate speech, because there would be no other good explanation as to what altered her attitudes. If an accumulation of this kind of evidence were to indicate that children take on identity-prejudicial attitudes as a result of exposure to hate speech, this would provide evidence to support the hypothesis that hate speech makes a distinct causal contribution to the prevalence of attitudes, either conscious or unconscious, about the legitimacy and normality of identity-based social hierarchies.

The point that I am making here isn't premised on the implausible claim that in general, or in typical cases, hate speech is the *main* factor (or the *only* factor) which influences children towards accepting the normality or legitimacy of identity-based inequality. For one thing, some children will absorb identity-prejudice through exposure to relatively subtle manifestations of it in their family members, in words and deeds. And we should also allow that in most cases in which hate speech *does* play a key role in inculcating prejudice in a particular child, the child's attitudes, as she matures, will typically be reinforced by other factors. More generally, it seems plausible to suppose that for (most) children who acquire identity-prejudicial attitudes, hate speech will just be one contributing factor among others in this. My point here is about what kinds of things it is possible to learn from observing hate speech's effects on particular child subjects, which it is not possible (to all practical purposes) to learn from observing hate speech's effects on particular adult subjects. Individual child subjects are sometimes insulated, in a way that adults cannot

be, from the other confounding causal factors that can influence people towards accepting the normality or legitimacy of identity-based inequalities. Because of this it will be easier with children, than with adults, to acquire evidence of any distinct influence that hate speech does have in normalizing and legitimating identity-based social inequalities.[48]

3.3 Assigning responsibility to the hate speaker

Here is a second reason why the modified CLAN Hypothesis is a better foundation on which to build a harm-based justification for BANS. Set aside my main contention in the previous section, for the sake of argument, and suppose we acquired evidence that provided a similar degree of empirical support for both the LAN *and* CLAN Hypotheses. With regards to the CLAN Hypothesis, we would have no reason to refrain from ascribing the hate speaker responsibility for the identity-prejudicial attitudes inculcated by her speech. If an adult influences a child's attitudes, the adult bears responsibility for this influence – both *moral* culpability, i.e. liability to be blamed, and *legal* responsibility, i.e. accountability for resultant harm – if anyone does. Of course there are complications and caveats around this, but in general, children are not responsible for being influenced by adults towards attitudes that result in harmful outcomes.

By contrast, when it comes to the LAN Hypothesis, the problem of how responsibility should be ascribed is much more complicated. And this is because adults are, outside of rare cases, like brainwashing, responsible – that is to say, culpable, and where practical stakes are involved, accountable – for their attitudes, even when those attitudes causally stem from the influence of other people. If person A's communication has an influence on B's attitudes, and if B is a responsible agent, then in the normal run of cases, B – and not A – is culpable and accountable for bad consequences that result from B's attitudes.[49]

---

[48] We can imagine cases in which a child is exposed to speech that expresses contempt for groups that *don't* occupy a subordinate position in a *de facto* hierarchy, e.g. in a hypothetical egalitarian social order where *no* groups occupy such a position, or in speech that is contemptuous towards privileged groups. In such cases, the speech's influence on the child wouldn't contribute to the kind of structural harms that I have been emphasizing. But it may still contribute to harmful outcomes, e.g. by influencing the child towards performing harmful acts, or by undermining the child's own self-respect. The fact that we are focusing on hate speech's contribution to structural harms, and how this might be involved in making a case for BANS, is consistent with thinking we might have other kinds of harm-based justifications for regulating hate speech in these other kinds of cases.

[49] The idea here is that if you're a normal adult, "you have a mind of your own", and are thus normally responsible for how you respond to other people's influence; see Thomas Nagel, 'Personal rights and public space', *Philosophy & Public Affairs* 24(2) (1995): 83-107, 96. This idea is integral to several influential accounts of the grounds of the right to free speech (e.g. Thomas Scanlon, 'A theory of freedom of expression', *Philosophy & Public Affairs* 1(2) (1972): 204-26); it is fully compatible with recognizing exceptional cases, e.g. of provocation, in which normal adults have diminished responsibility for how they are influenced by others (see L. W. Sumner, 'Incitement and the regulation of hate speech in Canada: a philosophical analysis', in I. Hare and J. Weinstein (Eds), *Extreme Speech and Democracy* (Oxford: Oxford University Press, 2009), pp. 204-20, 215ff); and there are reasons to think its

Again, this is not the case with children. Children don't bear this kind of general responsibility for their mental lives and how they respond to the influence of others: at least, not in the same range of cases, nor to the same degree as adults. This is what we standardly suppose, at any rate, both in our informal ethical blaming practices, and as a matter of legal doctrine.[50] All of this is consistent with the point from §2.2, that A's contribution to an aggregate harm, x, suffices in principle to justify us in holding A legally accountable for x. The point being made here is that there is an exception to this general thesis about when we can ascribe responsibility to people whose conduct contributes to aggregate harms. If the mechanism via which A makes her contribution to x is through communicative acts that influence another responsible agent, B, to behave in ways that lead to x, then A's contribution to x *isn't* sufficient to justify holding A accountable for x.[51]

One might worry that judgements about causation and culpability are being run together in what I'm saying. Our interest in the causal origins of social hierarchy isn't only about who can be blamed. We also want to understand the causal processes at work, independently of who can be held accountable for them. Still, the disparities in responsibility between adults and children aren't only relevant here with regards to questions of assigning blame. They also have a bearing on how we characterize the causal character of these modes of influence.

Imagine a case in which a teacher is indoctrinating his class of primary-school aged children. Given the cognitive disparities, the children will have limited ability to resist the teacher's influence, or to influence his attitudes in turn. Because of this it is plausible to characterize the inculcation of attitudes in these children as a matter of a certain *agent*, the teacher, acting upon a group of *patients*, the children. By contrast, in a situation where a community of agents with broadly comparable cognitive abilities are engaged in ongoing communication with each other, in a multidirectional network of cross-cutting influences, it generally isn't plausible to characterize this as a process of certain agents acting upon certain patients. The

---

structuring role in free speech theory can be retained despite the general limitations in our control over our cognition (see Robert Mark Simpson, 'Intellectual agency and responsibility for belief in free speech theory', *Legal Theory* 19(3) (2013): 307-30).

[50] Here is how the point was stated in a landmark contemporary U.S. Supreme Court case addressing issues around the capital punishment of adolescents. "Developments in psychology and brain science continue to show fundamental differences between juvenile and adult minds. For example, parts of the brain involved in behavior control continue to mature through late adolescence… [Juveniles'] actions are less likely to be evidence of "irretrievably depraved character" than are the actions of adults". See *Roper v. Simmons*, 543 U.S. 551 (2005), at 570.

[51] Evan Simpson presents an analysis of the case for regulating hate speech which emphasizes the responsibilities of listeners. On his view we can (or should be able to) expect listeners to be reasonable in how they respond to the influence of hate speakers, and, roughly, this consideration problematizes most kinds of legal regulation of hate speech; see 'Responsibilities for hateful speech', *Legal Theory* 12(2) (2006): 157-77.

correspondence between this and the two contrasting versions of our hypotheses should be clear. The CLAN hypothesis represents the harmful effects of hate speech in a manner such that if the hypothesis were substantiated, the hate speaker could straightforwardly be ascribed responsibility for causally contributing to the relevant harms. Whereas the LAN hypothesis does not. Given how the LAN hypothesis represents hate speech's contribution to the socially-mediated harm, an attempt to pin responsibility for this harm onto the hate speaker will lead to an implausible over-attribution of responsibility, i.e. to nearly everyone involved in the wider social ecosystem, or else it will involve some sort of *ad hoc* confinement of responsibility to the hate speaker alone.

These sorts of distinctions matter in the critical analysis of social hierarchy. There are differences between acts of subordination performed by particular agents, and processes of subordination that are structural – differences in the underlying causal mechanisms, and in how they can be counteracted.[52] In the way that they try to counteract identity-based social hierarchy, advocates of BANS aren't just recognizing and resisting structural injustices. They are trying to pinpoint and counteract a *specific contribution* to structural injustice, for which hate speech is allegedly responsible. If we are seeking to vindicate that critical project, the CLAN hypothesis presents us with a more viable characterization, than the LAN hypothesis, of the causal processes involved in the perpetuation of unjust social hierarchy, and of hate speech's role in this.

4. Summary and policy implications

Many progressives believe that communicative factors contribute to *de facto* social hierarchy, and that hate speech plays an important role in this, in a way that can justify BANS, at least in principle. In order to substantiate these convictions and defend the restriction of hate speech – even instances of it that aren't used to threaten or harass particular people – we need evidence that shows how all hate speech contributes to the harms of social hierarchy, in a way such that hate speakers bear some responsibility for the harms. The most promising hypothesis, in seeking such evidence, is the CLAN Hypothesis: hate speech influences children's attitudes in a way that legitimates and normalizes identity-based inequalities. Hate speech may influence adults too, but it will be easier to acquire evidence of the mechanisms of influence, and to hold hate speakers responsible for the outcomes, if we focus on its effect on children. In arguing for

[52] To reiterate the earlier point from Young, the essential difference is that in structural processes of subordination, "in most cases it is not possible to trace which specific actions of which specific agents cause which specific parts of the structural… outcomes"; see *Political Responsibility and Structural Injustice*, p. 7.

these merits of the CLAN Hypothesis, I don't mean to suggest that no other material, institutional, or communicative factors are involved in the inculcation of prejudicial attitudes in children, besides the influence of hate speech. Our question was whether there is any distinctive contribution that hate speech might be making to the structural harms of identity-based social hierarchy, alongside whatever other causal factors are involved in this. If hate speech is in fact making such a contribution, and if we are looking for evidence of this in order to provide an especially robust justification for BANS, then the CLAN Hypothesis warrants particular attention.

The restriction of hate speech is an established part of most liberal democratic legal systems outside the U.S. Those who want to see this evidentially vindicated should be pursuing collaborative inquiry with empirical researchers. If hate speech does contribute to social hierarchy – *all* hate speech, even those instances of it that aren't used to harass, etc. – then systematic evidence of this should be attainable. Cortese and Delgado and Stefancic gesture in this direction, but there are limitations in the evidence they cite. Nevertheless, examining hate speech's influence on children, as these authors do, is a promising approach. There may be other kinds of arguments for BANS, built around the aim of preventing harm to children, besides the one that I have been exploring here.[53] However, the most credible case for restricting hate speech will be one that simultaneously substantiates the claim that hate speech contributes to identity-based social hierarchies, but without implausibly downplaying the causal primacy of material and institutional factors in underwriting social structures. In order to develop and substantiate that case, we should be asking whether there is evidence that exposure to hate speech impacts children's attitudes in a way that legitimates and normalizes identity-based inequality.

For all I have said, one could still defend a hard-line free speech thesis, which would see BANS as illegitimate even if we did have an evidentially-backed account of hate speech's harmful influence on children. I won't try to offer an assessment of this view of free speech here, except to register one point. If our rationale for restricting hate speech adverts to its influence on children, this allows us to sidestep some prominent free-speech-based objections to BANS. Consider the views of authors like Weinstein and Heinze, that freedom to engage in hate speech is entailed by an essential condition of democratic legitimacy;[54] or consider Baker's view, that "the state only respects people's autonomy if it allows people… to

[53] For example, if it is the case hate speech fosters a climate of hatred towards members of a particular group, in a way that increases the incidence of acts of violence and discrimination against members of this group (see footnotes 26 and 28), then it may be that children comprise a significant portion of those harmed as a result of this.

[54] Weinstein, 'Extreme speech, public order, and democracy'; Heinze, *Hate Speech and Democratic Citizenship*.

express their own values", regardless of "how this expressive content harms other people".[55] These claims trade on the notion that it is illegitimate to impinge upon people's autonomy by trying to control what ideas they're exposed to. But this wouldn't always be a reason to oppose content-based restrictions on speech whose underlying justification was to limit what kinds of messages children are exposed to. In short, the kind of justification for restricting hate speech that I've been proposing is better-placed than some other justifications to address some free-speech based objections.

I will conclude by briefly considering what the policy implications might be if our primary justification for regulating hate speech is one that adverts to its malign influence on children. One set of issues will be about the policies governing social institutions involved in children's care, education, and socialization. We rightly expect these institutions to shield children from hate speech, by adults and by other children, although at higher levels we also expect social sciences and humanities education to enable children to intelligently reckon with the social reality of the attitudes animating hate speech. Beyond this, schools may be the only institutions with any hope of countering the influence of hate speech by parents to children in the home, and in a society committed to reforming identity-based social hierarchies, this is a *pro tanto* reason to think that overtly anti-discrimination values should structure not only the institutional culture of schools, but also certain parts of the curriculum. So far as that is the case, it creates complications in religious educational institutions, in which various kinds of identity-based prejudices – of a sexist, homophobic, or aggressively religiously chauvinistic kind – are more likely to be unofficially tolerated or inculcated as part of children's religious instruction. A serious commitment to shielding children from the influence of hate speech shouldn't suddenly lapse when hate speech occurs under the banner of religious instruction. But equally, regulatory bodies policing these requirements should strive for nuance and contextual-sensitivity in deciding exactly where the avowals of a devout conscience shade into hate speech.

The remaining question is how children's exposure to hate speech can be limited in public spaces generally, outside of institutional contexts. In regards to this question it is useful to distinguish two ways of imposing general legal restrictions on a given type of public expression. Consider the difference between how Holocaust denial laws function in some European states, and how laws function in many states to regulate the public broadcast of adults-only entertainment, including violent and sexually explicit cinema. In both cases the legal restraints are *general*, in that they apply to instances of the relevant communicative acts irrespective of whether, in any instance, they are being used to harass or in some other way

[55] C. Edwin Baker, 'Autonomy and hate speech', in I. Hare and J. Weinstein (eds.), *Extreme Speech and Democracy* (Oxford: Oxford University Press, 2009), pp. 139-57, 142.

target particular individuals. But obviously there is an important difference. In properly regulated contexts, where measures have been taken to constrain children's access, it is permissible to broadcast adults-only entertainment, whereas in jurisdictions with Holocaust denial laws, there is no cordoned-off public arena where Holocaust denial is permitted. It is prohibited regardless of whether it is expressed to random people on the street or to like-minded allies in a clubhouse for extremists.

If our case for restricting hate speech is linked to evidence of its influence on children, then the policies enacted by BANS might bear more of a resemblance to regulations governing adult entertainment, than to prohibitions on Holocaust denial. The primary aim of the intervention would be to limit hate speech's influence on children. Of course it wouldn't cancel out hate speech's influence entirely. Children occupy public spaces in various ways, and in different ways at different ages. The way that children encounter extremist ideas online is variable, and as long as the internet remains relatively free and open, children will sometimes encounter hate speech on it, whether by stumbling across it accidentally, or looking for it out of a curiosity towards the hidden and illicit. But still, in contexts where hate speech is protected children do encounter it in public spaces, and there is at least a *pro tanto* case for limiting their exposure to such encounters. Most legal systems still impose some age-based regulations on the distribution of adult entertainment, in a way that reflects the commonplace view that such material has a negative influence on children. And while the stakes might be higher (or at any rate different) with hate speech, there is reason to think the aim of removing hate speech from public spaces where children will be susceptible to its influence is most effectively pursued by a regulatory approach that borrows from this policy area. This suggestion is radical in that it proposes something different to the prevailing approaches in anti-hate speech law. But it is also mild in the sense that it's likely to be more agreeable to those who think that these prevailing approaches run unacceptably close to being outright prohibitions on certain forbidden opinions.