"It Ain't So Much the Things We Don't Know That Get Us in Trouble. It's the Things We Know that Ain't So" [1]: The Dubious Intellectual Foundations of the Claim that "Hate Speech" Causes Political Violence

Gordon Danning, J.D.*

Abstract

The United States is an outlier in its legal protection for what is commonly termed "hate speech." Proponents of bringing American jurisprudence closer to the international norm often argue that hate speech causes violence, particularly political violence. However, such claims largely rest on assumptions which are inconsistent with social scientists' understanding of the causes of political violence, including that ethnic identity and ideological salience are more often the result of violence than a cause thereof; that violence during conflict is generally unrelated to the conflict's ostensible central cleavage; and that violence is generally instrumental and elite-driven, rather than spontaneous and "bottom-up." Therefore, censorship of hate speech cannot be justified by the argument that such censorship is necessary to prevent or forestall political violence.

* Gordon Danning, J.D., was the History Research Fellow at the Foundation for Individual Rights in Education from January 2017 until January 2019.

1. Robert McHenry, *Knowledge in U.S.: I Know I'm Right and You're Wrong*, CHI. TRIB. (Mar. 20, 2005), http://www.chicagotribune.com/news/ct-xpm-2005-03-20-0503200191-story.html (quoting Artemus Ward).

TABLE OF CONTENTS

I. INTRODUCTION

Over the last several years in a wide range of countries, individuals have been arrested or punished for what the relevant authorities have deemed "hate speech."[2] Although the First Amendment has been interpreted as barring such prosecutions in the United States,[3] many scholars have advocated bringing American jurisprudence on that issue closer to the international norm,[4] where hate speech prosecutions are considered perfectly permissible.[5]

The debate over the propriety of censoring or punishing hate speech is, of course, a comparatively timeworn one,[6] and the arguments in favor of censoring or punishing hate speech are many, including, but not limited to, claims that hate speech causes emotional harm to the listener[7] and that it undermines

2. *See infra* note 5 and accompanying text.

3. Erik Bleich, *Freedom of Expression Versus Racist Hate Speech: Explaining Differences Between High Court Regulations in the USA and Europe*, 40 J. ETHNIC & MIGRATION STUD., 283, 284 (2014) ("In the USA, it is virtually impossible to secure a conviction for racist expressions . . . .").

4. *See, e.g.*, JEREMY WALDRON, THE HARM IN HATE SPEECH 16 (2012) (arguing that hate speech legislation will safeguard "an open and welcoming atmosphere in which all have the opportunity to live their lives, raise their families, and practice their trades or vocations"); Eric Posner, *The World Doesn't Love the First Amendment*, SLATE (Sept. 25, 2012, 4:10 PM), http://www.slate.com/articles/news_and_politics/jurisprudence/2012/09/the_vile_anti_muslim_video_and_the_first_amendment_does_the_u_s_overvalue_free_speech_.html (stating that outside the United States, "the rest of the world . . . see[s] no sense in the First Amendment").

5. *See, e.g.*, Jewish Community of Oslo v. Norway, Communication No. 30/2003, Opinion, Committee on the Elimination of Racial Discrimination [CERD], ¶¶ 2.1, 2.7, 10.5 (Aug. 15, 2005), http://www.worldcourts.com/cerd/eng/decisions/2005.08.15_Jewish_community_of_Oslo_v_Norway.htm (finding that the Norway Supreme Court's annulment of hate speech charges against neo-Nazis violated International Convention on the Elimination of All Forms of Racial Discrimination); TBB-Turkish Union v. Germany, Communication No. 48/2010, Opinion, Committee on the Elimination of Racial Discrimination [CERD], ¶ 12.8 (Feb. 26, 2013), http://www.worldcourts.com/cerd/eng/decisions/2013.02.26_TBB_v_Germany.pdf (finding that failure to prosecute dissemination of ideas based upon racial superiority or hatred and containing elements of incitement to racial discrimination violated Articles 4 and 6 of the International Convention on the Elimination of All Forms of Racial Discrimination); Adan v. Denmark, Communication No. 43/2008, Opinion, Committee on the Elimination of All Forms of Racial Discrimination [CERD], ¶¶ 2.1, 7.7 (Aug. 13, 2010), http://www.worldcourts.com/cerd/eng/decisions/2010.08.13_Mohamad_Adan_v_Denmark.pdf (finding that failure to prosecute individual based on statement that most Somalis carry out genital female mutilation as something quite natural violated International Convention on the Elimination of All Forms of Racial Discrimination).

6. *See, e.g.*, Frederick Schauer, *The Sociology of the Hate Speech Debate*, 37 VILL. L. REV. 805, 806 & n.4 (1992) (stating the amount of law review articles written on the hate speech debate is enormous).

7. *See, e.g.*, R. v. Keegstra, [1990] 3 S.C.R. 697, para. 64 (Can.) ("It is indisputable that the emotional damage caused by words may be of grave psychological and social consequence."); MARI J. MATSUDA ET AL., WORDS THAT WOUND: CRITICAL RACE THEORY, ASSAULTIVE SPEECH, AND THE FIRST AMENDMENT 15 (1993) (acknowledging that the fight over the first amendment "is a fight for a constitutional community where 'freedom' does not implicate a right to degrade and humiliate another human being any more than it implicates a right to do physical violence to another"); Charles R. Lawrence III, *If He Hollers Let Him Go: Regulating Racist Speech on Campus*, 1990 DUKE L. J., 431, 453

the democratic process.[8] But, perhaps the most appealing single argument in favor of censoring or punishing hate speech is the claim that it causes violence, including hate crime,[9] communal violence,[10] and even genocide.[11]

However, despite this lengthy pedigree, arguments that hate speech causes violence are highly suspect, and fall far short of the showing needed to justify the use of the state's coercive power to intrude upon the fundamental human right to freedom of thought and expression because most claims that

("When one is personally attacked with words that denote one's subhuman status and untouchability, there is little (if anything) that can be said to redress either the emotional or reputational injury."); Caleb Yong, *Does Freedom of Speech Include Hate Speech?*, 17 RES PUBLICA 385, 388 (2011) (stating that when defamatory speech wrongfully harms someone "to such an extent that, *even when free speech values are taken seriously*, restrictions on defamatory speech are justified").

8. *See, e.g.*, WALDRON, *supra* note 4, at 4; Roger Errera, *French Law and Racial Incitement: On the Necessity and Limits of the Legal Response*, *in* UNDER THE SHADOW OF WEIMAR: DEMOCRACY, LAW, AND RACIAL INCITEMENT IN SIX COUNTRIES 39, 51 (Louis Greenspan & Cyril Levitt eds., 1993) ("Fundamentally, such laws are necessary if we want to defend the basic civility of our society.").

9. *See, e.g.*, L.W. Sumner, *Incitement and the Regulation of Hate Speech in Canada: A Philosophical Analysis*, *in* EXTREME SPEECH AND DEMOCRACY 204, 209 (Ivan Hare & James Weinstein eds., 2010) ("The two broader social conditions to which hate messages are most frequently said to contribute are the social inequality of target minorities and violence against members of those minorities."); Kevin Boyle, *Hate Speech—The United States Versus the Rest of the World*, 53 ME. L. REV. 487, 501 (2001) ("Hate speech can kill, as too many examples plucked from neo-Nazi violence in Germany to the Timothy Evans of this world to the Rwanda genocide demonstrate.").

10. I use the term, "communal violence" as a shorthand for all forms of identity-based violence, including forms commonly referred to as "ethnic," "religious," or "sectarian." *See* ASHUTOSH VARSHNEY, ETHNIC CONFLICT AND CIVIC LIFE: HINDUS & MUSLIMS IN INDIA 4–5 (2d rev. ed. 2003). For an argument that "ethnic violence" is the more appropriate term, *see* Stuart J. Kaufman, *Ethnicity as a Generator of Conflict*, *in* ROUTLEDGE HANDBOOK OF ETHNIC CONFLICT 91, 91–92 (Karl Cordell & Stefan Wolff eds., 2d ed. 2016) ("[W]hat these cases all have in common is that the groups involved are primarily ascriptive – that is, membership in the groups is typically assigned at birth and is difficult to change.").

11. *See, e.g.*, Laurence Hauptman, *Group Defamation and the Genocide of American Indians*, *in* GROUP DEFAMATION AND FREEDOM OF SPEECH: THE RELATIONSHIP BETWEEN LANGUAGE AND VIOLENCE 9, 11 (Monroe H. Freedman & Eric M. Freeman eds., 1995) ("In relation to indigenous groups and colonizers, group defamation for centuries has resulted in direct and indirect policies of mass extermination."); Stephen J. Roth, *The Laws of Six Countries: An Analytical Comparison*, *in* UNDER THE SHADOW OF WEIMAR: DEMOCRACY, LAW, AND RACIAL INCITEMENT IN SIX COUNTRIES, *supra* note 8, at 177, 202 ("In the shadow of Weimar, and even more in the shadow of the holocaust, we must understand 'clear and present danger' differently from before and must act on the realization that words in themselves can create danger, or certainly are the beginning of a danger."); Audrey Golden, Comment, *Monkey Read, Monkey Do: Why the First Amendment Should Not Protect the Printed Speech of an International Genocide Inciter*, 43 WAKE FOREST L. REV. 1149, 1161 (2008) (stating that a Nazi and Rwandan Hutu "incited genocide through printed hate speech in newspapers"); William A. Schabas, *Hate Speech in Rwanda: The Road to Genocide*, 46 MCGILL L. J. 141, 171 (2000) ("A well-read and well-informed *genocidaire* will know that at the early stages of planning of the 'crime of crimes', his or her money is best spent not in purchasing machetes, or Kalatchnikovs, or Zyklon B gas, but rather investing in radio transmitters and photocopy machines. Genocide is prepared with propaganda . . . aimed at preparing the 'willing executioners' for the atrocious tasks they will be asked to perform.").

hate speech causes violence are premised upon assumptions which are inconsistent with social scientists' understanding of the causes of political violence.[12] In Part II, I discuss the challenges inherent in defining hate speech, and briefly summarize the current state of the law in the United States and abroad regarding the free speech protections provided to those who use language that is deemed to be hate speech.[13] In Part III, I illustrate the ways in which common arguments that hate speech causes political violence are contradicted by the current academic understanding of the causes of political violence, including that ethnic identity and ideological salience are more often the result of violence than a cause thereof; that violence during conflict is generally unrelated to the conflict's ostensible central cleavage; and that violence is generally instrumental and elite-driven, rather than spontaneous and "bottom-up," as one would expect if hate speech were a major cause of violence.[14] Finally, in Part IV, I discuss policy implications of what I argue are fruitless efforts to silence hate speech as a means of preventing violence.[15]

II. "HATE SPEECH": DEFINITIONS AND GENERAL PRINCIPLES

"Hate speech" is more a descriptive term than a legal one; as a recent article notes, "'[h]ate speech' seems to be whatever people choose it to mean. It lacks any objective criteria whatsoever[,]"[16] and in fact, most laws which are commonly referred to as hate speech bans do not use the term.[17] Instead, they outlaw particular forms of speech, which observers subsequently denominate hate speech.[18]

For example, Article Four of the International Convention on the Elimination of All Forms of Racial Discrimination (ICERD) requires that its state parties:

> (a) Shall declare an offence punishable by law all dissemination of ideas based on racial superiority or hatred, incitement to racial discrimination, as well as all acts of violence or incitement to such acts against any race or group of persons of another colour or ethnic origin, and also the provision of any assistance to racist activities, including the financing thereof;

12. *See infra* Part III.
13. See infra Part II.
14. See infra Part III.
15. See infra Part IV.
16. Roger Kiska, *Hate Speech: A Comparison Between the European Court of Human Rights and the United States Supreme Court Jurisprudence*, 25 REGENT U. L. REV. 107, 110 (2013).
17. *See infra* notes 19–25 and accompanying text.
18. *See infra* notes 19–25 and accompanying text.

> (b) Shall declare illegal and prohibit organizations, and also organized and all other propaganda activities, which promote and incite racial discrimination, and shall recognize participation in such organizations or activities as an offence punishable by law.[19]

The International Covenant on Civil and Political Rights (ICCPR) also includes a hate speech provision, albeit a narrower one; Article 20, paragraph 2 thereof mandates that "[a]ny advocacy of national, racial or religious hatred that constitutes incitement to discrimination, hostility or violence shall be prohibited by law."[20]

Various regional agreements include, or have been interpreted to include, similar restrictions on speech.[21] For example, Article 13 of the American Convention on Human Rights (ACHR) requires parties to make criminal "any advocacy of national, racial, or religious hatred that constitute incitements to lawless violence or to any other similar action against any person or group of persons on any grounds including those of race, color, religion, language, or national origin."[22] Furthermore, both the European Convention for the Protection of Human Rights and Fundamental Freedoms (ECHR) and the African Charter on Human and Peoples' Rights (ACHPR) include provisions guaranteeing freedom of expression, but also limitations on those rights; the ECHR states that "[t]he exercise of these freedoms . . . may be subject to such formalities, conditions, restrictions or penalties as are . . . necessary . . . for the protection of the reputation or rights of others,"[23] while the African Charter cautions that "[t]he rights and freedoms of each individual shall be exercised with due regard to the rights of others, collective security, morality and common interest."[24] In regard to the ECHR, the European Court of Human Rights held that it permits prosecutions for violations of national statutes which criminalize the "offence of inciting to national, racial and religious hatred, discord

19. G.A. Res. 2106 (XX) A, annex, International Convention on the Elimination of All Forms of Racial Discrimination, at art. 4 (Dec. 21, 1965).

20. International Covenant on Civil and Political Rights art. 20, ¶ 2, *adopted* Dec. 19, 1966, T.I.A.S. NO. 92-908, 999 U.N.T.S. 171 (entered into force Mar. 23, 1976) [hereinafter ICCPR].

21. *See infra* notes 22–24 and accompanying text.

22. Organization of American States, American Convention on Human Rights art. 13, ¶ 5, *opened for signature* Nov. 22, 1969, O.A.S.T.S. No. 36, 1144 U.N.T.S. 123 (entered into force July 18, 1978) [hereinafter ACHR].

23. Council of Europe, European Convention for the Protection of Human Rights and Fundamental Freedoms art. 10, *opened for signature* Nov. 4, 1950, E.T.S. No. 5, 213 U.N.T.S. 221 (entered into force Sept. 3, 1953).

24. Organization of African Unity, African Charter on Human and Peoples' Rights art. 27, ¶ 2, *adopted* June 27, 1981, 1520 U.N.T.S. 217 (entered into force Oct. 21, 1986).

or intolerance."[25]

In the United States, of course, there is no blanket prohibition on hate speech, as the United States Supreme Court reaffirmed in 2017 when it struck down the United States Patent and Trademark Office's (USPTO) refusal to register a trademark for a band whose name is often used as an anti-Asian slur.[26] The Court reiterated longstanding doctrine when it stated that "[s]peech that demeans on the basis of race, ethnicity, gender, religion, age, disability, or any other similar ground is hateful" but is nevertheless protected under the First Amendment to the United States Constitution.[27]

However, in practice there is some overlap between American and non-American jurisprudence in this area because speech in the United States that "is directed to inciting or producing imminent lawless action and is likely to incite or produce such action" is unprotected,[28] and sentencing enhancements for crimes that are motivated by the victim's race, religion, or other group status are not barred by the First Amendment.[29] In that sense, American jurisprudence, in its practical effect, is not inconsistent with the ACHR's call for criminalization of "any advocacy of national, racial, or religious hatred that constitute incitements to lawless violence,"[30] nor with that portion of Article 20 of the ICCPR which mandates that "[a]ny advocacy of national, racial or religious hatred that constitutes incitement to . . . violence shall be prohibited by law."[31]

The only other case[32] in which the Supreme Court has held that the First Amendment permits punishment for what can be described as hate speech is *Virginia v. Black*.[33] There, a defendant was convicted for violating a Virginia statute that outlawed "cross burning with intent to intimidate" when he burned a cross on the lawn of an African-American family.[34] The Court had previously overturned a Minnesota cross burning statute in *R.A.V. v. City of St.*

25. Smajić v. Bosnia and Herzegovina, App. No. 48657/16, at 11 (Eur. Ct. H.R. Jan. 16, 2018), http://hudoc.echr.coe.int/eng?i=001-180956.

26. *See* Matal v. Tam, 137 S. Ct. 1744, 1765 (2017) (holding that the disparagement clause of the Lanham Act violated the First Amendment).

27. *Id.* at 1764.

28. Brandenburg v. Ohio, 395 U.S. 444, 447 (1969).

29. *See* Wisconsin v. Mitchell, 508 U.S. 476, 479 (1993) (upholding a battery statute that permitted enhanced sentencing when the defendant intentionally selected the victim because of the victim's race).

30. ACHR, *supra* note 22.

31. ICCPR, *supra* note 20.

32. In *Beauharnais v. Illinois*, the Supreme Court famously upheld a conviction under a statute that forbade speech portraying.

33. Virginia v. Black, 538 U.S. 343, 363 (2003).

34. *Id.* at 362.

Paul,[35] but the Court upheld the Virginia law because it, unlike the statute at issue in *R.A.V.*, required an intent to intimidate,[36] and hence the speech at issue was akin to a "true threat," wherein the speaker "means to communicate a serious expression of an intent to commit an act of unlawful violence to a particular individual or group of individuals."[37] Moreover, while the statute at issue in *R.A.V.* outlawed only cross burning that "arouse[d] anger, alarm or resentment in others on basis of race, color, creed, religion or gender,"[38] the statute at issue in *Black* outlawed all cross burning that was intended to intimidate on any basis.[39] Hence, the statute at issue in *Black*, unlike that at issue in *R.A.V.*, did "not single out for opprobrium only that speech directed toward 'one of the specified disfavored topics.'"[40]

That being said, the class of expression that can be described as hate speech and which is punishable under American law is far narrower than that which is punishable elsewhere.[41] For example, the ICERD calls on states to outlaw not just incitement to violence, but also the mere "dissemination of ideas based on racial superiority or hatred";[42] the ICCRP calls on states to punish incitement to "hostility" based on protected status.[43] Furthermore, both the ICERD and ICCPR expressly call for the prohibition of incitement to discrimination;[44] and under the ECHR, "inciting to hatred does not necessarily entail a call for an act of violence, or other criminal acts."[45]

In addition, while the United States Supreme Court has refused to countenance the suppression of speech that cannot be shown to have an almost

35. R.A.V. v. City of St. Paul, 505 U.S. 377, 396 (1992).

36. *Virginia*, 538 U.S. at 360, 363 ("Respondents do not contest that some cross burnings fit within this meaning of intimidating speech, and rightly so. As noted in Part II, *supra*, the history of cross burning in this country shows that cross burning is often intimidating, intended to create a pervasive fear in victims that they are a target of violence. . . . The First Amendment permits Virginia to outlaw cross burnings done with the intent to intimidate because burning a cross is a particularly virulent form of intimidation. Instead of prohibiting all intimidating messages, Virginia may choose to regulate this subset of intimidating messages in light of cross burning's long and pernicious history as a signal of impending violence.").

37. *Id.* at 359. The Court overturned the portion of the statute that made burning a cross prima facie evidence of the intent to intimidate. *Id.* at 364–65.

38. *Id.* at 361 (quoting *R.A.V.*, 505 U.S. at 391).

39. *Id.* at 347.

40. *Id.* at 362 (quoting *R.A.V.*, 505 U.S at 391).

41. *See infra* notes 35–40.

42. *See* G.A. Res. 2106 (XX) A, *supra* note 19.

43. *See* ICCPR, *supra* note 20.

44. *See* G.A. Res. 2106 (XX) A, *supra* note 19; ICCPR, *supra* note 20.

45. Vejdeland v. Sweden, App. No. 1813/07, ¶ 55 (Eur. Ct. H.R. May. 9, 2012), http://hudoc.echr.coe.int/eng?i=001-109046.

immediate negative effect, many other jurisdictions take a very different approach.[46] For example, "the Canadian approach to hate speech focus[es] on gradual long-term effects likely to pose serious threats to social cohesion rather than merely on immediate threats to violence,"[47] and the hate speech jurisprudence of the European Court of Human Rights rests substantially on the "fear . . . that the spread of hateful views may generate or reinforce hatred in the community and ultimately result in hateful attitudes, discrimination or violence."[48]

As mentioned above, those who seek to censor hate speech set forth several rationales other than the claim that hate speech engenders violence.[49] This article, however, critically examines only that rationale.[50] Moreover, while the term hate speech has no single, precise meaning, all of the authorities discussed hereinafter either implicitly or explicitly employ a definition which is consistent with the initial phrase of Article 20 of the ICCPR, to wit, "advocacy of national, racial or religious hatred."[51]

III. CLAIMS THAT HATE SPEECH CAUSES VIOLENCE ARE OFTEN BASED ON ASSUMPTIONS WHICH ARE INCONSISTENT WITH ACADEMIC UNDERSTANDINGS OF THE CAUSES OF POLITICAL VIOLENCE

Many arguments that hate speech causes violence do not explicitly state a mechanism whereby hate speech produces that effect.[52] To the extent that

46. *See* Bleich, *supra* note 3.

47. Michel Rosenfeld, *Hate Speech in Constitutional Jurisprudence: A Comparative Analysis*, 24 CARDOZO L. REV. 1523, 1543 (2002) (discussing R. v. Keegstra, [1990] 3 S.C.R. 687 (Can.)).

48. Stefan Sottiaux, *'Bad Tendencies' in the ECtHR's 'Hate Speech' Jurisprudence*, 7 EUR. CONST. L. REV. 40, 47 (2011). *But see* Antoine Buyse, *Dangerous Expressions: The ECHR, Violence and Free Speech*, 63 INT'L & COMP. L.Q. 491 (2014) (arguing that there are substantial divisions within the Court regarding the extent to which direct causation of violence is necessary to render speech unprotected).

49. *See supra* notes 6–8 and accompanying text.

50. For a critique of the empirical underpinnings of other arguments in favor of hate speech regulation, see John T. Bennett, *The Harm in Hate Speech: A Critique of the Empirical and Legal Bases of Hate Speech Regulation*. 43 HASTINGS CONST. L.Q. 445 (2016).

51. *See* ICCPR, *supra* note 20.

52. Perhaps the most telling example of the poverty of the standard discussion of causal mechanisms is illustrated in Richard Wilson's criticism of the Rwanda International Criminal Tribunal's trial court decision in the prosecution of the perpetrators of the RTLM "hate radio" broadcasts: "In identifying the mechanisms through which propaganda exerted its causal force, [the tribunal] mixed its metaphors, combining the image of spreading gasoline with Nuremberg's portrayal of a propagandist injecting poison into the mind of a civilian population." Richard Ashby Wilson, *Inciting Genocide with Words*, 36 MICH. J. INT'L L., 277, 290 (2015) (quoting Prosecutor v. Nahimana, Case No. ICTR99-52-T, Judgment, ¶ 1078 (Dec. 3, 2003)). Those "mechanisms" are of course not mechanisms at all, but merely colorful metaphors which function to disguise the fact that the court could not make a coherent causal argument. *See* Susan Benesch, *The Ghost of Causation in International Speech Crime*

those arguments specify a causal sequence, it is the following: Hate speech creates hatred in the listener and hatred is then expressed in violence.[53] However, that view is inconsistent with the current academic understanding of the causes of political violence.[54] In particular, those arguments are inconsistent with the understanding that ethnic identity and ideological salience are endogenous to violence;[55] that violence during a conflict is generally unrelated to the conflict's ostensible central cleavage;[56] and that violence is generally instrumental and elite-driven, rather than spontaneous and "bottom-up."[57]

A. Group Identity is Often Endogenous to Violence

As noted above, the implicit causal claim made by many who contend that hate speech causes violence is that hate speech engenders hatred among the members of one group towards members of another group, and then hatred motivates violence by members of the first group against members of the second.[58] Implicitly, that argument presumes the existence of distinct, pre-existing group identities, because it assumes that some individuals identify as members of the first group, while others identify as members of the second.[59] However, that assumption is a highly questionable one, for it is subject to possible endogeneity; i.e, the fact that in causal analysis, "the values our explanatory variables take on are sometimes a consequence, rather than a cause, of our dependent variable."[60]

Cases, *in* PROPAGANDA, WAR CRIMES TRIALS AND INTERNATIONAL LAW 254, 254–57 (Predrag Dojčinovič ed., 2012) (stating that "the ICTR welded its own causal link between certain speech acts and thousands of deaths, since causation was not proved," and arguing that judges in international criminal tribunals make unsupported causal claims, despite the fact that causation is not an element of incitement to genocide, as a means of compensating for lack of alternative "adequate means of identifying incitement to genocide").

53. *See, e.g.*, Calvin R. Massey, *Hate Speech, Cultural Diversity, and the Foundational Paradigms of Free Expression*, 40 UCLA L. REV. 103, 156 (1992).

54. Laia Balcells, *Rivalry and Revenge: Violence against Civilians in Conventional Civil Wars*, 54 INT'L STUD. Q., 291, 291 (2010) (noting that, in the view of early scholars, "civil conflicts were seen as the result of existing political cleavages, and violence as the consequence of these divisions[,]" but that understanding has been supplanted by more recent scholars, generally "using more systematic research methods than the previous generation of scholars"); Benjamin A. Valentino, *Why We Kill: The Political Science of Political Violence Against Civilians*, 17 ANN. REV. POL. SCI. 89, 91 (2014) ("[T]he new research has overturned the once widely held view that large-scale violence against civilian populations was irrational, random, or the result of ancient hatreds . . . .").

55. *See supra* Section III.A.

56. *See supra* Section III.B.

57. *See supra* Section III.C.

58. *See supra* note 53 and accompanying text.

59. *See, e.g.*, James D. Fearon & David D. Latin, *Violence and the Social Construction of Ethnic Identity*, 54 INT'L ORG. 845, 857 (2000).

60. GARY KING, ROBERT O. KEOHANE & SIDNEY VERBA, DESIGNING SOCIAL INQUIRY 185

An example of the perils of endogeneity is seen in the historiography surrounding the political success of Nazism:

> One of the great puzzles of political analysis for an earlier generation of political scientists was the fall of the Weimar Republic . . . . One explanation . . . was that the main cause was the imposition of proportional representation as the mode of election in the Weimar Constitution. . . . The underlying explanation involved a causal mechanism with the following links in the causal chain: proportional representation was introduced and enabled small parties with narrow electoral bases to gain seats in the Reichstag (including parties dedicated to its overthrow, like the National Socialists). As a result, the Reichstag was stalemated and the populace was frustrated. This, in turn, led to a coup by one of the parties. But further study . . . indicated that party fragmentation was not merely the result of proportional representation. . . . [S]cholars found that societies with a large number of groups with narrow and intense views in opposition to other groups—minority, ethnic, or religious groups, for instance—are more likely to adopt proportional representation, since it is the only electoral system that the various factions in society can agree on.[61]

The failure of initial claims regarding the political success of the Nazi Party to consider the possibility of endogeneity is echoed in the assumption of hate speech opponents that hate speech causes violence by appealing to communal identities.[62] That assumption is inconsistent with current social scientific understanding of the relationship between identity and violence, which is that, in societies that are experiencing violent conflict, identity is often endogenous to violence.[63] In other words, rather than identities causing violence, the opposite is the case: violence causes individuals to adopt oppositional identities by increasing the salience of those identities.[64] Therefore,

(1994).

61. *Id*. at 189–90.

62. *See infra* note 64 and accompanying text.

63. *See infra* note 64 and accompanying text.

64. LEE ANN FUJII, KILLING NEIGHBORS: WEBS OF VIOLENCE IN RWANDA 102 (2009) (noting that in Rwanda, "group-level hatreds and fears seem to be the product, not the producer, of violence"); VAMIK VOLKAN, BLOODLINES: FROM ETHNIC PRIDE TO ETHNIC TERRORISM 25 (1997) (arguing that group identities become salient for individuals only when they are under threat); Laia Balcells & Abbey Steele, *Warfare, Political Identities, and Displacement in Spain and Colombia*, 51 POL. GEOGRAPHY 15, 16 (2016) ("Most people hold multiple identities, but during conflict some identities become more salient than others . . . ."); Fearon & Laitin, *supra* note 59, at 846 (stating that case

the standard assumption that hate speech causes violence rests on a false premise.[65]

A close correlate of these findings regarding the endogeneity of identity to violence is the understanding that ideological appeals, and the salience thereof, are also often endogenous to violence.[66] Thus, Stathis Kalyvas discussed "[t]he frequent endogeneity of ideology to the war"[67] and Matthew Isaacs found that, while

studies of communal violence in India, Sudan, Sri Lanka, and Northern Ireland indicate that "[v]iolence has the effect, intended by the elites, of constructing group identities in more antagonistic and rigid ways"); R. Brian Ferguson, *Tribal Warfare and "Ethnic" Conflict*, 29 CULTURAL SURVIVAL Q. 18, 18–19 (2005) (stating group loyalty is a function of conflict, rather than conflict being a function of group loyalty); Stathis N. Kalyvas & Matthew Adam Kocher, *Ethnic Cleavages and Irregular War: Iraq and Vietnam*, 32 POL. & SOC'Y 183, 186, 190 (2007) ("[E]thnic cleavages are further activated and deepened by the war, rather than war merely reflecting already deep ethnic cleavages . . . . [Hence,] ethnic affiliation [is] partially endogenous to the course of the war itself."); Kaufman, *supra* note 10, at 92 ("Furthermore, identities do sometimes change, with new ones emerging and old ones disappearing, especially in times of crisis."); Jeffrey Stevenson Murer, *Ethnic Conflict: An Overview of Analyzing and Framing Communal Conflicts from Comparative Perspectives*, 24 TERRORISM & POL. VIOLENCE 561, 567–68 (2012) ("[T]he crisis comes to be explained in ethnicized or racialized terms. The crisis is not ethnic in nature, but the explanation of it is. That is, there is a crisis, and the actors involved invoke, or observers ascribe ethnicity as the organizing coherence of a group. As the ethnicized explanation of crisis conditions begins to gain cultural and public currency, there is increased pressure for all members of the community—including liberal and moderately tolerant individuals—to think and act in an increasingly ethnically defined manner."); Duško Sekulić, Garth Massey & Randy Hodson, *Ethnic Intolerance and Ethnic Conflict in the Dissolution of Yugoslavia*, 29 ETHNIC & RACIAL STUD. 797, 812 (2006) (noting that, in the 1985, 1993, and 2001 censuses in Croatia, the percentage of respondents declaring their nationality as ''Yugoslav'' dropped from 10.8% in 1985 to 2.3% in 1993 at the beginning of Croatian War and then to 0.0% in 2001); Nils B. Weidmann, *Violence "From Above" or "From Below"? The Role of Ethnicity in Bosnia's Civil War*, 73 J. POL. 1178, 1188 (2011) (noting that the results of author's research "provide preliminary evidence for the activation of ethnic enmity by violence"). *But see* Robert Hislope, *From Expressive to Actionable Hatred: Ethnic Divisions and Riots in Macedonia*, *in* IDENTITY CONFLICTS: CAN VIOLENCE BE REGULATED? 149, 149 (J. Craig Jenkins & Esther E. Gottlieb eds., 2007) (arguing that "in some societies, ethnic hatreds are quite palpable and do shape the pattern of political dynamics[,]" but recognizing that "such hatreds are insufficient to spark violence"); Tobias Ide, *Why Do Conflicts Over Scarce Renewable Resources Turn Violent? A Qualitative Comparative Analysis*, 33 GLOBAL ENVTL CHANGE 61, 68–69 (2015) ("[T]he simultaneous presence of two structural conditions (negative othering and low power differences) and one triggering condition (recent political change) is sufficient for the violent escalation of renewable resource conflicts.").

65. *See, e.g.*, *supra* notes 64 and accompanying text. This phenomenon, of course, helps resolve the mystery of why neighbors who formerly got along well, ultimately turn on one another once conflict begins. *See, e.g.*, Prosecutor v. Brđanin, Case No. IT-99-36-T, Judgment, ¶ 80 (Int'l Crim. Trib. for the Former Yugoslavia Sept. 1, 2004) ("Within a short period of time, citizens who had previously lived together peacefully became enemies . . . ."). That transformation is caused not by hate speech, but by the initial violence which co-occurs with hate speech. *See, e.g.*, *supra* note 64 and accompanying text.

66. See Matthew Isaacs, *Sacred Violence or Strategic Faith? Disentangling the Relationship Between Religion and Violence in Armed Conflict*, 53 J. PEACE RES. 211, 212 (2016).

67. STATHIS N. KALYVAS, THE LOGIC OF VIOLENCE IN CIVIL WAR 45 (2006).

> [a]nalysis confirms that religion and violence are broadly correlated[,] . . . [there is] no evidence that prior religious rhetoric encourages organizations to participate in violence or to increase the intensity of violent tactics. On the contrary, . . . violent actors adopt religious rhetoric to solve the logistical challenges associated with violence, including access to mobilizing resources and recruitment and retention of members.[68]

Therefore, the assumption that hate speech catalyzes violence by communicating an ideology of hatred, thereby convincing previously neutral actors that a particular group is a threat,[69] is almost certainly a specious one; instead, it is the violence itself which causes violent groups to adopt identity-based rhetoric.[70]

Evidence that this is the case can be seen in Scott Straus's finding that the effects of RTLM radio station's notorious hate broadcasts in Rwanda were conditional on the pre-existence of violence.[71] Furthermore, to the extent that propaganda is correlated with increased levels of either hatred or violence, that correlation exists only in communities which are already sympathetic to the claims therein.[72] For example, Adena, *et al.*, found that the effect of Nazi anti-Semitic propaganda varied according to the predisposition of listeners because it was most effective in engendering anti-Semitic behavior in areas where anti-Semitism was historically high, but had a negative effect in places where anti-Semitism was historically low;[73] Jason Chan, Anindya Ghose, and Robert Seamans found that broadband internet availability increases racial hate crimes in the United States only in those areas with higher levels of racism;[74] Maria Petrova and David Yanagizawa-Drott's review of the literature indicates that propaganda which targets minorities is most effective when it is aligned with the predispositions of its audience;[75] and Suranjan Weeraratne

68. Isaacs, *supra* note 66.

69. *See supra* note 53 and accompanying text.

70. Isaacs, *supra* note 66.

71. Scott Straus, *What Is the Relationship between Hate Radio and Violence? Rethinking Rwanda's "Radio Machete,"* 35 POL. & SOC'Y 609, 630–32 (2007).

72. *See infra* note 73.

73. Maja Adena et al., *Radio and the Rise of the Nazis in Prewar Germany*, 130 Q.J. ECON. 1885, 1885 (2015).

74. Jason Chan, Anindya Ghose & Robert Seamans, *The Internet and Racial Hate Crime: Offline Spillovers from Online Access*, 40 MIS Q. (SPECIAL ISSUE) 381, 395 (2016).

75. Maria Petrova & David Yanagizawa-Drott, *Media Persuasion, Ethnic Hatred, and Mass Violence: A Brief Overview of Recent Research Advances*, *in* ECONOMIC ASPECTS OF GENOCIDES, OTHER MASS ATROCITIES, AND THEIR PREVENTION 274, 284 (Charles H. Anderton & Jurgen Brauer eds., 2016) (explaining that "[w]hen propaganda is aligned with population predispositions, persuasion appears especially effective[,]" and suggesting that "community level beliefs and behaviors are important

found that elite-orchestrated campaigns in Indonesia which scapegoated minorities resulted in anti-Chinese riots only when the rhetoric resonated at the local level.[76]

In rebuttal, some might argue that the fact that the perpetrators of genocide in places like Germany and Rwanda employed racist propaganda implies "that the perpetrators thought it was important[,]"[77] which in turn might indeed seem to create the inference that there is a causal relationship between hate propaganda and violence: why else would the perpetrators employ propaganda, if they did not think it would help them implement genocide?

However, that inference is mistaken; the observation that practitioners of propaganda "thought it was important" in fact raises a different question: important for whom? The interests of those charged with implementing state propaganda were not necessarily identical to those of the regime or of its leader, as decades of research on the principal-agent problem makes manifest.[78] In regard to Germany, the Nazi regime was characterized by extensive bureaucratic infighting.[79] This is to be expected in a dictatorial regime because a key to a tyrant remaining in power is to keep his essential supporters off balance so that they are less able to unite to depose and replace him.[80] Moreover, there were particularly intense internal rivalries within the Nazi regime between Joseph Goebbels and other high officials such as Joachim Rittentrop and Alfred Ernst Rosenberg, including disputes over the content of propaganda.[81] Given that regime insiders in personalist regimes generally must curry favor with the leader in order to maintain their status within the

for propaganda affecting individual participation in genocide").

76. Suranjan Weeraratne, *Ethnic Entrepreneurs and Collective Violence: Assessing Spatial Variations in Anti-Chinese Rioting Within Jakarta During the May 1998 Riots* (World Inst. for Dev. Econ. Research, Working Paper No. 2010/55, 2010), https://www.wider.unu.edu/publication/ethnic-entrepreneurs-and-collective-violence.

77. Susan Benesch, *The New Law of Incitement to Genocide: A Critique and a Proposal*, U.S. HOLOCAUST MEM'L MUSEUM 4 (Feb. 2009), https://www.ushmm.org/m/pdfs/20121009-benesch-new-incitement-law.pdf.

78. *See generally* JAN-ERIK LANE, COMPARATIVE POLITICS: THE PRINCIPAL-AGENT PERSPECTIVE (2007).

79. IAN KERSHAW, HITLER: A BIOGRAPHY 323 (2008); ROGER MANVELL & HEINRICH FRAENKEL, HEINRICH HIMMLER: THE SINISTER LIFE OF THE HEAD OF THE SS AND GESTAPO 329 (2007).

80. BRUCE BUENO DE MESQUITA & ALASTAIR SMITH, THE DICTATOR'S HANDBOOK 61–65 (2011).

81. DIETRICH ORLOW, THE LURE OF FASCISM IN WESTERN EUROPE: GERMAN NAZIS, DUTCH AND FRENCH FASCISTS, 1933-1939, 122 (2009); TOBY THACKER, JOSEPH GOEBBELS: LIFE AND DEATH 166 (2009); Cooper C. Graham, *"Sieg im Westen" (1941): Interservice and Bureaucratic Propaganda Rivalries in Nazi Germany*. 9 HIST. J. FILM, RADIO & TELEVISION, 19, 21 (1989) ("Dr. Goebbels and von Brauchitsch had already wrangled several times about their respective spheres of authority.")

regime,[82] and given Adolf Hitler's well-known personal anti-Semitism, the leaders of the regime's propaganda arm would almost certainly have produced anti-Semitic propaganda regardless of whether they considered it efficacious.[83]

In addition, it is a mistake to assume that a genocidal regime's purpose in employing hate propaganda is to engender support for a policy of mass violence; instead, it is equally likely that the regime's purpose is to increase the likelihood of regime survival by delegitimizing internal rivals:

> One of the central themes in the propaganda campaigns used in Nazi Germany, Rwanda, and Serbia is that the enemy must be destroyed before they destroy those native to the country. This message not only identifies and demonizes an external enemy but also warns that any dissent from government policies is part of a plot to collaborate with that enemy. Thus, even if propaganda does nothing to augment ethnic prejudices, it might still be in the interest of the regime, which seeks to defend its grip on power against other factions.[84]

Therefore, the claim that perpetrators' use of hate propaganda implies that they believe that propaganda is effective in promoting genocide rests on a false assumption about the motives, interests, and incentives of those perpetrators.[85]

In addition, these arguments also assume a level of efficacy of propaganda that is not supported by the available evidence.[86] For example, a study of the effectiveness of Nazi indoctrination found that those efforts helped to foster hatred, but that it was schooling, rather than propaganda, which was most effective.[87] Even in Rwanda, most of the studies published to date cast doubt

82. *See* JESSICA L. P. WEEKS, DICTATORS AT WAR AND PEACE 8 (Robert J. Art et al eds., 2014) ("[W]hile a key assumption of selectorate theory is that small-coalition regime insiders believe that they will lose their privileged positions under a new ruler, this assumption is inaccurate for many small-coalition regimes.").

83. *See supra* KERSHAW, *supra* note 79, at 42–43 (highlighting the development of Hitler's anti-semitism).

84. Donald P. Green & Rachel L. Seher, *What Role Does Prejudice Play in Ethnic Conflict?*, 6 ANN. REV. POL. SCI. 509, 516 (2003).

85. *See id.* at 510 ("When tracing the causes of genocide or ethnic civil war, scholars tend to refer obliquely to longstanding hatreds, nationalist ideologies, or animus generated by propaganda campaigns.").

86. For a criticism of international tribunals' failure to integrate social scientific findings regarding the effectiveness of propaganda, see generally RICHARD WILSON, INCITEMENT ON TRIAL: PROSECUTING INTERNATIONAL SPEECH CRIMES (2017).

87. Nico Voigtländer & Hans-Joachim Voth, *Nazi Indoctrination and Anti-Semitic Beliefs in Germany*, 112 PROC. NAT'L ACAD. SCI. U.S., 7931, 7931 (2015) ("As a result, Germans who grew up under the Nazi regime are much more anti-Semitic than those born before or after that period: the

on the effect of the hate propaganda employed there:

> [R]ecent empirical social science studies cast doubt upon the international tribunal's account of the role of propaganda and the media in 1994 Rwanda. On the basis of one hundred interviews of convicted perpetrators in a Kigali prison, Rwandan cultural anthropologist Charles Mironko found that many ordinary villagers either did not receive genocidal radio transmissions or did not interpret them in the way they were intended. Mironko therefore urges caution in ascribing a causal link between RTLM broadcasts and genocidal killings: "[T]his information alone did not cause them to kill." Scott Straus's more quantitative study of the relationship between radio and violence in Rwanda both corroborates and extends Mironko's study. Straus identifies a number of flaws in the [International Criminal Tribunal for Rwanda]'s reasoning and fact-finding: RLTM's coverage was very uneven, especially in rural areas, and only ten percent of the population owned a radio in 1994; the initial violence did not correspond with areas of broadcast coverage and the most extreme and inflammatory broadcasts came after most of the killings had been carried out. Straus complements his quantitative analysis with 200 perpetrator interviews, and these revealed that radio listeners did not necessarily internalize the elements of anti-Tutsi propaganda. Perhaps most crucially, no respondent cited the radio broadcasts as the most important reason for their participation in the genocide.
>
> Both Mironko and Straus's respondents reported that that peer pressure from male neighbors and kin exerted more influence on their participation in killing than did government and radio propaganda. Like Mironko, Straus infers that the radio broadcasts functioned as a device to coordinate attacks and were meant primarily for local authorities, who played the main role in mobilizing citizens directly: "Radio did not cause the genocide or have direct, massive effects. Rather, radio emboldened hard-liners and reinforced face-to-face mobilization. . . ."
>
> . . . Economist David Yanagizawa-Drott . . . is an outlier in his finding that approximately 10 per cent of the participation in the genocide can be attributed to the radio broadcasts, corresponding to an estimated 50,000 murders. At this point we can reliably say only that the

share of committed anti-Semites, who answer a host of questions about attitudes toward Jews in an extreme fashion, is 2-3 times higher than in the population as a whole.").

> empirical evidence on the effect of propaganda is mixed, and until more conclusive evidence is available, it would be prudent to approach with circumspection ICTR judges' forceful claims about a direct connection between speech acts and violence.[88]

Significantly, even the findings of the cited outlier might be spurious; a reanalysis of the data relied upon therein found no evidence that Radio Rwanda broadcasts increased participation in the genocide.[89]

Also, those who make the claim that public hate propaganda campaigns cause violence rarely specify why regimes which engage in such mass violence need public support for their policies in the first place.[90] Certainly, Germany did not; because it was a one-party, autocratic state, its leader needed only to satisfy the demands of a small number of core supporters in order to stay in power.[91] Therefore, its leader could safely ignore the demands and opinions of the citizenry.[92] Similarly, when the Guatemalan government engaged in genocidal acts against indigenous groups as part of a counterinsurgency campaign, it made no attempt to disseminate hateful speech among the general Guatemalan population; instead, it addressed such speech to its soldiers and to communities within the rural regions where the insurgency was

88. Wilson, *supra* note 52, at 301–02 (footnotes omitted) (first citing Richard Carver, *Broadcasting and Political Transition: Rwanda and Beyond*, *in* AFRICAN BROADCAST CULTURES, 188, 192 (Richard Farndon & Graham Furniss, eds., 2000); then quoting Charles Mironko, *The Effect of RTLM's Rhetoric of Ethnic Hatred in Rural Rwanda*, *in* THE MEDIA AND THE RWANDA GENOCIDE 125, 134 (Allan Thompson ed., 2007); then citing SCOTT STRAUS, THE ORDER OF GENOCIDE: RACE, POWER, AND WAR IN RWANDA (2d ed. 2006); then quoting Straus, *supra* note 71, at 131; and then citing David Yanagizawa-Drott, *Propaganda and Conflict: Evidence from the Rwandan Genocide*, 129 Q. J. ECON. 1947, 1985–86 (2014)).

89. *See generally* Gordon Danning, *Did Radio RTLM Really Contribute Meaningfully to the Rwandan Genocide?: Using Qualitative Information to Improve Causal Inference from Measures of Media Availability*, 20 CIVIL WARS, 529 (2018).

90. *See* Yangaizawa-Drott, *supra* note 88, at 1948 ("Elites in control of autocratic states have repeatedly used mass media—often under their direct control—with the intention to induce citizen support of and participation in violence against certain groups.").

91. BRUCE BUENO DE MESQUITA, *ET AL*, THE LOGIC OF POLITICAL SURVIVAL 475 (2003); *see also* BENJAMIN A. VALENTINO, FINAL SOLUTIONS 66 (2013) ("A few leaders cannot implement mass killing alone, but perpetrators do not need widespread social support in order to carry it out."). *But see* MICHAEL MANN, THE DARK SIDE OF DEMOCRACY ix (2005) (arguing that emerging democracies which define their political community in ethnic terms are at greater risk of "murderous ethnic cleansing" than are authoritarian regimes, while acknowledging that institutionalized democracies are at the lowest risk of all); *see also* Rogers Brubaker & David D. Laitin, *Ethnic and Nationalist Violence*, 24 ANN. REV. SOC. 423, 434 (1998) (discussing the role of ethnic outbidding, whereby "two or more parties identified with the same ethnic group compete for support, . . . each seeking to demonstrate to their constituencies that it is more nationalistic than the other, and each seeking to protect itself from the other's charges that it is 'soft' on ethnic issues").

92. *See supra* note 91 and accompanying text.

most active.[93] Nor are Germany and Guatemala outliers; rather, they are the norm because as a rule:

> [E]lites can manipulate violence for their own gain and at great cost to the public because they do not need public cooperation. Even large-scale violence against civilians does not require the direct participation of large numbers of armed men, and elites can easily reward the small numbers they do need with private incentives.[94]

Finally, even if advocates of censorship are correct, and there is indeed a causal relationship between hate propaganda and violence, the causal mechanism involved is not what advocates apparently believe it to be. The evidence is clear that in the very narrow circumstances in which researchers have found a correlation between media exposure and violence, they have found that any causal effect operates not by increasing enmity, but by helping to solve logistical and coordination challenges,[95] or by signaling to those who are violence-prone that the authorities approve of violence and will likely not punish those who engage in violence.[96]

---

93. Frank Smyth, *Painting the Maya Red: Military Doctrine and Speech in Guatemala's Genocidal Acts*, U.S. HOLOCAUST MEM'L MUSEUM 3 (2009), https://www.ushmm.org/confront-genocide/speakers-and-events/all-speakers-and-events/speech-power-violence ("The Army further developed colloquial speech to disseminate the same ideas down to non-commission officers and soldiers.").

94. Valentino, *supra* note 54, at 98 (citing JOHN MUELLER, THE REMNANTS OF WAR 1 (2004)).

95. Catie Snow Bailard, *Ethnic Conflict Goes Mobile: Mobile Technology's Effect on the Opportunities and Motivations for Violent Collective Action*, 52 J. PEACE RES.323, 323 (2015) (finding "mixed support" for hypotheses that "mobile phone availability primarily increases a group's opportunities to engage in violent collective action as a result of decreased organizational costs due to diminished communication costs" and that "mobile phone availability makes violent collective action more likely as a result of increasing a group's motivation to organize, thanks to enabling more efficient communication about shared grievances between group members"); Jan H. Pierskalla & Florian M. Hollenbach, *Technology and Collective Action: The Effect of Cell Phone Coverage on Political Violence in Africa*, 107 AM. POL. SCI. REV.207, 207 (2013) (determining that the availability of cell phone coverage increases the probability of violent conflict by allowing political groups to overcome collective action problems more easily and to improve coordination); Wilson, *supra* note 52, at 302("[S]ocial science research offers quite a different model of violence than that asserted by the ICTR judges: instead of positing a relationship between leaders' speeches and popular genocidal acts, it points toward speeches as a form of communication between elites, who then recruited on a personal or kin basis."); Yanagizawa-Drott, *supra* note 85, at 1973 n.27 ("Such reinforcing effects [of propaganda] could arise when radio facilitates coordination, but also if information transmission among neighbors influence how strongly beliefs are updated.").

96. DONALD L. HOROWITZ, THE DEADLY ETHNIC RIOT 343–52 (2001); Hollie Nyseth Brehm, *State Context and Exclusionary Ideologies* 60 AM. BEHAV. SCIENTIST, 131, 133 (2016) ("[M]any 'foot soldiers' who have committed genocide likely would not have acted if not for a state-led (or state-supported) ideology . . . . While each participant—or even each member of the political elite—was not necessarily motivated by an exclusionary ideology, state-driven ideologies certainly supported such action and helped legitimate the violence for actors on the ground.")(citations omitted); Scott Poynting & Barbara Perry, *Climates of Hate: Media and State Inspired Victimisation of Muslims in*

B. Political Violence is Generally Unrelated to a Conflict's Ostensible Central Cleavage

Causal claims that hate speech causes violence are inconsistent with studies of the microdynamics of violence, which demonstrate that even in civil wars, most of the violence which takes place is not associated with the "master cleavage" that ostensibly drives the conflict, but rather with other, localized disputes[97]:

> [It is] oft-noted . . . [that] conflicts and violence "on the ground" often seem more related to local issues rather than the "master cleavage" that drives the civil war at the national level. . . . [I]ndividuals and local communities involved in the war tend to take advantage of the prevailing situation to settle private and local conflicts whose relation to the grand causes of the war or the goals of the belligerents is often tenuous.[98]

This phenomenon—that the perpetration of violence has little relation to the conflict's master cleavage—is true even of wars that are ostensibly based on ethnic divisions.[99]

Canada and Australia Since 9/11, 19 CURRENT ISSUES CRIM. JUST. 151, 151 (2007) ("[N]egative media portrayals, together with discriminatory rhetoric, policy and practices at the level of the state create an enabling environment that signals the legitimacy of public hostility toward the Muslim communities.").

97. *See infra* note 97 and accompanying text.

98. KALYVAS, *supra* note 67, at 364–75 (citing numerous examples); *see* Kalyvas & Kocher, *supra* note 64, at 205, 208 ("[E]xisting empirical research provides little support for the validity of the exogenous cleavages claim. . . . Endogenous cleavages emerge first out of revenge and second out of a myriad of local cleavages, which are activated by the civil war."); Stephen C. Lubkemann, *Migratory Coping in Wartime Mozambique: An Anthropology of Violence and Displacement in 'Fragmented Wars'*, 42 J. PEACE RES. 493, 493 (2005) ("[P]olitical alignment during the Mozambican civil conflict (1977–92) was shaped largely by family- and community-level struggles rather than national politics . . . . [T]he means of violence of the two national parties to the civil conflict were appropriated by local actors in service to their own agendas."); Thomas M. McKenna, *Murdered or Martyred? Popular Evaluations of Violent Death in the Muslim Separatist Movement in the Philippines*, *in* DEATH SQUAD: THE ANTHROPOLOGY OF STATE TERROR 199 (Jeffrey A. Sluka ed., 2000) ("Exclusive attention to the official politics of resistance to state terror ignores the internal political complexities of such movements, especially the often camouflaged conflicts between local-level concerns of civilian supporters and the 'national' interests of movement leaders."); *see also* Balcells, *supra* note 54 (noting that, in the view of early scholars, "civil conflicts were seen as the result of existing political cleavages, and violence as the consequence of these divisions[,]" but that that understanding has been supplanted by the findings of more recent scholars, who generally use "more systematic research methods than the previous generation of scholars").

99. *See* KALYVAS, *supra* note 67, at 371 (citing JAN T. GROSS, REVOLUTION FROM ABROAD: THE SOVIET CONQUEST OF POLAND'S WESTERN UKRAINE AND WESTERN BELORUSSIA (1988); then citing MOHAND HAMOUMOU, ET ILS SONT DEVENUS HARKIS (1993); and then citing PAUL RICHARDS, FIGHTING FOR THE RAIN FOREST: WAR, YOUTH AND RESOURCES IN SIERRA LEONE 6 (1996));

This understanding of the locus and local source of the actual violence that is perpetrated in the course of political violence clearly undermines the argument that hate speech is a cause of ethnic violence, because if most violence in ostensible "ethnic wars" is not related to the supposed cleavage between the two groups, then hate speech which targets one group logically cannot be the cause of that violence.[100]

C. Violence is Generally Instrumental and Elite-Driven, Rather than Spontaneous and "Bottom-Up"

The argument that hate speech causes violence by engendering hate assumes that political violence is a hate-driven, bottom-up phenomenon. However, most scholars find that political violence is instead a project of elites, and is driven by instrumental motives, rather than by hatred or other normative factors.[101] Thus, scholars of genocide agree that "the choice of genocide

---

GEOFFREY ROBINSON, "IF YOU LEAVE US, HERE WE WILL DIE": HOW GENOCIDE WAS STOPPED IN EAST TIMOR 1 (2010) (arguing that the Indonesian central government could not carry out mass violence without local militia, which were often driven by local factors rather than the master cleavage of the conflict); *see also* Rogers Brubaker, *Ethnicity Without Groups*, 43 EUR. J. SOC. 163, 176 (2002) ("What is represented as ethnic conflict or ethnic war—such as the violence in the former Yugoslavia, may have as much or more to do with thuggery, warlordship, opportunistic looting and black-market profiteering than with ethnicity."); John Mueller, *The Banality of "Ethnic War,"* 25 INT'L SECURITY 42, 42 (2000) ("Specifically, insofar as it is taken to imply a war of all against all and neighbor against neighbor—a condition in which pretty much everyone in one ethnic group becomes the ardent, dedicated, and murderous enemy of everyone in another group—ethnic war essentially does not exist. . . . [E]thnic warfare more closely resembles nonethnic warfare, because it is waged by small groups of combatants, groups that purport to fight and kill in the name of some larger entity.").

100. *See* Kalyvas & Kocher, *supra* note 64, at 204, 206 ("We identify two major theoretical claims about the ways in which cleavages are connected with violence. The first posits that violence in civil wars flows primarily from preexisting and deep animosities . . . . The link between prewar polarization and violence implies an underlying theory of action in two steps: (1) a person is victimized because of her membership in a group that (2) is targeted because of its position on the dimension that motivates the conflict. In this formulation, prewar polarization explains both why a group is targeted and why its members are victimized. This link is usually assumed rather than subjected to empirical investigation. . . . However, this inference is based on a premise akin to that of ecological fallacy: in the absence of individual-level data about particular acts of violence, we tend to extrapolate from the aggregate down to the individual level. This extrapolation can be and often is fallacious.").

101. MADELINE K. ALBRIGHT & WILLIAM S. COHEN, PREVENTING GENOCIDE: A BLUEPRINT FOR U.S. POLICYMAKERS 36 (2008) ("[M]ass atrocities are generally perpetrated when underlying risk factors . . . are exploited by opportunistic elites seeking to amass power and to eliminate competitors."); KEITH SOMERVILLE, RADIO PROPAGANDA AND THE BROADCASTING OF HATRED: HISTORICAL DEVELOPMENT AND DEFINITIONS 217 (2012) (noting that the media's interpretation of the 2007 post-election violence in Kenya as "ethnic" served the interests of elites who wanted the violence to be seen as spontaneous and driven by ethnic enmity in order to disguise their role in organizing it); Shiping Tang, *The Onset of Ethnic War: A General Theory*, 33 SOC. THEORY 256, 259 (2015) ("[M]anipulation by the elite of ethnic identity, fear, and hatred is the most crucial process driving ethnic politics toward ethnic war."); Valentino, *supra* note 54, at 96 (stating that scholarly consensus is that "elites promote violence against civilians to obtain private political or material benefits or to achieve ideological

emerges over time"[102] as "the best available strategy" for obtaining state goals.[103] For example, even the Nazi extermination of Jews was adopted only after less extreme measures—such as forced resettlement—were rendered impossible by military and political setbacks.[104] Similarly, communal riots are generally understood to take place "in a context of intense political mobilization or electoral competition in which riots are precipitated as a device to consolidate the support of ethnic, religious, or other culturally marked groups by emphasizing the need for solidarity in face of the rival communal group."[105]

---

goals"). *But see* Green & Seher, *supra* note 84, at 523 (noting that, "[i]n part, this [majority view] . . . reflects an inherent historiographic bias in accounts that emphasize elite maneuvering over the harder-to-measure dynamics in mass opinion and behavior."); *accord* Stathis N. Kalyvas, *The Urban Bias in Research on Civil Wars*, 13 SECURITY STUD. 160, 169 (2004) ("Because rural-based movements and peasants do not usually leave behind many written sources, their actions are neglected or imputed to other actors who are seen as representing or manipulating them—depending on the author's political preferences.").

102. Scott Straus, *"Destroy Them to Save Us": Theories of Genocide and the Logics of Political Violence*, 24 TERRORISM & POL. VIOLENCE 544, 551 (2012).

103. *Id.* at 554; *see also* VALENTINO, *supra* note 91, at 3 ("[M]ass killing is most accurately viewed as an instrumental policy—a brutal strategy designed to accomplish leaders' most important ideological or political objectives and counter what they see as their most dangerous threats."); ERIC D. WEITZ, A CENTURY OF GENOCIDE: UTOPIAS OF RACE AND NATION 236–37 (2003) (noting that in Germany, Serbia, Cambodia, and the Soviet Union, the decision to engage in mass violence emerged at a time of crisis, when each regime perceived violence as necessary for achieving the regime's long-term, utopian goals); H. Zeynup Bulutgil, *War, Collaboration, and Endogenous Ethnic Polarization: The Path to Ethnic Cleansing*, *in* RETHINKING VIOLENCE 57 (Erica Chenoweth & Adria Lawrence, eds., 2010) (finding that ethnic cleansing typically occurs when a state is involved in conflict with another state, and that other state allies itself with ethnic minorities within the first state); FUJII, *supra* note 64, at 129, 185–86 (noting that Rwandans who joined the Interahamwe did not exhibit greater hatred or fear of Tutsis than those who did not join, but rather were recruited because of their personal connections with self-interested politicians); V.P. Gagnon, Jr., *Ethnic Nationalism and International Conflict: The Case of Serbia*, 19 INT'L SECURITY 130, 132 (1994/1995) ("[V]iolent conflict along ethnic cleavages is provoked by elites . . . . [B]y constructing individual interest in terms of the threat to the group, endangered elites can fend off domestic challengers . . . ."); Ernesto Verdeja, *The Political Science of Genocide: Outlines of an Emerging Research Agenda*, 10 PERSP. ON POL. 307, 310 (2012) ("[G]enocide develops as other strategies and policies are considered inadequate for addressing whatever 'threat' leaders perceive.").

104. DAVID CESARANI, FINAL SOLUTION: THE FATE OF THE JEWS 1933–1949, XXXV (2016) ("The Jews paid the price for German military failure. The preferred solution to the 'Jewish question' from 1939 to 1941 was a combination of forced emigration and expulsion. . . . German's defeat in Russia in 1941 not only removed the option of ejecting millions of Jews from areas under German control, it had a domino effect across the continent."); WEITZ, *supra* note 103, at 129.

105. PAUL BRASS, THE PRODUCTION OF HINDU-MUSLIM VIOLENCE IN CONTEMPORARY INDIA 15 (2003); *see also* Steven Wilkinson, VOTES AND VIOLENCE: ELECTORAL COMPETITION AND ETHNIC RIOTS IN INDIA 4 (2004) (noting that town-level electoral incentives account for where Hindu-Muslim violence breaks out); Raheel Dhattiwala & Michael Biggs. *The Political Logic of Ethnic Violence: The Anti-Muslim Pogrom in Gujarat, 2002*, 40 POL. & SOC'Y 483, 483 (2012) (stating that killing was less likely where the Hindu nationalist Bharatiya Janata Party (BJP) was weakest, but was even less likely where the BJP was strong; it was most likely where the party faced the greatest electoral competition); Samsu Rizal Panggabean & Benjamin Smith, *Explaining Anti-Chinese Riots in Late 20th Century Indonesia*, 39 WORLD DEV. 231, 232 (noting that anti-Chinese riots in May 1998 were a

Therefore, rather than being the spontaneous eruption of hate "from below," as is assumed by those who maintain that hate speech foments hatred and thence violence, most political violence is instrumental, and orchestrated from above, albeit often with the assistance of violence-prone members of the populace—what Paul Brass terms, "violence specialists"—who play specific roles in sparking unrest on behalf of elites.[106] This implies that neither hate speech nor hatred is a cause of political violence.[107]

IV. POLICY IMPLICATIONS

Prevention of violence is, of course, a central function of government.[108] But, so too is the preservation of liberty.[109] The censorship and punishment of ostensible hate speech carries with it a profound threat to liberty, because "[t]he reality is that governments are most often inclined to exercise their censorship powers on behalf of the powerful and other oppressive voices in society and seldom on behalf of the weak and vulnerable."[110] As Michel Rosenfeld observes:

---

frame-shifting strategy employed by security forces to distract public attention from their failure to control anti-government student demonstrations). *But see* YUKHI TAJIMA, THE INSTITUTIONAL ORIGINS OF COMMUNAL VIOLENCE: INDONESIA'S TRANSITION FROM AUTHORITARIAN RULE 8–10 (2014) (arguing that (1) repression by state security forces during authoritarian rule in Indonesia left some communities dependent on the state to maintain intercommunal security; (2) other communities which experienced less repression developed their own informal institutions to maintain security; (3) during the transition away from authoritarianism, communities in the former group experienced higher levels of communal violence until they developed informal institutions to keep the peace); Ashutosh Varshney & Joshua R. Gubler, *Does the State Promote Communal Violence for Electoral Reasons?*, 11 INDIA REV. 191, 198 (2012) (discussing methodological challenges of research on the state's role in fomenting communal violence and concluding that "it is well-known, and certainly true, that Indian states have not been unfailingly committed to their constitutional role of keeping peace. . . . But, it does not follow that the states are *always* interested in, or capable of instigating, riots for the sake of the electoral objectives of the ruling party"); Chris Wilson, *Provocation or Excuse?: Process-Tracing the Impact of Elite Propaganda in a Violent Conflict in Indonesia*, 17 NATIONALISM & ETHNIC POL. 339, 339 (2011) (arguing that "[i]t is widely recognized that many cases of violent ethno-religious conflict are preceded, if not caused, by incitement by politicians or other powerful individuals" but concluding that rioters are often not provoked by elite propaganda but rather are acting out of their own interest).

106. PAUL BRASS, THEFT OF AN IDOL: TEXT AND CONTEXT IN THE REPRESENTATION OF COLLECTIVE VIOLENCE 13-16 (1997).

107. *See id.*

108. Steven J. Heyman, *The First Duty of Government: Protection, Liberty and the Fourteenth Amendment*, 41 Duke L.J. 507, 509 (1991) ([T]he congressional debates on the Fourteenth Amendment show that establishing a federal constitutional right to protection was one of the central purposes of the Amendment.").

109. *Id.* at 510 ("With its roots in the common law tradition and social contract theory, the right to protection in life, liberty, and property became a central principle of American constitutional thought by the time of the Revolution.").

110. Carver, *supra* note 88, at 192.

> Whereas in Nazi Germany hate speech was perpetrated by the government as part of its official ideology and policy, in contemporary democracies it is by and large opponents of the government and, in a wide majority of cases, members of marginalized groups with no realistic hopes of achieving political power who engage in hate speech. Moreover, in some cases those punished for engaging in hate speech have been members of groups long victimized by racist policies and rhetoric, prosecuted for uttering race based invectives against those whom they perceive as their racist oppressors. Thus, for example, it is ironic that the first person convicted under the United Kingdom's Race Relations Law criminalizing hate speech was a black man who uttered a racial epithet against a white policeman.[111]

Indeed, the history of hate speech censorship is filled with prosecutions of ordinary individuals whose speech carries with it very little apparent risk of engendering violence.[112] Finally, many hate speech censorship efforts target speech which would seem to merit protection under even quite narrow conceptions of freedom of speech, such as Salmon Rushdie's THE SATANIC VERSES,[113] information about political issues,[114] and advocacy of changes in

111. Rosenfeld, *supra* note 47, at 1525. The first prosecution under the 1965 Race Relations Act that alleged racist speech alone was against a speaker at a "black power" meeting. Avrom Sherr, *Incitement to Racial Hatred in England*, *in* UNDER THE SHADOW OF WEIMAR, *supra* note 8, at 63, 70. Subsequently, four members of the Universal Coloured People's Association were convicted for a series of speeches in which they were alleged to have advocated that black nurses give the wrong injections to white patients. *Id.*; *see* Henry Louis Gates, Jr., *Let Them Talk*, 209 THE NEW REPUBLIC 37 (1993) (noting that among the first casualties of Canada's hate speech law was BLACK LOOKS: RACE AND REPRESENTATION by Bell Hooks which was confiscated as anti-male hate speech); *Case of Jaume Roura and Enric Stern*, GLOB. FREEDOM OF EXPRESSION, COLUM. UNIV., https://globalfreedomofexpression.columbia.edu/cases/case-of-jaume-roura-and-enric-stern (last visited Oct. 14, 2018) (showing that sentence was upheld against anti-monarchy separatists who burned photograph of monarchs at protest where act was both hate speech and an incitement to violence); Tom Phillips, *Pu Zhiqiang Given Three-Year Suspended Sentence*, THE GUARDIAN (Dec. 22, 2015, 21:21), https://www.theguardian.com/world/2015/dec/22/pu-zhiqiang-chinese-human-rights-lawyer-sentenced-to-three-years (discussing civil rights lawyer who was given three-year suspended sentence for "inciting ethnic hatred" and "disturbing public order").

112. A lengthy list of such prosecutions can be found at the author's Harvard Dataverse page. *See* Gordon Danning, *Hate Speech Prosecutions*, HARVARD DATAVERSE (Sep. 22, 2018), https://doi.org/10.7910/DVN/PGVSXB/N4FRNN.

113. Stephanie Farrior, *Molding the Matrix: The Historical and Theoretical Foundations of International Law Concerning Hate Speech*, 14 BERKELEY J. INT'L. L. 1, 61 (1996) ("In considering the French periodic report in 1989, several members of the Committee [on the Elimination of All Forms of Racial Discrimination] expressed the view that France should not have permitted publication or distribution of the controversial book by Salman Rushdie, THE SATANIC VERSES. Mr. Aboul-Nasr noted that Article 4(a) and (b) of the Convention outlaw activities inciting racial discrimination and hatred, and, 'in his view, the book had already incited hatred in the form of demonstrations and counter-demonstrations as well as events involving mosques.'").

114. *See, e.g.*, Benjamin Ducol et al., *Assessment of the State of Knowledge: Connections Between*

the law.[115]

Given this manifest risk and clear history of abuse, if hate speech is to be punished or censored at all on the ground that it increases the risk of violence, it should be punished or censored only in exceptionally narrow circumstances.[116] First, even those who contend that hate speech predicts violence are clear that that risk attaches only to very narrow categories of hate speech.[117] For example, Antoine Buyse points out that it is only "[t]he instigation of fear among one's own group, rather than hatred against the other, has been found to be a key mechanism in such processes leading to violence[,]"[118] and Rhiannon Neilson contends that the traditional emphasis on "dehumanization" as a precursor[119] to genocide is misplaced, because "dehumanization is found to exist in a variety of instances that do not lead to aggression or violence. . . . '[T]oxification' [is] a more precise early warning sign."[120] Similarly, Jonathan Leader Maynard and Susan Benesch, while

Research on the Social Psychology of the Internet and Violent Extremism 24 (Can. Network for Research on Terrorism, Sec., & Soc'y, Working Paper No. 16-05, 2016) ("In relation to radicalization, the Internet allows extremist groups to disseminate their messages and ideologies. This radical content has the potential to inspire radicalization. On one hand, they may produce a sort of 'awakening' within individuals who are becoming aware of issues for the first time. Muslims in the West may be introduced to events in areas such as Iraq, Syria, Chechnya, and Palestine.").

115. *Derby Men Jailed for Giving Out Gay Death Call Leaflets*, BBC NEWS (Feb. 10, 2012, 16:49), http://www.bbc.co.uk/news/uk-england-derbyshire-16985147 (discussing three men convicted for distributing a leaflet entitled "Death Penalty?", which showed an image of a wooden mannequin hanging from a noose, quoted Islamic texts, and said capital punishment was the only way to rid society of homosexuality).

116. *See infra* notes 116–30 and accompanying text.

117. *See generally* Jonathan Leader Maynard & Susan Benesch, *Dangerous Speech and Dangerous Ideology: An Integrated Model for Monitoring and Prevention*, 9 GENOCIDE STUD. & PREVENTION: AN INT'L J., no. 3, 2016 at 70 (suggesting an integrated framework to "help identify the sorts of speech and ideology" that increase the potential for violence).

118. Antoine Buyse, *Words of Violence: "Fear Speech," or How Violent Conflict Escalation Relates to the Freedom of Expression*, 36 HUM. RTS. Q. 779, 785 (2014).

119. The fact that hate speech can *predict* violence does not mean that it *causes* violence, because causal and predictive claims are analytically distinct from one another. *See* Galit Shmueli, *To Explain or To Predict?*, 25 STAT. SCI. 289, 290 (2010); *see also* Andrew Gelman, *How to Do a Descriptive Analysis Using Regression Modeling?*, STATISTICAL MODELING, CAUSAL INFERENCE, & SOC. SCI. (Mar. 7, 2017, 9:00 AM), http://andrewgelman.com/2017/03/07/descriptive-analysis-using-regression-models ("[R]egression [is] a predictive tool that will only give causal inferences under strong assumptions."). For example, the killing of journalists is often a precursor to, and hence a predictor of, increased levels of government repression. Anita R. Gohdes & Sabine C. Carey, *Canaries in a Coal-Mine? What the Killings of Journalists Tell Us About Future Repression*, 52 J. PEACE RES. 157, 157, 171–72. Yet, the killing of journalists obviously does not cause the subsequent repression. *See id.* at 171. Similarly, a regime that holds animus towards a group will often engage in both hateful rhetoric and violence toward that group, but the former does not cause the latter; they are both the result of the animus. Hence, the fact that increased levels of hate speech can predict the onset of mass categorical violence does not necessarily demonstrate that hate speech is a cause of that violence. *See* Section III.C.

120. Rhiannon S. Nielson, '*Toxification' as a More Precise Early Warning Sign for Genocide than*

maintaining that ideology and speech can catalyze violence, argue that "[b]oth speech and the ideology that underpins it can be dangerous (in the sense of promoting violence) without being hateful, and can also be hateful without being dangerous."[121]

Second, preventing violence cannot form a legitimate reason for censoring or punishing hate speech in any circumstances that do not involve a severe, ongoing crisis.[122] This is because the onset of collective violence is triggered not by increases in hate, but by specific events that signal or create a political threat to a group that is in power, such that perpetrators come to believe that "extreme violence is necessary to protect one's country, one's core political project, and one's primary political community against a fundamental, imminent, and usually, future danger."[123] Indeed, this understanding is consistent with such statements of international criminal law as the trial court judgment of the International Criminal Tribunal for Rwanda in the case of the singer Simon Bikindi, which found him guilty of incitement to commit genocide for

Dehumanization? An Emerging Research Agenda, 9 GENOCIDE STUD. & PREVENTION: AN INT'L J., no. 1, 2015 at 83, 83. *But see* Timothy Williams & Rhiannon Neilsen, *"They Will Rot the Society, Rot the Party, and Rot the Army": Toxification as an Ideology and Motivation for Perpetrating Violence in the Khmer Rouge Genocide?*, TERRORISM & POL. VIOLENCE (forthcoming 2016) (manuscript at 1), https://doi.org/10.1080/09546553.2016.1233873 (finding that toxification as a genocidal ideology was present in the Khmer Rouge discourse, but was not a motivating factor for individual perpetrators).

121. Maynard & Benesch, *supra* note 117, at 71 (emphasis omitted).

122. *See infra* notes 122–125 and accompanying text.

123. SCOTT STRAUS, MAKING AND UNMAKING NATIONS: WAR, LEADERSHIP, AND GENOCIDE IN MODERN AFRICA 55, 58 (2015) ("[T]he typical scenario in genocide cases is that members of the inferior category seek to change the political dispensation, which in turn cements the perception of the narrative's defenders that the interests of the two populations are inherently antagonistic and zero-sum."); *see also* JACQUES SEMELIN, PURIFY AND DESTROY: THE POLITICAL USES OF MASSACRE AND GENOCIDE 8–22 (Cynthia Schoch trans., 2007); Benjamin E. Goldsmith et al., *Forecasting the Onset of Genocide and Politicide: Annual Out-of-Sample Forecasts on a Global Dataset, 1988–2003*, 50 J. PEACE RES. 437, 439 ("[T]here is no instance of the onset of genocide (or politicide) in our time frame in which the state is not also experiencing another form of instability in the same year."); Barbara Harff, *No Lessons Learned from the Holocaust? Assessing Risks of Genocide and Political Mass Murder Since 1955*, 97 AM. POL. SCI. REV.57, 61–62 (finding that political upheaval, defined as "an abrupt change in the political community caused by the formation of a state or regime through violent conflict, redrawing of state boundaries, or defeat in international war" is a necessary, but not sufficient, condition for genocide or politicide); Straus, *supra* note 102, at 546 ("[S]trategies of mass violence are developed in response to real and perceived threats to the maintenance of political power."); Verdeja, *supra* note 103, at 310 ("[G]enocide develops as other strategies and policies are considered inadequate for addressing whatever 'threat' leaders perceive."). However, Wilson and Lillie conducted a study where participants were exposed to propaganda speeches of eight types: "calls for revenge, extreme nationalist sentiments, negative stereotyping other groups, dehumanizing language, demands for justice, references to past atrocities, victimization of his own group, and warnings of a direct violent threat to his group. . . . [O]nly the speeches that called for revenge . . . led the participants to morally justify violence. Richard Ashby Wilson & Christine Lillie, *Does Propaganda Incite Violence?* INST. LETTER, (Inst. for Advanced Study, Princeton, N.J.), Summer 2015, at 5.

statements made at the height of the genocide, but which declined to convict him of crimes related to his statements at a political rally months before the genocide commenced.[124] In fact, neither the Nuremberg nor Rwanda tribunals convicted defendants of crimes based on hate speech uttered before violence was implemented, but rather only for speech that took place when violence was already ongoing.[125]

Finally, the social science of political violence is clear that, to the extent that hate speech plays a causal role in increasing violence at all, it is only speech by elites that does so.[126] This is a fact that is explicitly recognized by many authorities on the subject,[127] and is also implicit in the understanding that violence is instrumental.[128] Moreover, even well-recognized "triggers" of mass violence, such as a "symbolically significant violation" by an out-group, which would seemingly result in quintessential "hate-based" violence, do not result in violence without the support of community elites.[129] Therefore, prosecutions of ordinary citizens for hate speech should be categorically barred.[130]

124. Prosecutor v. Bikindi, Case No. ICTR 01-72-T, Judgement, ¶¶ 183, 426 (Dec. 2, 2008) ("Accordingly, the Chamber finds that the Prosecution has failed to prove that Bikindi's actions at that meeting in early 1994 constituted anti-Tutsi propaganda or were a motivating factor in anti-Tutsi violence.").

125. U.S. HOLOCAUST MEM'L MUSEUM, HATE SPEECH AND GROUP-TARGETED VIOLENCE: THE ROLE OF SPEECH IN VIOLENT CONFLICTS 3 (2009), https://www.ushmm.org/m/pdfs/20100504-hate-speach-violence.pdf (last visited Oct. 20, 2018); *see also* Gregory S. Gordon, *Speech Along the Atrocity Spectrum*, 42 GA. J. INT'L & COMP. L. 425, 446–47 (2014) (arguing that counter-speech is ineffective as a solution to the threat posed by hate speech only in the very late stages of the path to genocide).

126. *See supra* note 101.

127. *See, e.g*, Carver, *supra* note 88, at 191 ("This balancing of rights is difficult even in circumstances where the hatred is being vocalized by minority groups with only limited power to carry out their threats."); Susan F. Hirsch, *Putting Hate Speech in Context: Observations on Speech, Power, and Violence in Kenya*, U.S. HOLOCAUST MEM'L MUSEUM 10–11 (2009), https://www.ushmm.org/m/pdfs/20100423-speech-power-violence-hirsch.pdf ("[W]hether those who speak the words of hate also hold power is crucially important in determining the potential effects of those words . . . ."); Maina Kiai, *Speech, Power and Violence: Hate Speech and the Political Crisis in Kenya*, U.S. HOLOCAUST MEM'L MUSEUM 5 (2009), https://www.ushmm.org/m/pdfs/20100423-speech-power-violence-kiai.pdf ("It does matter who speaks for speech to move to violence."); *accord* Golden, *supra* note 11, at 1174, 1184 (arguing that the United States should extradite those accused of hate speech in connection with mass atrocities where the accused held a monopoly on information at the time). *But see* T. Camber Warren, *Explosive Connections? Mass Media, Social Media, and the Geography of Collective Violence in African States*, 52 J. PEACE RES. 297, 297 (2015) (finding that centralized mass communication is associated with reduced levels of collective violence, whereas social media penetration is associated with increased levels of collective violence).

128. *See supra* Section III.C.

129. Scott Straus, *Triggers of Mass Atrocities*, POL. & GOVERNANCE, Oct. 2015, at 13.

130. See examples *supra* note 111.

IV. CONCLUSION

As Justin La Mort notes, there is an inherent danger in government efforts to prevent and deter violence:

> No one wants to allow perpetrators to escape punishment. This does not mean that in striving towards "never again" we sacrifice free speech as a casualty of war. Freedom of speech is "the indispensable condition of nearly every other form of freedom." A vague or overly expansive interpretation of incitement will be abused and misused by dictators in silencing artists, journalists, and genuine political opposition. A limited, well-defined interpretation will still allow for the intended purpose of prevention and punishment of genocide, yet respect the basic tenets of free expression.[131]

Achieving that goal requires setting aside the inevitable emotional responses to past atrocities and clearly analyzing the true benefits and costs of imposing a regime of hate speech censorship. Hence, legal scholars and advocates alike have a responsibility to familiarize themselves with the academic research in fields outside the law which casts light on those costs and benefits. Both the victims of violence and the victims of repression of conscience are owed nothing less.

131. Justin La Mort, *The Soundtrack to Genocide: Using Incitement to Genocide in the* Bikindi *Trial to Protect Free Speech and Uphold the Promise of Never Again*, 4 INTERDISC. J. HUM. RTS. L. 43, 44 (2009) (footnotes omitted) (quoting Palko v. Connecticut, 302 U.S. 319, 327 (1937)).