BERKMAN KLEIN CENTER
FOR INTERNET & SOCIETY
AT HARVARD UNIVERSITY

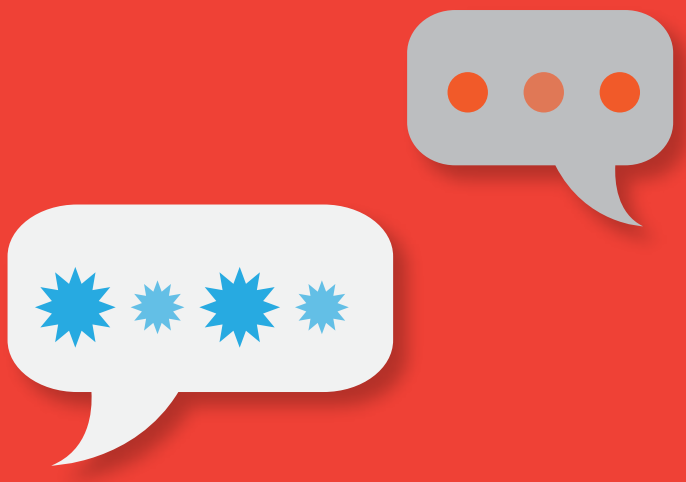
Research Publication No. 2016-20
December 2016

Defining Hate Speech

Andrew F. Sellars

23 Everett Street • Second Floor • Cambridge, Massachusetts 02138
+1 617.495.7547 • +1 617.495.7641 (fax) • http://cyber.law.harvard.edu/ •
cyber@law.harvard.edu

December 2016



Defining Hate Speech

Andrew F. Sellars

BERKMAN KLEIN CENTER
FOR INTERNET & SOCIETY
AT HARVARD UNIVERSITY

DEFINING HATE SPEECH

Andrew F. Sellars*

Suggested Citation: Sellars, Andrew F. Defining Hate Speech (December 8, 2016). Berkman Klein Center Research Publication No. 2016-20. Available at: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2882244

Harmful Speech Online Project
23 Everett Street | Second floor | Cambridge, Massachusetts 02138
+1 617.495.7547 | +1 617.495.7641 (fax)
http://cyber.harvard.edu

harmfulspeech@cyber.harvard.edu

* Formerly the Corydon B. Dunham First Amendment Fellow, Berkman Klein Center for Internet & Society at Harvard University. Currently the Director of the BU/MIT Technology and Cyberlaw Clinic at Boston University School of Law.
Thanks to Susan Benesch for many thoughtful inputs for this essay, and to my Berkman Klein colleagues Amar Ashar, Rob Faris, Amy Johnson, J. Nathan Matias, and Niousha Roshani, for the thought-provoking discussions that helped push and shape my thoughts here. Thanks also to Berkman Klein interns Prashanth Bhat, Daisy Joo, and Ofra Klein for their comments and edits.

Table of Contents

Introduction

Few pairs of words evoke such a diverse range of feelings, perspectives, and reactions as "hate speech." Calls are made for it to be embraced,[1] tolerated,[2] ridiculed,[3] targeted for counter-speech,[4] blocked on websites,[5] actionable in a civil lawsuit,[6] made criminally illegal,[7] or the basis of war crimes prosecution,[8] with no shortage of shading in between. If one can find a single point of agreement, it is perhaps that the topic is ripe for rigorous study. And with a massive amount of today's speech happening through the Internet, the trend is toward studying hate speech online.[9]

But just what is hate speech, and how will we know it when we see it online? For the great depth of discussion about the harms of it, how it is spread, the appropriate public and private responses to it, and how it should be reconciled with theories of free expression, surprisingly little work appears to have been done to define the term "hate speech" itself. Without a clear definition, how will scholars, analysts, and regulators know what speech should be targeted? In other areas of social science there has been great work done to unpack the complexity of content analysis, and the inherent context and biases that must be addressed in such analysis.[10] But when talking of hate speech, a shocking degree of the discussion — be it academic[11] or in public discourse[12] — looks solely to finding specific words or phrases that the observer believes signal the presence of hate speech. Is that a sound strategy?[13]

This essay attempts to draw out a framework by which the concept of "hate speech" can be identified and understood from a mixture of sources. It also, I hope, begins a comparison of two different theoretical fields — free speech theory and critical race theory — both of which seem to address hate speech, but without much discourse between each. Ultimately, I aim to identify several traits of hate speech that, when combined and applied to a text, can be used to isolate "hate speech" worthy of that highly loaded definition.

In writing this I do not aim to summarize the body of literature on hate speech. I do not seek to develop a theory for what hate speech does to a society; answer whether and how it should be responded to through social, political, legal, or technological channels; or summarize the role of digital communication technologies in

1 Jonathan Rauch, *The Case for Hate Speech*, The Atlantic (Nov. 2013), http://www.theatlantic.com/magazine/archive/2013/11/the-case-for-hate-speech/309524/.

2 Lee C. Bollinger, The Tolerant Society (1986).

3 Julien Mailland, The Blues Brothers *and the American Constitutional Protection for Hate Speech: Teaching the Meaning of the First Amendment to Foreign Audiences*, 21 Mich. St. Int'l L. Rev. 443, 459–60 (2013).

4 Iginio Gagliardone et al., UNESCO, Countering Online Hate Speech (2015), *available at* http://unesdoc.unesco.org/images/0023/002332/233231e.pdf.

5 Anita Bernstein, *Abuse and Harassment Diminish Free Speech*, 35 Pace L. Rev. 1, 25–27 (2014).

6 Richard Delgado, *Words That Wound: A Tort Action for Racial Insults, Epithets, and Name-Calling*, 17 Harv. C.R.–C.L. L. Rev. 133 (1982).

7 Mari J. Matsuda, *Public Response to Racist Speech: Considering the Victim's Story*, 87 Mich. L. Rev. 2320 (1989).

8 Gregory S. Gordon, *Formulating a New Atrocity Speech Offense: Incitement to Commit War Crimes*, 43 Loyal U. Chi. L.J. 281 (2012).

9 *See, e.g.*, Danielle Keats Citron, Hate Crimes in Cyberspace (2014); Arthur Gaus, *Trolling Attacks and the Need for New Approaches to Privacy Torts*, 47 U.S.F. L. Rev. 353 (2012); but see Lynn Edelman & Jon Deitrich, *Extremist Speech and the Internet: The Continuing Importance of Brandenburg*, 4 Harv. L. & Pol'y Rev. 361, 362–63 (2010) (demonstrating some skepticism towards the exceptionalism of the Internet).

10 *See*, e.g., Klaus Krippendorff, Content Analysis: An Introduction to its Methodology 22–25 (2004).

11 *See* Claire Hardaker, *Misogyny, Machines, and the Media, or: How Science Should Not Be Reported* (May 27, 2016), http://wp.lancs.ac.uk/drclaireh/2016/05/27/misogyny-machines-and-the-media-or-how-science-should-not-be-reported/ (reviewing bad syntax analysis on online hate speech by a professional research organization).

12 *See* Lori Tharps, *"Reprint Reporting" and Race*, Colum. Journalism Rev. (Sept. 25, 2013), http://www.cjr.org/minority_reports/reprint_reporting_and_race.php (criticizing the trend toward searching for and publishing Twitter posts with slurs as the heart of a news story).

13 Perhaps unsurprisingly, my view is no. *See* notes 115–28 *infra* and accompanying text.

changing hate speech. I hope only to help clarify what it is that scholars seek to study, and explain why the definitions used in such study must be approached with nuance.

The essay begins with a review of regulation of hate speech in the United States, and how such speech was defined in the brief period when such speech was considered punishable consistent with the First Amendment. As will be seen, such regulation presented serious tensions with free speech principles. Those theories of free speech are next introduced to draw out why such punishment was so fraught with difficulty. Part II of the essay looks to existing attempts to define hate speech in academia, law, and the regulation of online platforms. Part III explains why any solution or methodology that purports to present an easy answer to what hate speech is and how it can be dealt with is simply not a product of careful thinking. But that said, there are a series of criteria that may be used to help guide a researcher to identify a discrete corpus of hate speech for further study.

The Theoretical Context of Hate Speech

America's Brief Turn to Hate Speech Proscription

There cannot be serious doubt that expressed hatred based on a person's immutable characteristics, ethnic background, or religious identity causes a harm.[14] But every principle of freedom of expression recognizes times when governments should protect speech in spite of the harm that it causes.[15] There is also always an upper limit the protection of speech: every implemented system of freedom of expression finds some point where speech has such a strong nexus to harm that it can be prohibited.[16] Those three simple boundaries aside, all opinions as to the proper balance fracture.

Quite famously, the United States does not make hate speech *per se* illegal under any definition, while many other nations do.[17] These dueling perspectives are often shorthanded to the American approach versus the European approach, though this is far too simplistic a generalization.[18] There was a brief moment, however, when the United States Supreme Court did uphold a facial hate speech proscription against a First Amendment challenge.[19] The story of America's turn toward, and then away from, law-driven hate speech regulation serves as an effective illustration of the tensions at play between speech and the harm it causes.[20]

As James Whitman notes in his comparative review of dignity-related laws in France, Germany, and the United States, differing approaches to the common goal of egalitarianism have led to the absence of a protectable interest in one's honor in the United States.[21] But as he notes, the related concept of one's reputation is protected in America through defamation law.[22] Indeed,

14 *See*, e.g., Jeremy Waldron, The Harm in Hate Speech (2012); Delgado, *supra* note 6, at 136–49.

15 *See infra* notes 59–100 and accompanying text.

16 Michael Herz & Peter Molnar, *Introduction*, *in* The Content and Context of Hate Speech at 1, 4 (Michael Herz & Peter Molnar eds., 2012).

17 *See*, *e.g.*, James Q. Whitman, Enforcing Civility and Respect: Three Societies, 109 Yale L. Rev. 1279, 1281 (2000).

18 C. Edwin Baker, *Hate Speech*, *in* The Content and Context of Hate Speech, *supra* note 16, at 57, 59. Milkos Haraszti refers to this in a more nuanced way as the difference between "two very different minimums," a "minimum regulation" approach versus a "minimum of mutual respect" approach. Milkos Haraszti, *Foreward: Hate Speech and the Coming Death of the International Standard before It Was Born (Complaints of a Watchdog)* in The Content and Context of Hate Speech, *supra* note 16, at xiii.

19 Or, more formally, as a Fourteenth Amendment challenge as a violation of his substantive due process rights to free speech. *See* Beauharnais v. Illinois, 343 U.S. 250, 251–52 (1952).

20 I seek only to illustrate the tensions here; I leave to others whether this turn was normatively correct. *See*, *e.g.*, Waldron, *supra* note 14, at 34–64.

21 Whitman draws a distinction between the French and German approach of "leveling up" the dignity of all citizens to that of the classic aristocracy to the American approach of "leveling down" all its citizens, so that they "all stand together on the lowest rung of the social ladder." Whitman, *supra* note 17, at 1282.

22 I am using the concepts of honor and reputation somewhat interchangeably here. It is worth noting, as Whitman does, that the United States blends the concepts of "honor" and "reputation" in a way that European governments do not. Whitman, *supra* note 17, at 1292.

American judges speak of the rights in defamation law with the same rhetoric often used to call for hate speech proscription, defending "[t]he right of a man to the protection of his own reputation" as essential to "our basic concept of the essential dignity and worth of every human being — a concept at the root of any decent system of ordered liberty."[23] It is therefore not surprising to see many attempts to proscribe hate speech in the United Sates under the framing of defamation law, through a "group libel" claim.[24]

The historical high point for this argument was the 1952 United States Supreme Court decision *Beauharnais v. Illinois*,[25] a case concerning an Illinois statute that prohibited the publication of any piece of media that "portrays depravity, criminality, unchastity, or lack of virtue of a class of citizens, of any race, color, creed or religion," if the publication exposed that class of citizens to "contempt, derision, or obloquy which is productive of breach of the peace or riots."[26] The government of Illinois portrayed this as a "form of criminal libel" law.[27]

Joseph Beauharnais was charged under this statute after he published a racist screed that advocated against black families moving into predominantly white neighborhoods in Chicago, specifically because of the social ills that he alleged would follow.[28] Following the adjudicatory framework of libel law, Beauharnais challenged his prosecution on free speech grounds — specifically, that his statements were true, and, as the Illinois Constitution provides, true statements are a defense to libel so long as they are published "with good motives and for justifiable ends."[29] Beauharnais sought to introduce evidence that when an African-American family "moves into a district or block, real estate values immediately go down," and that "in white districts in Chicago the number of offenses reported were much smaller in number than those reported" in districts with black residents.[30]

Of course, his proffered evidence is illogical. Beauharnais's real estate factoid, even if true and causally connected, would only show that he was not alone in his racism. As for his crime statistics, a long body of criminology scholarship demonstrates why higher crime rates among African Americans are better attributed to the racial biases found throughout the criminal justice system, which repeat themselves in reports of crime.[31] A jury hearing such evidence may have found his conclusion to be false.[32] But the courts did not even allow such evidence to reach a jury. Instead, the Illinois Supreme Court held that any evidence of truth was irrelevant, because Beauharnais lacked any "good motives" or "justifiable ends."[33] The United States Supreme Court allowed that judgment to stand.[34]

It is easy to agree with the courts in *Beauharnais* at a normative level, but as a methodology for determining whether speech is actionable as hate speech it is uncomfortable and confusing. Under the approach in *Beauharnais*, it is a judge alone who decides which motives are "good,"

23 Rosenblatt v. Baer, 383 U.S. 75, 92 (1966) (Stewart, J., concurring). *See* Robert C. Post, *The Social Foundations of Defamation Law: Reputation and the Constitution*, 74 Cal L. Rev. 691, 707 (1986).

24 *See, e.g.*, Whitman, *supra* note 17, at 1292–93; Charles Lawrence, *If He Hollers Let Him Go: Regulating Racist Speech on Campus*, 1990 Duke L.J. 431, 463–64 (1990).

25 343 U.S. 250 (1952).

26 *Id.* at 251.

27 People v. Beauharnais, 97 N.E.2d 343, 346 (Ill. 1951).

28 *See id.*

29 *See* Ill. Const. art. II § 4 (1870), *as amended*, Ill. Const. art. I § 4 (1970). The modern limitation on defamation doctrine imposed by the First Amendment has done away with the "truth plus motives" defense in favor of more speech-protective defenses, but it still creeps into defamation cases from time to time. *See* Thomas Edward Powell II, *The Truth Will Not Set You Free in Nebraska: Actual Malice and Nebraska's "Truth Plus Motives" Defense*, 72 Neb. L. Rev. 1236 (1993).

30 *Beauharnais*, 97 N.E.2d at 346.

31 For a modern review of this scholarship, see Michael Rocque, *Racial Disparities in the Criminal Justice System*, 2011 Race & Justice 292 (2011).

32 At common law, a defamation defendant would be strictly liable for publishing a falsehood, except in a few particular areas. This, too, changed with the overlay of constitutional principles of defamation law beginning in the 1960s. Hon. Robert D. Sack, Sack on Defamation § 5.1 (4th ed. 2010). Beauharnais today could argue that this is a protected opinion based on disclosed facts. *See, e.g.*, Standing Comm. on Discipline v. Yagman, 55 F.3d 1430 (9th Cir. 1995).

33 *Beauharnais*, 97 N.E.2d at 347.

34 *Beauharnais*, 343 U.S. at 266.

and once a judge makes such a determination, all consideration of truth or merit is beside the point.[35] This approach does avoid putting racial questions before a jury, a famously perilous path for justice,[36] but what of the racist judge? What guides, limits, or informs the judge's determination? Could there ever be speech that demeans a group but nevertheless has "good motives"? Would Beauharnais have been allowed to offer evidence if he had published the same notion in a dispassionate academic setting?[37] What if he had targeted an individual instead of a group?[38] How is this different than adopting an "official truth," which scholars have long considered an anathema to the American approach to speech?[39] This whole line of questioning feels peculiar, because the true harm in Beauharnais's words was not his purported facts at all, but his opinion and its expression. Opinions are not so easily subjected to the true/false analysis of libel law.[40]

Subsequent cases could have worked to resolve these tensions, had the Illinois legislature not opted to frame their hate speech law as one akin to defamation. But because it was a defamation case, its future application was subject to the United States Supreme Court's series of decisions limiting that doctrine under the First Amendment.[41] The *Beauharnais* decision is now generally considered dead letter.[42] Under modern First Amendment scholarship, legal sanction of hate speech is limited to a few specific contexts, such as speech that directly incites criminal activity[43] or specific threats of violence.[44]

But this was not an easy truce. America reopened the debate on hate speech following the "Skokie controversy" of 1977, where a neo-Nazi group sought permission to hold a march through Skokie, Illinois, a predominantly Jewish suburb of Chicago.[45] While subject to robust discussion and debate at the time,[46] the enduring memory of the case is the symbolic stance that America took towards this public expression of hatred[47] — allowing the (very small) group of individuals spouting hateful rhetoric to speak, while also allowing the speakers to be surrounded by (often much larger) groups of individuals heckling, gawking at, or shouting down the hateful speakers, with law enforcement perilously guarding

35 Contemporaneous scholarship suggests that the "good motives" provision in many state defamation laws was both confusing, and presented interesting procedural challenges as to whether and in what order litigants had to offer evidence of motive and truth. *See generally* B. Sidler, *The Requirement of "Good Motives and Justifiable Ends,"* 43 Chi.-Kent L. Rev. 92 (1966).

36 *See generally* Brian J. Seer & Mark Many, *Racism, Preemptory Challenges, and the Democratic Jury: The Jurisprudence of a Delicate Balance*, 79 J.L. & Criminology 1 (1988).

37 Matsuda, *supra* note 7, at 2364–65 (questioning how hate speech law should impact the "dead-wrong social scientist").

38 *See Beauharnais*, 343 U.S. at 300–01 (Jackson, J., dissenting) (exploring this question).

39 *See* Am. Communication Ass'n v. Douds, 339 U.S. 382, 442–43 (1950) (Jackson, J., dissenting in part) ("The danger that citizens will think wrongly is serious, but less dangerous than atrophy from not thinking at all.… The priceless heritage of our society is the unrestricted constitutional right of each member to think as he will. Thought control is a copyright of totalitarianism, and we have no claim to it.").

40 *See* Robert Post, *Racist Speech, Democracy, and the First Amendment*, 32 William & Mary L. Rev. 267, 298 (1991) (hereinafter Post, *Racist Speech*); *see* Milkovich v. Lorain Journal Co., 497 U.S. 1 (1990) (discussing differences between statements of facts and opinions in defamation law).

41 This included the "good motives" limitation to the defense of truth in defamation actions. *See* Garrison v. Louisiana, 379 U.S. 64, 70–73 (1964); New York Times Co. v. Sullivan, 376 U.S. 254 (1964).

42 *See* Nuxoll v. Indian Prairie Sch. Dist., 523 F.3d 668, 672 (7th Cir. 2008).

43 Brandenburg v. Ohio, 395 U.S. 444 (1969); *see also* Edelman & Deitrich, *supra* note 9, at 370.

44 Watts v. United States, 394 U.S. 705 (1969); *see also* Elonis v. United States, 135 S. Ct. 2001 (interpreting the scope of the federal threats statute).

45 The situation played out over a series of cases in both state and federal court. *See* Collin v. Smith, 447 F. Supp. 676 (N.D. Ill.), *aff'd* 578 F.2d 1197 (7th Cir.) *cert. denied* 439 U.S. 916 (1978); Village of Skokie v. National Socialist Party, 366 N.E.2d 347 (Ill. App. 1977), *modified* 373 N.E.2d 21 (Ill. 1978).

46 Bollinger, *supra* note 2, at 14–15 ("Few legal disputes in the last decades caught the public eye with such dramatic power as did that case. For well over a year, as the case moved ponderously through the courts, it was seldom out of the news and often on the front pages of newspapers when it was in the news.").

47 "Symbolic" because despite earning the right to march, the neo-Nazis never actually marched in Skokie. Bollinger, *supra* note 2, at 27.

the border between the two factions.[48]

The trend starting at *Beauharnais*, running through Skokie, and leading to today can be viewed as an increasing division between legal and normative responses to hate speech.[49] The Court in *Beauharnais* saw the two regulatory forces[50] in complete alignment — a defendant in a libel case could only defend himself if he could show a presence of "good motives," and the goodness of the motives were determined by the presiding judicial figure. At the same time, no clear definition of "hate speech" can emerge from this formulation, given its high level of subjectivity. The Skokie case presents a near-complete division between law and norms; even those who defended the legal right of the neo-Nazis to march condemned the ideas of Nazism in the strongest normative terms: "heaping scorn and ridicule on the group to an extent limited only by shortcomings of imagination and eloquence."[51] Whether this is the better approach for hate mitigation or not,[52] the definition under this approach is equally subjective and elusive. It is the general public that decides, collectively but subjectively, what is or is not hate speech. The same problems of bias and blindness arise, but in a more social and structural manner.

The subjectivity of hate speech also emerges in the circumstances where American law *does* allow for its punishment. In fact, if prosecutors or law enforcement subjectively desire to punish hate speech they often have the tools to do so under the laws as they exist today. A critical review of cases will find examples where courts use the tremendous breadth of criminal and tort law to find a means for sanctioning the person, even when their speech is constitutionally protected.[53] And if the speaker committed a crime, the prosecutor can in some cases apply hate crimes escalations, and the First Amendment doctrine tolerates such escalation.[54] If the speech happens in a workplace, an educational institution, or the speaker is a government employee, discrimination laws provide an avenue for punishment of speech directly.[55] Because these laws are not designed to target hate speech specifically, their drafters have not labored to cast clear definitions as to what speech qualifies, but because they are so broad, a figure of authority can bring them to bear against speech they subjectively find hateful.

48 The Skokie case itself was vividly depicted in this way in the 1980 comedy *The Blues Brothers*, where the brothers confront the scene in their iconic car, leading Jake Blues to grumble "I hate Illinois Nazis" and drive at the group, forcing them to jump into a river for safety, to the cheers of onlookers. Julien Mailland argues that this scene should be used by jurists seeking to teach the differences between approaches to hate speech in France and the United States. Julien Mailland, The Blues Brothers *and the American Constitutional Protection for Hate Speech: Teaching the Meaning of the First Amendment to Foreign Audiences*, 21 Mich. St. Int'l L. Rev. 443, 451 (2013).

49 *See, e.g.*, Bollinger, *supra* note 2, at 12.

50 This "forces" analysis is drawn from Lawrence Lessig. Lawrence Lessig, Code Version 2.0 120–38 (2006).

51 Vincent Blasi, *The Teaching Function of the First Amendment*, 87 Columbia L. Rev. 387, 389 (1987) (hereinafter Blasi, *Teaching Function*). Indeed, looking purely at normative responses, the United States is often more hostile to many forms of hate speech than nations that make such speech illegal. Baker, *supra* note 18, at 60.

52 Proponents of a normative approach to hate speech regulation note that open and public confrontation can be more effective than a punishment scheme that only triggers when the government can see the speech in question. Baker, *supra* note 18, at 71, 73–75. A focus on speech may even distract law enforcement away from other, perhaps more harmful, forms of hate-motivated activity. Nadine Strossen, *Regulating Racist Speech on Campus: A Modest Proposal*, 1990 Duke L.J. 484, 495–507 (1990). Critics often rebut by noting that the historical record suggests that allowing speech to prevent violent action is problematic; escalating racist speech often accompanies escalating racist violence, especially in the all-too-often case when the population does not call out the speech. *See* Matsuda, *supra* note 7, at 2352 n.166.

53 *See, e.g.*, Fisher v. Carrousel Motor Hotel, Inc., 424 S.W.2d 627, 630 (Tex. 1967) (a restaurant patron was allowed to recover damages because, in the process of having racial slurs hurled at him, a waitress forcibly pulled a plate from his hands, thus committing a physical battery).

54 Wisconsin v. Mitchell, 508 U.S. 47 (1993).

55 *See, e.g.*, Imperial Diner, Inc. v. State Human Rights Appeal Bd., 417 N.E.2d 231 (N.Y. 1980); Harris v. Harvey, 605 F.2d 330 (7th Cir. 1976). The constitutionality of these types of discrimination actions has not been fully explored. See Eugene Volokh, *Freedom of Speech vs. Workplace Harassment Law – A Growing Conflict*, http://www2.law.ucla.edu/volokh/harass/ (last visited Nov. 27, 2016).

This is not cause for a feeling of relief, for the converse is also true. When hatred is not seen by those in a position of power they will often refuse to use the legal tools at their disposal. To take one example, a staggering volume of harmful speech online — particularly sexist and racist speech — occurs in a form that would almost certainly constitute actionable threats under federal law.[56] But federal law enforcement appears devotionally uninterested in pursuing such crimes, and local law enforcement usually lacks the resources and knowledge to meaningfully investigate such activity.[57] Mari Matsuda notes numerous studies where overt and undoubted incidents of hatred against minority groups are treated by government and law enforcement as not important, mere "pranks," or otherwise worthy of casual dismissal.[58]

Much of the debate around hate speech seems hopelessly tangled in this subjective analysis, but in order to come to more observable definitions, it is worth exploring the countervalue that is consciously or subconsciously informing these subjective definitions: freedom of expression. A brief review of the underlying theories of freedom of expression, and how they inform the procedural and structural approach American courts take to speech issues, can better frame the definitions that follow.

Speech Theory

Kent Greenawalt's *Free Speech Justifications* provides a leading exposition on the various theories of freedom of expression, their merits, and their detractors.[59] Greenawalt draws a principle of freedom of expression that he identifies as distinct from a "minimal principle of liberty," that is, the basic premise in a liberal democracy that citizens should be free to do what they wish, and that the government should only be allowed to act when the actions of a person harm the rights of another.[60] If that alone provided the basis for action, governments would have all the reason they need to regulate hate speech, however it is defined.[61] A theory of freedom of expression helps to explain why many governments, online platforms, and scholars are reluctant to take that step, in spite of the harms it causes.

Greenawalt notes that there are several different theories proffered to explain the interest in freedom of expression, and while there have been some attempts to unify the theories into a single, coherent whole,[62] the prevalent thought is that there are a series of different, valid justifications, with the array being greater than the sum of its parts.[63] Greenawalt divides the field into two general halves, "consequentialist" principles, which defend speech because of productive results of such protection, and "nonconsequentialist" principles, which defend speech due to its inherent value.[64] This section will review some of the dominant consequentialist themes that inform discussions of hate speech.

"The Marketplace of Ideas"

The dominant proffered justification — and the one most frequently expressed in the decisions of the United States Supreme Court[65] — is the "marketplace of ideas" theory, which posits that society will be better able to progress if government is kept out of the business of adjudicating what is true versus false, valid versus invalid, or acceptable versus abhorrent. Credit to this idea is usually given to John Stuart Mill, though its

56 *See* 18 U.S.C. § 875(c); *Last Week Tonight with John Oliver: Online Harassment* (HBO television broadcast June 21, 2015) (detailing specific threats women have received online that clearly fit the legal definition of true threats).

57 Citron, *supra* note 9; *see also* Ann Merlan, *The Cops Don't Care About Violent Online Threats. What Do We Do Now?*, Jezebel (Jan. 29, 2015), http://jezebel.com/the-cops-dont-care-about-violent-online-threats-what-d-1682577343.

58 Matsuda, *supra* note 7, at 2327.

59 Kent Greenawalt, *Free Speech Justifications*, 89 Columbia L. Rev. 119 (1989) (hereinafter Greenawalt, *Free Speech Justifications*). Greenawalt, it should be noted, has also written about racial insults and epithets, specifically. *See* Kent Greenawalt, *Insults and Epithets: Are They Protected Speech?*, 42 Rutgers L. Rev. 287 (1990).

60 *Id.* at 120–23.

61 Frederick Schauer, Free Speech: A Philosophical Enquiry 10 (1982).

62 *See, e.g.* Martin H. Redish, Freedom of Expression: A Critical Analysis (1984).

63 Greenawalt, *Free Speech Justifications*, *supra* note 59, at 126–27.

64 *Id.* at 127–30.

65 *See* Ronald J. Krotoszynski, Jr., The First Amendment In Cross-Cultural Perspective 21–24 (2006) (noting influences of both the marketplace theory and the self-governance theory, though the former seems to dominate the outcomes of the Court).

roots go back to John Milton's *Areopagitica* in 1644.[66] Frederick Siebert said this of Milton's theory:

> Milton was confident that Truth was definite and demonstrable and that it had unique powers of survival when permitted to assert itself in a "free and open encounter." […] Let all with something to say be free to express themselves. The true and sound will survive; the false and unsound will be vanquished. Government should be kept out of the battle and not weigh the odds in favor of one side or the other. And even though the false may gain a temporary victory, that which is true, by drawing to its defense additional forces, will through the self-righting process ultimately survive.[67]

Under this framework Joseph Beauharnais would have been at liberty to make his argument about racially segregated neighborhoods in Chicago, dubious evidence and all, and others would have been allowed to respond to him by critiquing his evidence and providing evidence of their own. And in the end the correct view would overtake Beauharnais, and maybe even change his mind personally. These principles saved the *Pittsburgh Courier* from a libel lawsuit from Beauharnais himself, after the newspaper called him "a sinister character in Chicago who is more dangerous than the nation's worst gangster," and one who "conducts a vicious and risky business — the promotion of racial hatred, with biased whites as his steady clients."[68]

Marketplace advocates in the hate speech debate also point to counterexamples of idea regulation, such as the disparate treatment of speech concerning the massacre of Armenians in 1915. In Turkey, writers have been prosecuted under hate speech laws for calling the action genocide, whereas in Switzerland a politician was prosecuted for denying the Armenians were the victims of genocide.[69] When countries construct "official truths," those truths can change as you move from country to country, presenting a perilous situation for those who wish to speak on global platforms.

The marketplace theory has no shortage of criticism. Greenawalt identifies several critiques, ranging from the deeply philosophical to the more pragmatic.[70] Chief among the latter are concerns about the state and makeup of the "marketplace" today.[71] This can be especially problematic in issues concerning the rights and dignities of marginalized groups, given the dominance of majoritarian groups in mass media.[72] And while the Internet can help diversify the possible spectrum of voices as a structural matter,[73] there is still great concern that even with those affordances power is still concentrated in majoritarian groups.[74]

More systemically, the emergence of social science concepts like bounded rationality and cognitive bias raise more fundamental problems for the theory, and whether we can trust individuals to act reasonably or rationally when engaging in debates about what is true, valid, or acceptable.[75] The marketplace theory presents a framework where the sides engaged in intellectual agon are able and willing to receive arguments to the contrary and reconsider their position, or that society can see the spectrum of debate and consider their position openly, instead of finding the evidence and truth that already supports their preexisting feelings on a

66 *See* John Milton, Areopagitica (1644), *available at* https://www.dartmouth.edu/~milton/reading_room/areopagitica/text.shtml.

67 Frederick Siebert, *The Libertarian Theory of the Press*, *in* Four Theories of the Press 39, 44–45 (Frederick Siebert et al. eds. 1956).

68 Beauharnais v. Pittsburgh Courier Publ'g Co., 243 F.2d 705, 706 (7th Cir. 1957).

69 Haraszti, *supra* note 18, at xiii, xvii.

70 Greenawalt, *Free Speech Justifications*, *supra* note 59, at 132–41.

71 *See, e.g.*, C. Edwin Baker, Media, Markets, and Democracy (2002).

72 Owen Fiss, The Irony of Free Speech 50–78 (1996).

73 *See generally* Yochai Benkler, The Wealth of Networks: How Social Production Transforms Markets and Freedom (2006).

74 *See, e.g.*, Albert-László Barabási & Réka Albert, *Emergence of Scaling in Random Networks*, 286 Science 509 (1999). For a review of the scholarly back-and-forth on this question, *see* Yochai Benkler et al., *Social Mobilization and the Networked Public Sphere*, 32 Pol. Comm. 594 (2015).

75 Greenawalt, *Free Speech Justifications*, *supra* note 59, at 134–36.

matter.[76] Both social science and informal personal experience call into question that premise.[77] Deeper examination and empirical study of the marketplace theory is sorely needed, and thanks to the more observable nature of Internet-mediated communication, is now underway.[78]

And even accepting the validity of this theory, one may find justification for restricting hateful speech on the grounds that it excludes, coerces, or frightens some would-be participators because of their affiliation with the targeted group, thus denying them access to the marketplace (as audience or speakers or both). This line of argumentation is usually ascribed first to Catharine MacKinnon, who observed the silencing effect that pornography can have on women, but has since been applied to the hate speech context particularly.[79] Those in favor of an affirmative role of the government in setting that marketplace under the marketplace theory[80] can justify some regulation of hate speech when its effect excludes speakers from the marketplace, but any attempt to prohibit topics, positions, or values on their merit is typically a nonstarter.

Democratic Self-Governance

Second in popularity among the consequentialist justifications of free speech is the "self-governance theory." This viewpoint is associated with Alexander Meiklejohn, who drew a theory of free speech by looking at both the First Amendment and the Speech and Debate Clause of the Constitution to suggest that the true goal of the constitutional protection for speech is to ensure that the public can freely and fairly debate all political issues.[81] He drew an ideological goal for public discourse as the town hall governance style in New England, where, as he put it, "what is essential is not that everyone shall speak, but that everything worth saying shall be said."[82] To borrow once again from the *Beauharnais* case, under this theory Joseph Beauharnais should be allowed to argue for the exclusion of African Americans from Chicago neighborhoods as a matter of policy, others could challenge his conclusions or epistemology, and the body politic would then be free to accept or reject his argument.[83]

Meiklejohn's theory is relatively more tolerant of state regulation of the press, though the regulation contemplated is more like the imposition of *Robert's Rules of Order* than any promotion or suppression of any given ideology.[84] Therefore, one can find reasons both to tolerate or to regulate hate speech under this framework, or tolerate some forms of hateful speech, while blocking others because, as Meiklejohn puts it, they are not "worth saying."[85] On the one hand, the concern of the silencing effect raised above has even stronger salience when considering who has access to the political process, and who can elect and persuade legislators to pass laws en-

76 For a review of some of the scholarship on these points, *see* Yochai Benkler et al., *supra* note 74.

77 *See supra* notes 71–74. This is not to mention that little of the harmful speech observed online feels as though it has little-to-no relationship to the honest offer of proof in a factual debate. Citron, *supra* note 9, at 197–99.

78 *See, e.g.*, Jennifer Brundidge, *Encountering "Difference" in the Contemporary Public Sphere: The Contribution of the Internet to the Heterogeneity of Political Discussion Networks*, 60 J. of Comm. 680 (2010).

79 *See, e.g.*, Catharine MacKinnon, *Difference and Dominance: On Sex Discrimination*, in Feminism Unmodified (1987); *see also* Post, *Racist Speech*, *supra* note 40 at 275 (noting the harm of racist speech to the marketplace of ideas); Fiss, *supra* note 72, at 16 ("In this context, the classic remedy of more speech rings hollow. Those who are supposed to respond cannot."). For a recent article applying this logic to the Internet, *see* Anita Bernstein, *Abuse and Harassment Diminish Free Speech*, 35 Pace L. Rev. 1 (2014).

80 *See, e.g.*, Jerome A. Barron, *Access to the Press: A New First Amendment Right*, 80 Harv. L. Rev. 1641 (1967).

81 Alexander Meiklejohn, Political Freedom: The Constitutional Powers of the People at 16–17 (1960).

82 Meiklejohn, *supra* note 81, at 26; *see also* John Doe No. 1 v. Reed, 561 U.S. 186, 223 (2010) (Scalia, J., concurring in judgment) (noting the influence of the New England town meeting in open governance law).

83 Post, *Racist Speech*, *supra* note 40.

84 For what it's worth, *Robert's Rules of Order* would likely sanction many forms of hateful speech. The rules stress that debate decorum not resort to personal attack. "The measure, not the member, is the subject of debate." Robert's Rules of Order 392 (11th ed. 2011).

85 My thanks to Susan Benesch for this framing observation. See Meiklejohn, *supra* note 81, at 26.

suring greater tolerance.[86] One may justify hate speech regulation on the ground that it ensures true inclusion in self-governance. On the other hand, protecting speech under this theory helps to ensure that counter-majoritarian positions will be heard, and that citizens feel an "obligation to test our thinking" about what is assumed as true.[87] Indeed, other scholars have shown how an ill-defined law against hate speech can actually be used *against* marginalized groups by majoritarian powers, to further exclude them from the political process.[88]

"The Tolerant Society"

A less dominant justification, but one with special salience in the context of harmful speech, is the "tolerance theory" most famously espoused by Lee Bollinger.[89] Bollinger's *The Tolerant Society* takes on several key questions related to extremist and hateful speech, including the question raised above as to why we are so reluctant to punish it under law, while we see no problem with normatively shunning speech, often quite aggressively.[90] Bollinger notes that the advantage of this "nonlinear structure" can allow us to better identify and respond to the underlying hatred that is expressed in hate speech. "Free speech provides a discrete and limited context in which a general problem manifests itself and in which that problem can usefully be singled out for attention."[91] And against a natural inclination of intolerance of other ideas or people, tolerated expression of other ideas "demonstrates powerfully, more powerfully than a general injunction to be appropriately tolerant in all circumstances ever would, to ourselves and others, a commitment to exercise moderation throughout social intercourse."[92]

Allowing for extremist speech forces society to confront, and thus not forget, that intolerance exists, it can rest within all of us, and we will always have work to do in order to address the spectrum of harms that societal intolerance can cause.[93] Under this framework, Beauharnais would be legally allowed to publish his screed, but only so that we all to confront the fact that people still feel this way in our society,[94] and we would be free (indeed, encouraged) to call out this racism for what it is, and commit ourselves to the ongoing struggle for equality and tolerance.

Responses to Bollinger have been mixed. Vincent Blasi noted that Bollinger's theory in some ways "is based on an expansive conception of the role of government in shaping the attitudes of the citizenry," placing it in tension with other major speech principles.[95] Matsuda focuses on harm experienced in the process of Bollinger's approach, noting that "[t]olerance of hate speech is not a tolerance borne by the community at large. Rather, it is a psychic tax imposed on those least able to pay."[96] This concern is augmented considering law enforcement's present reluctance to enforce existing laws that address threats and discrimination.[97] Moran notes that the paradigm case Bollinger adopts for his work, the Skokie case, may not resemble more recent examples of hate speech, as in that case nearly all who were involved disagreed with the viewpoints of the Nazi speakers, and in more recent cases (such as those concerning racist speech against African Americans) the intolerance can often be imputed to society more generally.[98]

Bollinger has responded to some of these objections in later scholarship.[99] Beyond his own responses, there may be some small evidence of Bollinger's theory in action. It is worth noting that the United States, despite famously avoid-

86 *See supra* note 79 and accompanying text.
87 Meiklejohn, *supra* note 81, at 73.
88 *See* Herz & Molnar, *supra* note 16, at 3.
89 *See* Bollinger, *supra* note 2.
90 *Id.* at 35–36, 109.
91 *Id.* at 121.
92 *Id.* at 123.
93 *Id.* at 129.
94 Which, sixty years later, is still true. *See* Daniel Denvir, *It's Mostly White People Who Prefer to Live in Segregated Neighborhoods*, CityLab (June 25, 2015), http://www.citylab.com/housing/2015/06/its-mostly-white-people-who-prefer-to-live-in-segregated-neighborhoods/396887/.
95 Blasi, *Teaching Function*, *supra* note 51, at 413.
96 Matsuda, *supra* note 7, at 2323.
97 *See supra* notes 57; Matsuda, *supra* note 7, at 2338.
98 Mayo Moran, *Talking About Hate Speech: A Rhetorical Analysis of American and Canadian Approaches to the Regulation of Hate Speech*, 1994 Wisc. L. Rev. 1425, 1452 n.113 (1994).
99 Lee C. Bollinger, *The Tolerant Society: A Response to Critics*, 90 Columbia L. Rev. 979 (1990).

ing legal punishment for hate speech, has been the source of what seems to be the deepest and most voluminous literature addressing hate speech as a topic of public concern.[100] The contours of acceptable behavior are debated openly and vigorously every day, and the recent increase of attention to issues misogyny, racism, and other intolerance online could well be seen as exposing a state of affairs that always existed, but was never discussed. Under a theory like Bollinger's, this could be our system working itself to a more tolerant place. The human cost of this process, however, is real, and its depth remains to be seen.

Other Theories and Procedural Values

These are just three of several theories Greenawalt examines in his sweeping review of the justifications for freedom of expression. Others include the value in having a society overtly accommodate a plurality of interests and perspectives,[101] the importance of speech as a check on abuses of governmental power,[102] and the value exposure to a variety of opinions can be for autonomous development.[103] Some of these militate against suppression of hate speech on their own, and some instead serve as warnings against how laws against hate speech could be abused when in the wrong hands. Many of the same concerns and responses addressed to the theories above can be brought to these justifications as well.

Moving from theory to praxis, an underappreciated aspect of the American approach to freedom of expression is the process by which courts address free speech issues, and the procedures that are implemented to safeguard speech. Though the importance of this concept to the drafting of the First Amendment is frequently debated,[104] concerns over prior restraint of speech are undoubtedly a hallmark of free speech jurisprudence in the United States.[105] Courts are extremely reluctant to endorse regimes of speech enforcement that remove speech before it is actually adjudicated to be illegal or actionable, even when the speech falls into a proscribable category.[106] A series of cases around obscenity prosecution in the United States have ensured that judgments about what speech is unlawful cannot be made by law enforcement agents in the field,[107] that governments cannot interfere with the dissemination of allegedly unlawful speech without judicial review,[108] and that seizure of a book or film to obtain evidence of illegal speech cannot effectively remove the speech from circulation.[109] This procedural protection ensures against over-censorship (accidental or deliberate), but also safeguards free speech principles against the exigencies of any given period, when governments may be strongly inclined to censor speech for precisely the wrong reasons.[110]

Similar motivations guide the general approach

100 Friedrich Kübler, *How Much Freedom for Racist Speech?: Transnational Aspects of a Conflict of Human Rights*, 27 Hofstra L. Rev. 335, 347 (1998).

101 *See* Greenawalt, *Free Speech Justifications*, *supra* note 59, at 141–42.

102 This view is most commonly cited to Vincent Blasi, *The Checking Value in First Amendment Theory*, 1977 Am. Bar Found. Research J. 523 (1977); *see also* Greenawalt, *Free Speech Justifications*, *supra* note 59, at 142–43.

103 *See* Greenawalt, *Free Speech Justifications*, *supra* note 59, at 143–45.

104 *See generally* Leonard Levy, Emergence of a Free Press (1985).

105 Thomas I. Emerson, *The Doctrine of Prior Restraint*, 20 Law & Contemp. Problems 648 (1955).

106 *See, e.g.*, Ctr. for Democracy & Tech. v. Pappert, 337 F. Supp. 2d 606, 657 (E.D. Pa. 2004); Dawn C. Nunziato, *How (Not) to Censor: Procedural First Amendment Values and Internet Censorship Worldwide*, 42 Georgetown J. of Int'l Law 1123, 1128–29 (2011) (reviewing procedural safeguards under First Amendment jurisprudence).

107 Marcus v. Search Warrants of Property, 367 U.S. 717, 731–32 (1961).

108 Bantam Books, Inc. v. Sullivan, 372 U.S. 58, 69–71 (1963).

109 Ft. Wayne Books, Inc. v. Indiana, 489 U.S. 46, 62–65 (1989); Roaden v. Kentucky, 413 U.S. 496, 503–04 (1973).

110 For just a few of many articles on this concept, *see* Nunziato, *supra* note 106; Jack Balkin, *Old School/New School Speech Regulation*, 127 Harv. L. Rev. 1 (2014); Vincent Blasi, *The Pathological Perspective and the First Amendment*, 85 Colum. L. Rev. 449 (1985); Henry P. Monaghan, *First Amendment "Due Process"*, 83 Harv. L. Rev. 518, 537–38 (1970). I have previously written about this in the context of copyright infringement actions. *See* Andrew Sellars, *Seized Sites: The In Rem Forfeiture of Copyright Infringing Domain Names* (2011), http://papers.ssrn.com/sol3/papers.cfm? abstract_id=1835604.

that the Supreme Court takes to new forms of content regulation by the legislative and executive branches, often called the "categorical approach."[111] The approach is distinguished from a general balancing test, where a court would assess any given speech against the harms it causes.[112] This categorical approach further guards against improper use of speech laws, as it avoids the ideological predispositions or paranoia that can permeate a balancing test.[113]

Attempts to Define Hate Speech

This is a difficult background from which to draw an objective definition of hate speech that gets at the "bad" forms of speech while leaving out the "good." The wide variety of approaches to the term are daunting to summarize,[114] but by reviewing them and the context in which they sit, scholars may be better equipped to appreciate the complexity of what they endeavor to study. This section reviews existing attempts to define such speech, and draws out some key similar themes.

Can We Know It When We See It?

Before turning to definition, there is one last piece of complication that helps to explain why defining the most important or worst hate speech can be so difficult. Before the Supreme Court settled on a formal definition for obscenity, Justice Stewart famously summarized his feelings on identifying obscenity as "I know it when I see it."[115] What makes hate speech even more difficult than obscenity is that for hate speech Justice Stewart's statement is less likely to be true.

In fact, the consensus view appears to be that a wide array of different forms of speech could or could not fit a definition of "hate speech," depending on the speech's particular context, which rarely makes it into the definition itself.[116] Looking to content of speech, epithets and insults may be easy to define and identify, but an epithet devoid of context may lead a scholar to see hate speech where the speaker, recipient, and subject of discussion may not.[117] Coded speech can be especially corrosive to a group's dignity, but can be hard to see unless one knows to look for it.[118] As Henry Louis Gates, Jr. puts it, it is wrong "to spend more time worrying about speech codes than coded speech."[119] The rhetoric of hatred employed can be quite varied, even within one type of hate speech.[120] To make matters worse, certain online hate groups have resorted to using steganography in their online communications, putting symbolic markers on online speech in order to identify a hate speech target to other members of the group,[121]

111 *See* Gregory P. Magarian, *The Marrow of Tradition: The Roberts Court and Categorical First Amendment Speech Exclusions*, 56 Wm. & Mary L. Rev. 1339 (2015).

112 *See* Leslie Kendrick, *Content Discrimination Revisited*, 98 Va. L. Rev. 231 (2012).

113 John Hart Ely, *Flag Desecration: A Case Study On the Roles of Categorization and Balancing in First Amendment Analysis*, 88 Harv. L. Rev. 1482, 1501 (1975).

114 As Robert Post put it when addressing "racist speech," the term "probably has as many different definitions as there are commentators, and it would be pointless to pursue its endlessly variegated shades of meaning." Robert Post, *Racist Speech*, *supra* note 40, at 271.

115 Jacobellis v. Ohio, 378 U.S. 184, 197 (1964) (Stewart, J., concurring). Catharine MacKinnon has noted that the phrase is "more interesting than it is usually taken to be," as it is both a reflection of how Justice Stewart reaches the question, and a reflection of his position of power – "his seeing determines what obscenity is in terms of what he sees it to be." Catharine MacKinnon, *Pornography, Civil Rights, and Speech*, 20 Harv. C.R.-C.L. L. Rev. 1, 3, (1985).

116 *See, e.g.*, Bhikhu Parekh, *Is There a Case for Banning Hate Speech?*, in The Content and Context of Hate Speech, *supra* note 16, at 37, 40.

117 See Delgado, *supra* note 6, at 179–80.

118 See Parekh, *supra* note 116, at 41.

119 Henry Louis Gates, Jr., *War of Words: Critical Race Theory and the First Amendment*, in Speaking of Race, Speaking of Sex: Hate Speech, Civil Rights, and Civil Liberties 17, 47 (Henry Louis Gates, Jr. et al. eds. 1995).

120 *See* Priscilla Marie Meddaugh & Jack Kay, *Hate Speech or "Reasonable Racism?"* The Other in Stormfront, 24 J. Mass Media Ethics 251 (2009) (looking at the various rhetorical tactics with one white supremacist group).

121 *See* Cooper Fleishman & Anthony Smith, *(((Echoes))), Exposed: The Secret Symbol Neo-Nazis Use to Target Jews Online*, Tech.mic (June 1, 2016),

or adopting benign terms as code words for racial slurs.[122] Once identified, online tools can surface such speech quickly, but one would have to know to look for it first.

Identifying which types of hateful expression are the most harmful can be especially elusive, for the nexus between any given hateful speech and a harmful consequence can be so hard to identify *ex ante*. The situation with the *Innocence of Muslims* video in 2012 is a powerful example. The video was created under false pretenses by Nakoula Basely Nakoula, casting and producing a film called *Desert Warrior* before taking the film to another group to add in vitriolic anti-Islam content over the original footage.[123] The video was so poorly produced and overdubbed that I suspect most viewers (or certainly most American viewers) would never take it seriously, and the film lingered in obscurity online for months, before another figure, notorious anti-Islamic figure Terry Jones, drew attention to it for an event on the anniversary of the September 11th attacks. Islam opponent Morris Sadek saw it through Jones, dubbed the film into Arabic (and thus removed the obvious overdubs in the original), and then shared it with colleagues in Egypt. Once it was in Egypt, al Nas Television personality Khaled Abdullah used the film to promote his longstanding narrative that the United States is at war with Islam, and riots ensued.[124]

Whose speech along the chain is worthy of identification as "hate speech?" Nakoula for creating the film? Jones for amplifying its message with intent to incite hatred? Sadek for misleading Arabic audiences into thinking it was anything more than a hokey overdub of a pathetic movie? Abdullah for using the film as part of a longer narrative of hatred? What of the news coverage that extensively debated the film and, in so doing, showed clips of the film again and again? What of the litany of scholars who reviewed and discussed the film when a curious turn of events found the film being examined in federal court in a copyright dispute?[125] All of these people repeated the hateful message in a modified way. Whose dissemination or modification should be the focus?[126]

Given these difficulties, it is no surprise that most scholarship on hate speech starts not with a definition, but with examples.[127] This is useful as an empathetic framing — researchers must not forget that there are people directly affected by the definitional choices that are made in this space[128] — but the lack of definitions in scholarship translates to uncertain definitions in law and social science research, and even more uncertain application of principles in online spaces. To the extent these groups have endeavored to identify hate speech, they are reviewed below.

Academic Attempts

Academics define hate speech for a variety of different ends, and their particular motivations often drive the approach they take to a definition. Some do not overtly call for legal sanction for such speech and seek merely to understand the phenomenon;[129] some do seek to make the speech illegal, and are trying to guide legislators and courts to effective statutory language;[130] some are in between.[131] No doubt all definers here

https://mic.com/articles/144228/echoes-exposed-the-secret-symbol-neo-nazis-use-to-target-jews-online#.KANtlJ0Pk.

122 Nikhil Sonnad, *Alt-Right Trolls are Using These Code Words for Racial Slurs Online*, Quartz (Oct. 1, 2016), http://qz.com/798305/alt-right-trolls-are-using-googles-yahoos-skittles-and-skypes-as-code-words-for-racial-slurs-on-twitter/.

123 Garcia v. Google, Inc., 786 F.3d 733 (9th Cir. 2015) (en banc).

124 *See* Rebecca MacKinnon & Ethan Zuckerman, *Don't Feed the Trolls*, Index on Censorship (Dec. 3, 2012), http://www.indexoncensorship.org/2012/ 12/dont-feed-the-trolls-muslims/.

125 *Garcia*, 786 F.3d 733; Rebecca Tushnet, *Performance Anxiety: Copyright Embodied and Disembodied*, 60 J. Copyright Soc'y 209 (2013).

126 This concern was central to a European Court of Human Rights case, *Jersild v. Denmark*, where a journalist who interviewed members of a racist group in Denmark was charged as an aider and abettor of the hate speech made by the interviewees. A divided ECHR reversed the conviction. 298 Eur. Ct. H.R. (ser. A) (1994).

127 *See, e.g.*, Waldron, *supra* note 14, at 2–3; Matsuda, *supra* note 7, at 2320–21.

128 Matsuda, *supra* note 7, at 2327.

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

130 *See, e.g.*, Matsuda, *supra* note 7.

131 *See* Parekh, *supra* note 116, at 46 (defining hate

feel as though some hate speech will fall outside their definition, but one assumes each intended to cover the most prevalent or egregious examples.[132]

Many assessments of the academic landscape begin with Richard Delgado's highly influential article *Words that Wound*.[133] Delgado begins with a review of a tension between cases like the Skokie case, where racist speech is awarded full protection, and cases that have allowed for compensation for victims of racial harassment in the workplace and elsewhere.[134] Delgado focuses on racism in particular, and makes a detailed argument for why law should sanction racist speech. In crafting a proposed tort for racist speech, Delgado proposed a definition that would require the plaintiff to prove: (1) *that "[l]anguage was addressed to him or her by the defendant that was intended to demean through reference to race;"* (2) "that the plaintiff understood as intended to dem*ean through reference to race; and"* (3) "*that a reasonable person would recognize as a racial insult.*"[135] Unpacking Delgado's definition, it is notably elusive on criteria for content, and instead focuses primarily on intent, impact, and objective perception: the intent of the speaker to "demean through reference to race," the impact that the tort victim understood the speech as it was intended, and that a "reasonable person" (a popular, if controversial, abstract figure in tort and criminal law[136]) could identify the speech as a racial insult, one assumes making that assessment in the context in which it is spoken.

Mari J. Matsuda builds upon Delgado's work, and looks at hate speech as a criminal matter, placing the issues of hate speech in the context of greater structural analysis of law and inequality which she calls a study of "outsider jurisprudence,"[137] or a critical examination of how law can be "both a product and a promoter of racism."[138] Matsuda also engages directly with critics who raised objections under the First Amendment, and crafts an approach that she believed satisfies those objections. Her approach requires racist speech to be actionable if: (1) *the message is "of racial inferiority;"* (2) *the message is "directed against a historically oppressed group;"* and (3) *the message is "prosecutorial, hateful, and degrading,"* which Matsuda later clarifies has an intent-element within it.[139] Compared to Delgado, Matsuda appears to adopt a more overt content formulation in the first element of this definition. To be actionable the speech must "den[y] the personhood of target group members," and treat all members of the targeted group as "alike and inferior."[140] Matsuda also limits her definition to speech targeting historically oppressed or subordinated groups. Finally, the speech is actionable if it "is, and intended as" speech that is harmful.[141]

Calvin Massey engaged directly with Matsuda's definition as part of a longer assessment of how hate speech fits into a theory of freedom of expression as a theoretical matter.[142] Massey notes that "most definitions tend to prejudice the discussion, by defining the term in a way that shapes, if not predetermines, the outcome, or by using terms laden with subjectivity."[143] He settles on a definition for study that "hate speech is *any form of speech that produces the harms which advocates for suppression ascribe to hate speech: loss of self-esteem, economic and social subordination, physical and mental stress, silencing of the victim, and effective exclusion from the political arena.*"[144] He notes that his approach "treat[s] all racists the same; polite, civil and unconscious racists are considered here to be no less malignant than the vulgar, nasty, and

speech that should be subject to sanction, but arguing that "law must be our last resort").

132 This review also no doubt overlooks many other efforts to define hate speech, but proceeds in the hope that the examples presented here can sketch the appropriate contours.

133 Delgado, *supra* note 6.

134 *Id.* at 133 (citing Contreras v. Crown Zellerbach, Inc., 565 P.2d 1173 (Wash. 1977)).

135 Delgado, *supra* note 6, at 179.

136 *See, e.g.*, Victoria Nouse, *After the Reasonable Man: Getting Over the Subjectivity/Objectivity Question*, 11 New Crim. L. Rev. 33 (2008).

137 Matsuda, *supra* note 7, at 2323.

138 *Id.* at 2325 (citing Derrick Bell, And We Are Not Saved (1987)).

139 *Id.* at 2357–58.

140 *Id.* at 2358.

141 *Id.*

142 Massey, *supra* note 129.

143 *Id.* at 105 n.2.

144 *Id.*

brutal ones."[145] This suits his framing as a theoretical study, instead of an effort to create a workable framework for punishment.

Mayo Moran also engaged with Matsuda's definition, and looked to both public debate of hate speech and regulation of such speech in the United States and Canada.[146] Moran was skeptical that one could present an effective description of the problem, noting that "no pre-theoretical description exists for anything, much less for complex social and cultural phenomena. So, describing the problem already involves choices and commitments that favor a certain way of seeing the world, that makes some arguments more persuasive than others, some cases more relevant, and some facts easier to 'find.'"[147] Nevertheless, she sketches "certain elements" that occur in definitions in other literature and in areas of law,[148] and settles on a definition for study as "*speech that is intended to promote hatred against traditionally disadvantaged groups.*"[149] Moran does not adopt a content element in her definition, and is distinct from Matsuda's on intent, in that it looks to speech that "promotes hatred" instead of speech that is itself hateful. This could therefore extend study to speech that is more subtle and coded, so long as it is still intended to "promote hatred."[150] The use of the word "promote" instead of "incite" is also noteworthy, in that it gives the impression of the less imminent standard than "incitement" normally carries.

Kenneth Ward engaged in further analysis a few years later, and specifically looked to times where judges uphold speech restrictions with the goal of promoting more speech overall.[151] He defined hate speech as "*any form of expression though which speakers primarily intend to vilify, humiliate, or incite hatred against their targets.*"[152] This can be seen as a blend of the intents put forth by Matsuda and Moran, covering both speech that is directly hateful and speech that incites hatred in others.[153] There is also a degree of magnitude in Ward's definition, noting that a speaker should be seen as employing hate speech if "their attacks are so virulent that an observer would have great difficulty separating the message delivered from the attack against the victim."[154] This is in some ways a reflection of a "no redeeming purpose" element, discussed further below."[155]

Susan Benesch has looked to a particular subset of hate speech that is more directly linked to the incitement of mass violence, which she calls "dangerous speech."[156] Rather than aim for an elements-based definition, Benesch looks to five "variables" that are relevant for determining the dangerousness of the speech, whether (1) *there is a "powerful speaker with a high degree of influence;"* (2) *there is a receptive audience with "grievances and fear that the speaker can cultivate;"* (3) *a speech act "that is clearly understood as a call to violence;"* (4) *a social or historical context that is "propitious for violence, for any of a variety of reasons;"* and (5) *an "influential means of dissemination."*[157] Benesch's formulation is noteworthy in that it spends considerably more energy on the context in which speech is happening, and what sorts of societal or environmental descriptors or qualities should be of most concern.[158] Her formulation is specifically targeted toward incidents of mass violence and the speech that causes such violence, but similar contextual descriptors could likely be developed for the specific harms other scholars seek to study.

145 *Id.*

146 Moran, *supra* note 98.

147 *Id.* at 1428.

148 *Id.* at 1429.

149 *Id.* at 1430.

150 See *id.*

151 Kenneth D. Ward, *Free Speech and the Development of Liberal Virtues: An Examination of the Controversies Involving Flag Burning and Hate Speech*, 52 U. Miami L. Rev. 733 (1998).

152 Ward, *supra* note 151, at 765.

153 See *id.*

154 *Id.* at 766.

155 See infra notes 312–19.

156 Susan Benesch, *Proposed Guidelines for Dangerous Speech*, Dangerous Speech Project (Feb. 23, 2013), http://dangerousspeech.org/guidelines/.

157 *Id.* An earlier version of Benesch's formulation rearranged some of these elements, and looked also to whether the "marketplace of ideas" is still functioning. See Susan Benesch, *Vile Crime of Inalienable Right: Defining Incitement to Genocide*, 48 Va. J. Int'l L. 485, 519–25 (2008).

158 *See generally* Jonathan Leader Maynard & Susan Benesch, *Dangerous Speech and Dangerous Ideology: An Integrated Model for Monitoring and Prevention*, 9 Genocide Studies & Prevention 70 (2016).

Bhikhu Parekh engaged directly with the question of whether one could develop a workable definition of hate speech for the purposes of regulation. He drafted a list of eleven examples of speech that may or may not be hate speech and then developing a framework for triaging those examples.[159] He began by noting the diversity in the examples he created:

> Some of them express or advocate views but do not call for action. Some are abusive or insulting but not threatening. Some express dislike of a group but not hatred, and some of those that do are so subtle as not to be obviously abusive or insulting. Some take a demeaning or denigrating view of a group but wish it no harm and even take a [patronizingly] indulgent attitude toward it.[160]

He went from there to distill "three essential features" to form a definition of hate speech: (1) "*it is directed against a specified or easily identifiable individual or, more commonly, a group of individuals based on an arbitrary or normatively irrelevant feature;*" (2) *the speech* "*stigmatizes the target group by implicitly or explicitly ascribing to it qualities widely regarded as undesirable;*" and (3) "*because of its negative qualities, the target group is viewed as an undesirable presence and a legitimate object of hostility.*"[161] Interestingly, Parekh's definition appears to lack any requirement that the speaker intend harm, and thereby sweeps in forms of expression that would raise serious issues under free speech theories, such as the criminology scholarship about disparate reported crime rates between different races discussed above, which discusses differences between races in a dry, scientific manner.[162]

Finally, in 2014, Alice Marwick and Ross Miller took an extensive review of existing definitions, including those by Massey and Ward, and came up with three general elements that are used to define hate speech: (1) *a content-based element*, (2) *an intent-based element*, and (3) *a harms-based element*.[163] On the content side, they look specifically to the use of symbolism to convey a hateful message, and the Supreme Court's assessment of such symbolism in *R.A.V. v. City of St. Paul*.[164] On the intent element, they look specifically toward a subjective intent "only to promote hatred, violence or resentment against" a marginalized group or member of a marginalized group, "merely because of the status of the minority."[165] On the harms formulation, Marwick and Miller largely draw from Massey, and add on a requirement that the recipient must subjectively experience the harm.[166] Perhaps most interestingly, Marwick and Miller's definition supposes a "no redeeming purpose" element similar to Ward's above, noting several times that their definition should extend to speech that is "intended only to promote hatred," targeting of a person "merely because of the status of the minority," and done "without communicating a legitimate message."[167]

Legal Attempts

A growing number of countries seek to regulate speech directly, and in so doing have endeavored to define such speech in their criminal codes.[168] The similarities and differences between these definitions bear an interesting reflection of how each society values different aspects of hate speech regulation.[169] Jeremy Waldron, author of one of the current leading books to defend hate speech regulation, largely looks to

159 Parekh, *supra* note 116, at 38–39.

160 *Id.* at 40.

161 *Id.* at 40–41.

162 *See supra* note 31.

163 Alice Marwick & Ross Miller, Online Harassment, Defamation, and Hateful Speech: A Primer of the Legal Landscape (2014), *available at* http://ir.lawnet.fordham.edu/cgi/viewcontent.cgi?article=1002&context=clip.

164 505 U.S. 377 (1992); *see also* Charlotte Taylor, *Hate Speech and Government Speech*, 12 U. Pa. J. Const. L. 1115 (2010).

165 Marwick & Miller, *supra* note 163, at 17.

166 *Id.* at 17.

167 *Id.*

168 Haraszti, *supra* note 18, at xiv.

169 I select British and formerly-British-colonial governments for this section, as the relatively similarity in the systems to the United States can better lend to easy comparison. Of course, such a quick review does not fully explain the similarities and differences between these approaches and how they are enforced, but I hope the review can surface some interesting priorities for approaching these definitions.

these definitions to frame his argument.[170]

To begin with the now-defunct American example, the Illinois statute that was initially upheld in the *Beauharnais* decision stated that it would be unlawful for any person to make a public[171] dissemination or presentation "*which publication or exhibition portrays depravity, criminality, unchastity, or lack of virtue of a class of citizens, of any race, color, creed or religion, which said publication or exhibition exposes the citizens of any race, color, creed or religion to contempt, derision, or obloquy or which is productive of breach of the peace or riots.*"[172] There appear to be two independent ways one can be liable under this statute: (1) disseminating speech that portrays a group as having a bad virtue and thereby expose a member of a group to experience "contempt, derision, or obloquy;" or (2) disseminating similar speech that causes a breach of the peace. This all should be understood within the greater common law libel framework, which usually included within it an intent element.[173]

Canada has provisions which prohibit hate speech in a way similar to the former American law. The Criminal Code of Canada punishes anyone who "*willfully promotes hatred against any identifiable group,*" but specifically excludes from this definition statements that are proven by the defendant to be true,[174] statements that are offered "in good faith," when expressing "an opinion on a religious subject," statements that are "relevant to the public interest, the discussion of which was for the public benefit," or if "in good faith," the person was pointing out other hate speech "for the purpose of removal."[175] Canada also separately prohibits those who "*communicat[e] statements in a public place*" speech that "*incite[] hatred against any identifiable group,*" but only when such incitement "*is likely to lead to a breach of the peace.*"[176] Canada also specifically punishes anyone "who advocates or promotes genocide," with "genocide" defined as acts "committed with intent to destroy in whole or part any identifiable group."[177] Targeted groups for all three actions can include groups identified by color, race, religion, national or ethnic origin, age, sex, sexual orientation, or mental or physical disability.[178]

The United Kingdom has had a specific prohibition against hate speech since the Race Relations Act of 1965, which prohibited incitement to discrimination or incitement to racial hatred.[179] Notably, this law originally required a showing that the defendant intended to incite hatred, but this element was later removed.[180] Under the current law, the Public Order Act of 1986 (as amended), prohibits a person who disseminates or displays any speech that is "*threatening, abusive or insulting,*" if "*he intends thereby to stir up racial hatred,*" or if "*having regard to all the circumstances racial hatred is likely to be stirred up thereby.*"[181] This formulation accounts for deliberate harm as well as negligent harm, a more objective measurement that does not require interrogation into the defendant's motives.

Australia has a blend of federal and state laws that address hate speech. Federally, the Racial Discrimination Act of 1975 prohibits *actions that are "reasonably likely, in all the circumstances, to offend, insult, humiliate or intimidate another person or group of people" when the act is "done because of the race, [color] or national or ethnic origin of the other person or of some or all of the people in the group.*"[182] As with Canada, the statements must be made in public in

170 Waldron, *supra* note 14 at 8 ("By 'hate speech regulation,' I mean the regulation of the sort that can be found in Canada, Denmark, Germany, New Zealand, and the United Kingdom….")

171 Interestingly, the statute only applied if the dissemination was public. Private dissemination of hate speech was not included. *See* People v. Simcox, 40 N.E.2d 525, 526 (Ill. 1942).

172 *Beauharnais*, 343 U.S. at 251.

173 *See* Note, *Defamation*, 69 Harv. L. Rev. 875, 902–03 (1956).

174 This is an interesting switch in burdens compared to how the United States approaches burdens of truth and falsity in defamation cases, at least when the speech is on a matter of public concern. *See* Philadelphia Newspapers, Inc. v. Hepps, 475 U.S. 767, 776 (1986).

175 Canada Criminal Code § 319(3).

176 *Id.* § 319(1).

177 *Id.* § 318.

178 *Id.* § 318(4).

179 *See* Race Relations Act of 1965, ch. 73 § 6.

180 Matsuda, *supra* note 7, at 2346–47.

181 Public Order Act 1986 § 18(1).

182 Racial Discrimination Act 1975 § 18C(1).

order to be actionable.[183] Interestingly, the State of Victoria, Australia, does allow for punishment for private communication, if done to *"incite[] hatred against, serious contempt for, or revulsion or severe ridicule of" a person or class of persons*,[184] but specifically excludes artistic performances, statements made for "any genuine academic, artistic, religious, or scientific purpose," "any purpose that is in the public interest," or "publishing a fair and accurate report of any even or matter of public interest."[185]

There are also efforts to adopt international standards. Most significant is the International Covenant on Civil and Political Rights (ICCPR), which (in keeping with the theme of this essay) also has within it conflicting rules. On the one hand, Article 19 of the ICCPR makes clear that "[e]veryone shall have the right to hold opinions without interference" and that "[e]veryone shall have the right to freedom of expression." On the other, Article 20 of the ICCPR states that "[a]ny advocacy of national, racial or religious hatred that constitutes incitement to discrimination, hostility or violence shall be prohibited by law."[186] Toby Mendel at the Centre for Law and Democracy argues that the two can be reconciled against each other, but admits that courts have not done a good job doing so.[187] One notable effort to harmonize the two is "the Rabat Plan of Action," a multistakeholder processes convened by the U.N. Office of the High Commissioner of Human Rights. The Rabat Plan sought to develop a six part test for assessing when speech is severe enough to warrant punishment under Article 20, which looks to: (1) *the social and political context in which the statement is made*; (2) *the position or status of the speaker in society*; (3) *the specific intent to cause harm*; (4) *the degree to which the content of the speech was "provocative and direct," and the "nature of the arguments deployed in the speech"*; (5) *the extent and reach of the speech and the size of the audience*; and (6) *the likelihood of effectively inciting harm*.[188]

Finally, the European Union's "Framework Decision on combating certain forms and expressions of racism and xenophobia by means of criminal law"[189] has become newly relevant, as it serves as the basis of the definition for the new cooperative agreement reached between private online platforms and the European Commission to police hate speech on the websites' platforms.[190] In the European Union regulatory fabric a "framework decision" is not itself a binding statement of law, but is a strong indication of how relevant EU and EU nation laws should be interpreted on this point, and now may well become the basis of hate speech regulation for a large section of the popular Internet.[191] The framework defines hate speech as one of three things: (1) *"Public incitement to violence or hatred directed against a group of persons or a member of such group defined on the basis of race, [color], descent, religion or belief, or national or ethnic origin;"* (2) *the same, when done through "public dissemination or distribution of tracts, pictures, or other material;"* and (3) *"publicly condoning, denying or grossly trivializing crimes of genocide, crimes against humanity, and war crimes [as defined in EU law], when the conduct is carried out in a manner likely to incite violence or hatred against such group or a member of such group."*[192] This "genocide denial" approach to hate speech regulation is reflected in the laws of several countries.[193]

Attempts by Online Platforms

While governments have a large role to play in defining hate speech, especially outside of the United States, perhaps the most active space in adjudicating definitions of hate speech comes from no government at all. Private online platforms, like Facebook, Twitter, YouTube, and others, routinely draft definitions of hate speech

183 *Id.* § 18C(2).

184 Racial and Religious Tolerance Act 2001 § 8(1) (Victoria, Aus.).

185 *Id.* § 11 (Victoria, Aus.).

186 International Covenant on Civil and Political Rights, *available at* http://www.ohchr.org/en/professionalinterest/pages/ccpr.aspx.

187 Toby Mendel, *Does International Law Provide for Consistent Rules on Hate Speech?*, in The Content and Context of Hate Speech, *supra* note 16, at 417.

188 Rabat Plan of Action ¶ 22.

189 Framework Decision 2008/913/JHA (Nov. 28, 2008), *available at* http://eur-lex.europa.eu/legal-content/EN/TXT/?uri=uriserv%3Al33178 (hereinafter "EU Framework Decision on Racism and Xenophobia").

190 *See infra* note 208 and accompanying text.

191 *See* EU Framework Decision on Racism and Xenophobia, *supra* note 189.

192 See *id.*

193 *See, e.g.*, Parekh, *supra* note 116, at 41.

for use in moderating their online platforms, and these definitions can also provide insight for scholars.

The objectives of private online platforms, and how they regulate, are worth exploring before turning to the definitions themselves.[194] It is now unremarkable to observe that the regulation of online platforms has some key structural differences than the regulation of governments.[195] In the United States, platforms have profound liberty to set rules around speech as they choose. They are not bound to protect the speech of their users under First Amendment, as they are not state actors, and indeed are immunized from a wide array of liability for speech generated by their uses under specific federal laws.[196] What seems to motivate inclusion of hate speech regulation is a desire not to engender controversy on their platforms, be that for liability reasons or because litigating to enforce their legal safe harbors is often expensive.[197]

It is also important to consider the platforms' greater options and subtler motivations in articulating rules for hate speech. Technology affords many possible responses to hate speech, which can extend to deleting content, modifying content, blocking users, making content invisible to some but not all users, and even more creative experimentation such as temporary bans or internal quasi-judicial resolution among users.[198] Most, if not all, platforms always reserve the right to delete or modify content or remove users at a platform's sole discretion.[199] It is therefore unnecessary — and perhaps unwise — for a platform to neatly delineate the precise scope of unacceptable behavior on their platform. Platforms want the flexibility to respond at will.

This approach, however, puts the regulation of speech by online platforms precisely at odds with the thoughtful development of structural and procedural safeguards for speech discussed above, which help to ensure against arbitrary enforcement, or overreacting to speech that touches on the pathological hysteria of the era.[200] Scholars have noted that unclear speech rules on online platforms can lead to many bad results.[201] And because the decision-making around these questions can happen behind closed doors — in fact, often through use of large teams of underpaid laborers in other countries — there are really two sets of relevant standards one should study: the public declaration of the rule, which has a vague or ceremonial role, and the actual operational document hidden from public view.[202] The impression of a

194 Though as others have noted, the relationship between platforms and governments is nuanced, fluid, and still provides a good deal of regulation capabilities for state actors. Jack Goldsmith & Tim Wu, Who Controls the Internet? Illusions of a Borderless World (2006).

195 *See generally* Lawrence Lessig, Code 2.0 (2006).

196 *See generally* Adam Holland et al., *Intermediary Liability in the United States*, Global Network of Internet & Society Research Centers (Feb. 18, 2015), https://publixphere.net/i/noc/page/OI_Case_Study_Intermediary_Liability_in_the_United_States.

197 See *id.*; *see also* David S. Ardia, *Free Speech Savior or Shield for Scoundrels: An Empirical Study of Intermediary Immunity Under Section 230 of the Communications Decency Act*, 43 Loyola L.A. L. Rev. 373 (2010).

198 See Justin Reich, *Civic Education and Video Games: The League of Legends Tribunal* (May 23, 2013), http://www.edtechresearcher.com/2013/05/civic_education_and_video_games_the_league_of_legends_tribunal/ (describing how Riot Games formerly adopted a mixture of technological and social responses, including a user-staffed tribunal, on their platform to cut bullying and offensive behavior).

199 *See, e.g.*, *Terms of Use*, Instagram, https://help.instagram.com/478745558852511 (last updated Jan. 19, 2013), ("We may, but have no obligation to, remove, edit, block, and/or monitor Content or accounts containing Content that we determine in our sole discretion violates these Terms of Use."); Statement of Rights and Responsibilities, Facebook (last updated Jan. 30, 2015), https://www.facebook.com/terms (hereinafter "Facebook Statement of Rights") ("We can remove any content or information you post on Facebook if we believe that it violates this Statement of our policies.").

200 *See supra* notes 104–13 and accompanying text.

201 *See, e.g.*, Erica Nowland et al., Account Deactivation and Content Removal: Guiding Principles and Practices for Companies and Users (Sept. 2011), https://cdt.org/files/pdfs/Report_on_Account_Deactivation_and_ Content_Removal.pdf.; Jillian York, *Guns and Breasts: Cultural Imperialism and the Regulation of Speech on Corporate Platforms* (March 17, 2016), http://jilliancyork.com/2016/03/17/guns-and-breasts-cultural-imperialism-and-the-regulation-of-speech-on-corporate-platforms/.

202 Jeffrey Rosen, *The Delete Squad*, The New Re-

user is that the adjudication is seamless and mechanized, but in fact there are many humans on the other side of the machine.[203]

The platforms are also subject to informal pressure from governments, who would no doubt love to have this freedom and flexibility in their own regulation, however contrary it would be to the public's interest. In the United States, the pressures placed by the Executive Branch around the Wikileaks "Cablegate" memos[204] and the platforms hosting the *Innocence of Muslims* video[205] have required courts to develop new theories of First Amendment doctrine to address the soft power coercion of governments on platforms and intermediaries.[206] In Europe, the coerced cooperation manifests in the European "code of conduct," to address hate speech, which obligates companies to prohibit hate speech on their platform, respond quickly to reports of hate speech, provide regular updates to the member countries about enforcement statistics, and to promote counterspeech on the platforms targeted against hate speech.[207] In the European example at least, some attempt is made to define when this soft pressure should trigger.[208] The United States pressure is based solely on the subjective interest of the relevant actors in the Executive Branch.

Turning to specific platforms, the social video website YouTube has both a "Terms of Service" and a set of "Community Guidelines," the latter of which is incorporated by reference into the formal agreement between YouTube and its users. All discussion of what constitutes hate speech is confined to the Community Guidelines, while the Terms of Service mention "offensive content" when it makes clear that the platform is not liable for the offensive content of its users[209] (which is both in the contract and generally true under United States law[210]). The Community Guidelines begin by first insisting that a user "respect the YouTube community," explaining "[w]e're not asking for the kind of respect reserved for nuns, the elderly, and brain surgeons. Just don't abuse the site."[211] This is later explained in detail in a later section specifically addressing "hateful content,"[212] which immediately opens with a statement articulating the tensions between free speech and hate speech regulation: "Our products are platforms for free expression *But we don't support content that promotes or condones violence against individuals or groups based on race or ethnic origin, religion, disability, gender, age, nationality, veteran status, or sexual orientation/gender identity, or whose primary purpose is inciting hatred on the basis of these core characteristics."*[213] The list of protected characteristics has many elements similar to the laws described above.[214] The nexus to violence is in the definition, but it extends to

PUBLIC (April 29, 2013), https://newrepublic.com/article/113045/free-speech-internet-silicon-valley-making-rules; Adrian Chen, *The Laborers Who Keep Dick Pics and Beheadings Out of Your Facebook Feed*, WIRED (Oct. 23, 2014), http://www.wired.com/2014/10/content-moderation/.

203 Mary Gray, *The Paradox of Automation's "Last Mile,"* SOCIAL MEDIA COLLECTIVE RESEARCH BLOG (Nov. 12, 2015), https://socialmediacollective.org/2015/11/12/the-paradox-of-automations-last-mile/.

204 See Michael Lambert, *The State as Soft Power – The Intermediaries Around Wikileaks* (Feb. 18, 2015), https://mlonml.com/2015/02/18/the-state-as-soft-power-the-intermediaries-around-wikileaks/.

205 Michelle Quinn, *Google Decides to Leave Video on YouTube*, POLITICO (Sept. 14, 2012), http://www.politico.com/story/2012/09/google-decides-to-leave-video-on-youtube-081245.

206 Backpage.com, LLC v. Dart, 807 F.3d 229 (7th Cir. 2015).

207 *Code of Conduct on Countering Illegal Hate Speech Online*, EUROPEAN COMMISSION (May 31, 2016), http://ec.europa.eu/justice/fundamental-rights/files/hate_speech_code_of_conduct_en.pdf.

208 The Code of Conduct adopts the EU Framework Decision identified above. *See supra* notes 189–93 and accompanying text.

209 *See Terms of Service*, YouTube, https://www.youtube.com/ static?template=terms § 5(D) (last updated June 9, 2010) (the user must "acknowledge that [they] will be exposed to [c]ontent that is inaccurate, offensive, indecent, or objectionable" agree to waive any claims against YouTube for the content); *id.* § 10 (the user must "specifically acknowledge that YouTube shall not be liable for content or the defamatory, offensive, or illegal conduct of any third party").

210 *See generally* Holland, *supra* note 196.

211 *Community Guidelines*, YouTube, https://www.youtube.com/yt/ policyandsafety/communityguidelines.html (last viewed Nov. 27, 2016).

212 *Id.*

213 *Id.*

214 The one atypical inclusion may be discrimination based on veteran status, though this is likely a reflection of existing U.S. law on this point. *See, e.g.*, 38 U.S.C. § 4212.

either "promot[ing]" or "condon[ing]" such violence, thus extending beyond incitement as the concept is understood under First Amendment doctrine.[215] YouTube also introduces intent, but makes it an optional requirement.

Much like YouTube, the social media platform Twitter adopts a bifurcated structure to its regulation, with a Terms of Service that disclaims liability for offensive content on the platform, and refers users a set of "Twitter Rules" to discuss platform norms.[216] Within a section of those rules entitled "abusive behavior," Twitter specifically prohibits "hateful conduct," defined as "*promot[ing] violence against or directly attack[ing] or threaten[ing] other people on the basis of race, ethnicity, national origin, sexual orientation, gender, gender identity, religious affiliation, age, disability, or disease.*"[217] Twitter also makes clear that it does not allow accounts "*whose primary purpose is inciting harm towards others on the basis of these categories.*"[218]

The social media platform Facebook adopts a similar structure to its regulation of hate speech. Its "Statement of Rights and Responsibilities" mirrors a standard terms of service,[219] and also references a set of "Community Standards" that are not directly incorporated to their terms as a condition of the agreement, but are part of a suite of other documents "[y]ou may also want to review."[220] Again, the terms only disclaim liability, and require that the user not "use Facebook to do anything … discriminatory."[221] On its Community Standards, Facebook identifies hate speech subject to removal from the platform as "*content that directly attacks people based on their race; ethnicity; national origin; religious affiliation; sexual orientation; sex, gender, or gender identity; or serious disabilities or diseases.*"[222] Beyond this, Facebook expressly bans "*[o]rganizations and people dedicated to promoting hatred against these protected groups.*"[223] These rules reflect a definition that includes a protected group and a specific harm related to the speech. Intent to cause harm is not directly part of Facebook's main definition, but is part of the critical definition of the types of hate groups that may be prohibited. Facebook also expressly considers "innocent" use of some forms of hate speech — specifically, "shar[ing] content containing someone else's hate speech for the purpose of raising awareness or educating others about that hate speech."[224]

But behind the curtain, a leaked copy of Facebook's internal guidelines for its outsourced content moderation team shows a much more granular and content-oriented definition framework for hate speech.[225] Facebook's 2012 "Abuse Standards" operation manual[226] asks content moderators to flag nine different forms of "hate content," including (1) *slurs or racial comments of any kind*; (2) *attacking based on a protected category*; (3) *hate symbols, either out of context or in the context of hate phrases or support of hate groups*; (4) *showing support for organizations and people primarily known for violence*; (5) *depicting symbols primarily known for hate and violence, unless comments are clearly against them*; (6) "*versus photos*" *[…] comparing two people (or an animal and a person that resembles that animal) side by side*; and

215 *See generally Brandenburg*, 395 U.S. 444.

216 *See Terms of Service*, Twitter, https://twitter.com/tos?lang=en (last updated Jan. 27, 2016).

217 *The Twitter Rules*, Twitter, https://support.twitter.com/articles/18311 (last viewed June 2, 2016).

218 *Id.*

219 *Statement of Rights and Responsibilities*, Facebook (last updated Jan. 30, 2015), https://www.facebook.com/terms (hereinafter "Facebook Statement of Rights"). Facebook does make clear that it reserves the right to delete any content that, in its view, violates any of its policies. See *id.* § 5(2).

220 See *id.*; *Community Standards*, Facebook, https://www.facebook.com/ communitystandards (last visited May 29, 2016) (hereinafter "Facebook Community Standards").

221 Facebook Statement of Rights, *supra* note 219, at § 15(2).

222 Facebook Community Standards, *supra* note 220.

223 *Id.*

224 *Id.*

225 *See* Adrian Chen, *Inside Facebook's Outsourced Anti-Porn and Gore Brigade, Where "Camel Toes" are More Offensive than "Crushed Heads,"* Gawker (Feb. 16, 2012), http://gawker.com/5885714/inside-facebooks-outsourced-anti-porn-and-gore-brigade-where-camel-toes-are-more-offensive-than-crushed-heads.

226 See oDesk, Abuse Standards 6.2 (2012), *available at* https://www.scribd.com/doc/81877124/Abuse-Standards-6-2-Operation-Manual. Facebook no doubt has updated its procedures since the time of this publication.

(7) *photo-shopped images showing the subject in a negative light*. In one of the few attempts to engage in some form of definitional balancing here, moderators are instructed that "humor overrules hate speech UNLESS slur words are present or the humor is not evident."[227]

Later slides articulate the categories along which hate-based content can be subject for action, including race, ethnicity, national origin, religion, sex, gender identity, sexual orientation, disability, or any serious disease. There appears also to be toleration built into Facebook's enforcement model for political speech: content that includes hate symbols should not be escalated when it concerns a public figure or head of state, but should be escalated when it includes an ordinary person or law enforcement officer.[228]

Facebook's enforcement suite perhaps is the best example out of all the definitions as to how definitions of hate speech should be balanced with other speech considerations. Carve-outs for humor, discussion of public figures, and speech that opposes hatred, with specific caveats for when the speech uses slurs, shows at least some attempt to accommodate speech values in the definition of what should be actionable, even if such accommodation would be unconstitutional if done by a state actor in the United States. This particular definition suite is also notable for the level of proxies to other identifiers of hate speech, including symbols,[229] organizations and people clearly oriented toward "violence," and overt forms of dehumanizing content.

Emerging Themes and Continuing Questions

Common Traits in Defining Hate Speech

Much as I would like to, I offer no single definition to govern all of the above-cited examples. From these myriad definitions, however, I believe some common themes and insights can be drawn. The following part reviews some of the congenerous concepts that appear throughout the definitions above.

A question naturally arises as to what to do with these character traits. The traits identified below do not form a single definition, but could be used as a scoring system to help improve a researcher's confidence that the speech in question is the type of speech that is likely to cause the harms that hate speech causes, is unlikely to have the redeeming cultural values of consequentialist justifications of freedom of expression, and thus is worthy of identification as "hate speech." Confidence scoring is a technique used in other areas of language study, and is a useful analytic tool, in that it embraces the fact that speech is inherently a complicated, multitudinous, and highly contextual human behavior, and that all categorization of speech can only be done by matters of degrees.[230] More qualitatively speaking, looking solely at one or two of these factors will likely result in a vastly overbroad corpus, with a lot of speech included that could not be fairly called "hate speech." Looking at several factors will probably drive a researcher closer to what most consider "hate speech," but may still identify speech for which there is good reason to keep it outside the definition. Speech that hits all of these criteria is likely to be speech that most countries and online platforms would define as "hate speech," and may even be actionable speech in the United States, as a true threat[231] or incitement to imminent lawless action.[232]

Finally, as I believe the criteria below make clear, the specific environment and context around any incident of hateful speech can be quite relevant when considering how society should

227 See *id.*

228 See *id.* at 4. Interestingly, obesity is expressly carved out.

229 Also identified by Marwick & Miller, *supra* note 163.

230 *See, e.g.*, Christine Pao et al., *Confidence Scoring for Speech Understanding Systems* (1998), *available at* https://groups.csail.mit.edu/sls/publications/1998/icslp98-confidence.pdf.

231 *See* United States v. Elonis, 135 S. Ct. 2001 (2015).

232 *Brandenburg*, 395 U.S. 444. It may be that only looking for speech that qualifies under every category seeks to prove too much. Some of these elements, like the requirement that the speech cause a second-level harm, could be too limiting. A consideration of all but one or two factors, with a reason as to why some were omitted, could still produce a valid working framework.

respond. I do not believe that one should take these criteria and develop a methodological approach that claims to capture and analyze all hate speech online, or even all hate speech on a specific platform or about a certain group. These traits may help surface specific case studies, but those studies should examine their unique context.

1 - Targeting of a Group, or Individual as a Member of a Group

This factor may be the only threshold factor, as it is the one that separates "hate speech" from any other form of harmful speech, such as bullying or threats. To meet the definition of hate speech, the speech should target a group or an individual as they relate to a group.

Which "groups" count has some variance. Some definitions use descriptors, such as "historically oppressed,"[233] "traditionally disadvantaged,"[234] or "minority."[235] Others prefer to list actionable groups.[236] Parekh's definition is notable in that it does not look for a defined group, but looks to see whether the speaker targets someone based on "an arbitrary or normatively irrelevant feature."[237]

By any of these frameworks, which groups make the list is a fascinating reflection of the values of the particular organization or context. Race, ethnicity, and religion appear most frequently; gender, sexual orientation, and gender identity appear somewhat frequently; veteran status, physical ability, and suffering from serious diseases only occasionally appear to make the list. While legal definitions frequently use such definitional lists, they are not without nuance.

Finally, any attempt to classify individuals will be an attempt to objectify what can be often complicated and at-times-subjective questions about identity.[238] Those who seek to study hate speech online must determine for themselves how they will handle the subjective–objective line in whichever groups they seek to study.

2 - Content in the Message that Expresses Hatred

As a basis for legal punishment, content-based definitions have a long climb under American law. The Supreme Court has made the "categorical approach," and its hostility toward content-based restriction of speech, a key ingredient to First Amendment doctrine.[239] Only a small list of types of speech are proscribable absent an extraordinary state interest, and the Supreme Court has made clear that they are not inclined to expand that list.[240] If the goal is to create a workable definition under law, as opposed to merely for study, most will be well-advised to avoid definitions that speak to message or content.

Ward expressly rejects a content limitation; "any form of expression" can be hate speech, if made with the intent to incite hatred.[241] Moran similarly looks solely toward speech intended to "promote hatred," however formulated.[242] Benesch looks more toward the likely effect that the speech has on the relevant audience."[243] Delgado's definition seems to avoid a content definition, though he does look to speech that makes a "reference to race,"[244] and suggests that one should be able to objectively identify the speech as an insult, though his definition does not specify whether that identification should be done based on content, context, or some combination of the two.[245] (Though Delgado's and Moran's definitions are more general, I believe they would still be considered content-based definitions under American law.[246])

233 Matsuda, *supra* note 7, at 2357–58.
234 Moran, *supra* note 98, at 1430.
235 Marwick & Miller, *supra* note 163, at 17.
236 *See, e.g.*, Canada Criminal Code § 318(4).
237 Parekh, *supra* note 116, at 40–41.
238 *See generally* Martha Minow, *The Supreme Court 1986 Term – Foreword: Justice Engendered*, 101 Harv. L. Rev. 10 (1987).
239 *See supra* notes 111–13 and accompanying text.
240 United States v. Stevens, 559 U.S. 460, 469 (2010).
241 Ward, *supra* note 151, at 763.
242 *See* Moran, *supra* note 98, at 1430.
243 Benesch, *supra* note 156.
244 *See* Delgado, *supra* note 6, at 179.
245 *Id.*
246 The definitions still require a judge or law enforcement to examine the content of the message to see if it meets the criteria, and are drafted to respond to the communicative impact of the speech in question. *See* Reed v. Town of Gilbert, 135 S. Ct. 2218, 2228 (2015); FCC v. League of Women Voters of Cal., 468 U.S. 364, 383 (1984).

Some scholars do propose a content-based definition. Matsuda looks for speech that promotes "racial inferiority," or "denies the personhood of target group members," and treats all members of the targeted group as "alike and inferior."[247] Parekh looks to speech that "stigmatizes the target group by … ascribing to it qualities widely regarded as undesirable."[248] This approach is also popular in foreign law; the United Kingdom looks to speech that is "threatening, abusive, or insulting."[249] Australia looks to actions (one assumes chiefly speech acts) that are likely to "offend, insult, humiliate, or intimidate others."[250] The Rabat Plan looks to see if the speech is "provocative and direct," and asks EU member states to look to the "nature of the arguments employed."[251] This could be either a content characterization or an analysis of the speech's tone. The statute at issue in *Beauharnais* sought to punish speech that "portrays depravity, criminality, unchastity or lack of virtue" in a group.[252] These can each give a framework for study, but to look solely at what is said without its context could lead many online scholars to make the same errors discussed above concerning coded speech, re-appropriation of slurs, and other ways seeming hate speech can be serving a very different purpose.[253] In the online space especially, context and "context collapse" as speech moves outside its intended audience are very important considerations.[254]

As an interesting alternative approach, Marwick & Miller look specifically to use of symbolism in hate speech, a place where the Supreme Court has provided an interesting distinction between what is and is not proscribable.[255] Facebook's internal content guidelines also put a strong emphasis on symbols, or references to key figures famous for hatred.[256] Symbols in this context are in many senses similar to an epithet, where one can assume a degree of intent in many situations, with some key exceptions.[257] Of course, the ability to scour the Internet for instances of a visual symbol is a notoriously difficult computer science problem.[258]

3 - The Speech Causes a Harm

This essay has largely taken harm as a given with hate speech. As a look to the harms of speech often permeate definitions, however, I will turn to these harms now.

Some definitions look specifically to extrinsic harms beyond the speech itself, most often physical violence. The European Union framework decision, Twitter's Terms of Service, and Benesch's study of dangerous speech all look to speech that causes a physical harm.[259] First Amendment scholarship looks to violence specifically in its incitement and true threats frameworks, and often grapples with violent responses to otherwise-lawful speech, sometimes referred to as the "heckler's veto" problem.[260]

Others, like Delgado, look instead to the myriad ways in which speech can cause harm, including deep structural considerations.[261] Delgado notes that targeting hatred due to immutable characteristics causes a more salient harm in part because the victim will also despair that they cannot change the attribute that gives rise to the hatred.[262] They can permeate and impact the victim's relationship with others, especially across racial, religious, gender, or ethnic divides.[263] Performance in work, relationships, and social and personal life will no doubt suffer due to the pervasive and withering damage that consistent denial of one's self-worth causes.[264]

247 *See* Matsuda, *supra* note 7, at 2358.

248 Parekh, *supra* note 116, at 40–41.

249 Public Order Act 1986 § 18(1).

250 Racial Discrimination Act 1975 § 18C(1).

251 Rabat Plan of Action ¶ 22.

252 *Beauharnais*, 343 U.S. at 251.

253 *See supra* notes 116–28 and accompanying text.

254 *See* Alice Marwick & danah boyd, *I Tweet Honestly, I Tweet Passionately: Twitter Users, Context Collapse, and the Imagined Audience*, 13 Data & Society 114 (2010).

255 *Compare* R.A.V., 505 U.S. 377 to Virginia v. Black, 538 U.S. 343 (2003).

256 *See supra* notes 225–29 and accompanying text.

257 *See generally* Delgado, *supra* note 6.

258 *See Tasks*, XKCD, http://xkcd.com/1425/ (last visited Nov. 27, 2016).

259 *See supra* notes 156, 191, and 217, and accompanying text.

260 *See, e.g.*, Terminiello v. City of Chicago, 337 U.S. 1 (1949); Feiner v. New York, 340 U.S. 315 (1951).

261 Matsuda, *supra* note 7, at 2338; see Post, *Racist Speech*, *supra* note 40, at 271–77.

262 Delgado, *supra* note 6, at 136–37.

263 *Id.* at 137.

264 Id. at 139–40.

Growing scholarship and attention is placed on the structural harms of such hateful treatment, and how feelings of diminished expectation or self-worth can pass through generations from parents to children.[265] Indeed, the harms of racist speech, sexist speech, anti-Semitic speech, and other forms of harmful speech may be difficult to analyze as a universal matter, given the specific contexts of each.[266] I doubt that any sincere scholar would argue that these harms do not exist, though they may well argue that some or all of these harms are necessary in order to protect the free speech principles at stake.

Massey looks almost exclusively to harm as the basis of his assessment of hate speech.[267] This is effective for Massey's purposes of discussion, but it does not elucidate clear standards for identification without a great deal of consideration and context, and is very ill-suited as a definition to be used for legal sanction, as it has no consideration of culpability. Most instead consider harms along with other factors. For example, the statute in *Beauharnais* punished those who, among other things, published speech that exposed a member of a group to "contempt, derision, or obloquy."[268]

The more a definition relies solely on harms, the more it risks sweeping in cases that would likely give pause to academics, lawyers, and online platform moderators alike. Many forms of highly important speech for self-governance can cause discomfort at best and harms at worst. A politician expressing indignation at governing principles like diversity or equality is important to hear and to know, and yet can certainly cause harm to marginalized groups.[269] Many of the current debates on college campuses about diversity, tolerance, and how universities should respond to confrontational topics and ideas are highly important discussions that may cause harm to the speakers and listeners involved. Before one would consider such discussion "hate speech," I would hope that one would look to other aggravating factors.

265 *Id.* at 147.
266 Matsuda, *supra* note 7, at 2331.
267 Massey, *supra* note 129, at 105 n.2.
268 Beauharnais, 343 U.S. at 251.
269 Matsuda, *supra* note 7, at 2334.

4 - The Speaker Intends Harm or Bad Activity

Many definitions look to whether a speaker of hate speech intends some harm or other bad activity, but what exactly the speaker should intend to constitute "hate speech" is subject to dispute.

Some definitions use a non-physical framework. Under Delgado the speaker should have an intent to demean,[270] with Ward this was an intent to "vilify, humiliate, or incite hatred,"[271] with Moran this is an intent to "promote hatred."[272] Matsuda does not include intent as part of her definition, but explains later that speech should be actionable if it "is, and is intended as, persecutorial, hateful, and degrading."[273] The statute in *Beauharnais*, coming from defamation law, likely had a tacit intent component.[274] Canada's law, similarly, looks to speech that willfully "promotes hatred."[275] The Rabat Plan of Action identifies an intent to cause harm as a key trait for Article 20 of the ICCPR.[276]

With others the intended harm needs to be more physical. Twitter targets conduct that "promotes violence" or "directly attacks" a group, suggesting an underlying intent.[277] Marwick and Miller make look to an intent to promote "hatred, violence or resentment."[278] Others have intent as an optional component. YouTube makes intent a component of their assessment, but not a mandatory one.[279] The United Kingdom does as well — the speaker can either intend to stir up racial hatred, or the circumstances can be such that it is likely to be stirred up.[280] Facebook looks to intent when determining whether to block groups dedicated to hate speech, but intent is not part of their general definition.

Intent seems to be the element that many point

270 Delgado, *supra* note 6, at 179.
271 Ward, *supra* note 151, at 765.
272 Moran, *supra* note 98, at 1430.
273 Matsuda, *supra* note 7, at 2358.
274 *See supra* notes 21–30 and accompanying text.
275 Canada Criminal Code § 319(3).
276 Rabat Plan of Action ¶ 22.
277 *See* The Twitter Rules, *supra* note 217,
278 Marwick & Miller, *supra* note 163, at 17.
279 *See supra* notes 209–15 and accompanying text.
280 Public Order Act 1986 § 18(1) (U.K.).

to when trying to work a definition that avoids the harder questions under speech theory. How can one truly say they are trying to move the needle in the marketplace of ideas, or submit genuine proposals for self-governance, if they intend only to hurt someone or some group?[281] An intent to harm does not answer the free speech objections raised by Bollinger,[282] but intent is a piece of many existing definitions of actionable speech in the United States,[283] so one assumes that it can help in this environment as well. Nevertheless, discerning intent will be extremely difficult to do well when studying online speech. The speaker's intent may be obscured, denied, or simply not disclosed. Speakers may lie about their intent to others; they may even by lying about their intent to themselves. That this factor can be so crucial to balancing speech principles and yet so hard to identify may be one of the more frustrating aspects of studying hate speech online.

5 - The Speech Incites Bad Actions Beyond the Speech Itself

Perhaps also a reflection of a tendency towards avoiding the free speech conflicts with identification of "hate speech," many definitions require that the speaker incite some other consequence as a result of the speech. Here again, however, what specifically one should be inciting is subject to debate.[284]

Some definitions look for incitement of other non-physical reactions. Several definitions look to incitement of "hatred,"[285] which could be seen as "harm" in element 3, or could be seen as an independent bad reaction. Other definitions look to violence. Benesch focuses on murder, ethnic cleansing, and other atrocities, and looks to speech that incites that specifically.[286] The statute at issue in *Beauharnais* and the law in Canada both look to speech that incited a "breach of the peace or riots."[287] Some, like the EU Framework Decision on racism and xenophobia, look to both.[288] And some, like the Rabat Plan of Action, contemplate simply inciting further harm as a relevant factor, suggesting by implication that the harm of speaking is not enough.[289]

"Incitement" is a term of art under American speech law, drawn from the famous hate speech case *Brandenburg v. Ohio*, and tends to only apply when the incitement is imminent, or almost inevitable.[290] It is not at all clear that the definitions here seek to put such a strong qualifier on the term. Courts looking at online speech have already struggled to decide how concepts like imminence in incitement doctrine should be understood with regard to online speech.[291]

Online researchers should keep in mind that it is not always apparent what speech does or does not incite, especially when considering concepts like inciting hatred, which may be undetectable or preconditioned in the audience. Even when the definition looks to violence, the *Innocence of Muslims* example teaches, even when one can establish a nexus from speech to violence, there may be many intervening steps and intervening speakers along the way.[292] When examining a particular incident, like the riots in response to the *Innocence of Muslims* video, careful forensic work should be done to track how the speech is altered, recontextualized, and appropriated as it goes from the initial speaker to the intended audience.[293]

281 *See, e.g.*, Citron, *supra* note 9.

282 *See supra* notes 89–92 and accompanying text.

283 *See Brandenburg*, 395 U.S. at 447–48 (incitement must be "directed to" producing imminent lawless action); *Virginia v. Black*, 538 U.S. 343; *Elonis*, 135 S. Ct. at 2013 (2015) (as a statutory matter, threats are only actionable under federal law if the defendant intends to threaten).

284 *See* Herz & Molnar, *supra* note 16, at 1, 4.

285 *See, e.g.*, Ward, *supra* note 151, at 765; YouTube Community Guidelines, *supra* note 211.

286 Benesch, *supra* note 156.

287 *Beauharnais*, 343 U.S. at 251; Canada Criminal Code § 319(1). It should be noted that Supreme Court decisions subsequent to *Beauharnais* also severely limited punishment of speech on such vague grounds as speech that may cause a "breach of the peace." Cox v. Louisiana, 379 U.S. 536, 551–52 (1965).

288 See EU Framework Decision on Racism and Xenophobia, *supra* note 189.

289 Rabat Plan of Action ¶ 22.

290 *Brandenburg*, 395 U.S. 447–48.

291 *See* Planned Parenthood v. Am. Coalition of Life Activists, 290 F.3d 1058 (9th Cir. 2002).

292 *See supra* notes 123–25 and accompanying text.

293 *See also* Susan Benesch, *Charlie the Freethinker: Religion, Blasphemy, and Decent Controversy*, 10 Religion & Hum. Rights 244, 252 (2015) (noting that in two famous cases where religiously provocative speech lead to physical violence, "ill-intentioned

6 - The Speech is Either Public or Directed at a Member of the Group

Several definitions look to audience, though the audience in question often depends whether the defining party seeks to examine public spread of hatred, or the effect of an insult hurled at an individual. Some, like Benesch, look for public declarations of hatred and the social harms that flow therefrom,[294] and some, like Delgado's, target the personal attack.[295] Canada's and Australia's hate speech laws make clear that only "public" statements are subject to punishment, though it would seem that public statements to a single individual of the protected group would be encompassed under these definitions.[296] Marwick and Miller also seem to cover both spaces, by requiring that the subject "experience" the harm of the speaker's message, which could be either direct or indirect.[297] On the other hand, Matsuda evades this somewhat; she requires that the message be "directed against a historically oppressed group," but it is not clear if that is a reference toward its content or its audience.[298]

Parekh contemplates a public audience for his definition for a hate speech crime, and specifically that the speech cause the public to view the target group "as an undesirable presence and a legitimate object of hostility."[299] Audience reactions such as these may be easier to observe and detect online, but it is a bit of a puzzle as to why a criminal defendant under Parekh's law would be not guilty if he or she delivered remarks to an unmoved audience.

The missing group from this trait is speech that is neither public nor targeted at the member of the group. The absence of this type of speech from the definitions above may be a reflection of where many see the limits of governance in a liberal democracy. With few exceptions, regulating speech "at the home" is a step too far for many regulators of speech, as it feels more like regulating freedom of thought itself.[300]

But for those that are hoping to study the larger effects of hate speech beyond what can be regulated, it may be wise to leave in private speech that is directed to those outside of the targeted group. First, if one is looking for where to intervene in hate speech to mitigate its lasting effect, the speech that happens privately may be the best target. Ingrained racism spread within the privacy of one's home or social circle may be the most calcified and hardest sentiment to dislodge. Second, the concepts of "public" vs. "private" speech, if they were clear before the Internet, are even more confused today. Some "private" conversations happen on public platforms such as Twitter, where other users can not only see what was said, but amplify those conversations to others, despite not being a party to what was initially discussed, just as a researcher may be accidentally lead through "context collapse" to see a racial slur where neither the sender nor recipient perceived it as such, a speaker may find that what was meant to be only a private conversation can be brought public through the architecture of the speaker's online platform.[301]

7- The Context Makes Violent Response Possible

Of the various academic definitions offered, Benesch's is the one that most directly addresses the question of context at length. Benesch looks to factors such as the power of the speaker, the receptiveness of the audience, and the history of violence in the area where the speech takes place.[302] If the focus is on speech that is actually likely to catalyze physical harm, this contextual analysis seems critical.

Others reference this indirectly. Delgado's use of a "reasonable person" standard for identifying the remark as a racial insult is a reference to a standard concept in tort and criminal law,

figures relentlessly publicized the mocking content, in ways that they knew were likely to [catalyze] violence, and among audiences they knew were likely to react with violence").

294 *See* Benesch, *supra* note 156.

295 *See* Delgado, *supra* note 6, at 179.

296 Canada Criminal Code § 319; Racial Discrimination Act 1975 § 18C(2).

297 Marwick & Miller, *supra* note 163, at 17.

298 Matsuda, *supra* note 7, at 2358.

299 Parekh, *supra* note 116, at 40–41.

300 *See, e.g.*, Stanley v. Georgia, 394 U.S. 557 (1969) (obscenity possession in the home is not punishable); *but see* Osborne v. Ohio, 495 U.S. 103 (1990) (refusing to extend *Stanley* to child pornography).

301 *See* Marwick & boyd, *supra* note 255.

302 Benesch, *supra* note 196.

where the liability or guilt of a person is made in reference to what a reasonable person would do or understand in those circumstances. In the analogous space of workplace harassment, the precise role of social context for a "reasonable person" test is still debated.[303] The United Kingdom looks to the "circumstances" around speech to see if "racial hatred is likely to be stirred up thereby."[304] The Rabat Plan of Action advises that EU member states look to "the social and political context," the status of the speaker, and the size of the audience.[305]

If a scholar seeks to look at a legal attempt to codify context in law, they may be well-served by looking to the example set by the United States approach to obscenity. American obscenity law famously has a contextual element to it, and how it has been applied in online environments presents an interesting analogy for those who seek to study or regulate hate speech online. Starting with the Supreme Court's decision in *Roth v. California*,[306] and cemented with the Court's *Miller v. California*, the test for determining whether a given work is obscene requires proof that, among other things, a person applying "contemporary community standards" would view the work as prurient.[307] The court in *Miller* directly avoided a nationwide standard because, as they put it "our Nation is simply too big and too diverse for this Court to reasonably expect that such standards could be articulated for all 50 States in a single formulation."[308] With the advent of the Internet, the Supreme Court was again asked to adopt a nationwide standard, because, as advocates put it at the time, "[o]nce a provider posts its content on the Internet, it cannot prevent that content from entering any community."[309] The Court again refused, as they felt the harm to speech too great if "any communication available to a nation wide audience will be judged by the standards of the community most likely to be offended by the message."[310] The same speech principles that limited obscenity to the places where it is actually viewed as obscene can inform some of the speech-oriented limitations to hate speech observation. As noted a few times above, it is important to bear in mind the limits of what one can see without the benefit of local context, though it is also important to note how speech can leave its original context and transform into something else that may in fact be worthy of study, especially on a global communications platform.

8 - The Speech Has No Redeeming Purpose

It would seem as though every definition is aware of its own limitations, and several choose to address this by trying to excise out the "good" speech that may have fallen into the original definition. Marwick and Miller state their general definition of hate speech as "speech that *carries no meaning other than* hatred towards a particular minority"[311] A version of this can also be seen in Ward's definition.[312] This lack of "good" purposes is a repeat theme in speech laws, including with obscenity[313] and many definitions of actionable harassment.[314] The difficulties in using a "no legitimate purpose" test for speech in the abstract, however, are well documented, as they involve deep subjective assessments on the part of the adjudicator, and the complex nature of most speech would require adjudicators to either overlook this nuance punish speech with some redeeming purpose, or embrace the complexity and thus exclude nearly all speech from the definition.[315] Twitter tries to engage with

303 See Melissa K. Hughes, *Through the Looking Glass: Racial Jokes, Social Context, and the Reasonable Person in Hostile Work Environment Analysis*, 76 S. Cal. L. Rev. 1437 (2003).

304 Public Order Act 1986 § 18(1) (U.K.).

305 Rabat Plan of Action ¶ 22.

306 354 U.S. 476 (1957).

307 413 U.S. 15 (1973).

308 *Id.* at 30.

309 *Reno*, 521 U.S. at 853 (quoting ACLU v. Reno, 929 F. Supp. 824, 844 (E.D. Pa. 1996)).

310 *Id.* at 877–78.

311 Marwick & Miller, *supra* note 163, at 16; *see also id.* at 17 (definitions should look to whether a speaker intends "*only* to promote hatred, violence, or resentment" (emphasis added).

312 *See* Ward, *supra* note 151, at 766 (noting that speech qualifying as hate speech should be "so virulent" that an observer would have difficulty separating the attack from the message).

313 *Miller*, 413 U.S. 15, 24 (1973).

314 *See, e.g.*, Del. Code Ann. tit. 11, § 1311(a)(1); Fla. Stat. § 784.048(1)(d).

315 *See* Eugene Volokh, *One-To-One Speech vs. One-To-Many Speech Criminal Harassment Laws, and "Cyberstalking,"* 107 Northwestern L. Rev. 731, 776–80 (2013).

this complexity of language by targeting for removal groups whose "primary purpose" is inciting harm, though this determination — like most in online platform regulation — is a subjective one made by Twitter alone.[316]

Canada tries to solve this puzzle by specifically exempting certain types of speech from its overall definition, including speech expressing "good faith" opinions on a religious subject, speech that is true, or made in the public interest.[317] The State of Victoria, Australia does a similar balancing act.[318] This compromise may split the baby. Proponents of hate speech regulation may be legitimately concerned that hate speech will simply disguise itself as scientific or religious analysis and thus avoid regulation. Opponents may worry that qualifications like "good faith" will wind up playing the same role as "good motives" did in *Beauharnais* — a judge will make a subjective judgment of goodness, and society will be left without a principled standard. Triaging through observation of speech online to identify "good faith" offerings will also have many of the same concerns as attempts to discern intent above.

Continuing Questions

Any assessment or classification inherently has within it value choices, prioritizations, and omitted elements.[319] At the end of her landmark article on defining hate speech, Matsuda noted several types of more nuanced or troubling speech that evaded her definition,[320] and I see similar limitations with my own framework.

One of the hard remaining questions from the definition framework above, or at least how such framework can be put to practice, is how to deal with the relative power of different groups in different online spaces. The definitions above largely don't account for the natural feeling of most scholars that, for example, speech that promotes the power of women online should not be thought of as a form of hate speech, but the same empowerment message directed toward men, at least as it is seen today in "men's rights activists" platforms, feels different in kind.[321] The Southern Poverty Law Center has gone as far as to identify certain "men's rights" platforms as hate sites.[322]

Matsuda notes the tensions presented by this contradiction, but ultimately finds that angry or hateful speech by a "subjugated group" should be seen as "a victim's struggle for self-identity in response to racism," and therefore, one presumes, not hate speech in many cases. That said, these power considerations could be fluid — Matsuda also notes that should the victim group change power and be placed "in a dominant or equalized position," the special protection for such speech against sanction should be lost.[323] That these standards can evolve will force scholars to stay on their toes as they observe communities.

I also worry that I have not fully addressed how to handle harmful *uses* of another's speech, both when done deliberately and when the disseminator subconsciously uses another's speech to substantiate a point that the original speech does not support. For an example of the latter, an Islamophobe may subconsciously gloss over dozens of news stories about hate crimes against Muslims, find the one story where a Muslim person was the aggressor, and post that story online as justification of their bigotry. I do not believe punishing that individual would change much of anything. I would hope that counterspeech and discourse can solve these sorts of problems, but I worry for the efficacy of the marketplace in such irrational spaces, for the reasons critics have noted above.[324]

And finally, I worry that in trying to deal with the hard tensions between speech theory and harmful speech, we may find ourselves repeating many of the same biases and errors that have led many to openly question the effica-

316 *See* The Twitter Rules, *supra* note 217.

317 Canada Criminal Code § 319(3).

318 *See* Racial and Religious Tolerance Act 2001 § 11.

319 Moran, *supra* note 98, at 1428.

320 *See generally* Matsuda, *supra* note 7, at 2361–70.

321 I would suspect that only select content from these platforms would be a form of "hate speech" subject to the criteria above, but I also suspect that scholars would more easily identify that speech when it is a man speaking of a woman, instead of vice-versa.

322 *Misogyny: The Sites*, S. Poverty L. Ctr. (March 1, 2012), https://www.splcenter.org/fighting-hate/intelligence-report/2012/misogyny-sites.

323 Matsuda, *supra* note 7, at 2361–62.

324 *See supra* notes 70–80 and accompanying text.

cy of free speech theory itself. Regulators are incredibly anxious to respond to hate speech, and they will look to scholarship to justify their decisions. As scholars, we owe it to those regulators to give a fair assessment. By now I hope it is clear that hate speech is highly context-sensitive, and many studies who look to how these issues play out on "the Internet," forget that it is not one grand corpus, where everyone is speaking to everyone. No one person sees the Internet; we only see what our framework and perspective leads us to see.[325] The technological affordances of the Internet may mean that content can slip between communities and spread the way no speech has ever had before, but that does not mean that each new audience will be looking at these issues in the same way.

Conclusion

In the above sections I have labored through many definitions and examined their theoretical shortcomings. All of this may feel as though I am trying to work scholars and regulators to such a confused place that they refuse to study hate speech altogether. This is not my intent at all. I only seek to illustrate that this is difficult, it is difficult for good reasons in light of the competing interests at play, and we should approach these questions in an intelligent manner. A person who says they have an easy solution to the problem of hate speech, or even how to observe and document hate speech, is simply not thinking hard enough. This issue is a critical one, and we must approach it critically. The definition of hate speech in a study or regulatory environment may be the most important part of the project's design. Looking through the eight traits identified above, and using several of them in designing an identification system will help ensure that the research targets what most would agree is the type of speech that is causing us so much concern, and the speech that deserves the most rigorous study.

325 *See, e.g.*, Ethan Zuckerman, Rewire: Digital Cosmopolitans in the Age of Connection (2013).