2011

Intermediaries and Hate Speech: Fostering Digital Citizenship for Our Information Age

Danielle Keats Citron
University of Maryland School of Law

Helen Norton
University of Colorado Law School

Citation Information

Danielle Keats Citron and Helen Norton, *Intermediaries and Hate Speech: Fostering Digital Citizenship for Our Information Age*, 91 B.U. L. Rev. 1435 (2011), *available at* https://scholar.law.colorado.edu/faculty-articles/178.

ARTICLE

INTERMEDIARIES AND HATE SPEECH: FOSTERING DIGITAL CITIZENSHIP FOR OUR INFORMATION AGE

DANIELLE KEATS CITRON & HELEN NORTON*

INTRODUCTION 1436
I. CIVIC ENGAGEMENT, CYBER HATE, AND INTERMEDIARIES' POTENTIAL FOR FOSTERING DIGITAL CITIZENSHIP 1442
A. *The Internet's Potential to Deepen Civic Engagement* 1443
B. *Cyber Hate's Potential to Imperil Digital Citizenship* 1447
C. *Intermediaries' Freedom to Challenge Digital Hate* 1453
II. IMPLEMENTING A CONCEPTION OF DIGITAL CITIZENSHIP: A TRANSPARENT COMMITMENT TO FIGHTING HATE 1457
A. *The Transparency Principle* 1457
B. *An Illustrative Definitional Menu* 1459
1. Speech that Threatens and Incites Violence 1460

* Danielle Citron is Professor of Law at the University of Maryland School of Law and an Affiliate Fellow of the Stanford Center on Internet and Society and of the Yale Information Society Project. Helen Norton is Associate Dean for Academic Affairs and Associate Professor of Law at the University of Colorado School of Law. This Article grew out of a true partnership with both authors contributing equally. The Authors wish to thank the organizers of the Stanford Center on Internet and Society's Roundtable on Hate Speech; Columbia University Computer Science Department's Distinguished Lecture Series; Stanford Law School's 7th Annual E-Commerce conference; Washington University School of Law's Faculty Workshop Series; and the Progress and Freedom Foundation's panel on Censorship, Hate Speech, and Dissent for the chance to present early versions of this work. Rachel Arnow-Richman, Brad Bernthal, Richard Boldt, Ryan Calo, Miriam Cherry, Julie Cohen, Lisa Fairfax, Lauren Gelman, Don Gifford, Michelle Harner, Melissa Hart, Leslie Meltzer Henry, Deborah Hellman, Sonia Katyal, R.A. Lenhardt, Steven Morrison, Scott Moss, Patrick O'Donnell, Paul Ohm, Frank Pasquale, Neil Richards, Robert Rhee, Wendy Seltzer, Steve Sheinberg, Jana Singer, Catherine Smith, Daniel Solove, David Super, Berin Szoka, Alexander Tsesis, Barbara van Schewick, Chris Wolf, and the participants at University of Colorado and University of Maryland Law School faculty colloquia gave us terrific suggestions. We are grateful to the numerous safety officers, communication directors, and in-house counsel who talked to us about their approach to hate speech – their help was invaluable. Alice Johnson, Ovais Anwar, John Buchanan, Jordan Bunch, David Cline, Adam Farra, Matt Haven, Lindsey Lanzendorfer, and Susan McCarty provided excellent research assistance. Molly Carter and the editors at the *Boston University Law Review* gave us invaluable feedback and support.

2. Speech that Intentionally Inflicts Severe Emotional Distress 1463
3. Speech that Harasses 1464
4. Speech that Silences Counter-Speech 1466
5. Speech that Exacerbates Hatred or Prejudice by Defaming an Entire Group 1467
III. RESPONDING TO HATE SPEECH 1468
A. *Removing Hateful Content* 1468
B. *Countering Hate Speech with Speech* 1471
C. *Educating and Empowering Community Users* 1476
1. Education 1477
2. Empowerment 1478
3. Architectural Choices 1482
CONCLUSION 1484

No longer confined to isolated corners of the web, cyber hate now enjoys a major presence on popular social media sites. The Facebook group Kill a Jew Day, *for instance, acquired thousands of friends within days of its formation, while YouTube has hosted videos with names like* How to Kill a Beaner, Execute the Gays, *and* Murder Muslim Scum. *The mainstreaming of cyber hate has the troubling potential to shape public expectations of online discourse.*

Internet intermediaries have the freedom and influence to seize this defining moment in cyber hate's history. We believe that a thoughtful and nuanced intermediary-based approach to hate speech can foster respectful and vibrant online discourse. We urge intermediaries to help address cyber hate by adopting accessible and transparent policies that educate users about their rights and responsibilities as digital citizens. Intermediaries' options include challenging hateful speech by responding with counter-speech and empowering community members to enforce norms of digital citizenship.

INTRODUCTION

The Facebook group *Kill a Jew Day* declared July 4, 2010 as the start of an eighteen-day period of violence "anywhere you see a Jew."[1] The group's profile featured a swastika and images of corpses piled on top of one another.[2] Group members commented that they could not "wait to rape the dead baby Jews."[3]

The *Kill a Jew Day* social network group is an example of the more than 11,000 websites, videos, and social network groups devoted to spreading hate.[4]

---

[1] Yaakov Lappin, *'Kill a Jew' Page on Facebook Sparks Furor*, JERUSALEM POST, July 5, 2010, at 5.

[2] *Id.*

[3] *Id.*

[4] Jesse Solomon, *Hate Speech Infiltrates Social-Networking Sites, Report Says*, CNN (Mar. 15, 2010, 4:37 PM), http://www.cnn.com/2010/TECH/03/15/hate.speech.social.

Neo-Nazi websites allow users to maneuver virtual nooses over digital images of black men.[5] Videos posted online urge viewers to murder "Muslim scum"[6] and to kill homosexuals.[7] Typing "I hate spics" into Google generates 45,300 results.[8]

The greatest increase in digital hate has occurred on social media sites.[9] Examples include the *How to Kill a Beaner* video posted on YouTube, which allowed players to kill Latinos while shouting racial slurs,[10] and the Facebook group *Kick a Ginger Day*, which inspired physical attacks on students with red hair.[11] Facebook has hosted groups such as *Hitting Women*,[12] *Holocaust Is a Holohoax*,[13] and *Join if you hate homosexuals*.[14]

---

networks/index.html.

[5] Maria Seminerio, *"Hate Filter" Tackles Racist Sites*, ZDNET (Nov. 12, 1998, 3:29 PM), http://www.zdnet.co.uk/news/networking/1998/11/12/us-report-andquothatefilterand quot-tackles-racist-sites-2069870/.

[6] Mark MacAskill & Marcello Mega, *YouTube Cuts Murder Race-Hate Clips*, SUNDAY TIMES (London) (Sept. 28, 2008), http://www.timesonline.co.uk/tol/news/uk/scotland/article4837923.ece.

[7] Theresa Howard, *Online Hate Speech: It's Difficult to Police*, USA TODAY, Oct. 2, 2009, at 4D.

[8] Petition for Inquiry Filed on Behalf of the National Hispanic Media Coalition at 10, In the Matter of Hate Speech in the Media, Before the F.C.C., Jan. 28, 2009.

[9] *See generally* SIMON WIESENTHAL CENTER, FACEBOOK, YOUTUBE + HOW SOCIAL MEDIA OUTLETS IMPACT DIGITAL TERRORISM AND HATE (2009) (providing screenshots of social media websites promoting hate). Hate groups recruit new members on popular social network sites like YouTube and Facebook. *Social Networks Are New Sites for Hate Speech*, REUTERS, May 13, 2009, http://www.pcmag.com/article2/0,2817,2347004,00.asp.

[10] aborn88, *How to Kill a Beaner*, YOUTUBE (June 1, 2008), http://www.youtube.com/watch?v=Dq-tUPOGp8w.

[11] Liz Nordlinger, *Cartman Started It*, ST. PETERSBURG TIMES (Fla.), Feb. 25, 2010, at 8; Matthew Moore, *Facebook 'Kick a Ginger' Campaign Prompts Attacks on Redheads*, TELEGRAPH (U.K.) (Nov. 22, 2008, 12:47 AM), http://www.telegraph.co.uk/news/worldnews/northamerica/canada/3498766/Facebook-Kick-a-Ginger-campaign-prompts-attacks-on-redheads.html.

[12] Phil Bradley, *Facebook Group: Hitting Women*, PHIL BRADLEY'S WEBLOG (Feb. 18, 2010), http://philbradley.typepad.com/phil_bradleys_weblog/2010/02/facebook-group-hitting-women.html (reporting that as of February 10, 2010 the Facebook page *Hitting Women* remained on Facebook); Julie Ross Godar, *Facebook and Hate Speech: Are You a Fan of Hitting Women?*, BLOGHER (Feb. 18, 2010, 5:39 PM). As of December 20, 2010, the Facebook group *Hitting Women* was no longer available.

[13] Corilyn Shropshire, *Facebook Wrestles with Anti-Semitism*, HOUS. CHRON., May 15, 2009, at 6.

[14] David Badash, *Facebook or Hate Book? Facebook Shuts Down Anti-Gay Hate Groups!*, THE NEW CIVIL RIGHTS MOVEMENT (Mar. 9, 2010), http://thenewcivilrightsmovement.com/facebook-or-hate-book-facebook-shuts-down-anti-gay-hate-groups/successes/2010/03/09/8828.

Groups recognized cyberspace's potential to facilitate hate from its earliest days. In 1984, for example, the Aryan Nation sponsored a Usenet bulletin board featuring a "hit list," which included among its targets Alan Berg, a Jewish radio talk show host who had ignited the anger of the Order, an Aryan Nation spin-off group, by ridiculing the group on air.[15] Members of the Order murdered Berg in his driveway after the posting of the hit list.[16]

Even though cyber hate is not a new phenomenon, its recent growth is startling.[17] No longer isolated in little-known bulletin boards and websites, digital hate appears in the internet's mainstream. Digital hate's prevalence has considerable – and troubling – potential to shape public expectations of online discourse, especially as cyber hate penetrates social media populated with the young and impressionable. We thus face an important point in cyber hate's history and development: norms of subordination may overwhelm those of equality if hatred becomes an acceptable part of online discourse.

For these reasons, some scholars support governmental intervention to combat digital hate.[18] Governmental efforts to regulate hate speech based on its content, however, trigger important First Amendment and other concerns.[19] Given the challenges faced by regulatory solutions to the problem of digital hate, this Article focuses instead on the potential role of online intermediaries – private entities that host or index online content – in voluntarily addressing

---

[15] *See The Murder of Alan Berg: 25 Years Later*, DENVER POST, June 18, 2009, at A-01.

[16] *Id.*

[17] The Simon Wiesenthal Center has documented the extraordinary increase in online hate over the past ten years. SIMON WIESENTHAL CENTER, *supra* note 9, at 1 (discussing its growth from one website in 1995 to 10,000 today and the 25% increase in sites devoted to hate in the past year alone). This growth mirrors the escalating power and range of the technologies that facilitate the distribution of expression generally, including but not limited to digital hate. *See* Nathan Myhrvold, *Moore's Law Corollary: Pixel Power*, N.Y. TIMES, June 7, 2006, at G3 (explaining that the speed and breath of computing power doubles every eighteen months). As Microsoft founder Bill Gates explains of the information age, "we're always in a time of utter change, maybe even accelerating change." John Markoff, *Gates's Lieutenants Look Ahead, Hoping to Avoid Other Companies' Mistakes*, N.Y. TIMES, June 17, 2006, at C1.

[18] For various proposals to modify the First Amendment standards to be applied to governmental regulation of online hate speech, see Jennifer L. Brenner, *True Threats – A More Appropriate Standard for Analyzing First Amendment Protection and Free Speech When Violence Is Perpetrated over the Internet*, 78 N.D. L. REV. 753, 783 (2002); John P. Cronan, *The Next Challenge for the First Amendment: The Framework for an Internet Incitement Standard*, 51 CATH. U. L. REV. 425, 428 (2002); Nancy S. Kim, *Web Site Proprietorship and Online Harassment*, 2009 UTAH L. REV. 993, 997 (urging courts to impose tort liability upon website sponsors "for creating unreasonable business models" by failing to adopt "reasonable measures" to prevent foreseeable harm of online harassment).

[19] *See, e.g.*, R.A.V. v. City of St. Paul, 505 U.S. 377, 391 (1992) (holding that a city ordinance that prohibited expression that "arouses anger, alarm or resentment in others . . . on the basis of race, color, creed, religion, or gender" impermissibly discriminated on the basis of viewpoint in violation of the First Amendment (internal quotation marks omitted)).

cyber hate and its attendant harms.[20] Internet intermediaries[21] wield considerable control over what we see and hear today, akin to that of influential cable television and talk radio shows. Examples include search engines like Google, Microsoft, and Yahoo!; browsers like Mozilla; social network sites like Facebook, MySpace, and Formspring.me; micro-blogging services like Twitter; video-sharing sites like YouTube; and newsgathering services like Digg.[22] As more and more expression appears online, these intermediaries increasingly impact the flow of information.

Importantly, intermediaries have enormous freedom in choosing whether and how to challenge digital hate, as intermediaries' response to online speech remains largely free from legal constraint in the United States.[23] Not only are intermediaries free from First Amendment concerns as private actors, they are also statutorily immunized from liability for publishing content created by others as well as for removing that content.[24]

---

[20] Although this Article focuses only on private intermediaries' voluntary responses to cyber hate, we do not discount the possibility that government might have a role to play regarding the perpetrators of digital hate in at least some circumstances. *See, e.g.*, Danielle Keats Citron, *Cyber Civil Rights*, 89 B.U. L. REV. 61, 86-95 (2009) [hereinafter Citron, *Cyber Civil Rights*] (exploring law's coercive role in deterring and remedying cyber harassment); Danielle Keats Citron, *Law's Expressive Value in Combating Cyber Gender Harassment*, 108 MICH. L. REV. 373, 404-14 (2009) [hereinafter Citron, *Law's Expressive Value*] (documenting the expressive value of a cyber civil rights agenda in addressing cyber gender harassment).

[21] *See* 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 LOY. L.A. L. REV. 373, 386 (2010) (suggesting that intermediaries fall into three general categories: communication conduits, content hosts, and search/application providers). This Article focuses on voluntary measures available to a specific subset of internet intermediaries – content hosts and search/application providers – given their unique role in hosting online communities and in linking individuals to them. We leave to the side intermediaries that primarily serve as communication conduits (such as internet service and broadband providers) that we see as more akin to the phone company or the postal service in that they carry, but do not typically mediate, expressive content.

[22] This Article refers to sites that enable the production and sharing of digital content in mediated social settings as "social media." *See* Danielle Keats Citron, *Fulfilling Government 2.0's Promise with Robust Privacy Protections*, 78 GEO. WASH. L. REV. 822, 824 n.12 (2010) (explaining that social media include social-network sites, video-sharing sites, photo-sharing sites, and the like).

[23] This is not necessarily true outside of the United States, as many countries' laws require intermediaries to moderate content and to ensure its compliance with substantive restrictions. *See, e.g.*, Wendy Seltzer, Remarks, *The Politics of the Internet,* 102 AM. SOC'Y INT'L L. PROC. 45, 45-47 (2008) (describing how some countries require internet service providers, search engines, and other intermediaries to prevent in-country users from reaching certain sites).

[24] *See infra* note 111 and accompanying text.

This situation invites important and challenging questions about whether and how intermediaries might thoughtfully exercise their freedom and influence to shape on-line expression. Indeed, a number of intermediaries have begun to consider such questions, motivated by concerns about the potential business, ethical, and instrumental costs of digital hate. This has led many intermediaries to include hate speech prohibitions in their Terms of Service (TOS) agreements and Community Guidelines.

This Article proposes that intermediaries who feel a responsibility to challenge digital hate might also understand that responsibility to include fostering digital citizenship.[25] As we use the term in this Article, a commitment to digital citizenship seeks to protect users' capability to partake freely in the internet's diverse political, social, economic, and cultural opportunities, which informs and facilitates their civic engagement.[26] In short, a commitment to digital citizenship aims to secure robust *and* responsible participation in online life.

Intermediaries can foster digital citizenship by inculcating norms of respectful, vigorous engagement.[27] Just as law can be an "omnipresent

---

[25] For arguments that intermediaries can also play a central role in responding to defamation and other reputational harms, see David S. Ardia, *Reputation in a Networked World: Revisiting the Social Foundations of Defamation Law*, 45 HARV. C.R.-C.L. L. REV. 261, 264 (2010); Daniel H. Kahn, *Social Intermediaries: Creating a More Responsible Web Through Portable Identity, Cross-Web Reputation, and Code-Backed Norms*, 11 COLUM. SCI. & TECH. L. REV. 176, 195-96 (2010).

[26] *See* Jennifer Gordon & R.A. Lenhardt, *Rethinking Work and Citizenship*, 55 UCLA L. REV. 1161, 1185 (2008) (arguing that citizenship requires a person's ability to participate in society in a meaningful manner).

[27] In Lawrence Lessig's estimation, social norms may often regulate behavior as effectively as law. Lawrence Lessig, *The Law of the Horse: What Cyberlaw Might Teach*, 113 HARV. L. REV. 501, 507 (1999); Lawrence Lessig, *The New Chicago School*, 27 J. LEGAL STUD. 661, 669-70 (1998) [hereinafter Lessig, *The New Chicago School*] (explaining that institutions can shape behavior through the development of social norms, as well as through law, markets, and architecture). Nancy Kim similarly characterizes cyber harassment primarily as a failure of website operators' business norms, suggesting that tort law should encourage operators to engage in a range of preventive behaviors to deter online harassment. *See* Kim, *supra* note 18, at 996.

Here we note, but do not take part in, the debate over whether norms or law are more appropriate or effective in the context of internet governance. *Compare* Ardia, *supra* note 25, at 264 (proposing that online community governance through norms may often better protect users from reputational harms than defamation law), *and* Kahn, *supra* note 25, at 195-96 ("[W]e should account for the coming growth of norms in our decisions about when and how to regulate, as the new growth of norms may sometimes obviate (or occasionally exacerbate) the need for regulation."), *with* Mark A. Lemley, *The Law and Economics of Internet Norms*, 73 CHI.-KENT L. REV. 1257, 1260-61 (1998) (questioning the claim that reliance on norms is more effective than regulation in achieving cyberspace governance), *and* Neil Weinstock Netanel, *Cyberspace Self-Governance: A Skeptical View from Liberal Democratic Theory*, 88 CALIF. L. REV. 395, 401-02 (2000) (concluding that selective

teacher,"[28] intermediaries' voluntary actions can educate users about acceptable behavior. Their inaction in the face of online hate plays a similar role: intermediaries' silence can send a powerful message that targeted group members are second-class citizens.[29]

Specifically, we suggest that intermediaries can valuably advance the fight against digital hate with increased transparency – e.g., by ensuring that their efforts to define and proscribe hate speech explicitly turn on the harms to be targeted and prevented. This requires intermediaries to engage in thoughtful conversations with stakeholders externally and internally to identify the potential harms of hate speech (and its constraint) that *they* find most troubling. The more intermediaries and their users understand why a particular policy regulates a certain universe of speech, the more likely they can apply that policy in a way that achieves those objectives.

Not only can well-developed and transparent policies effectively acknowledge and address the meaningful distinctions between hate speech and other expression, intermediaries may also respond to hateful speech in ways other than simply removing it. Indeed, intermediaries' choices among available options – removing speech, responding with counter-speech, and empowering and educating community members to advance norms of digital citizenship themselves – may reflect the varying ways in which their different activities might facilitate the spread of online hate and thus undermine digital citizenship.

To be sure, self-governance is not without its shortcomings.[30] But because regulatory approaches to cyber hate are largely unavailable due to First Amendment constraints, intermediaries' voluntary efforts permit the development of flexible and nuanced solutions tailored to specific contexts.[31]

---

governmental regulation of cyberspace will better realize liberal democratic ideals than cyberspace self-governance). In evaluating comparative costs and benefits, that debate largely assumes the freedom to choose between law and norms in a particular context. This Article, in contrast, focuses on a context where such a choice is often unavailable because law – i.e., government regulation of online hate speech – is constrained by the First Amendment.

[28] Olmstead v. United States, 277 U.S. 438, 485 (1928) (Brandeis, J., dissenting).

[29] MARY ANN GLENDON, RIGHTS TALK: THE IMPOVERISHMENT OF POLITICAL DISCOURSE 101-05 (1991) (exploring how silence can provide misleading lessons about social responsibility ethos).

[30] *See, e.g.*, Lemley, *supra* note 27, at 1260-61 (discussing limitations of reliance on norms for cyberspace governance); Netanel, *supra* note 27, at 401-02 (discussing the advantages of selective governmental regulation of cyberspace over cyberspace self-governance).

[31] Joanne Scott & Susan Sturm, *Courts as Catalysts: Re-thinking the Judicial Role in New Governance*, 13 COLUM. J. EUR. L. 565, 566 (2006) ("New governance moves away from the idea of specific rights elaborated by formal legal bodies and enforced by judicially imposed sanctions. It locates responsibility for law-making in deliberative processes which are to be continually revised by participants in light of experience, and provides for

Scholars have suggested that "soft" approaches may be especially helpful when addressing issues that are particularly complex and politically intractable.[32] This is certainly true of hate speech, which involves challenging clashes between key commitments to free expression, autonomy, equality, and dignity. Soft approaches also promote solutions that reflect intermediaries' different business models, which offer varying services from which users can choose.[33]

This Article has three Parts. Part I summarizes the internet's potential for deepening civic engagement, as well as the substantial threats to that potential posed by digital hate. After describing the legal and political barriers to regulatory approaches to this problem, it explains that promising solutions nonetheless remain. More specifically, it documents the freedom and influence that intermediaries enjoy in shaping online expression generally and in addressing digital hate specifically.

Part II turns to implementation. It offers a range of recommendations for how intermediaries might exercise their power over cyber hate. We set forth an illustrative spectrum of possible hate speech definitions – grounded in terms of cyber hate's potential threats to digital citizenship as well as other specific harms – from which intermediaries might choose when developing their policies.

Part III then explores the variety of ways in which an intermediary might respond to speech that violates its policy. These include not only removal, but also engaging in or facilitating counter-speech, as well as educating and empowering users with respect to digital citizenship. We conclude that a thoughtful intermediary-based approach to hate speech can foster digital citizenship without suppressing valuable expression.

I. CIVIC ENGAGEMENT, CYBER HATE, AND INTERMEDIARIES' POTENTIAL FOR FOSTERING DIGITAL CITIZENSHIP

This Part starts by briefly recounting the internet's potential for deepening civic engagement and then summarizes the substantial threats to such engagement posed by digital hate. After identifying the legal and political

---

accountability through transparency and peer review.").

[32] *See id.* at 571 (describing the value of "normatively motivated inquiry and remediation by relevant non-judicial actors" in situations that involve unusual uncertainty or complexity).

[33] *See* Orly Lobel, *The Renew Deal: The Fall of Regulation and the Rise of Governance in Contemporary Legal Thought*, 89 MINN. L. REV. 342, 388, 389, 391 (2004) ("The governance model aims to create a flexible and fluid policy environment that fosters 'softer' processes that either replace or complement the traditional 'hard' ordering of the regulatory model[, such as] . . . social labeling, voluntary corporate codes of conduct, private accreditation, and certification by nongovernmental actors. . . . Flexibility implies variation in the communications of intention to control and discipline deviance. Less coercive sanctions can promote flexibility in implementation and compliance.").

barriers to governmental solutions to this problem, it explains that promising solutions remain in the form of voluntary measures by interested intermediaries.

A. *The Internet's Potential to Deepen Civic Engagement*

Among the many reasons to celebrate the internet's growth is its potential to enhance civic engagement, which in turn facilitates democratic functions. Democracy is often said to work best when citizens build networks of social interaction and trust.[34] Civic engagement allows people to see their lives as entwined with others.[35] In turn, people learn "habits of cooperation and public-spiritedness."[36] Civic engagement reinforces Alexis de Tocqueville's notion of "self-interest well understood" – that is, the capacity to consider the interests of others in addition to one's own[37] – and encourages "responsible citizenship."[38]

Although citizenship often describes a legal status enjoyed by members of a body politic,[39] citizenship can refer more broadly to participation in community life,[40] which may not be explicitly political but may ultimately further political participation.[41] Citizenship extends beyond the legal dimension to include "all the relationships . . . involved in membership in a community."[42] Citizenship "provides what the other roles cannot, namely an integrative experience which brings together the multiple role-activities of the contemporary person and demands that the separate roles be surveyed from a more general point of view."[43]

---

[34] ROBERT D. PUTNAM, BOWLING ALONE: THE COLLAPSE AND REVIVAL OF AMERICAN COMMUNITY 137-47 (2000).

[35] JOHN STUART MILL, *Considerations on Representative Government, reprinted in* THREE ESSAYS 143, 196-98 (Oxford Univ. Press 1975).

[36] PUTNAM, *supra* note 34, at 338.

[37] ALEXIS DE TOCQUEVILLE, DEMOCRACY IN AMERICA 501 (Harvey C. Mansfield & Delba Wintrop eds. and trans., 2000) (1835).

[38] RICHARD DAGGER, CIVIC VIRTUES: RIGHTS, CITIZENSHIP, AND REPUBLICAN LIBERALISM 104 (1997).

[39] *Id.* at 99.

[40] MILL, *supra* note 35, at 196.

[41] Indeed, public participation and civic engagement are often viewed as essential for members of a democracy to form a citizenry. JÜRGEN HABERMAS, BETWEEN FACTS AND NORMS: CONTRIBUTIONS TO A DISCOURSE THEORY OF LAW AND DEMOCRACY 307-08 (William Rehg trans., The MIT Press 1996) (1992); *see also* MILL, *supra* note 35, at 196-97 (explaining that a citizen is someone who develops his faculties through active engagement in public life). For John Dewey, citizen participation in communal life constituted the "idea of a democracy." JOHN DEWEY, THE PUBLIC AND ITS PROBLEMS 148-50 (1927).

[42] John Dewey, *The School as Social Centre*, 3 THE ELEMENTARY SCH. TCHR. 73, 76 (1902).

[43] SHELDON S. WOLIN, POLITICS AND VISION: CONTINUITY AND INNOVATION IN WESTERN POLITICAL THOUGHT 434 (1960).

Online activity can facilitate civic engagement and political participation. Neighborhood communities combine offline activities with online ones.[44] Companies encourage employees to use social network sites to deepen workplace relationships.[45] Workers can therefore discuss issues in person and in online postings.[46] Student organizations meet face-to-face in classrooms and in social network groups.[47]

Mediating institutions like schools, workplaces, churches, and community centers have traditionally given expression to the idea of citizenship.[48] This is especially so for institutions cultivating norms of trust across lines of social division, often referred to as "bridging ties."[49] Tocqueville emphasized the importance of townships and civic associations in enabling citizens to acquire the skills and habits of dialogue.[50] John Dewey found schools uniquely

---

[44] Amitai Etzioni, *On Virtual, Democratic Communities*, *in* COMMUNITY IN THE DIGITAL AGE: PHILOSOPHY AND PRACTICE 225, 228-29 (Andrew Feenberg & Darin Barney eds., 2004) (explaining that the town of Blacksburg, Virginia has an online community called Blacksburg Electronic Village where various groups and neighborhoods post meetings, share information, and interact). Examples abound of online political engagement, including the use of social media to raise campaign funds and organize voters during the 2008 Presidential election. Miki Caul Kittilson & Russell J. Dalton, *Virtual Civil Society: The New Frontier of Social Capital?*, POL. BEHAV. (Oct. 7, 2010) (forthcoming), *available at* http://www.springerlink.com/content/740r3560j640080t/; *see also* Nathaniel J. Gleicher, *MoneyBombs and Democratic Participation: Regulating Internet Fundraising*, 70 MD. L. REV. (forthcoming 2011) (manuscript at 5-6), *available at* http://ssrn.com/abstract=1695552.

[45] Jacob Christensen, *Managing Mondays: Facebook, a Viable Workplace Tool?*, LINKED 2 LEADERSHIP BLOG (Apr. 5, 2010), http://linked2leadership.com/2010/04/05/mm-facebook-a-workplace-tool/; *Two on Facebook . . . FNN Video and Employee Groups*, ONE DEGREE (Jan. 11, 2008), http://www.onedegree.ca/2008/01/two-on-facebook.html.

[46] For a discussion of the relationship between workplace relationships and civic engagement, see CYNTHIA ESTLUND, WORKING TOGETHER: HOW WORKPLACE BONDS STRENGTHEN A DIVERSE DEMOCRACY. Estlund suggests, however, that the rise of internet technology in the workplace may weaken rather than strengthen these bonds. *Id.* at 36-38.

[47] Popular social network sites like Facebook and MySpace were originally organized around existing institutions like schools, universities, and towns to enhance existing social connections. FELICIA WU SONG, VIRTUAL COMMUNITIES: BOWLING ALONE, ONLINE TOGETHER 22 (Digital Formations No. 54 2009).

[48] BENJAMIN R. BARBER, STRONG DEMOCRACY: PARTICIPATORY POLITICS FOR A NEW AGE 267 (1984).

[49] ESTLUND, *supra* note 46, at 107-08; *see also* BARBER, *supra* note 48, at 268. Not all associations contribute to liberal conceptions of democracy, however. MICHAEL J. SANDEL, DEMOCRACY'S DISCONTENT: AMERICA IN SEARCH OF A PUBLIC PHILOSOPHY 314-15 (1996). Some groups pursue distinctly illiberal aims, as this Article explores.

[50] TOCQUEVILLE, *supra* note 37, at 65, 496-97 (highlighting the importance of townships and civil associations because they allow citizens to "govern society" in the "restricted sphere that is within his reach"); *see also* SANDEL, *supra* note 49, at 333-35, 343 (extolling municipal parks, schools, libraries, community development corporations, and local retail

situated to teach children and adults about the social meaning of community[51] because they brought diverse people together in ways that "introduce deeper sympathy and wider understanding."[52] For Cynthia Estlund, the workplace serves as an important site for the formation of social and political views because it permits informal discourse among people "who are both *connected* with each other, so that they are inclined to listen, and *different* from each other, so that they are exposed to diverse ideas and experiences."[53]

Similarly, online intermediaries have potential to serve as mediating institutions that give expression to the idea of citizenship.[54] They can *extend* workplaces, schoolhouses, and community centers to digital spaces,[55] supplementing real-space exchanges of information and opinion with virtual ones. Online intermediaries also play an indispensable role in bringing together minority or marginalized groups in different geographic locations.[56] As Anupam Chander has noted, cyberspace can help "give members of minority groups a fuller sense of citizenship – a right to a practice of citizenship that better reflects who they are."[57]

---

establishments because they bring together rich and poor in public places and in public pursuits); Charlotte Garden, *Labor Values Are First Amendment Values: Why Union Comprehensive Campaigns Are Protected Speech*, 79 FORDHAM L. REV. 2617, 2656-58 (2011).

[51] *See* DEWEY, *supra* note 41, at 200.

[52] Dewey, *supra* note 42, at 83; *see* Harry C. Boyte, *A Different Kind of Politics: John Dewey and the Meaning of Citizenship in the 21st Century*. 12 GOOD SOC'Y, no. 2, 2003 at 1, 7. Dewey enlisted schools in the battle against bigotry: intolerance would lose force if exposed to "the ideas of others." Dewey, *supra* note 42, at 77.

[53] ESTLUND, *supra* note 46, at 123. She also emphasized the workplace's potential for enforcing civility and equality, which in turn allows diverse voices to be heard. *Id.* at 121-22.

[54] For an insightful discussion of schools as crucial speech-facilitating institutions, see Joseph Blocher, *Institutions in the Marketplace of Ideas*, 57 DUKE L.J. 821, 856-59 (2008).

[55] *See* SONG, *supra* note 47, at 4 (explaining that social media providers mediate practices of businesses, schools, and associations). A 2007 study found that Facebook cultivates bridging social capital. Nicole B. Ellison et al., *The Benefits of Facebook "Friends:" Social Capital and College Students' Use of Online Social Network Sites*, 12 J. COMPUTER-MEDIATED COMM. 1143, 1161-62 (2007); *see also* Sebastian Valenzuela et al., Lessons from Facebook: The Effect of Social Network Sites on College Students' Social Capital 33 (Apr. 5, 2008) (unpublished manuscript), *available at* http://online.journalism.utexas.edu/2008/papers/Valenzuela.pdf (finding that Facebook users come from diverse backgrounds, contrary to the popular myth that they are typically female, upper-middle class students). Intermediaries of course also support bonding ties – those involving groups of similar backgrounds.

[56] SONG, *supra* note 47, at 74.

[57] Anupam Chander, *Whose Republic?*, 69 U. CHI. L. REV. 1479, 1481 (2002) (reviewing CASS SUNSTEIN, REPUBLIC.COM (2001)). For an innovative conception of transnational cultural citizenship, see Sonia K. Katyal, *The Dissident Citizen*, 57 UCLA L. REV. 1415, 1467-75 (2010).

In these and myriad other ways, users of online intermediaries can participate in community life.[58] When we speak of "digital citizenship" in this Article, we refer to the ways in which online activity has the potential to deepen civic engagement.[59]

Of course, that the internet carries the promise of fostering digital citizenship does not mean that such promise inevitably will be realized.[60] Online communications can instead foster isolation and disengagement.[61] Timothy Zick explains that networked technologies can interfere with expression in public spaces by distracting people from face-to-face interactions.[62] Robert Putnam questions whether the internet will generate norms of trust given its facilitation of anonymous interactions that lack wider social context.[63] The next Section focuses more specifically on the perils to such engagement posed by cyber hate.

---

[58] Timothy Fort defines "mediating institutions" to mean "communities which socialize their members," and "require individuals to grasp their responsibilities to others, at least within their group, so that a person's very identity is developed." Timothy L. Fort, *The Corporation as Mediating Institution: An Efficacious Synthesis of Stakeholder Theory and Corporate Constituency Statutes*, 73 NOTRE DAME L. REV. 173, 174-75 (1997); *see also* Andrew Crane et al., *Stakeholders as Citizens? Rethinking Rights, Participation, and Democracy*, 53 J. BUS. ETHICS 107, 108 (2004) (describing various conceptions of corporate citizenship as including "corporations as citizens; corporations as administrators of citizenship; and stakeholders as citizens").

[59] The term "digital citizenship" can mean different things depending upon the community and audience. For instance, some political scientists have used the term to refer broadly to the "ability to participate in society online," arguing that disadvantaged groups cannot fully participate as citizens due to their limited access to the internet. KAREN MOSSBERGER, CAROLINE J. TOLBERT & RAMONA S. MCNEAL, DIGITAL CITIZENSHIP: THE INTERNET, SOCIETY, AND PARTICIPATION 1 (2008). Intermediaries, too, have invoked the concept of digital citizenship as an aspiration for civil online behavior. Indeed, we first encountered this term in conversations with those in the industry who had already identified the facilitation of "digital citizenship" as a goal. Interview with Hemanshu Nigam, former Chief Safety Officer, MySpace (June 22, 2010).

[60] Evgeny Morozov's work explores the related, though distinct, question of how democratic freedoms can be threatened by governmental abuse of networked technologies. *See generally* EVGENY MOROZOV, THE NET DELUSION: THE DARK SIDE OF INTERNET FREEDOM (2011). Anupam Chander has also explored this question with great insight in *Googling Freedom*, 99 CALIF. L. REV. 1 (2011).

[61] SONG, *supra* note 47, at 23.

[62] TIMOTHY ZICK, SPEECH OUT OF DOORS: PRESERVING FIRST AMENDMENT LIBERTIES IN PUBLIC PLACES 304 (2009). Professor Zick explains that new technologies contribute to the phenomenon of "absent presence" where people occupy personal "technology bubbles" and disconnect from others who physically surround them. *Id.*

[63] PUTNAM, *supra* note 34, at 175-76; *see also* Chander, *supra* note 57 at 1480 ("Which of these possible uses of the Internet – the Internet as a tool for discovery and education, or the Internet as an echo chamber – will find more adherents is an empirical question that we may not yet be able to answer.").

B. *Cyber Hate's Potential to Imperil Digital Citizenship*

Online activities can pose dangers that work to undermine civic engagement. The internet facilitates anonymous and pseudonymous discourse, which can just as easily accelerate destructive behavior as it can fuel public discourse.[64] It provides a cheap and easy way to reach like-minded individuals located at disparate geographic locations, removing barriers that often limit group activity. [65] Search engines ensure access to, and the persistence of, online content of all types – including hateful content. Hate groups exploit these and other online attributes to spread, legitimize, and entrench hateful messages that imperil participation in community life.[66]

Cyber threats and calls for violence can undermine political and civic engagement. History[67] and social science[68] confirm that hate speech may

---

[64] Social science research suggests that people may behave more aggressively when they believe that they cannot be observed and caught. Citron, *Cyber Civil Rights*, *supra* note 20, at 82.

[65] *See* Kyu Ho Youm, *First Amendment Law: Hate Speech, Equality, and Freedom of Expression*, 51 J. COMM. 406, 406 (2001) (book review) (describing reports by Don Black – the "godfather of the Internet racist movement" – that the internet dramatically increased his ability to disseminate his views compared to his previous reliance on traditional print media).

[66] *See* ADAM G. KLEIN, A SPACE FOR HATE: THE WHITE POWER MOVEMENT'S ADAPTATION INTO CYBERSPACE 55 (2009) (describing "information laundering" to mean "the legitimizing factor of an interconnected information superhighway of web directories, research engines, news outlets, and social networks that collectively funnel into and out of today's hate websites"). Klein continues:

> For information-seekers, the result of this funneling process is a wider array of perspectives, and thus, a broader understanding of any given topic. However, for propaganda-providers like the white power movement, the same process inadvertently lends the credibility and reputation of authentic websites to those illegitimate few to which they are nonetheless connected. Such is the case with many of today's leading search engines like Google, that unwittingly filter directly into hate websites, or public networks like YouTube, which host their venomous content everyday.

Id.

[67] *See* Mari J. Matsuda, *Public Response to Racist Speech: Considering the Victim's Story*, 87 MICH. L. REV. 2320, 2352 n.166 (1989) (describing history of escalating racist violence that accompanies racist speech); Alexander Tsesis, *Dignity and Speech: The Regulation of Hate Speech in a Democracy*, 44 WAKE FOREST L. REV. 499, 509-15 (2009) [hereinafter Tsesis, *Dignity and Speech*] (describing history of anti-Semitic and racist speech that incited or escalated violent acts); Alexander Tsesis, *The Empirical Shortcomings of First Amendment Jurisprudence: A Historical Perspective on the Power of Hate Speech*, 40 SANTA CLARA L. REV. 729, 740-55 (2000) (detailing the relationship between hate speech and acts of violence against Jews, Native Americans and African-Americans).

[68] *See, e.g.*, David Kretzmer, *Freedom of Speech and Racism*, 8 CARDOZO L. REV. 445, 463 (1987) (describing social science demonstrating the importance of speech as a precondition to acts of racial violence or scapegoating).

facilitate acts of violence against members of targeted groups.[69] For instance, digital hatred helped inspire the 1999 shooting of African-Americans, Asian-Americans, and Jews in suburban Chicago by Benjamin Smith, a member of the white supremacist group World Church of the Creator (WCOTC) that promotes racial holy war.[70] Just months before the shootings, Smith told documentary filmmaker Beverly Peterson that: "It wasn't really 'til I got on the internet, read some literature of these groups that . . . it really all came together."[71]

More recently, the Facebook group *Kick a Ginger Day* urged members to get their "steel toes ready" to attack individuals with red hair.[72] The site achieved its stated goal: students punched and kicked children with red hair, with dozens of Facebook members claiming credit online for the attacks.[73]

Aside from producing physical harm, online calls for violence and threats can silence members of targeted groups.[74] Consider a California teenager's

---

[69] Hate speech that takes the form of "fighting words" may sometimes provoke violent responses from its targets in addition to inciting violence against them. *See* Charles R. Lawrence III, *If He Hollers, Let Him Go: Regulating Racist Speech on Campus*, 1990 DUKE L.J. 431, 452; Ronald Turner, *Regulating Hate Speech and the First Amendment: The Attractions of, and Objections to, an Explicit Harms-Based Analysis*, 29 IND. L. REV. 257, 298-300 (1995) (describing violence provoked by use of the n-word or other face-to-face uses of particular racial or religious epithets).

[70] Christopher Wolf, *Racists, Bigots and the Law on the Internet: Internet Hate Speech and the Law*, ANTI-DEFAMATION LEAGUE, http://www.adl.org/Internet/Internet_law3.asp (last visited Apr. 5, 2011). Smith killed former Northwestern University basketball coach Ricky Byrdsong and Indiana University student Won Joon Yoon and wounded six Orthodox Jews and three African-Americans. Elizabeth Brackett, *The Hate Crimes Question*, PBS ONLINE NEWS HOUR (Aug. 11, 1999), http://www.pbs.org/newshour/bb/law/july-dec99/hate_8-11.html. The internet helped make the WCOTC one of the fastest growing hate groups in the United States. *Id.*

[71] *The Consequences of Right-Wing Extremism on the Internet: Inspiring Extremist Crimes*, ANTI-DEFAMATION LEAGUE, http://www.adl.org/internet/extremism_rw/inspiring.asp (last visited Apr. 5, 2011). WCOTC's website operator at the time of the rampage confirmed that Smith sent him "about five" email messages "congratulating" him on the group's websites and indicating that he regularly read them. *Id.*

[72] Nordlinger, *supra* note 11, at 8; Moore, *supra* note 11.

[73] Nordlinger, *supra* note 11, at 8.

[74] *See* Richard Delgado & David Yun, *The Neoconservative Case Against Hate-Speech Regulation – Lively, D'Souza, Gates, Carter, and the Toughlove Crowd*, 47 VAND. L. REV. 1807, 1822-23 (1994) (explaining how our culture has developed a host of narratives about overcoming hurt feelings while ignoring hurtful words that undermine victims' ability to respond and to mobilize effectively against hate); Lawrence, *supra* note 69, at 452 ("When racial insults are hurled at minorities, the response might be silence or flight rather than a fight, but the preemptive effect on further speech is just as complete as with fighting words."); Netanel, *supra* note 27, at 426 ("Individuals may develop deep feelings of attachment and loyalty to virtual communities and may be devastated by perceived wrongs within those communities. In such instances, exit is far from costless."); Steven H. Shiffrin,

experience with internet hate speech. Commenters (later discovered to be students at the teenager's high school) on the student's website repeatedly threatened him in homophobic ways.[75] One wrote: "F-----, I'm going to kill you."[76] Another wrote: "If I ever see you I'm . . . going to pound your head in with an ice pick."[77] Others wrote "You are an oversized f----- . . . . I just want to hit you in the neck" and "I hate f--s . . . . You need to be stopped."[78] The student's father shut down the site and, on the advice of the police, kept his son from attending school during the investigation.[79]

In a similar vein, Kathy Sierra, a well-known programmer, maintained a popular blog on software development called "Creating Passionate Users."[80] In 2007, anonymous posters verbally attacked Ms. Sierra on her blog and two other websites.[81] On her blog, commenters suggested that she deserved to have her throat slit, be suffocated, sexually violated, and hanged.[82] On another blog, posters uploaded doctored photographs of Ms. Sierra: one picture featured her with a noose beside her neck; another depicted her screaming while being suffocated by lingerie.[83] After the attacks, Ms. Sierra canceled speaking engagements and feared leaving her home.[84] As she explained, "my blog was in the Technorati Top 100 [at the time of the attack]. I have not blogged there – or anywhere – since."[85]

Racist Speech, Outsider Jurisprudence, and the Meaning of America, 80 CORNELL L. REV. 43, 86 (1994) ("[R]acial vilification can create a repressive environment in which the speech of people of color is chilled or not heard."); Mike Adams, *Facebook Devolves into Dark Web of Anonymous Hate Speech*, NATURALNEWS (Aug. 26, 2010), https://www.naturalnews.com/029572_Facebook_hate_speech.html (stating that hate speech on Facebook has caused individuals who would otherwise be participating in the public discourse to close their accounts).

[75] Kim Zetter, *Court: Cyberbullying Threats Are Not Protected Speech*, WIRED BLOG: THREAT LEVEL (Mar. 18, 2010, 3:23 PM), http://www.wired.com/threatlevel/2010/03/cyberbullying-not-protected/.

[76] *Id.*

[77] D.C. v. R.R., 106 Cal. Rptr. 3d 399, 405 (Ct. App. 2010).

[78] *Id.* at 407

[79] *Id.* at 446 (Rothschild, J., dissenting).

[80] Dahlia Lithwick, *Fear of Blogging: Why Women Shouldn't Apologize for Being Afraid of Threats on the Web*, SLATE (May 4, 2007, 7:20 PM), http://www.slate.com/id/2165654/.

[81] *Id.*

[82] *Id.*; Greg Sandoval, *Blogger Cancels Conference Appearance After Death Threats*, CNET NEWS BLOG (Mar. 26, 2007), http://news.cnet.com/8301-10784_3-6170683-7.html.

[83] Jessica Valenti, *Women: How the Web Became a Sexists' Paradise*, GUARDIAN (London), Apr. 6, 2007, at 16; Sandoval, *supra* note 82.

[84] *Blog Death Threats Spark Debate*, BBC NEWS (Mar. 27, 2007), http://news.bbc.co.uk/go/pr/fr/-/2/hi/technology/6499095.stm.

[85] Kathy Sierra, Comment to *CCR Symposium: A Behavioral Argument for Stronger Protections*, CONCURRING OPINIONS (Apr. 18, 2009, 2:25 PM), http://www.concurringopinions.com/archives/2009/04/ccr_symposium_a_1.html#comments.

Consider too the posters on a white supremacist website who targeted Bonnie Jouhari, a civil rights advocate and mother of a biracial girl.[86] The site showed a picture of Ms. Jouhari's workplace exploding in flames next to the threat that "race traitors" are "hung from the neck from the nearest tree or lamp post."[87] Posters included bomb-making instructions and a picture of a hooded Klansman holding a noose.[88] Ms. Jouhari and her daughter have withdrawn from public life.[89] They do not have driver's licenses, voter registration cards, or bank accounts for fear of creating a public record of their whereabouts.[90]

Cyber hate can undermine targeted group members' capability for civic engagement in other ways apart from threatening or inciting violence. It can convey the message that a group in the community is "not worthy of equal citizenship."[91] R.A. Lenhardt explains that hate speech undermines group members' ability to belong and participate in processes crucial to community life.[92] Online hate can thus denigrate group members' basic standing in society and deprive them of their "civic dignity."[93]

In this way, cyber hate can inflict serious psychological injury, including fear, stress, feelings of inferiority, and depression.[94] Recall the attacks upon

---

[86] Ryan Wilson, HUDALJ 03-98-0692-8 ¶¶ 2-3, 6 (July 19, 2000). For an excellent description and analysis of the case, see Catherine E. Smith, *Intentional Infliction of Emotional Distress: An Old Arrow Targets the New Hate Hydra*, 80 DENV. U. L. REV. 1, 35-48 (2002).

[87] Wilson, HUDALJ 03-98-0692-8, at ¶¶ 9-11.

[88] *Id.* at ¶¶ 9, 15

[89] DeWayne Wickham, *They Suffer for Doing Right Thing*, USA TODAY (May 16, 2000, 8:39 AM), http://www.usatoday.com/news/opinion/columnists/wickham/wick093.htm (explaining that Ms. Jouhari and her daughter have moved four times to ensure that posters do not find them).

[90] *Id.*

[91] Jeremy Waldron, *Dignity and Defamation: The Visibility of Hate*, 123 HARV. L. REV. 1596, 1601 (2010).

[92] R.A. Lenhardt, *Understanding the Mark: Race, Stigma, and Equality in Context*, 79 N.Y.U. L. REV. 803, 844-48 (2004); *see also* KENNETH L. KARST, BELONGING TO AMERICA: EQUAL CITIZENSHIP AND THE CONSTITUTION 3 (1989) (offering a principle of equal citizenship that suggests people are "entitled to be treated" as "respected, responsible, and participating member[s]"). Jennifer Gordon and R.A. Lenhardt's theory of belonging focuses on formal and informal pathways to the genuine possession and exercise of citizenship in the United States, including political participation, work, and education. Gordon & Lenhardt, *supra* note 26, at 1186-88.

[93] Waldron, *supra* note 91, at 1607.

[94] *See* Matsuda, *supra* note 67, at 2332 (describing the harm of "[t]he spoken message of hatred"); Shiffrin, *supra* note 74, at 86 (describing how racist speech inflicts harm on its individual victims by inspiring self hatred, isolation, and emotional distress); *see also* Kretzmer, *supra* note 68, at 466 (describing how hate speech may trigger insecurity, self hatred, humiliation, isolation, and other psychological harm); Lawrence, *supra* note 69, at 462 (describing how racial epithets and harassment cause deep emotional scarring in the form of anxiety and fear); *cf.* Citron, *Law's Expressive Value*, *supra* note 20, at 388-90

Ms. Jouhari: Ms. Jouhari suffered headaches and anxiety, and her daughter was diagnosed as suffering from severe post-traumatic stress disorder.[95] Indeed, young people can feel such psychological harms intensely, as electronic media exert a powerful influence on children and teenagers who have not yet reached full cognitive development.[96] Not only are children particularly vulnerable to hate's emotional harms, they are also less able to fight back.[97]

Hate speech may further degrade public discourse by skewing society's assessment of members of certain racial, religious, or other groups and of their ideas.[98] Charles Lawrence, for example, argues that racism "trumps good ideas that contend with it in the market, often without our even knowing it."[99] By devaluing targeted group members' expression, hate speech can produce a process defect in the marketplace of ideas.[100]

Moreover, because hate speech may inspire or deepen prejudice, it can lead to discriminatory decisions about jobs, housing, and other life opportunities.[101] Stigma, often exacerbated or inspired by hate speech, can render targeted group members dishonored and erect significant barriers to full acceptance into the wider community.[102] Not only does such bigotry impose tangible costs on targeted group members who suffer the effects of discriminatory decisions, it more broadly undermines society's commitment to equality and dignity.[103]

(exploring how cyber gender harassment produces anxiety and other forms of emotional distress).

[95] Ryan Wilson, HUDALJ 03-98-0692-8, at 24-25 (July 19, 2000).

[96] Michele L. Ybarra et al., *Linkages Between Internet and Other Media Violence with Seriously Violent Behavior by Youth*, 122 PEDIATRICS 929, 933 (2008).

[97] *See* Richard Delgado, *Words that Wound: A Tort Action for Racial Insults, Epithets, and Name-Calling*, 17 HARV. C.R.-C.L. L. REV. 133, 147 (1982).

[98] As Charles Lawrence powerfully observes, the notions of "racial inferiority of non-whites infects, skews, and disables the operation of the market (like a computer virus, sick cattle, or diseased wheat)." Lawrence, *supra* note 69, at 468.

[99] *Id.*

[100] *Id.* Websites and other online actors may exacerbate this process defect by enabling hate groups to link exclusively to hateful content, creating "echo chambers" of extreme positions, which can harden and encourage the development of even more extreme views. CASS R. SUNSTEIN, REPUBLIC.COM 2.0 145 (2007).

[101] *See* Delgado & Yun, *supra* note 74, at 1813 (maintaining that hate speech feeds discriminatory decision-making by reinforcing stereotypes); Kretzmer, *supra* note 68, at 505. In this way, prejudice and bigotry fostered by hate speech can produce conscious and unconscious behavioral consequences and thus intensify their targets' disadvantage. Jerry Kang, *Trojan Horses of Race*, 118 HARV. L. REV. 1489, 1539-40 (2005) (examining the role played by racial stereotypes in mass media in creating and maintaining biases that result in discriminatory decision-making).

[102] Lenhardt, *supra* note 92, at 844-48; *see also* ERVING GOFFMAN, STIGMA: NOTES ON THE MANAGEMENT OF SPOILED IDENTITY 2-5 (1963) (discussing stigma as creating a spoiled social identity).

[103] *See* Delgado, *supra* note 97, at 142 (explaining that racist speech undermines "society

In turn, search engines ensure the persistence of cyber hate and its costs to digital citizenship. Because search engines reproduce information cached online, targets of hate speech cannot depend upon time's passage to alleviate the damage that online postings cause.[104] For this reason, Jeremy Waldron contends that cyber hate produces a "permanent disfigurement" of group members.[105] In all these ways, cyber hate threatens to undermine digital citizenship.[106]

In our opinion, the threats posed by online hate to digital citizenship are sufficiently substantial to demand a response. Regulatory solutions, however, face considerable constitutional and political barriers. Governmental efforts to regulate hate speech based on its content, for example, trigger significant First Amendment concerns.[107]

---

as a whole").

[104] Danielle Keats Citron, *Mainstreaming Privacy Torts*, 98 CALIF. L. REV. 1805, 1813 (2010); Jeffrey Rosen, *The End of Forgetting*, N.Y. TIMES, July 25, 2010, at MM30.

[105] Waldron, *supra* note 91, at 1601, 1610.

[106] Although our discussion here focuses on the harm to civic engagement posed by digital hate, we recognize that such hate speech inflicts other moral and instrumental harms as well.

[107] *See, e.g.*, R.A.V. v. City of St. Paul, 505 U.S. 377, 391 (1992) (holding that city ordinance that prohibited expression that "arouses anger, alarm or resentment in others . . . on the basis of race, color, creed, religion, or gender" impermissibly discriminated on the basis of viewpoint in violation of the First Amendment). Indeed, as some thoughtful commentators have observed, regulatory efforts to constrain hate speech not only face constitutional challenges, but may impose instrumental costs of their own. For example, visible hate speech can remind readers and listeners of bigotry's prevalence and the need to enforce existing antidiscrimination laws. Shiffrin, *supra* note 74at 89. It may perform a powerful teaching function in exposing the poverty of a hate group's beliefs. Kingsley R. Browne, *Title VII as Censorship: Hostile-Environment Harassment and the First Amendment*, 52 OHIO ST. L.J. 481, 542 (1991) (arguing that permitting hate speech can contribute to the elimination of prejudice because such speech will expose the poverty of those beliefs). Others suggest that hateful expression may play a role in preventing violence by allowing speakers to let off steam. Vincent Blasi, *The Teaching Function of the First Amendment*, 87 COLUM. L. REV. 387, 408 (1987) (reviewing LEE C. BOLLINGER, THE TOLERANT SOCIETY (1986)). *But see* Dhammika Dharmapala & Richard H. McAdams, *Words That Kill? An Economic Model of the Influence of Speech on Behavior (with Particular Reference to Hate Speech)*, 34 J. LEG. STUD. 93, 132 (2005) (discussing how raising the costs of engaging in hate speech may deter hate crime rather than increase the rate of hate crime). Refraining from regulating hate speech may avoid making martyrs of – and thus invigorating and multiplying – hateful speakers. Graham Hughes, *Prohibiting Incitement to Racial Discrimination*, 16 U. TORONTO L.J. 361, 365 (1996) (suggesting that regulation creates martyrs and converts to the cause of hatred); Larissa Barnett Lidsky, *Where's the Harm?: Free Speech and the Regulation of Lies*, 65 WASH. & LEE L. REV. 1091, 1099-1100 (2008) (concluding that punishing Holocaust denial will paradoxically entrench that view and inspire stronger belief in conspiracy theories). So, too, the expression of hate speech might foster a certain capacity of mind, enabling us to confront our biases, master our irrational passions, and develop further tolerance ourselves. LEE C.

Apart from the First Amendment difficulties confronted by governmental efforts to regulate digital hate, such efforts face considerable political challenges as well, as demonstrated by the experience of those who proposed legislation to address discriminatory conduct far beyond the realm of pure expression. For example, the Hate Crimes Prevention Act – which criminalizes bias-motivated crimes of violence – was enacted only after years of effort.[108] Along the same lines, legislation to prohibit job discrimination on the basis of sexual orientation has been introduced in Congress in various forms since 1975 but has yet to be enacted.[109]

For these reasons, calls for governmental responses to digital hate face substantial challenges. As the next Section explains, however, promising alternatives remain available.

C. *Intermediaries' Freedom to Challenge Digital Hate*

Internet intermediaries enjoy enormous freedom to decide whether and how to shape online expression. The First Amendment, of course, protects speech only from governmental restriction and thus does not govern private actors' decisions to remove or filter online expression.[110] At the same time, federal law immunizes "provider[s] or user[s] of interactive computer services" from liability arising from content created by others and from requirements to remove "offensive" speech.[111] Intermediaries thus enjoy wide latitude to make

---

BOLLINGER, THE TOLERANT SOCIETY: FREEDOM OF SPEECH AND EXTREMIST SPEECH IN AMERICA 142, 173 (1986).

[108] Matthew Shepard and James Byrd, Jr. Hate Crimes Prevention Act, Pub. L. No. 111-84, 123 Stat 2190 (2009) (to be codified at 18 U.S.C. § 249).

[109] H.R. REP. NO. 110-406 pt. 1, at 2 (2007).

[110] *See, e.g.*, Green v. Am. Online (AOL), 318 F.3d 465, 472 (3d Cir. 2003) (finding that private company AOL is not subject to constitutional free speech guarantees and has not been transformed into a state actor simply because it "provides a connection to the Internet on which government and taxpayer-funded websites are found"); Langdon v. Google, Inc., 474 F. Supp. 2d 622, 631 (D. Del. 2007) (ruling that Google, Yahoo!, and Microsoft are private companies not subject to constitutional free speech guarantees even though they may work with state actors like public universities).

[111] 47 U.S.C. § 230(c) (2006); *see also* Zeran v. Am. Online, Inc., 129 F.3d 327, 330 (4th Cir. 1997) (barring claims against an online service provider under § 230 because defendant did not create the allegedly tortious content). Intermediaries can incur liability for content that they create themselves (e.g., for their own postings that are defamatory or threatening), or for publishing content that violates copyright law. 47 U.S.C. § 230(e)(1); *see also* Wendy Seltzer, *Free Speech Unmoored in Copyright's Safe Harbor: Chilling Effects of the DMCA on the First Amendment*, 24 HARV. J.L. & TECH. 171, 228 (2010) (noting that § 230 "specifically excludes intellectual property and criminal claims from its protections"). For instance, Internet Service Providers (ISPs) and website operators can incur liability under the Digital Millennium Copyright Act for refusing to take down content that they have been notified violates copyright law, whereas they enjoy immunity from liability for defamatory postings created by others. *Id.* at 175.

all sorts of decisions – including none at all – with respect to others' hate speech.[112]

A number of intermediaries have begun to consider such questions, variously motivated by concerns about the potential business, moral, and instrumental costs of digital hate. Some intermediaries see digital hate as a potential threat to profits.[113] MySpace,[114] for instance, sees its aggressive approach to hate speech – and, indeed, to a wide range of potentially offensive speech in addition to hate speech – as essential to securing online advertising for its customer base.[115] According to MySpace's former Chief Safety Officer Hemanshu Nigam, its approach stems from its sense of "what the company stood for and what would attract advertising and revenue."[116] Nigam explains that because kids and adults use MySpace, the company wanted to ensure a "family friendly" site, which could only be accomplished by taking down content that "attacked an individual or group because they are in that group and . . . made people feel bad."[117] As Nigam suggests, voluntary efforts to address hate speech may serve an intermediary's bottom line by creating market niches and contributing to consumer goodwill.[118]

---

[112] *See* Seltzer, *supra* note 111, at 228 (explaining that internet service providers can thus "set their own terms of service – choosing to maintain 'family-friendly' environments, attempting to build communities, or taking a hands-off, anything goes approach").

[113] Such intermediaries may explain their actions under the traditional "shareholder primacy" view that understands the corporation's primary (and perhaps exclusive) objective as maximizing shareholder wealth. *See, e.g.*, Mark J. Roe, *The Shareholder Wealth Maximization Norm and Industrial Organization*, 149 U. PA. L. REV. 2063, 2065 (2001); *For Whom Corporate Managers* Are *Trustees: A Note*, 45 HARV. L. REV. 1365, 1367-69 (1932). Along these lines, intermediaries' sense of the bottom-line benefits of addressing hate speech can be shaped by consumers' – i.e., users' – expectations.

[114] Although Facebook has now overtaken MySpace in popularity, MySpace remains popular. *Search Results for Myspace*, QUANTCAST CORP., http://www.quantcast.com/search?q=myspace (last visited Apr. 16, 2011) (noting that MySpace is the 35th most popular site in the United States).

[115] MySpace prohibits content that "promotes or otherwise incites racism, bigotry, hatred or physical harm of any kind against any group or individual . . . [or] exploits people in a sexual or violent manner." *MySpace.com Terms of Use Agreement*, MYSPACE.COM, http://www.myspace.com/help/terms (last visited Apr. 7, 2011).

[116] Interview with Nigam, *supra* note 59.

[117] *Id.*

[118] *Id.*; *see also* Paul Alan Levy, *Stanley Fish Leads the Charge Against Immunity for Internet Hosts – But Ignores the Costs*, CONSUMER L. & POL'Y BLOG (Jan. 8, 2011), http://pubcit.typepad.com/clpblog/2011/01/stanley-fish-leads-the-charge-against-immunity-for-internet-hosts-but-ignores-the-costs.html (arguing that websites that fail to provide protections against abuse will find "that the ordinary consumers whom they hope to serve will find it too uncomfortable to spend time on their sites, and their sites will lose social utility (and, perhaps more cynically, they know they will lose page views that help their ad revenue)").

Some intermediaries are motivated to address digital hate based on their sense of their own corporate social responsibility.[119] Indeed, many intermediaries explicitly invoke broad social responsibility principles when describing their services and their mission.[120] For example, Google explains that in offering the platform Blogger to users, it "want[s] to be socially responsible."[121] For this reason, it admonishes users that they can utilize "Blogger to express [their] opinions, even very controversial ones," but that they cannot "cross the line by publishing hate speech."[122]

---

[119] Such decisions may be justified as a matter of corporate law under the social entity theory of the corporation, which permits corporate decision-makers to consider and serve the interests of all the various constituencies affected by the corporation's operation. *See* Lisa M. Fairfax, *Doing Well While Doing Good: Reassessing the Scope of Directors' Fiduciary Obligations in For-Profit Corporations with Non-Shareholder Beneficiaries*, 59 WASH. & LEE L. REV. 409, 412 (2002).

[120] Yahoo! lists its company values as including "an infectious sense of mission to make an impact on society and empower consumers in ways never before possible. We are committed to serving both the Internet community and our own communities." *Yahoo! – What We Value,* YAHOO! INC., http://docs.yahoo.com/info/values/ (last visited Apr. 7, 2011). Yahoo! also states that it "empower[s] people through corporate social responsibility programs, products, and services to make a positive impact on their communities." *Overview*, YAHOO! INC., http://pressroom.yahoo.net/pr/ycorp/overview.aspx (last visited June 1, 2011). Google makes clear that it will pursue policies that may conflict with short-term shareholder economic gain but that it believes will benefit shareholders in the long term, and which may include benefits other than economic ones. Google Inc., Initial Public Offering Letter: *'An Owner's Manual' for Google's Shareholders* (Form S-1/A) (Aug. 18, 2004), *available at* http://investor.google.com/corporate/2004/ipo-founders-letter.html. Microsoft states:

> As a global company, we are accountable to millions of customers and stakeholders around the world. As we work to meet their needs, we are committed to creating value for our partners, employees, and wider society, and to managing our business sustainably. This commitment gives focus to our corporate citizenship work and helps us measure our performance over time.

Our Commitments, MICROSOFT, http://www.microsoft.com/about/corporatecitizenship/en-xf/our-commitments/ (last visited Apr. 7, 2011). Following the "Goals" link from that page brings a description of a link titled "Corporate Governance," which states that "considering the interests of other stakeholders – employees, customers, partners, suppliers, and the many communities around the world where we do business – is important to achieving the long-term interests of Microsoft shareholders." *Goals*, MICROSOFT, http://www.microsoft.com/about/corporatecitizenship/en-xf/our-commitments/goals/ (last visited Apr. 7, 2011).

[121] Rachel Whetstone, *Free Expression and Controversial Content on the Web*, THE OFFICIAL GOOGLE BLOG (Nov. 14, 2007, 3:58 PM), http://googleblog.blogspot.com/2007/11/free-expression-and-controversial.html.

[122] Google Blogger requires users to refrain from promoting "hate or violence towards groups based on race, ethnicity, religion, disability, gender, age, veteran status, or sexual orientation/gender identity." *Blogger Content Policy*, GOOGLE, http://www.blogger.com/content.g (last visited Apr. 7, 2011). Google further admonishes Blogger users: "don't write a blog saying that members of Race X are criminals or advocating violence against

Other intermediaries have invoked similar values in response to certain types of online hatred. After Facebook took down the *Kill a Jew Day* page in May 2010,[123] its spokesperson Andrew Noyes explained:

> Unfortunately ignorant people exist and we absolutely feel a social responsibility to silence them on Facebook if their statements turn to direct hate. That's why we have policies that prohibit hateful content and we have built a robust reporting infrastructure and an expansive team to review reports and remove content quickly.[124]

***

As this Part documents, intermediaries have the ability to decide whether and how to shape online expression. Many have elected to use that freedom to challenge online hate speech. Of course, many others have not. Indeed, some intermediaries base their business on tolerating or encouraging cyber hate. This is true, for instance, of the social network site Hate Book, which urges its users to "Post something you hate!"[125]

This Article addresses those intermediaries interested in combating their users' cyber hate. We urge them to consider the ways in which their services can be used to enrich as well as to endanger civic engagement.[126] In so doing, we recognize that a focus on the effects on civic engagement is not the only – nor necessarily the best – way of understanding the harms of hate speech. Nonetheless, we identify a commitment to digital citizenship as among the justifications for developing thoughtful approaches to hate speech, and one that could motivate interested intermediaries as well. In the remainder of this Article, we examine more specifically how intermediaries might address those challenges in developing and implementing hate speech policies.

This Article discusses intermediaries' choices in light of the freedom that they enjoy under current law. Indeed, Congress has encouraged intermediary involvement, providing immunity for intermediaries who take down "offensive material."[127] We note, however, that some thoughtful commentators challenge that status quo, arguing that select intermediaries should be treated as monopolies and thus subject to greater regulation.[128] Some of this discussion

---

followers of Religion Y." *Id.*

[123] *See supra* notes 1-3 and accompanying text.

[124] Lappin, *supra* note 1. Facebook is "sensitive to content that includes pornography, bullying, hate speech, and actionable threats of violence." *Id.*

[125] *Hate Book*, HATE BOOK, http://www.hatebook.com/tos.php (last visited Apr. 7, 2011).

[126] As Neil Netanel wrote of intermediaries who exclude individuals from their networks due to their race or gender, these intermediaries' actions "work a fundamental impairment not only of 'netizenship,' but also of citizenship in territorial polities." Netanel, *supra* note 27, at 457.

[127] 47 U.S.C. § 230(c)(2) (2006).

[128] *See, e.g.*, Oren Bracha & Frank Pasquale, *Federal Search Commission? Access, Fairness, and Accountability in the Law of Search*, 93 CORNELL L. REV. 1149, 1180-82 (2008) (deeming search engine Google a natural monopoly deserving of public regulation).

relates to net neutrality debates over the regulation of broadband network operators that we do not address in this Article.[129] To the extent that some urge greater regulation of social media and search engine intermediaries discussed here, their concerns do not stem from such intermediaries' attention to hate speech issues.[130]

II. IMPLEMENTING A CONCEPTION OF DIGITAL CITIZENSHIP: A TRANSPARENT COMMITMENT TO FIGHTING HATE

As explained above, significant moral and policy justifications support intermediaries who choose to engage in voluntary efforts to combat hate speech. Indeed, many intermediaries already choose to address online hatred in some way.[131] In this Part, we urge intermediaries – and others – to think and speak more carefully about the harms they hope to forestall when developing hate speech policies.

A. *The Transparency Principle*

We believe that intermediaries can valuably advance the fight against digital hate with more transparency and specificity about the harms that their hate speech policies address, as well as the consequences of policy violations. With more transparency regarding their specific reasons for choosing to address digital hate, intermediaries can make behavioral expectations more understandable.[132] Without it, intermediaries will be less effective in expressing what it means to be responsible users of their services.

In a series of articles, Frank Pasquale has argued that an Internet Intermediary Regulatory Council should oversee search engines and carriers, assisting the FCC and FTC in carrying out their present missions. Frank Pasquale, *Trusting (and Verifying) Online Intermediaries' Policing*, *in* THE NEXT DIGITAL DECADE: ESSAYS ON THE FUTURE OF THE INTERNET 347, 348 (Berin Szoka & Adam Marcus eds., 2010), *available at* http://nextdigitaldecade.com/read-book/now. Pasquale argues that such regulation of Google is warranted given "Google's dominance of the general search market," the company's indispensable role in economic, cultural, and political life, and the opacity of its practices that immobilize consumer voice options. *Id.*

[129] *See, e.g.*, BARBARA VAN SCHEWICK, INTERNET ARCHITECTURE AND INNOVATION 222-51 (2010).

[130] *See, e.g.*, DAWN C. NUNZIATO, VIRTUAL FREEDOM: NET NEUTRALITY AND FREE SPEECH IN THE INTERNET AGE, at xiv-xv (2009) (arguing that Congress should pass a law, or require the FCC, to prohibit broadband providers from blocking legal content or applications and from engaging in various forms of discrimination and prioritization of packets and that perhaps law should regulate powerful search engines such as Google as well).

[131] *See supra* notes 113-124 and accompanying text.

[132] Past calls for transparency from these entities have focused on legitimacy concerns regarding stealth marketing and undisclosed political and cultural biases. These have included Frank Pasquale, *Beyond Innovation and Competition: The Need for Qualified Transparency in Internet Intermediaries*, 104 NW. U. L. REV. 105, 155 (2010) (discussing

Indeed, those intermediaries that address hate speech in their Terms of Service (TOS) agreements or Community Guidelines rarely define key terms like "hateful" or "racist" speech with specificity.[133] The terms of service of Yahoo!, for instance, requires users of some of its services to refrain from generating "hateful, or racially, ethnically or otherwise objectionable" content without saying more.[134] Microsoft's gaming service Xbox Live warns users that they may not publish content that "incites . . . hatred [or] bigotry"[135] or that is "related to or suggestive of hate speech (including but not limited to racial, ethnic, or religious slurs)."[136] Some intermediaries attribute their reluctance to address digital hate to the difficulties in defining such speech.[137]

We do not pretend that we can make hard choices easy, nor do we advocate for a particular definition of hate speech. We recognize that intermediaries' decisions will turn on their available resources, business interests, and varied

---

particular institutional solutions); Frank Pasquale, *Internet Nondiscrimination Principles: Commercial Ethics for Carriers and Search Engines,* 2008 U. CHI. LEGAL F. 263, 268-69 (discussing regulation of search engines and social networks).

[133] TOS agreements typically include not only an intermediary's hate speech policy (if any), but also its privacy policies, which typically notify users that they can opt-out of the collection of personally identifiable information. Commentators have criticized TOS privacy policies on the grounds that users do not pay attention to them and thus do not make meaningful choices about their privacy, which can lead to the collection and use of personal information. Danielle Citron, *The Boucher Privacy Bill: A Little Something For Everyone Yet Nothing for All?*, CONCURRING OPINIONS (June 13, 2010, 11:37 AM), http://www.concurringopinions.com/archives/2010/06/the-boucher-privacy-bill-a-little-something-for-everyone-yet-nothing-for-all.html. Ryan Calo's scholarship thoughtfully responds to critiques of notice provisions. *See, e.g.*, M. Ryan Calo, *Against Notice Skepticism*, 87 NOTRE DAME L. REV. (forthcoming 2012), *available at* http://papers.ssrn.com/sol3/papers.cfm?abstract_id=1790144#%23.

[134] *Yahoo! Terms of Service*, YAHOO! INC., http://info.yahoo.com/legal/us/yahoo/utos/utos-173.html (last visited Apr. 7, 2011). These policies apply to some of Yahoo!'s services other than its search engine, such as its Flickr photo-sharing service. *Id.* For instance, Yahoo! explains that its Answer application is not "a soapbox to vent personal frustrations or rant about issues. We are a community of people with diverse beliefs, opinions, and backgrounds, so please be respectful and keep hateful and incendiary comments off Yahoo! Answers." *Yahoo! Answers Community Guidelines*, YAHOO! INC., http://answers.yahoo.com/info/community_guidelines (last visited Apr. 7, 2011).

[135] *Xbox LIVE Terms of Use*, MICROSOFT, http://www.xbox.com/en-US/legal/LiveTOU (last visited Apr. 7, 2011).

[136] *Xbox LIVE Code of Conduct*, MICROSOFT, http://www.xbox.com/en-US/legal/codeofconduct (last visited Apr. 7, 2011).

[137] For instance, Twitter's Director of Program Development remarked: "What counts as name calling? There are sites that do employ teams of people that do that investigation . . . but we feel that's a job we wouldn't do well." Anick Jesdanun, *On the Internet, Free Speech Is No Guarantee*, HAMPTONROADS.COM (July 21, 2008), http://hamptonroads.com/2008/07/internet-free-speech-no-guarantee.

assessments of their corporate social responsibility.[138] Instead, we hope to encourage intermediaries – and others – to think and speak more carefully about the harms they hope to forestall when developing hate speech policies. The more intermediaries and users understand *why* a particular policy prohibits a certain universe of speech, the more they will be able to execute the policy in a way that achieves those objectives.[139] This understanding will require intermediaries to engage in thoughtful conversations with stakeholders both externally and internally to identify the particular potential harms of hate speech – and the harms of its constraint – that they find most troubling.

No matter the particular definition of hate speech that intermediaries choose, an accessible and transparent policy can help users develop a better appreciation of their responsibilities as they work, debate, and connect with others online. Hard judgment calls will inevitably remain, regardless of how an intermediary chooses to define hate speech. But those decisions – however difficult – can be made in a more principled way when an intermediary grounds its policy's hate speech definition and application in terms of the specific harms it seeks to avoid. In the next section, we explore a spectrum of definitions available to intermediaries to guide them in this effort.

B. *An Illustrative Definitional Menu*

We propose that an intermediary's voluntary efforts to define and proscribe hate speech should expressly turn on the harms to be targeted and prevented. Rather than identifying new harms, here we rely on thoughtful commentary about First Amendment controversies over proposed governmental regulation of hate speech in outlining a menu of possible harm-based definitions.

---

[138] *See* Frank Pasquale, *Asterisk Revisited: Debating a Right of Reply on Search Results*, 3 J. BUS. & TECH. L. 61, 73 (2008) (recognizing that resource constraints will limit intermediary duties, but recommending some such duties nevertheless); Frank Pasquale, *Rankings, Reductionism, and Responsibility*, 54 CLEV. ST. L. REV. 115, 117 (2006) (cautioning against too easy acceptance of reductionist presentations of reality by intermediaries).

[139] Along these lines, Lisa Fairfax has documented the extent to which an institution's written commitments – such as a corporation's rhetoric evincing a responsibility to groups and interests beyond their shareholders – encourage and shape its actual behavior. Lisa M. Fairfax, *Easier Said Than Done? A Corporate Law Theory for Actualizing Social Responsibility Rhetoric*, 59 FLA. L. REV. 771, 776 (2007) (debunking "the notion that corporate [social responsibility] rhetoric has no connection to actual practice [and] demonstrat[ing] the manner in which such rhetoric can be used strategically" to shape behavior). Fairfax continues: "when an individual expresses a commitment to a given idea or principle, the human preference for consistency generates internal and external pressures to engage in behavior consistent with that commitment." *Id.* She also points to social psychology literature showing that the "more often someone makes a commitment, the more likely she is to engage in corresponding behavior." *Id.* at 777.

1. Speech that Threatens and Incites Violence

Intermediaries may define prohibited hate speech as that which threatens or encourages violence against individuals. In an area where consensus is exceedingly rare, most commentators seem to agree that these harms are sufficiently serious to warrant prohibiting such speech.[140] Indeed, the United States Supreme Court has held that speech that constitutes a "true threat"[141] or intentional incitement to imminent violence[142] is unprotected by the First Amendment and within the government's power to regulate.

Whether certain speech is likely to incite imminent violence or will lead reasonable people to fear violence will vary with the content and context of the expression.[143] Key factors in making such evaluations include the clarity with which the speech advocates violence and the specificity with which individuals are identified as potential targets. As courts have noted, for example, the inclusion of a target's personal identification information can contribute to a reasonable person's conclusion that the expression communicates the intent to inflict bodily harm upon the target.[144]

---

[140] *See, e.g.*, John T. Nockleby, *Hate Speech in Context: The Case of Verbal Threats*, 42 BUFF. L. REV. 653, 708 (1994) (urging government regulation of only that universe of hate speech perceived by the listener as threatening violence); Frederick Schauer, *Uncoupling Free Speech*, 92 COLUM. L. REV. 1321, 1349 (1992) (defining actionable hate speech as "first, utterances intended to and likely to have the effect of inducing others to commit acts of violence or acts of unlawful discrimination based on the race, religion, gender, or sexual orientation of the victim; and, second, utterances addressed to and intended to harm the listener (or viewer) because of her race, religion, gender, or sexual orientation").

[141] *See* Virginia v. Black, 538 U.S. 343, 365-66 (2003) (holding that the First Amendment permits states to prohibit individuals from burning crosses only when it is done "with the intent to intimidate."); Watts v. United States, 394 U.S. 705, 705-08 (1969) (defining a true threat as that which a reasonable person would consider an expression of the speaker's intent to inflict bodily harm).

[142] *See* United States v. White, 610 F.3d 956, 957, 962 (7th Cir. 2010) (holding that, under Supreme Court precedent, internet speech in which the poster intends to request or solicit a violent crime is not protected by the First Amendment, and declining to dismiss the government's solicitation case based on the defendant's website that "posted personal information about a juror who served on the Matthew Hale jury, along with postings calling for the use of violence on enemies of white supremacy").

[143] Of course causation remains a challenging issue even under some of the narrower definitions of hate speech. But although we may not be able to say with certainty that certain statements will actually lead to violence, we can be more confident in stating that certain speech will reasonably lead targets to fear such violence.

[144] *See* Planned Parenthood of Columbia/Willamette, Inc. v. Am. Coal. of Life Activists, 290 F.3d 1058, 1080 (9th Cir. 2002) (en banc) (holding that the Nuremberg Files' website could be characterized as an unprotected true threat, where the site listed the names, addresses, and license plate numbers of abortion providers, with the names of those who had been murdered lined through in black, and the names of those wounded highlighted in grey); *see also* United States v. Fullmer, 584 F.3d 132, 156 (3d Cir. 2009) (concluding that animal rights activists' website included expression that instilled fears in its targets and thus could

These factors can help intermediaries determine whether certain situations should be characterized as threats of, or incitement to, violence. Posters on a Yahoo! bulletin board, for instance, listed names of specific Arab-Americans alongside their home addresses, telephone numbers, and the suggestion that they are "Islamic terrorists."[145] There, the targeted individuals notified Yahoo!, which immediately took down the postings.[146] Neo-Nazi Hal Turner's blog postings offer another illustration of targeted speech that threatens or incites violence. A jury convicted Turner in a criminal case based on his postings saying that Judges Frank Easterbrook, Richard Posner, and William Bauer "deserve to be killed," along with the targets' photographs, work locations, and a picture of their courthouse modified to show the locations of "anti-truck bomb barriers."[147]

Intermediaries could also define hate speech as that which urges violence against groups as well as specific individuals. For example, Turner's website also urged readers to murder "illegal aliens": "We're going to have to start killing these people . . . . I advocate using extreme violence against illegal aliens. Clean your guns . . . . Find out where the largest gathering of illegal aliens will be near you . . . and then do what has to be done."[148] In response to similar concerns, Facebook explained that neo-Nazi and other hate groups calling for violence against gypsies,[149] Jews,[150] and even red-headed people[151] violated its hate speech policy.

To be sure, definitional challenges remain under a policy that constrains only hate speech that threatens or incites violence against specific individuals or groups. Of course, some situations present more difficult questions than others. For example, would a reasonable person understand certain online

---

be prosecuted as true threats unprotected by the First Amendment).

[145] Tom Spring, *Digital Hate Speech Roars*, PC WORLD, (Sept. 21, 2001, 7:00 PM), http://www.pcworld.com/article/63225/digital_hate_speech_roars.html.

[146] *Id.* Another web hosting company took down sites proclaiming that minorities should be hanged. Raphael Cohen-Almagor & Sharon Haleva-Amir, *Bloody Wednesday in Dawson College – The Story of Kimveer Gill, or Why Should We Monitor Certain Websites to Prevent Murder*, 2 STUD. IN ETHICS, L. & TECH. J. no. 3, 2008 at 1, 22-23.

[147] James Joyner, *Hal Turner and the Limits of Free Speech*, OUTSIDE THE BELTWAY (Aug. 16, 2009), http://www.outsidethebeltway.com/hal_turner_and_the_limits_of_free_speech/; *see also* Tom Hays, *NJ Blogger Convicted of Threatening III Judges*, ASSOCIATED PRESS, Aug. 13, 2010, *available at* http://www.boston.com/news/local/connecticut/articles/2010/08/13/nj_blogger_convicted_of_threatening_ill_judges.

[148] Susy Buchanan & David Holthouse, *Extremists Advocate Murder of Immigrants, Politicians*, S. POVERTY L. CTR. INTELLIGENCE PROJECT (Mar. 30, 2006), http://www.splcenter.org/intel/news/item.jsp?aid=49.

[149] Robin Pomeroy, *Facebook Pulls Italian Neo-Nazi Pages After Outcry*, REUTERS (Nov. 14, 2008), http://www.reuters.com/article/idUSTRE4AD3KZ20081114.

[150] Lappin, *supra* note 1, at 5.

[151] *See supra* notes 72-73 and accompanying text (discussing Facebook's decision to take down the *Kick a Ginger Day* groups).

speech – such as the use of certain cultural symbols, like nooses, burning crosses, and swastikas[152] – to communicate a true, if implied, threat? As the Supreme Court has observed with respect to cross-burning, some symbols in certain contexts – but not in all contexts – effectively express frightening threats.[153] But contextual inquiry is as inevitable as it is difficult under *any* definition of hate speech. Focusing on the specific harms to be prevented can help us sharpen and justify our inquiry in a principled way.

Some online actors specifically prohibit users from threatening or inciting violence in a manner that helpfully explains their community norms. For instance, Beliefnet, a website devoted to providing information on a wide variety of topics related to faith and spirituality, defines hate speech to mean "speech that may cause violence toward someone (even if unintentionally) because of their age, disability, gender, ethnicity, race, nationality, religion or sexual orientation."[154] The policy explains that unlike mere insults, speech "that may cause violence" includes that which advocates violence against protected class members or states that such violence is "acceptable [or] . . . deserved . . . perhaps by characterizing them as guilty of a heinous crime, perversion, or illness, such that violence may seem allowable or inconsequential."[155] Further boosting its value to users, the policy discusses the reasons underlying the rule,[156] its relationship to free speech guarantees,[157]

---

[152] *See, e.g.*, Timothy Zick, *Cross Burning, Cockfighting, and Symbolic Meaning: Toward a First Amendment Ethnography*, 45 WM. & MARY L. REV. 2261, 2346-49 (2004) (describing the use of context and cultural meaning to determine whether cross-burning communicates threats of violence or instead political protest).

[153] Virginia v. Black, 538 U.S. 343, 365-66 (2003) (holding that the First Amendment permits states to prohibit individuals from burning crosses "with the intent to intimidate"). Alexander Tsesis argues that cultural symbols of hate, like burning crosses or swastikas, are effective at intimidation because such symbols trigger in victims a well-grounded fear of physical violence. *See* Tsesis, *Dignity and Speech*, *supra* note 67, at 503-04 ("Destructive messages are particularly dangerous when they rely on historically established symbolism, such as burning crosses or swastikas, in order to kindle widely shared prejudices.").

[154] *Hate Speech and the Beliefnet Community*, BELIEFNET, http://www.beliefnet.com/Skipped/2004/06/Hate-Speech.aspx (last visited Apr. 8, 2011).

[155] *Id.*

[156] The website explains that it developed this policy because of its concern that certain forms of hate speech can, and have, inspired violent acts. *Id.*

[157] *Id.* ("Hate speech is legal in the United States. Americans may choose to read or engage in hate speech. Likewise, Americans may choose to gather in groups where they mutually agree upon standards of conduct that do not include hate speech. As a private website, Beliefnet is a choice for those who want civil discussion that is free of hate speech. When speech could incite harm to individuals, harm to the Beliefnet community, or harm to Beliefnet, it is appropriate for us to place limits on it. If you wish to engage in hate speech, there are numerous options available on the Internet. This is not one of them.").

its application to certain challenging contexts (e.g., discussions of homosexuality),[158] and specific practical guidelines for its use.[159]

2. Speech that Intentionally Inflicts Severe Emotional Distress

Along the same lines, intermediaries might define hate speech to include that which intentionally inflicts severe emotional distress. Although this inquiry too is inevitably context-specific, a body of tort law illuminates factors that courts use in determining if speech amounts to intentional infliction of emotional distress.[160] As Benjamin Zipursky explains, "[o]ver decades and even centuries, courts recognized clusters of cases" that constituted extreme and outrageous behavior outside the norms of decency.[161] These most often involve expression that is individually targeted, especially threatening or humiliating, repeated, or reliant on especially sensitive or outrageous material.[162]

---

[158] *Id.* ("We recognize that many faith groups are engaged in important debate about homosexuality and its relationship to faith. We encourage members to discuss this topic on Beliefnet and have created specific forums for this debate. . . . You may express the belief that homosexuality is wrong, or that it is sinful. . . . You may not advocate violence against anyone because of their sexual orientation.").

[159] *Id.*

[160] Scholars have cautioned that the First Amendment requires a very narrow understanding of this tort to ensure that government does not constrain offensive speech on the basis of viewpoint. *See, e.g.*, Eugene Volokh, *Freedom of Speech and the Intentional Infliction of Emotional Distress Tort*, 2010 CARDOZO L. REV. DE NOVO 300, 300-03; Christina Wells, *Regulating Offensiveness:* Snyder v. Phelps, *Emotion, and the First Amendment*, 1 CALIF. L. REV. CIRCUIT 71, 72 (2010). Along these lines, the Supreme Court has held that the First Amendment prohibits the tort's application to a defendant who "addressed matters of public import on public property, in a peaceful manner, in full compliance with the guidance of local officials." Snyder v. Phelps, 131 S. Ct. 1207, 1220 (2011).

[161] Benjamin Zipursky, Snyder v. Phelps, *Outrageousness, and the Open Texture of Tort Law*, 60 DEPAUL L. REV. (forthcoming 2011) (manuscript at 31), *available at* http://ssrn.com/abstract=1687688.

[162] *See* Citron, *Cyber Civil Rights*, *supra* note 20, 87-88; Smith, *supra* note 86, 35-48; *see also* Nadine Strossen, *The Tensions Between Regulating Workplace Harassment and the First Amendment: No Trump*, 71 CHI.-KENT L. REV. 701, 716-17 (1995) (suggesting that proscribable harassment under Title VII focus on workplace speech that directly targets a particular individual and that is so extreme that it amounts to intentional infliction of emotional distress). Along these lines, intermediaries' policies might also address defamatory hate speech. As the Supreme Court's First Amendment doctrine makes clear, the harms of defamatory speech – i.e., culpably false statements of fact that damage the target's reputation – are sufficiently great to justify its regulation by the government under certain circumstances. *See, e.g.*, N.Y. Times Co. v. Sullivan, 376 U.S. 254, 301-02 (1964); *see also* BOLLINGER, *supra* note 107, at 186 (explaining that his tolerance theory permits the regulation of libel because it targets an individual for harm and the purposes of toleration are not served by insisting that an individual – rather than the community as a whole – bear

Recall, for example, Bonnie Jouhari's experience with digital hate.[163] There, an administrative law judge determined that the website operator intentionally inflicted emotional distress on Jouhari and her daughter through "a relentless campaign of domestic terrorism."[164]

3. Speech that Harasses

Intermediaries might choose to define hate speech as that which would rise to the level of actionable harassment if it occurred at work or in school. Although harassment in the employment and education contexts does not parallel that in cyberspace in important respects,[165] internet intermediaries remain free to consider these efforts when crafting their own policies.

Longstanding anti-harassment principles permit government to regulate harassing speech at work or at school if such speech is sufficiently severe or pervasive to create a discriminatory educational or workplace environment.[166] Factors relevant to assessing whether verbal or written conduct meets this standard include "the frequency of the discriminatory conduct; its severity; whether it is physically threatening or humiliating, or a mere offensive utterance;" and whether it inflicts psychological harm.[167]

In the educational context, for example, verbal or written conduct violates Title IX's statutory prohibitions on discrimination by federally funded educational activities when the "harassment is so severe, pervasive, and objectively offensive that it can be said to deprive the victims of access to the educational opportunities or benefits provided by the school."[168] Along these

---

the harm of speech activity).

[163] *See supra* notes 86-90 and accompanying text.

[164] Ryan Wilson, HUDALJ 03-98-0692-8, at 19 (July 19, 2000).

[165] *See* Helen Norton, *Regulating Cyberharassment: Some Thoughts on* Sexual Harassment 2.0, 87 DENV. U. L. REV. ONLINE 11, 11-15 (2010) (identifying the difficulties in extending the First Amendment analysis applicable to governmental regulation of harassment at work and school to proposed government regulation of cyber harassment).

[166] *See* R.A.V. v. City of St. Paul, 505 U.S. 377, 389 (1992) (describing Title VII's regulation of harassing speech in the workplace as permissible under the First Amendment as proscribing "sexually derogatory 'fighting words,'" within "Title VII's general prohibition against sexual discrimination in employment practices"); Wisconsin v. Mitchell, 508 U.S. 476, 487 (1993) (explaining that the Court in *R.A.V.* "cited Title VII['s prohibition of sexual harassment] as an example of a permissible content-neutral regulation of conduct").

[167] Harris v. Forklift Sys., Inc., 510 U.S. 17, 23 (1993) (identifying factors relevant to a conclusion that workplace harassment rises to the level of a Title VII violation).

[168] Davis v. Monroe Cnty. Bd. of Educ., 526 U.S. 629, 633 (1999) (interpreting Title VI's prohibition on sex discrimination by federally funded educational activities); *see also* Racial Incidents and Harassment Against Students at Educational Institutions; Investigative Guidance, 59 Fed. Reg. 11448, 11449 (Mar. 10, 1994) (interpreting Title VI's prohibition on race and national origin discrimination by federally funded activities to include "harassing conduct (e.g., physical, verbal, graphic, or written) that is sufficiently severe,

lines, Bryn Mawr College defines harassment to include "verbal behavior such as unwanted sexual comments, suggestions, jokes or pressure for sexual favor; [and] nonverbal behavior such as suggestive looks or leering," and offers as examples "[c]ontinuous and repeated sexual slurs or sexual innuendoes; offensive and repeated risqué jokes or kidding about sex or gender-specific traits; [and] repeated unsolicited propositions for dates and/or sexual relations."[169] The College of William and Mary prohibits "conduct that is sufficiently severe, persistent or pervasive enough so as to threaten an individual or limit the ability of an individual to work, study, or participate in the activities of the College" and defines such conduct to include "making unwanted obscene, abusive or repetitive telephone calls, electronic mail, instant messages, or similar communications with intent to harass."[170]

***

In selecting an appropriate definition of hate speech, intermediaries may draw insight from longstanding First Amendment principles. Indeed, much of the speech described above in Subsections 1 through 3 can be regulated by the government under the First Amendment in certain contexts.[171] That the courts have held that such expression has limited constitutional value suggests that voluntary regulation by private intermediaries may impose comparatively few costs.[172] But as private actors, intermediaries remain unconstrained by the Constitution and are thus legally free to choose to respond to a wider universe of hate speech. The remainder of this Section briefly explores some additional possibilities.

---

pervasive or persistent so as to interfere with or limit the ability of an individual to participate in or benefit from the services, activities or privileges provided by a recipient").

[169] Letter from Jane McAuliffe, President, Bryn Mawr Coll., to William Creeley, Dir. of Legal and Pub. Advocacy, Found. for Individual Rights in Educ. (July 17, 2010), *available at* http://www.thefire.org/index.php/article/12035.html; *see also* Emory University Residence Life & Housing Standards & Policies 4-5, http://www.emory.edu/HOUSING/FORMS/form_ugrad.html (follow "download" hyperlink beside "Residence Life & Housing Policies") (last visited Apr. 8, 2011) (defining prohibited harassment to include "objectionable epithets, demeaning depiction or treatment, and threatening or actual abuse or harm").

[170] COLLEGE OF WILLIAM AND MARY, STUDENT HANDBOOK 20 n.3 (2010), *available at* http://www.wm.edu/offices/deanofstudents/services/studentconduct/documents/studenthandbook.pdf.

[171] *See supra* notes 141-142 and accompanying text (explaining that true threats and incitement can be prosecuted without running afoul of the First Amendment); *supra* note 160 (explaining that certain speech that intentionally inflicts severe emotional distress can trigger civil liability without running afoul of the First Amendment); *supra* notes 166-167 and accompanying text (explaining that verbal harassment in the workplace that is sufficiently severe or pervasive to alter the terms and conditions of employment can trigger civil liability without running afoul of the First Amendment).

[172] *See supra* notes 110-112 and accompanying text.

4. Speech that Silences Counter-Speech

Intermediaries may define hate speech as including that which silences or devalues its targets' counter-speech. Examples include slurs, insults, and epithets that shut down reasoned discourse, rather than facilitate it. In so doing, intermediaries might draw from private universities' extensive experience in regulating speech of this type, since they – like internet intermediaries – are unconstrained by the First Amendment yet for institutional reasons generally remain deeply attentive to free speech as well as antidiscrimination concerns.

Some private universities, for example, go beyond the anti-harassment requirements of Titles VI and IX in identifying a certain set of community norms to be protected from disruptive speech.[173] Such policies often emphasize a spirit of academic freedom that requires not only a commitment to free discourse, but also an understanding that certain expression can actually undermine that discourse.[174]

Colgate University, for example, articulates its commitment to intellectual inquiry and debate by prohibiting "acts of bigotry" because they "are not part of legitimate academic inquiry."[175] The University emphasizes the contextual nature of this inquiry, noting that prohibited bigotry "has occurred if a reasonable person would have found the behavior offensive and his or her living, learning, or working environment would be impaired," while reserving the right to "discipline offensive conduct that is inconsistent with community standards even if it does not rise to the level of harassment as defined by federal or state law."[176]

---

[173] As is true with virtually any proposed definition of hate speech, these efforts are not without controversy, as some argue that even private universities' efforts to address hate speech too often unwisely interfere with the unfettered flow of expression. *See, e.g.*, Azhar Majeed, *Defying the Constitution: The Rise, Persistence, and Prevalence of Campus Speech Codes*, 7 GEO. J.L. & PUB. POL'Y 481, 483-84 (2009) (criticizing public and private university efforts to regulate hate speech on campus); Nadine Strossen, *Regulating Racist Speech on Campus: A Modest Proposal*, 1990 DUKE L.J. 484, 488-89.

[174] *See, e.g.*, J. Peter Byrne, *Racial Insults and Free Speech Within the University*, 79 GEO. L.J. 399, 416 (1991) (arguing that the university is "a distinct social entity, whose commitment to enhancing the quality of speech justifies setting minimum standards for the manner of speech among its members").

[175] COLGATE UNIVERSITY STUDENT HANDBOOK 2010-2011, at 112-13 (2010), *available at* http://www.colgate.edu/portaldata/imagegallerywww/939d3f45-4876-4ef5-b567-1082dd4c58e4/ImageGallery/Student_handbook_2010.pdf.

[176] *Id.* at 115. Colgate offers the following as potential examples of impermissible harassment: "using ethnic, racial, religious or other slurs to refer to a person, or jokes or comments that demean a person" on protected bases; "creating or displaying racially, ethnically, religiously offensive pictures, symbols, cartoons, or graffiti;" and "phone calls, emails, text messages, chats or blogs that offend, demean, or intimidate another" on protected bases. *Id.*

Other proposals would similarly permit private universities to punish slurs, insults, and epithets (normally protected by the First Amendment from regulation by public actors), but would otherwise allow speech that invites a response and rational discourse. For example, Peter Byrne argues that access to free speech on campus "should be qualified by the intellectual values of academic discourse," permitting universities to bar racial insults but not "rational but offensive propositions that can be disputed by argument and evidence."[177] He argues that "[r]acial insults have no status among discourse committed to truth. They do not aim to establish, improve, or criticize any proposition."[178] Instead, racial insults simply communicate irrational hatred designed to make the target feel less worthy.[179] Along these lines, intermediaries remain free to define prohibited hate speech as that which shuts down, rather than facilitates, reasoned discourse – e.g., slurs, insults, and epithets.

5. Speech that Exacerbates Hatred or Prejudice by Defaming an Entire Group

An intermediary might choose to focus on speech that more broadly contributes to bigotry and prejudice by defaming an entire group.[180] Jeremy Waldron, for example, seeks to return to an understanding of group defamation's harms as including visible signs that group members may "be subject to abuse, defamation, humiliation, discrimination, and violence."[181] Mari Matsuda similarly characterizes Holocaust denial as a false statement of fact that defames the dead.[182] MySpace apparently adopts a definition along

---

[177] Byrne, *supra* note 174, at 400; *see also id.* at 415 ("[U]niversities do believe that racial insults are a meritless form of speech that poisons the atmosphere on campus for learning and discussion.").

[178] *Id.* at 419.

[179] *Id.*

[180] For additional proposals along these lines, see, for example, YAMAN AKDENIZ, RACISM ON THE INTERNET 7 (2009) (emphasizing virulent, inflammatory language that is likely to inspire hatred in defining hate speech); Kretzmer, *supra* note 68, at 454 (urging that responsive hate speech policy focus on "threatening, abusive or insulting" speech that "is likely in the circumstances to stir up hatred against a racial, ethnic, or national group"); Matsuda, *supra* note 67, at 2357 (defining hate speech as constituting any message of inferiority, "directed at a historically oppressed group," that is "persecutorial, hateful, and degrading").

[181] Waldron, *supra* note 91, at 1599 (arguing for hate speech regulations that promise that groups will not suffer these injuries).

[182] Matsuda, *supra* note 67, at 2366-67. Jeremy Waldron also urges that we abandon our limited understanding of actionable defamation as concerning false facts about specific individuals, and would instead include the defamation of an entire group through falsehoods. Waldron, *supra* note 91, at 1607-09. For these reasons, he reminds us of the Anti-Defamation League's founding to stop the defamation of the Jewish people because of "the danger that anti-Semitic signage would become an established feature of the landscape

these lines, prohibiting content that targets a group in a way that would make group members "feel bad."[183] Hemanshu Nigam thus describes MySpace's decision to remove Holocaust denial sites as an "easy" call under this conception of hate speech.[184] Other intermediaries take similar approaches.[185]

***

As the discussion above demonstrates, private intermediaries unconstrained by the First Amendment have a wide range of choices when defining hate speech. An intermediary's choice among them depends on a variety of unique institutional values: its assessment of the relative costs of hate speech and its constraint; empirical predictions about what sort of speech is indeed likely to lead to what sorts of harms; its business interests (which, in turn, may be shaped by users' demands and expectations); and the breadth of its sense of corporate social responsibility to address digital hate.

By identifying a spectrum of possible approaches, this Part has sought to provide a framework within which to have these conversations and to make these choices. As the next Part explores, intermediaries also have a wide range of available options when responding to hate speech that violates their chosen policy.

III. RESPONDING TO HATE SPEECH

Many intermediaries have already identified and deployed a number of responses to hate speech. In this Part, we identify promising efforts, critique others, and offer recommendations.

A. *Removing Hateful Content*

The removal of hateful content is the most powerful tool at intermediaries' disposal. Some intermediaries aggressively enforce their hate speech policies by removing offending language, blocking access to sites, or terminating user accounts. For instance, MySpace actively looks for and then deletes pages

---

and that Jews would have to lead their lives in a community whose public aspect was permanently disfigured in this way." *Id.* at 1610.

[183] Interview with Nigam, *supra* note 59.

[184] *Id.*

[185] Under the title "Don't be sexist, racist, or a hater," Digg describes its hate speech policy as: "Would you talk to your mom or neighbor like that? Digg defines hate speech as speech intended to degrade, intimidate, or incite violence or prejudicial action against members of a protected group. For instance, racist or sexist content may be considered hate speech." *Community Guidelines*, DIGG, http://about.digg.com/guidelines (last visited Apr. 8, 2011). YouTube appears to take a similar definitional approach. *YouTube Community Guidelines*, YOUTUBE, http://www.youtube.com/t/community_guidelines?gl=GB&hl=en-GB (last visited Apr. 8, 2011) ("We encourage free speech and defend everyone's right to express unpopular points of view. But we don't permit hate speech (speech which attacks or demeans a group based on race or ethnic origin, religion, disability, gender, age, veteran status, and sexual orientation/gender identity).").

and/or bans users who "promote[] or otherwise incite[] racism, bigotry, hatred or physical harm of any kind against a group or individual" or who "exploit[] people in a sexual or violent manner."[186] Other intermediaries apparently define removable hate speech more narrowly – for example, only where threats of violence are involved.[187]

Removal's enormous power counsels against its overuse, as speakers' access to certain communities can depend upon the cooperation of intermediaries. While intermediaries can prominently display websites and blogs, they can also prevent people from accessing them.[188] Thoughtful and effective responses thus do not, and should not, always require removal.

In our view, intermediaries should consider blocking forms of hate speech that satisfy certain of the narrower definitions described in Part II.B – that is, expression that is more directly related to threats of, or incitement to, violence and intentional infliction of emotional distress, and for these reasons may be

---

[186] *Terms & Conditions*, MYSPACE, http://www.myspace.com/help/terms (last visited Apr. 8, 2011); *see also* Nora Flanagan, *Social Networking: A Place for Hate?*, IMAGINE 2050 (May 19, 2009), http://imagine2050.newcomm.org/2009/05/19/social-networking-a-place-for-hate (describing MySpace's strict enforcement of its hate speech policy); Interview with Nigam, *supra* note 59. MySpace employs forum moderators who "keep an eye out for anti-semitism and derogatory comments." Michael Arrington, *MySpace Wants to Avoid this Whole Holocaust Denial Thing*, TECHCRUNCH BLOG (May 12, 2009), http://techcrunch.com/2009/05/12/myspace-wants-to-avoid-this-whole-holocaust-denial-thing/. Its terms of service explains that it "expressly reserves the right to remove your profile and/or deny, restrict, suspend, or terminate your access to all or any part of the MySpace Services if MySpace determines, in its sole discretion, that you have violated this Agreement." *Terms & Conditions*, MYSPACE, http://www.myspace.com/help/terms (last visited Apr. 8, 2011).

[187] *See supra* notes 149-151 and accompanying text (discussing Facebook's removal of pages threatening violence like *Kick a Ginger*). Assuming that Facebook understands removable hate speech to mean only that which threatens violence, it should say so more clearly in its actual hate speech policy, which instructs users not to "post content that: is hateful, threatening, or pornographic; incites violence; or contains nudity or graphic or gratuitous violence." *Statement of Rights and Responsibilites*, FACEBOOK, http://www.facebook.com/terms.php (last visited June 1, 2011).

[188] Video-sharing services and social network sites can remove content, precluding users from seeing them. Social media services can ban users by blocking their IP addresses. *Cf. Google and Internet Control in China: A Nexus Between Human Rights and Trade?: Hearing Before the Cong.-Exec. Comm'n on China*, 111th Cong. 68-76 (2010) (prepared testimony of Rebecca MacKinnon) (exploring the Chinese government's efforts to censor its citizens' online activities, including through the use of IP address blocking). Search engines can refuse to sell advertising to companies and thus limit their visibility to customers engaging in relevant searches. *See* Floyd Norris, *France Calls Google a Monopoly*, N.Y. TIMES, July 2, 2010, at B1 (describing how Google refused to sell online advertising to French company Navx, which lets French drivers know where the police operate radar traps, because "Google found Navx's business distasteful" – thus search terms like "radar trap" no longer yielded advertisements for the company's product, whose sales "plunged").

subject even to government regulation under the First Amendment.[189] Removal may be especially appropriate where counter-speech is unlikely to eliminate the harms posed by the hateful expression.

Calls for violence strike at the very heart of digital citizenship. They can inspire actual physical attacks.[190] Threats of violence also violate principles of digital citizenship even if they do not directly lead to actual violence, such as YouTube's *How to Kill a Beaner* or *Execute the Gays* videos,[191] because they deny group members the opportunity to engage in activities free from fear. In our view, Facebook and YouTube appropriately removed these and similar postings as soon as they received notice of them.[192] Threats and encouragement of violence undermine their targets' security and peace of mind, without facilitating discourse. Moreover, intermediaries generally can surgically remove threats of violence with little risk to other speech.[193]

Online hate that inflicts severe emotional distress accomplishes a similar denial of digital citizenship. For instance, recall that persistent and menacing online harassment coerced a California teenager into closing his website and leaving his school.[194] Similar results followed the attacks on Kathy Sierra: she shut down her well-known blog after anonymous posters uploaded doctored photographs, revealed her home address and Social Security number, and threatened violence.[195] This type of online hate has little chance of generating

---

[189] *See supra* notes 141-142 and accompanying text.

[190] *See, e.g.*, Nordlinger, *supra* note 11, at 8; Moore, *supra* note 11.

[191] Howard, *supra* note 7, at 4D.

[192] *Id.* (reviewing Facebook and Google/YouTube policies on removal of hate speech); Lappin, *supra* note 1 (chronicling Facebook's removal of "Kill a Jew" pages).

[193] Intermediaries also employ other strategies in addition to removal. For instance, they might accompany the removal of speech that violates their TOS with other sanctions, including warnings followed by temporary or permanent banning of individual users found in violation. *Cf.* David A. Hoffman & Salil K. Mehra, *Wikitruth Through Wikiorder*, 59 EMORY L.J. 151, 182 & n.162 (2009) (discussing Wikipedia's warning of users to stop certain behaviors and placement on probation as well as banning of users).

[194] *See supra* notes 74-79 and accompanying text.

[195] *See supra* notes 80-85 and accompanying text. Chris Locke operated the blog where the threatening comments and doctored photographs were posted. Chris Locke, *Re Kathy Sierra's Allegations*, THE EGR WEBLOG (Mar. 27, 2007, 3:16 AM), http://www.rageboy.com/2007/03/re-kathy-sierras-allegations.html. He summarized his reaction to the posts in this way:

> [T]here were a couple posts – the ones Kathy mentions in her post – that were over the top. I didn't think for a minute that they were "threatening" – and again, they were not my doing – but when I saw mail from her objecting to them, I nuked the entire site rather than censor any individual.
>
> I was a conference host on the Well 15 years ago where the core ethos was acronymized to YOYOW – You Own Your Own Words. This has remained a guiding principle for me ever since. I will not take responsibility for what someone else said, nor will I censor what another individual wrote. However, it was clear that Sierra was upset, so it seemed the best course to make the whole site go away.

counter-speech – it seems designed to cause deep distress, not to generate dialogue.

Even when content is appropriately removed, however, acknowledging its deletion can support a commitment to transparent and accountable enforcement.[196] For example, Facebook readers could see Facebook's acknowledgment that it took down *Kick a Ginger Day* and *Kill a Jew Day*.[197] And although Google's search engine does not take down hateful content in the United States as a matter of policy, it alerts web users of content removal when the law of another country requires it to block in-country users from certain sites otherwise available on the internet.[198]

B. *Countering Hate Speech with Speech*

Rather than – or in addition to – the removal of online hatred, intermediaries can counter digital hate with speech of their own. Google offers an instructive – if rare – example. In 2004, the number-one Google result for a search of "jew" was the URL jewwatch.com, a site featuring anti-Semitic content.[199] In response, a Jewish activist asked people around the Web to link the word "jew" to a Wikipedia article so that search engine users would more likely see that article at the top of search results rather than the Jew Watch site, a practice known as a "Googlebomb."[200] Neo-Nazi sites, in turn, launched a counter-

---

Id. In our view, Locke, as the blog operator and relevant intermediary, should have taken down the posts as soon as he saw them in light of their clear potential to threaten and inflict fear and distress without offering any genuine opportunity for dialogue. Locke's "You Own Your Own Words" philosophy, applied there, seems ironic given that the posters wrote anonymously and thus avoided owning their own words to avoid bearing responsibility for the threats and doctored photographs. *See id.*

[196] *See State AG Questions Research on Child-to-Child Online Bullying*, WASHINGTON INTERNET DAILY (Warren Comm'ns News, Inc., Washington, D.C.), Dec. 12, 2008.

[197] *See 'Kill a Jew Day': Spike in Virulent Anti-Jewish Facebook Pages*, THE NEW YORK BLUEPRINT (Oct. 3, 2010), http://nyblueprint.com/articles/view.aspx?id=796. Unfortunately, however, "Kill a Jew" groups continue to appear on Facebook. *See* Anomaly100, *Two Dozen 'Kill a Jew Day' Pages Found in the Last Seven Days*, FREAKOUTNATION (Oct. 3, 2010, 6:09 PM), http://freakoutnation.com/2010/10/03/two-dozen-facebook-kill-a-jew-day-pages-found-in-the-last-seven-days.

[198] Seltzer, *supra* note 23, at 46-47 ("[W]hen sites are blocked at a search-engine level, it is up to the search providers to notify their end-users. If they do not, the page disappears invisibly. In most engines, pages simply disappear from listings, leaving searchers unaware that a site they never saw is gone. . . . Among the major search engines, only Google gives indication when it removes results from a search page because of legal demands."). Of course, Google's current policy might stem from its objection to those countries' restrictive laws. Whatever the reasons underlying its policy, we still laud the company for the transparency of those decisions to remove content.

[199] *See* JEW WATCH, http:/jewwatch.com (last visited Apr. 8, 2011). *See generally Jew Watch*, WIKIPEDIA, http://en.wikipedia.org/wiki/Jew_Watch (last visited Apr. 8, 2011).

[200] John Brandon, *Dropping the Bomb on Google*, WIRED (May 11, 2004), http://www.

Googlebomb, leading the results back to Jew Watch.[201] Individuals asked Google to remove Jew Watch entirely from its search results.[202]

After the story drew significant media and interest-group attention, Google announced that it would not change its software to eliminate Jew Watch in its results pages.[203] It explained that it chose not to change its algorithms because it "views the comprehensiveness of [its] search results as an extremely important priority," and it does not "remove a page from [its] search results simply because its content is unpopular or because we receive complaints concerning it."[204]

Instead, Google inserted its own advertisement entitled "Offensive Search Results" on top of its page where the link to Jew Watch appeared among other search results.[205] Google explained the company's understanding that the Jew Watch site may be offensive and "apologize[d] for the upsetting nature of the experience you had using Google."[206] Google assured readers that it did not

---

wired.com/culture/lifestyle/news/2004/05/63380. "Googlebombing" refers to a practice in which users can artificially inflate a page's search ranking by linking to a page in as many other pages as possible. James Grimmelmann, *The Google Dilemma*, 53 N.Y.U. SCH. L. REV. 939, 942-43 (2008-09); *see* Bracha & Pasquale, *supra* note 128, at 1167-88 (describing search engines' capacity to manipulate their results).

[201] Grimmelmann, *supra* note 200, at 943.

[202] *Id.*

[203] *An Explanation of Our Search Results*, GOOGLE, http://www.google.com/explanation.html (last visited Apr. 8, 2011).

[204] *Id.* Apparently, however, Google does change search results for at least some purposes. Consider the example of a merchant who deliberately engaged in bad behavior because the sheer volume of negative mentions that then appeared on consumer advocacy websites improved his ranking in search results. David Segal, *A Bully Finds a Pulpit on the Web*, N.Y. TIMES, Nov. 28, 2010, at BU1. In response, Google changed its algorithm to penalize sites that others link because it provided "extremely poor user experience." Amit Singhal, *Being Bad to Your Customers Is Bad for Business*, THE OFFICIAL GOOGLE BLOG (Dec. 1, 2010, 12:06 PM), http://googleblog.blogspot.com/2010/12/being-bad-to-your-customers-is-bad-for.html. In a blog posting, Google explained that it developed an algorithmic solution to ensure that "being bad is, and hopefully will always be, bad for business in Google's search results." *Id.* For another example, see David Segal, *The Dirty Little Secrets of Search*, N.Y. TIMES, Feb. 13, 2011, at BU1 (discussing Google's changes in search results to counter the effects of manipulative efforts to maximize J.C. Penney's search result rankings).

[205] Google Search for "Jew", GOOGLE, http://www.google.com/search?q=jew (last visited Mar. 26, 2011).

[206] *An Explanation of Our Search Results*, *supra* note 203. If, however, you type "jew" into Google's German version, google.de, Jew Watch does not appear at all. Grimmelmann, *supra* note 201, at 948. At the bottom of the results page, a notice explains that Google has removed three results from the page. *Id.* Google changed its results because German law criminalizes incitement of hatred against minorities. *Id.* at 947. For a discussion of whether and how countries that have experienced genocide may take more aggressive approaches to hate speech, see Jennifer M. Allen & George H. Norris, *Is Genocide Different? Dealing*

endorse the views expressed by Jew Watch.[207] Google's explanation added that readers "may be interested in some additional information the Anti-Defamation League has posted about this issue."[208] To date, however, Jew Watch continues to appear prominently in a Google search of "jew."[209]

Google similarly inserted an explanatory advertisement after images of the First Lady, altered to resemble a monkey, prominently appeared among the results of Google image searches for "Michelle Obama."[210] After Google posted its advertisement, a Chinese blog that had recently featured the image took it down, saying, "I am very sorry for this article."[211]

This kind of intermediary counter-speech is, however, far from routine. For instance, although Google has a "Report Offensive Image" function, it rarely responds to such reports, and Google has to date bought "Offensive Search Results" advertisements in only the cases discussed here.[212] Such counter-speech by intermediaries thus remains extremely rare.[213]

---

with Hate Speech in a Post-Genocide Society, 7 J. INT'L L. & INT'L REL. (forthcoming 2011), *available at* http://ssrn.com/abstract=1640812.

[207] *An Explanation of Our Search Results*, *supra* note 203.

[208] *Id.*

[209] Google Search for "Jew", *supra* note 205 (showing Jew Watch as the second result).

[210] Saeed Ahmed, *Google Apologizes for Results of 'Michelle Obama' Image Search*, CNN (Nov. 25, 2009, 12:05 PM), http://www.cnn.com/2009/TECH/11/25/google.michelle.obama.controversy-2/index.html. A Google forum user flagged the picture. *Id.* Initially, Google de-indexed the website that posted the photograph on the grounds that "it could spread a malware virus." *Id.*

[211] Bianca Bosker, *Michelle Obama Pictures UPDATE: Offensive Image REMOVED, Google 'SORRY'*, HUFFINGTON POST (updated Mar. 18, 2010, 5:12 AM), http://www.huffingtonpost.com/2009/11/24/michelle-obama-photo-goog_n_368760.html.

[212] *See* Barry Schwartz, *Report Offensive Images on Google Does Not Work*, SEARCH ENGINE ROUNDTABLE (Apr. 13, 2010, 7:54 AM), http://www.seroundtable.com/archives/022010.html. Google's inaction on other cases has sparked much criticism. *See, e.g.*, Esra'a Al Shafei, *Google Apologizes for Offending Jews Through Search Results*, MIDEAST YOUTH (Mar. 21, 2007), http://www.mideastyouth.com/ 2007/03/21/google-apologizes-for-offending-jews-through-search-results.

[213] Google has bought ads in at least one other instance outside of the context of hate speech: a Google user searching for "suicide" will encounter Google ads featuring suicide prevention resources. Noam Cohen, *'Suicide' Query Prompts Google to Offer Hotline*, N.Y. TIMES, Apr. 5, 2010, at B6. The "icon of a red phone and the toll-free number for the National Suicide Prevention Hotline" appear over the linked results in a way that is "different and more prominent than an advertisement." *Id.* Google has also provided the telephone number for national poison control in searches like "poison emergency." *Id.* MySpace has gone further than putting up advertisements when users write about suicide. Interview with Nigam, *supra* note 59. As Hemanshu Nigam explained, when MySpace identified, or received notice of, users noting a desire to commit suicide, it would contact the National Suicide Prevention hotline and local police to recruit help for the users. *Id.* According to Nigam, MySpace's intervention helped prevent ninety-three suicides in 2009. *Id.*

To be sure, we recognize – and remain concerned by – the possibility that counter-speech may shine a spotlight on, and thus bring more attention to, digital hate. But silence in response to digital hate carries significant expressive costs as well. When powerful intermediaries rebut demeaning stereotypes (like the Michelle Obama image) and invidious falsehoods (such as Holocaust denial), they send a powerful message to readers. Because intermediaries often enjoy respect and a sense of legitimacy, users may be inclined to pay attention to their views.[214] With counter-speech, intermediaries can demonstrate by example what it means to treat others with respect and dignity.

Moreover, such counter-speech can expose digital citizens to diverse views, piercing the insularity of hateful messages that may lead to more extreme views. This is just the sort of strategy Cass Sunstein alludes to in his book, *Republic.com 2.0*, where he calls for "self-conscious efforts by private institutions" to expose citizens to diverging views.[215] He urges intermediaries to adopt best practices that expose citizens to different perspectives on public issues, such as through "creative use of links to draw people's attention to multiple views."[216]

By challenging hate speech with counter-speech, intermediaries can help transform online dialogue by documenting the continuing existence of racism and other forms of hatred while concomitantly rebutting it. In this way, intermediary action may help develop the qualities of tolerance advocated by Lee Bollinger,[217] while repairing the public discourse by speaking for silenced or devalued targets. Intermediaries could play a valuable role in challenging hate without defusing the safety valve and other attributes of permitting the

---

[214] For additional arguments of the value of counter-speech by powerful speakers in response to hate, see Corey Brettschneider, *When the State Speaks, What Should It Say? The Dilemmas of Freedom of Expression and Democratic Persuasion*, 8 PERSP. ON POL. 1005, 1005 (2010) (urging that "a proper theory of the freedom of expression obligates the legitimate state" to respond to hateful but protected speech by emphasizing the importance of respect for equality and dignity); Helen Norton, *Campaign Speech Law with a Twist: When Government Is the Speaker, Not the Regulator*, 61 EMORY L.J. (forthcoming 2011) (urging government to engage in political speech on contested ballot measures that counters that of powerful private speakers); Charlotte H. Taylor, *Hate Speech and Government Speech*, 12 U. PA. J. CONST. L. 1115, 1188 (2010) (urging government – generally prohibited by the First Amendment from banning hate speech – to engage in counter-speech to "help forge consensus about the nature of social practices" and "shift the ground under the hateful speaker's feet, robbing her of her confidence that she can invoke an entire system of subordination by using a few cheap words").

[215] SUNSTEIN, *supra* note 100, at 191.

[216] *Id.* at 192, 200-01, 208 (calling for radio stations, television stations, and newspapers to provide links to diverse views on their online sites).

[217] BOLLINGER, *supra* note 107, at 172-73 (arguing that tolerating the expression of hatred may actually enhance our intellectual capacities and embolden civic courage).

expression of hateful views.[218] Importantly, intermediaries that respond to hate speech through forceful counter-speech or in some other way short of removal appear to trigger few, if any, of the expressive concerns about intermediaries' voluntary measures identified above.[219]

In some respects, Facebook's response to Holocaust denial groups illustrates a missed opportunity for meaningful counter-speech. Facebook vigorously defended its refusal to take down the sites on the grounds that such refusal allows people to see that the sites' proponents are "stupid."[220] Facebook, however, could have explained to its users through counter-speech *why* it views those sites as "stupid." Many other instances of hate – from demeaning characterizations of groups[221] and individuals[222] to falsehoods meant to inspire hate[223] – similarly invite intermediaries' counter-speech.

To be sure, an intermediary's ability to respond to cyber hate will inevitably depend on available resources. Indeed, cyber hate's exponential growth could overwhelm intermediaries interested in engaging in counter-speech. Nonetheless, the ability to automate functions like searching for key terms and inserting prepared responses may help cut down on costs.[224]

Given limited resources, intermediaries might attend carefully to hate speech targeted at children given electronic media's profound impact on children's behavior and views.[225] Indeed, some hate sites are designed specifically to influence youths.[226] *MartinLutherKing.org*, a hate site, is

---

[218] *See, e.g.*, Blasi, *supra* note 107, at 408 (highlighting Bollinger's recognition "of the safety-valve function of letting discontent surface"); Hughes, *supra* note 107, at 365 (suggesting that hate speech regulation creates martyrs and converts to the cause of hatred); Lidsky, *supra* note 107, at 1099-1100 (concluding that punishing Holocaust denial will paradoxically entrench that view and inspire stronger belief in conspiracy theories).

[219] *See supra* notes 110-124 and accompanying text.

[220] Michael Arrington, *Facebook Remains Stubbornly Proud of Position on Holocaust Denial*, TECHCRUNCH (May 12, 2009), http://techcrunch.com/2009/05/12/facebook-remains-stubbornly-proud-of-position-on-holocaust-denial.

[221] *See, e.g.*, *Common Pro-N----- Arguments*, N-----MANIA, http://niggermania.com/tom/niggerarguments/niggerargumentstextpagetwo.htm (last visited Mar. 27, 2011) ("We hate n[-----]s" because they are a "failed ape species.").

[222] *See, e.g.*, *The Beast as Saint: The Truth About "Martin Luther King, Jr."*, http://www.martinlutherking.org/thebeast.html (last visited Apr. 8, 2011) (arguing that Dr. Martin Luther King, Jr. was an academic cheat, communist, and sex addict).

[223] *See, e.g.*, JEW WATCH, *supra* note 199.

[224] Those costs would be comparatively minor in instances where an intermediary can automate counter-speech. Although an intermediary would need to incur the fixed cost of designing, or purchasing, responsive software, it would incur virtually no expenditures for the software's implementation in future cases. *Cf.* Danielle Keats Citron, *Technological Due Process*, 85 WASH. U. L. REV. 1249, 1284-85 (2008).

[225] *See* Ybarra et al., *supra* note 96, at 929.

[226] LORRAINE TIVEN & PARTNERS AGAINST HATE, HATE ON THE INTERNET: A RESPONSE GUIDE FOR EDUCATORS AND FAMILIES 20, 21 (2003) ("[J]ust as fashion editors and e-book

directed at students researching the civil rights movement.[227] Neo-Nazi adherent Vincent Breeding (credited by Partners Against Hate as creating and maintaining the site) writes: "If you are a teacher or student, I hope you will take a stand for right and wrong and use this information to enlighten your peers."[228] The Klan has Youth News video games available online; some are ethnic cleansing games.[229] Hateful games aimed at young people have seeped into the mainstream where they are either hosted or reviewed on regular gaming sites.[230] Because children are particularly vulnerable to influence, intermediaries might thus be quicker to challenge hate speech that targets them.[231] Hate speech can teach children that prejudice is socially acceptable. Hate speech that condones violence against group members, notably ethnic cleansing games,[232] is especially troubling because video games that engage in fantasies about killing group members can desensitize children to violence and promote violent behavior.[233] Counter-speech – and, indeed, sometimes the removal of such speech altogether – is thus especially important with respect to hate speech that targets children.

C. *Educating and Empowering Community Users*

Just as we see with other mediating institutions like schools, workplaces, and churches, intermediaries can help develop an understanding that citizenship – here, digital citizenship – should include attention to the dignity and safety of other users. Educators, supervisors, and pastors have long played this sort of role with regard to bullying – they endeavor to teach children and adults alike how to treat others with respect.[234] Intermediaries can play a similar role with regard to online hatred.

---

publishers have started reaching out to elementary school children and teens . . . so have hate groups." (quoting Tara McKelvey, *Father and Son Target Kids in a Confederacy of Hate*, USA TODAY, Jul. 16, 2001, at 3D)).

[227] *Id.* at 20.

[228] *Id.*

[229] SIMON WIESENTHAL CENTER, *supra* note 9 (displaying screenshots of video games based in hate speech)

[230] *Id.*

[231] *See* Shiffrin, *supra* note 74, at 89 ("As an African American father once said to me when I spoke about the contribution of racist speech to the democratic dialogue, 'Tell that to my seven-year-old daughter.'").

[232] *See supra* notes 9-10 and accompanying text (discussing video games posted on YouTube and neo-Nazi social network sites).

[233] Social science research demonstrates the significant harm caused by exposing children to violence. *See, e.g.*, Amitai Etzioni, *On Protecting Children from Speech*, 79 CHI.-KENT L. REV. 3, 36-37 (2004).

[234] *See, e.g.*, Susan Engel & Marlene Sandstrom, *There's Only One Way to Stop a Bully*, N.Y. TIMES, July 23, 2010, at A23.

1. Education

Intermediaries' educational efforts can take a variety of forms. For example, intermediaries can valuably educate their users about digital citizenship norms by more transparently explaining their enforcement decisions. They can offer examples of instances when they did, and did not, remove contested content, along with their reasoning. Intermediaries with similar priorities could join forces in drafting a set of principles and explanatory examples.[235] Just as the preceding Part urged greater transparency and specificity when identifying the harms to be targeted – and thus the objectives to be achieved – by a particular hate speech policy and definition, this Part highlights the value of greater transparency when explaining the reasons behind certain decisions enforcing these policies.

For example, Facebook, MySpace, YouTube, AOL, and other intermediaries currently devote significant staff and energy addressing abuse complaints.[236] Yet their actual practices – that is, what decisions they actually make and how – remain unclear.[237] As part of a commitment to transparent policy implementation, they could explain the grounds of certain decisions, including the definition of hate speech that they employed and specific examples of the harms that they sought to forestall in rendering those decisions. The more clearly and specifically that intermediaries identify and explain their approach to hate speech, the more informed users' choices will then be about the sort of online community with which they choose to interact. The Beliefnet policy discussed in Part II provides a helpful illustration.[238]

Intermediaries can also engage in efforts to educate the public more broadly about hate. For instance, YouTube's Safety & Security Center features information and links to resources developed by the Anti-Defamation League (ADL) to help internet users respond to and report offensive material and

---

[235] This recalls the international standards organization for the World Wide Web – the W3C group – that identifies voluntary standards. *See W3C Mission*, WORLD WIDE WEB CONSORTIUM, http://www.w3.org/Consortium/mission (last visited Apr. 8, 2011). We thank Neil Richards, Berin Szoka, and Chris Wolf for their helpful thoughts on this notion.

[236] *See, e.g.*, *Fourth Law and Information Society Symposium: Hate Versus Democracy on the Internet*, FORDHAM LAW EVENT CALENDAR (Mar. 26, 2010), http://law2.fordham.edu/ihtml/cal-2uwcp-calendar_viewitem.ihtml?idc=10320.

[237] Indeed, clearer and more transparent policies might have averted the situation where Facebook pulled Sarah Palin's controversial-but-not-hateful posting about proposals to build a mosque near Ground Zero after a number of users responded to a campaign encouraging them to click the "Report Note" hyperlink indicating the posting as hate speech. When Palin questioned the action, Facebook put it back up, apologized for pulling the comment, and promised to modify their process for taking down postings. *See* Brian Ries, *My Facebook War with Palin*, THE DAILY BEAST (JULY 23, 2010, 11:10 AM), http://www.thedailybeast.com/blogs-and-stories/2010-07-23/palins-facebook-ground-zero-mosque-post-how-it-disappeared/full/. Transparent policies might thus have additional salutary effects: the prevention of user manipulation of intermediaries' reporting tools.

[238] *See supra* notes 154-159 and accompanying text.

extremist content that violates YouTube's Community Guidelines on hate speech.[239] It includes tips from the ADL on how to confront hate speech, including flagging offensive videos for review by the YouTube team, posting videos or comments that oppose the offensive point of view, and talking to friends, family, and teachers about what they have seen.[240] As one more of the many ways that intermediaries might help educate users about the impact and treatment of cyber hatred, intermediaries might also consider funding cyber literacy campaigns to teach students about digital citizenship.

2. Empowerment

Empowering users to respond to hate speech on their own sites and to report Terms of Service violations can help communicate and enforce community norms and expectations of digital citizenship.[241] As Clay Shirky observes:

> Any group trying to create real value must police itself to ensure it isn't losing sight of its higher purpose . . . . Governance in such groups is not just a set of principles and goals, but of principles and goals that have been internalized by the participants. Such self-governance helps us behave according to our better natures.[242]

Note, however, that such efforts are most likely to be effective when intermediaries have educated their users and enforcement personnel about the specific harms to be addressed by their specific hate speech policy.

How can an intermediary help its users internalize norms of digital citizenship? As Shirky explains, communities that permit "mutually visible action among the participants, credible commitment to shared goals, and group members' ability to punish infractions" create contexts in which users "can do a better job both in managing the resource and in policing infractions than can markets or government systems designed to accomplish the same goals."[243]

---

[239] *Safety Center: Hateful Content*, YOUTUBE, http://www.google.com/support/youtube/bin/answer.py?hl=en&answer=126264 (last visited Apr. 8, 2011).

[240] *Id.*

[241] *See, e.g.*, Jon M. Garon, *Wiki Authorship, Social Media, and the Curatorial Advantage*, 1 HARV. J. SPORTS & ENT. L. 95, 99 (2010) ("By expanding opportunity for interaction and fostering behavioral norms of trust among users, these communications tools can expand the reach of social networks for mutual advantage.").

[242] CLAY SHIRKY, COGNITIVE SURPLUS: CREATIVITY AND GENEROSITY IN A CONNECTED AGE 165 (2010); *see also id.* at 177 ("Unlike personal or communal value, public value requires not just new opportunities for old motivations; it requires governance, which is to say ways of discouraging or preventing people from wrecking either the process or the product of the group.").

[243] *Id.* at 113; *see also* ELINOR OSTROM, GOVERNING THE COMMONS: THE EVOLUTION OF INSTITUTIONS FOR COLLECTIVE ACTION 88-102 (1990) (identifying the factors key to community regulation of common resources to include institutions well-equipped to gather information about the resource, forums to discuss its management, community participation in developing and enforcing the rules, and appropriate and graduated sanctions to discipline

Intermediaries will likely have greater success setting norms if they contain code designed to foster social governance, such as reputation scoring systems.[244]

The Wikipedia experience provides a powerful example of such dynamics in action to foster effective online norms of good behavior. As Jonathan Zittrain explains, Wikipedia's key distinguishing attributes – and one that may explain much of its success – included its initial core of editors who shared a "common ethos" and then shared those behavioral norms with new users "through informal apprenticeships as they edited articles together."[245] These norms include administrators' power to create locks to prevent misbehaving users from editing and to ensure that articles prone to vandalism are not subject to changes by unregistered or recently registered users.[246] Users acquire such administrative powers "by making lots of edits and then applying for an administratorship" – that is, by demonstrating their compliance with community norms.[247]

Moreover, Wikipedia enlists volunteer editors called "Third Opinion Wikipedians" who resolve disputes between editors.[248] As David Hoffman and Salil Mehra document, Wikipedia's guidelines urge Third Opinion Wikipedians to "read the arguments, avoid reckless opinions, be civil and nonjudgmental, offer neutral opinions, and monitor the page after offering an opinion."[249] Wikipedia also permits users to report impolite, uncivil, or other

---

abuse).

[244] Amazon, eBay, Craigslist, and other commercial sites permit users to rate other users or to flag potential misbehavior. Kahn, *supra* note 25, at 198-201 (describing Wikipedia and eBay's use of community norms to police users' behavior); Kim, *supra* note 18, at 1016. Daniel Kahn similarly observes that sites like Wikipedia and eBay that successfully rely on community norms to encourage good behavior share a few key characteristics: "the sites provide easy methods for users to view each others' reputational information"; "reputational information is reciprocal: those who wish to comment on others' behavior must also open themselves to being rated"; "the sites do not merely expect norms to emerge in a vacuum, but instead contain code designed to help foster social governance"; and "they give users incentives to opt into the norm system and to take it seriously." Kahn, *supra* note 25, at 202-03. In response to Neil Netanel's assertion that the internet is not the sort of environment in which norms can generally shape behavior, Netanel, *supra* note 27, at 432, Kahn replies that "the Web is no longer simply too big to handle norms" because social intermediaries enable the formation of smaller communities of manageable size. Kahn, *supra* note 25, at 235.

[245] JONATHAN ZITTRAIN, THE FUTURE OF THE INTERNET – AND HOW TO STOP IT 134-35 (2008). These norms include "the three-revert rule," in which "an editor should not undo someone else's edits to an article more than three times in one day." *Id.* at 135.

[246] *Id.*

[247] *Id.* at 135-36.

[248] Hoffman & Mehra, *supra* note 193, at 172-73.

[249] *Id.* (explaining that Third Opinions are provided under separate headings from the original disputes). Wikipedia has an Arbitration Committee, whose elected members adjudicate disputes between users. *Id.* at 154. Hoffman and Mehra explain that while the

difficult communications with editors in its Wikiquette alerts notice board.[250] On the non-binding Wikiquette alerts page, users seek advice, informal mediation, or referrals to a more appropriate forum.[251] The Wikiquette alerts page also explicitly asks those who have benefited from the process to contribute to other alerts.[252] The Wikipedia model may prove helpful to intermediaries when devising systems for responding to user's abuse reports about cyber hate.

Intermediaries could rely on users to help them identify and respond to hateful content. Currently, many intermediaries depend upon users to report prohibited content, which their employees then address. YouTube's global communications director Scott Rubin has explained that the company cannot "prescreen" content because "[t]here are 20 hours of video uploaded to our site every minute." [253] According to Rubin, YouTube counts on its community to "know the guidelines and to flag videos that they believe violate guidelines."[254] YouTube also offers to its users a Safety Mode tool that blocks videos with objectionable material and encourages users to address hate speech appearing on their own profiles.[255] It reminds users that they can remove others hateful comments from their videos and moderate comments on their channels.[256]

A few intermediaries even allow users to make initial decisions about whether material ought to appear online. For example, Mozilla, the developer of the web browser Firefox, allows users to personalize their browser with

---

Arbitration Committee generates norms, its task is to rule on specific cases and set forth concrete rules on how users should behave. *Id.* The Arbitration Committee has sanctioned users who make "homophobic, ethnic, racial or gendered attacks" or who are stalkers and harassers. *Id.* at 180. The Arbitration Committee can ban individuals from participation on all or part of the site or place them on probation. *Id.* at 182. Generally speaking, there is a 63% chance that the arbitrators caution the parties or impose probations, and a 16% chance that they will ban a party from the site. *Id.* at 184. In cases when either impersonation or anti-social conduct like hate speech occurs, the Administrative Committee will ban the user in 21% of cases. *Id.* at 189. Wikipedia's more than 1500 administrators, in turn, enforce those rules. *Id.* at 174. The Arbitration Committee publishes its final decisions. *Id.* at 177.

[250] *Id.* at 173.

[251] *Id.*

[252] *Id.*

[253] Howard, *supra* note 7, at 4D. Facebook similarly urges users to provide the company's team of professional reviewers with "accurate and detailed information" so that "you can help us locate and remove abuse on the site as quickly and efficiently as possible." Jessica Ghastin, *Responding to Abuse Reports More Effectively*, THE FACEBOOK BLOG (Oct. 14, 2009, 10:43 AM), http://blog.facebook. com/blog.php?post=144628037130.

[254] Howard, *supra* note 7, at 4D.

[255] Dan Raywood, *YouTube Safety Mode Introduced to Block Inappropriate Content*, SC MAG. (U.K.) (Feb. 15, 2010), http://www.scmagazineuk.com/youtube-safety-mode-introduced-to-block-inappropriate-content-but-claims-made-that-it-will-only-have-a-minor-impact/article/163784/; *Safety Center: Hateful Content*, *supra* note 239.

[256] *Safety Center: Hateful Content*, *supra* note 239.

artwork using an application called Personas.[257] Mozilla lets community members review users' Persona requests; once approved, the user's artwork is available for others to adopt.[258] Mozilla provides guidelines on artwork's potentially offensive and hateful content to its community members to assist them in their review of applications.[259] Mozilla, however, retains the ability to oversee the community members' decisions, especially when users contest those decisions.[260]

Intermediaries might also empower users in other ways: those who dispute hateful distortions might be provided a space to present their case and discuss it.[261] Google's news service, for example, has taken steps in this direction by permitting the subjects of news articles to reply to stories that include their name.[262] Along similar lines, search engines could offer discounted advertisement rates for counter-speakers targeted by digital hate, who could use that advertising space to directly respond to hate speech generated by a search engine's results. Just as Google itself placed "Offensive Search Results" ads,[263] it could provide discounted rates for other groups to do the same. A group like the NAACP could inexpensively purchase ads providing links to counter-speech about Dr. Martin Luther King in searches of his name to ensure that readers see their link alongside links to the neo-Nazi website *MartinLutherKing.org*.[264] Google could also award free online advertising to targeted groups as it does for certain charitable organizations.[265]

---

[257] *How to Create Your Own Persona*, MOZILLA, http://www.getpersonas.com/en-US/demo_create (last visited Apr. 8, 2011). Personas is a feature in the Firefox browser that allows a user to select simple-to-use themes, known as Personas, to personalize their browser and status bar. *Personas for Firefox*, WIKIPEDIA, http://en.wikipedia.org/wiki/Personas_for_Firefox (last visited Apr. 8, 2011). Over 220,000 Personas are available for users to choose from on the *GetPersonas.com* website. *Id.*

[258] E-mail from Julie Martin, Assoc. Gen. Counsel, Mozilla, to Danielle Citron, Professor of Law, Univ. of Maryland School of Law (Aug. 11, 2010) (on file with author).

[259] *Id.*

[260] *Id.*

[261] *See* Pasquale, *supra* note 138, at 62.

[262] ALEXANDER HALAVAIS, SEARCH ENGINE SOCIETY 136 (2009). Google News Service provides news to users, a service that is separate from its work as a search engine. With respect to its search engine services, Google has also considered "expos[ing] user reviews and ratings for various merchants alongside their results" to address the problem of high rankings for merchants with "extremely poor user experience." Singhai, *supra* note 204. It ultimately rejected that course of action because it "would not demote poor quality merchants in our results and could still lead users to their websites." *Id.*

[263] *See supra* notes 205-208 and accompanying text.

[264] *See supra* notes 227-228 and accompanying text (discussing racist website *MartinLutherKing.org* aimed at children researching the civil rights leader).

[265] *In Kind Advertising for Non-profit Organizations*, GOOGLE GRANTS, http://www.google.com/grants/ (last visited Apr. 8, 2011) (explaining its "unique in-kind donation program awarding free AdWords advertising to select charitable organizations" that share

3. Architectural Choices

Intermediaries can also help encourage the development of digital citizenship norms through architectural choices.[266] As Jaron Lanier reminds us, the web's anonymity – often extolled as an irreplaceable virtue – was neither an inevitable feature of net design,[267] nor necessarily a salutary one. Indeed, the internet's great communicative strengths – e.g., its ability to aggregate large numbers of speakers as well as disaggregate speakers' offline identities from their online voices – also magnify its capacity to empower certain socially destructive behaviors.[268] Anonymity is thus valuable when it enables speakers to avoid retaliation,[269] but not when it simply enables speakers to avoid responsibility for destructive behavior. For this reason, Lanier urges users: "Don't post anonymously unless you really might be in danger."[270]

Private intermediaries can play an important role in shaping these norms by discouraging anonymity in appropriate circumstances. For example, intermediaries might permit anonymity as a default matter, revoking it when users violate TOS agreements or Community Guidelines.[271] Or they might instead follow Facebook's lead.[272] Facebook requires every user to register under his or her real name and to provide an email address to assist Facebook in verifying his or her identity.[273] On Facebook, "there would be no pseudonymous role-playing, as in so many online social networks."[274] Facebook's philosophy is one of "radical transparency," which its founder

---

its "philosophy of community service to help the world in areas such as science and technology").

[266] *See* Lessig, *The New Chicago School*, *supra* note 27, at 662-63 (explaining that institutions can shape others' behavior through the development of social norms, as well as through law, markets, and architecture).

[267] JARON LANIER, YOU ARE NOT A GADGET 6 (2010) (as originally introduced, the web "emphasized responsibility, because only the owner of a website was able to make sure that their site was available to be visited").

[268] Citron, *Cyber Civil Rights*, *supra* note 20, at 63-65; *see also* Smith, *supra* note 86, at 59-60 (explaining how unique features of the internet exacerbate its power to spread hate).

[269] *See, e.g.*, Amy J. Schmitz, *"Drive-Thru" Arbitration in the Digital Age: Empowering Consumers Through Binding ODR*, 62 BAYLOR L. REV. 178, 202-04 (2010) (discussing how consumers may be more likely to challenge corporate misbehavior through online vehicles that offer some measure of anonymity and thus protection from retaliation).

[270] LANIER, *supra* note 267, at 21.

[271] Intermediaries might accomplish this strategy by requiring users to register with intermediaries, e.g., requiring credit card information or email address. We thank Julie Cohen for this insightful suggestion.

[272] DAVID KIRKPATRICK, THE FACEBOOK EFFECT: THE INSIDER STORY OF THE COMPANY THAT IS CONNECTING THE WORLD 13 (2010).

[273] Richard A. Posner, *Just Friends*, NEW REPUBLIC, Aug. 12, 2010, at 27 (reviewing KIRKPATRICK, *supra* note 272).

[274] *Id.*

Mark Zuckerberg believes will help make people more tolerant of each other's eccentricities.[275]

Facebook justifies its comparatively hands-off approach to hate speech[276] partly because it does not permit truly anonymous speech.[277] A Facebook employee asked: "Would we rather Holocaust denial was discussed behind closed doors or quietly propagated by anonymous sources? Or would we rather it was discussed in the open on Facebook where people's real names and their photo is associated with it for their friends, peers, and colleagues to see?"[278] Although, as discussed above, we think Facebook and other intermediaries in this context have missed valuable opportunities to engage in counter-speech, we urge more intermediaries to make architectural choices that discourage speakers from refusing to take responsibility themselves for their own hateful expression.

Although focusing on website operators rather than on intermediaries, Nancy Kim has similarly urged architectural designs that default to identified rather than anonymous postings, thus challenging the assumption that all postings should be afforded equal weight.[279] Along these lines, some newspapers and games have moved away from anonymous comments on their online versions.[280] Kim also offers thoughtful suggestions on how website

---

[275] *Id.*

[276] *See supra* notes 149-151, 220 and accompanying text (explaining that Facebook deems threats of violence to groups as prohibited hate speech worthy of removal and refuses to recognize Holocaust denial as prohibited hate speech).

[277] As Richard Posner explains, Facebook requires every user to register under his real name and to provide an email address to assist Facebook in verifying his identity. Posner, *supra* note 273, at 27.

[278] Chris Matyszczyk, *Facebook: Holocaust Denial Repulsive and Ignorant*, CNET NEWS BLOG (May 6, 2009, 1:04 PM), http://news.cnet.com/8301-17852_3-10234760-71.html.

[279] Kim, *supra* note 18, at 1016-17; *see also id.* at 1017 ("The point is not to make identified postings mandatory, but to make identified postings easier than slightly more burdensome anonymous postings.").

[280] *See, e.g.*, Levy, *supra* note 118 (explaining that because the New York Times, but not the Washington Post, devotes staff time to moderating comments before they appear on their blogs, "comments at the Times tend to be much more thoughtful – and hence worth reading – while comments on the Post's political blogs tend to be much more partisan and much more full of rant"); Roy Greenslade, *Paper Puts Up a Paywall for Comments*, GREENSLADE BLOG (July 13, 2010, 17:23), http://www.guardian.co.uk/media/greenslade/2010/jul/13/paywalls-us-press-publishing; Stephanie Goldberg, *News Sites Reining in Nasty User Comments*, CNN (July 19, 2000), http://articles.cnn.com/2010-07-19/tech/commenting.on.news.sites_1_comments-news-sites-credit-card?_s=PM:TECH (discussing news websites like the Huffington Post that require registration or real names as a condition of commenting); Richard Pérez-Peña, *News Sites Rethink Anonymous Online Comments*, N.Y. TIMES, Apr. 11, 2010, http://www.nytimes.com/2010/04/12/technology/12comments.html (discussing news websites that are considering real names or otherwise regulating comments to their online news content).

sponsors might design their systems to "slow down the posting process, encouraging posters to more carefully consider what they say before they press 'Send'" – for example, by requiring a waiting or cooling-off period before the post is published, during which the poster may choose to edit or remove the message.[281] These are just a few examples: the possibilities for community education, empowerment, and encouragement are substantial, especially as emerging technologies facilitate even more interactivity online. [282]

CONCLUSION

Troubled by the considerable harms posed by digital hate to civic engagement and thus digital citizenship, we nonetheless recognize the considerable legal and political barriers to governmental responses. For this reason here we leverage the interest and commitment to addressing digital hate already expressed by a number of intermediaries to explore promising alternatives, while noting users' potential role in shaping that interest and commitment through consumer demand.

To this end, we suggest that interested intermediaries can valuably advance the fight against digital hate with increased transparency – e.g., by ensuring that their voluntary efforts to define and proscribe hate speech explicitly turn on the harms to be targeted and prevented. We also urge them to consider the range of available responses to hateful speech that include not only removal, but also engaging in or facilitating counter-speech, as well as educating and empowering users with respect to digital citizenship. We remain optimistic that a thoughtful intermediary-based approach to hate speech can significantly contribute to norms of equality and respect within online discourse without sacrificing expression.

---

[281] Kim, *supra* note 18, at 1017.

[282] KLEIN, *supra* note 66, at 191-92 ("[T]he net generation must equip themselves with the new awareness that many websites *are not* what they appear to be . . . . In addition to asking questions and promoting awareness about the nature of media and information in cyberspace, a socially responsible net generation must acquire a mature understanding about the sinister elements that purvey that world, and where they lead.").