The Free Speech Balancing Act of Digital Intermediaries

An explication of the concept of content governance

A Dissertation
SUBMITTED TO THE FACULTY OF
UNIVERSITY OF MINNESOTA
BY

Brett Gregory Johnson

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

Amy Kristin Sanders, Adviser

May 2015

Acknowledgements

On the second day of my first year at the University of Iowa in August 2002, I knew I wanted to be a college professor. A 50-minute lecture from Reza Aslan—then the instructor for the course "Introduction to Islam," and now an internationally renowned expert on Islamic politics—was all it took. Professor Aslan was no professor. He was a storyteller. His lectures didn't convey information. They painted pictures. Simply put, I wanted to be Professor Aslan one day. Or at least come close. I don't know if I'll ever reach that level of awesomeness. Since that day, I've never been surer of what I wanted to end up doing with my life. For that, my first thank you of this dissertation must go to Reza. Thank you for giving me my calling.

Although I've known for almost 13 years what my general vocation would end up being, exactly what I would be a professor of and how I would get there have rarely been sure things for more than a few days in a row. Therefore, the rest of my thanks are for everyone who has helped me through this uncertain journey.

Thank you to my parents, Kim and Greg Johnson, for giving me the freedom and encouragement to choose my career path, combined with the appreciation for the hard work and drive it takes to be able to walk your own path. Thank you most of all for your endless love and support every step of the way.

Thank you to my sisters, Kelly and Katie Johnson, for being constant examples of how to be successful, loving and caring human beings. You are my best friends, and I am in constant awe of all that you do and are.

Thank you to the animals in my life: Licorice (1997-2011) and Dexter (2013-Present). When life gets you down, nothing is quite as sweet as petting a dog.

Thank you to all my extended family, grandparents, great-grandparents, aunts and uncles, great-aunts and great-uncles, first and second cousins (however many times removed), for giving me a network of love and support for 31+ years. You sweeten life.

Thank you to my in-laws, Pat and Chuck Hession, for your love and unwavering support of Kathleen and me during our years in Minneapolis. Your regular visits provided the perfect blend of fun, rest, discussing big ideas, and exploring new corners of the Twin Cities. Thank you to the enormous group of extended in-laws for making my already spectacular family network that much bigger and more supportive.

Thank you to everyone I have been able to call my friend ever since I could recognize the value of friendship, in every conceivable context. Each one of you has shaped me in the way only friends can do, by showing me that important blend of graciousness, kindness, and a glimpse into how life works for another human being. The time passed with a friend I saw yesterday is no more or less valuable than the time passed with a friend I haven't seen for years. It meant what it was supposed to mean at that moment in time and space.

I will say a specially thank you to those of you I have been able to call friends and colleagues over the last four years at the University of Minnesota. In this most crucial time of transformation from student to professor, you helped give me a personal and professional identity. Every day, you made me realize that the dream of a naïve 18-year-old of going into academia was still valid and incredibly worth it. You pushed me and challenged my beliefs, even when I wasn't expecting (or sometimes even wanting) to be pushed and challenged.

As a future educator, I owe an extra special thank you to my teachers. Each and every one of you. From pre-school to grad school.

Pre-school: Carol Ducette; Linda Scheetz.

Elementary school: Linda Rice (Kindergarten); Ms. Hazelhuhn (1st grade); Ms. Jensen (2nd grade); Linda Grigsby, Lisa Roberts and Stan Vanderlinden (3rd grade); Julie Piper and Lisa Roberts again (4th grade); Nancy Kohl and Dan Mascal (5th grade); David Quegg and Sarah Sobocinski (6th grade); Denny Thomas (music); Gwen Leslie (art); Mr. Dane and Melanie Alberts (P.E.); Ms. Gelman (library); Ms. Grey (elementary band); Doris Leslie (piano).

Junior high: Dr. Witinok (earth science); Ms. Chelf, Mr. Kuepker and another teacher whose name I can't remember (P.E.); Jerry Zinn (band); Janice Shields (music); Ms. Hawtry (reading); Ms. Mixdorf (home-ec); Mr. Piper (personal development); Ms. Gherke (personal development); Meg Corbin (language arts … twice); Kay Nigg (algebra); Lois Crowley (global studies); Mr. Remley (leadership); Mr. Huff (American studies); Sra. Phelps (Spanish); Sra. Uribe (Spanish … and later again in high school); Mr. Griffith (art); Mr. Norton (geometry); Mr. Bordewick (life science); and to all of my junior high coaches, especially coach Coleman.

High school: my first English 9 teacher whose name I can't remember; Rob and Rich Medd (band and jazz band); Wayne Thelander (orchestra); Zach Durlam (show choir band); Kate Hamm (theatre); Sra. Duffner (Spanish); Sr. Wood (Spanish); Sr. Rosenthal (Spanish); Nancy Pacha (Spanish … you deserve an especially huge thank you); Mr. Rogers (home-and-auto maintenance and drivers ed); Kathy Bresnahan (health and P.E.); another health and P.E. teacher whose name I can't remember; Mr. Stumpf

(P.E.); Gary Hollingsworth (P.E. … and for being an extraordinary track coach); Mr. Wymer (biology); Dr. Herman (AP biology); Mr. McLaughlin (chemistry); Ms. Muhly (algebra 2); Ms. Walker (pre-calc); Jon Bach (calculus); Gary Neuzil (psychology); LeShane Sadler (American studies); Mr. Abermeit (European history); Ms. Potter (world religions … twice!); Mr. Lindsey (English 9 and 10); Nate Frese (American humanities); Ms. Dole (Brit lit); Ms. Cox (Brit lit); Brady Schutt (AP government); Mr. Bancroft (economics); Taz Anthony (AP European history).

Undergraduate years: first, all of the TAs and instructors, some of whose names I can remember and some of whose I cannot; Reza Aslan (again, why not, for Introduction to Islam and Religion and Politics in the Middle East); Dr. French (Anthropology of Language); Dr. Filios (twice for Spanish seminars); Jim Drier and Marc Earnest (jazz); Dr. McBride (Zen Buddhism); Dr. Coneybear (International Relations); Amber Brian (Latin American Civilization); Morten Schlütter (Introduction to Buddhism); Jay Holstein and Ralph Keen (Judeo-Christian Tradition … and Ralph Keen again for Honors Senior Seminar … and again in grad school for Religion in the Modern Secular State); Arda Collins (Creative Writing); Dr. Vecera (Introduction to Psychology); Dr. Mutel (Astronomy); Dr. Huntington (Anthropology of Religion); Dr. Fales (Introduction to Philosophy); Dr. Klemm (Religion and Society); all of my professors from studying abroad in Murcia, Spain during the fall semester of 2004; Dr. Steele (Macroeconomics); Dr. Golder (European Politics); Dr. Turner (African-American Religious Perspectives); Dr. Verónica (Latinos in the United States); Dr. Golnick (Latin American Literature); Prof. Wing (Law in the Muslim World); Dr. Souaiaia (senior seminar adviser); Dr. Barbosa (Luso-Brazilian Culture … and twice again in grad school); Dr. Duarte

(Portuguese for Spanish Speakers … and again in grad school for Brazilian Literature); Dr. Rothman (Hispanic Linguistics).

Thank you to all of my professors at PUC-Minas in Belo Horizonte in 2008.

MA years: Roy Justis (Introductory Reporting and Writing); Steve Berry (Investigative Reporting, and Contemporary Problems in Journalism); Steve Bloom (Advanced Reporting and Writing, and my MA project adviser); Leo Eko (Law and Ethics in Mass Communication … I owe a huge thank you to Dr. Eko for first starting me down the path of studying mass communication law); Patrick Reipe (web design); John Kimmich-Javier (Advanced Photojournalism); Lisa Rose-Weaver (Global Journalism); Dr. Sosale (Globalization and Mass Communication); Don McLeese (Magazine Reporting and Writing); Connie Peterson (Publication Design); Chris Lammer-Heindel (Ethics); Bonnie Sunstein and Jane Singer (serving on my MA project committee). Also, thank you to Venise Berry and Leo Eko for my experience TAing for them, and to all of the students who helped me become a better teacher and grader.

And, finally, in my PhD program … first the professors: Jisu Huh, Brian Bix, Timothy R. Johnson, Dan Sullivan, Mary Rumsey, Marco Yzer, Liz Fry, Bob DelMas, and Lisa Hillbink. Now, those for whom I was a TA: Shayla Thiel-Stern, Nora Paul, and Kathy Hansen.

And now, some very big thank yous to the faculty who were most influential in shaping my graduate career:

To Amy K. Sanders: You were an incredible adviser, mentor and friend. You showed me how to find and put together the pieces that make research profound, relevant

and a joy to conduct. You have given me a strong foundation on which to build an exceptional career. Thank you for your guidance throughout this arduous process.

To Seth C. Lewis: Your ability to see and investigate interesting and big ideas is uncanny. Thank you for helping me develop my own version of that kind of meaningful research acumen.

To Brendan R. Watson: Thank you for keeping me grounded in this diverse, complex and richly traditional field known as mass communication.

To William McGeveran: Thank you for pushing my legal reasoning skills closer and closer toward the grand-master level that you so deftly demonstrate, where arguments become chess matches and every conceivable move and countermove is readily apparent before your eyes.

To Jane Kirtley: Thank you for many things. First, for never letting me forget that a free, fearless and robust press is the ultimate reason we study, teach and defend the First Amendment. Second, for opening up my eyes to the many avenues, small and large, hidden and overt, in which issues of mass communication law can be found every day. And third, thank you for improving my scholarly writing with just one sentence.

To Giovanna Dell'Orto: Thank you for being a mentor, friend, and fellow *peregrino*. You have been incredibly supportive in helping me develop and test my ideas, and you were undeniably instrumental in helping me take the next step in my career. Every interaction with you has been as encouraging and uplifting as a friendly "*buen camino*."

I must also say thank you to Twin Cities Metro Transit for giving me a cheap, green and reliable source of transportation throughout the last three of my four years in

Minneapolis. Even more valuable than saving money on gas and parking was the time I could spend on the bus every day working, studying or grading. My PhD would've taken another year to complete if not for the service you provided.

And, saving the best for last, thank you to Kathleen. Four years ago you took this crazy journey with me to Minneapolis without any guarantee that my four years of grad school would lead us both to anything remotely close to a good, successful life. What did we do? We made a truly amazing life together. We supported each other through many successes and a few disappointing failures. We pushed each other to work harder and seek new knowledge in places we never would have dared explore alone. We challenged each other's convictions, and then built new ones together. Without you, my life would be significantly more boring, less joyful, and very mediocre. Together, we can do anything.

Dedication

Once again—can’t say it enough—to Kathleen.

Abstract

This study explicates the concept of governance by mainstream online digital intermediaries such as Facebook, YouTube and Twitter over extreme user-generated content (UGC)—a.k.a. "content governance." The study synthesizes First Amendment theory and jurisprudence, as well as theories about the interconnected power roles of individuals and digital intermediaries, to explicate how such content is governed in an environment of global networked communication. Two key questions guide this explication: How and why do digital communication intermediaries respond to extreme UGC? What are the potential implications of their responses for public discourse in a system of networked communication? This study also examines ethical duties that digital intermediaries may have to protect speech or prevent harm. This synthesis of theories is applied to an empirical case-study analysis of how Facebook has changed its community guidelines throughout the 11 years of its existence. This analysis will look at examples of Facebook removing or not removing extreme UGC from its platform. The purpose of this analysis is to assess how the norms of freedom of expression are being negotiated in a networked communication environment facilitated by digital intermediaries.

Table of Contents

List of Figures

Chapter 1: Introduction

Context

Over the last two decades, Internet communication and globalization repeatedly have challenged legal and social limits of freedom of expression in the United States and around the world. In the mid-to-late 1990s, interest groups in the United States rallied around the cause of protecting children from accessing the vast amounts of pornography that were becoming readily available online.[1] To this day, governments around the world continue to fight a Sisyphean battle to stanch the online trafficking of images of child sexual abuse.[2] Incongruities in countries' defamation laws have upset the balance between protecting speech and preserving reputations across borders, particularly between the United States and virtually every other country on the planet.[3] The Internet's

---

[1] *See* Philip Elmer-DeWitt, *On a Screen Near You: Cyberporn*, 146 TIME 38 (July 3, 1995) (discussing a study that claimed that more than 80% of the Internet in 1995 was pornographic); Reno v. ACLU, 521 U.S. 844 (1997) (striking down parts of the 1996 Communications Decency Act (47 U.S.C. § 223) for being unconstitutionally vague and overbroad in its attempt to restrict "indecent transmissions" on the Internet, and classifying the Internet as more akin to print media than to broadcast in terms of government's ability to regulate it); the 1998 "Child Online Protection Act," 47 U.S.C. 231, a law that attempted to force commercial providers of online pornography to restrict access to minors (ultimately struck down as unconstitutional in Ashcroft v. ACLU, 542 U.S. 656 (2004)).

[2] *See* the 1996 "Child Pornography Protection Act," 18 U.S.C. § 2251 *et seq*., which attempted to extend federal prohibitions on images of child sexual abuse (18 U.S.C. §§ 1460-1466) to images that seem to depict minors yet are not made using real children (i.e. young adult actors portraying minors or animations of minors in explicit sexual contexts) (struck down as unconstitutionally vague and overbroad in Ashcroft v. Free Speech Coalition, 535 U.S. 234 (2002)); Sean O'Neill, *Police Are "Failing to Halt Spread of Online Child Abuse,"* THE TIMES (LONDON) (June 17, 2013) (quoting Michael Moran, head of Interpol's Crimes Against Children unit, as saying that "no police force in the world" was properly combating the spread of child pornography online); Charles Arthur, *Online Pornography: Cameron's "War" Muddles Two Separate Issues*, THE GUARDIAN (July 23, 2013) (arguing that people who view images of child sexual abuse do so through proxy servers or private networks, and "the only way to stop someone really determined to access [those] sites is to cut off the Internet").

[3] *See* the 2010 "Securing the Protection of our Enduring and Established Constitutional Heritage (SPEECH) Act, 28 U.S.C. § 4010 *et seq*., stipulating that U.S. courts shall not recognize foreign defamation judgments against U.S. citizens unless "the defamation law applied in the foreign court's adjudication provided at least as much protection for freedom of speech and press … as would be provided by the first amendment" (§ 4102(a)(1)(A)) or the U.S. defendant "would have been found liable for

facilitation of anonymous speech has lowered the social cost for speakers to inflict many types of harm through their online words.[4] The ease of sharing digital files of copyrighted works has facilitated contributory and vicarious infringement of copyright.[5] And speech that is legal (and even socially acceptable) in some parts of the world can infiltrate other corners of the world where the mere knowledge that it exists is so offensive that it can push people to the point of protest, civil unrest, violence or murder.[6]

These issues and events have created a web of context that has pulled scholars of mass communication, law, ethics, political science, international studies, and science-and-technology studies closer together in pursuit of answering several important questions: What does freedom of expression mean in our increasingly connected world? Can conflicting legal norms of freedom of expression be reconciled with each other as the Internet allows speech to penetrate borders? If so, how? If not, how can humanity best adapt to this new world? Are the widely publicized incidents of speech associated with

---

defamation by a domestic court" (§ 4102(a)(1)(B)); Trout Point Lodge, Ltd. v. Handshoe, 729 F.3d 481 (5th Cir. 2013) (holding that a Canadian libel judgment against a U.S. citizen could not be successfully enrolled in a U.S. court because the Canadian plaintiffs could not meet either of the burdens of the SPEECH Act); *cf* Dow Jones & Co. v. Gutnick, High Court of Australia [2002] HCA 56 (holding that "those who post information on the World Wide Web do so knowing that the information they make available is available to all and sundry without any geographic restriction," and that "publication" for the purposes of Australian defamation law occurs wherever the alleged defamatory writing is downloaded).

[4] *See* Lyrissa Barnett Lidsky, *Incendiary Speech and Social Media*, 44 TEX. TECH L. REV. 147 (2011); DANIELLE KEATS CITRON, HATE CRIMES IN CYBERSPACE (2015); Mary Anne Franks, *Unwilling Avatars: Idealism and Discrimination in Cyberspace*, 20 COLUM. J. GENDER & L. 224 (2011); Danielle Keats Citron, *Cyber Civil Rights*, 86 BOST. U. L. REV. 61 (2009); Yuval Karniel, *Defamation on the Internet—A New Approach to Libel in Cyberspace*, 2 J. INT'L MEDIA & ENT. L. 215 (2008).

[5] *See* A & M Records, Inc. v. Napster, Inc., 239 F.3d 1004 (9th Cir. 2001); Metro-Goldwyn-Mayer Studios, Inc. v. Grokster, Ltd., 545 U.S. 913 (2005).

[6] A prominent example of such speech is the controversy surrounding the cartoons of the Prophet Muhammad published in the Danish newspaper *Jyllands-Posten* in September 2006. *See, e.g.,* Adria Battaglia, *A Fighting Creed: The Free Speech Narrative in the Danish Cartoon Controversy*, 43 FREE SPEECH YEARBOOK 20 (2006); Stephanie Craft and Tayo Oyedeji, *United States: Journalism as a Prism of Culture Clash*, *in* READING THE MOHAMMED CARTOONS CONTROVERSY: AN INTERNATIONAL ANALYSIS OF PRESS DISCOURSES ON FREE SPEECH AND POLITICAL SPIN (Risto Kunelius, Elisabeth Eide, Oliver Hahn & Roland Schroeder eds., 2007).

harmful outcomes leading to a decrease in tolerance for liberal conceptions of freedom of expression, such as the regime of First Amendment jurisprudence in the United States? If so, to what extent? What are the implications for global democratic discourse of a world that has little or no tolerance for extreme and potentially harmful speech? Is such a world desirable or not? Why?

This study does not pretend to fully answer all of these questions. They are part of a large and ongoing global debate, to which this study will lend its perspective. Therefore, these questions will guide the analyses herein as they focus on one particular aspect of our globalized and networked world: How mainstream[7] digital intermediaries deal with extreme speech.

Extreme Speech in a New Communicative (and Regulatory) Environment

A Question of Balance

This study revolves around the following concept: Mainstream digital intermediaries (such as Facebook, Twitter and YouTube) engage in a constant balancing act between protecting their users' ability to speak freely on their platforms and preventing harms that may arise from users' speech. This balancing act is not easy.[8] In fact, finding the ideal point at which the greatest amount of speech is protected relative to the least amount of harm caused may be an impossible task for digital intermediaries. Yet, as intermediaries struggle to find that ideal point, the speech of every individual who uses these intermediaries is implicated. The robustness of the public discourse that takes

[7] *See infra* note 17.
[8] Josh Braun and Tarleton Gillespie, *Hosting the Public Discourse, Hosting the Public*, 5 JOURNALISM PRACTICE 383, 392 (2011)

place on these intermediaries is implicated. Therefore, individuals must understand—and this study seeks to explicate—how digital intermediaries govern the content that individuals publish to their platforms, and what the various ramifications are for this process of governance. This explication will require the synthesis of multiple theories from several distinct fields of study, including mass communication, First Amendment jurisprudence, and media ethics. It also requires a strong foundation based on the definitions of the two key concepts used throughout this study: extreme speech, and digital intermediaries.

Definition of Key Concepts

Extreme Speech

Devoid of context, the term "extreme speech" is vague. It can "describe a wide variety of expression, such as [H]olocaust denial, extreme pornography, and speech inciting hatred or likely to provoke public disorder,"[9] though it is arguably primarily used to describe "expression that is seen to be discriminatory or perpetuate[s] discriminatory attitudes."[10] Any of the examples cited in the opening paragraph above could constitute extreme speech, including defamatory speech or speech that violates copyright, depending on the context in which they arise. Indeed, because the definition of extreme speech depends so much on disparate and constantly evolving social norms,[11] coming up with a clear definition that serves this study is no easy task. However, what this study can

[9] Jacob Rowbottom, *Extreme Speech and the Democratic Functions of the Mass Media*, *in* EXTREME SPEECH AND DEMOCRACY (Ivan Hare & James Weinstein eds., 2009), 608.
[10] *Id*.
[11] Robert Post, *Hate Speech*, *in* EXTREME SPEECH AND DEMOCRACY (Ivan Hare & James Weinstein eds., 2009), 129.

do is place boundaries around broad categories of extreme speech, thereby finding a compromise between respecting the fluidity of their definition and identifying a recognizable focus for analysis. Therefore, this study focuses on extreme speech that 1) can reasonably be considered to carry a political or social message;[12] 2) is protected under the First Amendment; yet 3) could reasonably be considered to cause some type of harm.

The analysis of extreme speech in this study is steeped in the tradition of the First Amendment. This tradition has a distinct set of assumptions and biases that must be made clear from the outset. First, this study takes the perspective that more speech is better than less, and therefore the practice of restricting speech must be based on strong interests that outweigh the interests of promoting as much speech as possible.[13] Second, and in relation to the first point, this study takes the perspective that extralegal or social means of restricting speech are just as threatening as legal means to the values of promoting more speech.[14] This perspective will color the bulk of the analysis in this study, and it ties into the second focus of the analysis: the role digital intermediaries play in governing the speech that gets published on their platforms.

---

[12] The requirement that speech have "political" or "social" significance is loosely borrowed from the third part of the test for obscenity from *Miller v. California*, 413 U.S. 15, 24 (1973), whereby speech cannot be considered unprotected obscenity if it can reasonably be considered to have "serious literary, artistic, political, or scientific value." Although obscenity is not the focus of this study, the same broad idea from *Miller*, that undesirable speech can and often does have political or social value, is at the center of this study.

[13] These competing interests will be the focus of chapters 3 and 4.

[14] *See infra* notes 45-48.

Intermediaries

Each of the harmful types of speech listed above has one thing in common: to reach the public, it requires an intermediary, or a "platform,"[15] an online space devoted to facilitating communicative activity among individual users.[16] The analyses in this study focus particular attention on what could be considered "mainstream" digital intermediaries: Twitter, Facebook and YouTube, which together connect at least 1.35 billion users to each other and publish their various forms of content to the world.[17] The roles that these intermediaries play in facilitating the networked communication environment complicate traditional legal means of defining what speech is and how it might be regulated. Intermediaries make it easy for anyone with Internet access to publish his or her speech before a potentially global audience. Speech of all flavors, created by individuals, can be published on platforms: harmful speech, pornographic speech, entertainment, social or political commentary, inane banter, and addicting content that causes readers to easily waste an hour on a Saturday afternoon. And, of course, these types of speech are not mutually exclusive.

---

[15] Tarleton Gillespie, *The Politics of Platforms*, 12 NEW MEDIA & SOC'Y 347 (2010).
[16] *Id*. at 351.
[17] Caitlin Dewey, *Almost as Many People Use Facebook as Live in the Entire Country of China*, WASH. POST "THE INTERSECT" BLOG (Oct. 29, 2014) (citing Facebook's claim in its 2014 Q3 earnings report that it had more than 1.35 billion monthly active users) *available at* http://www.washingtonpost.com/news/the-intersect/wp/2014/10/29/almost-as-many-people-use-facebook-as-live-in-the-entire-country-of-china/; *About*, TWITTER (2015) (reporting that the company has more than 288 million monthly active users), *available at* https://about.twitter.com/company; *Statistics*, YOUTUBE (2015) (reporting that the platform has more than 1 billion users), *available at* https://www.youtube.com/yt/press/statistics.html. *See generally* CITRON, *supra* note 4. Although she does not explicitly label the intermediaries she analyzes, Citron places platforms into five categories throughout the course of her book when discussing where abusive speech occurs online: 1) Sites specifically devoted to hate messages, such as white supremacist websites; 2) sites that encourage the posting of potentially abusive UGC, such as TheDirty.com; 3) sites devoted to gossip; 4) sites devoted to trolling (*see* the section of chapter 4 titled "Abuse, Trolling and Gamergate"); and 5) mainstream sites, such as YouTube, Facebook and Twitter.

Figure 1-1 presents a very basic model of networked communication. Individuals, who are simultaneously speakers and audience,[18] publish their various forms of speech via a digital intermediary. The intermediary makes that speech available for other individuals, who then publish more speech, either in response to or completely unrelated to another individual's message. The process is continuous. For the sake of clarity, only four individuals are shown in this model to represent the hundreds of millions of individuals who speak via intermediaries. Also (and again for the sake of clarity), only one digital intermediary is shown in this model; however, individuals are likely to use multiple intermediaries, and the speech they publish through one intermediary has the potential to influence speech published through others (e.g. a YouTube video may get shared via Twitter or discussed on Facebook).

[18] *See, e.g.*, AXEL BRUNS, BLOGS, WIKIPEDIA, SECOND LIFE, AND BEYOND (2008); YOCHAI BENKLER, THE WEALTH OF NETWORKS: HOW SOCIAL PRODUCTION TRANSFORMS MARKETS AND FREEDOM (2006).

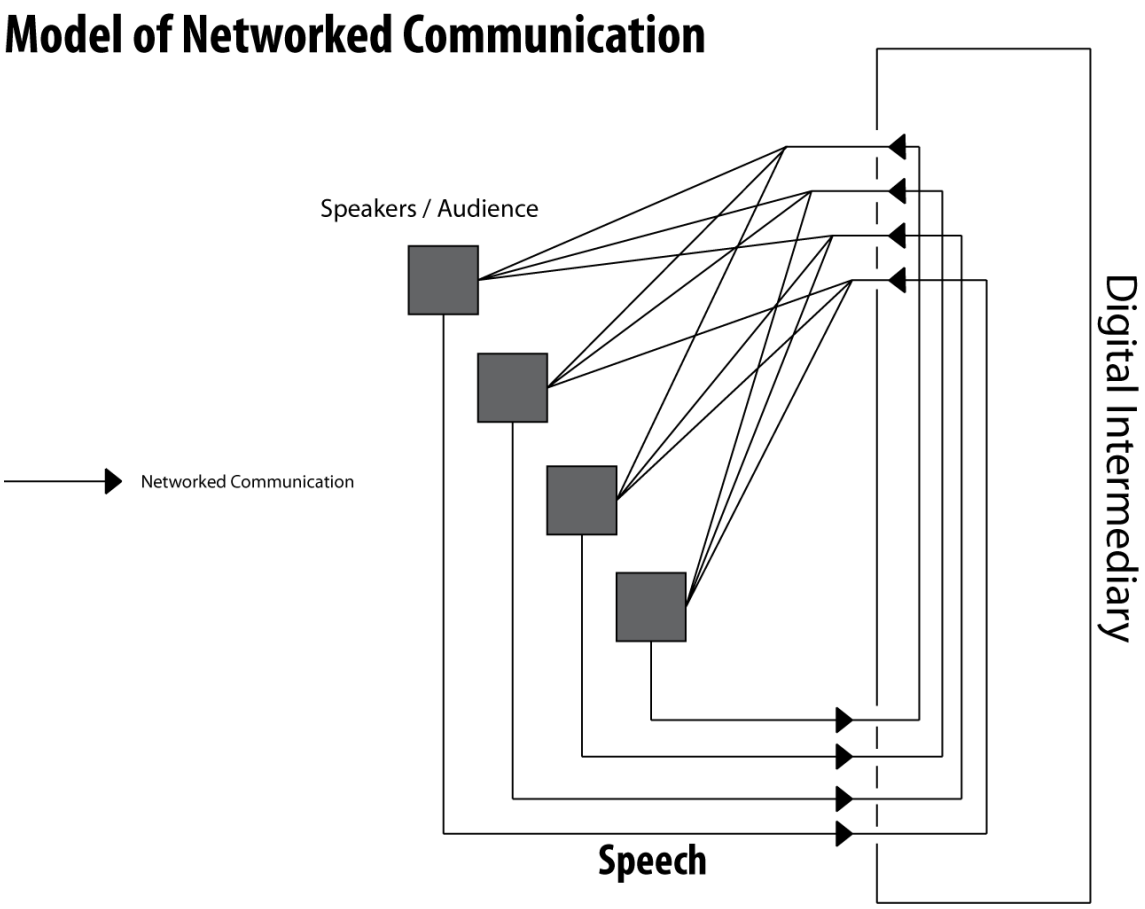

Figure 1-1: Networked Communication

What intermediaries giveth, they also taketh away. Just as they are the vehicles that propel speech across borders and into communities where it is not welcome, intermediaries often are agents of regulation of speech when that speech is so unwelcome that it becomes harmful. Intermediaries can move swifter than many state actors (at least in liberal democracies) to remove harmful speech from the public discourse, and, because they are not state actors, intermediaries can remove speech virtually without fear of consequences for violating an individual's "right" to speak on the platform.[19]

[19] *See* Braun and Gillespie, *supra* note 8; LAURA DENARDIS, THE GLOBAL WAR FOR INTERNET GOVERNANCE (2014); Colin M. Maclay, *Protecting Privacy and Expression Online*, *in* ACCESS

Digital intermediaries are not state actors. These intermediaries have a First Amendment right to govern the speech that they host on their platforms, while individuals have no constitutional claim against these intermediaries stemming from their acts of governance.[20] However, from the perspective of champions of freedom of expression, private mediation of harmful content raises several issues. First, just like the definition of extreme speech discussed above, the definition of what constitutes "harmful" speech is fluid and heavily dependent on both the context within which it arises and the sensibilities of the individuals or groups that it implicates. Under such a fluid definition for harmfulness, extralegal actions to police user-generated content (UGC) essentially may end up following the subjective early 20th century "bad tendency test" for determining undesirable speech.[21] Applying the broad guiding questions above

---

CONTROLLED: THE SHAPING OF POWER, RIGHTS, AND RULE IN CYBERSPACE (Ronald Deibert, John Palfrey, Rafal Rohozinski and Jonathan Zittrain eds., 2012), at 90.

[20] Miami Herald v. Tornillo, 418 U.S. 241 (1974) (holding that a Florida statute requiring newspapers to publish responses from individuals who believed they were attacked in the newspapers amounted to an unconstitutional prior restraint.) Adopting a negative approach to First Amendment jurisprudence, the Court determined that such a statute constituted government control over the editorial process of a free press. Although the Court acknowledged that an ideal press was a responsible press that provided a forum for diverse viewpoints, it nonetheless held that "press responsibility is not mandated by the Constitution, and like many other virtues it cannot be legislated," at 256. *See also See* Bruce W. Sanford and Jane E. Kirtley, *The First Amendment Tradition and Its Critics*, *in* THE PRESS (Geneva Overholser and Kathleen Hall Jamieson eds., 2005), at 268. But *cf* Red Lion Broadcasting Co. v. FCC, 395 U.S. 367 (1969). In *Red Lion*, a unanimous Supreme Court held that the now defunct Fairness Doctrine—which required broadcasters to allot equal time to discussion of competing issues—did not violate broadcasters' First Amendment right to "use their allotted frequencies continuously to broadcast whatever they choose, and to exclude whomever they choose from ever using that frequency" (at 386). Rather, the Court held that the medium of broadcast (via TV or radio), with its unique situation of depending upon the scarce availability of the electromagnetic spectrum, required regulations such as the Fairness Doctrine to prevent a monopoly of viewpoints from controlling all access to broadcast channels. The Court matter-of-factly averred, "It is the right of the viewers and listeners, not the right of the broadcasters, which is paramount" (at 390). Although such an audience-centric interpretation of the First Amendment may have found favor at the High Court, that interpretation strictly applied to broadcast rather than print, and the Court has stated that the Internet should be classified as more akin to the latter than the former (Reno v. ACLU, 521 U.S. 844, 870 (1997).

[21] Zechariah Chafee, *Freedom of Speech in War Time*, 32 HARV. L. REV. 932 (1919); Debs v. United States, 249 U.S. 211 (1919).

to the specific focus of digital intermediaries, this study asks the following: How is extreme speech defined in an era of networked communication? How do digital intermediaries govern the extreme and potentially harmful speech that users publish to their platforms? What are the potential implications of content governance to global discourse on matters of social and political importance? Answering these questions will require a theoretical framework based on a synthesis of literature from several fields.

This study is based on a somewhat radical idea: The fact that digital intermediaries are not state actors does not preclude an analysis of the social values of extreme or potentially harmful speech in a global and networked society, nor an appraisal of the social implications of the ways in which content governance may be affecting these values. This study is steeped in the tradition of the field of mass communication law, yet it is not a study about the law. Put differently, the focus is not an analysis of a specific new law or a proposal for a specific new law. Rather, this study uses the law as a lens to explicate the concept of content governance. The chapters outlined below will elaborate further on how this study plans to go about accomplishing this mission.

Overview of Chapters

Chapter 2: "Conceptualizing Private Governance in a Networked Society: A Review of Scholarship on Content Governance"

The goal of chapter 2 is to explore the relationship between digital intermediaries and the individual users who communicate via these intermediaries. Beginning the study here is crucial for developing a framework with which to analyze content governance.

The focus of this chapter will be on the nature of the so-called "networked economy"[22] and the role that it plays in shaping the function of digital intermediaries as both facilitators and regulators of the speech that individuals publish on their platforms. This chapter will also examine the position of state actors in this new communicative system, coming to the conclusion that all three actors in the system (individuals, intermediaries and state actors) are interdependent of one another within this system. This interdependence is the defining characteristic of the current individual-driven and intermediary-facilitated speech environment.

To explicate how regulation of speech operates within this system of interdependence, this chapter uses a theoretical framework that borrows from the field of Internet governance[23] and law professor Lawrence Lessig's concepts of legal and extralegal regulation.[24] Using this framework, this chapter takes the approach that the interdependent regulatory relationship among individuals, intermediaries and state actors is one whose definition is in constant flux.[25]

The second half of the chapter is devoted to incorporating affirmative First Amendment theories into the analysis of how intermediaries facilitate speech. Such

---

[22] *See* José van Dijck, *Users like you? Theorizing agency in user-generated content*, 31 MED. CULT. & SOC'Y 41 (2009); Ute Schaedel & Michel Clement, *Managing the Online Crowd: Motivations for Engagement in User-Generated Content*, 7 J. MED. BUS. STUD. 17 (2010); James G. Webster, *User Information Regimes: How Social Media Shape Patterns of Consumption*, 104 NW. U. L. REV. 593 (2010); Ramon Lobato, Julian Thomas & Dan Hunter, *Histories of User-Generated Content: Between Formal and Informal Media Economies*, 5 INT'L J. COMM. 899 (2011).

[23] *See*, e.g., DENARDIS, *supra* note 19; Malte Ziewitz and Christian Pentzold, *In Search of Internet Governance: Performing Order in Digitally Networked Environments*, 16 NEW MEDIA & SOC'Y 306, 307 (2014); Michel JG van Eeten and Milton Mueller, *Where Is the Governance in Internet Governance?* 15 NEW MEDIA & SOC'Y 720, 723 (2013).

[24] LAWRENCE LESSIG, CODE, VER. 2.0 (2006).

[25] *See*, e.g., TARLETON GILLESPIE, WIRED SHUT: COPYRIGHT AND THE SHAPE OF DIGITAL CULTURE (2007).

theories are generally consequentialist in nature and conceive of freedom of expression as an affirmative right: a right to speak due to the benefit speech brings to society, rather than a right to not have the government restrict one's speech.[26] These theories are valuable to the analysis in this chapter for two reasons. First, their essential focus on the social values of freedom of expression provides an analytical lens that portrays the role of intermediaries as facilitators of these values. Second, several versions of these theories view private actors that have great control over channels of speech as equally threatening to the robustness of the public discourse as state actors with the censorial power of law.[27] Thus, affirmative First Amendment theories have the ability to critically argue that the essential democratic function of intermediaries should be seen as more valuable or as a more important concern than intermediaries' own rights as private businesses (or, at the very least, that these concerns are equally important).

Chapter 3: "The Value of Extreme Speech in a Networked Society: A Perspective from First Amendment Theory and Jurisprudence"

The purpose of this chapter is to analyze how key First Amendment theories and doctrines define the notions of extremeness and harmfulness in the context of freedom of expression, and ultimately apply these definitions to how freedom of expression is exercised in a global system of networked communication. This chapter begins with the foundational question posed by law professor Frederick Schauer: Why must speech be

---

[26] *See, e.g.,* ALEXANDER MEIKLEJOHN, FREE SPEECH AND ITS RELATION TO SELF-GOVERNMENT (1948); CASS R. SUNSTEIN, DEMOCRACY AND THE PROBLEM OF FREE SPEECH (1993); OWEN FISS, THE IRONY OF FREE SPEECH (1996).

[27] *See, e.g.,* Lessig, *supra* note 24; Jack M. Balkin, *The Future of Free Expression in a Digital Age*, 36 PEPP. L. REV. 438 (2009) [hereinafter Balkin, *The Future of Free Expression*]; Jack M. Balkin, *Digital Speech and Democratic Culture: A Theory of Freedom of Speech for the Information Society*, 79 N.Y.U. L. REV. 1 (2004) [hereinafter Balkin, *Digital Speech and Democratic Culture*].

special?[28] In other words, why should this analysis pay so much attention to the potential effects of methods of Internet governance on speech ("content governance") when methods of Internet governance affect many other phenomena and seek to uphold other important social values? Like Schauer, this chapter concludes that speech must be considered special because of the multiple values that come from the broad category of speech, values that require us to carefully analyze the potential harms that can be experienced at the expense of these values.[29] I then proceed to an analysis of three of the most prominent First Amendment theories: marketplace of ideas theory, individual autonomy theory and tolerance theory.[30] Each of these theories, in its own way, answers the following questions: What makes speech harmful? What makes it so harmful that it should be regulated? How should that speech be regulated? At what point does regulation damage deliberative democracy? Answering these questions will help clarify the social values of extreme and potentially harmful speech.

Grasping clear definitions of harm is a task that calls for law professor Rodney Smolla's three-part model of harmful speech, which identifies physical harm, relational harm, and reactive harm.[31] These three classifications of harm are useful because they neatly categorize First Amendment doctrine. U.S. free speech jurisprudence considers physical harm the worst of the three types of harms, and certain tests have been devised to address this harm and to decide when the speech that causes it falls outside of

[28] Frederick Schauer, *Must Speech Be Special?* 78 NW. U. L. REV. 1284 (1983).
[29] *Id*. at 1304, 1306.
[30] Another important First Amendment theory, self-governance theory, is not discussed in this chapter because it is part of chapter 2's discussion of affirmative First Amendment theories. However, versions of self-governance theory will make an appearance in chapter 3 as perspectives for criticizing the three theories focused on in chapter 3.
[31] RODNEY A. SMOLLA, FREE SPEECH IN AN OPEN SOCIETY 48 (1992).

constitutional protection. These types of speech include fighting words,[32] true threats,[33] and incitement to imminent lawless action.[34] Relational harm involves speech that causes injury to social relationships (defamation), business relationships (fraud or false advertising), ownership interests (copyright) and confidentiality (leaking national security secrets).[35] Courts have devised legal tests to determine if and how such speech should be legally sanctioned. Finally, reactive harms include intentional infliction of emotional distress of public officials,[36] and tortious invasions of privacy, as well as any type of hate speech. Hate speech has been defined many ways by many different scholars,[37] but a generic definition for the purposes of this study may categorize hate speech and any speech that attacks and attempts to subordinate any group or class of people, typically spoken by a group with a higher level of social power than the targets of the speech. The

[32] Chaplinsky v. New Hampshire, 315 U.S. 568, 572-3 (1942) (defining unprotected fighting words as words said in another person's face that "by their very utterance inflict injury or tend to incite an immediate breach of the peace").
[33] Virginia v. Black, 538 U.S. 343, 360 (2003) (O'Connor, J., writing for the plurality) (defining unprotected true threats as speech that can be interpreted both objectively and subjectively as threatening).
[34] Brandenburg v. Ohio, 395 U.S. 444, 447 (1969) (defining unprotected incitement as "advocacy of the use of force or of law violation [that] is directed to inciting or producing imminent lawless action and is likely to incite or produce such action").
[35] SMOLLA, *supra* note 31.
[36] Hustler Magazine, Inc. v. Falwell, 485 U.S. 46 (1988).
[37] For various studies with various definitions of hate speech, *see generally* Clay Calvert, *Hate Speech and Its Harms: A Communication Theory Perspective*, 47 J. COMM. 4 (1997); Alexander Tsesis, *Dignity and Speech: The Regulation of Hate Speech in a Democracy*. 44 WAKE FOREST L. REV. 497 (2009); Richard Delgado and David H. Yun, *Pressure Valves and Bloodied Chickens: An Analysis of Paternalistic Objections to Hate Speech Regulation*, 82 CAL. L. REV. 871 (1994); Owen M. Fiss, *The Supreme Court and the Problem of Hate Speech*. 24 CAPITAL U. L. REV. 281 (1995); Stephanie Farrior, *Molding the Matrix: The Historical and Theoretical Foundations of International Law Concerning Hate Speech*. 14 BERKELEY J. INT'L L. 1 (1996); Jean-Marie Kamatali, *The U.S. First Amendment Versus Freedom of Expression in Other Liberal Democracies and How Each Influenced the Development of International Law on Hate Speech*. 36 OHIO N.U. L. REV. 721 (2010); Post, *supra* note 11; Tanya Katerí Hernández, *Hate Speech and the Language of Racism in Latin America: A Lens for Reconsidering Global Hate Speech Restrictions and Legislation Models*. 32 U. PA. J. INT'L L. 805 (2011).

targets of such speech typically include racial minorities, women, religious minorities, and homosexuals.

According to Smolla, speech that leads to reactive harms deserves the highest level of constitutional protection due to its tendency to implicate public figures or officials, or its tendency to involve important social issues and matters of public concern: factors which greatly outweigh the potential harms of the speech. However, although these types of speech receive strong legal protection, their harms are no less real to the people who suffer them. Content governance has the potential to fill the role of mitigating these harms.

In crafting a model of harmful speech that would trigger mechanisms of content governance, this study primarily focuses on speech with the potential to cause reactive harms and physical harms. It does not focus on speech that can cause relational harms, namely defamation. This decision was made for several reasons. First, except for certain torts of invasion of privacy, First Amendment jurisprudence has all but precluded private individuals from recovering damages for reactive harms caused by other individuals. This lack of legal options for mitigating reactive harms has created a vacuum that means of private governance are able to fill. Meanwhile, defamation remains a tort in which private individuals in the United States have viable options for recovery against their alleged defamers, thereby giving digital intermediaries little reason or incentive to mitigate defamatory claims on the behalf of individuals.

Second, physical harms are given greater attention in this study because they remain the most grievous type of harm regardless of which set of rules—First

Amendment jurisprudences or digital intermediaries' community standards—is doing the judging. However, the high standards that First Amendment jurisprudence places in front of state actors who wish to punish speech for its potential to cause physical harms can create a governance vacuum that digital intermediaries are able to fill. For example, posts on Facebook that advocate violent uprising may, in fact, lead some people to violently rise up against government officials. Chapter 3 will outline the reasons why the *Brandenburg* "imminent lawless action" standard likely would not find that such speech violated the law of incitement. However, the at-least perceived connection between the online speech and the physical harm caused could lead digital intermediaries to step in and remove the speech from its platforms in an attempt to prevent any further harm.

The discussion in this section of chapter 3 is framed around the following argument: tolerance theory, put forth by legal scholar Lee Bollinger in 1986,[38] should be revitalized as the preeminent theory of freedom of expression in a communication environment in which content governance has become a common tool for controlling free expression. Bollinger contends that by allowing extreme and hurtful views to be put forth into our public discourse, we are actively fighting our natural proclivity to want to be intolerant of these viewpoints—or any viewpoint we oppose, for that matter.[39] Tolerance theory is the best fit for the analyses within this study for several reasons. First, while other theories engage in apologetics of extreme speech as a necessary side effect of the central values the theories place upon freedom of expression, the central focus of

[38] LEE C. BOLLINGER, THE TOLERANT SOCIETY: FREEDOM OF SPEECH AND EXTREMIST SPEECH IN AMERICA (1986).
[39] *Id*. at 109.

tolerance theory is extreme speech (Bollinger calls it "extremist speech"). Bollinger, himself, does not give an explicit definition for extreme speech, but he does leave clues on how a definition can be formed. Extreme speech, Bollinger says, is what "nearly all of us believe immoral and vicious."[40] It "tend[s] to attract attention," and "is very often the product or the reflection of the intolerant mind at its worst and, as such, an illustration to us of what lies within ourselves."[41]

Second, tolerance theory posits that tolerance of extreme speech should end where significant harm begins.[42] Understanding harm and its relation to freedom of expression is important because content governance is about commercial intermediaries finding a balance between upholding the values of freedom of expression and mitigating the potential harms that individuals can cause through UGC. Individuals' ability to tolerate extreme speech represents the balancing point, and it is argued in this study that digital intermediaries play a crucial role in influencing individuals' tolerance by how they govern extreme speech. Third, tolerance theory holds that not only is it a natural tendency of *government* to censor, as proponents of individual autonomy and marketplace of ideas theory proclaim;[43] rather, it is the natural tendency of *every human being* to censor, and champions of freedom of expression must constantly be on guard against attempts from powerful non-state actors to censor speech.[44] This concept matches the argument from

---

[40] *Id*. at 124.
[41] *Id*. at 126.
[42] *Id*. at 192. Bollinger does not answer the question "When is speech so harmful it should be banned?" He does argue that "social" harm caused by allowing speech that society generally disapproves of is not sufficient to warrant proscription. *Id*. However, the principle of harm being the boundary of tolerance is an important one that fits the analysis in this chapter.
[43] SMOLLA, *supra* note 31, at 51.
[44] BOLLINGER, *supra* note 38, at 86.

Internet governance that non-state actors have great power to control speech, and that this power is worrisome due to the relative lack of transparency and standards employed in the process of controlling the speech.

Finally, tolerance theory is important because, at bottom, Bollinger's claim is that the act of tolerating extreme speech is an act of mental growth. Bollinger urges reflection on how society tends to invoke community norms to silence extreme, potentially harmful or otherwise undesirable speech so that society may collectively strengthen itself. Global networked communication continues to put more and more examples of detestable speech in front of our eyes, meaning that there has never been a more important time for society to strengthen its resolve and support for extreme speech.

The overall goal of this chapter is to distill the key values of freedom of expression from First Amendment theory and doctrine. This goal answers Lessig's call for a discussion on our values of freedom of expression as we continue to understand the relationship between content governance and freedom of expression. Lessig and many other legal scholars argue that we should focus on "constitutional values"[45] (others call them "goals,"[46] "principles,"[47] or "ideals"[48]). One such value, which Lessig argues is the preeminent value of the First Amendment, is to encourage mass participation in public discourse by individuals.[49] Similarly, law professor Jack Balkin argues that "a theory of

---

[45] *Supra* note 24, at 269.
[46] Ruth Walden, *A Government Action Approach to First Amendment Analysis*, 69 JOURNALISM Q. 65, 81 (1992).
[47] Adam Candeub, *The First Amendment and Measuring Media Diversity: Constitutional Principles and Regulatory Challenges*, 33 N. KY, L. REV. 373 (2006).
[48] Robert C. Post and Reva B. Siegel, *Democratic Constitutionalism*, *in* THE CONSTITUTION IN 2020 (Jack M. Balkin & Reva B. Siegel eds., 2009), at 30.
[49] LESSIG, *supra* note 24, at 269.

freedom of speech justified in terms of its potential contributions to representative self-government seems altogether too narrow in the age of the Internet."[50] The Internet, he argues, is proof that "the point of the free speech principle is to promote not merely democracy, but something larger: a *democratic culture*" defined by mass participation.[51] The argument of this chapter is that promoting a democratic culture through a commitment to facilitating individual speech on platforms and tolerating the extreme forms of speech that invariably come with such mass participation are the main values of freedom of expression by which the concept of content governance should be judged.

Chapter 4: "Heckler's Veto 2.0: Speakers' Rights v. Audience Rights in a Networked Society"

This chapter looks at the First Amendment doctrine of the "Heckler's Veto." A heckler's veto is "the suppression of speech by the government[] because of the possibility of a violent reaction by hecklers."[52] A heckler's veto occurs when "the state [hides] behind the unpleasant reaction of some portions of the public in order to silence a speaker" through the use of an instrument of law, such as a disorderly conduct statute.[53] The heckler's veto offers an analytical lens through which to examine the phenomenon of digital intermediaries governing extreme UGC out of a concern that the UGC threatens to violate certain community norms in some way, shape or form. The heckler's veto also offers an excellent analytical lens through which to address the conflict between

[50] Balkin, *The Future of Free Expression*, *supra* note 27.
[51] *Id*. (original emphasis). *See also* Balkin, *Digital Speech and Democratic Culture*, *supra* note 27, at 2.
[52] Ronald B. Standler, HECKLER'S VETO, Dec. 4, 1999, *available at* http://www.rbs2.com/heckler.htm.
[53] Cheryl A. Leanza, *Heckler's Veto Case Law as a Resource for Democratic Discourse*, 35 HOFSTRA L. REV. 1305, 1306 (2007).

speakers' "rights" and audience "rights" in the context of networked communication.[54] These rights have distinct histories in First Amendment theory and jurisprudence, yet the heckler's veto doctrine enshrines the principle that the rights of audiences end when members of an audience attempt to silence a speaker through an instrument of law.[55] Understanding the heckler's veto is important because it illustrates the ways in which the First Amendment protects extreme speech from being silenced by individuals who would seek to use community norms to goad the law into restricting speech.[56] Subsequently, understanding how extreme speech and community norms clash will be important to understanding forms of content governance, particularly the form that involves individuals putting pressure on intermediaries to remove or block speech that contravenes certain norms.

Chapter 5: "Facebook's Free Speech Growing Pains: A Case of Content Governance"

Chapter 5 explores the content governance practices of arguably the most popular digital intermediary: Facebook. Facebook's ubiquity in the global system of networked communication makes its content governance practices worthy of study; as of this

[54] "Rights," of course, are in quotes here because neither speakers nor audiences have a "right" regarding speech against a digital intermediary.
[55] Essentially, the heckler's veto doctrine straddles so-called negative First Amendment theories (such as the marketplace of ideas theory) that conceive of freedom of expression as a right *against* the government and so-called affirmative First Amendment theories (such as self-governance theory) that conceive of freedom of expression as a right that government can protect through proactive policies. *See generally* MEIKLEJOHN, *supra* note 26.
[56] *See* ROBERT POST, CONSTITUTIONAL DOMAINS 144 (1995) (contending that the main goal of the First Amendment is to promote "critical interaction" among groups with conflicting opinions and values, and that the way it does so is by preventing the government from invoking community norms to silence extreme speech).

writing, the social network claims to have more than 1.35 billion active monthly users,[57] making it the second most popular site on the World Wide Web.[58] Facebook is unique among digital intermediaries because it allows users to publish a broader array of content (including video, text and photos) with greater latitude as to the volume of that content (i.e. unlike Twitter, posts are not limited to 140 characters) compared to other mainstream intermediaries (namely Twitter and YouTube).[59] Facebook is a general social network, appealing not to one niche population but rather to users from many different nationalities, language groups, races, ages, religions and ideological backgrounds. This diversity both in types of UGC and among Facebook's users presents the company with a major challenge: to create a set of community standards that accounts for the nuances of diverse UGC and the dozens of norms of speech that its users from around the world bring to the network.

This chapter tackles the following research questions:

RQ1: How have Facebook's community standards changed from the origins of the social network until the standards' most recent update in March 2015?

RQ2: What instances have there been of Facebook controversially removing or not removing extreme UGC that seemed to contravene Facebook's community standards?

To answer these questions, this chapter will utilize the following methodological approaches. To answer the first research question, this chapter will analyze Facebook's

[57] Dewey, *supra* note 17.
[58] *The Top 500 Sites on the Web*, ALEXA (Mar. 20, 2015), *available at* http://www.alexa.com/topsites (Google.com is the Web's most popular site, according to Alexa).
[59] *See* Kate Crawford and Tarleton Gillespie, *What is a flag for? Social media reporting tools and the vocabulary of complaint*, NEW MEDIA & SOC'Y 1, 7 (2014) (discussing the policies of various digital intermediaries of allowing users to "flag" undesirable content, as well as the reasons why individuals flag content).

community standards or terms of service pages[60] as cached snapshots from the Internet Archive's "Wayback Machine."[61] These snapshots offer the ability to observe how the social network's speech guidelines have changed over its 11-year existence. The Wayback Machine generally caches webpages at a frequency of once every few weeks to once every few months, depending on how well linked-to the page is.[62] This chapter will examine every update of Facebook's terms of use from 2004 until March 2015, and four examples of Facebook's community standards from January 2011 to March 2015.[63] Two criteria will guide the analysis of these pages. First, the analysis will look for changes in policies over the course of 11 years. This analysis will require tracking which items get added to newer versions of the standards and subtracted from older versions, as well as noting if/how definitions of certain key terms (such as "hate speech") change over time. Second, several benchmarks will be used to assess how Facebook's community standards balance protection of individuals' speech with prevention of harm. These benchmarks include legal tests for distinguishing protected from unprotected speech (discussed in chapters 3 and 4), as well as Facebook's interests within the networked economy

[60] According to the Internet Archive, Facebook did not create a separate page outlining community standards until 2011. Before then, all stipulations for what constituted allowable versus unallowable content was contained in Facebook's "Terms of Service" page.
[61] *Wayback Machine*, INTERNET ARCHIVE, *available at* http://archive.org/web/.
[62] According to the Internet Archive's website, the Wayback Machine's "automated systems crawl the web every few months or so." Also, "Much of our archived web data comes from our own crawls or from Alexa Internet's crawls. Neither organization has a 'crawl my site now!' submission process. Internet Archive's crawls tend to find sites that are well linked from other sites. The best way to ensure that we find your web site is to make sure it is included in online directories and that similar/related sites link to you." *Frequently Asked Questions*, INTERNET ARCHIVE (Mar. 20, 2015), *available at* http://archive.org/about/faqs.php#The_Wayback_Machine.
[63] The terms of service are clearly marked with the date on which they are updated, allowing for an analysis of each update. The community standards are not marked with such a date, and therefore the pages that are analyzed are the first edition, the most recent update (March 15, 2015), and two randomly selected pages in between.

(discussed in chapter 2). Obviously, Facebook's community standards need not be as protective of speech as First Amendment jurisprudence, and the purpose of the analysis is not to make such an obvious argument. Rather, the goal of assessing Facebook's standards vis-à-vis legal standards is to understand which areas of speech Facebook protects or restricts more than others. The objective is to understand what criteria Facebook uses to find the balance between promoting speech and preventing harm.

To answer the second research question, I will compile news reports on Facebook's actions toward extreme content that appears to contravene its community standards. I will analyze these reports inductively, coding the key themes that emerge from reading them.[64] The theoretical framework guiding this analysis is Gillespie's politics of technology theory.[65] Together, the evolution of Facebook's community standards and the way in which news media have covered incidents of removals or non-removals of extreme UGC are part of an ongoing dialogue about how the norms of freedom of expression should be defined in a networked communication environment. In other words, Facebook is the technology, and the debate being analyzed in this study is

[64] The literature on qualitative textual analysis is fragmented and piecemeal (Elfriede Fürsich, *In Defense of Textual Analysis*, 10 JOURNALISM STUDIES 238 (2009). However, the method continues to be an important tool in mass communication research. It has been used to uncover polyvalent media messages, due especially to its essential focus on the context(s) in which media messages are both made and interpreted (NORMAN FAIRCLOUGH, DISCOURSE AND SOCIAL CHANGE (1992); Fürisch (2009); GIOVANNA DELL'ORTO, THE HIDDEN POWER OF THE AMERICAN DREAM: WHY EUROPE'S SHAKEN CONFIDENCE IN THE UNITED STATES THREATENS THE FUTURE OF U.S. INFLUENCE (2008)). The methodology of qualitative textual analysis commonly categorizes findings of media texts into key themes (DELL'ORTO (2008); THOMAS R. LINDLOF AND BRYAN C. TAYLOR, QUALITATIVE COMMUNICATION RESEARCH METHODS 246 (2011)). Often these themes are built from frames, which, in the qualitative literature, regularly take the form of "interpretative packages": "central organizing idea[s]" that give meaning to an issue and "make sense of relevant events" (William A. Gamson and Andre Modigliani, *Media Discourse and Public Opinion on Nuclear Power: A Constructionist Approach*, 95 AM. J. OF SOCIOLOGY, 1, 3 (1989)).

[65] GILLESPIE, *supra* note 25.

how that technology is being defined (by Facebook, itself, and by news media) in terms of that technology's role in facilitating individuals' expressive agency.

This analysis has several broader implications. First, although this analysis (by definition) cannot be generalized to describe all forms of content governance, it can serve as an illustrative example of how norms of freedom of expression are being defined for networked communication via digital intermediaries. Second, this analysis can shed light on the potential extent to which tolerance for extreme speech is changing in a networked communication environment. The findings of this study can lead to two potential hypotheses that can be tested using social scientific methods:[66] 1) Facebook is its own communicative arena that has its own norms of communication; 2) the norms of freedom of expression on Facebook diffuse into greater society, affecting individuals' tolerance for extreme speech outside of Facebook. Third, on a related note to the second broader implication just discussed, the examples from this case can serve as a proxy for assessing another of DeNardis' concerns regarding Internet governance and freedom of expression: the public's perception of what their civil liberties are in a networked communication environment where digital intermediaries have become the arbiters of expression.[67]

Chapter 6: "A Duty to Freedom: Conceptualizing Platform Ethics"

Law and ethics are ultimately related; "both [are] concerned with the advancement of some socially shared vision of the public good."[68] The difference between the two is that the "law sets a minimum standard below which our actions must

[66] *See* section on "Avenues for Future Research" in chapter 7, the concluding chapter of this dissertation.
[67] DENARDIS, *supra* note 19, at 157.
[68] Erik Ugland & Jennifer Henderson, *Who Is a Journalist and Why Does it Matter? Disentangling the Legal and Ethical Arguments*, 22 J. MASS MED. ETHICS 241, 242-3 (2007).

not fall," while "ethics sets a higher standard to which we ought to aspire."[69] The values of freedom of expression in a networked society cannot be defined by legal theory and jurisprudence alone. Ethical theory must contribute to the debate. This chapter begins with the position that "[m]oral decision making is a complex and uncertain business," and that "in any true moral dilemma, acting rightly necessarily involves overriding one or more prima facie duties—duties that would otherwise have moral force."[70] It is argued here that in content governance, intermediaries are faced with a "true moral dilemma": mitigating harm versus upholding the value of mass participation in an environment where individuals can maximize their expressive agency. This chapter is devoted to analyzing the moral dimensions of these two competing social values. The chapter puts forth the argument that digital intermediaries should abide by a primary duty to promoting freedom and mass participation in the global networked public discourse. The duty to prevent harms associated with online speech should be a secondary (though not completely unimportant) duty, which intermediaries should honor by following clear criteria that govern the definition, identification and means for dealing with harmful speech. This chapter applies these principles to Facebook's community standards discussed in chapter 5.

This chapter also has a broader goal: pushing the boundaries of the field of mass communication ethics by moving the ethical analysis outside the traditional practice of applying ethical principles to journalistic contexts. Digital intermediaries are not

---

[69] Michael Perkins, *International Law and the Search for Universal Principles in Journalism Ethics*, 17 J. MASS MED. ETHICS 193, 195 (2002).
[70] Christopher Meyers, *Reappreciating W. D. Ross: Naturalizing Prima Facie Duties and a Proposed Method*, 26 J. MASS MED. ETHICS 316, 318 (2011).

journalistic institutions, and they should not be submitted to a journalism-ethics analysis as if they were such institutions. Rather, a new understanding must be developed and theorized surrounding the ethical dilemmas facing digital intermediaries in the context of content governance. I propose that this emerging field be called "platform ethics."

Chapter 7 will conclude the study with suggestions on avenues for future research, as well as a broad discussion of how this study fits into the tradition of blending legal and social scientific research in the field of mass communication law.

The main goal of this study is to increase awareness about the values of freedom of expression, and to shed light on a trend (content governance) that has the potential to threaten those values. Thus, it is hoped that this study can provide some criteria upon which people can make sound judgments about issues involving content governance. If YouTube blocks a video from being shown in a certain part of the world, Facebook takes down a page based on popular demand, or Twitter deletes an account, the individuals who use these platforms need to be able to judge with confidence whether the move was desirable or not. Such criteria make up the new civic literacy necessary for deliberative democracy in a networked communication environment.

Chapter 2: Conceptualizing private governance in a networked society: An analysis of trends in and scholarship on content governance

Introduction

The purpose of this study is to explicate the concept of content governance: the control that digital communication intermediaries exercise over user-generated content (UGC). The particular focus of this explication is the governance of extreme UGC. Two key questions guide this explication: How and why do digital communication intermediaries respond to extreme UGC? What are the potential implications of their responses for public discourse in a system of networked communication?

Answering these questions requires a greater understanding of the structure and function of public discourse within networked communication. The emergence of the "networked public sphere"[1] has afforded individuals enormous potential to simultaneously create and consume content that is at once political, cultural, social and commercial in nature.[2] This system of networked communication has "produced a quantitative change in the number of entry points to the sphere of highly distributed expression such that … [t]he nation state has lost its complete control as the administrator of the freedom of expression."[3] Meanwhile, the power of the digital intermediaries that facilitate this networked communication environment has increased relative to individuals as the latter have become dependent on the former to exercise their creative

[1] YOCHAI BENKLER, THE WEALTH OF NETWORKS: HOW SOCIAL PRODUCTION TRANSFORMS MARKETS AND FREEDOM (2006).
[2] AXEL BRUNS, BLOGS, WIKIPEDIA, SECOND LIFE, AND BEYOND 19 (2008).
[3] Ejvind Hansen, *Freedom of Expression in Distributed Networks*, 10 TRIPLEC 741, 743 (2012).

agency.[4] Thus, although the right of individuals to freedom of expression is still defined vis-à-vis state actors, the functions of freedom of expression—distribution of content and access to information—depend on digital intermediaries in a networked environment.[5] Digital intermediaries allow more people to enjoy the functions of freedom of expression than perhaps at any time in history.[6] Therefore, put simply, the sheer ability of intermediaries to control these functions threatens individuals' ability to realize them.

In this context, threats to freedom of expression primarily come from two sources: state actors leaning on private actors to restrict expression;[7] and intermediaries restricting speech in reaction to complaints by users,[8] likely out of a concern for their own business interests.[9] According to law professor Yochai Benkler, if "constraint"—over any activity, but here referring to control over speech—is defined simply in terms of the effect entities have on reigning in the "relative capacity of individuals to be the authors of their lives," then "whether the sources of constraint are private actors or public law is irrelevant."[10] Within this framework of thinking, individuals have pride of place; any way in which individual communicative agency is unduly restricted is considered an undesirable outcome.

---

[4] *See, e.g.*, ACCESS CONTROLLED: THE SHAPING OF POWER, RIGHTS, AND RULE IN CYBERSPACE (Ronald Deibert, John Palfrey, Rafal Rohozinski and Jonathan Zittrain eds., 2012), at 7; JACK GOLDSMITH AND TIMOTHY WU, WHO CONTROLS THE INTERNET? ILLUSIONS OF A BORDERLESS WORLD 70 (2006); José van Dijck, *Users like you? Theorizing agency in user-generated content*, 31 MED. CULT. & SOC'Y 41, 54 (2009).
[5] LAURA DENARDIS, THE GLOBAL WAR FOR INTERNET GOVERNANCE (2014).
[6] *See, e.g.*, ELIOT KING, FREE FOR ALL: THE INTERNET'S TRANSFORMATION OF JOURNALISM (2010); BENKLER, *supra* note 1; BRUNS, *supra* note 2.
[7] Hansen, *supra* note 3, at 742; DENARDIS, *supra* note 5, at 213.
[8] Kate Crawford and Tarleton Gillespie, *What is a flag for? Social media reporting tools and the vocabulary of complaint*, NEW MEDIA & SOC'Y 1 (2014).
[9] DENARDIS, *supra* note 5, at 158.
[10] BENKLER, *supra* note 1, at 141.

However, exactly what constitutes "undue" restrictions is a major point of contention. The networked communication environment has given individuals the potential to create content that is harmful and destructive, leading some scholars to call for greater action by both state actors and digital intermediaries to mitigate these harms.[11] Individuals, digital intermediaries and state actors are thus in the midst of a struggle over how to define the norms of the term "freedom of expression" in a networked communication environment.[12] The concept of content governance is at the very center of this battle.

Governance of UGC is not only about control of the technologies that facilitate individual agency, but also assigning meaning to those technologies.[13] These technologies must be analyzed through a "techno-social lens,"[14] which sees society and technology as co-determining[15] and seeks to understand the human values programmed into technology.[16] The purpose of this chapter is to review scholarship on individual agency and intermediary control in an era of networked communication, and synthesize this literature to better understand how the concepts of freedom of expression and control

---

[11] *See, e.g.*, Mary Anne Franks, *Unwilling Avatars: Idealism and Discrimination in Cyberspace*, 20 COLUM. J. GENDER & L. 224 (2011); AMY GAJDA, THE FIRST AMENDMENT BUBBLE: HOW PRIVACY AND PAPARAZZI THREATEN A FREE PRESS (2015); DANIELLE KEATS CITRON, HATE CRIMES IN CYBERSPACE (2015).
[12] Manuel Castells, *Communication, Power and Counter-Power in the Network Society*, 1 INT'L J. OF COMM. 238, 258 (2007).
[13] Ganaele Langlois, *Participatory Culture and the New Governance of Communication: The Paradox of Participatory Media*, 14 TELEVISION & NEW MEDIA 91, 100 (2013); LAWRENCE LESSIG, CODE, VER. 2.0, 293 (2006).
[14] Hector Postigo, *Cultural Production and the Digital Rights Movement*, 15 INFO., COMM. & SOC'Y 1165, 1171 (2012).
[15] Leah Lievrouw, *New Media, Mediation, and Communication Study*, 12 INFO., COMM. & SOC'Y 303, 310 (2009).
[16] *See, e.g.*, TARLETON GILLESPIE, WIRED SHUT: COPYRIGHT AND THE SHAPE OF DIGITAL CULTURE (2007); Tarleton Gillespie, *The Politics of Platforms*, 12 NEW MEDIA & SOC'Y 347 (2010).

are being defined in this environment. The scope of this analysis is broad and focused at the institutional level, examining the potential that intermediaries and state actors have to control UGC and thereby shape norms of freedom of expression. Although content governance is primarily framed in terms of "private control over the flows of information and access to knowledge,"[17] the same principles gleaned from this study can—and should—also apply to understanding how extreme UGC is governed within the networked communication environment. Indeed, *greater* attention should be paid to how extreme UGC is governed because such speech can implicate important social and political issues, and because its potentially harmful nature can trigger its removal from the public discourse all too easily.[18]

The first section of this chapter reviews scholarship on individual agency and dependence in the context of networked communication. One goal of this analysis is to examine the debate over how power relationships among individuals, state actors and digital intermediaries have been "renegotiated" in today's networked communication system.[19] This process of renegotiation recasts the overarching theme of content governance as one of *inter*dependence among these three stakeholders, while acknowledging that theoretical power imbalances appear to exist within certain contexts of this system. The second goal of this section is to highlight that such power imbalances reinforce the argument that the networked communication system is a contested space,

---

[17] Langlois, *supra* note 13, at 93; Arne Hintz, *Challenging the Digital Gatekeepers: International Policy Initiatives for Free Expression*, 2 J. OF INFO. POL'Y 128 (2012).
[18] Ronald Deibert and Rafal Rohozinski, *Beyond Denial: Introducing Next-Generation Information Access Controls*, *in* DEIBERT, ET AL, *supra* note 4, at 4.
[19] van Dijck, *supra* note 4, at 46.

and that content governance plays an important role in how state actors, digital intermediaries and individuals define the norms of networked communication.

The second section of this chapter will place the concept of content governance of extreme UGC within the literature of so-called "affirmative" First Amendment theory, which considers the ultimate purpose of freedom of expression to be the maximization of individual participation within the public discourse.[20] The importance of incorporating this literature into the discussion is to introduce the concept of content governance to the many ideals of freedom of expression that come from the diverse body of scholarship on First Amendment theory. Affirmative First Amendment theory is the ideal place to start because it recognizes the interdependence of state, corporate and individual stakeholders within systems of mass communication. Namely, this body of theory envisions the primary purpose of state actors as curbing corporate power over the mass communication system in an effort to enhance the agency and participation of a diverse group of individuals within the system. In other words, this body of theory fits well with a "private governance perspective" that undergirds the field of Internet governance, and thereby content governance. This perspective treats "distinctions between public and private spheres as doubtful rather than as given" when it comes to an analysis of the power each sphere has over the communicative agency of individuals.[21]

---

[20] *See, e.g.*, CASS SUNSTEIN, DEMOCRACY AND THE PROBLEM OF FREE SPEECH (1993); OWEN FISS, THE IRONY OF FREE SPEECH (1996); Jack M. Balkin, *Digital Speech and Democratic Culture: A Theory of Freedom of Speech for the Information Society*, 79 N.Y.U. L. REV. 1 (2004).
[21] Stuart Macaulay, *Private Government, in* LAW AND SOCIAL SCIENCES (Lipson, L & Wheeler, S eds., 1986), 445-518, at 446.

A second important reason for incorporating this body of literature into the discussion of content governance is that it will aid the transition into the analysis in chapter three of the values of extreme speech as defined by so-called "negative" First Amendment theories. These theories conceive of freedom of expression as a negative right that individuals have against the government, thereby viewing state actors as categorical enemies to individual freedom of expression who must remain neutral when crafting any law or policy that would affect that freedom.[22] Such laws or policies invariably include efforts to curb corporate control over individuals' ability to enter into the public discourse championed by affirmative First Amendment theorists (such as through a right of access to broadcast media).[23] Generally, negative First Amendment theories contend that something of value is lost when government—no matter how well intentioned and no matter how extreme the speech (save for several rare exceptions)—attempts to dictate the norms of freedom of expression.[24]

Overall, the purpose of this chapter is to construct a theoretical framework with which to analyze the concept of content governance. Therefore, some of the concepts will necessarily be discussed in the abstract. The goal is for this framework to aid the empirical analysis of Facebook's community standards and practices of content governance in chapter 5.

---

[22] *See, e.g.*, Ronald A. Cass, *The Perils of Positive Thinking: Constitutional Interpretation and Negative First Amendment Theory*, 34 UCLA L. REV. 1405 (1987).
[23] *See, e.g.*, Jerome A. Barron, *Access to the Press: A New First Amendment Right*, 80 Harv. L. Rev. 1641, 1656 (1967).
[24] Cass, *supra* note 22.

Agency, Dependence and Contested Space in Networked Communication

Conceptualizing Content Governance in an Age of Individual Agency

The concept of content governance resides in a web of contexts and scholarly literatures.[25] The primary objective of this chapter is to establish the location of, and place boundaries around, the study of content governance within the broader field of private governance of communication. The literature of this broader field is not new. Scholars, predominantly from the critical-cultural vein of the field of mass communication with roots in philosophers Karl Marx[26] and Michel Foucault,[27] have contended that powerful private media corporations make up a regime of hegemonic control over the ability of individuals to participate in public discourse.[28] Meanwhile, scholars who hold a post-positivistic worldview readily criticize these theories for their relative lack of empirical proof regarding this claim.[29] This paradigmatic battle persists in the study of private governance in Internet communication, to the detriment of studying the concept of content governance. The goal of this chapter is to answer the call by communication professor José van Dijck to carefully synthesize theories from a broad spectrum of literature, both within and outside the field of mass communication, to

---

[25] van Dijck, *supra* note 4, at 42.
[26] *See, e.g.*, Karl Marx, *On Freedom of the Press: Censorship,* REINISCHE ZEITUNG, (May 15, 1842) (Brian Baggins and Sally Ryan, translators), *available at* <http://www.marxists.org/archive/marx/works/1842/free-press/ch05.htm>.
[27] *See, e.g.*, Michel Foucault, *The Subject and Power*, 8 CRITICAL INQUIRY 777 (1982).
[28] *See* ROBERT M. ENTMAN, DEMOCRACY WITHOUT CITIZENS: MEDIA AND THE DECAY OF AMERICAN POLITICS (1989); BEN BAGDIKIAN, THE NEW MEDIA MONOPOLY (2004); Daniel C. Hallin*, Hegemony: The American News Media from Vietnam to El Salvador, A Study of Ideological Change and its Limits*, *in* POLITICAL COMMUNICATION: APPROACHES, STUDIES, ASSESSMENTS, (David L. Paletz ed., 1986).
[29] *See, e.g.*, Melvin DeFleur, *Where Have All the Milestones Gone? The Decline of Significant Research on the Process and Effects of Mass Communication*, 1 MASS COMM. & SOC'Y 85 (1998); Annie Lang, *Discipline in Crisis? The Shifting Paradigm of Mass Communication Research*, 23 COMM. THEORY 10 (2013).

explicate the concept of content governance and lead scholars to a better understanding of the implications surrounding the concept.[30]

Content governance is nested within a relatively recent conception of the field of private governance of communication: Internet governance.[31] Internet governance is a broad field that encompasses issues such as data privacy, net neutrality, deep packet inspection, policing of child abuse images and any other issue involving how governmental or non-governmental actors constrain or regulate certain aspects of the Internet.[32] The object of study that connects these disparate topics is "the design and administration of the technologies necessary to keep the Internet operational[,] and the enactment of substantive policy around these technologies."[33] Broadly, the field of Internet governance recognizes a trend toward "privatization of authority" regarding key features of Internet technology,[34] and a big debate in the field revolves around the definition of that private authority. Much scholarly focus has been on international institutions involved in the global governance of the Internet, such as the Internet Corporation for Assigned Names and Numbers (ICANN), the Internet Governance Forum (IGF) and the World Summit on the Information Society (WSIS).[35] However,

---

[30] van Dijck, *supra* note 4.
[31] DENARDIS, *supra* note 5.
[32] *Id*.; Malte Ziewitz and Christian Pentzold, *In Search of Internet Governance: Performing Order in Digitally Networked Environments*, 16 NEW MEDIA & SOC'Y 306, 307 (2014); Michel JG van Eeten and Milton Mueller, *Where Is the Governance in Internet Governance?* 15 NEW MEDIA & SOC'Y 720, 723 (2013).
[33] DENARDIS, *supra* note 5, at 6.
[34] Deibert and Rohozinski, *supra* note 18, at 12; *see also* REBECCA MACKINNON, CONSENT OF THE NETWORKED: THE WORLDWIDE STRUGGLE FOR INTERNET FREEDOM (2012), at xxii.
[35] Laura DeNardis, *The Social Media Challenge to Internet Governance*, *in* SOCIETY AND THE INTERNET: HOW INFORMATION AND SOCIAL NETWORKS ARE CHANGING OUR LIVES (William Dutton and Mark Graham eds., 2014); Dawn C. Nunziato, *Freedom of Expression, Democratic Norms, and Internet Governance*, 52 EMORY L. J. 187 (2003).

some scholars have called for the label "Internet governance" to be applied more broadly to include studies on the "many real-world activities that actually shape and regulate the way the Internet works."[36]

This call by Internet governance scholars Michel JG van Eeten and Milton Mueller to expand the field of Internet governance is important because it opens the field up to studying the central role that digital intermediaries play in governing how individuals use networked communication to participate in a global public discourse.[37] In particular, the digital intermediaries that are the focus of this analysis are "single-firm industry platforms" that facilitate networked communication activities.[38] The metaphor of the platform is appropriate for describing these intermediaries. They prop up individual users, facilitating their agency;[39] yet the UGC that is published on those platforms ultimately is compiled in relation to a proprietary message set by the intermediary.[40] The companies that own these platforms determine the design and norms of the communicative activities that take place on them. However, companies do not create these norms in a vacuum. They also will respond to "pressures from … users that they choose to respect" when creating their terms of use—including their speech policies or "community standards."[41] Therefore, the process of creating norms that govern speech on

[36] van Eeten and Mueller, *supra* note 32, at 721.
[37] Langlois, *supra* note 13, at 93.
[38] K.C. Claffy and David D. Clark, *Platform Models for Sustainable Internet Regulation*, 4 J. OF INFO. POL'Y 463 (2014).
[39] Langlois, *supra* note 13, at 94.
[40] Joseph B. Walther and Jeong-woo Jang, *Communication Processes in Participatory Websites*, 18 J. OF COMPUTER-MEDIATED COMM. 2, 4 (2012).
[41] Claffy and Clark, *supra* note 38, at 466.

these platforms is somewhat dialogical in nature, though the exact extent to which individuals have a say over how these norms are created remains an open question.

These two concepts—agency and its facilitation—are the poles that bracket the battle for the meaning of freedom of expression in a networked society. Today's networked system of communication empowers individuals on an unprecedented scale.[42] Scholars have identified several factors that account for such unprecedented empowerment. First, the networked structure of the system increases agency by exponentially increasing the size of the audience that individuals are able to reach. For example, this networked structure has the potential to quickly turn individual agency into collective action, thereby affording enormous power to social movements.[43] Second, platforms offer user-friendly design, ensuring that more individuals with relatively low levels of technical literacy have the ability to contribute to online discourse.[44] In other words, the days of needing coding skills as a ticket to simply enter the online public discourse are long gone. The networked public sphere thereby has expanded. Third, the informal nature of production and consumption of content has become normalized.[45] The key characteristic of the system is that small-scale, amateur cultural production is

---

[42] Deibert and Rohozinski, *supra* note 18, at 3.
[43] Taso G. Lagos, Ted M. Coopman and Jonathan Tomhave, '*Parallel Poleis': Towards a Theoretical Framework of the Modern Public Sphere, Civic Engagement and the Structural Advantages of the Internet to Foster and Maintain Parallel Socio-Political Institutions*, 16 NEW MEDIA & SOC'Y 398, 409 (2014); Andrew J. Flanagin, Craig Flanagin and Jon Flanagin, *Technical Code and the Social Construction of the Internet*, 12 NEW MEDIA & SOC'Y 179, 182 (2010).
[44] Flanagin, Flanagin and Flanagin, *supra*, at 184; *see also* Ganaele Langlois, Fenwick McKelvey, Greg Elmer and Kenneth Werbin, *Mapping Commercial Web 2.0 Worlds: Towards a New Critical Ontogenesis*, 14 THE FIBRECULTURE J. 1, 3 (2009).
[45] Ramon Lobato, Julian Thomas & Dan Hunter, *Histories of User-Generated Content: Between Formal and Informal Media Economies*, 5 INT'L J. COMM. 899, 909 (2011).

becoming more visible and more institutional.[46] Therefore, individuals are asserting their position as key players in this communicative environment. Fourth, networked communication harnesses the "necessarily participatory" nature of this creation of culture.[47] For Benkler, "[h]ow culture is produced is … an essential ingredient in structuring how freedom and justice are perceived, conceived, and pursued."[48] Thus, freedom of expression, democratic participation and the creation of culture are all interdependent of one another, and all are activated in a system of networked communication.

However, some scholars remain skeptical of the extent to which individuals actually have communicative agency, and—even if they do have greater agency—the extent to which that agency actually makes a difference in public discourse. These scholars talk of a psychological rather than a real or empirically observable empowerment among individuals who participate in the system of networked communication.[49] Individuals have a "sense of agency" or a "sense of empowerment," they argue, rather than any actual ability to alter public policy through their participation in the public discourse.[50] Communication scholar Matthew Hindman argues that one must study how the Internet has redistributed power to multiple stakeholders, not simply (or specially) to individuals.[51] He laments the "popular enthusiasm" for the revolutionizing potential of technology, which he argues "has made a sober appraisal of

[46] *Id*. at 900.
[47] Postigo, *supra* note 14, at 1688.
[48] BENKLER, *supra* note 1, at 274.
[49] Louis Leung, *User-Generated Content on the Internet: An Examination of Gratifications, Civic Engagement and Psychological Empowerment*, 11 NEW MEDIA & SOC'Y 1327, 1329 (2009).
[50] Flanagin, Flanagin and Flanagin, *supra* note 43, at 186.
[51] MATT HINDMAN, THE MYTH OF DIGITAL DEMOCRACY (2009).

the Internet's complicated political effects more difficult."[52] In the context of journalism, mass communication professor Brendan Watson has documented that Twitter users did not provide any alternative perspectives to mainstream news media's coverage of the 2010 Deep Water Horizon oil spill that ravaged the U.S. Gulf Coast.[53] This finding questions the notion that individuals categorically will create new meaning within the public discourse when afforded the tools of networked communication. All told, the message of these scholars is that the extent to which networked communication affects individual communicative agency or individuals' capacity to shape or alter public discourse should not be thought of in extremes: networked communications are neither revolutionary nor are they undergoing a process of normalizing the status quo.[54] Rather, user agency is complex, and the only way to understand it better is through careful, nuanced study.[55]

Dependence

Individuals around the world are becoming increasingly reliant on cyberspace as their main source of consuming and sharing information,[56] and digital intermediaries are playing an increasingly indispensible role in facilitating the communicative agency of individuals.[57] Within this context of dependency, scholars seek to understand the role that

---

[52] *Id*. at 5.
[53] Brendan R. Watson, *Is Twitter an Alternative Medium? Comparing Gulf Coast Twitter and Newspaper Coverage of the 2010 BP Oil Spill*, 42 COMM. RESEARCH n.p. (2015).
[54] Scott Wright, *Politics as Usual? Revolution, Normalization and a New Agenda for Online Deliberation*, 14 NEW MEDIA & SOC'Y 244 (2012).
[55] van Dijck, *supra* note 4, at 42.
[56] Deibert and Rohozinski, *supra* note 18, at 7; *see also* Nunziato, *supra* note 35.
[57] GOLDSMITH AND WU, *supra* note 4, at 70.

platforms play in "steering" or "channeling" the agency of their users.[58] This process of channeling agency is the primary locus for the battle over the meaning of agency and freedom of expression in a networked communication environment, pitting powerful media against powerful speakers/audiences.[59] How these concepts—agency and freedom of expression—are defined within this environment could have great effect on the nature of democratic discourse. Communication scholar Ganaele Langlois sees the "agency-dependency" trade-off as having the potential to "pervert the very democratic ideals of free and unfettered communication on which the Internet is based."[60] Other scholars agree, contending that "the very technical features that currently appear to engender relative freedom can also be employed to exert strict control."[61]

Especially germane to this study is the part of Internet governance that deals with "the evolution of the technical and transactional infrastructures concealed beneath content and how these infrastructures potentially constrain the future of individual civil liberties" such as freedom of expression.[62] Communication professor Laura DeNardis contends that "individual freedom of expression [is] dependent on online infrastructures and the policies enacted to preserve both liberty and infrastructure reliability."[63] Specifically, individual expression has found itself in an environment where its ability to enter and make an impact on global public discourse is dependent on digital intermediaries, which

[58] van Dijck, *supra* note 4, at 43; Langlois, *supra* note 13, at 102.
[59] Lievrouw, *supra* note 15, at 307.
[60] Langlois, *supra* note 13, at 95.
[61] Flanagin, Flanagin and Flanagin, *supra* note 43, at 188.
[62] DeNardis, *supra* note 35, at 348.
[63] DENARDIS, *supra* note 5, at 17.

"have become the front lines of … governance issues in cyberspace."[64] Studying these intermediaries is essential to understanding their role in shaping norms of freedom of expression, as they "have increasingly become the arbiters of online expressive liberty."[65] This trend, DeNardis contends, "is highly controversial, contextually dependent, and … evolving."[66] She argues that such private mediation is ultimately a major concern: it "constrains what individuals can express because it requires permission and administration by an information intermediary."[67] DeNardis posits that governance of UGC published on digital platforms can take three forms, to which this study adds a fourth.

Discretionary Governance

First, commercial intermediaries may voluntarily exclude or remove from its platforms certain types of extreme speech. DeNardis calls this action "discretionary" governance.[68] This type of governance is rare, as intermediaries typically remove content in response to some form of direct pressure (from individuals or governments, see below).[69] However, Google notably followed this type of governance after the YouTube video "Innocence of Muslims" sparked violent protests throughout the Muslim world in

[64] *Id*. at 156.
[65] *Id*. at 157.
[66] *Id*. at 172.
[67] DeNardis, *supra* note 35, at 355.
[68] DENARDIS, *supra* note 5, at 158. DeNardis uses the term "discretionary *censorship*," but this study chooses to replace "censorship" with "governance" for two reasons. The first reason comes from a desire to highlight the central place of the term "governance" in this analysis, as well as its connection to scholarship on private governance. Second, the term censorship is problematic because censorship typically denotes state control over speech.
[69] *See* Crawford and Gillespie, *supra* note 8.

September 2012.[70] Despite its rarity, this type of governance is important to understand because it illustrates the unbridled potential that intermediaries have to restrict individuals' speech on their own volition, as well as the lack of a means to hold intermediaries accountable for potential abuses of this power.

The control that intermediaries have over online speech on their own may not be as insidious or anti-democratic as the control that state actors have in collusion with intermediaries.[71] Nevertheless, following Benkler's broad definition of power and constraint discussed above,[72] intermediary control over speech has the potential to threaten online public discourse.[73] Therefore, it needs to be better understood. Some scholars have adopted the metaphor of gatekeeping—once reserved most prominently for the editorial and information selection process of journalism[74]—to describe the function of intermediaries as the ultimate deciders of not only what information individuals can access, but also what UGC gets into the public discourse.[75] This metaphor is helpful because, at bottom, gatekeeping is an example of a "regime of control."[76] Indeed, the challenge facing journalists of how to sift through and "curate"[77] the abundance of information and UGC created in a networked environment provides a direct analogy for studying the same challenge faced by digital intermediaries. Journalism professor Jane Singer talks of the "secondary gatekeeping" function that online news sites play to judge

[70] DENARDIS, *supra* note 5, at 158.
[71] Deibert and Rohozinski, *supra* note 18.
[72] BENKLER, *supra* note 1.
[73] Ethan Zuckerman, *Intermediary Censorship*, *in* DEIBERT, ET AL, *supra* note 4, at 71.
[74] *See, e.g.*, PAMELA J. SHOEMAKER AND TIMOTHY VOS, GATEKEEPING THEORY (2009).
[75] Hindman, *supra* note 51, at 12.
[76] Jane B. Singer, *User-Generated Visibility: Secondary Gatekeeping in a Shared Media Space*, 16 NEW MEDIA & SOC'Y 55, 56 (2014).
[77] *See, e.g.,* Gillespie, *supra* note 16.

the value and quality of UGC in terms of its merit for redistribution.[78] The factors that weigh into this decision-making process include the appropriateness of the UGC and the effect that redistributing it will have on the news organization's bottom line.

Similarly, communication professors Josh Braun and Tarleton Gillespie study the gatekeeping function of online news sites in controlling user comments, which are often "unpolished, wide-ranging, and unpredictable."[79] In "imposing and justifying policies for managing what is sometimes an unruly dialogue," these news sites must ensure that their policies "not only be practical and enforceable, but also balance the economic, professional, and ideological aspirations of the news organization."[80] Importantly, "[t]he content policies and their enforcement must toe the line between avoiding legal liability, keeping an eye on the economic bottom line, and some kind of commitment to protecting their users' freedom of speech and the vibrancy of the public discourse they produce."[81] With all of these factors to consider, "[d]iscerning between valuable and invaluable speech, the political from the profane, measuring degrees of hatefulness and harmfulness is hard" for these organizations.[82] These conclusions from the study of news organizations managing readers' comments can be projected onto the broader issue of content governance. Digital intermediaries such as YouTube, Facebook and Twitter face potentially even more difficult decisions in governing UGC, given that the boundaries they are seeking to protect through content governance are broader and less well defined

[78] Singer, *supra* note 76, at 56.
[79] Josh Braun and Tarleton Gillespie, *Hosting the Public Discourse, Hosting the Public*, 5 JOURNALISM PRACTICE 383, 383 (2011).
[80] *Id*. at 384.
[81] *Id*. at 385.
[82] *Id*. at 392.

compared to the boundaries that journalistic institutions seek to protect. Digital intermediaries have their own message and image that they seek to project, yet that message and image must compete with the millions of other messages that individuals publish via these intermediaries every day.

Separating the concepts of intermediary control over UGC and the joint control intermediaries and state actors can impose on UGC—done here solely for the purposes of the present analysis—does not necessarily mean that these are, in fact, two separate concepts that deserve separate fields of study, nor that one type of control is worse than the other. In fact, the opposite is true: studying the control that intermediaries have over UGC will contribute to a greater understanding of their potential to work with state actors to control UGC. The important conclusion to note here is that economic concerns are arguably going to be the most important factor in determining intermediary practices of content governance.

Delegated Governance

Second, the efficiency with which commercial intermediaries can restrict speech offers governments a "back door" to state-sponsored censorship. Public officials, who could never successfully pass a law to ban a certain type of extreme speech, can pressure intermediaries to remove that same type of speech. DeNardis calls this action "delegated" governance over expression.[83] Other scholars have viewed this form of governance as potentially the most worrisome, due to the lack of transparency that exists (at least on the

[83] DENARDIS, *supra* note 5, at 213. As with the term "discretionary censorship," *supra* note 68, DeNardis also refers to this second term as "delegated censorship." I have changed the term to "delegated governance" for the same reasons as above.

part of governments)[84] to inform individuals about the nature and extent of this practice. In their 2012 collection of essays titled *Access Controlled*,[85] political science professor Ronald Deibert and colleagues warn that the threat to free speech today comes not from efforts by governments to directly censor or filter content that its citizens (or subjects) create (which was the subject of their 2008 collection *Access Denied*[86]). Rather, the main threat comes from the insidious nature with which governments and commercial platforms have teamed up to manage such content, such as through both parties entering into an agreement whereby the commercial platforms actively seek out and eliminate content at the behest of the government.[87] Deibert and his co-authors argue that such hybrid private-public governance of content is becoming the new norm in discussions of control over public discourse.[88] They argue that state actors "no longer fear pariah status by openly declaring their intent to regulate and control cyberspace" because they couch their reasons for such control in terms of protecting citizens from harm.[89] Meanwhile intermediaries are likely to heed the pressure from governments to ensure they can turn

[84] Intermediaries, for their part, have been more willing to be transparent about the requests they receive from governments to remove content. *See e.g. Requests to Remove Content from Governments*, GOOGLE TRANSPARENCY REPORT, http://www.google.com/transparencyreport/removals/government/?hl=en; *Removal Requests*, TWITTER TRANSPARENCY REPORT, https://transparency.twitter.com/removal-requests/2014/jan-jun.
[85] DEIBERT, ET AL, *supra* note 4.
[86] ACCESS DENIED: THE PRACTICE AND POLICY OF GLOBAL INTERNET FILTERING (Ronald Deibert, John Palfrey, Rafal Rohozinski and Jonathan Zittrain eds., 2008).
[87] An example of the second type of arrangement is the agreement between UK Prime Minister David Cameron's government and major Internet service providers (ISPs) in the United Kingdom to filter by default all content that the government considers "pornographic." *See The Internet and Pornography: The Prime Minister Calls for Action*, GOV.UK (July 22, 2013), *available at* https://www.gov.uk/government/speeches/the-internet-and-pornography-prime-minister-calls-for-action.
[88] Deibert and Rohozinski, *supra* note 4, at 11-12.
[89] *Id*. at 4.

an operating profit in the countries in question.[90] Rebecca MacKinnon, a former journalist and current project director at the Open Technology Institute, sees this ability of state actors and intermediaries to work together to remove, filter and monitor UGC as having the potential to threaten the ability of individuals to sustain democratic discourse through participation via platforms.[91]

Figure 2-1 presents a model of delegated content governance in the system of networked communication modeled in Figure 1-1 from Chapter 1 (the lines representing continuous communication have been lightened for the sake of clarity). Individuals continue to simultaneously publish and consume speech via an intermediary. However, once an individual publishes a message deemed out-of-bounds by a state actor, the latter will notify the intermediary about the allegedly infringing speech by "flagging" it.[92] Employees of the intermediary will review the speech to determine whether it does, in fact, need to be removed due to its violating a law or its potential disruption of business interests in the polity governed by the state actor. If the intermediary chooses to remove the speech, that speech (represented by the dashed lines), which was once visible to other individuals, will now become invisible via that particular intermediary.

[90] Zuckerman, *supra* note 73, at 80.
[91] MACKINNON, *supra* note 34, at 13-14.
[92] Crawford and Gillespie, *supra* note 8.

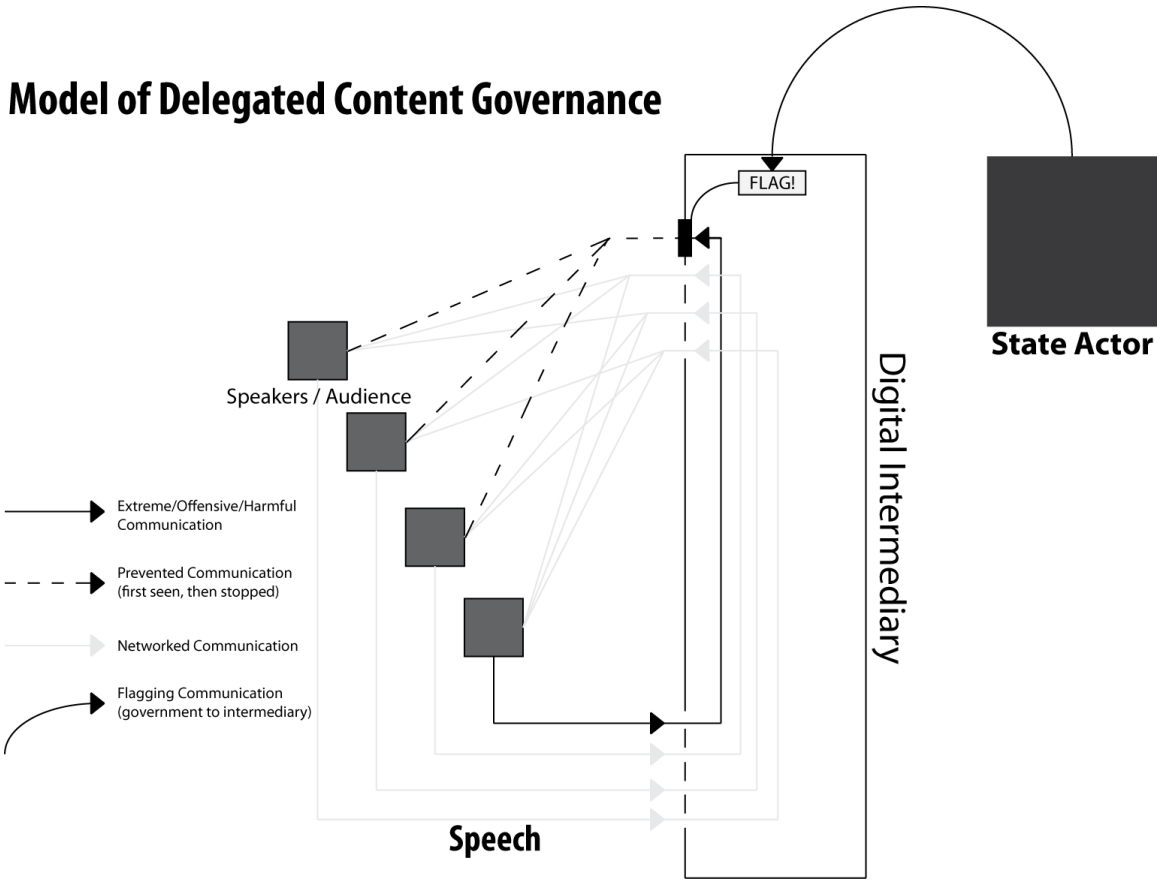

Figure 2-1: Delegated Content Governance

Governance through Legal Compliance

Third, violation of laws can trigger governance over UGC. For example, in the United States, the Digital Millennium Copyright Act (DMCA) governs illegal use of copyrighted material in UGC.[93] Under the DMCA's "notice-and-takedown" regime, rights-holders can give intermediaries a good-faith notification that their copyrighted works are allegedly being published on their platforms without the rights-holders' permission. The intermediary must remove the allegedly infringing material within 10

[93] 17 U.S.C. § 512.

business days to avoid liability for vicarious or contributory infringement.[94] Governments also can notify intermediaries when content that they are hosting is in violation of their nation's laws by sending the intermediaries official letters requesting the removal of allegedly illicit content from a platform.[95] This form of governance is very similar in theory to DeNardis's concept of "delegated" governance from the preceding paragraph. However, in practice, this third form of triggering content governance is distinct because 1) it is (generally speaking) transparent, and 2) it is relatively clear that content is, indeed, in violation of a country's laws. Certainly, this form of governance can be considered less a matter of private governance than simply a matter of intermediaries following the law in the countries in which they do business. However, understanding this form of governance is important because it highlights not only the legal but also the potential market incentive that intermediaries have to remove or block unlawful content in certain countries.

Governance by Crowd

Lastly—and this is the form of governance that this study adds to DeNardis's list—a "crowd" of individuals online may have the ability to force digital intermediaries to exclude or remove extreme or unpopular messages, which is something that legally

[94] 17 U.S.C. § 512(g)(2)(B). The intermediary must notify the alleged rights violators of the takedown (§ 512(g)(2)(C)), and these authors can file a counter notice with the intermediary claiming that their use of the original work was not infringing, (§ 512(g)(3)).

[95] An example of the first type of arrangement is the power of the Brazilian judiciary to send takedown notices to foreign platforms (such as Google's YouTube) requesting the removal of content that is allegedly defamatory or racist. *See* Raphael Spuldar, *On the Ground: São Paulo*, INDEX ON CENSORSHIP (Mar. 20, 2013), *available at* http://www.indexoncensorship.org/2013/03/on-the-ground-sao-paulo/.

they could not do in a traditional public forum such as a town square.[96] This form of governance is arguably the most important of the four to understand because mainstream digital intermediaries (Twitter, Facebook and YouTube) rely on users' "flagging" of undesirable content to be made aware of the content and choose whether to take action against it.[97] Individuals can also pressure intermediaries by publicizing their grievances over the undesirable content. For example, Facebook removed pages with misogynistic titles such as "Dropkicking sluts in the teeth" after a group of feminist activists ran a grassroots campaign pressuring companies to remove advertisements from Facebook if the social networking site did not remove the pages.[98] Figure 2-2 presents a model of this form of content governance. Flagging by individuals follows essentially the same process as flagging by state actors (see Figure 2-1). The obvious difference between the two models is the fact that in the former model, the flagging comes from within the community of speakers/audience served by the intermediary rather than from outside the community. Also, rather than having to make a decision on whether the flagged speech violates a particular law, intermediaries in the scenario modeled in Figure 2-2 must decide whether the flagged speech violates its own set of standards governing speech within the community it serves. This form of governance is important to understand because it involves certain groups pressuring intermediaries into following a particular

[96] *See e.g.* Cantwell v. Connecticut, 310 U.S. 296 (1940); Terminiello v. Chicago, 337 U.S. 1 (1949); Gregory v. Chicago, 394 U.S. 111 (1969); Cox v. Louisiana, 379 U.S. 536 (1965); Edwards v. South Carolina, 372 U.S. 229 (1963); Forsyth County, Ga. v. Nationalist Movement, 505 U.S. 123 (1992).
[97] Crawford and Gillespie, *supra* note 8.
[98] *See Facebook Agreement Statement*, WOMEN, ACTION, & THE MEDIA (May 28, 2013), *available at* http://www.womenactionmedia.org/fbagreement/. For more on the subject of how individuals can leverage platforms to remove controversial UGC by "flagging" the content, *see* Crawford and Gillespie, *supra* note 8.

set of community norms, which may exclude certain forms of extreme or potentially harmful speech.[99] Reaching a normative conclusion on whether such pressure to follow community norms is desirable will require a greater understanding of the competing values at stake in this form of governance: protecting extreme speech and preventing some form of harm.

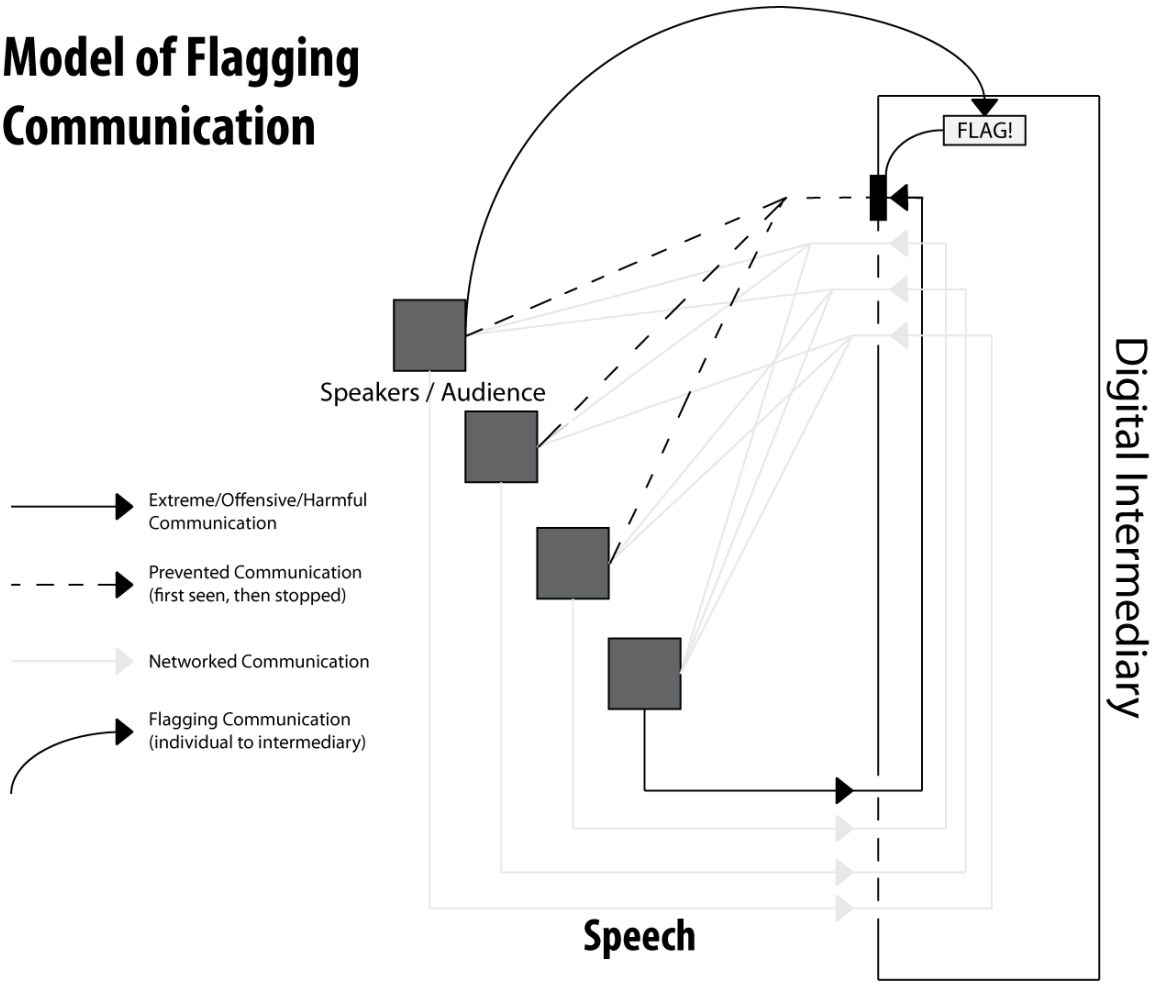

Figure 2-2: Flagging Communication

[99] *See* ROBERT POST, CONSTITUTIONAL DOMAINS 144 (1995) (contending that the main goal of the First Amendment is to promote "critical interaction" among groups with conflicting opinions and values, and that the way it does so is by preventing the government from invoking community norms to silence extreme speech).

Interdependence

Understanding the economic pressures that intermediaries face is important for conceiving of a system in which individuals, intermediaries and state actors are *inter*dependent upon one another, rather than one in which individuals are completely dependent on intermediaries. In the context of the networked economy, neither stakeholder can function very well without the other.[100] Individuals and intermediaries alike have a demand for participatory culture.[101] Motivating production is an important goal for platforms because "giving users more power over content … add[s] business value."[102] Namely, the creation of UGC often generates valuable information as a byproduct.[103]

Intermediaries that facilitate UGC sell the ideal of potential, promise, opportunity and the allure of fame,[104] as well as the prospects of entertainment and play, in exchange for commoditizing users' personal data and their content.[105] Scholars have been critical of this "Faustian trade-off."[106] Langlois and colleagues see it as evidence that "the power dynamics in commercial Web 2.0 [are] both repressive and productive."[107] They decry the proposition that "the ease of communication, connection and exploration of one's

---

[100] James G. Webster, *User Information Regimes: How Social Media Shape Patterns of Consumption*, 104 NW. U. L. REV. 593, 596 (2010).
[101] van Dijck, *supra* note 4, at 42.
[102] *Id*. at 46; *see also* Ute Schaedel and Michel Clement, *Managing the Online Crowd: Motivations for Engagement in User-Generated Content*, 7 J. MED. BUS. STUD. 17, 19 (2010).
[103] Greg Lastowka, *User-Generated Content and Virtual Worlds*, 10 VAND. J. ENT'MT & TECH. L. 893, 895 (2008).
[104] Schaedel and Clement, *supra* note 102, at 22-23.
[105] van Dijck, *supra* note 4, at 50.
[106] Michael Zimmer, *The Externalities of Search 2.0: The Emerging Privacy Threats when the Drive for the Perfect Search Engine Meets Web 2.0*, 13 FIRST MONDAY n.p. (2008).
[107] Langlois, et al, *supra* note 44, at 11.

interests can only take place through agreeing to terms of service and terms of use that allow for [surveillance of personal data] and the commercialization of user-generated content through advertising."[108] Within this system, digital intermediaries that host UGC likely will have a propensity to view offensive UGC as risky due to its potential to alienate users and lead to lost subscriptions, while UGC that infringes on copyright poses the risk of legal liability.[109]

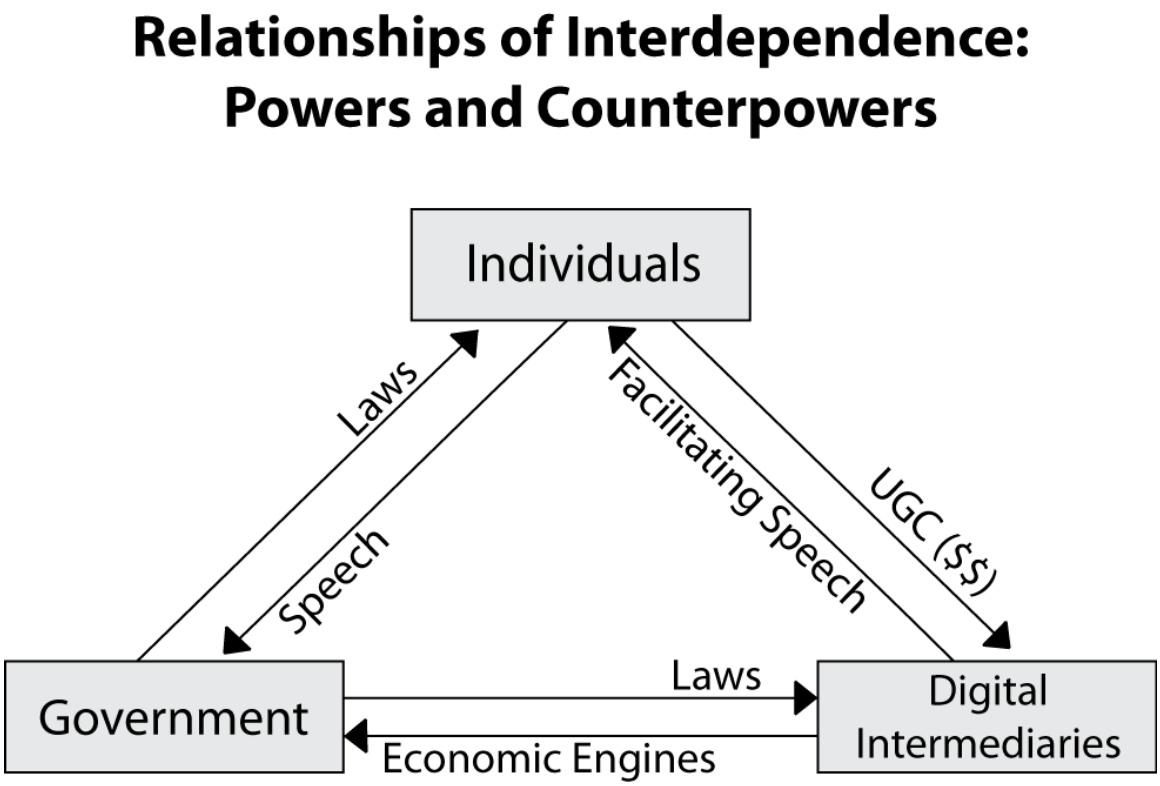

Figure 2-3: Interdependence in Networked Communication

Figure 2-3 presents a model of interdependence in a system of networked communication. The arrows represent pressure that each actor puts on the others.

[108] *Id*. at 2.
[109] Lobato, Thomas and Hunter, *supra* note 45, at 912.

Individuals can act as a checking agent against the government through their speech.[110] However, individuals are dependent on digital intermediaries to be able to realize this checking function, as these intermediaries facilitate individuals' speech. However, individuals provide intermediaries with a source of revenue through the commoditization of their UGC, meaning that intermediaries risk shutting down if the limits they put on individuals' ability to speak are too stringent. Governments can check the checking power of individuals through their ability to set legal boundaries around individuals' ability to speak. This power varies from polity to polity, with state actors in the United States having relatively little power (compared to many other countries) over individuals' speech due to the First Amendment. Governments can attempt to indirectly place restrictions on individuals' speech by putting legal pressure on digital intermediaries to restrict certain types of individuals' speech that gets published on their platforms. Examples of such legal pressure are the various laws around the world that define the parameters of the liability intermediaries face for third-party content. In the United States, Section 230 of the 1996 Communications Decency Act (CDA)[111] gives a relatively high degree of immunity to intermediaries for legal actions arising from third-party content. This high degree of immunity reflects the power that intermediaries exert over governments: intermediaries are economic engines that purportedly function best when unfettered by legal regulations. Congress recognized this power when it passed the CDA, writing in the preamble to the law that its intent was "to preserve the vibrant and

[110] *See, e.g.*, Vincent Blasi, *The Checking Value in First Amendment Theory*, 2 AM. BAR FOUND. RES. J. 521 (1977).
[111] 47 U.S.C. § 230 (1996).

competitive free market that presently exists for the Internet and other interactive computer services, unfettered by Federal or State regulation."[112]

Within this interdependent system, the power of speech is only as strong as its counterpowers allow it to be. In the United States, the First Amendment keeps the power of the government to regulate speech at a relatively constant (and low) level. Therefore, in a system of networked communication in the United States, the entities with the highest potential ability to check the power of individuals' speech are digital intermediaries. Unlike the power of government over individuals, the power of intermediaries is not necessarily constant. Multiple factors may make it wax and wane at any given time, including economic factors, the legal regime determining intermediary liability, and social norms on the acceptability of extreme speech.[113] Just as legal scholars seek to understand the parameters that First Amendment jurisprudence puts on government's ability to restrict speech, mass communication scholars must explore what parameters these extralegal factors place on digital intermediaries' ability to restrict speech. The best place to begin such exploration is within the body of so-called affirmative First Amendment theory.

Agency, Control and Affirmative First Amendment Theory

Content governance, as a subfield of Internet governance, must be discussed and debated among scholars of mass communication law and policy. Such a proposition is not controversial. Rather, it answers a broad call from law professor Enrique Armijo that the issues of networked communication should "now establish the frame within which all of

[112] 47 U.S.C. § 230(b)(2).
[113] LESSIG, *supra* note 13.

our public policy and academic debates concerning communications law and policy take place."[114] However, connecting content governance and mass communication law and policy is a tricky endeavor, and several guidelines must be established from the outset for how to properly synthesize these fields of study.

Although the traditional practice of mass communication law research is often to draw normative conclusions about an issue of freedom of expression based on one theoretical framework,[115] the proposal of this chapter is that one must abstain from drawing normative conclusions about content governance until the concept can be assessed from the perspective of multiple theories of freedom of expression. Scholars should follow such a cautious approach to avoid the pitfalls of adopting normative conclusions that are not based on solid reasoning of First Amendment jurisprudence. For example, some scholars have argued that state actor status should be ascribed to digital intermediaries.[116] The motives of these scholars are to make the actions taken by digital intermediaries to remove users' speech subject to the First Amendment, thereby strengthening the rights of individuals who use these platforms to speak and stripping the rights these intermediaries would otherwise have had to manage their networks. However, such a perspective amounts to reductionism,[117] thereby running the risk of alienating more traditionalist First Amendment scholars. This is no small risk. These

---

[114] Enrique Armijo, *Communication Law, Technological Change, and the New Normal*, 19 COMM. L. & POL'Y 401, 402 (2014).
[115] Matthew D. Bunker and David K. Perry, *Standing at the Crossroads: Social Science, Human Agency and Free Speech Law*, 9 COMM. L. & POL'Y 1 (2004).
[116] Nunziato, *supra* note 35; Sandra Braman & Stephanie Roberts, *Advantage ISP: Terms of Service as Media Law*, 5 NEW MEDIA & SOC'Y 422 (2003).
[117] Matthew D. Bunker, *Imperial Paradigms: First Amendment Theory, Legal Interdisciplinarity and Reductionism*, 3 COMM. L. & POL'Y 515 (1998).

scholars have arguably the greatest understanding of the benefits that extreme speech brings to society, benefits that can come under threat from extreme speech being restrained by private actors as much as state actors. These scholars must be made more aware of the issues of content governance.

Therefore, the proposal put forth here is that scholars should embrace the fact that digital intermediaries are not state actors, but rather powerful media institutions that have the ability to control the speech that individuals publish on their platforms. Such a perspective allows legal scholars to do two things. First, it allows issues of content governance to be assessed using so-called "affirmative" theories of freedom of expression, such as First Amendment scholar Alexander Meiklejohn's self-governance theory,[118] the "new realist" theories of the 1990s,[119] and emerging theories of freedom of expression that are centered on maximizing individual participation in an environment of networked communication.[120] The common threads running through these theories are that they all recognize the control that powerful non-governmental institutions (media corporations) have over individuals' ability to participate in public discourse, and they all recognize the legal potential that state actors have to increase the power of the latter in relation to the former. These theories propose various policies that state actors can follow to realize this goal, most of which are heavily criticized by proponents of so-called "negative" First Amendment theory (discussed in chapter 3). Although the validity and legal soundness of these proposed policies may be questionable, one can still accept the

[118] ALEXANDER MEIKLEJOHN, FREE SPEECH AND ITS RELATION TO SELF-GOVERNMENT (1948).
[119] SUNSTEIN, *supra* note 20; FISS, *supra* note 20.
[120] Nunziato, *supra* note 35; Balkin, *supra* note 20; Marvin Ammori, *First Amendment Architecture*, 2012 WISC. L. REV. 1 (2012).

premise that media institutions have a great deal of power to control public discourse. This premise is the starting place for understanding the interdependent nature of content governance from the perspective of legal theory. Second, by focusing on digital intermediaries as powerful institutions rather than state actors, legal scholars can use the values of freedom of expression distilled from First Amendment theory to assess the potential effects that these intermediaries have over the public discourse without having to subscribe to the faulty doctrinal issues involved in claiming that intermediaries should be considered state actors.

Affirmative Theory

Meiklejohn argues that the First Amendment's ultimate purpose is found "not the words of the speakers, but the minds of the hearers."[121] The government has no control over what people say, but government intervention can ensure that, as Meiklejohn puts it, "everything worth saying shall be said."[122] In other words, freedom of expression is valued for its essential contribution to deliberative democracy. Under this theory, not only do other First Amendment values—such as facilitating individual autonomy[123]—take a back seat to this primary value, these other values can be perceived as threats to the self-governance value. When this is the case, affirmative theorists argue that state actors are best equipped to deal with the threat. Even economics professor Ronald Coase, the free-market champion and founder of the field of law and economics, contends that "the case for government intervention in the market for ideas is much stronger than it is, in

[121] MEIKLEJOHN, *supra* note 118, at 25.
[122] *Id*. at 22.
[123] *See, e.g.*, C. Edwin Baker, *Scope of the First Amendment Freedom of Speech*, 25 UCLA L. REV. 964, (1978); C. EDWIN BAKER, HUMAN LIBERTY AND FREEDOM OF SPEECH (1989).

general, in the market for goods."[124] Put differently, these theorists contend that media corporations have a duty to maintain a diverse and robust public discourse, and if they fail to perform that duty, government not only can but must step in to force them to do so. Not only are affirmative theorists ready to propose state action to remedy threats to democratic participation from private power, they are also apt to diagnose problematic laws that improvidently grant power to private actors. Law professor Cass Sunstein, for example, argues that owners of shopping malls have the ability to exclude certain viewpoints from entering through their doors only because the laws of private property say they do, not because they have an incontrovertible First Amendment right to that ability.[125]

Both of these perspectives are in play when it comes to theorizing the values of and controls over freedom of expression in a networked society. Scholars in the field of "Cyberlaw" have argued about the question of whether the Internet is subject to the laws of the brick-and-mortar world since the Internet first went public.[126] The general consensus today is that laws can and should apply to the Internet in certain situations, but the bigger concern is how non-legal forms of regulation affect the workings of the Internet.[127] Lessig famously argued that social norms, the workings of the market, and the technological design of Internet infrastructure all have the power to regulate Internet

[124] Ronald Coase, *The Market for Goods and the Market for Ideas*, 62 THE AM. ECON. REV. 384, 389 (1974).
[125] SUNSTEIN, *supra* note 20, at 44.
[126] *See, e.g.*, Goldsmith and Wu, *supra* note 4; Lastowka, *supra* note 103; Andrea Braithwaite, *'Seriously, Get Out': Feminists on the Forums and the War(craft) on Women*, 16 NEW MEDIA & SOC'Y, 703 (2014).
[127] Ziewitz and Pentzold, *supra* note 32, at 312.

activities with potentially greater effect than law.[128] This shift of focus to the "law-like effects"[129] of these factors "systematically treat[s] technical decision-making as a source of Internet policy that, in its effects if not its source, now interpenetrates legal policy-making to create the communicative and informational environment in which we live."[130]

An example of this interpretation is Section 230 of the CDA, which has increased the salience among scholars of the "law-like effects" of intermediaries' ability to control UGC.[131] The statute stipulates that digital intermediaries that host content created by third parties shall not be treated as the publisher or speaker in regards to that content, and it grants such intermediaries immunity from civil liability if they voluntarily restrict access to or remove UGC, regardless of whether the UGC is constitutionally protected. The intent of Section 230 was to foster free speech, not stifle it by incentivizing intermediary control. However, the ability to choose whether to take action over UGC, without fear of legal liability in either instance, is a very powerful legal subsidy that digital intermediaries enjoy. Law professor Rebecca Tushnet, following the philosophy of law professor Jerome Barron[132] and Sunstein,[133] sees Section 230 as providing "dominant providers [that have] substantial market control"[134] with "substantial concentrations of

---

[128] LESSIG, *supra* note 13.
[129] Sandra Braman, *The Interpretation of Technical and Legal Decision-Making for the Internet*, 13 INFO., COMM. & SOC'Y 309, 309 (2010).
[130] *Id*. at 311.
[131] DAWN C. NUNZIATO, VIRTUAL FREEDOM: NET NEUTRALITY AND FREE SPEECH IN THE INTERNET AGE 36 (2009).
[132] Barron, *supra* note 23.
[133] SUNSTEIN, *supra* note 20.
[134] Rebecca Tushnet, *Power Without Responsibility: Intermediaries and the First Amendment*, 76 G.W. L. REV. 986, 994 (2008).

power over public discourse."[135] Because intermediaries "do not generally compete to protect user rights,"[136] Tushnet argues that limiting intermediaries' legal liability over UGC should require a concomitant limiting of their power to control speech.[137]

Exactly how such power should be limited has been the subject of much debate among scholars of Internet law. One approach argues that digital intermediaries should be treated as public forums.[138] The theory behind this approach is that digital intermediaries perform a service like that of a public park or square; in both environments, people use the service to publicly express their views. The implication of this approach is that any kind of regulation of speech on these platforms would become subject to First Amendment analysis. However, this approach has not been victorious in court. In *Cyber Promotions v. AOL*,[139] the U.S. District Court for the Eastern District of Pennsylvania held that AOL's email service was not the "functional equivalent" to a public forum. In other words, AOL was not acting as an agent supplying a forum for communication that state actors would normally make available. The court also held that AOL, unlike cable systems, did not control the "critical pathway" of communication, and thus the government could not regulate it[140]—unlike what the U.S. Supreme Court said government could do in 1994 with mandating "must-carry" provisions on cable providers.[141] Rather, the court held that AOL was one of multiple pathways to publishing information online. Some scholars extol this holding. Law professor Eric Goldman warns

[135] *Id*. at 993.
[136] *Id*. at 1004.
[137] *Id*. at 1009.
[138] Braman and Roberts, *supra* note 116; Nunziato, *supra* note 131.
[139] Cyber Promotions v. AOL, 948 F. Supp. 436 (E.D. Pa. 1996).
[140] *Id*. at 455.
[141] Turner Broadcasting System, Inc. v. FCC, 512 U.S. 622, 657 (1994).

that one should not "instinctively react negatively and emotionally to the specter of censorship," a term he uses to refer to the ability of companies that facilitate so-called virtual worlds (such as "Second Life") to remove or restrict the creations of individuals within those worlds.[142] "Converting private … providers into state actors could, paradoxically, limit speech rather than increase it," he writes.[143] "The enemy is not a vendor's private censorship of a customer, however irrational or short-sighted that may be. The real enemy is an emotional response to private censorship that trumps sound policy judgments."[144]

Other scholars suggest that the similarity between digital intermediaries and public forums should be embraced in the conventional sense, and that such a perspective should guide the formation of free-speech values vis-à-vis these intermediaries. Balkin argues that intermediaries are "public in the sense that their value as networks arises from public participation that produces network effects."[145] Law professor James Grimmelmann argues that the value of the Internet comes from its nature as a "semicommons."[146] "Without the private aspects, the Internet would collapse from overuse and abuse; without the common ones, it would be pointlessly barren," he writes. "But the two together are magical; their combination makes the Internet hum."[147] Balkin argues that "digital technologies change the social conditions in which people speak, … bring[ing] to light features of freedom of speech that have always existed in the

[142] Eric Goldman, *Speech Showdowns at the Virtual Corral*, 21 S. CLARA COMP. & HIGH TECH. L. J. 845, 848 (2005).
[143] *Id*. at 851.
[144] *Id*. at 853-4.
[145] Balkin, *supra* note 20, at 23.
[146] James Grimmelmann, *The Internet Is a Semicommons*, 78 FORD. L. REV. 2799 (2010).
[147] *Id*. at 2800.

background but now become foregrounded."[148] Such technologies, he argues, facilitate the purpose of freedom of speech, which "is to promote democratic culture" by affording individuals "a fair opportunity to participate in the forms of meaning making that constitute them as individuals."[149]

Synthesis

At their core, affirmative First Amendment theories put forth two arguments. First, they contend that the most important value of freedom of expression is mass participation by individuals in a self-governing democracy. Second, they contend that this ultimate value faces threats not only from state actors, but also from powerful private actors (namely large media conglomerates) that would seek to curb individuals' participation within public discourse due to its potential to compete with their own messages. Although some of the affirmative theorists discussed above take the additional step of proposing policies whereby state actors use laws to reign in the power of these private actors, this study does not take that step with them. Rather, this study seeks to use affirmative First Amendment theories to present a framework for understanding the potential threats that mainstream digital intermediaries pose to public discourse through content governance. Connecting the concepts from Internet governance with Lessig's theory of regulation helps bring the study of content governance into the ambit of legal scholarship and First Amendment theory.[150]

---

[148] Balkin, *supra* note 20, at 2.
[149] *Id*. at 3.
[150] LESSIG, *supra* note 13.

Lessig, Regulation and Affirmative Theory

Lessig calls for an analysis of regulation of speech through the lens of key First Amendment values. He proposes a model that involves four "modalities" of regulation: law, social norms, the marketplace, and the design of technologies that facilitate the activity being regulated.[151] Law regulates an activity either through threatening to punish the activity or codifying incentives that may lead people to engage in an alternative activity. Norms, often defined by a society's moral values, regulate an activity through either social stigmatization or encouragement. Markets regulate an activity by making it costly or by incentivizing an alternative activity. Finally, the design (Lessig calls it "architecture") of a technology that facilitates an activity will end up regulating the activity by only allowing it to be performed in the way permitted by the technology.[152] Importantly, all of these modalities are interdependent of one another.[153] All four interact with one another in the context of content governance:

Law: Section 230 of the 1996 Communications Decency Act gives commercial intermediaries immunity from tort liability over content created by third parties that others may consider "obscene, lewd, lascivious, filthy, excessively violent, harassing, or

---

[151] *Id*. at 124.

[152] Lessig gives the example of smoking to illustrate his four-part model. *Id*. at 122-3. Laws can make it more difficult to smoke in public places, thereby leading people to consider quitting smoking. Social stigmatization of smoking may lead people to quit lest they become social pariahs. The high cost of cigarettes may make the opportunity cost of smoking too high for many people, leading them to quit. Finally, the design of a cigarette makes smoking smelly and leads to lung cancer. If one wants to smoke but does not want to smell or get lung cancer, then one must choose to either not smoke, or use an alternative to smoking (like using a so-called "e-cigarette," which does make its user smell, though studies have yet to determine whether or not this new technology causes cancer).

[153] Continuing with the smoking example, the high cost of cigarettes is often the result of laws that determine the taxes governments collect on the product. The fact that cigarettes have a strong smell may lead to smoking being a socially stigmatized activity, which may in turn lead to laws that ban the activity in public places.

otherwise objectionable, whether or not such material is constitutionally protected."[154] Not only are intermediaries immune from liability as a result of the material being published on their platforms, they are also immune from liability for removing the content—essentially, from assuming control of it.[155] In other words, § 230 encourages content governance by promising that intermediaries will not be punished for it. Meanwhile, case law leads us to a fairly solid understanding that commercial intermediaries have a First Amendment right to manage content on their platforms.[156]

Norms: This modality is perhaps the most important to this study. Because law gives intermediaries so much latitude to manage UGC, these intermediaries must devise and enact their own policies to manage the content that gets published on their platforms.

---

[154] 47 U.S.C. § 230(c)(2)(A). However, this immunity is not unlimited: § 230(e)(1) stipulates that this statutory immunity will have no effect on matters of criminal law, namely on matters related to child pornography.

[155] *Id*., "*any action voluntarily taken in good faith to restrict access to or availability of material* that the provider or user considers …" *etc*. (emphasis added).

[156] Miami Herald v. Tornillo, 418 U.S. 241 (1974) (holding that a Florida statute requiring newspapers to publish responses from individuals who believed they were attacked in the newspapers amounted to an unconstitutional prior restraint.) Adopting a negative approach to First Amendment jurisprudence, the Court determined that such a statute constituted government control over the editorial process of a free press. Although the Court acknowledged that an ideal press was a responsible press that provided a forum for diverse viewpoints, it nonetheless held that "press responsibility is not mandated by the Constitution, and like many other virtues it cannot be legislated," at 256. *See also See* Bruce W. Sanford and Jane E. Kirtley, *The First Amendment Tradition and Its Critics*, *in* THE PRESS (Geneva Overholser and Kathleen Hall Jamieson eds., 2005), at 268. But *cf* Red Lion Broadcasting Co. v. FCC, 395 U.S. 367 (1969). In *Red Lion*, a unanimous Supreme Court held that the now defunct Fairness Doctrine—which required broadcasters to allot equal time to discussion of competing issues—did not violate broadcasters' First Amendment right to "use their allotted frequencies continuously to broadcast whatever they choose, and to exclude whomever they choose from ever using that frequency" (at 386). Rather, the Court held that the medium of broadcast (via TV or radio), with its unique situation of depending upon the scarce availability of the electromagnetic spectrum, required regulations such as the Fairness Doctrine to prevent a monopoly of viewpoints from controlling all access to broadcast channels. The Court matter-of-factly averred, "It is the right of the viewers and listeners, not the right of the broadcasters, which is paramount" (at 390). Although such an audience-centric interpretation of the First Amendment may have found favor at the High Court, that interpretation strictly applied to broadcast rather than print, and the Court has stated that the Internet should be classified as more akin to the latter than the former (Reno v. ACLU, 521 U.S. 844, 870 (1997).

Norms in content governance are constructed through a process of negotiation.[157] Sometimes, employees of digital intermediaries decide what kinds of content are subject to private governance. Other times, individual users define what constitutes undesirable content by pressuring intermediaries into following a set of certain community values.[158]

Market: As discussed above, the structure of the networked economy gives intermediaries an incentive to make their platforms a welcome experience for users.[159] The goal is to attract as many users as possible while scaring as few away as possible. This modality, then, is tightly connected to the norms modality: what sells will be what the community of users considers desirable. Thus, if community norms dictate that certain forms of extreme or potentially harmful speech "do not sell" (i.e. their presence may deter individuals from using a platform), that speech is not likely to survive due to market pressure.

Design: Whatever norms commercial intermediaries choose to implement and follow to govern content, they will be inscribed into the design (which, in this context, Lessig refers to as "code") of the platforms. For example, platforms may give users the ability to "flag" offensive content (i.e., notify the intermediary about the content).[160] The very act of removing content, or of "excommunicating" an individual who created objectionable content, is a function of the design of platforms.

---

[157] van Dijck, *supra* note 4, at 46.
[158] *See* Crawford and Gillespie, *supra* note 8.
[159] van Dijck, *supra* note 4, at 51; CITRON, *supra* note 11, at 202.
[160] Crawford and Gillespie, *supra* note 8.

Content Governance and Politics of Technology Theory

Technological design must be conceived as a product of social norms. Technology policy, according to communication professor Tarleton Gillespie, "is the construction and the legal authorization of sociotechnical systems designed to select out those activities we want to render impossible (and the converse, those we hope to encourage)."[161] Technologies "are the product of political choices and have political consequences that must be recognized and acknowledged."[162] Political choices and consequences imply an inherent conflict in the design and implementation of technology. Builders of technology "also build them in the rhetorical sense, drawing linguistic boundaries around them to indicate what is part of the [technology] and what is not, shaping how the relationship between elements can and will be characterized."[163] This rhetoric, the "interpretive flexibility" of technology, is always in flux.[164] Persuasion and technology go hand-in-hand because technology is, itself, an argument: an interpretation of how it should be used.

Answering the questions that guide this study is an act of interpreting and defining what the technology of communicative platforms should be and how they should function. This study synthesizes Internet governance with theories of content management and normative theories about the social value and limits of extreme expression. The aim of this synthesis is to explicate the concept of how digital intermediaries govern the extreme speech that gets published on their platforms. Key to

[161] GILLESPIE, *supra* note 16, at 10.
[162] *Id*. at 66.
[163] *Id*. at 75.
[164] *Id*. at 85.

this goal is an understanding of extreme speech. First Amendment theory and jurisprudence recognizes that extreme speech falls outside the "mainstream" of public discourse, yet it contends that the robustness of that discourse as a whole depends on the presence of extreme speech.[165] Extreme speech is the vanguard of all speech; by protecting extreme speech (but for a few exceptions), the law gives "breathing space"[166] for individuals to participate in a public discourse robust enough to test the boundaries of convention and decency.

Following the Trajectory of Broadcast

Another reason that this study calls upon affirmative theory is to connect the concept of content governance with the arguments of new realist scholars that once saw broadcast media as a threat to individual participation in public discourse. Barron, for example, argued that U.S. broadcasters wielded enormous power over the public discourse by exercising their First Amendment right to manage their programming.[167] This power imbalance came from the pervasiveness of the medium, the ability of broadcasters to reach large (potentially national) audiences, and the fact that the limited number of broadcasting channels were (and are even more so today) concentrated in the hands of a few large corporations. Famously, Barron contended this power was so great that U.S. courts should recognize a First Amendment right of giving individuals access to

[165] *See e.g.* Whitney v. California, 274 U.S. 357, 375 (1927) (Brandeis, J., dissenting); Brandenburg v. Ohio, 395 U.S. 444 (1969); NAACP v. Button, 371 U.S. 415 (1963); Texas v. Johnson, 491 U.S. 397 (1989); Snyder v. Phelps, 562 U.S. 09 (2011).
[166] New York Times Co. v. Sullivan, 376 U.S. 254, 270 (1964).
[167] Barron, *supra* note 23.

these otherwise closed channels of communication, lest the public discourse grow normalized and stagnant.[168]

Internet communications were seen as the antidote to this power imbalance, as individuals would have greater ability to communicate messages that could compete with those of large media conglomerates.[169] More than a decade before the invention of the World Wide Web, sociologist and technologist Ithiel de Sola Pool recognized the potential of electronic communication to be "expanders of human culture"[170] that could topple the monopolistic reign of large broadcast corporations.[171] However, Barron argues that networked communication has not vanquished the threat of corporate control over mass individual participation.[172] He concludes that the Internet, like broadcast, is a medium in which a few major players dominate traffic, thereby crowding out alternative perspectives.[173] Thus, following Barron's argument, if and when mainstream intermediaries scrub their platforms of extreme speech, access to such speech through fringe platforms becomes merely nominal.

Conclusion

The implications that content governance poses for mass participation in Internet communication should be framed in the context of this debate. For scholars like Pool and Barron, the issues presented by content governance are a step backward to the days of corporate power over public discourse through control of broadcast media. The only

---

[168] *Id*.
[169] *See*, *e.g.*, ITHIEL DE SOLA POOL, TECHNOLOGIES OF FREEDOM: ON FREE SPEECH IN AN ELECTRONIC AGE (1983); Jerome Barron, *Access Reconsidered*, 76 G. W. L. REV. 826 (2008).
[170] POOL, *supra*, at 226.
[171] *Id*. at 246-7.
[172] Barron, *supra* note 169.
[173] *Id*. at 843.

question that remains is how far backward this step is. Answering that question is the task of future research. In the meantime, this chapter calls for two things. First, it calls for greater awareness about the social values of extreme speech that face a potential threat from content governance. Scholars of mass communication law must vigorously defend and disseminate theories of freedom of expression that promote tolerance of extreme speech.[174] Second, it calls for greater transparency on the part of digital intermediaries on how UGC is governed on their platforms. Intermediaries, through their community standards,[175] do notify users what general categories of speech are permitted and what are not, but they do not open a window onto the process of determining one from the other. Intermediaries have a duty to open that window to maintain the robustness of the public discourse they facilitate.

[174] *See, e.g.*, LEE C. BOLLINGER, THE TOLERANT SOCIETY: FREEDOM OF SPEECH AND EXTREMIST SPEECH IN AMERICA (1986).
[175] Crawford and Gillespie, *supra* note 8.

Chapter 3: The value and limits of extreme speech in a networked society: A perspective from First Amendment theory and jurisprudence

The purpose of this study is to explicate the concept of content governance: the control that digital communication intermediaries exercise over user-generated content (UGC). The particular focus of this explication is the governance of extreme UGC. Two key questions guide this explication: How and why do digital communication intermediaries respond to extreme UGC? What are the potential implications of their responses for public discourse in a system of networked communication?

Chapter 2 analyzed content governance using so-called "affirmative" theories of the First Amendment. These theories generally promote the maximization of participation among individuals within public discourse, and propose that state actors have the ability to facilitate such participation through public policy measures aimed at limiting the power of corporate media to control that discourse.[1] Indeed, these theories conceive of corporate media of all stripes as powerful institutions that have a strong ability to limit individuals' participation in public discourse. However, from the perspective of affirmative theory, extreme speech still presents a quandary for issues of content governance. Some scholars[2] and politicians[3] have called for digital intermediaries to play a larger role in managing extreme or harmful speech on their platforms. The idea behind

---

[1] *See generally* ALEXANDER MEIKLEJOHN, FREE SPEECH AND ITS RELATION TO SELF-GOVERNMENT (1948); CASS SUNSTEIN, DEMOCRACY AND THE PROBLEM OF FREE SPEECH (1993); OWEN FISS, THE IRONY OF FREE SPEECH (1996); Jack M. Balkin, *Digital Speech and Democratic Culture: A Theory of Freedom of Speech for the Information Society*, 79 N.Y.U. L. REV. 1 (2004).
[2] DANIELLE KEATS CITRON, HATE CRIMES IN CYBERSPACE (2015); Mary Anne Franks, *Unwilling Avatars: Idealism and Discrimination in Cyberspace*, 20 COLUM. J. GENDER & L. 224 (2011).
[3] Nick Wingfield and Eric Lipton, *Google's Detractors Take Their Fight to the States*, N.Y. TIMES (Dec. 16, 2014).

these calls is that the law is either too slow or is faced with too many constitutional barriers to police extreme or harmful speech on its own, and therefore extralegal matters must be taken. Yet these calls essentially empower intermediaries to take on a *greater* role in governing UGC, something that the affirmative theorists discussed in the last chapter generally would be leery of. Thus, this chapter picks up where chapter two left off by posing the following question: at what point does the power of intermediaries to control extreme UGC become too great? And, at what point does society lose out from having an online public discourse cleansed of extreme and potentially harmful speech?

The analysis in this chapter will consider the answers that three First Amendment theories—marketplace of ideas theory, individual autonomy theory and tolerance theory—have to the following questions: What makes speech extreme? What makes it so extreme that it should be regulated? What is the ultimate benefit of speech—especially extreme speech—to democracy? The purpose of analyzing these theoretical questions is to distill the values of extreme speech and understand the limits (framed in terms of state action) of such speech—namely, at what point does protected extreme speech become harmful speech that could be prohibited? The main reason for defining the values and limits of extreme speech is that digital intermediaries and state actors often use the prevention of some sort of "harm" as a pretext for content governance. This one factor in common between these two distinct actors can allow a parallel analysis to take shape, whereby the concept to be explored is the balance between the social values of preventing harm and promoting extreme forms of speech. Thus, this chapter approaches the core focus of this study—the balancing act that digital intermediaries perform to promote

speech and prevent harm—from the perspective of traditional First Amendment theory and jurisprudence.

Digital intermediaries are not—nor should they be considered—state actors subject to the purview of the First Amendment. The First Amendment clearly states: "*Congress* shall make no law … abridging freedom of speech, or of the press,"[4] whereby Congress has been interpreted to refer to all government entities (and *only* government entities) throughout the United States through the doctrine of incorporation instilled in the Fourteenth Amendment.[5] Indeed, the entire U.S. Constitution is a treatise on the relationship between government and its citizens, not on the relationship between private parties (except, of course, for the Thirteenth Amendment, which prohibits the enslavement of one human being by another).[6] Therefore, but for this exception, a private actor can only be considered in violation of another's rights if the private actor is acting as an instrument of the state.[7] Digital intermediaries, for all the control they may harbor over public discourse, do not make the public discourse available as an agent of the state.[8]

---

[4] U.S. CONST. amend. I (emphasis added).
[5] Gitlow v. New York, 268 U.S. 652, 666 ("For present purposes we may and do assume that freedom of speech and of the press — which are protected by the First Amendment from abridgment by Congress — are among the fundamental personal rights and 'liberties' protected by the due process clause of the Fourteenth Amendment from impairment by the States").
[6] U.S. CONST. amend. XIII ("Neither slavery nor involuntary servitude, except as a punishment for crime whereof the party shall have been duly convicted, shall exist within the United States, or any place subject to their jurisdiction."). *See* Wilson R. Huhn, *The State Action Doctrine and the Principle of Democratic Choice*, 34 HOF. L. REV. 1379, 1388 (2006).
[7] *See, e.g.,* Lugar v. Edmondson Oil Co., 457 U.S. 922, 941 (1982) (holding, for example, that a private party's joint participation with state officials in the seizure of disputed property is sufficient to characterize that party as a 'state actor' for purposes of the Fourteenth Amendment").
[8] *See, e.g.,* Cyber Promotions v. AOL, 948 F. Supp. 436 (E.D. Pa. 1996) (holding that AOL's email service did not provide the "functional equivalent" to a public forum, nor did it amount to a "critical pathway" of communication, and thus the government could not regulate it).

However, such a distinction does not preclude the possibility of using negative First Amendment theory—theory that conceives of freedom of expression as a right against the government—to guide an analysis of the social values at stake and the limits that should be considered when intermediaries govern extreme UGC. Such a proposition is neither heretical nor novel. Smolla has argued that First Amendment jurisprudence can "serve as a model for private institutions and organizations" so that they "may choose to embrace freedom of speech as a preferred organizational value."[9] The private institutions that Smolla had in mind were private universities tasked with developing speech codes for their students and faculty,[10] with Smolla implying that private universities share the same function as their public counterparts (preparing young adults to be citizens), which would necessitate equal treatment of speech. However, unlike in Smolla's example, the degree to which private institutions resemble or function like public institutions should not matter in the type of "guiding" analysis found in this chapter. What matters is that digital intermediaries have power to control individuals' participation in public discourse, and therefore they merit an analysis of the values and limits that they employ to control that participation as if they were state actors. Negative First Amendment theories are a valuable tool in conducting such an analysis because of their binary structure: the rights of individuals to freedom of expression—and the values that justify such freedom, even to extreme lengths—are pitted against the power of another actor. With First Amendment theory, of course, that other actor is the government. The present analysis requires that digital intermediaries be transposed into the place of state actors.

[9] RODNEY A. SMOLLA, FREE SPEECH IN AN OPEN SOCIETY 45 (1992).
[10] *Id*.

This chapter will begin with a discussion of the value of speech, broadly conceived. It will then move into a discussion of the limits of extreme speech, also broadly conceived. Guiding these discussions will be an analysis of First Amendment doctrine as established by major court cases regarding the values and limits of extreme speech in specific situations.

The chapter will then proceed to an analysis of arguably the two most commonly cited negative theories: marketplace of ideas theory and individual autonomy theory.[11] In this first section, the general precepts of these theories as well as the main criticisms against them will be discussed. In particular, as noted in the overarching goal of this chapter, these theories will be analyzed in terms of how they conceive of the values and limits of extreme speech. The limits will be framed in terms of how each theory conceives of the potential harmfulness of speech.

The second half of this section will analyze tolerance theory, including its general precepts, its conceptions of the values and limits of extreme speech, and its main criticisms. The argument put forth in this chapter is that tolerance theory should be revitalized to further the understanding of the values and limits of extreme speech in an era of networked communication, for several reasons. First, the central focus of tolerance theory *is* extreme speech.[12] Second, tolerance theory expands upon John Stuart Mill's idea that tolerance of extreme speech is a societal issue, not simply a legal one.[13] Third,

---

[11] MATTHEW D. BUNKER, CRITIQUING FREE SPEECH: FIRST AMENDMENT THEORY AND THE CHALLENGE OF INTERDISCIPLINARITY 2, 11 (2001).
[12] LEE C. BOLLINGER, THE TOLERANT SOCIETY: FREEDOM OF SPEECH AND EXTREMIST SPEECH IN AMERICA 4 (1986).
[13] JOHN STUART MILL, ON LIBERTY 8 (1859/2001); BOLLINGER, *supra*, at 13.

the notion of tolerance is something that has been and can be empirically measured.[14] Therefore, even though tolerance theory—like all First Amendment theories—is normative in nature, its primary focus can become the foundation for future research on attitudes toward extreme speech and content governance in a networked communication system. Finally, tolerance theory is an important theory in the context of content governance because, at bottom, it argues that showing tolerance for extreme speech is a mental civic exercise. Therefore, once the social values of extreme speech have been established, the only way they will survive in an arena of public discourse where the threat of regulation is high is through a collective civic faith in those values.

Major criticisms of each of these three theories will also be presented. The goal of doing so is to separate the valuable core concepts of these theories from the major pitfalls that doom each theory from being able to stand on its own. These core concepts will be the main ingredients for a grand synthesis of a theory of freedom of expression for an era of networked communication in which private regulation of speech has become more and more common.

This chapter will conclude with a synthesis of the theories discussed in this chapter, and a proposal for a theory of the values and limits of extreme speech in the context of content governance. This theory will combine elements of Bollinger's tolerance theory and Baker's liberty model to create a theory that encourages tolerance

---

[14] *See generally* JULIE L. ANDSAGER, ROBERT O. WYATT, & EARNEST L. MARTIN, FREE EXPRESSION IN 5 DEMOCRATIC PUBLICS: SUPPORT FOR INDIVIDUAL AND MEDIA RIGHTS (2004); Dennis Chong, *How People Think, Reason, and Feel about Rights and Liberties*, 37 AMER. J. POL. SCI. 867 (1993); James L. Gibson & Richard D. Bingham, *On the Conceptualization and Measurement of Political Tolerance*, 76 AMER. POL. SCI. REV. 603 (1982); Jennifer L. Lambe, *Dimensions of Censorship: Reconceptualizing Public Willingness to Censor*, 7 COMM. L. POL'Y 187 (2002).

among both individuals and digital intermediaries, and that extolls the values of honoring individual autonomy in a system that empowers individuals' communicative potential like no other time in history. This theoretical analysis thus places content governance in the context of the broader discussion of how extreme speech has been regarded in First Amendment theory and jurisprudence for at least the last century.

Values and Limits

Why Speech?

Before getting into the discussion of negative First Amendment theory, the analysis must first ask a fundamental question: why speech?[15] What gives the act of speaking greater protection in U.S. law than other actions? Answering this question is important because it helps get past the so-called "argument from coincidence,"[16] the idea that freedom of speech must be revered simply because the First Amendment has conferred such a strong negative right. Answering this question is also important because certain harms caused by speech can be as harmful as—if not more harmful than—certain harms caused by conduct, yet "we are unwilling to disable ourselves from dealing with harmful, offensive, obnoxious, dangerous behavior in general in the way that we are with reference to speech."[17] Champions of freedom of expression must answer for these harms and give as strong a reason as possible for why "the [F]irst [A]mendment requires a likelihood of harm much higher than we otherwise require."[18] Beginning the analysis with this question does two things. First, it sets a tone that any theoretical analysis of

[15] Frederick Schauer, *Must Speech Be Special?* 78 NW. U. L. REV. 1284 (1983).
[16] *Id*. at 1298.
[17] *Id*. at 1303.
[18] *Id*.

freedom of expression should follow a high degree of rigor.[19] It gets beyond the "accepted assumptions, traditional metaphors, and standard platitudes" about the values of free speech, which Schauer argues are "clearly inadequate to confront the questions we must ask when trying to determine the extent to which … the [F]irst [A]mendment [should] encompass a wide range of activities seemingly so far from the comprehension of the classical free speech theorists that the relevance of classical theory has become attenuated."[20] Second, this question allows us to incorporate the perspectives of multiple First Amendment theories into the analysis. Schauer contends that "there need not be anything wrong with a multi-valued theory,"[21] as "it is unlikely that any *one* theory can explain the concept of free speech, and no reason necessarily exists to suppose that it could."[22] Or should. We would be wise to recognize that freedom of expression is made up of a "bundle of interrelated principles sharing no common set of necessary and sufficient defining characteristics."[23] This bundle includes a "unique mix of self-expression, self-realization, [and] capacity for influencing political change" as justifying the special protection for speech.[24] Law professors Daniel Farber and Philip Frickey

---

[19] *Id*. ("As we reject many of the classical platitudes about freedom of speech and engage in somewhat more rigorous analysis, trying to discover why speech—potentially harmful and dangerous, often offensive, and the instrument of evil as often as of good—should be treated as it is, our intuitions about the value of free speech, solid as they may be, are difficult to reconcile with this analysis.")
[20] *Id*. at 1288.
[21] *Id*. at 1303.
[22] Frederick Schauer, *Categories and the First Amendment: A Play in Three Acts*, 34 VAND. L. REV. 265, 277 (1981).
[23] *Id*. *See also* SMOLLA, *supra* note 9, at 2 ("There is no logical reason … why the preferred position of freedom of speech might not be buttressed by multiple rationales.")
[24] Schauer, *supra* note 15, at 1304.

agree, seeing free speech as "a powerful idea precisely because it appeals to so many diverse values."[25]

Therefore, speech is special because of its potential to be a great equalizing force in society. To be certain, social divisions by race, gender, class and access to communication resources mean that not everyone has the ability to speak—or be heard—with equal power.[26] Yet speech gives everyone the opportunity to exercise his or her autonomy by not only contributing a message to society, but also determining which messages are most valuable for society. Reverence for freedom of speech places society's faith in that autonomy, trusting it over any power (governmental or private) that would seek to manage the public discourse in a way that would negate that autonomy. Put differently, speech is special (especially in the United States) because it is the area of individual activity that involves a low amount of government involvement relative to individuals' potential to bring about social change. The theories analyzed herein will present their own interpretations about how freedom of expression can bring about such potential social change.

Extremeness and Harm

The next step in this analysis is to address the types of "extreme" speech that populate the borders of both social and legal acceptance. Truly, "[d]rawing constitutional

---

[25] Daniel A. Farber and Philip P. Frickey, *Practical Reason and the First Amendment*, 34 UCLA L. REV. 1615, 1643 (1987).

[26] *See, e.g.,* Clay Calvert, *Hate Speech and Its Harms: A Communication Theory Perspective*, 47 J. COMM. 4 (1997); Catherine MacKinnon, *Pornography, Civil Rights and Speech*, 20 HARV. CIV. RTS.-CIV. LIB. L. REV. 1 (1985); Cass Sunstein, *Pornography and the First Amendment*, 1986 DUKE L. J. 589 (1986); Richard Delgado, *Campus Antiracism Rules: Constitutional Narratives in Collision*, 85 NW. U. L. REV. 343 (1991); Katharine Gelber, *"Speaking Back": The Likely Fate of Hate Speech Policy in the United States and Australia*, *in* SPEECH & HARM: CONTROVERSIES OVER FREE SPEECH (Ishani Maitra and Mary Kate McGowan eds., 2012).

lines of inclusion and exclusion is vastly complicated by the multiple purposes and effects particular communications can have."[27] Defining extreme speech in terms of its potential to cause harm can help make these lines—if not brighter—at least less complicated to draw. However, setting these boundaries requires solid criteria as to what makes speech harmful. The criteria that this analysis will use come from Smolla's three-part model of harms that speech can cause: physical harm, relational harm, and reactive harm.[28] Each of these three harms has its own body of case law that sets the boundaries of the possible legal actions that can be taken against the speech. The following subsections will review those bodies of case law.

Physical Harm

U.S. free speech jurisprudence considers physical harm the worst of the three types of harms that could potentially be caused by speech, making it one area where clear exceptions have been devised to address this harm and to decide when the speech that falls outside of constitutional protection.[29] These types of speech include fighting words, true threats, and incitement to imminent lawless action. The fighting words doctrine comes from the 1942 case *Chaplinsky v. New Hampshire*.[30] In that case, Chaplinsky, a Jehovah's Witness, was convicted under a state breach of peace statute for calling a city marshal "a God damned racketeer and a damned Fascist."[31] The high court upheld Chaplinsky's conviction, and in so doing crafted the First Amendment exception for

[27] Kent Greenawalt, *Speech and Crime*, 1980 AM. B. FOUND. RES. J. 645, 784 (1980).
[28] SMOLLA, *supra* note 9, at 48.
[29] *Id*.
[30] 315 U.S. 568 (1942).
[31] *Id.* at 569.

fighting words, which the Court defined as words said in another person's face that "by their very utterance inflict injury or tend to incite an immediate breach of the peace."[32] Courts have seldom used the fighting words doctrine since *Chaplinsky*.[33]

The incitement standard was refined in the 1969 case *Brandenburg v. Ohio*.[34] That case involved a leader of a Ku Klux Klan group (Brandenburg) being convicted under an Ohio criminal syndicalism law for speaking racist messages to a frenzied crowd. The law prohibited "advocat[ing] … the duty, necessity, or propriety of crime, sabotage, violence, or unlawful methods of terrorism as a means of accomplishing industrial or political reform."[35] In a per curiam decision, the Supreme Court reversed Brandenburg's conviction, holding that "the constitutional guarantees of free speech and free press do not permit a State to forbid or proscribe advocacy of the use of force or of law violation except where such advocacy is directed to inciting or producing imminent lawless action and is likely to incite or produce such action."[36] The imminent lawless action standard narrowed the definition of unlawful incitement from the "bad tendency"[37] and "clear and present danger"[38] standards cited by the Court earlier in the twentieth century. Concurring in *Brandenburg*, Justice Douglas lamented "how easily [the] 'clear and present danger' [standard] is manipulated to crush what Brandeis called [t]he

---

[32] *Id*. at 572-3.
[33] Though for an interesting analysis of how the fighting words doctrine could apply in Internet communication, *see* Clay Calvert, *Fighting Words in the Era of Texts, IMs and E-Mails: Can a Disparaged Doctrine Be Resuscitated to Punish Cyber-Bullies?*, 21 DEPAUL J. ART TECH. & INTELL. PROP. L 1 (2010).
[34] 395 U.S. 444 (1969).
[35] *Id*. at 445.
[36] *Id*. at 447.
[37] Schenck v. U.S. 249 U.S. 39 (1919); Debs v. U.S., 249 U.S. 211 (1919); Abrams v. U.S., 250 U.S. 616 (1919); Gitlow v. New York, 268 U.S. 652 (1925).
[38] Cantwell v. Connecticut, 310 U.S. 296 (1940); Terminiello v. Chicago, 337 U.S. 1 (1949).

fundamental right of free men to strive for better conditions through new legislation and new institutions by argument and discourse."[39] He decried that the standard had been "manipulated,"[40] and "twisted and perverted" for political ends.[41] The test was applied and further refined soon after *Brandenburg* in *Hess v. Indiana*,[42] in which the Court held that a vague command shouted to protestors to "take the fucking streets later" did not amount to an incitement to imminent lawless action. In a *per curiam* opinion, the Court held that Hess' speech was not directed toward any specific individual or group, nor did it exhibit an intent to commit any imminent lawless action.[43]

Law professor Kent Greenawalt identifies four parts to the incitement standard: the extent of the lawlessness of the action the speech is advocating; who the speech is being directed at; the likelihood of the action occurring; and the imminence of the action occurring.[44] Each of these elements provides a layer of protection to speech that has the potential to lead to physical harm. Some scholars have argued that networked communication allows extreme speech to surpass each of these protective layers. For example, the requirement that the communication be directed immediately at an angry audience may no longer be a sufficient condition for the incited lawless action to be imminent. Law professor Lyrissa Lidsky argues that an inflammatory message posted on social media can target multiple audiences (both intended and unintended) who may be

[39] Brandenburg v. Ohio, 395 U.S. 444, 452 (1969) (Douglas, J., concurring).
[40] *Id*. at 453.
[41] *Id*. at 454.
[42] 414 U.S. 105 (1973).
[43] *Id*. at 108-9.
[44] Greenawalt, *supra* note 27.

more likely than even a restive mob to imminently commit a violent illegal act.[45] Others, however, urge greater restraint.[46] It is true that a message that did not directly advocate for a specific lawless action to imminently occur could still unwittingly spur an unintended audience to commit such an action. However, the fact that such a result may be highly likely should not be reason enough to suppress speech. The very idea that *any* message—especially a political one—could incite someone to violence should galvanize society to maintain its standard of only outlawing the rare case of *directly* inciting imminent lawless action, online or off.[47]

The extent of the "true threat" exception to the First Amendment is currently (as of this writing) under scrutiny at the U.S. Supreme Court. The contention stems from how courts should interpret the standard enunciated in *Virginia v. Black*, in which the Supreme Court held that a Virginia statute criminalizing cross burning was unconstitutionally overbroad.[48] Justice O'Connor wrote that to convict a person of threatening another by burning a cross, the state of Virginia must consider whether a reasonable person would have found the burning to be threatening, as well as whether the accused intended for the burning to be threatening.[49] The latter half of the two-prong test is the more difficult: the government must prove that the "speaker means to communicate

[45] Lyrissa Barnett Lidsky, *Incendiary Speech and Social Media*, 44 TEX. TECH L. REV. 147, 149 (2011).
[46] Lynn Adelman and Jon Deitrich, *Extremist Speech and the Internet: The Continuing Importance of* Brandenberg, 4 HARV. L. & POL'Y REV. 361 (2010). *See also* L. A. Powe, Jr., *Brandenburg: Then and Now*, 44 TEX. TECH L. REV. 69 (2011).
[47] *Id*.
[48] 538 U.S. 343 (2003).
[49] *Id*. at 360 (O'Connor, J., writing for the plurality).

a serious expression of an intent to commit an act of unlawful violence to a particular individual or group of individuals."[50]

Since *Black*, courts have given competing interpretations of how the true threat standard should be applied in networked communication. In *U.S. v. Bagdasarian*, the U.S. Court of Appeals for the Ninth Circuit held that vague references to assassinating President Obama posted on an online message board could not reasonably be interpreted to be threatening speech.[51] However, in *Planned Parenthood of the Columbia/Willamette, Inc. v. American Coalition of Life Activists*—decided a year before *Black*—a divided Ninth Circuit sitting *en banc* held that online "wanted posters" that glorified the deaths of and encouraged violent attacks on doctors who performed abortions amounted to an unconstitutional true threat.[52]

On December 1, 2014, the U.S. Supreme Court heard oral arguments in the case of *Elonis v. United States*,[53] thereby revisiting the role of the subjective standard to the true threat exception. The U.S. Court of Appeals for the Third Circuit held that the subjective standard was unique to the facts in *Black*, and that O'Connor's plurality opinion could not be read to require a subjective standard across all circuits.[54] Appellant Anthony Elonis was convicted of "transmitting in interstate commerce communications containing a threat to injure the person of another" under 18 U.S.C. § 875(c).[55] He had posted comments on his Facebook page alluding to a desire to savagely murder his

---

[50] *Id*.
[51] 652 F. 3d 1113 (9th Cir., 2011).
[52] 290 F.3d 1058 (9th Cir. *en banc*, 2002).
[53] 134 S. Ct. 2819 (*cert. granted* 2014).
[54] United States v. Elonis, 730 F.3d 321 (3rd Cir. 2013).
[55] *Id*. at 326.

estranged wife and cause bodily harm to kindergarten students, law enforcement officers and former coworkers.[56] Elonis contended that his speech was protected because *Black* required proof of the subjective intent of the accused speaker to deliver a threat, and he argued his speech was a mimicking of rap lyrics that gave him catharsis after his wife left him.[57]

However, the Third Circuit held that *Black* did not afford Elonis' speech the protection of a subjective intent standard. The court distinguished Elonis's Facebook posts from the cross burning in *Black*, holding that the symbolic nature of the cross-burning required the state to prove that the intent behind it was threatening rather than, for example, a political expression of solidarity among Klansmen.[58] Meanwhile, the meaning behind Elonis's Facebook posts was not nearly as ambiguous.[59] The Third Circuit also pointed out that in *Black*, the Virginia statute being challenged already contained a requirement that the subjective intent of a threat be proven. Therefore, the test for true threats enunciated in *Black* only applied to the Virginia statute, and it does not force other statutes that do not have a subject intent requirement (such as Section 875) to suddenly adopt such a requirement.[60] Moreover, the Third Circuit held that the spirit of Section 875(c) was sufficiently met using only an objective standard. The court wrote, "Limiting the definition of true threats to only those statements where the speaker subjectively intended to threaten would fail to protect individuals from 'the fear of

[56] *Id*.
[57] *Id*. at 327.
[58] *Id*. at 330.
[59] *Id*.
[60] *Id*. at 329.

violence' and the 'disruption that fear engenders,'" which is the chief goal of the federal statute.[61]

In sum, these three standards (fighting words, incitement and true threats) separate the worst of speech-related harms from the body of protected speech in the United States. They are a recognition that speech can cause significant harm and that the First Amendment is not an absolute, while simultaneously establishing incredibly high standards that are lauded for how strongly they protect extreme speech. Although recent cases have shown that networked communication is changing the ways in which extreme speech can lead to physical harm, these standards have yet to be substantially weakened.

Relational Harm

Relational harm, according to Smolla, involves speech that causes injury to social relationships (defamation), business relationships (fraud or false advertising), ownership interests (copyright) and confidentiality (leaking national security secrets).[62] This section will focus only on jurisprudence regarding harms caused by defamation due to 1) the tradition of strong First Amendment protection afforded to allegedly defamatory speech, and 2) the close relationship between defamation and Smolla's third category: reactive harms.[63]

The U.S. Supreme Court constitutionalized defamation law in *New York Times v. Sullivan* by requiring public-official plaintiffs to prove that libelous statements about them were made with "actual malice": the "knowledge that it [the statement] was false or

[61] *Id*. at 330.
[62] SMOLLA, *supra* note 9, at 48.
[63] *See infra*, notes 78-85.

[made] with reckless disregard of whether it was false or not."[64] Justice Brennan gave the following reasoning for the Court's unanimous decision in *Sullivan*: "if newspapers, publishing advertisements dealing with public issues, … risk liability [upon a jury's evaluation of the speaker's state of mind], there can also be little doubt that the ability of minority groups to secure publication of their views on public affairs and to seek support for their causes will be greatly diminished."[65] That same philosophy extends to public figures: those "who are not public officials, but [are] involved in issues in which the public has a justified and important interest,"[66] thereby making them subject to public scrutiny. Private individuals, however, generally are afforded greater leeway in pursuing defamation suits; depending on the law of the state where they bring the suit, private individuals may not have to prove actual malice to win their case.[67]

A litany of other statutory[68] and common law[69] protections—too numerous and nuanced to list here—exist to further buttress the protection afforded to defamation

---

[64] New York Times v. Sullivan, 376 U.S. 254, 280 (1964).
[65] *Id*. at 300.
[66] Curtis Publishing Co. v. Butts, 388 U.S. 130, 132 (1967).
[67] Gertz v. Welch, 418 U.S. 323 (1974).
[68] These primarily include state laws that seek to prevent so-called "strategic lawsuits against public participation" or "SLAPPs." Generally, these laws allow would-be defendants the ability to move to dismiss a defamation lawsuit against them at early stages of legal action, thereby placing upon plaintiffs the burden of proving that their suit likely would prevail on the merits if the case were to go to trial. The purpose of these laws is to eliminate the potential silencing of speech out of fear of the potentially high costs of defending a defamation suit that the defendant would most likely win anyway. For more on the state of anti-SLAPP laws in the United States, *see Anti-SLAPP Laws*, Reporters Committee for Freedom of the Press, *available at* http://www.rcfp.org/browse-media-law-resources/digital-journalists-legal-guide/anti-slapp-laws-0.
[69] These primarily include state law privileges that predate the *Sullivan* actual malice standard and protect the press from liability for defamation. For example, the fair comment privilege protects speakers who publish criticism of artistic works from liability for defamation as long as the critical comments are accurate, *see* Restatement (Second) of Torts, §§ 606-607. Another example is the privilege of neutral reportage, whereby "[l]iteral accuracy [in reporting] is not a prerequisite," and journalists should enjoy "immunity from defamation suits where the journalist believes, reasonably and in good faith, that his report

defendants. One such statute, section 230 of the 1996 Communications Decency Act[70]—discussed in the previous chapter—extends that protection to the digital intermediaries that host potentially defamatory speech by granting them immunity from civil liability if they voluntarily restrict access to or remove UGC.[71] The law has been a boon for harmful speech online, which can proliferate due to speakers' ability to cloak themselves in anonymity and intermediaries' potential disincentive to remove the allegedly infringing speech.[72] At the same time, the law has the potential to indirectly lead to the restriction of speech by intermediaries who will remove extreme speech out of a concern for its bottom line without the fear of being held liable for its removal.[73]

Smolla categorizes defamation as a relational harm due to its close similarity with other business-related harms such as copyright infringement; essentially, it is a harm against property rights. Constitutional scholar Robert Post argues that defamation law in the United States is built on the metaphor that "reputation is capital."[74] Reputation is the fruit "of one's own endeavors."[75] Reputation works hand-in-hand with American

---

accurately conveys the charges made," Edwards v. National Audubon Society, Inc., 556 F.2d 113, 120 (2nd Cir. 1977).

[70] 47 U.S.C. § 230.

[71] 47 U.S.C. § 230(c)(2). However, immunity does not extend to third-party content that violates criminal law, such as obscenity or child pornography (§ 230(e)(1)), or that violates intellectual property laws (§ 230(e)(2)).

[72] *See generally* 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 (2010). *See* Zeran v. America Online, Inc., 129 F.3d 327 (4th Cir. 1997), *cert denied*, 524 U.S. 937 (1998) (holding that AOL did not materially contribute to hosting allegedly defamatory statements posted by an anonymous user to AOL's message board service claiming that Zeran was selling apparel that disparaged the 1995 Oklahoma City bombing).

[73] *See* Rebecca Tushnet, *Power Without Responsibility: Intermediaries and the First Amendment*, 76 G.W. L. REV. 986, 1011 (2008).

[74] Robert C. Post, *The Social Foundations of Defamation Law: Reputation and the Constitution*, 74 CAL. L. REV. 691, 693 (1986).

[75] *Id*. at 694 (internal quotations omitted).

capitalism; it can be spent or invested to build up one's fortune, which, in turn, can be invested back into one's good reputation. Post argues that in the United States, the "purpose of the law of defamation is to protect individuals within the market by ensuring that their reputation is not wrongfully deprived of its proper market value."[76] Prior to *New York Times v. Sullivan*, Post's theory was corroborated at common law.[77] However, despite its close connection to property harms, defamation in several ways straddles the line between relational harm and reactive harm. The analysis in the following section will illustrate how, thereby connecting defamation with the murky implications for networked communication that are associated with reactive harms.

Reactive Harm

Smolla's third category, reactive harm, includes intentional infliction of emotional distress, tortious invasions of privacy, and any type of hate speech. The Supreme Court has raised the standards for plaintiffs suing under the first two categories by imputing the actual malice standard from *Sullivan* into many of these torts, due in large part to their similarity to the tort of defamation.[78] Hate speech has been defined many different ways by many different scholars, but a generic definition for the purposes of this study categorizes hate speech as any speech that attacks and attempts to subordinate any group or class of people, typically spoken by a group with a higher level of social power than

[76] *Id*. at 695.
[77] SUNSTEIN, *supra* note 1, at 39.
[78] Cantrell v. Forest City Publishing, 419 U.S. 245 (1975) (holding that plaintiffs must prove actual malice to successfully recover for the tort of false light invasion of privacy); Hustler v. Falwell, 485 U.S. 46 (1988) (holding that a public figure plaintiff must prove actual malice to successfully recover for the tort of intentional infliction of emotional distress).

the targets of the speech.[79] The targets of such speech typically include racial minorities, women, religious minorities, and homosexuals. Generally, hate speech is only punishable if it contravenes one of the few First Amendment exceptions listed above.[80] According to Smolla, speech that leads to reactive harms deserves the highest level of constitutional protection due to its tendency to implicate public figures or officials, or its tendency to involve important social issues and matters of public concern: factors which greatly outweigh the potential harms of the speech.[81]

Reactive harms are likely considered the least worrisome of harms caused by speech because generally they are considered less tangible than physical or relational harms.[82] These latter two categories implicate life and property, while reactive harms can be reduced to "hurt feelings."[83] Certainly, harm to one's mental wellbeing is nothing

---

[79] For various studies with various definitions of hate speech, *see generally* Calvert, *supra* note 26; Alexander Tsesis, *Dignity and Speech: The Regulation of Hate Speech in a Democracy*. 44 WAKE FOREST L. REV. 497 (2009); Richard Delgado and David H. Yun, *Pressure Valves and Bloodied Chickens: An Analysis of Paternalistic Objections to Hate Speech Regulation*, 82 CAL. L. REV. 871 (1994); Owen M. Fiss, *The Supreme Court and the Problem of Hate Speech*. 24 CAPITAL U. L. REV. 281 (1995); Stephanie Farrior, *Molding the Matrix: The Historical and Theoretical Foundations of International Law Concerning Hate Speech*. 14 BERKELEY J. INT'L L. 1 (1996); Jean-Marie Kamatali, *The U.S. First Amendment Versus Freedom of Expression in Other Liberal Democracies and How Each Influenced the Development of International Law on Hate Speech*. 36 OHIO N.U. L. REV. 721 (2010); Robert Post, *Hate Speech*, *in* EXTREME SPEECH AND DEMOCRACY (Ivan Hare and James Weinstein eds., 2009); Tanya Katerí Hernández, *Hate Speech and the Language of Racism in Latin America: A Lens for Reconsidering Global Hate Speech Restrictions and Legislation Models*. 32 U. PA. J. INT'L L. 805 (2011).

[80] *See* R.A.V. v. St. Paul, 505 U.S. 377 (1992) (holding that a St. Paul, Minn. ordinance banning symbolic speech (such as cross-burning) that is hateful "on the basis of race, color, creed, religion or gender"—a content-based restriction of speech—was unconstitutionally under-inclusive); National Socialist Party of America v. Village of Skokie, 432 U.S. 43 (1997) (holding that delays in issuing parade permits to Nazis were, in and of themselves, a content-based restriction on the Nazi Party's speech); Snyder v. Phelps, 562 U.S. 09 (2011) (holding that allowing an individual, even a private citizen such as Mr. Snyder, to sue for civil damages from emotional distress intentionally inflicted by lawful social speech would lead to a chilling effect on such speech).

[81] SMOLLA, *supra* note 9, at 48.

[82] C. Edwin Baker, *Scope of the First Amendment Freedom of Speech*, 25 UCLA L. REV. 964, 998 (1978) (hereinafter Baker, *Scope of the First Amendment*).

[83] Hustler Magazine, Inc. v. Falwell, 485 U.S. 46, 53 (1988) ("in the world of debate about public affairs, many things done with motives that are less than admirable are protected by the First Amendment"); at 55

trivial. Many scholars who crusade for greater regulations (particularly in the United States) against hate speech point out that the damage such speech causes to the wellbeing of minorities leads to physical (hence, more important) ailments such as anxiety and depression, which in turn may make minorities retreat from participating in society.[84] However, the inability to conceive of a legal test that would show a "direct causal link"[85] between speech and mental harms (like with true threats, incitement, or actual malice), weighed against Smolla's argument that speech associated with reactive harms often implicates public officials, makes speech that causes reactive harms the least deserving of an exception from First Amendment protection.

Harms v. Value

It is important that society understands the harms of speech because these harms must be juxtaposed with the societal value of expansive free speech rights. "The evils posed by 'harmful' speech are likely to appear real to the political branches," Smolla writes, but "[t]he interests served by allowing such speech to remain free … will often appear unreal."[86] To increase the realness of the value of speech in the face of the realness of its potential harms, one must identify and appreciate the ideational value of

---

("'Outrageousness' in the area of political and social discourse has an inherent subjectiveness about it which would allow a jury to impose liability on the basis of the jurors' tastes or views, or perhaps on the basis of their dislike of a particular expression. An 'outrageousness' standard thus runs afoul of our longstanding refusal to allow damages to be awarded because the speech in question may have an *adverse emotional impact* on the audience" (emphasis added)). *Cf* RESTATEMENT (SECOND) OF TORTS § 46 cmt. j (1965): "some degree of transient and trivial emotional distress is a part of the price of living among people."

[84] Calvert, *supra* note 26; Caroline West, *Words the Silence? Freedom of Expression and Racist Hate Speech*, *in* SPEECH & HARM: CONTROVERSIES OVER FREE SPEECH (Ishani Maitra and Mary Kate McGowan eds., 2012).

[85] Clay Calvert, Kara Carnley, Brittany Link and Linda Riedmann, *Conversion Therapy and Free Speech: A Doctrinal and Theoretical First Amendment Analysis*, 20 WM. & MARY J. WOMEN & L. 525 (2014).

[86] SMOLLA, *supra* note 9, at 41.

the speech. For physical harms, that task is relatively easy: unprotected speech must involve a "confluence of lack of ideational content, [physical] harm to a targeted recipient, and likelihood of ensuing physical violence."[87] For reactive harms, the task becomes more difficult. Here, Smolla puts forth his "emotion principle," which claims that speech has both emotional and intellectual effects. Under the emotion principle, speech cannot be banned due to its emotional component alone; the intellectual component must be factored in, and even the slightest intellectual value will tip the scale in favor of protecting the speech.[88] Thus, banning speech to prevent harm "may not be satisfied by the outrage or moral opprobrium that a majority of the populace attaches to the activity. Crimes must have victims, … and the victimization must be palpable, something beyond generalized disgust or disquiet over another's conduct."[89]

At the same time, the question should be asked: how broadly should one define the social value of speech? To propose a potential limit to protected speech, Smolla gives the example of a racial slur written on a bathroom stall: "It states no fact, offers no opinion, proposes no transaction, attempts no persuasion. It contains no humorous punch line, no melodic rhythm, no color or shape or texture that might pass as art or entertainment. It offers only hate for hate's sake, with no mental gloss other than the feeble minimum intellectual current necessary to power the use of words."[90] However, should exceptions be defined so narrowly? Is there any point outside of the well-

---

[87] Frederick Schauer, *Intentions, Conventions, and the First Amendment: The Case of Cross-Burning*, 2003 SUP. CT. REV. 197, 205 (2003).
[88] SMOLLA, *supra* note 9, at 46.
[89] *Id*. at 10.
[90] *Id*. at 167-8.

established legal tests discussed above at which the cost of the harm outweighs the benefit of the speech to an intolerable degree? Or is harm only legally recognized when speech has *no* value?

Sunstein concedes that "the line is sometimes thin between restrictions based on 'harm' and restrictions based on viewpoint of content."[91] However, he holds that the primary factor that should determine whether speech is protected "is whether the speech is a contribution to social deliberation, not whether it has political effects or sources."[92] Thus, Sunstein distinguishes a misogynist tract from pornographic movies, a racist speech to a crowd from face-to-face racial harassment, and a tract in favor of white supremacy from a racial epithet.[93] However, Sunstein points out that even within each of these categories, not all hateful words are equal in their potential to cause reactive harm. He writes, "It is obtuseness—a failure of perception or empathetic identification—that would enable someone to say that the word 'fascist' or 'pig' or even 'honky' produces the same feelings as the word 'nigger.'"[94] A deeper, simpler point can be made from Sunstein's argument: although the many examples of extreme speech listed above receive strong legal protection due to their theoretical social value, their harms are no less real to the people who suffer them.

The networked communication environment has mutated the relationship between speech and harm. The Internet's facilitation of anonymous speech has lowered the social

---

[91] SUNSTEIN, *supra* note 1, at 174.
[92] Cass Sunstein, *Free Speech Now*, 59 U. CHI. L. REV. 255, 309 (1992).
[93] *Id*.
[94] SUNSTEIN, *supra* note 1, at 186.

cost for speakers of inflicting all sorts of harm through their online words.[95] Networked communication has seen the generation of new categories of harmful speech, such as "revenge porn" (which involves posting nude images online of an ex-romantic partner to spite her or him) and "cyber-harassment" (which involves persistently inflicting substantial emotional distress against an individual through online communications).[96] The reach, permanence and anonymity of Internet communication have the potential to amplify the harms of both traditional hate speech and these new breeds of speech.[97] In Internet communication, the reactive harms associated with hate speech have the potential to morph into physical harms when they take the form of cyber-harassment or abuse of an individual. In other words, Smolla's clean lines distinguishing harms are becoming strained.

It sounds callous to say that any of the harms discussed above—physical, relational or reactive—no matter how real and severe the harm may seem to the victim of the harm, will most likely not outweigh the nebulous, theory-based social values of that speech. It is no less callous to point out that this outcome is the result of the subjectivity of the harms felt and the lack of (or difficulty in successfully meeting the requirements of) a legal test connecting speech directly with harm. Nevertheless, in this battle between the values and harms of speech, deference must always be given to speech. Deference does not mean absolute protection, but rather the presumption that extreme speech is

---

[95] *See* Barnett Lidsky, *supra* note 45; KEATS CITRON, *supra* note 2; Franks, *supra* note 2; Danielle Keats Citron, *Cyber Civil Rights*, 86 BOST. U. L. REV. 61 (2009); Yuval Karniel, *Defamation on the Internet—A New Approach to Libel in Cyberspace*, 2 J. INT'L MEDIA & ENTMT L. 215 (2008).
[96] KEATS CITRON, *supra* note 2.
[97] *Id*. *See also* Franks, *supra* note 2, at 228.

protected until a legal test can show that the speech has caused significant harm. In the sections below, three theoretical justifications for this position will be discussed: marketplace of ideas theory, individual autonomy theory, and tolerance theory. The purpose of doing so is to make the theoretical justifications for not punishing those who speak harmfully a little less nebulous. These justifications also will reiterate how important it is to give deference to extreme speech when the potential to silence it through extra-legal means (such as through content governance) exists.

Negative Theory

With a general notion of the doctrinal parameters highlighting the limits of extreme speech, the analysis now proceeds to a discussion of how negative theories of freedom of expression regard these values and limits. This analysis is just the beginning of a discussion that must be had over how digital intermediaries can practice content governance in a way that protects the values of individuals' expression of extreme speech while mitigating the harms of extreme speech that cross well-established limits. The binary nature of negative theories—pitting speakers against state actors—makes them useful tools for analyzing the regulation of freedom of expression, even when the regulators are not state actors. Obviously, these theories are deeply rooted in the state action doctrine,[98] and the goal of this analysis is not to brazenly disregard that storied doctrine by concluding that First Amendment jurisprudence should be applied to digital intermediaries. Rather, the present analysis is based on the following argument: if state actors are substituted in these theories for any institution that has the power to regulate

[98] *Supra* notes 4-8.

speech, the values of extreme speech that the theories purport to be lost from regulation do not change.

Marketplace of Ideas

The first brand of negative First Amendment theory that will be discussed here, the marketplace of ideas theory, holds that freedom of speech principally acts as a means to attaining truth. The roots of the theory date back to English author John Milton's *Areopagitica*, in which the English author and philosopher wrote, "And though all the winds of doctrine were let loose to play upon the earth, so Truth be in the field, we do injuriously by licensing and prohibiting to misdoubt her strength. Let her and Falsehood grapple; who ever knew Truth put to the worse in a free and open encounter?"[99] Two centuries later, English philosopher John Stuart Mill expressed his own take on Milton's philosophy, writing, "if [an opinion] is not fully, frequently, and fearlessly discussed, it will be held as a dead dogma, not living truth."[100] The Millian conception of free speech is preoccupied with the notion that truth always requires falsity.[101] According to Mill, "[a]ll silencing of discussion is an assumption of infallibility."[102] Truth without competing falsity "is but one superstition the more, accidentally clinging to the words which enunciate a truth."[103]

---

[99] JOHN MILTON, AREOPAGITICA 32 (1644), *available at* http://books.google.com/books?id=6bJDAAAAcAAJ&printsec=frontcover&dq=JOHN+MILTON,+AREOPAGITICA&hl=en&sa=X&ei=Ueo5VPvCDNSuyATm3oDgCA&ved=0CB8Q6AEwAA#v=onepage&q=JOHN%20MILTON%2C%20AREOPAGITICA&f=false.
[100] MILL, *supra* note 13, at 34.
[101] *Id*. at 25.
[102] *Id*. at 19.
[103] *Id*. at 34.

Thus, the general conclusion from Mill is that the strength of accepted truth comes from two distinct yet related sources: the truth's ability to withstand challenges from alternative ideas, and the truth's steadfastness in not assuming its infallibility. The value of extreme speech is found in the latter half of the model. Extreme ideas may not be the best alternatives to displace an accepted truth; rather, they are a test of the accepted truth's denial of infallibility. In other words, Mill contends that there is a social value in acknowledging the potential of "heretical" opinions to become truth, and a concomitant detriment in denying that potential.[104] However, Mill is not a triumphalist. He writes that "the dictum that truth always triumphs over persecution is one of those pleasant falsehoods which men repeat after one another till they pass into commonplaces, but which all experience refutes. History teems with instances of truth put down by persecution."[105]

Under Mill's perspective, extreme speech can be conceived as a necessary foil for society to accept a certain truth. However, it is important that the kind of "extreme" speech Mill was referring to be put into context. Mill's focus is on challenges to political and religious truths: Catholicism versus Protestantism, good taxes versus bad taxes, just wars versus unjust.[106] Former law professor Jeremy Ofseyer argues that scholars today anachronistically and irresponsibly use Mill's conception of heretical speech to justify categorical protection for all speech, including harmful speech such as incitement to

[104] *Id*. at 32. ("The greatest harm done is to those who are not heretics, and whose whole mental development is cramped, and their reason cowed, by the fear of heresy.")
[105] *Id*. at 28.
[106] *Id*. at 21.

riot.[107] Part of the problem, Ofseyer argues, is that scholars have misinterpreted Mill's famous "harm principle."[108] Mill does propose that "[t]he only freedom which deserves the name, is that of pursuing our own good in our own way, so long as we do not attempt to deprive others of theirs, or impede their efforts to obtain it."[109] Consequently, "the only purpose for which power can be rightfully exercised over any member of a civilized community, against his will, is to prevent harm to others."[110] However, although Ofseyer concedes that Mill "does not discuss in detail the proper limits on harmful speech,"[111] he interprets the harms that the harm principles seeks to avoid as harms to other rights that individuals possess.[112] For example, he interprets Mill to argue that speech used to commit fraud or conspiracy—and even, in certain circumstances, *sedition*—should be banned in the name of protecting individuals' autonomous rights.[113] Although this interpretation may seem no less confusing than any other, Ofseyer argues that it fits with Mill's view that individual autonomy is the greatest source of individual rights.[114]

Part of the problem in understanding the harm principle comes from the tricky question of whether freedom of speech is an outward right or an inward right. Mill concludes it is the latter: "The liberty of expressing and publishing opinions may seem to … belong[] to that part of the conduct of an individual which concerns other people; but, being almost of as much importance as the liberty of thought itself, and resting in great

---

[107] Jeremy Ofseyer, *Taking Liberties with John Stuart Mill*, 1999 ANN. SURV. AM. L. 395, 428 (1999).
[108] *Id*. at 401.
[109] MILL, *supra* note 13, at 16.
[110] *Id.* at 13.
[111] Ofseyer, *supra* note 107, at 411.
[112] *Id*. at 429.
[113] *Id*. at 430.
[114] *Id*. at 433.

part of the same reasons, is practically inseparable from it."[115] Thus, Mill conceives of speech as being more strongly connected to the thoughts from which the speech was formed than to the public to which the speech will be addressed. This conception is important because it connects speech closely (though not absolutely) with the autonomy of the speaker. Therefore, any "harm principle" argument that would prohibit speech that causes harm to another must be framed in terms of autonomy. Not only must the harm be to the autonomy of another, but it must be greater than the harm caused to the autonomy of the speaker by being prohibited from speaking.

The marketplace of ideas metaphor has become a powerful theory that fits well with the general *laissez-faire* philosophy of the U.S. political, legal and economic system.[116] However, scholars continue to criticize several of the theory's most central tenets. Particularly criticized is the lack of a causal link connecting free speech to both greater knowledge among individuals and the ultimate outcome of the realization of truth.[117] Schauer points out that a justification for why truth will always win out over falsity "is noticeably absent from all versions of the argument from truth."[118] He blames this flaw on the naïveté of the Enlightenment philosophy of continuous social progression.[119] Although Schauer does concede that the marketplace model may be more valid in the long run, he posits that the predominant risk of the model is that in the short run "false views may, despite their falsity, be accepted by the public, who will then act in

---

[115] MILL, *supra* note 13, at 15-16. But *cf* MILL, *supra* note 13, at 17: "it is impossible to separate the cognate liberty of speaking and of writing."
[116] *See* Frederick Schauer, *The Exceptional First Amendment*, *in* AMERICAN EXCEPTIONALISM AND HUMAN RIGHTS (Michael Ignatieff ed., 2005).
[117] FREDERICK SCHAUER, FREE SPEECH: A PHILOSOPHICAL INQUIRY 15 (1982).
[118] *Id*. at 25.
[119] *Id*. at 26.

accordance with those false views."[120] Ultimately, Schauer considers "the tired metaphors of the marketplace of ideas and the search for truth … as stage props" for a broader and more difficult issue: the "debate over how much the values of free speech would have to yield in the face of exigent public concerns."[121]

Meanwhile, the late law professor C. Edwin Baker takes issue with the idea that Mill's conception of truth is a subjective one (i.e. one that society accepts); he argues that the marketplace theory only works if truth is conceived of in a Miltonian sense: as objective.[122] In finding problems with each conception of truth, Baker attacks the whole of marketplace theory. Subjective truth, Baker argues, requires justification as to why it is the "best" possible truth.[123] Since all of the criteria required to assess which truth is the best would be relative to one another, it would be impossible to know which truth to accept as the best. But objective truth has its own problems. It assumes that all people value the same thing, which is erroneous according to the argument from social constructivism that all human beings are products of their perceptions of their social surroundings.[124] This conclusion leads back to truth being subjective, at which point Baker closes the circle of his argument. Even if one were to accept that the model works with a relative conception of truth, Baker argues that the marketplace theory fails because

[120] *Id*. at 28.
[121] Schauer, *supra* note 15, at 1285.
[122] C. EDWIN BAKER, HUMAN LIBERTY AND FREEDOM OF SPEECH 12-13 (1989).
[123] *Id*. at 6.
[124] *Id*.

it assumes that members of society actually *want* the truth.[125] Individuals may value entertainment or expedience, if not outright falsity, rather than the truth.

Finally, the marketplace model is criticized for the means by which society would attain truth: the assumption that all people are rational and have perfect knowledge of every possible idea to choose from.[126] Baker argues that there are simply too many choices of ideas to choose from, let alone to digest with the proper rational faculties to determine whether it is worthy of being considered true.[127] On a related note, as discussed above, many affirmative First Amendment theorists and critical legal theorists argue that social cleavages along the lines of race,[128] gender[129] and access to means of communication[130] preclude the marketplace of ideas from ever being a place where all ideas can be accessed, processed and judged equally, no matter how rational human beings may be. Despite these serious criticisms, marketplace of ideas theory retains great purchase and utility among U.S. jurists[131] and legal scholars.[132] One reason may be the fact that the theory gives individuals pride of place, thereby connecting it with another theory that complements it: individual autonomy theory.

---

[125] *Id*.
[126] *Id*. at 7.
[127] *Id*.
[128] Delgado, *supra* note 26.
[129] MacKinnon, *supra* note 26.
[130] Jerome A. Barron, *Access to the Press: A New First Amendment Right*, 80 Harv. L. Rev. 1641, 1656 (1967).
[131] *See, e.g.,* Red Lion Broadcasting Co. v. FCC, 395 U.S. 367, 390 (1969) ("It is the purpose of the First Amendment to preserve an uninhibited market-place of ideas in which truth will ultimately prevail"); Abrams v. U.S., 250 U.S. 616, 630 (1919) (Holmes, J., dissenting) ("the best test of truth is the power of the thought to get itself accepted in the competition of the market, and that truth is the only ground upon which their wishes safely can be carried out"); New York Times v. Sullivan, 376 U.S. 254, 279 N19 (1964) ("Even a false statement may be deemed to make a valuable contribution to public debate, since it brings about 'the clearer perception and livelier impression of truth, produced by its collision with error'").
[132] BUNKER, *supra* note 11.

Individual Autonomy

One alternative to the marketplace of ideas model is to abandon its essential consequentialist perspective. Individual autonomy theory—also referred to as the "liberty model" by Baker[133]—does exactly that. Like the marketplace of ideas model, individual autonomy theory has its roots in the Enlightenment ideal of entrusting good governance to individuals and their ability to think, act and make important decisions as autonomous, rational beings.[134] Law professor Martin Redish contends that the purpose of government is to facilitate self-fulfillment among individuals.[135] Similarly, Baker argues that the "fundamental purpose" of the First Amendment is to facilitate individual self-fulfillment, thereby allowing individuals to participate in social and political change.[136] Individual autonomy theory is not completely free from of consequentialism, but its consequentialist bent is located at the micro level (changes in the individual) rather than the macro level (societal changes). Unlike the truth-centric perspective of the marketplace of ideas, Baker's liberty model does not focus on the ideal outcomes of freedom of expression. He argues that his liberty model "manifests a deep, democratic faith in people by providing for a more realistic method of [social and political] change from 'the bottom up.'"[137] Similarly, Redish holds that freedom of expression has both an intrinsic value in its

[133] BAKER, *supra* note 122.
[134] SCHAUER, *supra* note 117, at 60.
[135] Martin H. Redish, *The Value of Free Speech*, 130 U. PA. L. REV. 591, 627 (1982).
[136] BAKER, *supra* note 122, at 51.
[137] *Id*. at 91.

allowing individuals to be in control of their own destinies, and it has an instrumental value in facilitating the development of individuals' human faculties.[138]

For Baker, the liberty model is so fundamentally focused on the individual that he contends that "freedom of speech is a right of individuals, not market-oriented institutions or corporations."[139] To a certain extent, this perspective allies Baker with some of the affirmative theorists discussed in Chapter Two. Baker goes so far as to argue that "the government can engage in structural regulation to reduce … private censorship," so long as the "response to the private threat must not abridge individuals' freedom of speech."[140] Baker maintains that this fundamental distinction between individual and institutional speakers extends to his conception of the press as being a means to serve individuals. "The individual's constitutionally protected speech interest is not in the press as a profit-making unit, as a means of production," Baker writes, "but as a consumption good, that is, as a means to communicate what the individual chooses."[141] Although this interpretation may not be doctrinally sound according to the jurisprudence of today's Supreme Court,[142] Baker's interpretation of media as a means to serving individual self-fulfillment could be extended to the function of digital intermediaries as facilitators of mass participation in public discourse. In other words, in certain circumstances content governance could be conceived (under Baker's interpretation) as an affront to individual self-fulfillment, to the extent that such disregard for individual autonomy could

[138] Redish, *supra* note 135.
[139] BAKER, *supra* note 122, at 271.
[140] *Id*. at 270.
[141] *Id*.
[142] *See* Citizens United v. Federal Election Com'n, 130 S. Ct. 876 (2010) (holding that "[t]he Government may regulate corporate political speech through disclaimer and disclosure requirements, but it may not suppress that speech altogether").

necessitate, at most, state action preventing intermediaries from such action, or, at least, an ethical duty on the part of intermediaries to uphold the value of individual autonomy. (The construction of this ethical duty will be the subject of chapter 5.)

Baker's liberty model is skeptical of the use of state action in the name of mitigating harm. For him, the "key aspect distinguishing harms caused by protected speech acts from most other methods of causing harms is that speech harms occur only to the extent people 'mentally' adopt perceptions or attitudes."[143] This conception follows Smolla's hierarchy of harms, with reactive (or mental) harms being considered the least problematic compared to physical or relational harms.[144] This fact does not necessarily mean the reactive harms are less important or do less damage than a physical harm, but rather that it is impossible for anyone—let alone a lawmaker—to know the extent of the harm taking place within an individual's mind.

For Baker, speech is most harmful when it causes harm to the very thing that his theory seeks to honor: autonomy. However, he distinguishes between formal autonomy and substantive autonomy: formal autonomy refers to an individual's sheer ability to make choices for him or herself;[145] substantive autonomy refers to an individual's "actual capacity and opportunities to lead the best, most meaningful, self-directed life possible."[146] For example, he writes, "hate speech does not interfere with or contradict anyone else's formal autonomy even if [it] does cause injuries that sometimes include

---

[143] Baker, *Scope of the First Amendment*, *supra* note 82, at 998.
[144] SMOLLA, *supra* notes 78-85.
[145] C. Edwin Baker, *Autonomy and Hate Speech*, *In* EXTREME SPEECH AND DEMOCRACY (Ivan Hare & James Weinstein eds., 2009), 142.
[146] *Id*. at 143.

undermining others' substantive autonomy."[147] In fact, Baker transfers that notion of formal autonomy to the listener in an attempt to argue that policing harmful speech does greater harm to individual autonomy than the speech ever could. He argues that "outlawing acts of the speaker in order to protect people from harms that result because the listener adopts certain preconceptions or attitudes disrespects the responsibility and freedom of the listener."[148] Echoing Baker, law professor John Garvey argues that such "[p]aternalistic restrictions [on speech] can be justified only on the assumption that the state is best able to choose on the individual's behalf, that is to say, only on the assumption that the individual's choice in the matter in question is not entitled to the same respect, and to the same constitutional protection, as the preference that the majority establishes for him."[149]

The main criticism against individual autonomy theory is its propensity to border on hedonism.[150] As stated above,[151] Baker's assumption is that any social benefits that come from speech will come as externalities to a system of freedom of expression that focuses solely on the role of and benefits for individual speakers. This assumption suffers from the same flaw as the assumption from marketplace of ideas theory that a regime of free speech will lead society to truth: its incompleteness belies its fundamental conclusion. Individual autonomy theory will play a key role in developing a theory of

---

[147] *Id*.
[148] *Id*.
[149] John H. Garvey, *Freedom and Choice in Constitutional Law*, 94 HARV. L. REV. 1756, 1771 (1981).
[150] BUNKER, *supra* note 11, at 13.
[151] *Supra* note 137.

freedom of expression in an era of networked communication, but it must be supplemented. Tolerance theory is that supplement.

Tolerance Theory

Unlike the marketplace of ideas and individual autonomy theories analyzed above, tolerance theory, proposed by Bollinger,[152] does not focus directly on an ideal outcome of speech nor on the nature of the speakers and their relation to government actors. Rather, Bollinger's focus is on the speech itself, and namely its potentially extreme manifestations. The purpose of his theory, Bollinger argues, is to show that "defining what appears, at least, to be the periphery is to gravitate toward considering the most fundamental issues about the principle [of freedom of expression] itself."[153] He contends that it is partly due to the fact that extreme speech is protected "that the free speech idea holds such a peculiar and powerful fascination for us."[154] Bollinger's central argument is that "society adds something important to its identity, [and] is significantly strengthened, by … acts of extraordinary tolerance."[155] Law professor Steven Smith holds that this "something important" can be found in both the prudential and intrinsic values of tolerance.[156] The prudential value of tolerance can be found in the (admittedly somewhat naïve) notion that as networked and mass communication shrink our world, the ability to tolerate the most extreme types of speech will allow human beings to tolerate

[152] BOLLINGER, THE TOLERANT SOCIETY, supra note 12.
[153] *Id*. at 4.
[154] *Id*. at 6. *See* Schauer, *supra* note 15.
[155] BOLLINGER, supra note 12, at 9.
[156] Steven D. Smith, *The Restoration of Tolerance*, 78 CALIF. L. REV. 304 (1990).

virtually everyone else on the planet.[157] The intrinsic value of tolerance (and this is the value on which Bollinger focuses most heavily) amounts to a quasi-religious experience, an ascetic quality of denying an impulse to be intolerant toward an extreme viewpoint.[158]

Bollinger, himself, does not give an explicit definition for extreme (or, as he calls it, "extremist") speech, but he does leave clues on how a definition can be formed. Extreme speech, Bollinger says, is what "nearly all of us believe immoral and vicious."[159] It "tend[s] to attract attention,"[160] and "is very often the product or the reflection of the intolerant mind at its worst and, as such, an illustration to us of what lies within ourselves."[161] Bollinger wrestles with escaping a tautological cycle of defining extreme speech: that it is extreme because it falls outside the general values of society, and it falls outside those values because it is extreme. This tautology can be broken if extreme speech is defined in terms of its harmfulness, or at the very least the reasonable perception of its being harmful. Bollinger is not naïve when it comes to acknowledging the real harms that speech can cause, as he argues that legal scholars have a tendency to "understate the risks and harms of speech and … overstate its benefits."[162] Yet he urges caution in taking this acknowledgement too far, writing:

> Whatever verbal formulation is ultimately used as a starting point for free speech analysis, it must be flexible enough to permit, and perhaps even invite, consideration of the wide variety of social harm speech can cause, while also strong enough to reflect the important institutional role of free speech, that the central purpose of the enterprise is to push the boundary of

[157] *Id*. at 334.
[158] *Id*. at 336.
[159] BOLLINGER, supra note 12, at 124.
[160] *Id*.
[161] *Id*. at 126.
[162] *Id*. at 237.

> toleration beyond what would be considered normal by the usual standards of the society.[163]

Thus, although he does not give a specific standard to measure when speech becomes too harmful to be tolerated, he seems to imply that the harms must do real physical damage to an individual, rather than merely upset society's standard of decency. Indeed, it is society's intolerant nature underlying that very standard of decency that Bollinger seeks to soften through his theory. This theory of harm puts Bollinger in the same camp as both Smolla and Baker, giving credence to the notion that only speech with a propensity to cause physical harm should be considered worthy of proscription.

In expounding tolerance theory, Bollinger attempts to answer Schauer's call for more rigorous analysis on why speech is given such a revered place in society that its extreme and often socially harmful nature is given a high level of protection.[164] "Why is it," Bollinger asks, "that one form of coercion or punishment—legal restraints—has essentially been removed from our general armory of possible responses to speech we hate and fear and believe dangerous to the values we cherish?"[165] The "threat of government abuse" of power to restrain speech "is an argument that is seriously overplayed in twentieth century life," he argues.[166] Instead, the "real threat to liberty of speech … rests within the general population of citizens instead of officialdom alone."[167] Government is but one "means by which this impulse [to censor] is executed, but the

---

[163] *Id*. at 192.
[164] Schauer, *supra* note 15.
[165] BOLLINGER, supra note 12, at 13.
[166] *Id*. at 79.
[167] *Id*. at 80.

impulse rests elsewhere in the recesses of human nature."[168] Human beings, Bollinger contends, have a "deep and profound difficulty in controlling a desire to censor or suppress any difference of belief, opinion, or way of thinking."[169] Bollinger is not the first legal theorist to put forward this argument. Law professor Thomas Emerson writes, "Any society, and any institution in society, naturally tends toward rigidity."[170] Smolla has called censorship "a social instinct."[171] Affirmative First Amendment theorists such as Sunstein decry an "artificial distinction" between intrusion on a speaker by government and by private actors.[172] The roots of this philosophy lie in Mill, who calls for precautions being needed as much against the tyranny of the majority in society as against any abuse of government power.[173] "[S]o natural to mankind is intolerance in whatever they really care about," Mill writes.[174]

For Bollinger, government is not necessarily a threat to free speech. Quite the opposite, in fact: by following an ethic of tolerance, government can act as a guide for its citizens on how to uphold and protect the values of extreme speech. Because "the power of social intolerance exceeds that of legal intolerance,"[175] the law must set an example for society by tolerating extreme speech to an extreme degree. Bollinger writes, "If it is a tendency of human nature to overreact in the use of legal restraints against speech activity, we must expect that tendency to manifest itself in the form of nonlegal coercion

---

[168] *Id*. at 86.
[169] *Id*. at 92.
[170] Thomas I. Emerson, *Toward a General Theory of the First Amendment*, 72 YALE L. J. 877, 884 (1963).
[171] SMOLLA, *supra* note 9, at 4.
[172] SUNSTEIN, *supra* note 1, at 48.
[173] MILL, *supra* note 13, at 8.
[174] *Id*. at 12.
[175] BOLLINGER, supra note 12, at 110.

as well."[176] He then argues that the converse to this statement is true: if the law reduces the tendency to restrain extreme speech, society will follow suit.[177]

Tolerance theory suffers from two major criticisms. The first is its proposed causal relationship between tolerating speech and tolerating other activity. Law professor David Strauss writes, "It is not at all clear that people who are forced to tolerate speech they abhor will become more tolerant in other contexts; they might easily become *less* tolerant. … Indeed [Bollinger's theory] would become an argument for suppression."[178] In other words, like marketplace of ideas theory and individual autonomy theory, tolerance theory is criticized for lacking the ability to empirically prove that the ultimate goal that the theory propounds will actually happen. The second criticism attacks the elitist nature of tolerance theory. Strauss calls the theory an "imposition of a regime … on the ignorant, intolerant masses by an elite that alone understands both the virtues of tolerance and the way to manipulate institutions in order to achieve it."[179]

Despite these knocks against it, tolerance theory remains appealing due to its simplicity, and due to its addressing of the relationship between social and legal constraints upon speech. Figure 3-1 models this relationship. The first set of concentric circles represents a state of low legal tolerance for extreme speech relative to social tolerance. In the second set of circles, legal tolerance for extreme speech increases (for example, to a level consistent with First Amendment jurisprudence). This increase leads

---

[176] *Id.* at 109.
[177] *Id*.
[178] David A. Strauss, *Why Be Tolerant? (reviewing Lee C. Bollinger, The Tolerant Society: Freedom of Speech and Extremist Speech in America (1986)*, 53 U. CHI. L. REV. 1485, 1499 (1986) (emphasis original).
[179] *Id*. at 1500.

to the outcome postulated by tolerance theory: social tolerance for extreme speech will concomitantly increase with legal tolerance. The catalyzing agent that makes this process work is knowledge, spread among society, of the benefits of extreme speech. As society understands more and more why legal tolerance exists to such a great extent, social tolerance will expand to achieve the same goals as legal tolerance. Social tolerance is fluid, and it has the potential to attain the same level as legal tolerance. However, given the many types of speech that many diverse sectors of society will find harmful or offensive, it is unlikely that social tolerance will ever quite match the expansive nature of legal tolerance under a regime of First Amendment jurisprudence.

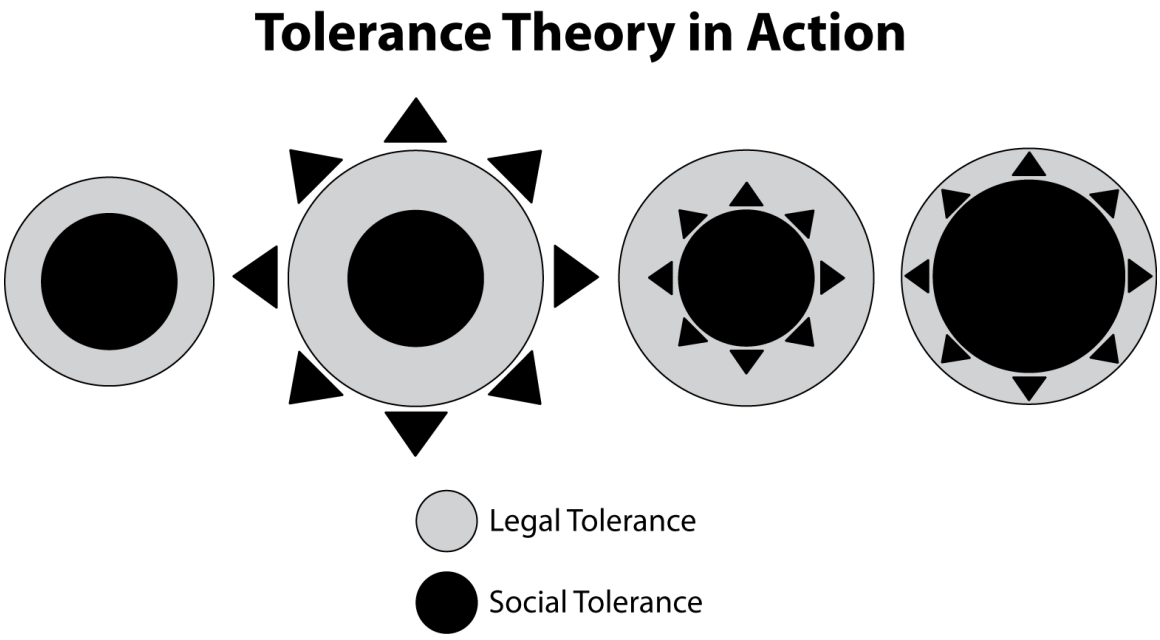

Figure 3-1: Relationship between Legal Tolerance and Social Tolerance in Tolerance Theory

Synthesis: A Theory of Free Speech for Networked Communication

Schauer is correct: no one theory can explain or justify the exceptional protection of free speech under the First Amendment.[180] This position is especially true in a system of networked communication, where digital technologies facilitate the potential for speech of all sorts from individuals and institutions, including the uplifting, the entertaining, the revolutionary, the offensive, and the harmful. This section will synthesize the theories analyzed and discussed to this point to arrive at a theory of freedom of expression for an era of networked communication.

Foundation in Tolerance

For several reasons, Bollinger's tolerance theory offers the ideal foundation for a networked-communication theory of free speech. First, the Internet has the potential to bring people closer to all types of speech, including speech that they may disagree with or find offensive. Although some may choose to ignore such speech, others have the ability to utilize tools that platforms make available to them to try to remove the speech from the platforms. Of particular concern and relevance is the ability of individuals to flag speech as inappropriate according to the community guidelines of the platforms[181] (see Figure 2-2 from chapter 1). Second, digital intermediaries have the legal authority to govern speech on their platforms. As discussed in the previous chapter, these intermediaries have an incentive in the networked economy to remove speech that would offend users so

[180] *Supra* notes 21-24.
[181] Kate Crawford and Tarleton Gillespie, *What is a flag for? Social media reporting tools and the vocabulary of complaint*, NEW MEDIA & SOC'Y 1 (2014).

much that the users would leave the platforms.[182] These two elements combined—intermediaries' incentive to remove extreme speech and individuals' ability to activate this incentive—pose a threat to speech in an online public discourse controlled by intermediaries. The third reason tolerance theory should form the foundation for a theory of freedom of expression in an era of networked communication is that it focuses on this very threat: individuals' ability to force extreme or unpopular speech out of public discourse.[183] The fact that digital intermediaries can be willing accomplices for intolerant individuals makes tolerance theory all the more important in an era of networked communication.

The goal of tolerance theory is not for extreme viewpoints to be *accepted* as truth in society, but simply that society *allows* them into the public discourse. What happens next is up to other theories to sort out. A revised version of the marketplace of ideas theory is the next best step in the blend of First Amendment theories proposed for networked communication: a marketplace of ideas model that is not concerned with an ultimate goal of truth, but rather with pure competition among all types of speech. Tolerance would make individuals more likely to engage in such a model, encouraging them to counter extreme speech with more speech rather than flagging it or pressuring intermediaries to remove it. Indeed, a fine line separates outcompeting extreme speech with more speech and using means made available by private institutions (essentially, a form of speech) to eradicate the speech from public discourse. For example, Mill writes, "There is a limit to the legitimate interference of collective opinion with individual

[182] CITRON, *supra* note 2, at 202.
[183] *Supra* notes 167-169.

independence: and to find that limit, and maintain it against encroachment, is as indispensable to a good condition of human affairs, as protection against political despotism."[184] This passage highlights one of the key principles of Mill's essay: to understand the "dealings of society with the individual in the way of compulsion and control," one must recognize the equal power of "physical force in the form of legal penalties, or the moral coercion of public opinion" to limit individual liberty.[185]

The difference between the marketplace of ideas model and Mill's fear of a "moral coercion of public opinion" is, simply put, more speech. More speech can lead to a greater diversity of voices in the public discourse. More speech can breed greater awareness among individuals of what ideas will be accepted in society and what will not. More speech can create a system in which extreme speech seems less scary and threatening to society. For example, despite its gross offensiveness and potential to cause serious reactive harm, a hypothetical video posted to Facebook of someone shouting racist epithets should stay on Facebook to allow society the opportunity to observe and identify the racist vitriol that still exists within our midst. It should be a phenomenon around which people rally with more speech denouncing racism.[186] It should not be something to fear and sweep under the rug by having Facebook remove it. Thus, if

---

[184] MILL, *supra* note 13, at 9.
[185] *Id*. at 13.
[186] In March 2015, the Oklahoma University chapter of the Sigma Alpha Epsilon fraternity was closed down after a video appeared online depicting several of its members using racial slurs to express their desire that an African American would never become a member of their fraternity. *See Two Oklahoma Students Expelled Over Racist Chant Video*, BBC NEWS (Mar. 10, 2015), *available at* http://m.bbc.com/news/world-us-canada-31821397. The video sparked further debate about the issues of race and white privilege at U.S. colleges and universities. A witness to the event filmed the video, not the members of the fraternity. However, had the members of the fraternity been the ones to publish the video, the resulting public outcry certainly would have been no different. Thus, such a video would of necessity need to be publicly aired, not removed from view.

tolerance theory works—and, admittedly, that is no small if—then greater tolerance of speech will lead to a more robust online public discourse because it will break the chain of private governance of extreme speech on digital platforms.

What About Harm?

Digital intermediaries have several goals in their operations, but two stick out: promoting speech and minimizing harm.[187] Indeed, as pointed out in chapter 2, intermediaries have an incentive to do both in the context of the networked economy. Because speech on platforms can be commoditized through the tracking of the personal data of both individual speakers and audience members, more speech is better economically speaking.[188] However, extreme and potentially harmful speech can lead individual users to stop using the platforms, which in turn can lead intermediaries to lose revenue.[189] Therefore, it makes no sense from an economic perspective for intermediaries to be tolerant, because they appear to have a greater incentive to police harmful speech than to tolerate all speech, including the harmful.

However, tolerance can be a prudent policy for digital intermediaries. Digital intermediaries have not been successful in creating an environment in which speech is

---

[187] *See* Josh Braun and Tarleton Gillespie, *Hosting the Public Discourse, Hosting the Public*, 5 JOURNALISM PRACTICE 383 (2011); Tarleton Gillespie, *The Politics of Platforms*, 12 NEW MEDIA & SOC'Y 347 (2010).

[188] *See, e.g.,* Ganaele Langlois, Fenwick McKelvey, Greg Elmer and Kenneth Werbin, *Mapping Commercial Web 2.0 Worlds: Towards a New Critical Ontogenesis*, 14 THE FIBRECULTURE J. 1 (2009); José van Dijck, *Users like you? Theorizing agency in user-generated content*, 31 MED. CULT. & SOC'Y 41, 46 (2009); Ute Schaedel & Michel Clement, *Managing the Online Crowd: Motivations for Engagement in User-Generated Content*, 7 J. MED. BUS. STUD. 17 (2010); James G. Webster, *User Information Regimes: How Social Media Shape Patterns of Consumption*, 104 NW. U. L. REV. 593 (2010); CITRON, *supra* note 2, at 202.

[189] Ramon Lobato, Julian Thomas & Dan Hunter, *Histories of User-Generated Content: Between Formal and Informal Media Economies*, 5 INT'L J. COMM. 899 (2011).

overly restricted in the name of preventing harm. Rather, online public discourse is an environment in which intermediaries remove some speech for being extreme or harmful,[190] while failing to act on reported instances of personal abuse.[191] In other words, the objectives of protecting extreme political or social speech and preventing harm are moving in the opposite direction from their ideal outcomes. Following a policy of tolerating extreme speech while policing abuse can help to reverse this trend. Intermediaries would show the world a clear distinction between these types of speech, highlighting how the former is often very valuable for public discourse, while the latter harms its specific targets as well as the general health of the public discourse.[192]

Conclusion

All of the First Amendment theories discussed above envision a link between speech and utopia. Marketplace of ideas theory sees speech as leading society to truth. Individual autonomy theory sees speech as the means by which all individuals can truly realize their natural ability to autonomously reason. Tolerance theory argues that accepting extreme speech will lead individuals to become more enlightened. For each of these theories, greater protection for speech is a step closer to utopia, and more restriction is a step away from it.

Utopia is impossible. But that not mean these theories are invalid. At best, they are imperfect. At worst, misguided. Instead of thinking of these theories in terms of

---

[190] *See generally* pp. 169-188 of chapter 4.
[191] A leaked memo sent to staff members at Twitter, the company's CEO Dick Costello famously stated, "We suck at dealing with abuse." Nitasha Tiku and Casey Newton, *Twitter CEO: "We suck at dealing with abuse,'* The Verge (Feb. 4, 2015), *available at* http://www.theverge.com/2015/2/4/7982099/twitter-ceo-sent-memo-taking-personal-responsibility-for-the. *See also* Caitlin Dewey, *Twitter Takes Death Threats Seriously … at Least When They're Directed at Its Own Employees*, Wash. Post (Sept. 9, 2014).
[192] *See* pp. 178-181 of chapter 4.

impossible utopian outcomes, one should view them in the simplest of terms: as reasons for why more speech is better than less. They are reasons why more control over speech (whether by state or private actors) is worse than less control. They are reasons why encouraging mass participation of individuals in public discourse is the most desirable outcome, and banishing individuals and their ideas from that discourse should be avoided. As digital intermediaries provide individuals with greater tools than ever before with which to participate in public discourse, while simultaneously presenting more opportunities to control individuals' speech, these theories have become more important than ever.

Chapter 4: "Heckler's Veto 2.0: Audience rights and agency in a networked society"

Introduction

The purpose of this study is to explicate the concept of content governance: the control that digital communication intermediaries exercise over user-generated content (UGC). The particular focus of this explication is the governance of extreme UGC. Two key questions guide this explication: How and why do digital communication intermediaries respond to extreme UGC? What are the potential implications of their responses for public discourse in a system of networked communication?

Chapter 4 now focuses on the conflicting concepts of speakers' rights and audiences' rights in First Amendment theory and jurisprudence. The extent to which audiences have a right to hear or avoid certain speech depends greatly on context, as this chapter will show. Arguably, the conflict between speaker rights and audience rights is best conceptualized within the context of so-called "hostile audience cases,"[1] in which the threat of a "heckler's veto" by a hostile audience against a speaker is present. A heckler's veto is "the suppression of speech by the government[] because of the possibility of a violent reaction by hecklers."[2] A heckler's veto occurs when "the state [hides] behind the unpleasant reaction of some portions of the public in order to silence a speaker" through the use of an instrument of law, such as a disorderly conduct statute.[3]

---

[1] Ashutosh Bhagwat, *Associational Speech*, 120 YALE L. J. 978, 1011 (2011).
[2] Ronald B. Standler, HECKLER'S VETO (Dec. 4, 1999), *available at* http://www.rbs2.com/heckler.htm. *See also* Berger v. Battaglia, 779 F.2d 992, 1001 (4th Cir. 1985) (defining the heckler's veto as "the successful importuning of government to curtail 'offensive' speech at peril of suffering disruptions of public order."
[3] Cheryl A. Leanza, *Heckler's Veto Case Law as a Resource for Democratic Discourse*, 35 HOFSTRA L. REV. 1305, 1306 (2007).

This chapter argues that understanding the conflict between speakers and audiences at a broad level—and especially understanding the doctrine of the heckler's veto—is crucial to understanding how crowds can suppress unpopular speech in networked communications.

The central argument of this chapter is that the legal principle of the heckler's veto is an essential analogy for understanding the dynamics of content governance—particularly the decisions of digital intermediaries to remove extreme or allegedly harmful content from their platforms due to popular pressure. This chapter will follow the same "comparative" argumentative structure as chapter 3: legal theory and doctrine will be discussed within its original context (vis-à-vis state actors), with key concepts then being transposed onto a system in which digital intermediaries play a powerful role in governing individual expression. Like in chapter 3, the purpose of this argumentative structure is not to argue that a functional equivalence between digital intermediaries and state actors should be established with regard to control over speech.[4] Rather, the purpose is to identify the principal values that shape the doctrine in its original context and apply those values to a discussion of content governance. The ultimate question that gets asked is thus: what mechanisms and inherent values are missing in a system of communication governance where the heckler's veto is likely (and perhaps incentivized) to take place?

[4] Such a line of analysis is interesting, but ultimately fruitless for the purpose of establishing any kind of legal force prohibiting digital intermediaries from discriminating against content as if it were the publicly appointed manager of a public forum. *See, e.g.,* Cyber Promotions v. AOL, 948 F. Supp. 436 (E.D. Pa. 1996) (holding that AOL's email service did not provide the "functional equivalent" to a public forum, nor did it amount to a "critical pathway" of communication, and thus the government could not regulate it).

The central argument in this chapter is based on the premise that the heckler's veto is a crucial concept within tolerance theory.[5] Hostile audience cases are the supreme test of tolerance by state actors toward extreme speech. Faced with the potential for popular disapproval of a message to turn violent, state actors must refrain from acquiescing to the popular will and protect the rights of speakers to share their unpopular message rather than punish them for angering the audience. This doctrine reflects a deep respect for the value of unpopular messages within American democracy. It removes the ability of audiences to take matters into their own hands and, through violent reaction, compel state action to silence speech with which they disagree or find offensive.

However, audiences can and do take their disdain for offensive or harmful messages into their own hands within networked communication. In a system of content governance, individuals can activate intermediary controls to remove extreme speech from mainstream platforms through several means. First, individuals can "flag" content published on platforms for allegedly violating one or more of the community guidelines of the platform.[6] Second, individuals can mobilize mass online protests against the intermediaries for allowing such extreme speech on their platforms, often in a way that

[5] LEE C. BOLLINGER, THE TOLERANT SOCIETY: FREEDOM OF SPEECH AND EXTREMIST SPEECH IN AMERICA (1986).

[6] *See, e.g.,* Kate Crawford and Tarleton Gillespie, *What is a flag for? Social media reporting tools and the vocabulary of complaint*, NEW MEDIA & SOC'Y 1 (2014). This method is the exclusive means by which YouTube and Twitter manage extreme content on their respective platforms. Facebook relies on flagging by users to alert it to extreme content, yet it also proactively monitors pages for extreme content that may violate its community standards. Crawford and Gillespie argue that Facebook's more "conservative" approach (as they refer to it) to governing content is due to its hosting a variety of types of content, including photo, video and text, at 7.

financially affects the intermediaries.[7] Third, individuals can manifest their discontent toward the extreme speech in the physical world by protesting (sometimes violently) against the speech, potentially forcing intermediaries to remove the controversial UGC to prevent further harm.[8] Each of these scenarios presents digital intermediaries with a set of conflicting interests not unlike those facing police in heckler's veto situations. Again, the comparison is not perfectly symmetric—no one has a right to prevent a digital intermediary from removing his or her message (unpopular, offensive, harmful or otherwise) from the intermediary's platform. The comparison being made here is not one of strict functional equivalence, whereby intermediaries are acting under color of law.[9] Rather, two different systems of governance over speech are being compared as if they were systems from two countries. Under such an analytical framework, "functional equivalence" refers to the similar challenges that each system of governance faces.[10] Understanding these challenges and how each system responds to them can reveal how each system values offensive, harmful or unpopular speech.

This chapter begins by broadly discussing the concept of the competition between audience rights and freedom of expression. This section builds on the analysis in chapter

---

[7] An example of this scenario is the pressure put on Facebook by activists to remove sexist and misogynist pages from the social networking site. *See Open Letter to Facebook*, WOMEN, ACTION, & THE MEDIA (May 21, 2013), *available at* http://www.womenactionmedia.org/facebookaction/open-letter-to-facebook/. *See infra* notes 241-242.

[8] The worldwide violence surrounding the "Innocence of Muslims" video on YouTube is an example of such a scenario. *See* Eva Galperin, *YouTube Blocks Access to Controversial Video in Egypt and Libya*, ELECTRONIC FRONTIER FOUNDATION (Sept. 12, 2013), *available at* https://www.eff.org/deeplinks/2012/09/youtube-blocks-access-controversial-video-egypt-and-libya.

[9] *Supra* note 4.

[10] *See, e.g.,* John C. Reitz, *How to Do Comparative Law*, 46 AM. J. COMP. L. 617 (1998); Ralf Michaels, *The Functional Method of Comparative Law*, *in* OXFORD HANDBOOK OF COMPARATIVE LAW (Mathias Reimann & Reinhard Zimmermann eds., 2006).

3, focusing on the issue of harm and offensiveness of speech, but from the perspective of the audience. The section contains an analysis of the various ways in which issues of audience rights can be conceived within First Amendment jurisprudence, focusing on case law that deals with speech that deeply offends audiences, and protection of minors from sexual or indecent speech within certain contexts. It will then look at Joel Feinberg's theory of "profound offense" and how it squares with jurisprudence regarding offensive speech. This section has two goals: 1) identifying the benefits of freedom of expression from an audience's perspective; and 2) understanding the theory and jurisprudence of offensive speech from an audience's perspective. Both will aid in understanding the development of heckler's veto case law.

The chapter turns next to an analysis of the jurisprudence of the heckler's veto. This section begins with an analysis of the foundational U.S. Supreme Court decisions in which the heckler's veto principle took shape. These cases span from 1940 to 1969, a period of political and social foment that was the crucible for the Civil Rights Movement and for an expansion in free speech rights.[11] The purpose of this analysis is to show how the Court dealt with the competing issues of protecting unpopular speech, preventing incitement to riot, and expounding upon police the duty to uphold the former while ensuring the latter. Next, the section will analyze federal cases involving hostile audiences from the last decade to highlight some of the major challenges facing heckler's veto jurisprudence, particularly challenges relating to the application of content-neutral laws to quell speech in hostile-audience situations. The section then looks at examples of

[11] *See, e.g.,* HARRY KALVEN, JR., THE NEGRO AND THE FIRST AMENDMENT (1965); GERALD N. ROSENBERG, THE HOLLOW HOPE: CAN COURTS BRING ABOUT SOCIAL CHANGE? (2008).

state disorderly conduct and breach of the peace statutes. The purpose of this analysis is to highlight that neither the case law nor the statutory law that set the parameters of the heckler's veto doctrine are crystal clear. Such lack of clarity is an important challenge found in heckler's veto scenarios both on street corners and in online content governance.

Section four will briefly review several First Amendment theories discussed in chapter 3, focusing on how they can be applied to understanding the concept of the heckler's veto. The goal of this section is to highlight a unique tension within the heckler's veto principle between negative and affirmative theories of the First Amendment. Section five will apply the values that shape the heckler's veto principle to the concept of content governance, focusing on scholarship analyzed in chapter 2. The goal of this section is to use the heckler's veto principle to illustrate how digital intermediaries respond to pressure from individuals to remove certain types of offensive or unpopular speech that enters the public discourse that they facilitate. As with the rest of this study, the goal of this analysis is to understand the values that the system of online content governance places on speech. In this particular analysis, the focus is on the competing values of speakers' ability to express unpopular and offensive speech and audiences' ability to utilize their own agentive powers to restrict such speech.

Audience Rights: Broadly Conceived

The concept of audience rights generally deals with the issue of which types of speech should or should not be allowed, within specific contexts, from an audience-centric rather than speaker-centric perspective. The principal criterion for settling these issues is the effect the speech will have on a particular audience, positive or negative

though primarily the latter. Using such a criterion places the discussion of audience rights within the context of issues of harmful speech. However, within First Amendment jurisprudence on issues related to harmful speech,[12] such an audience-centric perspective takes a back seat to traditional speaker-centric jurisprudence in all but a few contexts. Therefore, literature that advocates placing primacy on audience rights when it comes to adjudicating harmful speech tends to fall outside the mainstream of First Amendment theory, and doctrine only recognizes audiences' rights against harmful speech in a narrow set of circumstances (discussed below). This section reviews common themes in issues of audience rights theory and doctrine.

Issues of audience rights can be conceived in three ways. First, audience rights include the extent to which individuals have the right to receive certain information from government. Second, audience rights include the extent to which individuals have the right to hear the speech of other individuals or private entities. Third, audience rights involve the extent to which individuals have the right to avoid the speech of other individuals or private entities, and enforce that right either through prior restraints or after-the-fact punishment. These three conceptions of audience rights are connected in two ways. First, they all involve the receipt of information, the ability of audiences to use that information to function as engaged citizens within democratic society, and the determination of what types of information are integral to this function. Second (and related to the first point), they all involve the audience's own *speech* as much as an audience's *receipt* of speech. The receipt of information from government or private

[12] *See generally* Chapter 3.

sources is a crucial step for audiences to develop messages of their own to speak.[13] Also, when an audience is able to restrict the speech of others due to its right not to hear that speech, exercising that right is, at its core, a message spoken by the audience.[14]

The focus of this section—and certainly this entire chapter—is the third conception of audience rights: the extent to which audiences have (or at least perceive to have[15]) a right to have restrictions placed on speech so that they may not have to encounter it. This conception often conflicts with the rights of individuals to hear speech, the rights of individuals to receive information from government, and the rights of speakers to deliver the speech. Because of its focus on the harm or offense caused by speech, this third conception of audience rights intersects the discussion of harmful speech.[16] However, this analysis differs from the analysis in chapter 3 because it is looking at two competing rights: a speaker's right "both to seek out an audience and to be annoying, provocative, and offensive in public places";[17] and the extent to which an audience has the right to restrict the speaker's right.

The Right (Not) to Hear Speech

Meanwhile, in issues of audiences' rights to either hear or not hear information from individuals or other private entities, the factors that determine the extent of those rights are relatively less neatly defined. Firstly, the dominant strands of First Amendment theory discussed in chapter 3 give far greater deference to audiences' rights to hear

---

[13] Thomas I. Emerson, *Legal Foundations of the Right to Know*, 1 WASH. U. L. Q. 1, 2 (1976).
[14] *Id*.
[15] Audiences may lack a constitutional *right* to have speech restricted in most circumstances, but they may have many *interests* in restricting speech.
[16] *See generally* Chapter 3.
[17] Emerson, *supra* note 13, at 23.

speech rather than avoid it, due to the fact that the former is more compatible with these strands' primary focus on the rights of speakers.[18] Another one of the difficulties in neatly categorizing this context of audience rights issues is that the central factor (the conflict between the benefits and harms of a given message) is not only subjectively defined, but it also is coextensive with the definitions of the audiences in question. For example, any benefit of a message by the Ku Klux Klan is likely only best appreciated by other white supremacists, while the harm of the message is likely only best appreciated by non-whites. Other factors that can be difficult to untangle include the medium through which the message is delivered, whether the audience includes minors, and whether the rights of adult members of the audience to receive speech outweigh the interests of other audience members to prevent the purported harm to minors that certain speech may cause.

The U.S. Supreme Court has held in several major cases that audiences have no legal protection against the offensive messages of others in public places. In *Cohen v. California*,[19] a divided[20] Court held that California could not punish Paul Robert Cohen for walking through a Los Angeles courthouse wearing a jacket bearing the words "Fuck the Draft." Justice Harlan wrote that "the mere presumed presence of unwitting listeners or viewers does not serve automatically to justify curtailing all speech capable of giving offense."[21] If such a regime against offensive speech were to exist, Harlan wrote, it

---

[18] *See* C. EDWIN BAKER, HUMAN LIBERTY AND FREEDOM OF SPEECH (1989).
[19] 403 U.S. 15 (1971).
[20] The division in the 5-4 decision was over the issue of whether Cohen's wearing of the jacket was speech or conduct, the latter being punishable under the California statute in question, CAL. PEN. CODE § 415.
[21] Cohen, at 21.

"would effectively empower a majority to silence dissidents simply as a matter of personal predilections."[22] Harlan advised that "[t]hose in the Los Angeles courthouse could effectively avoid further bombardment of their sensibilities simply by averting their eyes."[23] This "avert-your-eyes" line of reasoning reflects the dominant jurisprudence toward offensive messages delivered in public places. It acknowledges that the offense it may cause is very real to some, but it maintains that the legal response to such speech is to respect the social value of its message over any offense it may cause.

Similarly, in *Texas v. Johnson*,[24] a divided Court held that a Texas law punishing "severe acts of physical abuse" of the American flag in a manner "intentionally designed to seriously offend other individuals"[25] was inconsistent with the First Amendment.[26] Writing for the Court, Justice Brennan posited, "If there is a bedrock principle underlying the First Amendment, it is that the government may not prohibit the expression of an idea simply because society finds the idea itself offensive or disagreeable."[27] However, in a rather confusing dissent, Chief Justice Rehnquist argued the Texas law did not violate the First Amendment because it did not punish Gregory Lee Johnson for his viewpoint (disgust with America) but rather the manner in which he expressed it (burning the flag). Rehnquist wrote:

> [I]n no way can it be said that Texas is punishing him because his hearers—or any other group of people—were profoundly opposed to the message that he sought to convey. Such opposition is no proper basis for

[22] *Id*.
[23] *Id*.
[24] 491 U.S. 397 (1989).
[25] *Id*. at 411.
[26] *Id*. at 399.
[27] *Id*. at 414.

> restricting speech or expression under the First Amendment. It was Johnson's use of this particular symbol, and not the idea that he sought to convey by it or by his many other expressions, for which he was punished.[28]

Rehnquist then ponders the question of why matters of deeply offensive expression should be left to courts to rule on, rather than legislatures: "Surely one of the high purposes of a democratic society is to legislate against conduct that is regarded as evil and profoundly offensive to the majority of people—whether it be murder, embezzlement, pollution, or flag burning."[29] Rehnquist's argument presupposes a second argument: that flag burning is conduct, not speech, and is therefore subject to regulation. This argument begs the question: at what point does the context of expressing the message—which may be the primary harmful factor in the overall expression of the message—become unprotected conduct rather than speech? At least with flag burning, the majority in *Johnson* holds that this activity does not cross this line.

But what about instances in which speech offends not simply conventional decency, but the very identity of minority communities? Legal theorist Joel Feinberg argues that the Nazis' use of the swastika in Skokie would have been "gratuitous," and "in no obvious way necessary to the content of the advocacy" of racial superiority.[30] Just as Johnson could have expressed his anti-American message by, say, shouting through a bullhorn rather than burning the flag, the Nazis could have simply handed out leaflets while wearing normal clothing. Only, to Feinberg, the context of the swastika is so offensive to Jews whose ancestors were killed under its banner that its use crosses the

[28] *Id*. at 432 (Rehnquist, C.J., dissenting).
[29] *Id*. at 435 (Rehnquist, C.J., dissenting).
[30] *Id*. at 88.

line from expression to harassment, i.e. regulable conduct.[31] However, in striking down a city ordinance invoked to prevent the National Socialist Party of America from marching in Skokie, Illinois (a town populated with numerous Holocaust survivors) due to its overbreadth, the U.S. Court of Appeals for the Seventh Circuit quoted the following passage from the U.S. Supreme Court case *Street v. New York*:[32]

> [A]ny shock effect . . . must be attributed to the content of the ideas expressed. It is firmly settled that under our Constitution the public expression of ideas may not be prohibited merely because the ideas are themselves offensive to some of their hearers.[33]

When they are the targets of such hateful speech, audiences are only able to invoke clearly established exceptions to First Amendment rights to freedom of expression to restrict the speech or punish the speaker. These exceptions, enumerated in chapter 3, include: fighting words,[34] true threats,[35] and incitement to imminent lawless action.[36] Audiences also have the right to invoke tort law to punish speech that is allegedly harmful to them, such as defamation or invasion of privacy. Protection from the harms of obscenity[37] and false advertising[38] are also available to audiences.

Meanwhile, laws that restrict speech prior to its being spoken in the name of preventing harm to an audience are very rare, confined almost exclusively to the interest in protecting minors. Indecency standards for broadcasting make up one such area of law,

---

[31] JOEL FEINBERG, THE MORAL LIMITS OF THE CRIMINAL LAW: OFFENSE TO OTHERS (1985). *See infra* note 63.
[32] 394 U.S. 576, 592 (1969).
[33] Collin v. Smith, 578 F.2d 1197, 1206 (7th Cir. 1978).
[34] Chaplinsky v. New Hampshire, 315 U.S. 568 (1942).
[35] Virginia v. Black, 538 U.S. 343 (2003).
[36] Brandenburg v. Ohio, 395 U.S. 444 (1969).
[37] Miller v. California, 413 U.S. 15 (1973).
[38] Virginia Pharmacy Board v. Virginia Consumer Council, 425 U.S. 748, 776 (1976) (Stewart, J., concurring).

if not the chief area. In the landmark case *FCC v. Pacifica*,[39] a divided Supreme Court held that the FCC's regime for restricting indecent speech[40] from broadcast from 6 a.m. to 10 p.m. did not violate the First Amendment rights of broadcasters. Distinguishing *Cohen*, Justice Stevens wrote that "broadcasting is uniquely accessible to children, even those too young to read. Although Cohen's written message might have been incomprehensible to a first grader, Pacifica's broadcast could have enlarged a child's vocabulary in an instant."[41] Although the general prohibitions of *Pacifica* remain intact, the FCC is presently (as of this writing) revisiting its policy on whether so-called "fleeting expletives" should be subject to sanctions following the 2012 *FCC v. Fox* case.[42]

Indecency standards are unique in that they make up a relatively well-defined area of law that is specific to a particular medium: broadcasting. Government enjoys a significant interest in regulating broadcast due to the scarce nature of the electromagnetic spectrum (a public good) upon which broadcasting relies.[43] Courts have thwarted attempts to protect minors from exposure to "indecent" material by legislation that seeks to transpose indecency standards to other media (namely the Internet), because they have found not found a similar significant government interesting in regulating these media.[44] Minus this interest, courts follow the standard that a law that bans speech for the sake of protecting minors from it, yet in doing so prevents adults from accessing such otherwise

[39] 438 U.S. 726 (1978).
[40] Defined as "language that describes, in terms patently offensive as measured by contemporary community standards for the broadcast medium, sexual or excretory activities and organs." *Id*. at 772, FN 7.
[41] *Id*. at 749.
[42] FCC v. Fox Television Stations, Inc., 132 S. Ct. 2307 (2012).
[43] Red Lion Broadcasting Co. v. FCC, 395 U.S. 367, 389 (1969).
[44] Reno v. ACLU, 521 U.S. 844, 878 (1997).

lawful speech, is unconstitutional.[45] As Justice Felix Frankfurter famously put it, such laws are akin to "burn[ing] the house to roast the pig."[46]

One final audience-rights issue to consider is whether Frankfurter's famous line applies to multiple adult audiences who are affected in drastically distinct ways by offensive speech. In *Snyder v. Phelps*, the U.S. Supreme Court held that Albert Snyder, the father of a Marine killed in Iraq, could not bring an action of intentional infliction of emotional distress (IIED) against the Topeka, KS-based Westboro Baptist Church for picketing near his son's funeral with signs containing such hateful messages as "Thank God for Dead Soldiers" and "You're Going to Hell."[47] Writing for the 8-1 majority, Chief Justice John Roberts held that the Church's fundamentally social and political message was a matter of public concern and therefore entitled to First Amendment protection.[48] Snyder had contended that the context of the Church's message—using his son's funeral as a platform—should strip the Church of its First Amendment protection because it transformed its speech into a matter of private concern rather than public concern.[49] However, Roberts held that "[t]he fact that Westboro spoke in connection with a funeral … cannot by itself transform the nature of Westboro's speech."[50] Not only was the Church's message a matter of public concern, but Roberts also noted that the Church

---

[45] Bolger v. Young Drug Products Corp., 463 U.S. 60, 74 (1983) (holding that "[t]he level of discourse reaching a mailbox simply cannot be limited to that which would be suitable for a sandbox"). *But cf.* Ginsberg v. New York, 390 U.S. 629, 639 (1968) (holding that "[t]he well-being of its children is of course a subject within the State's constitutional power to regulate," and that a statute requiring stores to restrict access of "sex materials" to children—though not adults—comprises such a subject).
[46] Butler v. Michigan, 352 U.S. 380, 383 (1957).
[47] Snyder v. Phelps, 131 S.Ct. 1207, 1213 (2011).
[48] *Id*. at 1219.
[49] *Id*. at 1217.
[50] *Id*.

made sure to abide by state and local requirements to picket only on public property more than 1,000 feet away from the funeral rather than immediately outside the funeral.[51]

The lone dissenter in the case, Justice Samuel Alito, sympathized with Snyder's "context" argument, arguing that Westboro's speech was "part of a cold and calculated strategy to slash a stranger as a means of attracting public attention."[52] Moreover, Alito distinguished *Snyder* from *Hustler v. Falwell*,[53] whose precedent of subjecting IIED claims to the *New York Times*[54] actual malice standard formed the jurisprudential foundation for the majority's reasoning. *Hustler* involved a public figure (the Rev. Jerry Falwell), while Alito contended that Snyder was clearly a private individual.[55] "[T]he caricature [in *Huslter*] does not have the same potential to wound as a personal verbal assault on a vulnerable private figure," Alito wrote.[56] What is interesting about Alito's dissent is his application of Kantian ethics to a legal analysis that favors Snyder and other private individuals caught in similar future situations.[57] To Alito, the Westboro Baptist Church robbed Snyder of his dignity by treating him as a means to an end (the Church's campaign) rather than an end in himself. To Alito, such an affront to individual dignity is a graver harm than the harm done to the public by punishing an organization for speaking on a matter of public concern, thereby potentially depriving that public of hearing the Church's message due to a chilling effect from the threat of litigation. In other words,

---

[51] *Id*. at 1218.
[52] *Id*. at 1227 (Alito, J., dissenting).
[53] 485 U.S. 46 (1988).
[54] New York Times Co. v. Sullivan, 376 U.S. 254 (1964).
[55] Snyder v. Phelps, 131 S.Ct. 1207, 1228 (2011) (Alito, J., dissenting).
[56] *Id*.
[57] KANT, I. (1785/2005). GROUNDWORK FOR THE METAPHYSIC OF MORALS, Trans: J. Barrett (2005), *available at* http://www.earlymoderntexts.com/pdfs/kant1785.pdf, at 29.

when speech has a broad audience and a specific one, and the speakers' goal is to reach the broad audience at the expense of causing harm to the specific one, Alito would have no problem with the specific audience taking action against the speakers at the expense of the broader audience. To better understand the reasoning behind this argument, one must explore the nature of the potential harm that such speech could cause to a private individual such as Snyder.

Feinberg and "Profound Offense"

Issues of audience rights often involve a disconnect between the law's deference to offensive speech and audiences' willingness to stand idly by and "take it." Feinberg explores this disconnect in his four-volume work *The Moral Limits of the Criminal Law*, particularly Volume II, titled *Offense to Others*.[58] In this volume, Feinberg delves into the question of when certain actions—including both conduct and speech—are deserving of criminal punishment. When it comes to speech, Feinberg contends that only speech that "personally harasses" an individual can be subject to legal sanction; if not, the person should simply "leave the provocation behind," which "is what the law should require of him, if he can do it without loss or hardship."[59] In other words, Feinberg agrees with the U.S. Supreme Court that the fighting words doctrine should be the line separating protected from unprotected speech.[60]

Still, Feinberg wrestles with the conundrum of whether and how the law should treat speech that does not personally harass yet still causes a profound offense. An

[58] FEINBERG, *supra* note 31.
[59] *Id*. at 91.
[60] *Id*. *See infra*, notes 148-154.

offense is profound, according to Feinberg, when the thing that individuals "feel to be violated or affronted is something they hold precious (human dignity, solidarity with martyred kinsmen)."[61] "Profound offense cannot be avoided by averting one's eyes," Feinberg argues, as the sheer knowledge that the speech is taking place is enough to cause profound offense.[62] For Feinberg, hateful speech, such as the marching of Nazis in Skokie, Illinois, or cross burnings by the Ku Klux Klan, can cause profound offense. "The main distinguishing feature[] of the swastika and K.K.K. emblems is their deliberate association with actual historic atrocities—lynchings, tortures, mass killings committed to vindicate the alleged prerogatives of a master race," he contends.[63] The offense felt by Jews at the sight of—or even the imagined presence of—the swastika would consist of "a complex mental state, compounded of moral indignation and disapproval, resentment …, and perhaps some rage or despair."[64]

Although devising an objective test for determining the level of speech's offensiveness may be impossible, Feinberg argues that there are six factors that contribute to offensiveness,[65] five of which are germane to the present discussion.[66] These include: 1) the social value of the speech; 2) the extent to which the speech

[61] FEINBERG, at 60.
[62] *Id*. at 59.
[63] *Id*. at 93.
[64] *Id*. at 87.
[65] *Id*. at 44. Feinberg's focus is on "conduct," which, in his conception of the term, includes speech. To avoid the confusion of jurisprudential differences between conduct and speech, the discussion in this paragraph will focus only on Feinberg's factors as they relate to speech.
[66] Feinberg's first of the six factors, which he calls the "personal importance" of the offending action, is based on the tenuous logic that the more personally important the action is to the actor, the more objectively reasonable the action is.

expresses an opinion on social or political matters;[67] 3) whether individuals have alternative opportunities to express their message in a less offensive place or manner; 4) the extent to which the speech is motivated by malice or spite; and 5) whether speakers deliberately chose to convey their message in a location that would amplify the offensiveness of their message.[68] Apart from the narrow doctrinal exceptions to First Amendment protection listed above, First Amendment jurisprudence does not entertain a calculus based on these five factors. Speech with minimal social value under Feinberg's scheme receives the same protection as speech with much greater social value. Nevertheless, Feinberg's factors remain valuable for the present analysis because they can be viewed as factors that lead to heckler's veto scenarios, whether on street corners or in content governance.

The First Amendment does not take away the realness of the harms (whether profound harm or mere offense) suffered by audiences who would wish to restrict the speech that caused them.[69] Audiences are not going to roll over and willfully accept such harmful speech simply because the law protects it. In some instances in which the speech in question is expressed in public, audiences may react against the speakers with counter-speech—which itself may even escalate into violence. Such scenarios pitting speakers'

---

[67] FEINBERG, at 44. Here, Feinberg contends, "Expressions of opinion … must be presumed to have the highest social importance in virtue of the great social utility of free expression and discussion generally, as well as the vital personal interest most people have in being able to speak their minds fearlessly. No degree of offensiveness in the expressed opinion itself is sufficient to override the case for free expression, although the offensiveness of the manner of expression, as opposed to its substance, may have sufficient weight in some contexts." *See* Feinberg's argument on this point discussed *supra* note 31.

[68] Here it appears rather obvious that Feinberg is referring to the Skokie case.

[69] *See, e.g.,* Clay Calvert, *Hate Speech and Its Harms: A Communication Theory Perspective*, 47 J. COMM. 4 (1997); DANIELLE KEATS CITRON, HATE CRIMES IN CYBERSPACE (2014); MARI J. MATSUDA, ET AL., WORDS THAT WOUND: CRITICAL RACE THEORY, ASSAULTIVE SPEECH, AND THE FIRST AMENDMENT 74 (1993).

rights against audience rights have their own doctrine with its own set of cases and theory: the heckler's veto.

The Heckler's Veto: Jurisprudence and Practice

Consider a scenario in which two groups with opposing viewpoints meet in a public forum in the United States. Here, several interests from each "stakeholder" group in the scenario potentially come into conflict. First, the group that initially sought access to the public forum wishes—indeed, it has the right—to speak its message.[70] Second, the opposition group wishes—indeed, it too has the right—to speak a counter message,[71] and thereby attempt to outcompete the initial group's message in this microcosm of the marketplace of ideas. Neither the initial speakers nor their opposition want to have their right to speak trampled upon. However, it is not hard to conceive of a situation where the message of the initial group is so offensive to the opposition that it spurs the opposition to react not merely vociferously but violently to defend their honor or dignity from attack by the initial group's message.[72] Therefore, third, law enforcement officials must ensure that the clash of opposing messages does not escalate into a physical confrontation that could endanger lives—of the protestors, the counter-protestors, or innocent bystanders—or that could cause damage to surrounding property.[73] These officials essentially must referee the competition of messages, making sure that each side's First Amendment right to

[70] *See, e.g.,* Capitol Square Review & Advisory Bd. v. Pinette, 515 U.S. 753 (1995) (holding that a city advisory board could not prevent the Ku Klux Klan from burning a cross in a public forum, either under the Establishment Clause, or under the Speech Clause (Thomas, J., concurring)).
[71] *See, e.g.,* Marcavage v. City of Philadelphia, 778 F.Supp.2d 556 (E.D. Pa. 2012), (holding that a counter-protestor has a First Amendment right to speak a message to counter another group, so long as the message does not infringe upon the rights of the original group).
[72] FEINBERG, *supra* note 61.
[73] *See, e.g.,* Cottonreader v. Johnson, 252 F. Supp. 492, 497 (M.D. Ala. 1966).

speak is honored, while preventing either group from violating the rights of the other. Ultimately, these officials are left with a choice: stop the speech of the initial group to quell the potential violence, or protect the speakers of the initial group from their opposition and any violent reaction that may ensue. The theory and jurisprudence that surround this choice are the focus of this chapter.

This section begins by examining six cases decided by the U.S. Supreme Court that deal with speakers who confronted an audience that was hostile toward their message. These six cases are by no means an exhaustive sample of cases appearing before the Supreme Court that deal with hostile audiences. However, these cases sufficiently illustrate the principle of the heckler's veto. Three of these cases preceded the Civil Rights Movement, with the other three occurring during the heart of the Civil Rights Movement.[74] The speaker prevailed in all but one of the cases; however, that one case (*Feiner v. New York*[75]) represents bad law in the line of cases.[76] The cases will be discussed briefly in chronological order to analyze the development of the heckler's veto principle. Following this analysis, the close relationship of the heckler's veto principle to both the *Brandenburg* incitement standard and the fighting words doctrine will be discussed.

---

[74] There is no agreed upon definition of when the "heart" of the Civil Rights Movement began, as the NAACP won significant legal victories as early as the first quarter of the 20th century. However, historian Jacquelyn Dowd Hall argues that the Movement "accelerated" in the 1950s and 1960s, no doubt due in great part to *Brown v. Board of Educ. of Topeka*, 347 U.S. 483 (1954), and the Montgomery bus boycott in 1955. Jacquelyn Dowd Hall, *The Long Civil Rights Movement and the Political Uses of the Past*, 91 J. AMER. HIST. 1233, 1244 (2005). *See also* ALDON D. MORRIS, THE ORIGINS OF THE CIVIL RIGHTS MOVEMENT: BLACK COMMUNITIES ORGANIZING FOR CHANGE 14 (1984).

[75] 340 U.S. 315 (1951).

[76] *See infra* note 109 and accompanying text.

Formation of the Doctrine

In 1940, the Supreme Court handed down its decision in the first hostile audience case examined here: *Cantwell v. Connecticut*.[77] Cantwell, a Jehovah's Witness, was convicted of committing common law incitement to a breach of the peace.[78] Cantwell and his two sons had stood on the corner of a public street that was known to be populated predominantly by Catholics and played a record that attacked Catholicism. At one point, Cantwell played the record for two passersby. The men, both Catholic, testified in the Court of Common Pleas of New Haven County that they "were tempted to strike Cantwell unless he went away."[79] The Connecticut State Supreme Court upheld Cantwell's conviction, holding that Cantwell's charge was not breach of the peace on his own part, but rather "invoking or inciting others to breach of the peace."[80] A unanimous U.S. Supreme Court overturned Cantwell's conviction on the grounds that his speech did not create a "clear and present danger … to public safety, peace, or order[.]"[81]

However, writing for the Court, Justice Owen Roberts paid great deference in *dicta* to states' interests in protecting order by punishing speech acts that exceeded the clear and present danger standard. "One may … be guilty of the offense [of incitement to breach of the peace] if he commit acts or make statements likely to provoke violence and disturbance of good order, even though no such eventuality be intended,"[82] he wrote. "The danger in these times from the coercive activities of those who in the delusion of

[77] 310 U.S. 296 (1940).
[78] *Id.* at 300.
[79] *Id*. at 303.
[80] *Id*.
[81] *Id*. at 308.
[82] *Id*. at 309.

racial or religious conceit would incite violence and breaches of the peace in order to deprive others of their equal right to the exercise of their liberties, is emphasized by events familiar to all. These and other transgressions of those limits the States appropriately may punish."[83] Two conclusions can be drawn from Roberts's *dicta*. First, it does not give speakers the strong protection for messages of incitement that would later be solidified in *Brandenburg*. Thus, *Cantwell* leaves the door open for future convictions of speakers who incite a hostile audience to a violent reaction. On a similar note, it also acknowledges the difficulty that both law enforcement and courts would face in balancing the competing interests of protecting speech and maintaining order, as a bright-line test for speech that would incite hostile reactions across any set of facts did not exist. Second, Roberts's *dicta* is an acknowledgement that the Court was in the midst of a tumultuous period in which incendiary speech that proposed radical ideas was becoming the norm.[84]

The Court next decided a hostile audience case nine years later in *Terminiello v. Chicago*.[85] In this case, controversial priest Arthur Terminiello delivered a racist, anti-Semitic and anti-communist speech at a Chicago auditorium. Terminiello's mere presence at the auditorium led to a gathering of about 1,000 protestors, who picketed, cursed at Terminiello, threw bottles and stones at police, and threw stones and bricks

[83] *Id*. at 310.
[84] *See* KALVEN, *supra* note 11; G. Edward White, *The First Amendment Comes of Age: The Emergence of Free Speech in Twentieth-Century America*, 95 Mich. L. Rev. 299 (1996); Norman L. Rosenberg, *Another History of Free Speech: The 1920s and the 1940s*, 7 LAW & INEQ. 333 (1988).
[85] 337 U.S. 1 (1949).

through the auditorium windows.[86] Like Cantwell, Terminiello was arrested and later convicted on charges of inciting disorderly conduct.[87] The U.S. Supreme Court reversed Terminiello's conviction in a 5-4 decision, holding that the city ordinance under which Terminiello was charged was unconstitutional. Justice Douglas wrote for the Court, "[Speech] may indeed best serve its high purpose when it induces a condition of unrest, creates dissatisfaction with conditions as they are, or even"—and here Douglas used the exact words of the ordinance used to charge Terminiello—"stirs people to anger."[88] Douglas continued, "Speech is often provocative and challenging. It may strike at prejudices and preconceptions and have profound unsettling effects as it presses for acceptance of an idea."[89] Douglas's language in favor of protecting speech is noticeably stronger than Roberts's from *Cantwell*. It acknowledges that the very potential of riling up a hostile audience is one of the key characteristics of speech worth protecting under the First Amendment.

However, like Roberts, Douglas held that protection of provocative messages is not absolute. As Roberts did in *Cantwell*, Douglas held that states must prove that a speaker's message is so provocative that it presents a clear and present danger—a fairly high, though still nebulous, standard—before they could punish the speaker.[90] The relative ambiguity of the clear and present danger standard again left the door open for future courts to recognize a state's interest in preventing speakers from inciting their

[86] *Id*. at 15.
[87] *Id*. at 3.
[88] *Id*. at 4.
[89] *Id*.
[90] *Id*.

hecklers to commit violence. Justice Jackson's dissent shows just how wide open the Court left the door. Jackson's chief problem with the majority's decision in *Terminiello* was its naïveté. He agreed with the majority's "generalized approbations of freedom of speech with which, in the abstract, no one will disagree." But, he argued, "the local court that tried Terminiello was not indulging in theory."[91] Thus, Jackson continued, the majority "fixes its eyes on a conception of freedom of speech so rigid as to tolerate no concession to society's need for public order."[92] He argued that the danger of such a national standard was that local communities that have expressed a strong interest in protecting order would fall into chaos, especially in urban areas.[93] He interpreted Terminiello's actions as "a local manifestation of a world-wide and standing conflict between two organized groups of revolutionary fanatics, each of which has imported to this country the strong-arm technique developed in the struggle by which their kind has devastated Europe."[94] Similarly, Chief Justice Vinson's strong dissent in *Terminiello* is indicative of a broader conflict of ideologies confronting the Court during this time, a conflict deeply influenced by both the European wartime experience and the start of the Cold War and Red Scare.[95] On one side was an interpretation of freedom of speech as quintessentially anti-totalitarian. This philosophy was based on Meiklejohn's vision of freedom of expression as forming the essential link between liberty and democracy.[96] On the other side was a real preoccupation with the potential of incendiary speech to lead to

[91] *Id*. at 13 (Jackson, J., dissenting).
[92] *Id*. at 14 (Jackson, J., dissenting).
[93] *Id*. at 23 (Jackson, J., dissenting).
[94] *Id*.
[95] White, *supra* note 84, at 343; Rosenberg, *supra* note 84, at 360.
[96] White, *supra* note 84, at 331.

government overthrow and revolution.[97] The clear and present danger standard allowed enough gray area for competing jurists to make cases in favor or either greater protection for or suppression of provocative speech, much to the lament of Meiklejohn.[98]

Indeed, this conflict of ideologies was so intense that speakers did not always prevail in hostile audience cases. In the most notable example, *Feiner v. New York*,[99] Vinson's dissenting arguments from *Terminiello* won the day. In that case, Feiner, a white college student, stood on a box at a public street corner in Syracuse, New York, and through a loudspeaker told a crowd of about 80 people, composed of both blacks and whites, that the mayor of Syracuse was "a champagne-sipping bum" who did "not speak for the negro people," that "President Truman [was] a bum," and that "negroes don't have equal rights" and therefore "they should rise up in arms and fight for their rights."[100] The police testified at Feiner's trial that they "were concerned with the effect of the crowd on both pedestrian and vehicular traffic," because the "crowd was restless and there was some pushing, shoving and milling around."[101] The police also testified that an angry person in the crowd told them that if they did not get "that son-of-a-bitch" off the box, he would.[102] After Feiner refused the officers' third order to stop speaking, the police arrested him and charged him with disorderly conduct. The Supreme Court upheld Feiner's conviction in a 6-3 decision. Writing for the majority, Vinson held that the Court must respect "the interest of the community in maintaining peace and order on its

[97] *Id*. at 343.
[98] Rosenberg, *supra* note 84, at 360.
[99] 340 U.S. 315 (1951).
[100] *Id*. at 330.
[101] *Id.* at 317.
[102] *Id*. at 324.

streets," and that the Court "cannot say that the preservation of that interest here encroaches on the constitutional rights of [Feiner]."[103] Vinson gave special deference to law enforcement in his decision, writing, "It is one thing to say that the police cannot be used as an instrument for the suppression of unpopular views, and another to say that, when as here the speaker passes the bounds of argument or persuasion and undertakes incitement to riot, they are powerless to prevent a breach of the peace."[104] In a concurring opinion to *Feiner*, published in the record for the case *Niemotko v. Maryland*,[105] which was heard the same day as *Feiner*, Justice Frankfurter appeared to strengthen the Court's deference toward law enforcement. "It is not a constitutional principle that, in acting to preserve order, the police must proceed against the crowd, whatever its size and temper, and not against the speaker," Frankfurter wrote.[106]

The majority's argument was met with sharp dissents from Justice Black, the notorious First Amendment absolutist, and Justice Douglas, author of the *Terminiello* decision. Black refuted Vinson's argument that the police were in the right to arrest Feiner in the name of maintaining order. Rather, to maintain order, Black argued that "[the police's] duty was to protect petitioner's right to talk, even to the extent of arresting the man who threatened to interfere."[107] Echoing Black, Douglas wrote, "Police censorship has all the vices of the censorship from city halls."[108]

---

[103] *Id*. at 320.
[104] *Id*. at 321.
[105] 340 U.S. 268 (1951).
[106] *Id*. at 289 (Frankfurter, J., concurring).
[107] Feiner, at 327 (Black, J., dissenting).
[108] *Id*. at 331 (Douglas, J., dissenting).

The reason why *Feiner* contrasts so sharply with *Terminiello* may be as simple as who was sitting on the Court for each case.[109] However, the aberrant status of *Feiner* may reflect a broader issue: the difficulty courts face in judging the actions of law enforcement officials when confronted with hostile-audience situations. These officials must make time-sensitive decisions on whether to silence speech based on subjective interpretations of the events playing out before them despite lofty constitutional principles that demand their restraint. This ideological struggle over the gulf between theory and practice will be illustrative when examining intermediaries' position as arbiters of content when faced with pressure from individuals to remove offensive content.

Three hostile audience cases from the Civil Rights era—*Edwards v. South Carolina*,[110] *Cox v. Louisiana*,[111] and *Gregory v. Chicago*[112]—moved the Court away from its *Feiner* decision and toward a stance of protecting speech in the face of hostile opposition. The relatively similar facts from all three cases can be summarized together. All three cases involved a group of African-Americans who marched on public property to protest segregation. In each case, the protestors confronted large crowds of angry

---

[109] The five Justices who comprised the majority in *Terminiello* were Douglas, Reed, Rutledge, Black, and Murphy. Murphy's replacement, Tom C. Clark, essentially provided the swing vote in *Feiner*. Clark routinely voted in the same bloc as Vinson and Frankfurter, *see* MICHAL R. BELKNAP, THE VINSON COURT: JUSTICES, RULINGS, AND LEGACY 58 (2004). Justice Reed, who voted with the majority in *Terminiello*, switched his position and voted in the majority in *Feiner*. Reed's switch could have been due to his strong friendship with Jackson and Frankfurter before his years on the Court, *see* John D. Fassett, *The Buddha and the Bumblebee: The Saga of Stanley Reed and Felix Frankfurter*, 35 J. SUP. CT. HIST. 166 (2010). Thus, the fact that two cases with very similar sets of facts could be decided differently due to (what appears to be) simple voting patterns highlights the contentious nature of the ideological battle between free speech and public order in hostile audience cases at this time.

[110] 372 U.S. 229 (1963).

[111] 379 U.S. 536 (1965).

[112] 394 U.S. 111 (1969).

whites. In each case, some of the African-American protestors were arrested and charged with violating disorderly conduct or breach of the peace statutes, with police officers testifying that they made the arrests because the crowds were growing restive and the protestors repeatedly disobeyed the officers' orders to disperse. In each case, the African-American protestors prevailed. In *Cox*, Justice Goldberg adopted Black and Douglas' argument that the police's duty was to maintain order by protecting the speakers, not appeasing the angry hecklers. "[The police] could have handled the crowd," Goldberg wrote.[113] Black expounded the argument from his *Feiner* dissent in his concurrence in *Gregory*: "[U]nder our democratic system of government, lawmaking is not entrusted to the moment-to-moment judgment of the policeman on his beat."[114]

However, this lack of trust in the "moment-to-moment judgment" of police does not mean police must always stay inactive when faced with two competing groups. In another Civil Rights era case from Alabama, a federal judge enjoined police in Greenville, Alabama from failing to provide protection to African-American protesters; the police had refused to protect a group of picketing African-Americans from physical attacks by a mob of angry whites.[115] This case can be considered a period piece; it is likely that the Greenville police willingly refrained from offering protection to the African-American protesters out of spite for the protesters' message rather than out of reverence for extreme "survival-of-the-fittest" neutrality. In other words, such subjective cruelty is reasonably unlikely to occur today under similar circumstances. However, the

[113] Cox v. Louisiana, 379 U.S. 536, 550 (1965).
[114] Gregory v. Chicago, 394 U.S. 111, 120 (1969) (Black, J., concurring).
[115] Cottonreader v. Johnson, 252 F. Supp. 492, 497 (M.D. Ala. 1966).

important takeaway from this case is that the status quo for police in hostile audience cases is to protect speakers from the hostile audience, rather than remaining completely neutral.

The fact that these cases deal with issues of civil rights—and that two of them come from the South—should not be overlooked. Professor Harry Kalven argued that these cases left the Supreme Court with a dilemma: "require that in the South the police go down with the Negro speakers[,]" or "permit the South one gigantic hecklers' veto[.]"[116] He argued that the Court chose the former option because, for one, it did not want to succumb to the will of segregationists after not doing enough to curb McCarthyism in the 1950s.[117] Indeed, Kalven argued that the Civil Rights Movement provided the impetus for the theoretical and doctrinal crystallization of the heckler's veto principle from its contentious beginnings in *Cantwell* and *Terminiello*.[118] Moreover, this historical context shows why defeating the heckler's veto is so important: political speech that is essential for the social transformation of the country is imperiled if hecklers can so easily force its suppression. No matter how unpopular the speech, and regardless of whether the hecklers are expressing the will of the majority, the value of that speech merits affirmative steps to protect it. Only if the speech itself—not the actions of the hecklers reacting to the speech—crosses a legally defined threshold and directly causes physical harm can that speech be punished.

---

[116] KALVEN, *supra* note 11, at 141.
[117] *Id*. at 114. *See also* L. A. Powe, Jr., *Brandenburg: Then and Now*, 44 TEX. TECH L. REV. 69, 73 (2011).
[118] KALVEN, *supra* note 11, at 141.

Challenges within the Doctrine

The heckler's veto doctrine appears to be pretty cut-and-dried following its mid-century formation and Civil-Rights-Era refinement. However, courts have been faced with challenging questions as to how the principles of the doctrine should be upheld in practice.

Arguably the biggest challenge facing the heckler's veto doctrine is the extent to which laws that regulate heckler's veto scenarios are content-neutral. Police may restrict a speaker from speaking a message that could rile up a hostile audience if the speaker violates a generally applicable, content-neutral law. For example, the U.S. Court of Appeals for the Ninth Circuit held that police restricting speech from loudspeakers after residents complained about the noise did not amount to a heckler's veto.[119] The court held that the fact that the message contained Christian and anti-gay themes and was directed at residents of a predominantly homosexual district of San Francisco was immaterial because the police were acting only on the noise complaints, not any potential complaints about the content of the message.[120]

Similarly, in *Ovadal v. City of Madison*,[121] police stopped Ralph Ovadal from displaying anti-gay signs on an overpass over a busy highway because they believed the messages were distracting drivers by making them angry, thereby putting the drivers at greater risk of getting into an accident.[122] The U.S. Court of Appeals for the Seventh Circuit held that the police had a legitimate interest in stopping the speech, and that the

[119] Rosenbaum v. City and County of San Francisco, 484 F.3d 1142 (9th Cir. 2007).
[120] *Id*. at 1159.
[121] 469 F.3d 625 (7th Cir. 2006).
[122] *Id*. at 627.

ordinance cited to stop the speech was content-neutral.[123] The court disagreed with Ovadal's argument that the police were simply kowtowing to a heckler's veto in the form of the angry drivers on the highway below his message, or that the ordinance in question vaguely gave police too much discretion to determine how much of a hazard his speech actually posed to traffic.[124]

Although Ovadal did not prevail, his case did raise an important point: the content-neutrality of a law used to stop speakers in a heckler's veto situation is not always clear. In *Bible Believers v. Wayne County*,[125] officers from the Wayne County Sheriff's Office threatened to issue citations for disorderly conduct to a group of Christians who were attempting to preach at the public Arab International Festival in Dearborn, Michigan. The group's speech comprised shirts that read "Only Jesus Christ Can Save You From Sin and Hell," as well as a severed pig's head on a pike and invectives calling the Prophet Muhammad a pedophile.[126] Several of the predominantly Muslim festival attendees, shouted angry responses at the Christian group, while some threw water bottles and trash at them, and one person physically pushed one of the group members to the ground.[127] A divided panel of the U.S. Court of Appeals for the Sixth Circuit held that the purpose of the threat of citations—to regulate the safety of festival attendees—was content-neutral, and therefore the threat of citations did not amount to a

[123] *Id*.
[124] *Id*. at 629.
[125] 765 F.3d 578 (6th Cir. 2014) (rehearing *en banc* granted, opinion vacated, Oct. 23, 2014).
[126] *Id*. at 584.
[127] *Id*.

heckler's veto.[128] The majority agreed with the district court's use of *Feiner* to hold that the officers "were not powerless to prevent a breach of the peace in light of the imminence of greater disorder that Plaintiffs' conduct created."[129] The majority held that the Bible Believers "*intended* to incite the crowd to turn violent,"[130] and therefore the officers' threats of citing the group for disorderly conduct unless it dispersed "was objectively necessary under the circumstances."[131] In dissent, Judge Eric L. Clay began by stating simply, "This is an easy case."[132] Clay contended that the officers here fell victim to the heckler's veto, as the Bible Believers violated neither the incitement standard nor the fighting words doctrine, either of which would have stripped their speech of constitutional (and thereby police) protection. Clay warned that the majority's holding that the Believers' speech was, in fact, incitement speech was the beginning of a precipitous slippery slope:

> It does not take much to see why law enforcement is principally required to protect lawful speakers over and above lawbreakers. If a different rule prevailed, this would simply allow for a heckler's veto under more extreme conditions. Indeed, hecklers would be incentivized to get *really* rowdy, because at that point the target of their ire could be silenced. More perniciously, a contrary rule would allow police to manufacture a situation to chill speech. Police officers could simply sit by as a crowd formed and became agitated. Once the crowd's agitation became extreme, the police could swoop in and silence the speaker. The First Amendment does not contain this large a loophole.[133]

---

[128] *Id*. at 590.
[129] *Id*. (internal quotations omitted).
[130] *Id*. (emphasis added).
[131] *Id*.
[132] *Id*. at 592 (Clay, J., dissenting).
[133] *Id*. at 595 (Clay, J., dissenting).

A rehearing of this case *en banc* was granted in October 2014, though a date for oral arguments has not been scheduled as of this writing. Among other issues, the court will address whether police could have used alternative means to handle potential violence (such as setting up a buffer zone between the speakers and the audience), as well as whether the officers should be entitled to qualified immunity for failing to utilize these alternative means and proceeding straight to stopping the speaker.[134]

Police also must follow content-neutral laws whenever they have to restrict the ability of an audience to speak its counter-message. They especially must follow content-neutral laws if the audience is not, in fact, hostile, but rather is seeking to peacefully convey a counter-message. Applying time-place-and-manner restrictions may become trickier when two competing groups have a right to speak in the same public place. In *Startzell v. City of Philadelphia*,[135] police were confronted with two competing messages: that of Philly Pride, a gay-rights group that had a non-exclusive permit to hold their gay pride event OutFest on Philadelphia's public streets; and that of Repent America, a Christian organization committed to preaching against homosexuality. Police allowed Repent America to preach on the same streets on which OutFest was being held, but they arrested the group's leader, Michael Marcavage, when he refused a police order to move his group to the perimeter of the permitted area so as not to interfere with a musical performance at the festival or block access to vending booths.[136] Two judges on a panel of the U.S. Court of Appeals for the Third Circuit held that the police order to for the

[134] Brief of Amici Curiae Center for Religious Expression and Alliance Defending Freedom in Support of Appellants, 2014 WL 7212987 (C.A.6) (Appellate Brief) (6th Cir. Dec. 12, 2014).
[135] 533 F.3d 183 (3rd Cir. 2008).
[136] *Id*. at 191.

group to move was a content-neutral restriction on the time, place or manner of the group's speech, and therefore the police did not violate Marcavage's First Amendment rights when they arrested him.[137] The majority disagreed with Marcavage's argument that the police were effectuating a heckler's veto on him by, in effect, supporting the message of OutFest over that of his own group.[138]

In a concurring opinion, Judge Walter K. Stapleton agreed that Repent America's rights were not violated, but only because one of the group's members had forfeited those rights be using fighting words against one of the OutFest members.[139] Stapleton contended that the police only would have been able to restrict Repent America's speech if fighting words had been used; otherwise such a restriction would have amounted to "favoritism shown to the OutFest supporters."[140] Even following content-neutral laws to restrict speech in a hostile-audience situation could not save the police from violating one of the group's constitutional rights, according to Stapleton. Rather, police must let the two messages compete with one another until one group clearly forfeits its constitutional protections: "Police may not, consistent with the First Amendment, silence protected speech based solely on their judgment that it is interfering with competing protected speech."[141]

An exception to the heckler's veto doctrine does not exist for speech that upsets an audience predominantly made up of minors. In *Center for Bio-Ethics Reform v. Los*

---

[137] *Id*. at 200.
[138] *Id*.
[139] *Id*. at 205 (Stapleton, J., concurring). The Repent American individual had "singled out a transgendered individual for abuse, repeatedly calling him a 'she-man,' telling him, 'The mirror lied to you this morning. Your shadow is showing,' and by suggesting that his sexual identity would send him to hell."
[140] *Id*. at 206 (Stapleton, J., concurring).
[141] *Id*.

Angeles County,[142] pro-life organization Center for Bio-Ethics Reform drove a van displaying enlarged photos of aborted fetuses around the perimeter of a Los Angeles junior high school. Fearing for the safety of some students who walked slowly across a busy street while staring at the van, the school's assistant principal called the Los Angeles County Sheriff's Department. The officers stopped and detained the drivers of the van, and then asked them to leave the area, citing California Penal Code § 626.8(a).[143] That law prohibits "[a]ny person [from coming] upon any school ground, or street, sidewalk or public way adjacent thereto … whose presence or acts interfere with the peaceful conduct of the school or its pupils."[144] A unanimous panel of the U.S. Court of Appeals for the Ninth Circuit held that the officers used the students' reaction to the speech as the pretext for stopping the Center's drivers and asking them to leave; in other words, the police were acting on a heckler's veto.[145] The court held that even though "[c]hildren may well be particularly susceptible to distraction or emotion in the face of controversial speech, and may not always be expected to react responsibly," it would "be an unprecedented departure from bedrock First Amendment principles to allow the government to restrict speech based on the listener reaction simply because the listeners are children."[146] However, the court held that the officers were entitled to qualified immunity in this case because they did not reasonably know that the Center had a clearly

[142] 533 F.3d 780 (9th Cir. 2008).
[143] *Id*. at 786.
[144] *Id*. at 791.
[145] *Id*. at 794.
[146] *Id*. at 790.

established First Amendment right to display its message within the presence of minors.[147]

Between Fighting Words and Incitement

Because the heckler's veto principle deals with the potential of speech to lead 1) an audience that *opposes* the speech in question to 2) engage in violent acts, the principle must be analyzed in conjunction with two of the major First Amendment exceptions discussed in chapter 3: fighting words and incitement. The fighting words doctrine, from *Chaplinsky v. New Hampshire*,[148] involves words said in another person's face that "by their very utterance inflict injury or tend to incite an immediate breach of the peace."[149] The incitement standard, established in the 1969 case *Brandenburg v. Ohio*,[150] provides that the State cannot punish "advocacy of the use of force or of law violation except where such advocacy is directed to inciting or producing imminent lawless action and is likely to incite or produce such action."[151]

Despite the fact that the Supreme Court has seldom used the fighting words doctrine since *Chaplinsky*, the doctrine remains a potential exception to free speech within First Amendment jurisprudence.[152] In order for the doctrine to be resurrected and applied to a heckler's veto case today as the rationale for ruling *against* a speaker, the key characteristic distinguishing fighting words from speech that merely angers an audience must be met: the speaker's speech must be uttered "face-to-face" with his or her

[147] *Id*. at 794.
[148] 315 U.S. 568 (1942).
[149] *Id*. at 572-3.
[150] 395 U.S. 444 (1969).
[151] *Id*. at 447.
[152] *See* Clay Calvert, *Fighting Words in the Era of Texts, IMs and E-Mails: Can a Disparaged Doctrine Be Resuscitated to Punish Cyber-Bullies?* 21 DEPAUL J. ART TECH. & INTELL. PROP. L 1 (2010).

opponent. It seems very unlikely that a court would extend that rather clear definition of proximity to also encompass the distances between speakers and audience experienced in the cases discussed above.[153] Meanwhile, *Brandenburg*, narrowly interpreted, deals with the power of speech to incite restive *sympathizers* of a speaker's message to imminent lawless action. However, such an interpretation is not an argument against supplanting the clear and present danger standard in *Terminiello* with the updated incitement standard of *Brandenburg*. Indeed, the speech in the *hostile* audience cases is not worth less and the threat to public order is no more dire than in cases such as *Brandenburg*[154] that involve a *sympathetic* audience. Therefore, cases such as *Terminiello* and *Cantwell* today likely would be judged according *Brandenburg*'s higher bar. However, as the following discussion of state disorderly conduct statutes will show, law enforcement officials continue to act subjectively when interpreting both imminent lawless action and their state's definition of disorderly conduct when confronted with heated situations involving both words and actions of speakers and counter-speakers.

State Disorderly Conduct Statutes

All 50 states have a statute proscribing disorderly conduct or breach of the peace.[155] These statutes are important to the present analysis because they highlight the potential difficulties that law enforcement faces in balancing the protection of extreme yet lawful speech and the maintaining of social order when confronted with restive,

---

[153] *See* Samantha Barbas, *Creating the Public Forum*, 44 AKRON L. REV. 809, 880 (2011); BOLLINGER, *supra* note 5, at 180.
[154] Hess v. Indiana, 414 U.S. 105 (1973).
[155] Many municipalities have ordinances similar to such state statutes, but this section will focus only on the state laws for the sake of brevity.

heckling crowds. As proposed throughout this study, such difficulties will reappear in the analysis of how digital intermediaries deal with this balancing process when faced with their own instances of extreme speech and heckling crowds. State disorderly conduct statutes serve as another example of the potential difficulty of upholding lofty ideals of freedom of expression in the on-the-street context, which will aid the discussion of this difficulty in the context of content governance.

Many of the statutes share important similarities. First, many of the laws are written in nearly identical language,[156] which underscores a relatively high degree of uniformity among states in one of their chief weapons against public disorder. Some use verbs and nouns pertaining only to unlawful action or conduct, not speech.[157] This separation of conduct and speech reflects Professor Emerson's conception of the "fundamental distinction" between speech and conduct: the government is able to regulate the latter, but its power to suppress the former is "in most respects non-existent."[158] Some statutes[159] that do list certain types of speech as punishable disorderly conduct narrowly define that speech within categories for which the U.S. Supreme Court has already crafted exceptions: obscenity,[160] fighting words,[161] true threats,[162] and

[156] *Cf.* CODE OF AL. § 13A-11-7; AK. STAT. 11.61.110; ARK. CODE § 5-71-207; AZ. STAT. 13-2904; DEL. TITLE 11, CH. 5.VII.A § 1301; GA. CODE § 16-11-39; HAW. REV. STAT. § 711-1101; IN. CODE 35-45; KS. STAT. 21-6203; KY. STAT. 525.060; ME. CRIM. CODE TITLE 17-A.2.21 § 501-A; MD. CODE § 10-201; MO. REV. STAT. § 574.010; N.Y. PEN. CODE § 240.20; ORC § 2917.11 (OHIO).
[157] TENN. CODE 39-17-305; UTAH STAT. § 76-9-102; CODE OF VA. § 18.2-415.
[158] Thomas I. Emerson, *Toward a General Theory of the First Amendment*, 72 YALE L.J. 877, 880-1 (1963).
[159] *See generally supra* note 156.
[160] Miller v. California, 413 U.S. 15 (1973).
[161] Chaplinsky v. New Hampshire, 315 U.S. 568 (1942).
[162] Virginia v. Black, 538 U.S. 343 (2003).

incitement to imminent lawless action.[163] A handful of statutes explicitly state that the First Amendment protects speech from punishment under charges of disorderly conduct or breach of the peace provided that the speech does not cross the boundaries set by the U.S. Supreme Court.[164]

Some of the state statutes could be interpreted as vague in the way they fail to separate vitriolic yet lawful speech from unlawful conduct. Florida's breach of the peace statute prohibits "such acts as are of a nature to corrupt the public morals, or outrage the sense of public decency."[165] Several states prohibit using "profane" or "obscene" language within the presence or hearing of children.[166] Texas prohibits the use of "abusive, indecent, profane, or vulgar language in a public place" in a way that tends to cause an immediate breach of the peace, without the requirement of children being present.[167] Such language appears to show that state statutes tend to conflate the adjectives vulgar, profane, indecent and obscene in their quotidian use (as descriptors for curse words) with the proper legal uses of the terms indecent and obscene.

Illinois' disorderly conduct statute describes disorderly conduct as "any act [done] in such unreasonable manner as to alarm or disturb another and to provoke a breach of

---

[163] Brandenburg v. Ohio, 395 U.S. 444 (1969).
[164] *See* CONN. STAT. § 53A-182; MASS. G.L. C. 272 § 53; ORS § 166.025 (OREGON).
[165] FL. STAT. TITLE XLVI, CH. 877.03.
[166] *See, e.g.,* S.C. CODE § 16-17-530(b) (2012) ("us[ing[obscene or profane language on any highway or at any public place or gathering or in hearing distance of any schoolhouse or church" constitutes misdemeanor disorderly conduct); O.C.G.A. [Georgia] 16-11-39(a)(4) (2010) ("A person commits the offense of disorderly conduct when [he or she,] [w]ithout provocation, uses obscene and vulgar or profane language in the presence of or by telephone to a person under the age of 14 years which threatens an immediate breach of the peace"); IDAHO STAT. 18-6409(1) ("uses any vulgar, profane or indecent language within the presence or hearing of children" constitutes misdemeanor disorderly conduct).
[167] TEX. PEN. CODE § 42.01(a)(1).

the peace."[168] The Minnesota disorderly conduct statute forbids "offensive, obscene, or abusive language tending reasonably to arouse alarm, anger, or resentment in others."[169] According to the Crimes against the Public Peace law of Nevada, "[e]very person who shall by word, sign or gesture willfully provoke, or attempt to provoke, another person to commit a breach of the peace shall be guilty of a misdemeanor."[170] In New Jersey, "[a] person is guilty of a petty disorderly persons offense if, in a public place, and with purpose to offend the sensibilities of a hearer or in reckless disregard of the probability of so doing, he addresses unreasonably loud and offensively coarse or abusive language, given the circumstances of the person present and the setting of the utterance, to any person present."[171] The North Carolina disorderly conduct statute defines disorderly conduct as "any utterance, gesture, display or abusive language which is intended and plainly likely to provoke violent retaliation and thereby cause a breach of the peace."[172]

Briefly, this chapter must address the right of the hecklers to heckle, or to engage in acts of counter-speech against a speaker. As stated in the introduction, this right is one of the factors police must take into account when working to control a scene in which two groups are dueling with opposing messages. Just as police must assess when the initial speaker's speech crosses the *Brandenburg* line and incites a hostile crowd to imminent lawless action, they must also judge when hecklers go too far and infringe upon the right of the initial speaker to speak. The California Supreme Court held in the 1970 case *In re*

[168] 720 ILCS 5/26-1.
[169] MINN. STAT. 609.72(3).
[170] NRS 203.030.
[171] NJ STAT. 2C:33-2b.
[172] NC STAT. § 14-288.4(a)(2).

Kay that "[a]udience activities, such as heckling, interrupting, harsh questioning, and booing, even though they may be impolite and discourteous, can nonetheless advance the goals of the First Amendment."[173] In that case, the court overturned the convictions of protestors who peacefully chanted and made noise during the speech of an unpopular congressman. The protestors had been convicted under a California law that forbade "willfully disturb[ing] or break[ing] up any [lawful] assembly or meeting."[174] The court held, "An unfavorable reception, such as that given Congressman [John V.] Tunney in the instant case, represents one important method by which an officeholder's constituents can register disapproval of his conduct and seek redress of grievances."[175]

This case reflects the law of California, which is known to be an outlier in its willingness to proscribe what often appear to be lawful types of speech.[176] However, it is not out of the question that the court's reasoning could be applied in other states. Indeed, many of the statutes mentioned above have similar provisions prohibiting the willful disturbance of lawful assemblies. It appears that as long as speech is met with peaceful speech and nothing more—not fisticuffs or thrown rocks or bottles, not true threats or fighting words or incitement to imminent lawless action—then law enforcement should allow each side to compete against the other with words and other forms of expression. However, although he concedes that "heckling or other interruption of the speaker may

---

[173] 1 Cal.3d 930, 939 (Calif. Sup. Ct. 1970).
[174] CALIF. PEN. CODE § 403.
[175] In re Kay, at 939.
[176] *See California's 930 New State Laws*, S.F. CHRONICLE (Jan. 5, 2015). For examples of unique or first-in-the-nation California legislation, *see, e.g.,* CALIF. PEN. CODE 647(j)(4) (criminalizing the publication of so called "revenge porn," signed into law Oct. 8, 2013); CALIF. SB 568 (signed into law Sept. 30, 2013, requiring companies that direct online services to minors, or have knowledge that minors are using their services, to allow minors to access and delete information that they posted).

be part of the dialogue," Professor Emerson does not hold heckling in high regard, likening it to "pure noise."[177] "[C]onduct that obstructs or seriously impedes the utterance of another, even though verbal in form, cannot be classified as expression," he argues. "It has the same effect, in preventing or disrupting communication, as acts of physical force. Consequently it must be deemed action and is not covered by the First Amendment."[178]

The purpose of reviewing the statutes above is not to argue that some of them are unconstitutionally vague on their face. Indeed, when the U.S. Supreme Court strikes down one state law for not affording adequate protections for speech, it does not strike down every similar-looking law in the country. It is up to legislators to revise outdated and constitutionally questionable laws. Instead, a parallel can be drawn between, on the one hand, the conflicting opinions between *Feiner* and the other hostile audience cases, and, on the other hand, conflicting opinions on how to regulate disorderly speech based on police officers' interpretations of state disorderly conduct statutes. Although *Brandenburg* now serves as the limit beyond which incitement speech becomes unlawful, one can argue that the imminent lawless action standard, as high a bar as it may be, remains a subjective standard, potentially even for the police who rely on vague statutory language for guidance on how to handle incitement-type situation.

At least one federal district court has said that the vagueness of a disorderly conduct ordinance can exacerbate the likelihood that police will succumb to a heckler's veto and punish speakers for rabblerousing. In *Goldhamer v. Nagode*,[179] the U.S. District

---

[177] THOMAS I. EMERSON, THE SYSTEM OF FREEDOM OF EXPRESSION 338 (1970).
[178] *Id*.
[179] 611 F.Supp.2d 784 (N.D. Ill. 2009), *vac'd by* Schirmer v. Nagode, 621 F.3d 581 (7th Cir. 2010).

Court for the Northern District of Illinois held that part of the language of Chicago's disorderly conduct ordinance was unconstitutionally vague. The language at issue stipulated that a person is guilty of disorderly conduct if he or she fails to disperse following a police order when "three or more persons are committing acts of disorderly conduct in the immediate vicinity … [that] are likely to cause substantial harm or serious inconvenience, annoyance or alarm."[180] The court held that the vagueness of this part of the statute could lead to arbitrary enforcement of the ordinance by police[181]—certainly, such arbitrary enforcement could be particularly problematic for heckler's veto scenarios. The U.S. Court of Appeals for the Seventh Circuit vacated this decision due to the speakers' lack of standing to permanent injunction against the city ordinance in question.[182] However, the concern over the potential of police to take advantage of vague language persists from *Goldhamer*, even if its precedent does not.

A recent empirical study by a team of legal scholars and social scientists posited that "culturally motivated cognition … influence[d] individuals' perceptions of facts essential to distinguishing "speech" from "conduct" for purposes of the First Amendment."[183] The study focuses specifically on what factors may lead people to form attitudes toward whether police should intervene to stop a protest from turning violent.[184] The authors argue that placing speech and conduct in distinctly separate categories was

[180] Chicago, Ill. Municipal Code § 8–4–010(d).
[181] Goldhamer, at 794.
[182] Schirmer v. Nagode, 621 F.3d 581, 585 (7th Cir. 2010) (although the court "recognize[d] that the failure-to-disperse provision can be misused to impose a heckler's veto," it held that plaintiffs' claim for violation of their civil rights in the instance in question "does not necessarily carry over to their facial challenge requesting an injunction against any enforcement of the failure-to-disperse provision").
[183] Dan M. Kahan, David A. Hoffman, Donald Braman, Danieli Evans & Jeffrey J. Rachlinski, *They Saw a Protest: Cognitive Illiberalism and the Speech-Conduct Distinction*, 64 STAN. L. REV. 851, 883 (2012).
[184] *Id*. at 862.

an exercise in "sophism and ad hocery."[185] A 1982 study highlighted that even members of the ACLU showed remarkably high levels of intolerance toward certain types of controversial speech, such as speech advocating the overthrow of the government or racist speech.[186] By combining these apparent natural attitudinal uncertainties toward incendiary speech, law enforcement's interest in maintaining public order, and the relatively vague statutory language defining when speech is punishable, one can make the case that the veto power of today's hecklers is not dead.[187]

The Heckler's Veto and First Amendment Theory

The heckler's veto case law can be interpreted as guaranteeing that "local government must take action to protect [a] speaker against a hostile crowd."[188] This doctrine is consistent with so-called affirmative theories of the First Amendment, which generally put forth the position that the government has a duty to foster and facilitate certain conditions through which citizens can exercise their First Amendment rights. As stated in chapter 2, Meiklejohn argues that the "point of ultimate interest" regarding the First Amendment's function in facilitating self-government "is not the words of the speakers, but the minds of the hearers."[189] Affirmative theorists from the present-day "New Realist" school[190] have interpreted Meiklejohn's theory to allow for government

---

[185] *Id*. at 856.
[186] James L. Gibson and Richard D. Bingham, *On the Conceptualization and Measurement of Political Tolerance*, 76 AMER. POL. SCI. REV. 603 (1982).
[187] *See* Alex Vitale, *Op-Ed: Disorderly Conduct? How Protesting Became a Crime in NYC*, THIRTEEN (Oct. 26, 2011), *available at* http://www.thirteen.org/metrofocus/2011/10/op-ed-disorderly-conduct-how-protesting-became-a-crime-in-nyc/.
[188] Leanza, *supra* note 3, at 1307.
[189] ALEXANDER MEIKLEJOHN, FREE SPEECH AND ITS RELATION TO SELF-GOVERNMENT 25 (1948).
[190] This term comes from MATTHEW D. BUNKER, CRITIQUING FREE SPEECH: FIRST AMENDMENT THEORY AND THE CHALLENGE OF INTERDISCIPLINARITY 129 (2001). For examples of such scholarship, *See, e.g.,*

intervention whenever it can ensure that, as Meiklejohn puts it, "everything worth saying shall be said."[191] Affording speakers protection from hostile audiences while speaking in public is an example of such intervention, and the rationale behind such "government regulation of speech [is that it] actually might promote free speech, and should not be treated as an abridgement at all."[192]

Standing counter to affirmative First Amendment theory, the theory of the marketplace of ideas holds that the First Amendment is not a means to effective democratic governance, but to attaining truth.[193] Autonomy theory holds that the primary purpose of the strong protections of the First Amendment is to honor individual autonomy and foster individuals' personal growth.[194] Although the former is consequentialist while the latter is not, these theories share the label of "negative" First Amendment theories due to their mistrust of government intervention over matters of free speech jurisprudence.

Undaunted, affirmative First Amendment theorists point out that markets can and do fail without government regulation. They argue that markets for goods and services are imperfect, and the government must intervene to fix imperfections, such as the

---

CASS R. SUNSTEIN, DEMOCRACY AND THE PROBLEM OF FREE SPEECH (1993); OWEN FISS, THE IRONY OF FREE SPEECH (1996).

[191] MEIKLEJOHN, *supra* note 189, at 22.

[192] Cass R. Sunstein, *Free Speech Now*, 59 U. CHI. L. REV. 255, 267 (1992).

[193] As stated in chapter 3: "John Stuart Mill wrote, "if [an opinion] is not fully, frequently, and fearlessly discussed, it will be held as a dead dogma, not living truth." JOHN STUART MILL, ON LIBERTY 34 (1859/2001).

[194] As stated in chapter 3: "[C. Edwin] Baker argues that the 'fundamental purpose' of the First Amendment is to facilitate individual self-fulfillment, thereby allowing individuals to participate in social and political change." BAKER, *supra* note 18, at 51.

formation of monopolies, lest the efficiency of the marketplace be undermined.[195] The same imperfections, particularly the formation of monopolies on opinion or access to channels of communication, are found in the marketplace of ideas analogy, and therefore the government likewise should intervene to ameliorate those imperfections. Law professor Owen Fiss concedes that "[t]he state has no corner on virtue," yet he believes that the state nevertheless can fulfill the function that markets alone fail to fulfill, which is to ensure that the "democratic needs of the electorate" are met.[196]

One of the most common manifestations of affirmative First Amendment theory is public forum doctrine. Traditional public forums are places "in which the right to free speech receives its strongest protection."[197] Public forums exist in part because the vast majority of speakers does not have the resources to disseminate messages through mainstream print and broadcast media.[198] Under public forum doctrine, public spaces such as parks, plazas and street corners are presumed open to speakers, and "the generosity and empathy with which such facilities are made available is an index of freedom."[199] Therefore, if the government is to properly reserve these spaces for use by speakers, it must protect the spaces to the greatest extent possible. This includes requiring police to shield the speakers from hostile audiences, as the case law discussed above has

---

[195] Owen M. Fiss, *Why the State?* 100 HARV. L. REV. 781 (1987).
[196] *Id*. at 788-789.
[197] Dawn C Nunziato, *The Death of the Public Forum in Cyberspace*, 20 BERKELEY TECH. L.J. 1115, 1116 (2005). *See* Schneider v. State (Town of Irvington), 308 U.S. 147 (1939), at 163 (holding that "the streets are natural and proper places for the dissemination of information and opinion; and one is not to have the exercise of his liberty of expression in appropriate places abridged on the plea that it may be exercised in some other place.")
[198] Barbas, *supra* note 153, at 810; Nunziato, *supra* note 197, at 1117.
[199] Harry Kalven, Jr., *The Concept of the Public Forum:* Cox v. Louisiana, 1965 SUP. CT. REV. 1, 11-12 (1965).

shown.[200] Such protection must occur regardless of the offensiveness of the speakers' message, and also regardless of the potential costs stemming from the level of violence with which an angry audience may be perceived to react. A divided U.S. Supreme Court held in 1992 that requiring the speakers to absorb the costs of police protection would amount to an unconstitutional prior restraint on speech.[201] Justice Harry Blackmun wrote that despite the potentially heavy costs incurred by local law enforcement to protect speakers, "[s]peech cannot be financially burdened, any more than it can be punished or banned, simply because it might offend a hostile mob."[202]

As noted at the introduction to this section, the heckler's veto doctrine straddles affirmative and negative theories. The requirement that law enforcement must intervene to protect speakers from hostile audiences in public spaces is, by all intents and purposes, an example of the law following affirmative theory; the police are essentially preventing the speakers from being unlawfully forced out of the marketplace of ideas. Yet the dueling messages of speakers and counter-speakers represent a marketplace of ideas in microcosm. Law professor Ashutosh Bhagwat elaborates on the competing theories of the hostile audience cases, which he describes as cases of "dissident speech with a strong associational flavor."[203] "Dissident organizations invariably will face public hostility—that is what makes them dissident—but … they play a critical role in self-governance by challenging established understandings and the predominance of the state," Bhagwat

[200] Indeed, all of the speakers in the heckler's veto cases (except Terminiello, who spoke in an auditorium) spoke on public streets, which are considered traditional public forums.
[201] Forsyth County, Ga. v. Nationalist Movement, 505 U.S. 123 (1992).
[202] *Id*. at 135.
[203] Bhagwat, *supra* note 1, at 1011.

argues, concluding that "the hostile audience cases are best understood as preventing not a heckler's veto against lone, unpopular speakers, but societal vetoes of unpopular associations."[204] Herein lies the conflict of theories. Left to the marketplace of ideas, the ideas of dissident groups may end up eradicated from the market. Indeed, the goal of hecklers is to do exactly that. However, as Bhagwat (not to mention Mill[205]) states, the value of dissident speech is that it tests the mettle of truth, or at least the conventional ideas that society takes to be true.[206] The government therefore must protect such valuable speech by intervening on its behalf.

Law professor Robert Post argues that by protecting such speech, the government is upholding the First Amendment value of fostering "a process of critical interaction" among individuals.[207] This critical interaction is brought about not simply because the First Amendment protects speech, but rather because it "shield[s] speakers from the enforcement of community standards."[208] Post defines community here as "a social formation that inculcates norms into the very identities of its members,"[209] and thus the First Amendment requires that government not choose which set of norms will dominate in society. Government must stay neutral, even if the norms being communicated are uncivil or have a propensity to cause hostile reactions from an audience. Post notes that this principle leads to what he calls the "paradox of public discourse"[210] He writes, "To the extent that a constitutional commitment to critical interaction prevents the law from

[204] *Id*. at 1012.
[205] *Supra* note 193.
[206] *Id*.
[207] ROBERT POST, CONSTITUTIONAL DOMAINS 143 (1995).
[208] *Id*. at 144.
[209] *Id*. at 149.
[210] *Id*. at 147.

articulating and sustaining a common respect for the civility rules that make possible the ideal of rational deliberation, public discourse corrodes the basis of its own existence."[211] Nevertheless, Post argues that one of the fundamental purposes of the First Amendment is to ensure the "separation of public discourse from the domination of civility rules that define the identity of communities."[212] This principle—that government cannot enforce community norms in public discourse—is essential for ensuring a robust public discourse in which many diverse voices can debate and compete with one another. It is this principle that is under threat from content governance. The following section will address the restriction of speech by individuals pressuring digital intermediaries. The analysis will be focused at a broad level as defined by this concept: the potential threat to public discourse through the imposition of community values of civility on speakers of extreme messages.

Transposing the Heckler's Veto Doctrine

Obviously, Internet application providers such as Google and Facebook are not the government. No one can enforce any right against digital intermediaries to have them refrain from removing speech from their platforms. However, as discussed in chapter 2, scholars recently have broached the question: how closely can digital intermediaries be analogized to government actors within the context of Internet communications?[213] The analogy hinges on two related issues: the extent to which the Internet is (or is perceived

---

[211] *Id*.
[212] *Id*. at 177.
[213] One scholar put the question this way: "[D]oes an over-emphasis on non-interventionist techniques enable intermediaries to possess unintentionally significant power in violation of the communicative rights of individual users?" Daithi Mac Síthigh, *The Mass Age of Internet Law*, 17 INFO. & COMM. TECH. L. 79, 86 (2008).

to be) a public forum; and the extent to which intermediaries have the power to control that forum. Law professor Dawn Nunziato argues that the distinction between state and non-state actors is "formalistic," and she calls for a theory of freedom of speech that considers how any powerful societal entities (state or non-state) regulate the free flow of information and expression.[214] Nunziato places some of the blame on a piece of federal legislation that has been hailed[215] as a boon for free speech: Section 230 of the Communications Decency Act.[216] Perhaps most significantly, Section 230 declares, "No provider … of an interactive computer service shall be held liable on account of any action voluntarily taken in good faith to restrict access to or availability of material that the provider … considers to be obscene, lewd, lascivious, filthy, excessively violent, harassing, or otherwise objectionable, whether or not such material is constitutionally protected."[217] Nunziato argues that Section 230 gives companies that control Internet communications an incentive to chill speech, often at the behest of vociferous individuals, who protest the fact that these intermediaries would be hosting such speech in the first place.[218] According to Nunziato's logic, by affording greater First Amendment protection to interactive computer services and angry crowds, at the expense of individuals who use those services to create extreme or controversial UGC, Section 230 grants the former two groups a big heckler's veto over such speech.[219]

---

[214] DAWN C. NUNZIATO, VIRTUAL FREEDOM: NET NEUTRALITY AND FREE SPEECH IN THE INTERNET AGE 36 (2009).
[215] 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 (2010).
[216] 47 U.S.C. § 230.
[217] 47 U.S.C. § 230 (c)(2).
[218] Nunziato, *supra* note 197, at 1129-30.
[219] *Id*.

Of course, the hecklers on Facebook are not the same hecklers that Terminiello, Feiner and Cantwell faced. However, an argument can be made that Internet intermediaries offer the functional (even if not quite the legal) equivalent of what the government offers: public forums. Unlike a street corner or a park, the Internet is not under government control, and by definition it is not a public forum. However, the Internet has expanded the potential for individuals to disseminate messages, leading some scholars to argue that the Internet acts a *de facto* public forum—in other words, the Internet is a public forum based solely on its function.[220]

Internet law scholar Jack Balkin gives credence to this argument by contending that digital technologies are changing the social conditions in which people speak.[221] Balkin further argues that the purpose of freedom of speech is to "promote a democratic culture," which he defines, with a hint of Meiklejohnian theory, as "a culture in which individuals have a fair opportunity to participate in the forms of meaning-making that constitute them as individuals."[222] Balkin posits that the Internet has the potential to bring about an era in which all individuals can, in fact, have such a fair opportunity to participate in culture. However, the powerful private entities that manage the Internet—which Balkin describes as "hybrids of content providers and conduits for the speech of others,"[223] arguing that "the 'publicness' of digital communications networks is merely a

[220] *Id*.
[221] Jack M. Balkin, *Digital Speech and Democratic Culture: A Theory of Freedom of Speech for the Information Society*, 79 N.Y.U. L. REV. 1, 2 (2004).
[222] *Id*. at 3.
[223] *Id*. at 21.

side effect of the use of private property by private actors"[224]—pose a threat to that potential.[225] He therefore calls for "administrative and legislative regulation of technology" and "judicial creation and recognition of constitutional rights" to combat this threat.[226]

Nunziato and Balkin follow in the tradition of scholars who do not agree with the bright-line distinction of the state action doctrine. Law professor Frank Michelman argues that "a categorical distinction between the dangers of private action and the dangers of state action [on speech] cannot deliver reliable answers."[227] Michelman criticized this "private-power/public-power" distinction in the context of regulation of pornography.[228] He argued that a pornography ban that is the product of the democratic legislative process is no different, and may be even better, than a ban on such speech that comes from a boycott carried out via the "despotism of so-called private, social, or market power."[229] Barron was one of the first affirmative theorists to call for government to intervene against broadcasters by allowing individuals equal access to channels of broadcast communication. Barron argued that "nongoverning minorities in control of the means of communication should perhaps be inhibited from restraining free speech (by the denial of access to their media) even more than governing majorities are restrained by the

[224] *Id*. at 19.
[225] *Id*. at 6.
[226] *Id*.
[227] Frank I. Michelman, *Conceptions of Democracy in American Constitutional Argument: The Case of Pornography*, 56 TENN. L. REV. 291, 313 (1988).
[228] *Id*. at 309.
[229] *Id*. at 313.

First Amendment."[230] All of these scholars take the purported effects of corporate control of media that political economy theorists in the field of mass communication studies have been decrying for decades[231] and analyze them within the context of First Amendment theory and jurisprudence.

These scholars' arguments inevitably run up against irreconcilable differences over whose First Amendment rights should be given more weight: individuals or the private institutions that facilitate Internet communications. The U.S. Supreme Court has chosen the latter. In *Miami Herald v. Tornillo*,[232] the Court held that a Florida statute requiring newspapers to publish responses from individuals who believed they were attacked in the newspapers amounted to an unconstitutional prior restraint. Adopting a negative approach to First Amendment jurisprudence, the Court determined that such a statute constituted government control over the editorial process of a free press. Although the Court acknowledged that an ideal press was a responsible press that provided a forum for diverse viewpoints, it nonetheless held that "press responsibility is not mandated by the Constitution, and like many other virtues it cannot be legislated."[233] The Court's 1997 decision in *Reno v. ACLU*,[234] which afforded Internet speech the same level of First

---

[230] Jerome A. Barron, *Access to the Press: A New First Amendment Right*, 80 Harv. L. Rev. 1641, 1656 (1967).
[231] *See* ROBERT M. ENTMAN, DEMOCRACY WITHOUT CITIZENS: MEDIA AND THE DECAY OF AMERICAN POLITICS (1989); BEN BAGDIKIAN, THE NEW MEDIA MONOPOLY (2004); Daniel C. Hallin*, Hegemony: The American News Media from Vietnam to El Salvador, A Study of Ideological Change and its Limits*, *in* POLITICAL COMMUNICATION: APPROACHES, STUDIES, ASSESSMENTS*,* (David L. Paletz ed., 1986); ROBERT MCCHESNEY, DIGITAL DISCONNECT: HOW CAPITALISM IS TURNING THE INTERNET AGAINST DEMOCRACY (2013).
[232] 418 U.S. 241 (1974).
[233] *Id*. at 256. *See* Bruce W. Sanford and Jane E. Kirtley, *The First Amendment Tradition and Its Critics*, *in* THE PRESS (Geneva Overholser and Kathleen Hall Jamieson eds., 2005), 268.
[234] 521 U.S. 844 (1997).

Amendment protection as newspapers, cemented the Court's interpretation that First Amendment protection of media companies' rights trumps any sort of protection of individuals' rights against those companies to be able to speak. Nevertheless, the principle of the heckler's veto may help today's affirmative theorists make a clearer case for their cause, as the examples below illustrate.

Hecklers and Suppression Online

The following are examples of incidents of online suppression of speech that resemble heckler's veto scenarios from traditional First Amendment jurisprudence. The examples can be separated into four categories: 1) individuals pressuring intermediaries to remove speech from their platforms; 2) a "rioters' veto;" 3) trolling and abuse; and 4) online shaming.

Individuals Pressuring Intermediaries

Facebook and Beheading Videos

In late April 2013, two videos depicting the beheading of three individuals, purportedly in Mexico, appeared on Facebook.[235] The social networking site initially refused to remove the videos in spite of formal requests made by individual members and humanitarian organizations. Facebook said of one of the videos:

> People are sharing this video on Facebook to condemn it. Just as TV news programs often show upsetting images of atrocities, people can share upsetting videos on Facebook to raise awareness of actions or causes. While this video is shocking, our approach is designed to preserve people's rights to describe, depict and comment on the world in which we live.[236]

[235] Leo Kelion, *Facebook U-turn after Charities Criticizes Decapitation Videos*, BBC NEWS: TECHNOLOGY (May 1, 2013), *available at* http://www.bbc.co.uk/news/technology-22368287.
[236] *Id*.

However, after pressure from members and interest groups increased, Facebook decided to remove the videos, saying it would "evaluate [its] policy and approach to this type of content."[237] At the time, Facebook's "Community Standards" page stated, "We understand that graphic imagery is a regular component of current events, but must balance the needs of a diverse community. Sharing any graphic content for sadistic pleasure is prohibited."[238] Facebook said in May 2013 that the videos did not meet its standards for graphic or gratuitous violence.[239] In mid-October 2013, it allowed the videos to be viewed on its site, again saying that people should be able to watch the videos to condemn them, and adding that it was considering a policy of including a warning alongside the link to the video.[240]

Facebook and Misogynist Pages

In another incident, also involving Facebook and taking place in early 2013, feminist organizations decried the existence of pages created on the social networking giant that glorified or made light of rape and domestic violence. Activist Soraya Chemaly, Jaclyn Friedman of the group Women, Action and the Media (WAM), and Laura Bates of the Everyday Sexism Project, published an open letter online on May 21, 2013, demanding that Facebook not tolerate "speech that trivializes or glorifies violence

---

[237] *Id*.
[238] *Community Standards*, INTERNET ARCHIVE (Dec. 15, 2012), *available at* http://web.archive.org/web/20121215024155/http://www.facebook.com/communitystandards.
[239] Kelion, *supra* note 235.
[240] Leo Kelion, *Facebook lets beheading clips return to social network*, BBC NEWS (Oct. 21, 2013), *available at* http://www.bbc.co.uk/news/technology-24608499.

against girls and women."[241] The open letter stated that pages had titles such as "Fly Kicking Sluts in the Uterus" and "Violently Raping Your Friend Just for Laughs," and images appeared on the network "of women beaten, bruised, tied up, drugged, and bleeding, with captions such as 'This bitch didn't know when to shut up' and 'Next time don't get pregnant.'" The women asked Facebook users to contact companies whose ads appeared on pages with such speech. In April 2013, Bates took a screenshot of a page titled "Drop kicking sluts in the teeth" and tweeted it to the beauty company Dove, whose ad appeared next to the page. In late May, the activists persuaded advertisers such as Nissan UK, *Jump* magazine, Desire Books and 15 other companies to pull ads from the social network.

Marne Levine, Facebook's vice-president for global public policy at that time, responded to the activists' demands in a May 28, 2013, blog post on the social network, promising that Facebook officials would "update the training for the teams that review and evaluate reports of hateful speech or harmful content on Facebook."[242] Levine wrote that Facebook would push for more accountability from "the creators of content that does not qualify as actionable hate speech but is cruel or insensitive by insisting that the authors stand behind the content they create." For example, this requirement would mean that "the creator of any content containing cruel and insensitive humor include his or her

[241] *Open Letter to Facebook*, WOMEN, ACTION, & THE MEDIA (May 21, 2013), *available at* http://www.womenactionmedia.org/facebookaction/open-letter-to-facebook/.
[242] *Controversial, Harmful and Hateful Speech on Facebook*, FACEBOOK SAFETY (May 28, 2013), *available at* https://www.facebook.com/notes/facebook-safety/controversial-harmful-and-hateful-speech-on-facebook/574430655911054.

authentic identity for the content to remain on Facebook." Levine did not promise that Facebook would remove any of the pages.

Brief Synthesis

These incidents will be addressed again in chapter 5, which focuses specifically on Facebook's balancing act between protecting freedom of speech and preventing harm. However, it is important to address some key points about these incidents and how they reflect heckler's veto situations. First, these incidents involve some individuals imposing a certain set of community norms on the speech of other individuals.[243] Of course, Facebook is not a state actor, and thus it can choose to require its users to follow whatever community norms the company desires. What is interesting here is that there are actually three sets of community norms involved in these incidents: the speakers,' the complainers,' and those of Facebook itself. The conflict in each incident was over which side could get Facebook to interpret that side's norms as more compatible with the company's own norms. The groups that posted the offensive speech in question did not speak up on their own behalf, despite the fact that their speech could be considered to have a political or social message (twisted though it may be). In other words, individuals who advocate for community norms that are less tolerant of offensive speech seem to have greater mobilization and greater leverage over Facebook than those who would publish more offensive speech. Again, this idea will be discussed in greater detail in chapter 5.

[243] POST, *supra* note 207.

Rioters' Veto

In early September 2012, a YouTube video with enigmatic origins titled "Innocence of Muslims" came to the attention of news media in Egypt. The video depicted the Islamic Prophet Muhammad as a war-hungry womanizer and pedophile. Although the creator was first reported to be a "Sam Bacile," that name was revealed to be one of the many aliases for the real author: Nakoula Basseley Nakoula, a Coptic Christian living in California who harbored strong anti-Islamic beliefs.[244] News media in Egypt broadcast the video with Arabic translations, and soon other broadcasters in the region were showing clips of the video.[245] Protests flared up throughout the Muslim world. On September 11, 2012, gunmen in Benghazi, Libya, stormed the American consulate and killed four people, including the American ambassador to Libya, J. Christopher Stevens. At the time, the close proximity between the broadcasting of the video and the Benghazi attack led many (including the Obama administration) to believe that the video had been the impetus for the attack.[246] Only later did the Obama administration admit that the attack on the consulate had been planned in advance, and any rage sparked by the inflammatory video was likely only a coincidence.[247]

---

[244] Adi Robertson, *From 'Desert Warrior' to 'Innocence of Muslims,' a Controversial YouTube Video Is Both Catalyst and Scapegoat*, THE VERGE (Sept. 17, 2012), *available at* http://www.theverge.com/2012/9/17/3346428/innocence-of-muslims-protests.
[245] *Id*.
[246] *Id*.
[247] *Pentagon to Review Video of Libya Attack*, CNN.COM (Sept. 12, 2012), *available at* http://news.blogs.cnn.com/2012/09/12/u-s-ambassador-to-libya-3-others-killed-in-rocket-attack-witness-says/comment-page-3/.

On September 12, 2012, Google, the owner of YouTube, announced it was blocking the video from being viewed in several countries.[248] Although the vast majority of these countries' governments formally requested that Google take down the video due to its violation of those countries' laws, Google notably removed the video from view in both Egypt and Libya without receiving a request to do so.[249] In a widely published press release following the removals of the video, Google stated the following:

> We work hard to create a community everyone can enjoy and which also enables people to express different opinions. This can be a challenge because what's OK in one country can be offensive elsewhere. This video—which is widely available on the Web—is clearly within our guidelines and so will stay on YouTube. However, we've restricted access to it in countries where it is illegal such as India and Indonesia as well as in Libya and Egypt, given the very sensitive situations in these two countries.[250]

The company contended that its decision to remove the video was consistent with YouTube's code of conduct on not permitting hate speech.[251] YouTube defines hate speech as "speech which attacks or demeans a group based on race or ethnic origin, religion, disability, gender, age, veteran status, and sexual orientation/gender identity."[252] The company further defines prohibited hate speech by stating the following: "Sometimes there is a fine line between what is and what is not considered hate speech.

---

[248] *Google Blocks Video Clips in Egypt, Libya Amid Concerns over Anti-Islam Film*, Al-Arabiya News (Sept. 13, 2012), *available at* http://english.alarabiya.net/articles/2012/09/13/237659.html.
[249] *Id*.
[250] Eva Galperin, *YouTube Blocks Access to Controversial Video in Egypt and Libya*, EFF (Sept. 12, 2012), *available at* https://www.eff.org/deeplinks/2012/09/youtube-blocks-access-controversial-video-egypt-and-libya.
[251] *Id*.
[252] *Community Guidelines*, YOUTUBE, *available at* http://www.youtube.com/yt/policyandsafety/communityguidelines.html.

For instance, it is generally okay to criticize a nation, but not okay to make insulting generalizations about people of a particular nationality."[253]

The "Innocence of Muslims" incident shares many characteristics with the heckler's veto scenarios discussed at the beginning of this chapter. Members of a hostile audience seeks to silence speech that offends them by resorting to violent actions, and the arbiter of the speech (here, Google) seeks to mollify the situation be removing the speech in certain locations. The offensive speech in question carried with it a distinctly social (and certainly political) message, thereby pitting the cost of the speech's offensiveness with its potential social benefit of contributing a social and political idea to the public discourse. Notwithstanding the fact that the video may have violated laws in some countries,[254] it undoubtedly threw the norms of freedom of expression of the United States and much of the Muslim world into stark contrast.

The "Innocence of Muslims" incident was not the first time speech created in the United States and published on the Internet caused outcry abroad. In 2000, a Frenchman sued Yahoo for facilitating an auction of Nazi memorabilia in France, where such goods are considered contraband.[255] Yahoo removed the auction of these items in January 2001.[256] Yet the incident led Internet engineers to realize that the World Wide Web did not necessarily have to be the same across the entire planet; in fact, French Internet entrepreneur Cyril Houri proved that the Yahoo could have engineered a way to block the

---

[253] Andrew Miga, *Google Says It Won't Take Down Anti-Muslim Clip*, AP (Sept. 15, 2012).
[254] *See, e.g.,* Pakistan Penal Code (Act XLV of 1860), Ch. XV, 295-C ("Use of derogatory remarks, etc., in respect of the Holy Prophet") *available at* http://www.pakistani.org/pakistan/legislation/1860/actXLVof1860.html.
[255] Yahoo!, Inc. v. La Ligue Contre Le Racisme, 169 F. Supp. 2d 1181 (N.D. Cal. 2001).
[256] JACK GOLDSMITH AND TIMOTHY WU, WHO CONTROLS THE INTERNET? ILLUSIONS OF A BORDERLESS WORLD 8 (2006).

Nazi memorabilia auction from being seen in virtually all of France.[257] Professors Jack Goldsmith and Tim Wu extolled this technological solution to a speech problem as a way to avoid the messiness of trying to come up with a legal solution to the problem.[258] This technological solution allows issues involving highly offensive speech, such as the "Innocence of Muslims" incident, to end in a more nuanced fashion: the offensive speech can be accessible in some places and inaccessible in others, rather than heckler's veto scenarios such as in *Cantwell* or *Terminiello* where the only options are either allowing offensive speech or suppressing it. Plus, inaccessible in this situation does not necessarily mean completely inaccessible; individuals in countries such as Egypt or Indonesia probably may be able to use proxy servers or virtual private networks to access the video if they desired.[259] This nuanced approach to dealing with this speech problem also allows Google to occupy a favorable middle ground, in which the company can claim to both be protecting free speech and practicing social responsibility by preventing harm in certain parts of the world (regardless of whether it is legally required to do so).

Despite this nuanced approach to dealing with speech that cuts across global norms of freedom of expression, the conflict between these norms does not end. Rather, the debate is heating up over how zealously Americans should defend their First Amendment right to publish offensive speech in an era when that speech can deeply affect individuals who live across oceans and abide by very different cultural norms. Law

---

[257] *Id*. at 7.
[258] *Id*. at 10.
[259] *See Circumvention for Internet Censorship and Filtering*, FRONTLINE: INTERNATIONAL FOUNDATION FOR THE PROTECTION OF HUMAN RIGHTS DEFENDERS (2007), *available at* https://equalit.ie/esecman/chapter2_6.html.

professor Noah Feldman argued in a guest column in *Newsday* that the global reach of Internet communication requires a reexamination of First Amendment jurisprudence.[260] Comparing the "Innocence of Muslims" incident to traditional hostile audience cases, Feldman contended:

> In principle, it should not be more difficult to predict the likelihood that an audience abroad will respond violently to a given statement than it should be to predict that a crowd gathered in front of the speaker will respond violently. At present, we rely on the informed judgment of law-enforcement officers who are on the scene, watching both the speaker and the audience that is on the verge of exploding. It seems possible that law-enforcement officers who were sufficiently well-informed about conditions elsewhere could make a similar judgment about a YouTube video.[261]

Elsewhere, Anthea Butler, associate professor of religious studies at the University of Pennsylvania and a guest columnist for *USA Today*, wrote, "While the First Amendment right to free expression is important, it is also important to remember that other countries and cultures do not have to understand or respect our right."[262] Sarah Chayes, former general counsel to the Joint Chiefs of Staff, argued in a widely published Op-Ed that Nakoula's video met *Brandenburg*'s "imminent lawless action" standard,[263] and therefore the state has the ability to force it offline.[264] Each of these authors denounced the deadly protests that followed the airing of the video, and none went so far as to say the video justified violence. However, although Chayes was the only one to argue that the video amounted to unprotected speech, all three authors agree that speech

---

[260] Noah Feldman, *Should the Internet Age Change Free Speech?* NEWSDAY (Sept. 28, 2012), *available at* http://www.newsday.com/opinion/oped/feldman-should-the-internet-age-change-free-speech-1.4054058.
[261] *Id*.
[262] Anthea Butler, *Why 'Sam Bacile' Deserves Arrest*, USA TODAY (Sept. 13, 2012).
[263] Brandenburg v. Ohio, 395 U.S. 444, 449 (1969).
[264] Sarah Chayes, *Does "Innocence of Muslims" Meet the Free-Speech Test?* L.A. TIMES (Sept. 18, 2012).

can no longer only be judged by one set of norms. Refraining from publishing speech deemed offensive by a foreign set of norms may be in Americans' best interest. Taking Post's perspective,[265] this chilling of speech by threat of riot amounts to an imposition of community norms by one group over another, and therefore it most certainly resembles a heckler's veto.

Abuse, Trolling and Gamergate

A new kind of heckler's veto has emerged online: trolling.[266] Unlike the examples of heckler's vetoes discussed so far—both in the physical and online worlds—this type of heckler's veto is not one that seeks to force the tone of public discourse to a tamer, less offensive level. Rather, this type of heckler's veto seeks to silence one voice or a particular set of voices in public discourse through upping the level of offensiveness to the point of abuse.[267] Instead of silencing a speaker by pressuring a digital intermediary to remove his or her speech due to its alleged violation of community norms, trolls create their own set of highly amplified fringe norms and force them into the online public discourse. The abusive practice of trolling disproportionately affects voices that are not completely members of mainstream discourse, namely female and minority voices.[268] Although the mechanism of the trolling variety of the heckler's veto is different from the more traditional variety, the end result is the same: the public discourse becomes less diverse and less robust.

---

[265] POST, *supra* note 207.
[266] *See* CITRON, *supra* note 69, at 52 (defining trolling as "pulling pranks targeting people and organizations, desecrating reputations, and revealing humiliating or personal information").
[267] *Id*. at 137, 171.
[268] *Id*. at 16. *See also* Mary Anne Franks, *Unwilling Avatars: Idealism and Discrimination in Cyberspace*, 20 COLUM. J. GENDER & L. 224 (2011); Anna North, *What Do We Know about Online Harassment?* N.Y. TIMES (Oct. 23, 2014).

One particularly illustrative example of this variety of heckler's veto is the case of female video game developers and reporters being abused on social media (predominantly on Twitter) in a saga that came to be known as "Gamergate." The Twitter hashtag #gamergate originated with several video game enthusiasts who used it to express outrage over allegations that video game developer Zoe Quinn was sleeping with a reporter who covered the industry.[269] The hashtag then started being used alongside death and rape threats against Quinn.[270] Then, Anita Sarkeesian, a feminist cultural critic who focuses on the video game industry, started receiving threats of death and rape on Twitter and had to cancel public speaking events due to bomb threats.[271]

The argument that trolling and online abuse constitute a heckler's veto borrows reasoning from critical legal theorists (including critical race theorists[272] and feminist legal scholars[273]), as well as certain new realist theorists,[274] who argue that hostile speech leads the targets of that speech to silence themselves for fear of more verbal or even physical reprisals. This silencing effect leads these scholars to argue that by allowing hostile speech, traditional First Amendment jurisprudence actually goes against the ultimate First Amendment value of promoting a diverse and robust public discourse. Of

---

[269] Nick Wingfield, *Feminist Critics of Video Games Facing Threats in 'GamerGate' Campaign*, N.Y. TIMES (Oct. 15, 2014).
[270] *Id*.
[271] *Id*.
[272] Charles R. Lawrence, *If He Hollers Let Him Go: Regulating Racist Speech on Campus*, 1990 DUKE L. J. 431 (1990); Richard Delgado, *Campus Antiracism Rules: Constitutional Narratives in Collision*, 85 NW. U. L. REV. 343 (1991). *See also* Calvert, *supra* note 69.
[273] Catherine MacKinnon, *Pornography, Civil Rights and Speech*, 20 HARV. CIV. RTS.-CIV. LIB. L. REV. 1 (1985); Cass Sunstein, *Pornography and the First Amendment*, 1986 DUKE L. J. 589 (1986).
[274] Owen M. Fiss, *The Supreme Court and the Problem of Hate Speech*, 24 CAPITAL U. L. REV. 281 (1995); Owen Fiss, *El Efecto Silenciador de la Libertad de Expresión*, 4 ISONOMÍA 17 (1996); FISS, THE IRONY OF FREE SPEECH, *supra* note 190, at 16.

course, this argument runs into trouble with traditional notions of First Amendment jurisprudence, due to its ultimate call for the law to prohibit hostile speech in the name of protecting the betterment of the overall public discourse. However, for the purposes of identifying trolling as a variety of heckler's veto, the argument is spot-on accurate. Professor Mary Anne Franks acknowledges this connection. She argues that the online world in which abuse has become prevalent is "a world in which only certain individuals enjoy the mythic degree of liberty … touted by cyberspace idealists, while others experience a loss of liberty and a re-entrenchment of physical restraints already unequally imposed upon them in the offline world."[275] Similarly, Professor Danielle Citron argues that stopping online abuse "would secure the necessary preconditions for free expression for targeted individuals."[276] Just as state actors must protect speakers from hostile audiences in brick-and-mortar heckler's veto scenarios out of an affirmative duty to preserve the robustness of the public discourse, intermediaries must protect speakers from abuse to preserve the diversity and robustness of the online public discourse.

This argument arrives at a fundamental point: preventing abuse and protecting freedom of speech are not necessarily mutually exclusive goals. Indeed, working to prevent abuse may help further the goal of protecting freedom of expression, for at least two reasons. The first reason is the argument discussed in the preceding paragraph: victims of abuse may be more likely to enter the public discourse as the abuse subsides. Second, preventing abuse will help decouple abusive speech from other forms of extreme speech (namely those that do not target specific individuals), rather than these two types

[275] Franks, *supra* note 268, at 246.
[276] CITRON, *supra* note 69, at 29.

of speech being conflated into one category of socially undesirable speech that individuals and intermediaries should work to eradicate. Chapter 3 showed that making this distinction is not always easy.[277] No matter, it is incumbent upon individuals and digital intermediaries to understand why the distinction exists, recognize the distinction in online public discourse, and work to preserve the distinction.

Shaming

A third variety of the heckler's veto in the online context is the concept of online shaming. Online shaming occurs when an online record of an individual's offensive or socially undesirable speech or actions is publicly pilloried and spread virally through a digital intermediary.[278] Privacy scholar Daniel Solove argues that online shaming is a variety of a longstanding human practice of "norm policing," whereby individuals publicly chastise others for their socially undesirable speech or actions in an attempt to get them to recognize social norms and ultimately alter their behavior.[279] However, Solove argues that online shaming is much more worrisome because "[h]aving a permanent record of norm violations is upping the sanction to a whole new level."[280] Online shaming is akin to a heckler's veto because it stigmatizes socially undesirable speech to an extreme extent, often resulting in the speaker suffering other forms of social punishment. For example, Justine Sacco, then senior director for corporate

---

[277] Cass Sunstein arguably came closest in making the distinction, arguing that a racists speech delivered to a crowd is distinct from a racial epithet delivered face-to-face (contending that the latter should be considered unprotected while the former should be protected). Cass Sunstein, *Free Speech Now*, 59 U. CHI. L. REV. 255, 309 (1992).
[278] DANIEL J. SOLOVE, THE FUTURE OF REPUTATION: GOSSIP, RUMOR AND PRIVACY ON THE INTERNET, 6 (2007).
[279] *Id*. *See also* LAWRENCE LESSIG, CODE, VER. 2.0, 287 (2006).
[280] SOLOVE, *supra* note 278, at 6.

communications for the media firm IAC, sent the following tweet on December 20, 2013, before boarding a flight to South Africa: "Going to Africa. Hope I don't get AIDS. Just kidding. I'm white!"[281] The public outcry was so strong that not only was Sacco fired from her job, but she has had difficulty building her reputation back up to its pre-tweet level.[282] In another example, a blog has been set up on the blogging service Tumblr called "Racists Getting Fired (and Getting Racists Fired)."[283] The site is home to posts of screenshots taken by vigilantes (sometimes anonymous, sometimes not) of racist comments (or at least what they consider to be racist comments) posted by individuals online, along with personal information about the individuals.[284] Sometimes the vigilantes contact the individuals' places of work to alert their employers in an attempt to get the individuals fired.[285] When the strategy works, they post about that as well.[286] Sometimes the individuals admit to their racist comments, while other times individuals claim that the posts about them are false and being carried out against them by others who have a grudge against them.[287]

At first glance, one may see online shaming as the potential price one must be prepared to pay for having a public presence online. Just as a person can reap the benefits of having a more positive or thought-provoking tweet or YouTube video go viral, one should also be subject to the consequences of a tweet like Sacco's. Indeed, Professor

---

[281] Jon Ronson, *How One Stupid Tweet Blew Up Justine Sacco's Life*, N.Y. TIMES MAGAZINE (Feb. 12, 2015).
[282] *Id*.
[283] *Available at* http://racistsgettingfired.tumblr.com/.
[284] India Rakusen, *Getting Racists Sacked: #FreeSpeechStories*, BBC NEWS TRENDING BLOG (Jan. 18, 2015).
[285] *Id*.
[286] *Id*.
[287] *Id*.

Post's theory of "reputation as property" would lend support to this idea.[288] Also, one could make the argument that online shaming is simply the marketplace of ideas at work in an online environment: desirable speech is simply pushing undesirable speech out of the market.[289] And who could blame any employer from firing someone after he or she found out about the employee's latent potential for racist outbursts? However, online shaming should be recognized as a heckler's veto for three reasons.

First, the practice of online shaming risks ensnaring (and thereby potentially chilling) speech of social and political significance.[290] For example, the second half of 2014 and early months of 2015 saw racial tensions hit a boiling point after white police officers killed unarmed black civilians Michael Brown in Ferguson, Missouri, and Eric Garner in Staten Island, New York. Analysts documented distinctly polarized pro-police and pro-Brown/Garner camps on social media, and noted that many of the pro-police messages had racist overtones.[291] Although one may disagree with the racist overtones, the message of the tweets had clear social and political significance,[292] which should be encouraged in times of great social turmoil such as what the country is dealing with today with racial issues. The threat of being fired due to vigilante justice for expressing such an

---

[288] Robert C. Post, *The Social Foundations of Defamation Law: Reputation and the Constitution*, 74 CAL. L. REV. 691, 694 (1986). Post theorizes that reputation, like one's fortune, is built of the "fruit of one's own endeavors." It can be spent or invested to build up one's fortune, which, in turn, can be invested back into one's good reputation. But it can also be subject to ruin through poor investment.

[289] *See, e.g.,* MILL, *supra* note 193.

[290] *See* Robert C. Post, *Racist Speech, Democracy, and the First Amendment*, 32 WM. & MARY L. REV. 267, 270 (1991) ("Any communication can potentially express the racist self, and thus no communication can ever be safe from … sanction").

[291] Mike Wendling, *#BBCtrending: Ferguson Exposes Twitter's Racial and Social Divisions Analysed*, BBC NEWS TRENDING BLOG (Dec. 4, 2014).

[292] One example of a tweet with both a social message and racist overtones was: "I would feel safer, any day, to encounter #DarrenWilson on the street, than to meet #MichaelBrown or half of those now protesting!" (Darren Wilson was the name of the police officer who shot Michael Brown.) *Id*.

opinion could lead to that opinion being chilled, and therefore the robustness of the online public discourse as a whole suffers. Second, shaming individuals may not address the root problems that spurred the undesirable speech (such as racism). Shaming may even exacerbate the problem by deepening the beliefs of the speakers that led them to make the comments in the first place.[293]

The third reason shaming acts like a heckler's veto comes from Bollinger's tolerance theory.[294] Chilling undesirable speech through shaming may harm society by not exposing individuals to the extreme ideas that would make them more aware of the true scope of public discourse. Law professor Jerry Kang argues that witnessing racist speech online may not only lead individuals to become more aware of the overt, hostile extent to which racism still exists in the United States, but also lead more moderate

---

[293] *See, e.g.,* Karen P. Leith & Roy F. Baumeister, *Empathy, Shame, Guilt, and Narratives of Interpersonal Conflicts: Guilt-prone People Are Better at Perspective Talking*, 66 J. OF PERSONALITY 1, 2 (1998) (arguing that "[w]ith shame, … the affective response of focusing on one's own distress may predominate, and this is less likely (as compared with taking the other's perspective) to produce beneficial consequences"); Leith & Baumeister, at 3 (arguing that shame may lead individuals "to ignore the problem, to deny one's responsibility, to avoid other people, or perhaps to lash out at one's accusers"); Richard H. Smith, J. Matthew Webster, W. Gerrod Parrott & Heidi L. Eyre, *The Role of Public Exposure in Moral and Nonmoral Shame and Guilt*, 83 J. OF PERSONALITY & SOC. PSYCH. 138, 157 (2002) (arguing: "The public exposure or the severity of the admonishment creates the experience of humiliation rather than shame. It may be that one of the reasons that shame is often linked with hostility is that shamed individuals, rather than feeling shamed, perceive the shaming as unjustified humiliation. They may recognize that they have committed a transgression, but they may also feel that the public exposure itself is unjustified. As a result, the focus of attention can shift to the perception of having been mistreated rather than on the person's own transgression. This shift may produce hostility rather than the negative self-appraisals associated with shame. There may be a tendency for any public exposure to seem unjustified and unnecessary from the biased point of view of the transgressor. If this is true, shaming will more often backfire on the shamer and create especially maladaptive, hostile feelings."). These arguments are similar to the theory that freedom of expression acts as a "safety valve" to allow those with extreme viewpoints a safe outlet for those viewpoints, while suppression "engender[s] hostility, resentment, fear and other divisive forces," Thomas I. Emerson, *Toward a General Theory of the First Amendment*, 72 YALE L. J. 877, 929 (1963).
[294] BOLLINGER, THE TOLERANT SOCIETY, *supra* note 5.

voices to talk openly about racial issues in American society.[295] Such a potential social benefit of tolerating extreme speech online leads Kang to argue that "[i]n the abstract, one cannot decide what is the greater threat: the private power of individuals making racist comments that flaunt social norms of equality, or the private power of virtual community hosts trying to enforce such norms."[296]

Assessment

Each of the examples above points to the notion that a fine line separates the natural workings of the marketplace of ideas from the silencing of unpopular speech through an amplified enforcement of a particular set of social norms. Each example involves its own distinct set of social norms, as well as distinct definitions of what constitutes acceptable and unacceptable speech according to those norms. Yet each variety of the heckler's veto threatens speech in its own way. In brick-and-mortar hostile audience scenarios, police must control (or at least attempt to control) a violent crowd rather than arrest the speaker. Intermediaries do not have the ability to send police to protect the targets of angry rioters because the targets have almost nothing to do with the speaker. Police have constitutional guidelines and a civic duty to direct them on when to protect speech and when to stop it. However, in spite of these standards, sometimes unclear state laws, the heat of the moment, or even the police's personal whims lead some officers to choose the latter course of action over the former. Intermediaries set their own, vague standards, and deal with cases vis-à-vis those standards as they arise.

---

[295] Jerry Kang, *Cyber Race*, 113 HARV. L. REV 1131, 1173 (2000). *See also* Post, *Racist Speech*, *supra* note 290, at 304-5.
[296] *Id*. at 1178.

Here, Post offers some guidance. As stated earlier, Post argues that one of the fundamental purposes of the First Amendment is to prevent government from imposing community norms of civility on public discourse.[297] Essentially, Post's position is that it is better to have a robust yet uncivil public discourse that leads to critical engagement among individuals with disparate beliefs than to have a public discourse that is cleansed of extreme positions out of a concern for civility.[298] The only difference between Post's theory and content governance is the entity that is enforcing the community norms: state actors versus private actors.

Bollinger's tolerance theory can help bridge the conceptual gap between hostile audience scenarios involving state-actor referees and those involving digital intermediaries. As discussed in chapter 3, Bollinger's theory deals specifically with instances of "extremist speech."[299] Bollinger criticizes both marketplace of ideas theory and affirmative First Amendment theory for their inability to adequately justify extremist speech, and he argues that if the value of such speech cannot be assessed in terms of how it contributes to truth or effective self-governance, then it should be assessed in terms of how it affects our character.[300] Tolerating extreme speech, Bollinger argues, "is intended and designed to perform a self-reformation function for the general community,"[301] a function that takes the form of citizens tempering their natural tendency to be intolerant

[297] *Supra* notes 207-212.
[298] *Id*.
[299] BOLLINGER, THE TOLERANT SOCIETY, *supra* note 5, at 9.
[300] *Id*.
[301] *Id*. at 134.

toward extremist messages. This function follows Mill's[302] assertion that "confrontation with falsehood" gives people "a 'livelier' sense of the truths they themselves already hold."[303] Bollinger does not focus so much on what is true as he does on the process citizens go through to examine the many possible "truths" around them. Within this process, government acts as a role model for citizens, facilitating the growth of their tolerance by not banning extremist speech.[304]

Digital intermediaries give extremists a wide reach, thereby affording them the potential to cause harm with their words on a larger scale than with a physically localized event.[305] This potential makes Bollinger's theory more important than ever. Just as we tolerate Nazis marching through Skokie, Illinois,[306] we must also tolerate "Innocence of Muslims" making its way around the Internet. The government can urge us to tolerate the former through First Amendment jurisprudence, but it has no formal power to make us tolerate the latter. The Nazis' speech can be forced out of the marketplace of ideas using counter speech. Yet for Bollinger, the marketplace is not a place that ideas should be forced out of; as tolerant citizens, we should never force speech completely out of the marketplace because, although the Nazis' speech may cause harm, such speech is integral to the process of our self-reformation. This perspective recognizes the calls for Facebook or Google to remove content as a phenomenon that looks less like a function of the

[302] *Supra* note 193.
[303] BOLLINGER, *supra* note 5, at 54.
[304] *Id*. at 86.
[305] CITRON, *supra* note 69.
[306] National Socialist Party of America v. Village of Skokie, 432 U.S. 43 (1977).

marketplace of ideas and more like what Bhagwat calls "societal vetoes of unpopular associations."[307]

To preserve the social benefit of some types of extreme speech and protect the victims of other, more directly harmful types, intermediaries need well-defined and uniform standards that distinguish these types of speech. The foundation for defining harmful speech should be the fighting words, incitement, and true threat exceptions to free speech as defined by the U.S. Supreme Court.[308] Intermediaries' calculations of whether some types of extreme speech should be considered incitement should not involve the speech's tendency to indirectly provoke hostile audiences to action. Instead, intermediaries should have clear standards distinguishing UGC that is extreme or uncivil from UGC that directly abuses a specific person, such as through revenge porn, cyber-bullying, trolling and cyber-harassment.

Conclusion

Mainstream online communication platforms such as Facebook and Twitter are becoming the locus of public discourse. What is said online has the potential to reach vast audiences and affect those audiences deeply. Hecklers no longer merely line the sidewalk across the police barricade from an incendiary speaker. Hecklers are also rioters who wreak havoc a world away from where a speaker uploads his or her inflammatory video to YouTube. Hecklers are the online crowds who petition Facebook to remove speech with which they disagree. Hecklers are the trolls who take to Twitter to intimidate women who speak up for a cause. Understanding how U.S. jurisprudence has treated hecklers

[307] Bhagwat, *supra* note 1.
[308] *Supra* notes 41-51.

throughout the last half-century is important for understanding how to judge hecklers today in networked communication. After some time brewing in the 1940s, 1950s and 1960s, the doctrine of the heckler's veto came to stand for the principle that state actors have a duty to protect speakers from hostile audiences who would seek to either actually do harm to the speakers, or threaten to do harm and thereby force law enforcement to silence the speaker. The doctrine of the heckler's veto teaches that individuals cannot take the easy way out with speech that they don't like, disagree with, or find offensive. Individuals cannot succumb to simply following community norms. At the same time, society cannot allow norm policing and trolling to be amplified to such an extent that it chills potentially valuable speech. For the most part, the state does not police speech online; that task is left to individuals and intermediaries. The doctrine of the heckler's veto has established the state's position of showing tolerance toward and a duty to protect unpopular speech. Individuals and digital intermediaries should heed Bollinger's theory of tolerance:[309] follow the example of government and show tolerance when faced with the potential to silence unpopular speech online.

This study of the changing norms of free speech in a context of greater intermediary and heckler control over speech is *not* an endorsement of the concept of "private censorship," a term that generically defines the ability of powerful private institutions to prevent important messages from reaching the public. Several of the scholars cited here embrace the term,[310] as do certain European scholars,[311] and American

[309] BOLLINGER, *supra* note 5, at 110.
[310] Nunziato, *supra* note 214.
[311] Mac Síthigh, *supra* note 213.

proponents of the political economy school of studying mass media.[312] The problem with "private censorship" is its semantic fallacy; by definition, only government can censor. Not only that, but loosely using the term "censorship" outside of its proper context is grossly offensive to those who must endure actual censorship at the hands of their government.[313] Intermediary "control" or "governance" should be considered its own issue, without conceptual ties to "censorship." Doing so will help the study of intermediary control retain legitimacy,[314] thereby leading to theoretically and doctrinally sound approaches to deal with the issue in both legal and social scientific research.

Popular condemnation of unpopular online speech, coupled with powerful intermediaries ready to answer to the demands of that popular condemnation, has more potential to quash speech than the marketplace of ideas should allow. Action must be taken to spare socially and politically valuable speech that is unpopular, extreme, or otherwise controversial from the whims of the heckling majority. For example, students should be taught about the nature of civil online discourse from a young age. In other words, "what free speech means in online communications" should be a topic that is taught as if it were part of a middle school civics course traditionally devoted to discussing free speech in public forums. Students should be taught what it means to be both a speaker sticking her or his neck out online, and part of the crowd that has the

---

[312] BAGDIKIAN; MCCHESNEY, *supra* note 231.
[313] Phillip Swann, *Fox Crosses the Line in Dish Fight*, TVPredictions.com, (Jan. 13, 2015), *available at* http://www.tvpredictions.com/dishfox011315.htm (arguing that by using the term "censorship" to refer to a contract dispute with satellite provider Dish Network, Fox News "shamefully dishonors people and societies that live under brutal reigns of repression that actually do prevent the dissemination of movies, books and other communications").
[314] *See, e.g.,* JAMES GRIMMELMANN, INTERNET LAW: CASES AND PROBLEMS, VER. 3.0 (2013) (Professor Grimmelmann made "intermediary control" one of the primary themes of his casebook).

power to silence the speaker through heckling. Citizens should have civic standards informing them when it is their responsibility to tolerate speech, and when they have the responsibility to mobilize the crowd to prevent harm.

Chapter 5: Facebook's Free Speech Balancing Act: A Case of Content Governance

Introduction

The purpose of this study is to explicate the concept of content governance: the control that digital communication intermediaries exercise over user-generated content (UGC). The particular focus of this explication is the governance of extreme UGC. Two key questions guide this explication: How and why do digital intermediaries respond to extreme UGC? What are the potential implications of their responses for public discourse in a system of networked communication?

All digital intermediaries engage in varying degrees of content governance. For example, in late 2014, Twitter deactivated the accounts of several users who proclaimed to be members of the so-called Islamic State in Iraq and Syria (ISIS).[1] This is an example of intermediaries removing content after users flagged it for violating the intermediaries' community standards.[2] As discussed in chapter 4, Google, the parent company of YouTube, removed the video "Innocence of Muslims" from view on the platform in several countries in September 2012.[3] This episode was an example of an intermediary removing content because government actors demanded its removal through a legal

[1] Jillian C. York, *Terrorists on Twitter*, SLATE (June 25, 2014), *available at* http://www.slate.com/articles/technology/future_tense/2014/06/isis_twitter_suspended_how_attempts_to_silence_terrorists_online_could_backfire.html.
[2] *See* Kate Crawford and Tarleton Gillespie, *What is a flag for? Social media reporting tools and the vocabulary of complaint*, NEW MEDIA & SOC'Y 1 (2014) (discussing the policies of various digital intermediaries of allowing users to "flag" undesirable content, as well as the reasons why individuals flag content).
[3] Adi Robertson, *From 'Desert Warrior' to 'Innocence of Muslims,' a Controversial YouTube Video Is Both Catalyst and Scapegoat*, THE VERGE (Sept. 17, 2012), *available at* http://www.theverge.com/2012/9/17/3346428/innocence-of-muslims-protests.

order,[4] as well as an example of an intermediary removing content in certain countries on its own volition.[5]

Content governance can vary in terms of the intermediary, the content in question and the reason for its removal. These variables make the total possible varieties of content governance too numerous to discuss in an entire dissertation, let alone one chapter. Therefore, this chapter will focus on how one specific digital intermediary—Facebook—governs extreme UGC. The choice to examine Facebook's content governance practices was made for several reasons.

As of this writing, Facebook is the second-most popular site on the World Wide Web,[6] with reportedly more than 1.39 billion active monthly users.[7] Women (58% of users) use Facebook only slightly more than men (42%), thus highlighting the relative gender parity of the social network.[8] In the United States, 71% of all adult Internet users were using Facebook in September 2014.[9] The U.S. Facebook population also has a vast age range: in January 2014, 9.8 million users (5.4% of all users) were 13-17 years old, 42 million (23.3%) were 18-24, 44 million (24.4%) were 25-34, 56 million (31.1%) were 35-54, and 28 million (15.6%) were over the age of 55.[10] U.S. Facebook users tend to vary

---

[4] LAURA DENARDIS, THE GLOBAL WAR FOR INTERNET GOVERNANCE 213 (2014).
[5] *Id*. at 158.
[6] *The Top 500 Sites on the Web*, ALEXA (Mar. 20, 2015), *available at* http://www.alexa.com/topsites (Google.com is the Web's most popular site, according to Alexa).
[7] *Company Info*, FACEBOOK (Mar. 21, 2015), *available at* http://newsroom.fb.com/company-info/.
[8] Albert Costill, *25 Amazing Facts about Facebook*, SEARCH ENGINE J. (Feb. 12, 2014), *available at* http://www.searchenginejournal.com/25-amazing-facts-facebook/88733/.
[9] *Social Networking Fact Sheet*, PEW RESEARCH CTR. (Sept. 2014), *available at* http://www.pewinternet.org/fact-sheets/social-networking-fact-sheet/.
[10] Ryan W. Neal, *Facebook Gets Older: Demographic Report Shows 3 Million Teens Left Social Network in 3 Years*, INT'L BUS. TIMES (Jan. 16, 2014), *available at* http://www.ibtimes.com/facebook-gets-older-demographic-report-shows-3-million-teens-left-social-network-3-years-1543092.

considerably in terms of political ideology.[11] Facebook's diversity extends well beyond the United States: as of March 2015, the company reports that 82.4% of its active users are outside the United States and Canada.[12] The platform is also diverse in terms of its function and the types of content that can be published on the site. Individuals use Facebook for a variety of reasons, such as entertainment, keeping up-to-date on the lives of friends and family, keeping up-to-date with news and current events, becoming civically engaged, and receiving support from people in their network.[13] Facebook is beginning to seriously challenge YouTube as the preferential site where users both upload and view videos.[14]

This collection of statistics makes Facebook an excellent subject for an analysis of content governance. It is not a niche platform that only caters to one target audience or to facilitating the publication of one type of UGC, thereby making it "the closest thing we have to a universal communication platform."[15] Its popularity and ubiquity mean that any changes it makes to its community guidelines will have a far-reaching effect on the norms of online freedom of expression. Facebook's users vary widely in terms of the social norms of freedom of expression that they come from, meaning that the social network faces the difficult task of seeking consensus among serious and potentially intractable

[11] *See* Timothy Macafee, *Some of These Things Are Not Like the Others: Examining Motivations and Political Predispositions Among Political Facebook Activity*, 29 COMPUTERS IN HUMAN BEHAVIOR 2766 (2013).
[12] *Company Info*, *supra* note 7.
[13] Aaron Smith, *6 New Facts about Facebook*, PEW RES. CTR. (Feb. 3, 2014), *available at* http://www.pewresearch.org/fact-tank/2014/02/03/6-new-facts-about-facebook/; Macafee, *supra*, note 11.
[14] Dan Kedmey, *Facebook Video Uploads Reportedly Overtake YouTube*, TIME (Dec. 9, 2014).
[15] Vindu Goel, *Facebook Clarifies Rules on What It Bans and Why*, N.Y. TIMES (Mar. 16, 2015).

differences in how users interpret the values and harms of certain types of speech.[16] The potential for lines separating abusive speech, political speech, hate speech and speech promoting terrorism to get blurred among this diverse group of users means that Facebook "walks a delicate line when it tries to ban violent or offensive content without suppressing the free sharing of information that it says it wants to encourage."[17] This chapter seeks to examine where Facebook—the virtually universal communication platform—draws this delicate line.

This chapter will review important changes that have been made to Facebook's community standards throughout its 11-year history. This analysis will involve comparing and contrasting cached snapshots of Facebook's community standards or terms of service pages[18] as compiled by the Internet Archive's "Wayback Machine."[19] Two criteria will guide the analysis of these pages. First, the analysis will look for changes in policies over the course of 11 years. This analysis will require tracking which items get added to newer versions of the standards and subtracted from older versions, as well as noting if or how definitions of certain key terms (such as "hate speech") change over time. Second, several benchmarks will be used to assess how Facebook's community standards balance protection of individuals' speech with prevention of harm. These benchmarks include legal tests for distinguishing protected from unprotected speech (discussed in chapters 3 and 4), as well as Facebook's interests within the

[16] *See* Michael Zimmer, *Facebook's Censorship Problem*, HUFFINGTON POST (June 22, 2011), *available at* http://www.huffingtonpost.com/michael-zimmer/facebooks-censorship-prob_b_852001.html.
[17] Goel, *supra* note 15.
[18] According to the Internet Archive, *infra*, Facebook did not create a separate page outlining community standards until 2011. A "Code of Conduct" page existed from May 2007 until 2011. Before then, all stipulations for what constituted allowable versus unallowable content was contained in Facebook's "Terms of Service" page.
[19] *Wayback Machine*, INTERNET ARCHIVE, *available at* http://archive.org/web/.

networked economy (discussed in chapter 2). Obviously, Facebook's community standards need not be as protective of speech as First Amendment jurisprudence, and the purpose of the analysis is not to make such an obvious argument. Rather, the goal of assessing Facebook's standards vis-à-vis legal standards is simply to understand which areas of speech Facebook protects or restricts more than others.

Next, this chapter will review examples of Facebook either removing or not removing extreme UGC that appeared to contravene the social network's community guidelines. The goal of this analysis is to give some empirical examples of how Facebook has governed UGC in accordance with its community standards. Methodologically speaking, these examples have much in common with cases in traditional First Amendment jurisprudence. Both sets of examples can be analyzed to illustrate broader principles of issues of freedom of expression. "Despite being a small sliver of overall activity, these events [of governing extreme speech on Facebook] have come to typify the dramatic struggles some users face" between protecting their ability to speak freely on Facebook and protecting themselves from harmful speech.[20] Similarly, First Amendment cases heard at the federal appellate level are outlier cases that do not reflect the "normal" course of interaction among speakers and audiences in the United States, yet these extreme cases represent the vanguard that ensures the protection of all communicative interactions, normal or fringe.[21] Therefore, the examples discussed herein are important because they create the contours of the constantly shifting norms of acceptable speech

[20] Luke Lancaster, *Facebook Updates Standards to Explain What It Will Remove*, CNET (Mar. 16, 2015), *available at* http://www.cnet.com/news/facebooks-updated-community-standards-explain-what-it-will-ban/.
[21] *See* Frederick Schauer & Richard Zeckhauser, *The Trouble with Cases*, *in* REGULATION VERSUS LITIGATION: PERSPECTIVES FROM ECONOMICS AND LAW (Daniel P. Kessler ed., 2011).

among users on Facebook.[22]

The chapter will conclude with a discussion on the broader implications of this analysis. Namely, this analysis can serve as an illustrative example of how norms of freedom of expression are being defined for networked communication via digital intermediaries. It also can shed light on the potential extent to which tolerance for extreme speech is changing in a networked communication environment.

Evolution of Facebook's Speech Codes and Community Standards

Digital intermediaries like Facebook must balance protecting users from harm and protecting users' ability to speak. They must distinguish between speech that causes actual harm from speech that merely offends. As clearly and transparently as possible, they must give definitions for each of these categories. They must determine when speech that contains a social or political message is worthy of protection and when it is liable for removal. They must walk a fine line between governing speech too strictly and too lightly, and they must explain their actions (or lack thereof) using strategic messages that simultaneously do not alienate champions of freedom of expression or advocates for stricter policing of harmful speech. They must justify how their standards allow some atrocious messages (such as the misogynist Facebook pages discussed in chapter 4) while they clamp down on arguably more innocuous and more socially valuable messages (such

[22] The analysis of examples of Facebook's controversial content governance is separate from the analysis of Facebook's terms of service and community standards because each analysis approaches content governance from a unique angle. The analysis of the evolution of Facebook's speech codes highlights the ambiguity with which Facebook has crafted policies that purport to both promote speech and protect users from harm. Meanwhile, the analysis of examples of content governance points to the subjectivity involved in judging the appropriateness of certain types of speech, which itself stems from the ambiguity of Facebook's speech codes.

as images of breastfeeding). Within this conflicted context, this study seeks to tackle the following research questions regarding Facebook's practices of content governance:

RQ1: How have Facebook's community standards changed from the origins of the social network until the standards' most recent update in March 2015?

RQ2: What instances have there been of Facebook controversially removing or not removing extreme UGC that seemed to contravene Facebook's community standards?

Like any website offering a service of any kind, Facebook has a page listing its terms and conditions that users must abide by. These terms include common matters such as personal jurisdiction, limitations on liability, and indemnity.[23] Facebook's policy regarding the privacy of users' data has received particular scrutiny among both journalists[24] and scholars.[25] The focus of this section, however, is on Facebook's terms and conditions as they apply to the speech that users are allowed to publish on Facebook.[26] This analysis is framed in terms of the chronological changes to Facebook's policies—as opposed to, say, an analysis framed in terms of the changes to particular categories of speech—because the goal of this chapter is to show the *evolutionary process* of Facebook's speech codes as a whole. This choice was made because such a process resembles the more protracted evolutionary process of First Amendment

[23] *See, e.g.*, Sandra Braman & Stephanie Roberts, *Advantage ISP: Terms of Service as Media Law*, 5 NEW MEDIA & SOC'Y 422 (2003); Specht v. Netscape Communications Corp., 306 F.3d 17 (2nd Cir. 2002).
[24] *See, e.g.*, Emily Steel & Geoffrey A. Fowler, *Facebook in Privacy Breach*, WALL ST. J. (Oct. 18, 2010); Issie Lapowsky, *Facebook Rolls Out Clearer Privacy Policy, but You Still Can't Control Your Data*, Wired (Nov. 13, 2014), *available at* http://www.wired.com/2014/11/facebook-revamps-privacy-policy/.
[25] *See, e.g.*, Benhard Debatin, Jennette P. Lovejoy, Ann-Kathrin Horn & Brittany N. Hughes, *Facebook and Online Privacy: Attitudes, Behaviors, and Unintended Consequences*, 15 J. COMPUTER-MEDIATED COMM. 83 (2009); danah boyd & Eszter Hargittai, *Facebook Privacy Settings: Who Cares?* 8 FIRST MONDAY n.p. (2010); Christian Fuchs, *The Political Economy of Privacy on Facebook*, 13 TELEVISION & NEW MEDIA 139 (2012).
[26] The section will not include an analysis of Facebook's copyright provisions, as they are standard and based on the Digital Millennium Copyright Act of 1998, 17 U.S.C. § 512.

jurisprudence in the United States since the 1930s. It is argued here that the nature of the evolutionary process, itself, will reveal Facebook's attempts to balance its support for freedom of expression and protecting users from harmful speech.

Methods

To assess potential changes to Facebook's community standards, this chapter examines cached copies of the pages that include Facebook's policies toward users' speech, as archived by the "Wayback Machine" of the Internet Archive project.[27] Assembling a body of cached pages requires one to enter a current URL into a search bar on the Wayback Machine's site. The service then sends back a list of dates at which its "bots" had taken a "snapshot" of the page corresponding to the URL. These cached pages look and function like live webpages (i.e. they are not JPEG or PDF files of the pages). However, not all the links on these pages will lead to pages that were cached on the same date (i.e. clicking on a link in a cached page will only open onto that link's page if the Wayback Machine also has a cached version of that page in its system).

Figure 5-1: Wayback Machine snapshots of Facebook's "Community Standards" page

[27] *Supra* note 19.

The process of collecting pages for this analysis began with entering the URL "https://www.facebook.com/communitystandards" (the current URL as of March 2015) into the Wayback Machine's search bar. The results indicate that this site only has been saved in the Machine's database after January 27, 2011 (see Figure 5-1: Wayback Machine snapshots of Facebook's "Community Standards). This does not necessarily mean that this site did not exist before that date, but rather that the Wayback Machine only began saving it at that date, or that the URL did not exist prior to the date. However, it is a strong clue that a specific page for Facebook's community standards did not exist before that date. Using this clue, the next URLs entered into the Machine were both "www.facebook.com" (the company's current root URL) and "www.thefacebook.com" (the company's original root URL). On the earliest possible versions of each of those sites, the link to the company's terms and conditions was clicked on. This link was chosen because no other link on the company's main page indicated that it contained anything about guidelines for users' speech. The earliest working link for "www.thefacebook.com/terms.php" was June 13, 2004 (only a few months after the company was founded). The latest update for this URL was June 28, 2005. After this date, the company switched its root URL from "www.thefacebook.com" to "www.facebook.com." Thus, the URL "https://www.facebook.com/terms.php" was entered into the Wayback Machine's search bar, which returned nearly 3,000 cached snapshots of this page from November 26, 2005 to the time of this writing (March 2015). The analysis of the "terms of use" page from May 24, 2007, revealed when Facebook created its first separate "Content Code of Conduct" page to supplement its "terms of

use" page. According to the Wayback Machine, this URL existed from May 2007 to August 2010. (See Figure 5-2. The site reports that it took snapshots of the page containing this URL until February 2013, but all snapshots after August 2010 show this URL returned a "Page Not Found" message.)

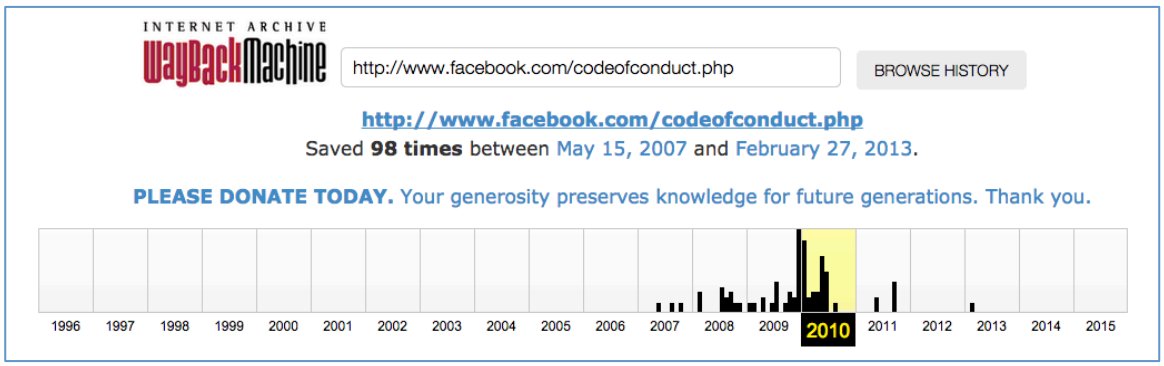

Figure 5-2: Wayback Machine snapshots of Facebook's short-lived "Code of Conduct" page

Combining the collection of cached pages from the URLs "https://www.thefacebook.com/terms.php," "https://www.facebook.com/terms.php," "http://www.facebook.com/codeofconduct.php," and "https://www.facebook.com/communitystandards," the method for collecting followed the following rationale. Links to snapshots of both sets of "Terms" pages were clicked on until a link showed a snapshot with an updated date (clearly visible on the top of the page). PDFs were made of each of these pages to facilitate the analysis (N = 14). Clicking through the links to snapshots of the "Code of Conduct" page revealed that this page was never updated during its entire lifespan[28] (thus, N = 1). Unlike the "Terms of Use" and "Statement of Rights and Responsibilities" pages, the "Community Standards" page does not contain dates of official updates. Thus, the method of selecting cached pages from the

[28] *Facebook Content Code of Conduct*, INTERNET ARCHIVE (May 24, 2007), *available at* http://web.archive.org/web/20100501072842/http://www.facebook.com/codeofconduct.php.

Wayback Machine to analyze involved randomly selecting a one page roughly per year from the Machine's timeline. The dates of cached pages analyzed here are: February 9, 2011 (the original page); December 15, 2012; November 9, 2013; and February 8, 2015 (n = 4) (see Appendix 2: Facebook's Community Standards).[29] Added to this collection was Facebook's March 15, 2015, major update to its community standards, which was not yet available on the Wayback Machine at the time of writing.[30]

The content analysis of these documents involved a close reading[31] of the portions of the "terms and conditions" and "community standards" pages that involved users' speech. This close reading involved comparing and contrasting the language used in the key sections being analyzed. Any language that was added, deleted, or changed within these sections was highlighted. The results and a discussion of this analysis are discussed in tandem below, in two separate sections: "Terms of Use/Rights and Responsibilities" and "Code of Conduct/Community Standards."

---

[29] The element of randomness comes from the lack of knowledge of the exact date that the Wayback Machine will return when one clicks on its timeline feature. Only four pages were chosen based on an assumption that the pages would not change drastically. The analysis herein validates that assumption.

[30] It should be noted here that using the Wayback Machine as a means to capture changes to Facebook's policies has its weaknesses. First, there is no guarantee that anyone actually saw any of the versions of these pages when they were live. In other words, the cached pages simply could be snapshots of a page that was only briefly published on Facebook's servers. Second, it is possible that changes that Facebook made to its pages may never have been cached, meaning that important changes to the social network's terms of use could be missing from the history according to the Wayback Machine. Despite these limitations, it is argued here that the Wayback Machine nevertheless gives researchers a sufficiently large and varied sample to accurately assess the changes to Facebook's terms and conditions over the last 11 years.

[31] The term "close reading" is widely used in legal research to refer to a critical, qualitative textual analysis of legal language (e.g. from a case or statute) whose goal is to uncover broader contextual meaning beyond the plain language. In the field of mass communication, this method of analysis and its ultimate goal is no different than a qualitative analysis of texts such as images or news reporting. *See, e.g.*, NORMAN FAIRCLOUGH, DISCOURSE AND SOCIAL CHANGE (1992); Elfriede Fürsich, *In Defense of Textual Analysis*, 10 JOURNALISM STUDIES 238 (2009); GIOVANNA DELL'ORTO, THE HIDDEN POWER OF THE AMERICAN DREAM: WHY EUROPE'S SHAKEN CONFIDENCE IN THE UNITED STATES THREATENS THE FUTURE OF U.S. INFLUENCE (2008); THOMAS R. LINDLOF AND BRYAN C. TAYLOR, QUALITATIVE COMMUNICATION RESEARCH METHODS 246 (2011); William A. Gamson and Andre Modigliani, *Media Discourse and Public Opinion on Nuclear Power: A Constructionist Approach*, 95 AM. J. OF SOCIOLOGY, 1, 3 (1989).

Results

Terms of Use/Rights and Responsibilities

Together, the "Terms" and "Standards" pages follow a broad trend of moving from general rules to more specific rules over time. Facebook also appears to improve the organization of the information on these pages over time. The earliest "terms" page (titled "Terms of Use") contains only basic guidelines regarding users' speech. The site contains the broad stipulation: "You understand and agree that Thefacebook may review and delete any content, photos or profiles (collectively 'Content') that in the sole judgement [*sic*] of Thefacebook violate this Agreement of which might be offensive, illegal, or that might violate the rights, harm, or threaten the safety of Members."[32] Regarding the protection of users from harm, the terms state: "Although Thefacebook cannot monitor the conduct of its members off the Web site, it is also a violation of these rules to use any information obtained from the Service in order to harass, abuse, or harm another person."[33] Two observations can be made here. First, "TheFacebook" has sole discretion over what content violates its policies, and it is not sharing its reasoning behind that discretion with its users at this time. Second, TheFacebook claims no responsibility to adequately police the content that users publish; in the words of legal scholars Sandra Braman and Stephanie Roberts, it is exercising "control without responsibility."[34] One specific prohibition that the terms state is that users will not create any profiles "that purport to represent an animal, place, inanimate objects, fictional character, or real

[32] *TheFacebook Terms of Use*, INTERNET ARCHIVE (June 13, 2004), *available at* http://web.archive.org/web/20040613185924/http://www.thefacebook.com/terms.php.
[33] *Id*.
[34] Braman & Roberts, *supra* note 23, at 438.

individual who is not you."[35] This "real name" policy continues to the present day (at the time of this writing), and, as will be shown below in accounts of several examples, it continues to be a thorn in the side for several groups of users.[36]

The "Terms of Use" update from June 28, 2005, contains the same language as the June 13, 2004 page regarding the company's ability to review and delete content, but it now includes a specific section titled "Member Conduct."[37] In this section, Facebook states that it prohibits "libelous, defamatory or otherwise unlawful material."[38] Abuse continues to be an important phenomenon for Facebook to control, though the company continues to convey that it has sole discretion over defining when content violates its policies. It prohibits users from posting "any content *that we deem* to be harmful, threatening, abusive, harassing, vulgar, obscene, hateful, or racially, ethnically or otherwise objectionable."[39] It reiterates in a separate paragraph that users are not allowed to "intimidate or harass another" user.[40]

An October 3, 2005, update to the "Terms of Use" page (with the company now called Facebook rather than TheFacebook) contained no changes to rules governing users' speech, though this update coincides with Facebook's decision to open its platform up to individuals over the age of 13, rather than 18.[41] Age will become an important

---

[35] *Id*.
[36] *See infra* notes 134-137.
[37] *TheFacebook Terms of Use*, INTERNET ARCHIVE (June 28, 2005), *available at* http://web.archive.org/web/20050630235618/http://www.thefacebook.com/terms.php.
[38] *Id*.
[39] *Id*. (emphasis added).
[40] *Id*.
[41] *Facebook Terms of Use*, INTERNET ARCHIVE (Oct. 3, 2005), *available at* http://web.archive.org/web/20060203041141/http://www.facebook.com/terms.php.

factor in how Facebook crafts its community standards.[42] A February 27, 2006, update similarly contained no changes to rules governing users' speech.[43]

Facebook's December 13, 2006, update adds several new provisions to the company's speech rules.[44] The list of Facebook's prohibited content now includes "any content *that we deem* to be harmful, threatening, unlawful, defamatory, infringing, abusive, inflammatory, harassing, vulgar, obscene, fraudulent, invasive of privacy or publicity rights, hateful, or racially, ethnically or otherwise objectionable."[45] The new terms also ban "content that would constitute, encourage or provide instructions for a criminal offense, violate the rights of any party, or that would otherwise create liability or violate any local, state, national or international law."[46] Three new observations can be made here. First, Facebook appears to acknowledge that the speech published on its platform can and does have major social and legal repercussions in the physical world.[47] Second, Facebook specifically singles out speech related to the category of "criminal activity" as a concern, yet it does not give more specific information on what such speech would look like. Third, Facebook combines elements of all three of Smolla's categories of harm into one amalgamation of prohibited speech: physical, relational and reactive

---

[42] *See Community Standards*, FACEBOOK (Mar. 15, 2015), *available at* https://www.facebook.com/communitystandards (the standards state that Facebook "restrict[s] the display of nudity because some audiences within [its] global community may be sensitive to this type of content - particularly because of their cultural background or age").
[43] *Facebook Terms of Use*, INTERNET ARCHIVE (Feb. 27, 2006), *available at* http://web.archive.org/web/20060301120239/http://www.facebook.com/terms.php.
[44] *Facebook Terms of Use*, INTERNET ARCHIVE (Dec. 13, 2006), *available at* http://web.archive.org/web/20070202024540/http://www.facebook.com/terms.php
[45] *Id*. (emphasis added).
[46] *Id*.
[47] *See* David R. Johnson & David Post, *Law and Borders: The Rise of Law in Cyberspace*, 48 STAN. L. REV. 1367 (1996).

harms.[48] The danger in lumping all of these harms together is that Facebook may cease to recognize the constitutional distinctions among these three harms and the level of protection that each should receive according to First Amendment jurisprudence.

The terms also prohibit "content that, *in the sole judgment of Company*, is objectionable or which restricts or inhibits any other person from using or enjoying the Site, or which may expose Company or its Users to any harm or liability of any type."[49] Like the prohibitions listed above, these terms make clear that Facebook is the ultimate arbiter when it comes to deciding between allowable and unallowable activity. Unlike the prohibitions listed above, the focus of these terms is much broader, encompassing issues of potential legal liability and the prevention of others' use of Facebook as well as objectionable content. This provision serves as a vague catchall to back up the already vague terms and conditions pertaining to users' speech.

The November 15, 2007 update of the "Terms of Use" page adds one important stipulation from the May update: "FACEBOOK DOES NOT PRE-SCREEN OR APPROVE FACEBOOK PAGES."[50] Here, Facebook is reiterating its longtime policy of not having responsibility over the content users create. The September 23, 2008 update contains no changes to the rules governing user speech.[51]

---

[48] RODNEY A. SMOLLA, FREE SPEECH IN AN OPEN SOCIETY 48 (1992)
[49] *Id*. (emphasis added).
[50] *Facebook Terms of Use*, INTERNET ARCHIVE (Nov. 15, 2007) (all-caps in original), *available at* http://web.archive.org/web/20080102211804/http://www.facebook.com/terms.php.
[51] *Facebook Terms of Use*, INTERNET ARCHIVE (Sept. 23, 2008), *available at* http://web.archive.org/web/20081120093758/http://www.facebook.com/terms.php

On May 1, 2009, the "Terms of Use" page becomes the "Statement of Rights and Responsibilities" page.[52] With this change in title, Facebook appears to be recognizing users' communicative agency via its platform while establishing a partnership with users to help ensure the company and its community enjoy continued growth and prosperity through harmonious relations. Along with the change in title, the page includes major changes to the organization of its policies, as well as an attempt to convey those policies in simple, less legal-sounding language. Under the subhead "Safety," the page states, "We [Facebook] do our best to keep Facebook safe, but we cannot guarantee it. We need your help to do that, which includes the following commitments."[53] Despite the attempt by the language to engage users as partners in maintaining the Facebook "community," the list of prohibited content remains virtually unchanged. No official attempt is made to either expand or contract the types of speech that Facebook will allow. The page mandates that users "will not bully, intimidate or harass any user," "will not post content that is hateful, threatening, pornographic, or that contains nudity or graphic or gratuitous violence," and "will not use Facebook to do anything unlawful, misleading, malicious, or discriminatory."[54] The August 28, 2009 update adds: "You will not develop or operate a third-party application containing alcohol-related or other mature content (including advertisements) without appropriate age-based restrictions."[55] This particular update is a reflection of Facebook's expansion into selling advertising, and it reveals a desire for

[52] *Facebook Statement of Rights and Responsibilities*, INTERNET ARCHIVE (May 1, 2009), *available at* http://web.archive.org/web/20090731004801/http://www.facebook.com//terms.php
[53] *Id*.
[54] *Id*.
[55] *Facebook Statement of Rights and Responsibilities*, INTERNET ARCHIVE (Aug. 28, 2009), *available at* http://web.archive.org/web/20100115012034/http://www.facebook.com/terms.php

commercial content posted on the site to comport with the social network's community standards.[56]

The December 21, 2009 update of the "Statement of Rights and Responsibilities" page includes two new provisions for Facebook pages.[57] Pages "look similar to personal profiles," and "are managed by people who have personal profiles," but they are designed "for businesses, brands and organizations" to build a public presence among Facebook's users.[58] Regarding user governance of pages, the new guidelines state, "You may not establish terms beyond those set forth in this Statement to govern the posting of content by users on a Page you administer, except you may disclose they types of content you will remove from your Page and grounds for which you may ban a user from accessing the Page."[59] They continue, "You will restrict access to your Page in order to comply with all applicable laws. For example, if your Page includes content not suitable for minors, you will use your Page to block minors from accessing your Page."[60] Thus, Facebook is giving users greater control over the governance of the pages they create, while simultaneously stipulating that Facebook's own speech guidelines are the ultimate source of authority on which users should base their governance practices. These provisions were moved to a separate "Page Terms" site in the April 22, 2010 "Statement

---

[56] Rupert Neate and Rowena Mason, *Networking Site Cashes in on Friends*, The Telegraph (Jan. 31, 2009), *available at* http://www.telegraph.co.uk/finance/newsbysector/mediatechnologyandtelecoms/4413483/Networking-site-cashes-in-on-friends.html.
[57] *Facebook Statement of Rights and Responsibilities*, INTERNET ARCHIVE (Dec. 21, 2009), *available at* http://web.archive.org/web/20100402021418/http://www.facebook.com/terms.php.
[58] *How Are Pages Different from Profiles?* FACEBOOK HELP CENTER (2015), *available at* https://www.facebook.com/help/217671661585622. *See also* Tiffany Black, *Facebook Profile vs. Facebook Page vs. Facebook Group*, ABOUT TECH (n.d.), *available at* http://facebook.about.com/od/Basics/fr/Facebook-Profile-Vs-Facebook-Page-Vs-Facebook-Group.htm.
[59] *Facebook Statement of Rights and Responsibilities*, INTERNET ARCHIVE (Dec. 21, 2009).
[60] *Id*.

of Rights and Responsibilities" update.[61] The "Rights and Responsibilities" page's rules governing users' speech do not change from August 25, 2010,[62] to January 30, 2015[63] (the most recent update as of this writing).

Code of Conduct/Community Standards

May 2007 – January 2015

On May 24, 2007, Facebook first unveiled its "Code of Conduct," which users could access via a link in the "Terms of Use" page.[64] As stated above, this page was never updated during its entire lifespan.[65] This new page (see Appendix 1: Content Code of Conduct) does not add anything new to the rules governing speech found in Facebook's "Terms" or "Rights and Responsibilities" pages. Rather, the "Content Code of Conduct" page can be considered an effort to move the rules governing speech onto one page, separate from other terms and conditions. As stated above, according to the Wayback Machine, Facebook's "Code of Conduct" page became its "Community Standards" page on or around January 27, 2011 (see Figure 5-1: Wayback Machine snapshots of Facebook's "Community Standards" page, and Figure 5-2: Wayback Machine snapshots of Facebook's short-lived "Code of Conduct" page).[66] All changes to the "Community Standards" page are highlighted in the Appendix (see Appendix 2:

---

[61] *Facebook Statement of Rights and Responsibilities*, INTERNET ARCHIVE (Apr. 22, 2010), *available at* http://web.archive.org/web/20100602022403/http://www.facebook.com/terms.php.
[62] *Facebook Statement of Rights and Responsibilities*, INTERNET ARCHIVE (Aug. 25, 2010), *available at* http://web.archive.org/web/20100929171235/https://www.facebook.com/terms.php.
[63] *Facebook Statement of Rights and Responsibilities*, INTERNET ARCHIVE (Jan. 30, 2015), *available at* http://web.archive.org/web/20150301012017/https://www.facebook.com/legal/terms.
[64] *Facebook Content Code of Conduct*, INTERNET ARCHIVE (May 24, 2007).
[65] *Supra* note 28.
[66] *Facebook Community Standards*, INTERNET ARCHIVE (Jan. 27, 2011), *available at* http://web.archive.org/web/20110127224041/https://www.facebook.com/communitystandards/.

Facebook's Community Standards). The present analysis will discuss important trends in those changes.

The first important change found in the Community Standards is how Facebook refers to itself: as a "global community"[67] rather than a "social utility."[68] Thus, the pretext for these updated standards comes from the notion that Facebook is a place or an experience rather than simply a tool for communication. Second, the Community Standards organize problematic areas of speech by category: Threats; Promoting Self-Harm; Bullying & Harassment; Hate Speech; Graphic Violence; Sex & Nudity; Theft, Vandalism, or Fraud; Identity & Privacy; Intellectual Property; and Phishing & Spam.[69] The types of speech that fall within these categories are not new to Facebook's rules (they're consistently found within the "Rights and Responsibilities" page), but their organization into specific categories conveys them with added seriousness and gives users the potential to better recognize these problematic types of speech. Third, Facebook includes instructions on how users can report abuse to the officials at the company. Here, Facebook conveys the caveat that not all offensive or disagreeable material will violate its Standards and therefore qualify for removal. Instead, Facebook states that it offers users tools to personally filter any content they find offensive.[70] These personal filtering tools are beyond the scope of the present analysis. Many scholars have conducted research on whether tools such as these contribute to creating a so-called "echo chamber" in which individuals only encounter and engage with content that they agree with or that

[67] *Facebook Community Standards*, INTERNET ARCHIVE (Feb. 9, 2011), *available at* http://web.archive.org/web/20110209013433/https://www.facebook.com/communitystandards//.
[68] *Facebook Content Code of Conduct*, INTERNET ARCHIVE (May 24, 2007).
[69] *Id*.
[70] *Id*.

they do not find offensive.[71] Their research poses interesting questions for how much such an echo chamber—to the extent that it exists—may affect individuals' tolerance toward extreme speech. Although these questions must be set aside for future research, for the present analysis it is important to point out that Facebook is willing to give users a more insular experience on its platform in the event that its community standards alone are not enough to create a safe environment for these users.

This basic organization of problematic speech does not change in future updates. Rather, updates add new categories of problematic speech, clarify certain categories with context or specific examples, or make categories more vague. The updated page cached on December 15, 2012,[72] contains numerous updates to the original Standards (see Appendix 2: Facebook's Community Standards). The page no longer refers to Facebook as a "global community," but rather claims that the "conversation that happens on Facebook—and the opinions expressed here—mirror the diversity of the people using Facebook."[73] Under "Violence and Threats," Facebook expands from simply prohibiting "credible threats" to banning any speech that could possibly result in "physical harm," a "direct threat to public safety," or the potential for "real-world violence" or "financial

---

[71] *See, e.g.*, Elanor Colleoni, Alessandro Rozza & Adam Arvidsson, *Echo Chamber or Public Sphere? Predicting Political Orientation and Measuring Political Homophily in Twitter Using Big Data*, 64 J. COMM. 317, 328 (2014) (finding: "If we look at Twitter as a social medium we see higher levels of homophily and a more echo chamber-like structure of communication. But if we instead focus on Twitter as a news medium, looking at information diffusion regardless of social ties, we see lower levels of homophily and a more public sphere-like scenario."); Jonathan Zittrain, *The Fourth Quadrant*, 78 FORDHAM L. REV. 2767, 2767 (2010) (arguing that social networking sites are platforms in which only certain content is "cultivate[d] and accelerate[d]."

[72] *Facebook Community Standards*, INTERNET ARCHIVE (Dec. 15, 2012), *available at* http://web.archive.org/web/20121215024155/http://www.facebook.com/communitystandards.

[73] *Id*.

harm."[74] Although these provisions do not completely follow the true threat or incitement standards found in First Amendment jurisprudence (nor do they provide the same high level of protection for speech as these standards), they do convey the notion that physical harm resulting from speech is the worst possible harm and therefore such speech has the potential to be regulated.[75] Under "Self-Harm," Facebook notes that it will work specifically with "suicide prevention agencies around the world"[76] rather than "the relevant authorities"[77] to offer users help when the company becomes aware of suicidal content.

Under "Hate Speech," the policy says that Facebook will distinguish "between serious and humorous speech."[78] The fact that Facebook is recognizing this distinction seems to reveal a firmer commitment on the part of the social network to protect users' speech, or at the very least offer somewhat clearer standards on what counts as protected versus unprotected in the category of hate speech. Interestingly, the update now defines hate speech as "attacks against others"[79] rather than "singl[ing] out individuals."[80] This change potentially broadens the definition of hate speech to include attacks against groups of people, and it also distinguishes hate speech from personal abuse or bullying against a specific individual. Facebook may also be attempting to sound more sensitive in the very language of its terms, as the term "disease"—one of the conditions from the

[74] *Id*.
[75] *See* SMOLLA, *supra* note 48.
[76] *Id*.
[77] *Facebook Community Standards*, INTERNET ARCHIVE (Feb. 9, 2011).
[78] *Facebook Community Standards*, INTERNET ARCHIVE (Dec. 15, 2012).
[79] *Id*.
[80] *Facebook Community Standards*, INTERNET ARCHIVE (Feb. 9, 2011).

original standards denoting classes of individuals protected from hate speech—is changed to the more generic and neutral term "medical condition."[81]

The category "Graphic Violence" is given the more generic label "Graphic Content" in the December 2012 update, and the types of speech prohibited under this category are also expressed in more general terms. Instead of "[s]adistic displays of violence against people or animals, or depictions of sexual assault"[82] being prohibited, "any graphic content [shared] for sadistic pleasure"[83] is prohibited under the December update. Meanwhile, the "Nudity and Pornography" category is outlined in more specific terms. Instead of there being a "strict 'no nudity or pornography' policy,"[84] Facebook now "aspires to respect people's right to share content of personal importance, whether those are photos of a sculpture like Michelangelo's David or family photos of a child breastfeeding."[85]

What is important about these changes to the policies governing the categories of hate speech, graphic content and nudity/pornography is that Facebook is choosing to focus on the intent of the speaker rather than simply the nature of the speech. On the one hand, this approach puts Facebook's standards more in step with legal tests distinguishing protected from unprotected speech in First Amendment jurisprudence. For example, one interpretation of Justice O'Connor's enunciation of the true threat test requires proof that a person accused of threatening someone actually intended the utterance to be

---

[81] *Id*.
[82] *Id*.
[83] *Facebook Community Standards*, INTERNET ARCHIVE (Dec. 15, 2012).
[84] *Facebook Community Standards*, INTERNET ARCHIVE (Feb. 9, 2011).
[85] *Facebook Community Standards*, INTERNET ARCHIVE (Dec. 15, 2012).

threatening.[86] Likewise, the test for obscenity requires the state to prove that the allegedly obscene speech has no "serious literary, artistic, political, or scientific value."[87] On the other hand, this approach gives no indication as to how intent will be determined, and it could reasonably be argued that "intent" under this approach would not be nearly as protective as the subjective intent standard from *Virginia v. Black*. It may more closely resemble the simple objective or "reasonable person" standard for true threats, which the U.S. Court of Appeals for the Third Circuit held was sufficient to prove a true threat in 2013.[88] The bottom line is that Facebook is still searching for that ideal point at which to balance protection of speech and prevention of harm. Although that point appears to be shifting more toward protection, the tremendous amount of gray area surrounding its standards keeps the scales tipped more toward prevention of harm.

The updated page cached on November 9, 2013,[89] contains an update only to Facebook's policy on Graphic Content (see Appendix 2: Facebook's Community Standards). This update retains key terms for delineating inappropriate content (e.g. depictions of violence shared for "sadistic effect or to celebrate or glorify violence), yet it also contains two caveats. In December 2012, the policies recognized that "graphic imagery is a regular component of current events."[90] The November 2013 update goes into greater detail, acknowledging that graphic content is sometimes shared to "raise awareness" about "human rights abuses or acts of terrorism."[91] Once again we see

[86] Virginia v. Black, 538 U.S. 343, 360 (2003) (O'Connor, J., writing for the plurality).
[87] Miller v. California, 413 U.S. 15, 24 (1973).
[88] United States v. Elonis, 730 F.3d 321 (3rd Cir. 2013).
[89] *Facebook Community Standards*, INTERNET ARCHIVE (Nov. 9, 2013), *available at* http://web.archive.org/web/20131109125448/https://www.facebook.com/communitystandards.
[90] *Facebook Community Standards*, INTERNET ARCHIVE (Dec. 15, 2012).
[91] *Facebook Community Standards*, INTERNET ARCHIVE (Nov. 9, 2013).

Facebook putting emphasis on the speaker's intent behind the speech, with the company acknowledging the power it affords its users to publish speech of profound social and political importance. But we also see an adaptation of Facebook's focus on intent: it places the responsibility on users to "carefully [choose] the audience for the [graphic] content" and to "warn their audience about the nature of the content … so that their audience can make an informed choice about whether to [view] it."[92] In other words, users must make their intent transparent to both Facebook and other users, while Facebook is declaring that it should not be responsible for guessing the intent behind users' speech. Meanwhile, the updated page cached on February 8, 2015,[93] adds a category of problematic speech that has been seen before in Facebook's "Terms" and "Rights and Responsibilities" pages: Regulated Goods (see Appendix 2: Facebook's Community Standards) such as firearms, alcohol, tobacco, or adult products. This addition is not so much an update as a reiteration of a policy.

March 2015 Update

On March 15, 2015, Facebook launched a completely redesigned version of its community standards[94] (see Appendix 2: Facebook's Community Standards). The most obvious update to the standards is the design of the page: instead of mere list of terms, the new site contains broad categories of goals ("Keeping You Safe," "Encouraging Respectful Behavior," "Keeping Your Account and Personal Information Secure," and "Reporting Abuse"), each with its own set of subcategories that are accessed by clicking

[92] *Id*.
[93] *Facebook Community Standards*, INTERNET ARCHIVE (Feb. 8, 2015), *available at* http://web.archive.org/web/20150209145145/https://www.facebook.com/communitystandards/.
[94] *Facebook Community Standards*, FACEBOOK (Mar. 15, 2015), *available at* https://www.facebook.com/communitystandards.

links in a sidebar next to the broad categories. For example, "Direct Threats," "Self-Injury," and "Bullying and Harassment" are under "Keeping You Safe," while "Nudity," "Hate Speech," and "Violence and Graphic Content" are under "Encouraging Respectful Behavior"[95] (see Figure 5-3 and Figure 5-4).

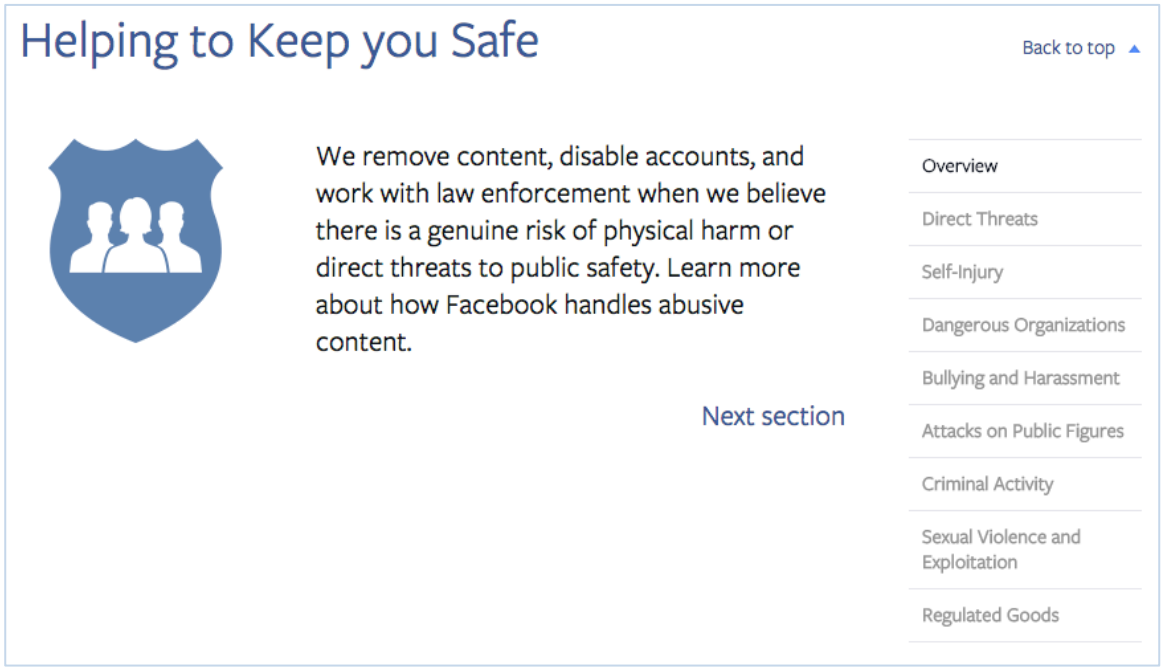

Figure 5-3: "Helping to Keep you Safe"

[95] *Id.*

Figure 5-4: "Encouraging respectful behavior"

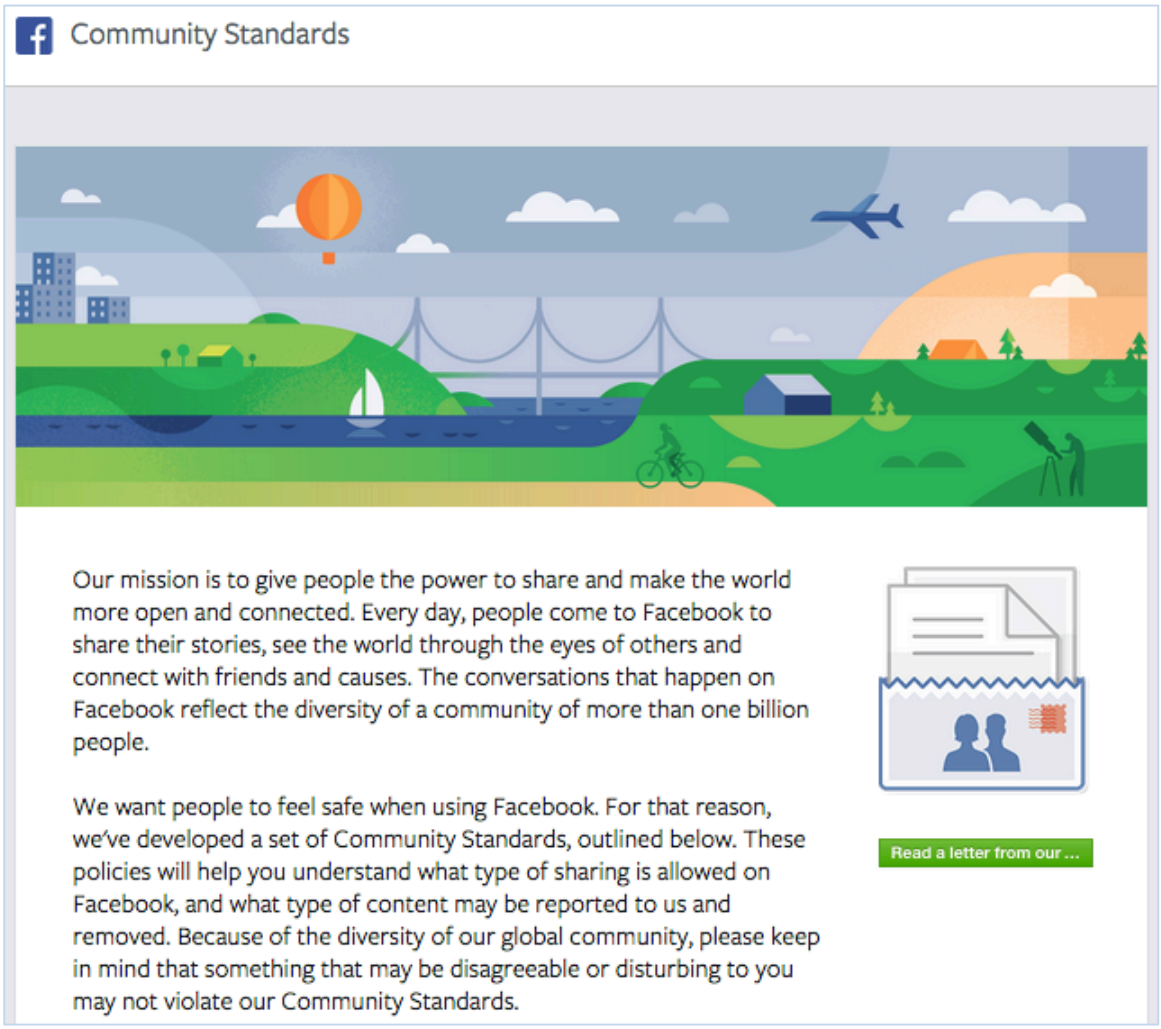

Figure 5-5: Opening to March 15, 2015 Update of Facebook's Community Standards

The new standards open by stating that Facebook is on a "mission … to give people the power to share and make the world more open and connected"[96] (see Figure 5-5). At the same time, Facebook "want[s] people to feel safe when using" the service.[97] These two goals frame each of the categories in the 2015 update. Virtually all of the core terms and definitions from past iterations of the standards have not changed, yet now they have specifications and contextual caveats added to them to highlight (as clearly as

[96] *Id.*
[97] *Id.*

possible) the boundary between the speech that Facebook hopes to champion and that which it hopes to quash. The section on "Direct Threats" adds the provision that Facebook "may consider things like a person's physical location or public visibility in determining whether a threat is credible."[98] This provision shows that Facebook's focus on speakers' intent is once again evolving, with the company declaring that it will assume some responsibility in determining the intent behind the speech by using the tools (i.e. data) that it readily has at its disposal.

The category of "Dangerous Organizations" is amended to specifically include groups that engage in terrorist activity or organized criminal activity. Under this category, "supporting" or "condoning" such activities is forbidden, though Facebook "welcome[s] broad discussion and social commentary on these general subjects."[99] The standards do not go into any greater detail on this category. "Bullying and Harassment" is updated with several examples of what may constitute this infraction, such as shaming, degrading or blackmailing private individuals.[100] A new prohibition specifically against the phenomenon of "revenge porn" is included.[101] A new section titled "Attacks on Public Figures" has also been added. Here, the standards stipulate that credible threats and hate speech directed at public figures will be removed just as they would be if directed at private individuals, although Facebook does "permit open and critical discussion" of public figures.[102]

---

[98] *Id*.
[99] *Id*.
[100] *Id*.
[101] *Id*.
[102] *Id*.

The new guidelines also state that sometimes it is a user's responsibility to make clear to Facebook and the Facebook community that its speech should be considered valuable. For example, under the update for the category of "Hate Speech," Facebook includes an exception for when individuals share "someone else's hate speech for the purpose of raising awareness or educating others about that hate speech. When this is the case, we expect people to clearly indicate their purpose, which helps us better understand why they shared that content."[103] This statement is perhaps the clearest declaration yet regarding Facebook's focus on users' intent. Users must not simply state a message, but rather they must provide any relevant context behind that message to ensure that it remains protected on Facebook's platform.

In the update for "Nudity," the standards offer perhaps the most absolute provisions of all, although even these provisions come with a caveat. The standards state, "we always allow photos of women actively engaged in breastfeeding or showing breasts with post-mastectomy scarring. We also allow photographs of paintings, sculptures, and other art that depicts nude figures."[104] However, an earlier part of the same section reads, "our policies can sometimes be more blunt than we would like and restrict content shared for legitimate purposes. We are always working to get better at evaluating this content and enforcing our standards."[105] In other words, Facebook is urging its users to realize that its system of speech governance is not perfect, as much as Facebook wants it to be. Ideally, users will be forgiving and not take the rejection of their content personally as

[103] *Id*.
[104] *Id*.
[105] *Id*.

Facebook continues to find the ideal balance between protecting speech and preventing harm to users.

One key strategy that Facebook appears to follow in its latest Community Standards is to portray state actors as the main enemies of freedom of expression. The standards read, "[W]e may have to remove or restrict access to content because it violates a law in a particular country, even though it doesn't violate our Community Standards."[106] The standards then place Facebook on the side of individual users and against state actors by stating, "We challenge requests that appear to be unreasonable or overbroad. And if a country requests that we remove content because it is illegal in that country, we will not necessarily remove it from Facebook entirely, but may restrict access to it in the country where it is illegal."[107] Thus, Facebook occupies a relative position as a champion of protecting freedom of expression. Even if Facebook does remove users' content in certain situations, it still is not as bad as government censors. But if government censors do put pressure on Facebook, the company must abide by their demands when legally required. Therefore, Facebook is portraying itself as just as much a victim of government censorship as the individuals whose content gets legally removed.

The updated standards make clear that they only "outline Facebook's expectations when it comes to what content is or is not acceptable in our community," while "countries have local laws that prohibit some forms of content."[108] On its face, this statement conveys a rather obvious fact. However, it is important to point out the distinction made in this statement between "our community" and "countries." In making

[106] *Id*.
[107] *Id*.
[108] *Id*.

this distinction, Facebook is placing itself in an advantageous position when it comes to supporting freedom of expression online. On the one hand, Facebook is claiming to set itself apart from the world of "flesh and steel" that philosopher John Perry Barlow vilified in 1996.[109] However, it is also acceding to reluctant participation in the legal regimes that inevitably do have control over online activity and speech, as legal scholars David Johnson and David Post averred.[110]

Facebook's position is that complying with government demands is better for speech in the long run. Mark Zuckerberg wrote in a March 15, 2015 post on Facebook's company blog, "If we ignored a lawful government order and then we were blocked, … people's voices would be muted, and whatever content the government believed was illegal would be blocked anyway."[111] Freedom of expression should not be considered a black-and-white issue, Zuckerberg argued; rather, "giving people a voice … is something that we must make incremental progress towards."[112]

Unsurprisingly, this philosophy reflects the evolution of Facebook's rules governing UGC. It is a constant work in progress. It is an experiment that combines elements of First Amendment theory and jurisprudence (e.g. for threats and incitement), media ethics, corporate social responsibility, and harnessing user agency to both encourage good speech and discourage the bad. It is dependent on users' intent and the context behind the speech. It will probably never be perfect, but that is ideal for

---

[109] John Perry Barlow, *A Declaration of the Independence of Cyberspace* (1996), *available at* https://projects.eff.org/~barlow/Declaration-Final.html.
[110] Johnson & Post, *supra* note 47.
[111] Mark Zuckerberg, *Today we released our latest Global Government Requests Report and updated our Community Standards*, FACEBOOK (Mar. 15, 2015), *available at* https://www.facebook.com/zuck/posts/10101974380267911?reply_comment_id=10101974442079041&total_comments=36&__fns&hash=Ac2U13_rwzPmNUEQ.
[112] *Id*.

Facebook. It can keep adjusting its social norms in an *ad hoc* process, always claiming to be serving the goals of both promoting speech and preventing harm.

Examples of Facebook's Controversial Content Governance

The March 15, 2015 update to Facebook's Community Standards may have added clarification on where the line is drawn when it comes to certain categories of problematic UGC, including some rather bold assertions regarding photos of breastfeeding and artistic nudes.[113] However, this latest update is far from black-and-white, and questions still remain regarding the boundaries of certain speech. Almost certainly, Facebook will continue to remove UGC. To better understand the issues involved in these potential removals, this chapter will examine some controversial examples of Facebook removing UGC prior to the March 2015 Community Standards update. These examples serve as a historical guide to Facebook's battle with updating its speech rules. They also highlight that Facebook's governance of UGC is lacking in transparency. Each example will be discussed in turn below. After all of the facts have been recounted, they will be analyzed and discussed in conjunction with what has been discussed so far about Facebook's evolving community standards.

Examples

• In February 2011, artists from Colorado and the New York Academy of Art who posted their nude or seminude artwork to Facebook later found that they had been removed.[114] New York Academy of Art had their account blocked for a week after

[113] *Supra* note 104.
[114] Miguel Helft, *Art School Runs Afoul of Facebook's Nudity Police*, N.Y. TIMES: BITS BLOG (Feb. 18. 2011), *available at* http://bits.blogs.nytimes.com/2011/02/18/art-school-runs-afoul-of-facebooks-nudity-police/.

several students' work had been taken down.[115] Although Facebook later apologized and reposted the works, the Academy took to its blog to criticize Facebook for being "the final arbiter—and online curator—of the artwork we share with the world."[116] Artist Richard T. Scott expressed why the removals were so concerning. "For figurative painters, Facebook has been a democratizing force, and it has been pivotal for my career" for getting works recognized by galleries, he said.[117]

• In July 2012, the social network removed photos of the German painter Gerhard Richter's painting called "Ema," featuring a blurry nude female, from the page of the Paris-based Pompidou Center.[118] Officials for the museum complained to Facebook, which restored the photos not long after.[119] A French art blog, "Les Notes de Véculture," called the incident "institutional puritanism."[120] The *Washington Post* reported on its blog "The Intersect" in March 2015 that a French schoolteacher was seeking to sue Facebook in France because the social network kept thwarting his attempts to post photos of Gustave Courbet's painting "L'Origine du Monde," which features an up-close depiction of female genitalia.[121]

---

[115] *Id*.
[116] *Id*.
[117] *Id*.
[118] Juliette Soulez, *Facebook Censors Pompidou's Gerhard Richter Nude, Fueling Fight Over "Institutional Puritanism,"* ARTINFO FRANCE (July 31, 2012), *available at* http://www.blouinartinfo.com/news/story/816583/facebook-censors-pompidous-gerhard-richter-nude-fueling-fight#.
[119] *Id*.
[120] *Id*.
[121] Caitlin Dewey, *Facebook Censored a Nude Painting, and it Could Change the Site Forever*, WASH. POST: THE INTERSECT BLOG (Mar. 9, 2015), *available at* http://www.washingtonpost.com/news/the-intersect/wp/2015/03/09/facebook-censored-a-nude-painting-and-it-could-change-the-site-forever/.

• Before Facebook's most recent update to its community guidelines, women had reported that they had photos depicting breastfeeding removed from Facebook.[122] In several cases, Facebook apologized to the women and reposted the photos.[123] When the *Huffington Post* contacted Facebook about the removals, the social network sent the following statement:

> We agree that breastfeeding is natural and we are very glad to know that it is important for mothers, including the many mothers who work at Facebook, to share their experience with others on the site. The vast majority of breastfeeding photos are compliant with our Statement of Rights and Responsibilities and Facebook takes no action on such content. However, photos which contain a fully exposed breast, do violate our terms and may be removed if they are reported to us. These policies are based on the same standards that apply to television and print media. It is important to note that photos upon which we act are almost exclusively brought to our attention by other users who complain about them being shared on Facebook.[124]

• In May 2014 (once again, less than a year before Facebook revised its community standards), a North Carolina woman who posted a photo to Facebook of her bare chest after undergoing a double mastectomy had those photos removed for violating Facebook's standards against nudity and pornography.[125] The woman said she posted the photo as a way to offer support to other women with breast cancer.[126] Facebook also removed photos posted by fashion photographer David Jay that were part of a project Jay

---

[122] *Kristy Kemp, Breastfeeding Advocate, Outraged When Nursing Photos Were Removed from Facebook*, HUFFINGTON POST (Apr. 5, 2013), *available at* http://www.huffingtonpost.com/2013/04/05/kristy-kemp-breastfeeding-photos_n_3021288.html.
[123] *Id*.
[124] *Id*.
[125] Jessica Firger, *Breast Cancer Survivor Battles Facebook over Mastectomy Photos*, CBS NEWS (May 9, 2014), *available at* http://www.cbsnews.com/news/breast-cancer-survivor-battles-facebook-over-mastectomy-photos/.
[126] *Id*.

had created showing topless women who had undergone double mastectomies.[127] Other women reported having photos of double mastectomies removed from Facebook, including a woman who posted a photo of her entire torso being tattooed following the surgery.[128] As noted above, Facebook responded by changing its policy to allow double-mastectomy photos as long as nipples could not be seen.[129]

• In September 2014, Facebook removed a photo that a North Carolina man had posted of his two-month-old son lying in a hospital bed and hooked up to machines.[130] The man reportedly had posted the photo to raise awareness of his son's heart condition and raise funds for the boy's impending surgery.[131] The man received a notice from Facebook saying that "scary, gory or sensational pictures" such as those of "vampires, zombies and dismembered bodies" are not allowed on the social network.[132] Facebook later apologized for the removal and the message, restored the photo, and donated $10,000 worth of advertisements to help the family raise funds, according to the report.[133]

• Some Native Americans have had difficulty getting Facebook to accept their legal names as "real" names under Facebook's policy of users not going by nicknames or aliases.[134] According to a March 2015 report from the BBC's "Trending" technology news blog, a South Dakota man of the Oglala Lakota tribe named Lance Browneyes had

---

[127] *Id*.
[128] Sara Gates, *Facebook Removes Photo of Breast Cancer Survivor's Tattoo, Users Fight Back*, HUFFINGTON POST (Feb. 20, 2013), *available at* http://www.huffingtonpost.com/2013/02/20/facebook-breast-cancer-tattoo-photo-double-mastectomy_n_2726118.html.
[129] Firger, *supra* note 125.
[130] *Facebook Rejects Photo of Baby Boy in Hospital, Calls It Too Graphic*, FOX 13 (Sept. 11, 2014), *available at* http://www.myfoxchicago.com/story/26505303/facebook-rejects-photo-of-baby-boy-in-hospital-calls-it-too-graphic.
[131] *Id*.
[132] *Id*.
[133] *Id*.
[134] Micah Luxen, *Facebook Challenges Legitimacy of Some Native Names*, BBC TRENDING (Mar. 3, 2015), *available at* http://www.bbc.com/news/blogs-trending-31699618.

his name changed to Lance Brown on Facebook.[135] Others such as Mike Raccoon Eyes Kinney and Lone Hill experienced similar problems.[136] Similarly, numerous drag queens reportedly have not been allowed to create profiles using their stage names (which they may argue are the names associated with their true identity), although they have been allowed to create business-oriented Pages using their stage names.[137]

• In November 2010, a Facebook page titled "Let's show these poppy burning bastards how many people want them deported," which called for the deportation of Muslims from the United Kingdom after Muslim protestors disrupted Remembrance Day events in London, was removed shortly after it was created.[138] Facebook refused to comment on whether it had removed the page, or whether the page's creator had deleted it, though a spokesperson for the social network released a statement saying that the company takes its community guidelines seriously and "react[s] quickly to reports of inappropriate content."[139] The page has since been restored, though the current version of the page (as of this writing) carries a disclaimer saying that it does not advocate for the deportation of all Muslims from the United Kingdom, only the ones who protested on Remembrance Day.[140]

---

[135] *Id*.
[136] *Id*.
[137] Lil Miss Hot Mess, *Say My Name: Facebook's Unfair "Real Names" Policy Continues to Harm Vulnerable Users*, SALON (Mar. 30, 2015), *available at* http://www.salon.com/2015/03/30/say_my_name_facebooks_unfair_real_names_policy_continues_to_harm_vulnerable_users/.
[138] Caitlin Fitzsimmons, *Anti-Muslim Facebook Page Now Removed*, ADWEEK: SOCIAL TIMES (Nov. 12, 2010), *available at* http://www.adweek.com/socialtimes/poppy-page-removed/326272.
[139] *Id*.
[140] *Deport the Muslims Who Ruined 2010 Remembrance Day*, FACEBOOK, *available at* https://www.facebook.com/pages/Deport-the-muslims-who-ruined-2010-remembrance-day/144232922290674.

• [Restated from Chapter 4]: In late April 2013, two videos depicting the beheading of three individuals, purportedly in Mexico, appeared on Facebook.[141] The social networking site initially refused to remove the videos in spite of formal requests made by individual members and humanitarian organizations. Facebook said of one of the videos:

> People are sharing this video on Facebook to condemn it. Just as TV news programs often show upsetting images of atrocities, people can share upsetting videos on Facebook to raise awareness of actions or causes. While this video is shocking, our approach is designed to preserve people's rights to describe, depict and comment on the world in which we live.[142]

However, after pressure from members and interest groups increased, Facebook decided to remove the videos, saying it would "evaluate [its] policy and approach to this type of content."[143] At the time, Facebook's "Community Standards" page stated, "We understand that graphic imagery is a regular component of current events, but must balance the needs of a diverse community. Sharing any graphic content for sadistic pleasure is prohibited,"[144] (see Appendix 2: Facebook's Community Standards). Facebook issued a statement in May 2013 saying that the videos did not meet its standards for graphic or gratuitous violence.[145] In mid-October 2013, it allowed the videos to be viewed on its site, again saying that people should be able to watch the

---

[141] Leo Kelion, *Facebook U-turn after Charities Criticizes Decapitation Videos*, BBC NEWS: TECHNOLOGY (May 1, 2013), available at http://www.bbc.co.uk/news/technology-22368287.
[142] *Id*.
[143] *Id*.
[144] *Facebook Community Standards*, INTERNET ARCHIVE (Dec. 15, 2012).
[145] Kelion, *supra* note 235.

videos to condemn them, and adding that it was considering a policy of including a warning alongside the link to the video.[146]

• [Restated from Chapter 4]: In early 2013, feminist organizations decried the existence of pages created on the social networking giant that glorified or made light of rape and domestic violence. Activist Soraya Chemaly, Jaclyn Friedman of the group Women, Action and the Media (WAM), and Laura Bates of the Everyday Sexism Project, published an open letter online on May 21, 2013, demanding that Facebook not tolerate "speech that trivializes or glorifies violence against girls and women."[147] The open letter stated that pages had titles such as "Fly Kicking Sluts in the Uterus" and "Violently Raping Your Friend Just for Laughs," and images appeared on the network "of women beaten, bruised, tied up, drugged, and bleeding, with captions such as 'This bitch didn't know when to shut up' and 'Next time don't get pregnant.'" The women asked Facebook users to contact companies whose ads appeared on pages with such speech. In April 2013, Bates took a screenshot of a page titled "Drop kicking sluts in the teeth" and tweeted it to the beauty company Dove, whose ad appeared next to the page. In late May, the activists persuaded advertisers such as Nissan UK, *Jump* magazine, and Desire Books and 15 other companies to pull ads from the social network.

Marne Levine, Facebook's vice-president for global public policy at that time, responded to the activists' demands in a May 28, 2013, blog post on the social network, promising that Facebook officials would "update the training for the teams that review

---

[146] Leo Kelion, *Facebook lets beheading clips return to social network*, BBC NEWS (Oct. 21, 2013), *available at* http://www.bbc.co.uk/news/technology-24608499.
[147] *Open Letter to Facebook*, WOMEN, ACTION, & THE MEDIA (May 21, 2013), *available at* http://www.womenactionmedia.org/facebookaction/open-letter-to-facebook/.

and evaluate reports of hateful speech or harmful content on Facebook."[148] Levine wrote that Facebook would push for more accountability from "the creators of content that does not qualify as actionable hate speech but is cruel or insensitive by insisting that the authors stand behind the content they create."[149] For example, this requirement would mean that "the creator of any content containing cruel and insensitive humor include his or her authentic identity for the content to remain on Facebook."[150] Levine did not promise that Facebook would remove any of the pages.[151]

Synthesis

These examples do not constitute an exhaustive list of Facebook's controversial governance of users' speech. Nevertheless, these examples point to Facebook's struggle with being both a platform that promotes free speech and facilitates social change, and that is a "safe" place for its hundreds of millions of users. These examples show that this struggle crosses categories of speech. Even a concept seemingly as benign as using a stage name or an abnormal-sounding name gets caught up (alongside beheading videos and nude sketches) in Facebook's quest to serve its community at the perfect intersection of freedom and safety. The problem here is that when users publish content to Facebook, only to have it removed, Facebook—despite its apologies and promises to do better—sends a message to these users: you are on the fringe of our community, and you may not even be welcome at all. As will be discussed below, the extent to which this message is a

[148] *Controversial, Harmful and Hateful Speech on Facebook*, FACEBOOK SAFETY (May 28, 2013), *available at* https://www.facebook.com/notes/facebook-safety/controversial-harmful-and-hateful-speech-on-facebook/574430655911054.
[149] *Id*.
[150] *Id*.
[151] *Id*.

problem depends on the extent to which Facebook is considered an important—if not essential—community for the world to be a part of.

Each of these types of speech at issue in the examples above is distinct. Nudity and glorifying terrorism (to pick just two examples) relay two distinct messages. In the legal world, each type of speech has its own set of parameters that define when state actors can proscribe it. These parameters do not exist in the world of content governance. Facebook can decide for itself what constitutes incitement to violence and what does not, without any obligation to define its criteria. The same goes for nudity: a fully nude sketch by an artist trying to make it big may be allowed while a seminude masterpiece may not be, or vice-versa; a titillating photo of a woman in a revealing swimsuit may be allowed, but a photo of a woman going topless to make a political statement may not be. In either case, Facebook is under no legal obligation to inform users about the criteria used to distinguish allowable from unacceptable (if specific criteria even exist). What nudity and glorifying terrorism *do* have in common is their potential to convey a message of social or political importance. This potential disappears when digital intermediaries remove images without due regard for that social or political importance. Facebook appears to be developing a system of giving speech its due regard by focusing on the intent of users who post controversial UGC. However, the examples recounted here—even though they occurred before Facebook's March 15, 2015 update added some clarity to its community standards—show that important social or political motives behind the speech did not always save it from being removed from Facebook. Indeed, some of the most important

social and political motives were found in some of the most offensive and contentious of speech: the beheading videos.

Two very powerful counterpoints can be made here that threaten to undermine the arguments made in this chapter, not to mention this entire dissertation. First, if speech gets removed from Facebook, Twitter or YouTube, what is to stop the speaker from publishing the speech elsewhere? For example, the website 4chan is notorious for allowing all sorts of UGC, especially UGC that would violate the community standards of Facebook, or another mainstream intermediary such as Twitter or YouTube.[152] Would it not be preferable for some intermediaries to govern their UGC more strictly while others govern it more loosely? Indeed, how a site governs its UGC may be a characteristic that endears it to certain users while repelling others. Like Pepsi versus Coca-Cola, consumers will choose intermediaries based on taste. Those who like gore and smut can go to 4chan. Those who like a safer environment can go to Facebook.

The response to this argument is two-fold. First, there is a clear distinction in scale between a mainstream intermediary such as Facebook (as of this writing the second most popular website on the World Wide Web, according to web analytics company Alexa)[153] and 4chan (the 707th most popular website on the Web, as of this writing).[154] Under the notion of the "long tail" model of audience fragmentation on the Internet—whereby audience "attention is clustered around a select few content options, followed by

---

[152] *See, e.g.*, *4chan Murder Pictures: David Kalac Arrested in Oregon*, BBC NEWS (Nov. 6, 2014), *available at* http://www.bbc.com/news/world-us-canada-29932087; Jon Kelly and Jude Sheerin, *The Strange Virtual World of 4chan*, BBC NEWS MAGAZINE (Aug. 31, 2010), *available at* http://www.bbc.com/news/magazine-10520487.
[153] *The Top 500 Sites on the Web*, *supra* note 6.
[154] *Site Overview: 4chan.org*, ALEXA (Mar. 26, 2015), *available at* http://www.alexa.com/siteinfo/4chan.org.

a long tail, in which the remaining multitude of content options each attract very small audiences that in the aggregate can exceed the audience for the 'hits'"[155]—this difference in scale is problematic for the public discourse. Depending on how far down the long tail 4chan (or other such "fringe" intermediaries) resides, a potentially logarithmically larger audience would be shielded from extreme speech through Facebook's content governance than the audience who sees it on 4chan. Second, this difference in scale creates gated communities in online space,[156] creating an online version of the conflict between exercising speech rights in public forums and exercising them in hybrid private-public places such as shopping malls. One of the criticisms some legal scholars have made against shopping malls is that they have replaced the free speech free-for-all of public squares with privately run zones that wall out controversial and extreme speech.[157] According to this argument, as more and more people tend to frequent shopping malls over traditional public forums, the capacity for a speaker to reach a large audience with his or her message is diminished. Similarly, in the world where a distinction between Facebook and 4chan is encouraged, Facebook becomes the sterile mall and 4chan becomes the unfamiliar public square, used mainly by the fringe of society. As important public discourse on matters of social and political significance has ventured onto digital intermediaries,[158] the diversity and robustness of that discourse depends on extreme voices being heard on mainstream intermediaries and not being banished to fringe sites.

---

[155] PHIL M. NAPOLI, AUDIENCE EVOLUTION: NEW TECHNOLOGIES AND THE TRANSFORMATION OF MEDIA AUDIENCES 5 (2010).
[156] Jonathan Zittrain, *A History of Online Gatekeeping*, 19 HARV. J. L. & TECH. 253 (2006).
[157] *See, e.g.*, CASS SUNSTEIN, DEMOCRACY AND THE PROBLEM OF FREE SPEECH 36 (1993).
[158] DENARDIS, *supra* note 4.

The second powerful counterargument to the concern over content governance stems from this last point: how much of a social and political difference does speech on digital intermediaries really make in the world? Indeed, it is difficult to measure the impact that speech published on digital intermediaries—particularly mainstream ones—has on a concept such as the robustness of public discourse. It is equally as difficult to measure the effect that the variable of robustness of public discourse has on outcomes such as political awareness or willingness to engage in a certain political behavior. In other words, one may be hard-pressed to argue that a robust public discourse even exists online. Therefore, why should digital intermediaries have any obligation to foster one, or even attempt to foster one? To better respond to this argument, this chapter must connect the issues of content governance to those of network (or net) neutrality.

Connection to Net Neutrality Debate

The examples of content governance recounted above are anecdotes. They are not necessarily indicative of a trend among digital intermediaries rampantly monitoring and removing UGC. Rather they are illustrative of the potential that intermediaries have to scrub undesirable content from their platforms. Although hard data indicating a clear trend (were they to exist) would be the ideal evidence for highlighting the potential threats content governance poses to online public discourse, citing anecdotal evidence to raise awareness of threats to free speech has relevant precedent. Proponents of net neutrality, including the FCC under the Obama administration, have used similar anecdotal evidence to successfully build their case for net neutrality.

Network Management and the Net Neutrality Debate

At its core, the issue over net neutrality has to do with whether Internet service providers (ISPs) should be allowed to control and differentiate the speed at which various Web applications are delivered to individual users.[159] Proponents of net neutrality argue that ISPs must not be allowed to engage in such "network management" practices because ISPs could use them to give priority service to certain applications (such as those who could afford to pay for faster service, or the ISPs' own applications) over others (competitors, those unable to afford faster service, or, potentially, those with political messages that the ISPs disfavor).[160]

In 2008, the FCC ordered broadband provider Comcast to cease engaging in network management practices that deliberately and discriminately slowed down the service provided by the peer-to-peer file-sharing site known as BitTorrent.[161] A majority of the Commission saw Comcast's practice as anti-competitive because it degraded the quality of service of Internet applications that rely on BitTorrent and that pose "a particular competitive threat to Comcast's video-on-demand ('VOD') service."[162] The majority held that the fact that Comcast conducted this discriminatory practice secretly only "compounded the harm."[163] Comcast's actions, the majority argued, highlighted that "the risk to the open nature of the Internet is particularly acute and the danger of network management practices being used to further anticompetitive ends is strong."[164]

---

[159] Timothy Wu, *Network Neutrality, Broadband Discrimination*, 2 J. TELECOMM. & HIGH TECH. L. 141 (2003).
[160] Scott Jordan & Arijit Ghosh, *A Framework for Classification of Traffic Management Practices as Reasonable or Unreasonable*, 10 TRANSACTIONS ON INTERNET TECH. 1 (2010).
[161] *In re* Comcast Corp. for Secretly Degrading Peer-to-Peer Applications, 23 FCC Rcd. 13028 (Aug. 20, 2008).
[162] *Id*. at 3.
[163] *Id*. at 1.
[164] *Id*. at 31.

Meanwhile, Comcast argued that its management practices were necessary to ensure that bandwidth-hogging applications using BitTorrent (particularly the illicit ones that allowed illegal distribution of copyrighted works) did not interfere with customers' quality of service.[165]

The U.S. Court of Appeals for the D.C. Circuit held in *Comcast Corp. v. FCC*[166] that the FCC lacked authority under the Telecommunications Act of 1996 to order Comcast to cease its network management practices. This ruling prompted the FCC to establish the Open Internet Order of 2010,[167] in which the Commission claimed it did, in fact, have the necessary statutory authority to establish a regime of net neutrality that would prohibit the types of discriminatory network management practices addressed in *In re* Comcast. However, in *Verizon v. FCC*,[168] decided in January 2014, a panel of the U.S. Court of Appeals for the D.C. Circuit struck down the FCC's Open Internet Order establishing net neutrality. Although the FCC lost the case, the defeat was not due to its use of anecdotal evidence, but rather to the improper way that the Commission used the 1996 Telecommunications Act to claim authority over regulating the Internet as if it were a common-carrier technology, like the telephone.[169] In fact, two judges on the three-judge panel held that "nothing in the record gives us any reason to doubt the Commission's determination that broadband providers may be motivated to discriminate against and among" Internet application providers.[170] In March 2015, the FCC officially reclassified

[165] *Id*. at 66.
[166] 600 F.3d 642 (D.C. Cir. 2010).
[167] *In re* Preserving the Open Internet, 25 FCC Rcd. 17905 (Dec. 23, 2010).
[168] 740 F.3d 623 (D.C. Cir. 2014).
[169] *Id*. at 628.
[170] *Id*. at 645.

the Internet as a public utility, thereby allowing it to be regulated as a common carrier. In the *In re* Protecting and Promoting the Open Internet Order,[171] the divided Commission justified its reclassification by yet again averring that Internet service providers have both the "incentive and ability" to discriminate traffic on its networks.[172]

Content Governance and Network Management: Similarities and Differences

Put simply, the private governance that was the impetus for net neutrality to be enforced and the private governance of content identified in this chapter are not equivalent. This chapter is not arguing that a net neutrality regime should be set up for digital intermediaries. ISPs and the digital intermediaries that are the focus of this study are both conduits for the speech of others, but they are different kinds of conduits. ISPs, per the language of the FCC from its March 2015 Order, do not engage in expressive activities.[173] Facebook, YouTube and Twitter *do* engage in expressive activities while simultaneously facilitating the activities of third parties. For example, they brand themselves, or they alter the design and graphics of their platforms. Because they engage in such expressive activities, digital intermediaries have a far stronger claim that their own First Amendment rights would be violated under a regulatory regime similar to net neutrality than ISPs do in their current predicament. Hence the position of this chapter.

However, when viewed from the perspective of affirmative First Amendment theory, the difference between ISPs' and platforms' *function* as conduits does not appear quite so stark, even if their legal definition as conduits remains so. In its March 2015

---

[171] *In re* Protecting and Promoting the Open Internet, FCC, GN Docket No. 14-28 (March 12, 2015), *available at* http://transition.fcc.gov/Daily_Releases/Daily_Business/2015/db0312/FCC-15-24A1.pdf.
[172] *Id*. at 28.
[173] *Id*. at 269.

order, the three-commissioner majority of the FCC held, "When engaged in broadband Internet access services, broadband providers are not speakers, but rather serve as conduits for the speech of others. The manner in which broadband providers operate their networks does not rise to the level of speech protected by the First Amendment."[174] The majority then cites *Turner Broadcast Systems v. FCC*[175] to contend that its net neutrality policy "serve[s] First Amendment interests of the highest order, promoting 'the widest possible dissemination of information from diverse and antagonistic sources' and 'assuring that the public has access to a multiplicity of information sources' by preserving an open Internet."[176] It later states that "the rules we adopt today ensure that the Internet promotes speech by ensuring a level playing field for a wide variety of speakers who might otherwise be disadvantaged."[177] In essence, the Commission is making the argument that one technology should be regulated so that a First Amendment *interest* (ensuring a robust public discourse) can flourish through the use of that technology.

Using such an argument from affirmative First Amendment theory is easy for the FCC to do here, as ISPs are now considered common carriers that do not have expressive capabilities of their own. As stated above, it would be impossible to make the same argument in favor of a regulatory regime against digital intermediaries that curtail the speech of others. Yet that has never been the goal of this study. Rather, this study argues that digital intermediaries have the potential to threaten a robust online public discourse,

[174] *Id*. at 268.
[175] 512 U.S. 622 (1994).
[176] *In re* Protecting and Promoting the Open Internet, *supra* note 171, at 268.
[177] *Id*. at 272.

in the same way that ISPs have the potential to threaten that discourse, according to the current FCC and other proponents of net neutrality. And, in the same way that the current FCC and proponents of net neutrality have done, this study argues that the principle of "incentive and ability," based on theory and anecdotal evidence, is sufficient for showing that content governance exists and is a problem that individuals should be aware of.

Thus, returning once again to the counterargument above on whether intermediaries should brand themselves based on the type of speech they allow, the issue is not the fear that the content that mainstream intermediaries remove will disappear completely from the Internet. Proponents of net neutrality would be hard pressed to argue that an edge provider that could not afford to pay for fast-lane service or that got deliberately blocked or had its service degraded by an ISP would disappear completely from the Internet. Their primary concern is that these sites would risk being left out of the mainstream Web, relegated to the fringes because no one dared wait the extra seconds for the site to load. The same concern is at the heart of content governance. Extreme speech is not at risk of disappearing from the entire Internet, just the mainstream Internet. Scholars and citizens alike need to decide whether that is a problem, and, if it is, how it should be handled.

Discussion

Facebook's strategy essentially boils down to a choice between two ideologies of governing speech. The first—and this is the approach that Facebook had chosen prior to the update of its community guidelines—is a utilitarian approach to speech. The argument behind this approach goes like this: By following a policy of allowing more

speech that comes close to violating its community guidelines than allowing less such speech, Facebook runs the risk of alienating users who are offended by that speech. These users may then choose to leave Facebook. If a critical mass of users leaves Facebook, the company will subsequently lose advertising revenue, and may eventually be forced out of business. Thus, although some individuals may become upset because their content was removed from Facebook (whether that content in fact violated Facebook's community guidelines or not), those removals are defensible out of a desire to preserve the social network's broader function of increasing individuals' communicative agency in the global public discourse.

This utilitarian approach has a ring of affirmative First Amendment theory to it. Mark Zuckerberg said in a March 15, 2015 post on Facebook's company blog that "threats of violence and bullying will be taken down" because they "are examples where one person exercising their [*sic*] voice may unfairly limit the voices of many others. Therefore, in the spirit of giving the most voice to the most people, we choose not to permit this content."[178] Of course, Zuckerberg's position specifically refers to Facebook's mission of encouraging more speech through preventing the types of abuse and harassment of users that would scare them away from using the social network as a platform for speaking. This mission is the same as the argument put forth in chapter 4 that abuse of individuals on digital intermediaries is a form of heckler's veto designed to suppress speech. The utilitarian approach to governing content shares the goal of promoting more speech at the expense of removing potentially offensive speech. However, Facebook's utilitarian approach promotes more speech out of a concern for

[178] Zuckerberg, *supra* note 111.

preserving the viability of the platform. This rationale lacks the central concern for individual dignity that is at the core of the argument from promoting speech through preventing abuse. This concept will be discussed in chapter 6.

The second ideological choice for governing content is tolerance. Tolerance is not synonymous with absolutism. Rather, according to Bollinger, tolerance involves an acknowledgement that the "real threat to liberty of speech … rests within the general population of citizens instead of officialdom alone."[179] The goal of tolerance is not for extreme speech to be *accepted* in society, but rather that society simply *allows* extreme speech into the public discourse. If a digital intermediary like Facebook is not going to allow certain extreme viewpoints into the public discourse it purports to host, then it is incumbent upon Facebook to explain its decision.

Therefore, tolerance requires two things of Facebook. First, it requires a firm commitment to protecting freedom of expression. This commitment goes beyond good public relations and Community Standards that say Facebook encourages free speech until that speech violates vague rules. Such a commitment means Facebook must state that promoting freedom of expression is its primary duty, and all harms that it seeks to avoid from extreme speech will be judged by specific, unambiguous standards. Second, it requires transparency. Such a commitment means Facebook must go beyond caveats and disclaimers and pleas for patience and forgiveness if speech is wrongly removed. For all the greater detail paid to clarifying what constitutes hate speech or acceptable nudity, the updated community guidelines still lack transparency. The March 2015 Community

[179] LEE C. BOLLINGER, THE TOLERANT SOCIETY: FREEDOM OF SPEECH AND EXTREMIST SPEECH IN AMERICA 80 (1986).

Standards talk of "dedicated teams working around the world to review things," and of some of these people being "the right person for review[ing]" certain categories of content.[180] Yet users are in the dark about what exactly happens between when content is flagged and when a decision is made on whether or not the content should be removed. Who are the people who review the flagged content? How are they trained? How do some people become experts in one category over another? Is the fate of content in the hands of one person, several or many?[181] In its "Facebook Principles," Facebook lists "Transparent Process" as principle 9 out of 10, averring that "Facebook should publicly make available information about its purpose, plans, policies, and operations."[182] Governance of user-generated content seems like an excellent place to make good on that principle.[183]

Conclusion

Facebook faces an unending struggle to find the balance between creating a communicative service where people feel "safe"[184] and promoting an arena of public

[180] *Facebook Community Standards*, FACEBOOK (Mar. 15, 2015).
[181] Journalist Adrian Chen has reported on the work of laborers responsible for reviewing flagged content for several unnamed digital intermediaries. Many of these laborers are based in the Philippines, due to a widespread knowledge of both English and U.S. cultural norms among Filipinos. The vast majority of their decisions involve more than just whether a photo of a breastfeeding mom shows too much nipple or whether a racist diatribe has a salvageable political message. Rather, these workers spend their days judging hardcore pornography, beheadings or other images of gore. The sheer volume and nature of flagged content may in fact necessitate that digital intermediaries employ hoards of day laborers that can stomach (even if barely) the job of reviewing that content. Adrian Chen, *The Laborers Who Keep Dick Pics and Beheadings Out of Your Facebook Feed*, WIRED (Oct. 21, 2014), *available at* http://www.wired.com/2014/10/content-moderation/.
[182] *Facebook Principles*, Facebook (Aug. 25, 2010), *available at* https://www.facebook.com/principles.php.
[183] To its credit, in early 2015 Facebook began releasing a so-called "Transparency Report," which consists of data on requests by governments around the world for information on certain users and for allegedly illegal UGC to be removed. *See Global Requests Report*, FACEBOOK (2015), *available at* https://govtrequests.facebook.com/. However, this report does not contain information on users flagging other users over extreme content, and it does not contain information on the process of how Facebook makes decisions on content governance.
[184] *Facebook Community Standards*, FACEBOOK (Mar. 15, 2015).

discourse where people "make the world more open and connected" as they "share their stories, see the world through the eyes of others and connect with friends and causes."[185] The line separating these two goals is constantly shifting. No matter how specific Facebook gets in trying to define that line, advocates for either goal will continue to put pressure on either side of the line.

The purpose of this analysis is not to argue that Facebook should become more protective of freedom of expression than it currently is. Nor, for that matter, does this study contend that Facebook should follow the standards of First Amendment jurisprudence in crafting its community guidelines. Rather, the ultimate conclusion reached from this analysis is that what Facebook's system of content governance lacks in absolute protection of speech should be made up for in transparency. The entire system of how Facebook governs UGC should be made clear to users.

The true significance of this study lies in its potential broader implications for the norms of public discourse on digital platforms. The questions that arise from this analysis are: Does Facebook (or other digital intermediaries) offer a "parallel universe" to public discourse conducted in public forums governed by First Amendment jurisprudence? Is Facebook becoming the shopping mall to the public square, complete with a public discourse that has become vapid? Implied in this question is an assumption that the social and political significance of public discourse in each of these spheres is measurable, and that it is something that can wax and wane, either independently or in relation to one another. Traditionally, scholarship in mass communication law has maintained that greater protection of freedom of expression is ultimately the best policy for building a

[185] *Id*.

robust democratic society.[186] That argument is fundamentally normative in nature; it is not based on empirical evidence, nor is it built to ever be open to falsifiability through empirical testing.[187] Nevertheless, this analysis should encourage researchers from the fields of mass communication and mass communication law to assess the relationships between speech, Facebook, social norms and democracy.

This analysis has an even broader implication: it can shed light on the potential extent to which tolerance for extreme speech is changing in a networked communication environment. Or, framed as a research question: Do the norms of expression on Facebook influence individuals' tolerance toward extreme speech in other arenas, such as public forums or other media? If so, how? And to what extent? The concluding chapter to this dissertation will discuss some of the possibilities for future research on this topic.

[186] *See* Matthew D. Bunker and David K. Perry, *Standing at the Crossroads: Social Science, Human Agency and Free Speech Law*, 9 COMM. L. & POL'Y 1 (2004).
[187] *Id*.

Chapter 6: A Duty to Freedom: Conceptualizing Platform Ethics

Introduction

The purpose of this study is to explicate the concept of content governance: the control that digital communication intermediaries exercise over user-generated content (UGC). The particular focus of this explication is the governance of extreme UGC. Two key questions guide this explication: How and why do digital communication intermediaries respond to extreme UGC? What are the potential implications of their responses for public discourse in a system of networked communication?

This chapter analyzes and discusses the ethical obligations digital intermediaries have to monitor, remove or leave be potentially harmful content created by users. This angle of analysis is important because it frames the activities of digital intermediaries to govern UGC in terms of duties to competing interests, namely: promoting and protecting freedom of speech, enhancing individual agency, preventing harm, obeying laws, and making money.[1] All of these duties cannot be served equally, and thus the purpose of this chapter is to examine the nature of these duties so that they can be measured against one another, and so that a normative ethical theory can be created that orders these duties in terms of their importance.

[1] These duties are loosely borrowed from Christians and Merrill's 2009 collection of short essays analyzing major ethical theorists, which are organized around five ethical "loyalties": to others, to self, to freedom, to authority, and to community. The duties listed in the paragraph above do not necessarily correspond exactly to Christians and Merrill's loyalties. However, these loyalties are useful for conceiving of the multiple duties that intermediaries must juggle; *see* Josh Braun and Tarleton Gillespie, *Hosting the Public Discourse, Hosting the Public*, 5 JOURNALISM PRACTICE 383 (2011).

Context

Following attacks in early January 2015 by Islamist extremists that ended in the deaths of nine staff members at the Parisian satirical weekly newspaper *Charlie Hebdo*, as well as the deaths of four hostages at a kosher supermarket, three police officers and one bystander, European leaders released a statement on their strategy for fighting terrorism. One of the measures that the statement called for was for digital intermediaries to play a larger role in this fight. The relevant part of the statement read:

> We are concerned at the increasingly frequent use of the Internet to fuel hatred and violence and signal our determination to ensure that the Internet is not abused to this end, while safeguarding that it remains, in scrupulous observance of fundamental freedoms, a forum for free expression, in full respect of the law. With this in mind, the partnership of the major Internet providers is essential to create the conditions of a swift reporting of material that aims to incite hatred and terror and the condition of its removing, where appropriate/possible.[2]

Such a call from public officials is not new. In a May 19, 2008 letter to Google CEO Eric Schmidt, Senator Joe Lieberman of Connecticut demanded that Google remove all "terrorist training" videos from YouTube. Propounding the belief that "Islamist terrorist organizations rely extensively on the Internet to attract supporters and advance their cause," Lieberman argued that "[b]y taking action to curtail the use of YouTube to disseminate the goals and methods of those who wish to kill innocent civilians, Google will make a singularly important contribution to this important national effort."[3] In September 2008, Google removed some (though not all, or even most) of the videos that

---

[2] *Joint Statement* (Jan. 11, 2015), *available at* https://www.bmi.bund.de/SharedDocs/Downloads/DE/Kurzmeldungen/gemeinsame-erklaerung.pdf?__blob=publicationFile

[3] *Lieberman Calls on Google to Take Down Terrorist Content*, U.S. SEN. COMMITTEE ON HOMELAND SEC. AND GOVT'L AFF. (May 19, 2008), *available at* http://www.hsgac.senate.gov/media/majority-media/lieberman-calls-on-google-to-take-down-terrorist-content.

concerned Senator Lieberman. In a statement, a YouTube spokesperson said that because such videos incited violence, they violated YouTube's terms of use. However, the YouTube official stated, "While we respect and understand his [Lieberman's] views, YouTube encourages free speech and defends everyone's right to express unpopular points of view."[4]

These episodes pose a provocative question: Do digital communication intermediaries have an ethical duty to stanch the flow of content published by extremists or terrorists in the name of preventing the spread of their ideologies and the perpetration of harms caused in their name? Do digital intermediaries have a duty to prevent the spread of *any* type of harmful speech that travels through their platforms? These questions beg an even broader question: To whom or to what—aside from themselves—do digital intermediaries have an ethical duty? This is an essential question in the context of content governance. The role that digital intermediaries play in both fostering and governing speech published on their platforms has received attention from legal scholars,[5] critical-cultural communication scholars,[6] and scholars in the field of Internet governance.[7] Although some ethicists have called for intermediaries to become regular

---

[4] Peter Whoriskey, *YouTube Bans Videos That Incite Violence*, WASH. POST (Sept. 12, 2008), *available at* http://www.washingtonpost.com/wp-dyn/content/article/2008/09/11/AR2008091103447.html.
[5] *See, e.g.*, Sandra Braman & Stephanie Roberts, *Advantage ISP: Terms of Service as Media Law*, 5 NEW MEDIA & SOC'Y 422 (2003); Rebecca Tushnet, *Power Without Responsibility: Intermediaries and the First Amendment*, 76 G.W. L. REV. 986 (2008); DAWN C. NUNZIATO, VIRTUAL FREEDOM: NET NEUTRALITY AND FREE SPEECH IN THE INTERNET AGE 36 (2009); Jack M. Balkin, *Old-School/New-School Speech Regulation*, 127 HARV. L. REV. 2296 (2014).
[6] José van Dijck, *Users like you? Theorizing agency in user-generated content*, 31 MED. CULT. & SOC'Y 41 (2009); Ganaele Langlois, *Participatory Culture and the New Governance of Communication: The Paradox of Participatory Media*, 14 TELEVISION & NEW MEDIA 91 (2013).
[7] LAURA DENARDIS, THE GLOBAL WAR FOR INTERNET GOVERNANCE (2014); JACK GOLDSMITH AND TIMOTHY WU, WHO CONTROLS THE INTERNET? ILLUSIONS OF A BORDERLESS WORLD (2006); REBECCA MACKINNON, CONSENT OF THE NETWORKED: THE WORLDWIDE STRUGGLE FOR INTERNET FREEDOM (2012).

subjects of ethical analysis,[8] so far this topic has not received attention from scholars of media ethics. This trend needs to change.

Argument and Roadmap

The purpose of this chapter is to construct an ethical framework that mainstream digital intermediaries should follow when faced with balancing the values of protecting individuals' expression and preventing harm. This chapter argues that digital intermediaries should follow a hierarchy of duties when confronted with such speech. Because digital intermediaries play an essential role in allowing individuals to engage in a diverse, global public discourse on matters of social and political importance,[9] their first duty should be to promoting freedom of expression and enhancing individual agency and autonomy. The duty of digital intermediaries to prevent harm should be subordinate to their duty to promote freedom of expression. This does not mean that digital intermediaries should not take harms seriously. Rather, they should follow a hierarchy when identifying and remedying harms, with abuse and harassment of specific individuals deemed the most harmful and deserving of policing by digital intermediaries, followed by direct and credible threats, incitement to reasonably foreseeable lawless action, and images of death and gore. The purpose of following such a hierarchy is to ensure that speech of social and political importance is tolerated in spite of the fact that

[8] Clifford G. Christians, *Media Ethics on a Higher Order of Magnitude*, 23 J. OF MASS MEDIA ETHICS 3 (2008); Jay Black, *An Informal Agenda for Media Ethicists*, 23 J. OF MASS MEDIA ETHICS 28 (2008); Bernd C. Stahl, *IT for a Better Future: How to Integrate Ethics, Politics and Innovation*, 9 J. OF INFO., COMM. & ETHICS IN SOC'Y, 140 (2011).
[9] YOCHAI BENKLER, THE WEALTH OF NETWORKS: HOW SOCIAL PRODUCTION TRANSFORMS MARKETS AND FREEDOM (2006); DENARDIS, *supra* note 7.

some may find the speech harmful or offensive, while speech of no political or social importance that concentrates its harm on specific individuals is eradicated.

The chapter will begin with a discussion of the differences and similarities between news organizations and digital intermediaries as institutions subject to ethical analysis. The fundamental similarity between these two sets of institutions—their essential function for deliberative democracy—opens up digital intermediaries to ethical perspectives and theories traditionally used in analyses of news organizations. However, the obvious difference between the institutions—the fact that news organizations predominantly[10] publish their own content while intermediaries facilitate third-party content—necessitates that the ethicality of digital intermediaries must be judged within its own unique set of principles. These ethical principles can be found in the philosophies on which laws on intermediary liability for third-party content are based. Therefore, the second section of the chapter will review scholarship and jurisprudence—both American and foreign—on the liability of digital intermediaries. In particular, this section will argue that conceptions of intermediary liability in the United States, the European Union, Brazil and India underscore a duty that mainstream digital intermediaries have to promoting individual freedom of expression and serving the democratic public discourse. These laws were chosen because of their salience: they represent the rules of the game for digital intermediaries seeking to serve billions of citizens of democratic polities. The final section of this chapter will apply this ethical framework to Facebook's March 15, 2015

[10] Of course, news organizations do have to manage comments published on their websites by readers (*see* Braun & Gillespie, *supra* note 1), and they also collect UGC from other platforms to publish on their own sites (*see* Jane B. Singer, *User-Generated Visibility: Secondary Gatekeeping in a Shared Media Space*, 16 NEW MEDIA & SOC'Y 55 (2014)).

update of its community standards (discussed in chapter 5). The purpose of this comparison is to identify the common thing missing from these codes in their current state: a clear and unambiguous commitment to freedom of expression for their users.

Platform Ethics Belongs in Media Ethics

In several respects, the concept of "platform ethics" proposed in this chapter does not fit neatly into the field of media ethics. A comparison of digital intermediaries with the institutions that are traditionally the focus of media ethicists (news organizations) reveals conceptual incongruities between the two. However, despite these incongruities, both sets of institutions share the essential function of facilitating democratic public discourse, thereby opening up digital intermediaries to analysis using theories and perspectives from the field of media ethics. To truly appreciate this core similarity and thereby accept platform ethics as a part of the broader field of media ethics, one must first understand the main differences between the two sets of institutions.

Differences

First, the two institutions compared in this analysis differ in terms of their uniformity. The press can be conceived as a monolithic social institution, albeit made up of many diverse actors.[11] Members of the press include public affairs reporters and celebrity gossip columnists, citizen journalists and foreign bureau chiefs, insightful op-ed writers and ranting bloggers, anyone with a smartphone and the entire network of CNN. Although each has his or her own perspective on the news of the day and style for how to deliver it, all pursue the same goal: collecting and relaying information of some modicum

[11] *See generally*, GENEVA OVERHOLSER AND KATHLEEN HALL JAMIESON, THE PRESS (2005); KATHLEEN HALL JAMIESON AND PAUL WALDMAN, THE PRESS EFFECT: POLITICIANS, JOURNALISTS, AND THE STORIES THAT SHAPE THE POLITICAL WORLD (2003).

of public concern to a particular audience.[12] Ethical debates revolve around the "best" way for these actors to fulfill this goal, focusing predominantly on how to achieve the ideals of reporting truth and minimizing harm.[13] Digital intermediaries, on the other hand, serve multiple functions in their capacity to facilitate individuals' speech. They can facilitate the work of citizen journalists (e.g. YouTube hosting videos of a protest recorded on smartphones), or they can be a location for subjects of journalistic intrigue (e.g. a politician's newsworthy Facebook message or an athlete's controversial tweet). They can propel messages of revolution and hope just as easily as they can help the spread of messages of harassment and hate. An ethical framework for digital intermediaries must be broad enough to encompass all of these functions, yet nuanced enough to respect their unique differences.

Second, on a related note, news organizations are made up of readily identifiable agents with potential ethical responsibility: reporters, editors, and perhaps executives or publishers.[14] Meanwhile, digital intermediaries are not made up of such readily identifiable agents due to the simple fact that intermediaries are not responsible for their own content. Coders, executives, and staff responsible for managing user complaints about other users' content are on the payrolls of digital intermediaries. However, someone sitting at a computer in San Francisco or the Philippines making snap decisions over whether UGC violates a platform's community guidelines does not have the direct,

---

[12] Seth C. Lewis, *The Tension Between Professional Control and Open Participation: Journalism and Its Boundaries*, 15 INFO., COMM. & SOC'Y 836, 851 (2012); Christopher Meyers, et al., *Professionalism, Not Professionals*, 27 J. OF MASS MEDIA ETHICS 189, 195 (2012).
[13] Clifford G. Christians and Kaarle Nordenstreng, *Social Responsibility Worldwide*, 19 J. OF MASS MEDIA ETHICS 3 (2004).
[14] Ian Richards, *Stakeholders Versus Shareholders: Journalism, Business, and Ethics*, 19 J. OF MASS MEDIA ETHICS 119 (2004).

authorial control over content that a reporter, editor or publisher has. Nor, for that matter, does any executive of a digital intermediary. The responsibility for reviewing and removing harmful UGC is shared among multiple actors within the organization, each with different roles to play in the process. Ideally, each of these actors would be the subject of its own unique set of ethical precepts. However, at the present time, all of these actors fall under the umbrella of the corporate identities of these intermediaries. Therefore, referring to digital intermediaries as monolithic wholes is the best way to analyze the concept of ethical governance of UGC. This proposition is hardly unorthodox or revolutionary; borrowing from literature on corporate social responsibility, some media ethicists have argued that media corporations "have the competency to develop values that are not just aggregations of the values of individuals in the organization"[15]

Third, although news organizations serve a group that is rather nebulous on its face (the "public," or the news organization's "audience"), in terms of media ethics this relationship is generally conceived as clear-cut and binary.[16] In this binary model, the public benefits from "good" or accurate reporting by news organizations, and it loses out from "bad" or inaccurate reporting. The model works one way: the news organization produces the message, and the public is affected by it. Platforms, on the other hand, serve as a conduit between speakers and audiences. Sometimes, the interests of these speakers and audiences are in sync, while other times they conflict with one another.[17] For example, if an intermediary removes a speaker's message from its platform, it is harming

[15] G. Stuart Adam, Stephanie Craft and Elliot D. Cohen, *Three Essays on Journalism and Virtue*, 19 J. OF MASS MEDIA ETHICS 247, 261 (2004).
[16] Meyers, et al, *supra* note 12.
[17] Tarleton Gillespie, *The Politics of Platforms*, 12 NEW MEDIA & SOC'Y 347 (2010).

the interest of that speaker. Whether the removed message harms the interests of the audience (or multiple audiences) depends on the nature of both the message and the audience(s). For instance, the removal of a racist post on Facebook may help the interests of an audience that despises racism, but it may—according to some First Amendment theorists[18]—harm the interests of society at large by not allowing exposure to a variety of messages, no matter how hateful.

Fourth, on a related note, news organizations (at least private, non-state organizations) generally have an adversarial relationship with governments. Their ideal job is to act like a "watchdog," constantly monitoring and checking government to ensure the public that it is not engaging in corrupt activities.[19] Digital intermediaries, on the other hand, can be adversarial toward governments, but only in their capacity to facilitate the speech of individuals that is critical of government actors. Other times, digital intermediaries can be coopted into aiding governments' agendas of security and public order. For example, the leak of classified documents by NSA contractor Edward Snowden exposed the NSA's practice of submitting National Security Letters (NSLs) to companies such as Google and Facebook, ordering that they secretly divulge information on users.[20] Such a practice resembles a "guard dog" function, whereby media are employed as protectors of prevailing power structures.[21] In the context of media ethics, such a practice reveals the conflict of duties that digital intermediaries serve. As Braun

[18] LEE C. BOLLINGER, THE TOLERANT SOCIETY: FREEDOM OF SPEECH AND EXTREMIST SPEECH IN AMERICA (1986); Robert C. Post, *Racist Speech, Democracy, and the First Amendment*, 32 WM. & MARY L. REV. 267 (1991); RODNEY A. SMOLLA, FREE SPEECH IN AN OPEN SOCIETY (1992).
[19] Vincent Blasi, *The Checking Value in First Amendment Theory*, 2 AM. BAR FOUND. RES. J. 521 (1977); George A. Donohue, Phillip J. Tichenor and Clarice N. Olien, *A Guard Dog Perspective on the Role of Media*, 45 J. OF COMM. 115 (1995).
[20] Balkin, *supra* note 5.
[21] Donohue, Tichenor and Olien, *supra* note 19, at 121; Adam, et al, *supra* note 15, at 273.

and Gillespie aptly put it, digital intermediaries "must toe the line between avoiding legal liability, keeping an eye on the economic bottom line, and some kind of commitment to protecting their users' freedom of speech and the vibrancy of the public discourse they produce."[22] This chapter takes the position that these platforms must make the commitment to freedom of speech and the public discourse their priority over preventing harm as they seek to find the ideal balance between these two goals. Indeed, as will be shown below in the analysis on similarities between digital intermediaries and news organizations, protecting freedom of speech and maintaining a vibrant public discourse is the *raison d'etre* of digital intermediaries.

Finally, the activity of these institutions that is subject to ethical scrutiny takes place within two distinct legal frameworks. The legal framework for U.S. news organizations is the First Amendment. Within this context, a relatively large gap exists between what news organizations are allowed to publish (legally speaking) and what they should or should not publish (ethically speaking).[23] The extent and nature of this gap is the focus of much (if not most) of the scholarship within the field of media ethics pertaining to U.S. media. Digital intermediaries, on the other hand, operate within a legal framework that sets parameters on their liability for the third-party speech that they host. In the United States, that framework takes the shape of Section 230 of the 1996 Communications Decency Act,[24] which, as will be discussed below, gives extensive immunity to digital intermediaries from civil liability for third-party content. Meanwhile,

[22] Braun & Gillespie, *supra* note 1, at 385.
[23] Black, *supra* note 8; Michael Perkins, *International Law and the Search for Universal Principles in Journalism Ethics*, 17 J. OF MASS MEDIA ETHICS 193 (2002).
[24] 47 U.S.C. § 230 (1996).

as will also be discussed below, other countries' legal systems afford less protection. However, even within these regimes of less protection for intermediaries, one can identify legal philosophies that hold that individuals must take pride of place in an intermediary-driven system of networked communication. These legal philosophies are an essential source for the construction of platform ethics.

Similarities

As stated above, the main similarity between news organizations and digital intermediaries is that each institution, in its own way, facilitates an essential function for democratic deliberation. This institutional similarity is crucial for the analysis of this chapter. Some scholars contend that media ethics should focus on the ethical practices of *media* institutions, broadly conceived, rather than merely *journalistic* institutions.[25] This perspective borrows from the work of political economy theorists[26] to make the argument that greater concentration of power in the ownership of mass media plays a major role in the ethicality of the activities of these media.[27] Of primary concern for this perspective is that corporate institutions, whose ultimate duty is to provide healthy returns to shareholders, "are short on the virtues of citizenship" necessary to understand media's democratic function.[28]

---

[25] Nick Couldry, *Why Media Ethics Still Matters*, *In* GLOBAL MEDIA ETHICS: PROBLEMS AND PERSPECTIVES (Stephen J. A. Ward ed., 2013) 13-29.
[26] *See* ROBERT M. ENTMAN, DEMOCRACY WITHOUT CITIZENS: MEDIA AND THE DECAY OF AMERICAN POLITICS (1989); BEN BAGDIKIAN, THE NEW MEDIA MONOPOLY (2004); Daniel C. Hallin*, Hegemony: The American News Media from Vietnam to El Salvador, A Study of Ideological Change and its Limits*, *in* POLITICAL COMMUNICATION: APPROACHES, STUDIES, ASSESSMENTS*,* (David L. Paletz ed., 1986); ROBERT MCCHESNEY, DIGITAL DISCONNECT: HOW CAPITALISM IS TURNING THE INTERNET AGAINST DEMOCRACY (2013).
[27] Adam, et al, *supra* note 15, at 255.
[28] *Id*. at 256.

Along those lines, some scholars borrow from the literature of business ethics to make the argument that media corporations not only can, but rather *should* be analyzed as moral agents. Ethicist Stephanie Craft, for example, contends that in conducting their general activities, businesses of any stripe "depend upon … a basic sense of community."[29] She writes, "Corporations … are part and parcel of the communities that created them, and the responsibilities that they bear are not the products of argument or implicit contracts but intrinsic to their very existence as social entities."[30] Having established that corporations in general can be conceived of as moral agents, Craft next argues that *media* corporations must necessarily be analyzed as moral agents because they "straddle two realms, business and public service, in ways that other corporations do not."[31] In fact, the Press Clause of the First Amendment, she contends, suggests that the Framers "thought of the press as an entity whose purpose was not solely or even predominantly profit generation, but public service."[32]

In other words, Craft is arguing that the essential function of media corporations to democracy naturally predisposes them to adhere to a duty to the communities they serve. It is probably naïve to think that these corporations actually consider this duty of primary importance, putting it above their duty to their shareholders. Nevertheless, at the very least, this duty has prudential value behind it. Economist Lawrence Souder argues that "[c]orporate-owned media … not only *should* be founded on sound media ethics;

[29] *Id*. at 259.
[30] *Id*. at 260.
[31] *Id*. at 262.
[32] *Id*. at 265.

they *must* be for their own good."[33] The following question remains: to whom should digital intermediaries keep an ethical duty that is at least co-extensive with their duty to their own bottom line? To speakers? To audiences that potentially could be harmed by speakers? To both? And if both, when would one duty trump the other? To answer these questions, this chapter turns to an examination of the issue of intermediary liability.

Ethical Principles and Intermediary Liability

With the potential for digital intermediaries to facilitate the spread of harmful messages of several different categories—including defamation, invasions of privacy, hate speech, threats, incitement to violence, personal harassment or abuse, and (though outside the analysis of this chapter) copyright infringement—debate has swirled globally around the extent to which intermediaries should be held liable for such facilitation. This section will show that the debate between those calling for greater protection for intermediaries and those calling for greater liability has been framed along the lines of concerns for freedom of expression, consumer protection and economic regulation. This section will give a brief overview and comparative analysis of the legal and philosophical foundations to the United States' approach to intermediary liability and a common global approach to intermediary liability, exemplified by legal regimes in the European Union, Brazil and India. The purpose of this analysis is to extract a framework for platform ethics from the legal context of intermediary liability. In particular, the analysis in this section will show that both American and international intermediary liability regimes conceive of digital intermediaries as having a duty to serve their users and create a

[33] Lawrence Souder, *A Free-market Model for Media Ethics: Adam Smith's Looking Glass*, 25 J. OF MASS MEDIA ETHICS 53, 54 (2010) (emphasis added).

vibrant public discourse (despite the fact that they have competing interpretations of how to uphold this duty). From this analysis, the argument will be made that platforms' primary duty should be to protect the ability of users to speak freely and preserve this vibrant public discourse.

U.S. Perspective

The U.S. approach to intermediary liability is enshrined in Section 230 of the 1996 Communications Decency Act (CDA).[34] The law grants digital intermediaries immunity from civil liability for content published by third parties on its platforms, even when they are notified of the presence of the content or when they choose to take control over the content and remove it in a "Good Samaritan" act.[35] Intermediaries are only obligated to remove content—under threat of criminal liability—that infringes on an author's copyright or that violates criminal law (e.g., images of child abuse).[36] Congress declared in the preamble to Section 230 that digital intermediaries deserve "a minimum of government regulation" because they "offer users a great degree of control over the information that they receive, as well as the potential for even greater control in the future as technology develops."[37] It called the Internet a "forum for a true diversity of political discourse,"[38] facilitated by digital intermediaries. It declared that its intent in passing the law was "to preserve the vibrant and competitive free market that presently exists for the

[34] 47 U.S.C. § 230 (1996).
[35] 47 U.S.C. § 230(c)(1).
[36] 47 U.S.C. § 230(e)(1).
[37] 47 U.S.C. § 230(a)(2).
[38] 47 U.S.C. § 230(a)(3).

Internet and other interactive computer services, unfettered by Federal or State regulation."[39]

U.S. case law involving Section 230 extolls the benefits of the provision to the public discourse in spite of its side effect of allowing potentially harmful speech to flourish. In *Zeran v. AOL*, the United States Court of Appeals for the Ninth Circuit stated that Section 230, quite simply, was a "policy choice … not to deter harmful speech through the separate route of imposing tort liability on companies that serve as intermediaries," thereby "maintain[ing] the robust nature of Internet communication."[40] In *DiMeo v. Max*, the U.S. District Court for the Eastern District of Pennsylvania averred that "we should expect such [harmful] speech to occur in a medium in which citizens from all walks of life have a voice."[41] Such strong statutory protection for speech parallels—if not exceeds—the exceptional constitutional protection of freedom of expression found in the United States.[42]

Non-U.S. Perspectives

In contrast to Section 230, intermediary liability regimes in the European Union, Brazil and India—three powerful liberal-democratic geopolitical entities that together account for more than one billion Internet users[43]—impose less immunity to digital

[39] 47 U.S.C. § 230(b)(2).
[40] Zeran v. America Online, Inc., 129 F.3d 327, 330 (4th Cir. 1997).
[41] DiMeo v. Max, 433 F. Supp. 2d 523, 533 (E.D. Pa. 2006).
[42] 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 (2010).
[43] *Internet Users per 100 People*, THE WORLD BANK (2015), *available at* http://data.worldbank.org/indicator/IT.NET.USER.P2. (The World Bank includes data on global Internet usage by country by percentage of total population. Therefore, Brazil's 51.6% of the population using the Internet equates to roughly 104 million people, India's 15% of the population using the Internet equates to roughly 190 million people, and the United States' 84.2% of the population using the Internet equates to roughly 267 million people); *Statistics*, INT'L TELECOMM. UNION, *available at* http://www.itu.int/en/ITU-

intermediaries in certain contexts. Namely, a "social theory of responsibility" in which "control capability [over content] implies co-responsibility" (morally, if not legally, speaking) distinguishes these regions from the United States.[44]

European Model

The regime of intermediary liability in the European Union is based on Directive 2000/31/EC (the so-called "e-Commerce Directive").[45] Recital 46 of the Directive states that digital intermediaries—here referred to as providers of an "information society service"—benefit from a limitation of liability if "upon obtaining actual knowledge or awareness of illegal activities [they] act expeditiously to remove or to disable access to the information concerned."[46] Recital 48 states that EU Member States may "apply duties of care" on digital intermediaries, "which can reasonably be expected from them and which are specified by national law, in order to detect and prevent certain types of illegal activities."[47] Various European courts have broadly interpreted "illegal activities" under this Directive to include content that causes economic harms (copyright or trademark infringement) as well as the types of content that are the focus of this chapter: those that cause relational or emotional harms,[48] such as defamation, violation of privacy and hate

---

D/Statistics/Pages/stat/default.aspx. (This site includes data on global Internet usage by region, estimating more than 460 million users in the European Union in 2014). Combined, the estimated number of Internet users from these two sources for these four polities comes to 1.021 billion.

[44] Tomas A. Lipinski, Elizabeth A. Buchanan, & Johannes J. Britz, *Sticks and Stones and Words that Harm: Liability vs. Responsibility, Section 230 and Defamatory Speech in Cyberspace*, 4 ETHICS & INFO. TECH. 143, 156 (2002).

[45] Directive 2000/31/EC.

[46] Directive 2000/31/EC (46).

[47] Directive 2000/31/EC (48).

[48] Smolla, *supra* note 18.

speech.[49] Thus, a duty of care is established when the alleged victim of such harms appropriately notifies the digital intermediaries of the presence of the content in question on their platforms, thereby putting these companies on the legal hook for removing it. Meanwhile, Article 15(1) of the Directive prohibits Member States from "impos[ing] a general obligation on providers … to monitor the information which they transmit or store, nor a general obligation actively to seek facts or circumstances indicating illegal activity."[50] Therefore, European officials can only—as they did following the *Charlie Hebdo* attacks—call on digital intermediaries to follow a *moral* obligation to police harmful speech published on their platforms.

The goal of the European model is to incentivize self-regulation by digital intermediaries by encouraging a proactive approach, whereby companies would actively screen user-generated content and remove the manifestly unlawful material before upset users have a chance to notify them and thus place them within the prospects of liability.[51] "Only when contents would be manifestly unlawful—so that intermediaries would not have to appreciate their lawfulness—would the latter be required to react and eventually take them down or restrict access to them."[52] As with Section 230, the philosophy behind the European approach to intermediary liability is that "private regulation is less

[49] Timothy Pinto, et al., *Liability of Online Publishers for User Generated Content: A European Perspective*, 27 COMM. LAWYER 5 (2010).
[50] Directive 2000/31/EC (Art. 15).
[51] Sophie Stalla-Bourdillon, *Sometimes One Is Not Enough! Securing Freedom of Expression, Encouraging Private Regulation, or Subsidizing Internet Intermediaries or All Three at the Same Time: The Dilemma of Internet Intermediaries' Liability*, 7 J. OF INT'L COMMERCIAL L. & TECH. 154, 164 (2012).
[52] *Id*. at 162.

dangerous than public regulation when it comes to the defence [*sic*] of freedom of expression."[53]

Brazilian Model

The Brazilian approach to intermediary liability is codified in the 2014 law known as the "Marco Civil da Internet."[54] Roughly translated as an "Internet Bill of Rights," the statute establishes, among other things, that Brazilian citizens have a right to net neutrality and to the responsible collection, use and storage of their personal data by all Internet application providers. Articles 18 and 19 of the Marco Civil provide that digital intermediaries have immunity from civil liability for third-party content unless they fail to remove defamatory or racist content after receiving a valid court order asking them to do so.[55] The Marco Civil also lists a special provision under Article 21 for the phenomenon of "revenge porn," whereby intermediaries will be held criminally liable if they either purposefully host or fail to remove revenge porn photos on their platforms.

---

[53] *Id*. at 164.
[54] Lei N$^{o.}$ 12.965 (April 23, 2014).
[55] One particular example of such a judicial move is illustrative. In September 2012, two videos appeared on YouTube alleging that Alcides Bernal, mayoral candidate for the city of Campo Grande in the Brazilian state of Mato Grosso do Sul, hated poor people, had unlawfully enriched himself, and had paid an ex-lover to abort a child he fathered, denied being the child's father after he was born, and then beat the child after finally admitting that he was the father. The people who posted the videos remain anonymous. Bernal filed a lawsuit against Google Brazil—owner and operator of YouTube in Brazil—for publishing defamatory electoral propaganda against him in the run-up to an election, a violation of Article 243 of the Brazilian Electoral Code. Wanting to uphold the rules for conducting free and fair elections as painstakingly defined in the Electoral Code, Judge Flávio Saad Perón of the 35th Electoral Zone of the municipality of Campo Grande—a division of the Regional Electoral Court ("Tribunal Regional Eleitoral" or TRE) of the state of Mato Grosso do Sul—ordered that the video be taken down. The head of Google Brazil, Fabio José Silva Coelho, refused to obey the order, citing a commitment to upholding the values of free speech. Judge Saad then ordered Coelho placed under house arrest for disobeying a judge's order—a violation of Article 347 of the Electoral Code—and ordered a 24-hour suspension of *all* Google and YouTube services in the state. The judge's order attracted national and international media attention on the otherwise ordinary and relatively insignificant election. Google Brazil released a statement saying it was "appealing the decision that ordered the removal of the YouTube video because, in being a platform, Google is not responsible for the content posted on its site." The company did not comment on Coelho's arrest. However, on September 26, 2012, Google removed the videos from YouTube.

The philosophical foundation to this approach to intermediary liability is based on the Brazilian Consumer Protection Code (CPC) of 1990,[56] which lays out numerous rights of consumer protection. The philosophical thrust of the CPC is that consumers deserve protection from businesses because ultimately consumers are the reason businesses are in business to begin with; in other words, consumers deserve a substantial amount of legal power over the businesses that are profiting off of them.[57] In the context of intermediary liability, the theory is that because online communication platforms profit off of users by commodifying both their content and their data, these platforms should ultimately respond to users when this venture turns harmful.[58]

Indian Model

Section 79 of India's Information Technology (IT) Act of 2000 (updated in 2008)[59] stipulates that intermediaries will not be held liable for third-party content except when it either materially contributes to the creation of the content, or if it receives actual knowledge that the content is unlawful. In 2011, the Indian government published the "Information Technology (Intermediary guidelines) Rules"[60] in its official gazette to further define what might make third-party content unlawful. This includes content that "is grossly harmful, harassing, blasphemous defamatory, obscene, pornographic, paedophilic [*sic*], libelous, invasive of another's privacy, hateful, or racially, ethnically

[56] Lei N$^{o.}$ 8.078 (Sept. 11, 1990).
[57] Guilherme M. Martins and João Vitor R. Longhi, *Internet Service Providers' Liability for Personal Damages on Social Networking Websites*, 11 U.S.-CHINA L. REV. 286, 308 (2014).
[58] *Id*.
[59] Act No. 21 of 2000 (Information Technology Act) § 79.
[60] *Notification*, THE GAZETTE OF INDIA: EXTRAORDINARY, Part II § 3(i) (April 11, 2011).

objectionable, disparaging, relating or encouraging money laundering or gambling, or otherwise unlawful in any manner whatever."[61]

According to the 2011 guidelines, intermediaries must follow "due diligence" to remove unlawful material.[62] The doctrine of due diligence comes from the realm of Indian business law and has many meanings, none of which has been clearly defined by a court or codified by a statute.[63] However, for the purposes of understanding the Indian philosophy behind intermediary liability, the doctrine of due diligence essentially means that once a company becomes aware that it is profiting off of the unlawful practices of a business partner, it must see to it that those unlawful business activities end.[64] To illustrate this principle, in 2008, the Delhi High Court *in dicta* condemned a website's placing the maximization of profits over "[s]afeguard[ing] … prevailing moral values" in regard to its business model of profiting off of spreading links to obscene material.[65] Thus, the philosophy behind the Indian model of intermediary liability bares striking resemblance to that of the Brazilian model, whereby dutiful treatment of consumers determines the ethical and legal standards of business operations.

Synthesis: Intermediary Liability and Platform Ethics

The easy conclusion to draw from a comparison of Section 230 with the European, Brazilian and Indian approaches to intermediary liability is that Section provides much greater protection of speech than the other three.[66] This easy distinction

---

[61] *Notification* § 3(2)(b).
[62] *Notification* § 3.
[63] Little & Co., *Due Diligence*, INT'L FINANCIAL L. REV., *available at*: <http://www.iflr.com/Article/2027418/Legal-due-diligence.html>.
[64] *See, e.g.*, James Grandolfo, *India*, 44 THE INT'L LAWYER 663 (2010).
[65] Avnish Bajaj v. State, 150 DLT 279 (May 2008) (India).
[66] Pinto, et al, *supra* note 49.

comes from the fact that the foreign approaches involve notice-based liability, whereas notice does not trigger liability under Section 230.[67] The more difficult conclusions are the normative ones: Should speech be protected at the expense of the harms it can cause from a lack of an incentive on the part of intermediaries to remove it? If so, does Section 230 strike the best balance between promoting free speech and reducing the potential harms caused by it? Several U.S. scholars have answered these questions "yes" and "no," respectively, calling for Section 230 to be amended to give victims of harmful speech the ability to seek damages from sites that intentionally traffic in such speech, particularly (though not limited to) sites that host revenge porn.[68] On the other hand, legal scholars Sandra Braman and Stephanie Roberts argue that digital intermediaries, in hypocritical and self-contradictory fashion, "do not want to be content providers but do want to control all content," effectively giving them "*control without liability*" under Section 230.[69] Similarly, legal scholars Rebecca Tushnet[70] and Dawn Nunziato[71] argue that Section 230 gives digital intermediaries *too much* of an incentive to control speech. The argument goes that if intermediaries are able, under Section 230, to proactively remove objectionable content without fear of liability, they will do so, to the detriment of

[67] Of course, under Section 230, notice-based liability does apply when interactive computer services host third-party images of child abuse or alleged infringements of copyrighted works (47 U.S.C. § 230(e)(1)).
[68] *See, e.g.*, Mary Anne Franks, *Unwilling Avatars: Idealism and Discrimination in Cyberspace*, 20 COLUM. J. GENDER & L. 224 (2011); AMY GAJDA, THE FIRST AMENDMENT BUBBLE: HOW PRIVACY AND PAPARAZZI THREATEN A FREE PRESS (2015); DANIELLE KEATS CITRON, HATE CRIMES IN CYBERSPACE (2015); Joshua N. Azriel, *Social Networking as a Communications Weapon to Harm Victims: Facebook, MySpace, and Twitter Demonstrate a Need to Amend Section 230 of the Communications Decency Act*, 26 JOHN MARSHALL J. OF COMP. & INFO. L. 415 (2009).
[69] Braman and Roberts, *supra* note 5, at 438 (emphasis added).
[70] Tushnet, *supra* note 5.
[71] NUNZIATO, *supra* note 5.

individuals' ability to speak freely on these platforms.[72] Tushnet accuses this legal regime of simultaneously supporting freedom and suppression,[73] and posits that "if we limit intermediary responsibility, … we should also limit intermediary power to control speech."[74]

In other words, digital intermediaries face a "tri-lemma" of sorts under the Section 230 regime: they can protect speech and be accused of not doing enough to prevent harm; they can remove harmful content at the request of individuals and be accused of not doing enough to protect speech, as well as spend significant resources to identify and remove the speech; or they can proactively remove objectionable speech and be accused of "private censorship."[75] To adequately balance these competing interests, intermediaries should follow a sound ethical theory. The European, Brazilian and Indian concepts of intermediary liability, when imputed into the U.S. philosophy on the subject, can be interpreted in a way that creates such an ethical theory.

At their core, the European, Brazilian and Indian regimes conceive of individual citizens as the most important stakeholders in a networked economy that thrives on facilitating public discourse. This core principle is the "thin" normative framework upon which these regimes then build up a "thick"[76] regime of intermediary liability, prescribing that citizens have rights against intermediaries for protecting them from the potentially harmful speech these platforms could host. The thin normative framework

---

[72] *Id*. at 2-3.
[73] Tushnet, *supra* note 5, at 1011.
[74] *Id*. at 1009.
[75] *Id*.; NUNZIATO, *supra* note 5.
[76] Edward H. Spence and Aaron Quinn, *Information Ethics as a Guide for New Media*, 23 J. OF MASS MEDIA ETHICS, 264 (2008).

upon which Section 230 is founded, as stated above, involves preserving the Internet as a "forum for a true diversity of political discourse."[77] Combining these thin normative frameworks, one arrives at the following conclusion: intermediaries have a primary duty to serve individual citizens to the extent that they use their platforms to create a vibrant public discourse. Therefore, digital intermediaries have a primary duty to preserve, promote and protect freedom of expression within the limits allowed them by law.

The notion of maintaining a primary ethical duty to freedom of expression has its critics. In his analysis of the ethical reasons for the magazine *Soldier of Fortune* to not publish classified ads seeking murders-for-hire, ethicist Scott Tomlinson entertains the position that the magazine may have a duty to publish the unquestionably harmful ads based on a duty to protect the broader value of promoting freedom of expression.[78] This duty follows the logic that only exercising the right to publish extreme speech will ensure that that right stays protected. However, Tomlinson concludes that refraining from publishing speech out of a moral concern for preventing harm to the targets of the speech "would not seriously impair free expression … because one instance of self-restraint would be extremely small and insignificant when compared to expression in general."[79] Tomlinson also argues that such moral self-restraint may actually *fulfill* a duty to freedom "because harming others and acting socially irresponsible could prove to involve a high degree of harm to individuals in the short term and to the First Amendment in the long

[77] 47 U.S.C. § 230(a)(3).
[78] Don E. Tomlinson, *Where Morality and Law Diverge: Ethical Alternatives in the* Soldier of Fortune *Cases*, 6 J. OF MASS MEDIA ETHICS 69 (1991).
[79] *Id*. at 77.

term."[80] Law professor Amy Gajda expands on this argument, positing that the U.S. legal culture affords too much protection for harmful speech, thereby inflating the value of such speech.[81] She argues that just like any economic bubble, this "First Amendment bubble" will eventually collapse, to the detriment of those who would wish to express unpopular, offensive or extreme yet less harmful ideas.[82]

This line of arguments is similar to contentions by critical legal theorists such as Charles Lawrence, who argues that the U.S. legal regime of exceptional protection of freedom of expression ultimately hurts the First Amendment value of creating an environment where many diverse ideas can flourish.[83] This value is undermined, according to Lawrence, because the prevalence of hateful and harmful speech will deter participation in public discourse by the targets of that speech, who naturally feel threatened and unwelcome.[84] Undoubtedly, speech can cause great harm. Traditional First Amendment theorists do not deny this fact, but they counter critical theorists such as Lawrence by arguing that any public benefit of hateful speech outweighs the potential harms it may cause. Ultimately, both groups lack the empirical evidence to back up their respective positions, thereby leaving to ethicists the question of how to "appropriately" wield the awesome powers of freedom of speech. Platform ethics changes the debate by focusing on how speakers should be *allowed* to wield their speech. The answer, this

[80] *Id*.
[81] GAJDA, *supra* note 68.
[82] *Id*.
[83] Charles R. Lawrence, *If He Hollers Let Him Go: Regulating Racist Speech on Campus*, 1990 DUKE L. J. 431 (1990).
[84] *Id*.

chapter proposes, is that digital intermediaries should have a primary duty to protect individuals' ability to speak.

This primary duty does not negate a duty to prevent harm caused by speech published on intermediaries' platforms. Rather, the duty to promote individuals' speech should be seen as co-extensive with the prevention of harm. By only removing content that is personally harassing, threatening or abusive to private individuals (but not to public figures) when notified of the presence of such content by users, digital intermediaries are promoting rather than stifling a robust public discourse. Figure 6-1 organizes several categories of extreme and potentially harmful speech in a hierarchy determined by the political or social significance of the speech, and the potential of the speech to chill the speech of other users. Abuse and harassment of private individuals have the greatest potential to keep others (i.e. the targets of the speech) from speaking, due to the likelihood they will incur more abuse and harassment by speaking up.[85] This type of speech also has little to no potential for political or social significance, due to the sheer fact that it involves a private individual. Therefore, digital intermediaries should move swiftly to remove such speech when users notify them of it. Meanwhile, the categories of extreme and potentially harmful speech at the top of the hierarchy have a greater tendency to implicate matters of public concern, due to the fact that they tend to deal with issues of political and social significance and involve entire groups of people rather than individuals. Concomitantly, they have a lesser potential for chilling the speech

[85] KEATS CITRON, *supra* note 68; see the section titled "Abuse, Trolling and Gamergate" in Chapter 4.

of individuals, due to the fact that identifiable individuals are not the direct targets of such speech.

The objective of devising this hierarchy is simply to give digital intermediaries such as Facebook a set of guidelines for governing speech. There certainly may be instances where images of graphic content or incitement speech should be removed, but intermediaries should follow strict and transparent protocol for governing such speech. Meanwhile, cases of individual abuse and harassment should not be tolerated. Cases of speech involving nudity have deliberately not been included in this hierarchy, due to the uniquely wide range of messages that can be associated with nudity (from sexual pandering to political statements).

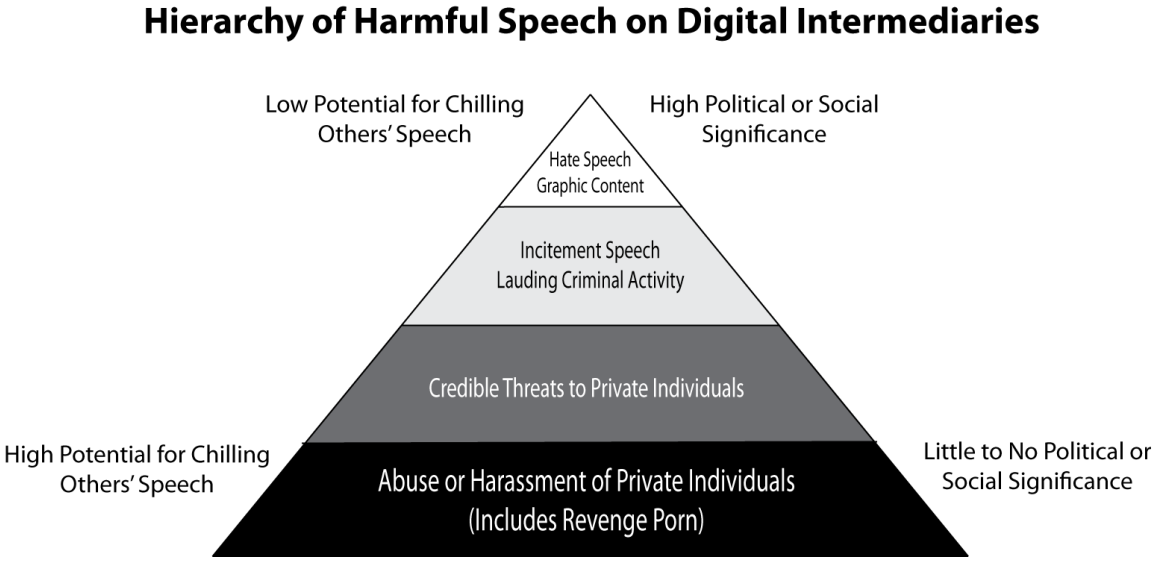

Figure 6-1: Hierarchy of Harmful Speech

Of course, the duty to follow various countries' laws proscribing the hosting of certain types of speech is an obvious exception to upholding a duty to protecting freedom of speech and individual autonomy. Although following such laws may violate the rule of keeping a primary duty to freedom of speech, doing so nevertheless upholds a duty to

respect individual autonomy. To a greater or lesser extent, laws are an expression of the cultural ethos from which they were enacted[86]—ideally, they are the expression of the democratic will of the people that they affect. However difficult it may be, changing laws is still a function of public agency, either via democratic deliberation or through revolution. If someone in Pakistan wants that country's law against blasphemy abolished, he or she can exercise his or her free will and lead others into the streets to protest and demand change, even if the likelihood of success is very low. Following DeNardis's concepts of Internet governance, such an example would involve an intermediary simply following the laws it is required to follow to operate within a given country.[87] It should not be held responsible for not standing up to laws that stifle speech that it did not pass and does not enforce.

However, a digital intermediary should not prevent content from being viewed in a country whose laws do not prohibit the content. Such a move is bereft of any respect for individuals' autonomous ability to consume the content and judge it on their own free will. Indeed, self-regulation through the following of ethical principles has the benefit of preventing state actors from imposing potentially more heavy-handed regulation. Self-regulation may have other benefits as an alternative to state or legal regulation. Law professor Monroe Price and communication scholar Stefaan Verhulst argue that self-regulation "has the apparent benefit of avoiding state intervention in sensitive areas of basic rights, such as freedom of speech and information, while offering standards for

[86] *See, e.g.*, Alf Ross, *Tû-Tû*, 70 HARV. L. REV. 812 (1957).
[87] DENARDIS, *supra* note 7.

social responsibility, accountability, and user protection from offensive material."[88] However, the authors caution that when it comes to matters of protecting freedom of expression, self-regulation can lead to "private censorship," which "can be more coercive and sweeping than public censorship."[89] In fact, they argue, "[t]he dangers of constitutional violation are particularly strong where the self-regulatory entity is acting in response to government or as a means of preempting its intervention."[90] Although no legal constitutional violation of free speech rights may take place in such a scenario, the extralegal restriction of speech may have a stronger and more sweeping effect on the robustness of online public discourse than a legal (i.e. state-sponsored) restriction. Digital intermediaries should be conscious of their power to bring about such violations of individuals' ability to speak freely, and they should do no more than laws require of them to remove harmful content.

Applying Platform Ethics to Facebook's Community Standards

Chapter 5 traced the evolution of Facebook's rules governing users' speech, from their early days as a part of the company's terms of use, to their current version: the March 2015 update to what are now called "Community Standards." Two themes were identified in this evolutionary process. First, Facebook has sought to find the ideal balancing point between protecting the ability of its users to speak on its platforms and preventing potential harms to other users that that speech could cause. In this balancing act, Facebook has declared that it will judge content based on the intent of the users who post it, as well as on the context of the content. Decisions on whether speech is allowable

[88] MONROE E. PRICE AND STEFAAN G. VERHULST, SELF-REGULATION AND THE INTERNET 9 (2005).
[89] *Id*.
[90] *Id*.

will be made on a case-by-case basis, with flexible rules guiding both users and Facebook on how those decisions are made. Second, Facebook has maintained that it takes no responsibility for either the harms it fails to prevent or the speech that it removes from its platforms. It asks its users for patience as it attempts to find the ideal balance between the two goals—if one even exists.

Community standards are important tools for guiding the ethical operations of digital intermediaries like Facebook. They "evolve according to changing norms. They influence the overall perception of what is appropriate in a society and, at the same time, are influenced by overall norms."[91] These standards "neither mirror public opinion nor necessarily lag or lead public opinion in terms of cultural norms," but rather simply "represent a temporarily agreed upon set of standards which serves as a way-station or *modus vivendi* as new modes of information are distributed."[92] It probably "is not desirable for providers to be too precise or to have exceedingly strict standards to measure compliance," as "[g]reater flexibility can make it easier to respond to changes in technology [and] modify expectations and outcomes."[93] Indeed, "the notion of harmful content—which is culturally diverse and subject to changing norms—is be[st] suited to self-regulation."[94]

The argument put forth in this chapter does not disagree that such flexibility has its advantages. The only difference that this chapter would seek to add to this self-regulation regime is that platforms should make a commitment to protect speech first—to

[91] *Id*. at 43.
[92] *Id*.
[93] *Id*. at 35.
[94] *Id*.

announce courageously to users that their default position will be to protect speech. Clearly, Facebook has not made such a commitment in its community standards. Although Facebook's decision to judge speech based on the intent of the speaker and the context of the speech is admirable and a step in the right direction for protecting users' speech on its platforms, its "we're not perfect" message is not. Facebook must clearly state that its primary duty as a digital intermediary is to facilitate the speech of its users, due to all the benefits that speech brings to society. It then must open up its process of governing speech to its users so that they can hold Facebook accountable to upholding that duty.

The purpose of this chapter was to build a set of ethical guidelines that maximizes respect for individual autonomy and protection of freedom of expression. The purpose of this chapter was not necessarily to argue that the policies and principles of these three mainstream intermediaries are misguided, or that there is a rampant trend going on of intermediaries removing unpopular or potentially harmful political speech that requires a revision in their policies. However, the sheer presence of such removals, combined with the imprecise language of the community guidelines of mainstream digital intermediaries such as Facebook, spells potential trouble for individuals' speech on these platforms. Granted, it is hard to blame platforms for using squishy language when crafting their speech policies. As communication professor Tarleton Gillespie notes, such vague policies are part of an "effort to limit [intermediaries'] liability not only to … legal charges but also more broadly to cultural charges of being puerile, frivolous, debased,

etc."[95] Intermediaries seek to be "rewarded for facilitating expression but not liable for its excesses."[96] They want to have their cake and eat it too: promote speech and police extremes when they can. Such a strategy does a disservice to all individuals who depend on these platforms to participate in public discourse.

In 1996, Internet philosopher John Perry Barlow argued in his manifesto "A Declaration of the Independence of Cyberspace" that the Internet should be a realm free from the laws of the world of "flesh and steel."[97] As the Internet evolved and became popularized through the invention of the World Wide Web, Barlow's extreme libertarian vision of the Internet became unrealistic.[98] But the spirit of Barlow lives on in those who lament the commercialization and corporatizing of the Internet, particularly by a handful of powerful companies such as Google and Facebook. Law professor Jonathan Zittrain points out that the private powers that have come to dominate the Internet have simultaneously expanded and put boundaries around individuals' online communicative agency.[99] In such a context, law professors David Johnson and David Post argue, "The strongest claim to control [online] comes from the participants themselves."[100] Of course, Johnson and Post are referring to who should be able to claim sovereignty over cyberspace in the legal sense: individuals versus governments. Yet the same philosophy holds true in the ethical sense: individuals should have pride of place in the online public

---

[95] Gillespie, *supra* note 17, at 357.
[96] *Id*. at 356.
[97] JOHN PERRY BARLOW, A DECLARATION OF THE INDEPENDENCE OF CYBERSPACE (1996), *available at* https://projects.eff.org/~barlow/Declaration-Final.html.
[98] LAWRENCE LESSIG, CODE, VER. 2.0 (2006); GOLDSMITH AND WU, *supra* note 7.
[99] Jonathan Zittrain, *A History of Online Gatekeeping*, 19 HARV. J. L. & TECH. 253 (2006); JONATHAN ZITTRAIN, THE FUTURE OF THE INTERNET—AND HOW TO STOP IT (2008).
[100] David R. Johnson and David Post, *Law and Borders: The Rise of Law in Cyberspace*, 48 STAN. L. REV. 1367, 1375 (1996).

discourse. Thus, the mainstream digital intermediaries through which individuals communicate should, first and foremost, reinforce a commitment to facilitating individuals' speech. Second, they should commit clearly to recognizing personal abuse as the most grievous of harms capable of occurring on their platforms, and they should fight such abuse as strongly as they protect all other forms of speech.

Conclusion

This chapter has highlighted the fact that digital intermediaries face several competing duties: to promote free speech, to prevent harm, to obey the law, and to make money. By reviewing relevant scholarship in the field of media ethics and the philosophies behind competing concepts of intermediaries' legal liability for third-party content, the chapter concludes that digital intermediaries should place their duty to promoting free speech above their other duties. They must do so because both their business model and deliberative democracy depend on the vibrant public discourse that individuals have created by using their services. Their duty to prevent harm caused by speech published on their platforms should be coextensive with their duty to protecting speech. Thus, only speech that personally threatens, harasses or abuses private individuals should be removed from their platforms upon notification by users of the presence of such speech. Intermediaries also should only remove speech as obligated by countries' laws—they should go no further.

At its core, an ethics for platforms is all about accountability. Digital intermediaries should be held accountable for the enormous power they have to control the speech of individuals who depend on these platforms to participate in global public

discourse.[101] This issue is common in the field of media ethics, where "[m]uch of the debate on media accountability has focused on efforts to neutralize the tension between journalistic autonomy and the need for a responsible press."[102] Ethicist Patrick Plaisance argues that "media accountability … is not a dilemma to be solved but a healthy tension to be managed. Although codes of ethics and correction boxes have their places, the media are accountable when they never stop seeking that uncomfortable balance with audience values."[103] Likewise, the concept of platform ethics proposed in this chapter will not solve the problems of harmful speech on digital intermediaries, nor will it sap the power of intermediaries and make them fully accountable to speakers. Rather, the hope of this theory is that it will protect speakers and the global public discourse by committing platforms to placing these values above all others.

[101] DENARDIS, *supra* note 7.
[102] Patrick L. Plaisance, *The Concept of Media Accountability Reconsidered*, 15 J. OF MASS MEDIA ETHICS 257, 266 (2000).
[103] *Id*.

Chapter 7: Discussion and Conclusion

"Mob Mentality?"

In an appearance on the March 11, 2015 episode of Comedy Central's "The Nightly Show" devoted to the issue of banning words on college campuses,[1] John Avlon, editor-in-chief of *The Daily Beast*, argued that individuals on social media are more concerned about being offended than about freedom of speech. "Social media is creating a mob mentality where people are all of a sudden acting as the police," Avlon said, "and they're going to pile on anyone that says anything that offends them."[2] Such a sanction, Avlon argued, "has the apparent velocity of force, … and what that's doing is having a chilling effect on discourse."[3]

Avlon's opinion is just that: opinion. However, it is an opinion that sits at the heart of this study. The First Amendment protects public discourse from government interference. But that same public discourse now faces potential threats from easily offended individuals and powerful private institutions (in the form of digital intermediaries) without the same First Amendment protection. This study has discussed and synthesized theories from several distinct schools of thought to give a theoretical shape to this phenomenon of "content governance." This concluding chapter briefly will review three key points to take away from this study. It will also address the main strengths and weakness of the study, as well as briefly summarize the original

---

[1] *Panel: Banning Words*, COMEDY CENTRAL: THE NIGHTLY SHOW WITH LARRY WILMORE (March 11, 2015), *available at* http://www.cc.com/video-clips/2nwm3v/the-nightly-show-panel---banning-words (at 5:38 of video clip). *Also available at* John Avlon, *The World Police & the Very Offended Internet Mob*, The Daily Beast (March 12, 2015), *available at* http://www.thedailybeast.com/articles/2015/03/12/the-word-police-the-very-offended-internet-mob.html.
[2] *Id*. (at 5:38 of video clip).
[3] *Id*. (at 5:55 of video clip).

contribution of this study to research in the field of mass communication law. Finally, and most importantly, the chapter will propose avenues of future research that the highly theoretical concepts from this study can inform.

Brief Overview

Three Key Takeaways

1. Improve literacy on individuals' interactions with digital intermediaries

Digital intermediaries have afforded individuals the ability to publish messages within a global public discourse. Individuals can bypass major organizations to create and share content that is entertaining, banal, thought-provoking, politically charged, revolutionary or reactionary. Yet individuals can also use these tools to publish speech that causes varying degrees of harm, from mere offense to personal abuse or harassment to instigating physical harms to persons and property. Finally, individuals can persuade intermediaries to remove speech they find too extreme or harmful by flagging it. Intermediaries may opt to remove that speech if it violates the standards that they create for the "community" of users that they foster.

Individuals must be cognizant of several characteristics of this system of communication. First, their communicative agency depends on the digital intermediaries who facilitate the networked platforms for communication. Under such a system of dependence, individuals must be aware of whether and to what extent digital intermediaries will manage the extreme speech that gets published on their platforms. Individuals also must understand both the values that extreme speech brings to society, as well as the limits demarcating when extreme speech becomes regulable harmful speech.

2. Build a culture of tolerance toward extreme speech

Understanding the values and limits of extreme speech involves building a culture of tolerance toward such speech. A culture of tolerance has two key attributes. First, tolerance requires individuals who may otherwise wish to exercise a "natural" tendency[4] to censor extreme speech through flagging it to refrain from doing so out of a desire to allow that speech the opportunity to compete in the marketplace of ideas.[5] Second, tolerance is an active and educational process. Tolerance is not blind acceptance of all speech simply for acceptance's sake. Tolerance involves accepting extreme speech out of a desire to improve one's mental faculties through active engagement with speech that is extreme, offensive or disagreeable. As the power of individuals and private institutions to stifle speech increases relative to the power of state actors to do so, building a culture of tolerance has become crucially important for the vitality of public discourse.

3. Encourage transparency in content governance

Digital intermediaries such as Facebook have continued to revise their community standards and rules for speech published on their platforms. Facebook's community standards seek to strike a balance between promoting speech and preventing harm to users. Facebook also has the legal authority to declare itself immune from liability for any harm it fails to prevent or any speech it removes in the name of preventing harm. It certainly would not be too audacious of a generalization to say that other mainstream digital intermediaries (such as YouTube or Twitter) do the same with their standards.

---

[4] *See, e.g.*, LEE C. BOLLINGER, THE TOLERANT SOCIETY: FREEDOM OF SPEECH AND EXTREMIST SPEECH IN AMERICA 92 (1986); JOHN STUART MILL, ON LIBERTY 8, 12 (1859/2001); Thomas I. Emerson, *Toward a General Theory of the First Amendment*, 72 YALE L. J. 877, 884 (1963); RODNEY A. SMOLLA, FREE SPEECH IN AN OPEN SOCIETY 4 (1992).

[5] *See generally* MILL, *supra*.

However, a great deal of the process by which intermediaries govern users' speech remains a mystery. Intermediaries depend on users' speech for their economic livelihood as much as individuals depend on intermediaries to participate in the public discourse. Therefore, intermediaries should make a clear commitment to protecting users' ability to speak on their platforms, which requires intermediaries to abide by an ethical duty to make their content governance process more transparent for their users.

Original Contribution to Scholarship

The theories and doctrines regarding freedom of expression discussed throughout this study are not, in and of themselves, original. What is original is how they have been synthesized and applied to the study of governance of extreme speech in networked communication. The perspective that this study takes—that more speech is better than less and that the goals of the First Amendment are ultimately beneficial to democratic deliberation in society—is not, in and of itself, original. What is original is how this perspective responds to scholars who call for digital intermediaries to play a greater role in protecting individuals from harmful speech.[6] Although there is a place for such content governance, this study cautions that it is a remedy whose impact on public discourse must be understood. This study also contributes original research to the field of mass communication ethics by arguing that field must embrace theoretical approaches to understanding and assessing the roles that digital intermediaries play in public discourse. Finally, the empirical analysis of the evolution of Facebook's rules governing users'

[6] *See, e.g.*, AMY GAJDA, THE FIRST AMENDMENT BUBBLE: HOW PRIVACY AND PAPARAZZI THREATEN A FREE PRESS (2015); DANIELLE KEATS CITRON, HATE CRIMES IN CYBERSPACE (2015); Mary Anne Franks, *Unwilling Avatars: Idealism and Discrimination in Cyberspace*, 20 COLUM. J. GENDER & L. 224 (2011).

speech is an original contribution to the field of mass communication and mass communication law. The social network's community standards are public knowledge, and major updates to them have garnered significant press attention. However, this study assesses those standards in terms of their evolution over Facebook's 11-year history, and in terms of their ability to successfully strike the balance between protecting speech and preventing harm.

Strengths and Weaknesses

The main strength of the research in this study is its ability to synthesize arguments from a wide and otherwise disconnected body of scholarship. The theories and jurisprudence of the First Amendment are too often the subject of an isolated body of scholarship. This isolation risks making these subjects the unapproachable province of elite thinkers and jurists, when in fact they belong (or are ideally conceived to belong) to the masses, equally and without exception. Calling upon the values of First Amendment theory and jurisprudence to analyze the norms of networked communication facilitated by digital intermediaries is an attempt to bring these values closer to the hoi polloi who communicate via these platforms.

The principal weakness of the research in this study is its abstractness. The theories and conclusions that this study puts forth are normative. To varying degrees, they rely conjectures and potentialities. They discuss a trend (content governance) that is indeed happening, yet they cannot say to what extent this trend is happening, and they cannot quantify the social harms of that trend. The root of this weakness lies in the choice to apply normative First Amendment theories to the study of content governance. This

abstractness also shows its flaws in the task of transposing First Amendment theories and values onto private actors. First Amendment analyses are generally black-and-white: they are framed as speaker versus a state actor. The problem with transposing First Amendment analyses onto private actors is that private actors come in all shades of gray.

However, these weaknesses are sources of strength for future research based on the conclusions arrived at in this study. Namely, the abstract conclusions from this study are ripe for more concrete and nuanced conceptualization through empirical analysis. The following section will address the future research agenda spawned from this study. It will give special attention to the paradigmatic traditions upon which this agenda will be based.

Law and Mass Communication Research

New Questions

This study opens the door for many new research questions. To what extent is content governance *perceived* as so powerful a control over speech that it rivals the evils associated with state censorship of speech? To what extent are individuals likely to take action against speech they find offensive or harmful by flagging it? What does "freedom of expression" even mean today, a time when individuals, state actors and digital intermediaries have varying degrees of agency within and control over the public discourse?

Answering these questions requires a research agenda that employs various types of social scientific and empirical analyses to complement traditional, documentary legal research. However, to truly understand the benefit of such a multi-method research

agenda, one must first understand the complementary relationship between legal research and research in mass communication.

The Legacy of Mass Communication Law in the Social Sciences

Mass communication law is crucial to the field of mass communication as a whole. The ultimate goal of the field of mass communication law is to further an understanding about freedom of expression in its many forms.[7] In their seminal 1981 chapter on legal research methods in mass communication, mass communication professors Don Gillmor and Everette Dennis declared that legal research in mass communication can achieve this goal by fulfilling one of five purposes: clarifying and explaining the law through analysis of procedure, precedent and doctrine; reforming old laws and creating new ones; providing a better understanding of how law operates in society; analyzing the political and social processes that shape law; and furnishing materials for legal education.[8] Researchers can utilize one of two methods to accomplish these ends, according to Gillmor and Dennis: "traditional legal research," involving exhaustive examination of cases and statutes; and "empirical and behavioral legal research, which employs the methods of social science while recognizing the unique circumstances and problems of law."[9] Each method operates by different rules and often pursues distinct outcomes. For example, unlike with social scientific research, scholarship employing the traditional method does not always pursue "knowledge for the sake of knowledge," but rather "an applied knowledge in keeping with the lawyer's

[7] Donald M. Gillmor and Everette E. Dennis, *Legal Research in Mass Communication*, *in* RESEARCH METHODS IN MASS COMMUNICATION (Guido H. Stempel, III & Bruce H. Westley eds., 1981) 341.
[8] *Id*. at 328-9.
[9] *Id*. at 321.

adversarial purpose" of winning a case.[10] Either method can be called upon, so long as the method is appropriate for answering the research question set before it.[11]

In 1986, the journal *Communications and the Law* published a special issue examining exactly what the relationship between research in mass communication law and research in the much broader field of mass communication should look like, and what sorts of research questions and corresponding methods could come out of this relationship. Overall, the issue called for an expansion in the use of multiple methods, both to increase the robustness of the field and to "diminish disciplinary isolation—both within and without the field of mass communication."[12] Dennis, himself, opened the issue with a call for media law scholars to "use other approaches to foster understanding of freedom of expression" beyond traditional legal research.[13] He warned that the traditional paradigm of legal research, involving (then, as now) "[n]arrow, adversarial studies or those that are no more than amicus briefs for media defendants in libel cases," will "do little to assure a rigorous analysis of complex rights in conflict."[14] For Professor F. Dennis Hale, this call meant giving greater attention to the third of Gillmor and Dennis's five purposes,[15] and meeting the "need for empirical research that measures the impact of the law of freedom of expression" on society.[16] Hale suggests studies that "measure the

---

[10] *Id*. at 331.
[11] *Id*. at 333.
[12] Robert E. Drechsel, *Mass Communication* of *the Law: Toward Theoretical Understanding of Journalists' Interacton with Judicial Sources*, 8 COMM. & L. 23, 33-34 (1986).
[13] Everette E. Dennis, *Frontiers in Communication Law Research*, 8 COMM. & L. 3, 6 (1986).
[14] *Id*. at 10.
[15] Gillmore and Dennis, *supra* note 8.
[16] F. Dennis Hale, *Impact Analysis of the Law Concerning Freedom of Expression*, 8 COMM. & L. 35, 35 (1986).

quantity and quality of specific legal activities in state and federal courts,"[17] assess the impact of administrative policies such as the now-defunct Fairness Doctrine on news content,[18] and measure the public opinion of legal doctrines[19] as furthering this mission. Professor David Pritchard charged researchers with employing social scientific theories, rather than the traditional normative theories of legal scholarship, to build "an understanding of how law actually works."[20] Professor Robert Drechsel, for his part, called on mass communication law scholars to "treat 'communication' as less an adjective than a noun—to expand the research agenda well beyond traditional scholarship on the principles, rules, and procedures of law that affect communication and to give more attention to the communication of law."[21]

Professors Jeremy Cohen and Timothy Gleason have conceived of a relationship between legal and mass communication research in which the First Amendment is considered a "paradigm."[22] Their main argument is that "a close familiarity with the *disciplines* of communication and law" enhances "a thorough understanding of the *concept* and *practice* of freedom of expression."[23] Similar to Gillmor and Dennis, Cohen and Gleason call for a research agenda that examines the relationship between law and mass communication.[24] Under the exceptional protections of the First Amendment, such an agenda may involve as much the study of the effects of extreme and potentially

---

[17] *Id*. at 42-43.
[18] *Id*. at 45.
[19] *Id*. at 48.
[20] David Pritchard, *A New Paradigm for Legal Research in Mass Communication*, 8 COMM. & L. 51, 56 (1986).
[21] Drechsel, *supra* note 12, at 23.
[22] JEREMY COHEN AND TIMOTHY GLEASON, SOCIAL RESEARCH IN COMMUNICATION AND LAW, 13, 110 (1990).
[23] *Id*. at 18 (original emphasis).
[24] *Id*. at 133.

harmful speech on society, as well as how certain actors respond to various guidelines and exceptions the law does impose (e.g. how common law privileges for media defendants in libel suits affect the frequency and outcome of libel suits against the press). However, the authors maintain that the end goal of such research need not and *should* not be the effectuation of change in First Amendment doctrine in this country.[25] Rather, the goal is to better understand the system in place so that the goals of normative theories of freedom of expression (the pursuit of truth, the realization of individual autonomy, training a tolerant mind) can be realized.

Today, scholars continue to echo Gillmor and Dennis's call for incorporating social scientific methods into research in mass communication law.[26] However, other legal scholars maintain that this endeavor will not bear fruit, as legal theories (and the doctrines that stem from those theories) are fundamentally normative and philosophical in nature, whereas social scientific theories are fundamentally empirical and falsifiable in nature.[27] In other words, legal theories are not testable in and of themselves, and transposing them into social scientific theories would strip them of their essential nature. Not only that, but the two paradigms pursue radically different goals. Research in mass communication law focuses on issues—such as libel, censorship and invasion of privacy—that are important because they attack the very heart of the legal tradition, not

[25] *Id*. at 98.
[26] *See generally* AMY REYNOLDS AND BROOKE BARNETT, COMMUNICATION AND LAW: MULTIDISCIPLINARY APPROACHES TO RESEARCH (2006).
[27] *See* Matthew D. Bunker and David K. Perry, *Standing at the Crossroads: Social Science, Human Agency and Free Speech Law*, 9 COMM. L. & POL'Y 1 (2004).

because they happen frequently.[28] In contrast, social scientists seek to reveal correlations and effects across a broad, generalizable population.[29]

These scholars are correct in their assessment of the rules and goals of the theories and methods of each paradigm.[30] However, the debate should not be over *whether* empirical theories and methods can be applied to legal research, but rather *when* they can be applied *appropriately*. For example, empirical studies that show a high correlation between an unpopular type of speech and the harm it is purported to lead to should not be invoked to make the argument that First Amendment jurisprudence should be revised and the unpopular speech made illegal (i.e. purpose number two of Gillmor and Dennis's five[31]). Such a practice violates the rules for what counts as a valid argument for restricting speech within First Amendment jurisprudence.[32] As Professor Jeremy Cohen puts it, "the object is to understand the application of the law as it is rather than to provide a rationale for what the law could or should be."[33] Not to mention, as Professor Clay Calvert and his coauthors point out, a regime of free speech jurisprudence where decisions of which speech to proscribe and which to allow come down to the results of

---

[28] COHEN AND GLEASON, *supra* note 22, at 13.
[29] *Id*.
[30] *Id.*; Anthony L. Fargo, *Social Science Research in Judges' First Amendment Decisions*, *in* REYNOLDS AND BARNETT, *supra* note 26, at 35.
[31] Gillmore and Dennis, *supra* note 8.
[32] State regulations of speech on the basis of content must pass the strict scrutiny standard of judicial review, whereby the state "must show that its regulation is necessary to serve a compelling state interest and that it is narrowly drawn to achieve that end," Perry Ed. Assn. v. Perry Local Educators' Assn., 460 U.S. 37, 45 (1983).
[33] Jeremy Cohen, *Degrees of Freedom: Parameters of Communication Law Research*, 8 COMM. & L. 11, 15 (1986).

scientific studies would be undesirable due to its reliance on imperfect, falsifiable and perhaps even manipulable conclusions.[34]

By the same token, citing normative legal theory to argue for the protection of extreme speech for the greater good of democratic society and public discourse will not wash away the empirically measurable harm that such speech can cause. The negative aspects of the U.S. free speech environment should lead researchers to figure out how people can best live within that environment, rather than how the environment can be torn down to make way for a new one. In other words, scholars can and should use social scientific methods to further Gillmor and Dennis's goal of studying the effects such a legal regime has on society. Such research could involve determining "whether something does or does not happen because of the law,"[35] such as whether the presence of state shield laws leads to more investigative reporting in those states.[36] Or, it could involve, as Professor Hale suggested, measuring public opinion of various aspects of U.S. media law.

Research on Freedom of Expression and Tolerance

One area of research that has been instrumental in the pursuit of understanding law from a social scientific perspective is the study of public attitudes toward the many extreme types of speech allowed under First Amendment jurisprudence. This body of research has reached a virtually uniform conclusion: "Although the notion of free speech

---

[34] *See generally*, Clay Calvert, Kara Carnley, Brittany Link and Linda Riedmann, *Conversion Therapy and Free Speech: A Doctrinal and Theoretical First Amendment Analysis*, 20 WM. & MARY J. WOMEN & L. 525 (2014).
[35] Cohen, *supra* note 33, at 16.
[36] *See, e.g.*, Vince Blasi, *The Newsman's Privilege: An Empirical Study*, 70 MICH. L. REV. 229 (1971); Laurence B. Alexander, *Looking Out for the Watchdogs: A Legislative Proposal Limiting the Newsgathering Privilege to Journalists in the Greatest Need of Protection for Sources and Informaton*, 20 YALE L. & POL'Y REV. 92 (2002).

is nearly irresistible to many people, certain specific examples of *how* other people may choose to enjoy that right may cause alarm and provoke intolerance among various segments of the public."[37] This conclusion is rather obvious; of course people are going to take issue with speech they disagree with or find offensive. What is more interesting is the assessment of the strength of individuals' intolerance: the extent to which they would be willing to see their lack of support for speech turn into some form of censorship of that speech.

Much of the research on attitudes toward speech (documented extensively in the 2004 work by mass communication professor Julie Andsager and colleagues)[38] focuses on individuals' support for *legal* sanctions against such forms of extreme speech as extremist political messages, pornography or other sexually charged speech, racism or sexism, and speech that could be damaging to national security. In other words, when the law is the exclusive remedy available to individuals to silence unpopular speech, the issue can be framed via the following question: "why are … individuals—regardless of their backgrounds—willing to sacrifice bits and pieces of their expressive rights, especially when each encroachment sets a precedent for further regulation?"[39]

Early Work

Historical context is important to research on attitudes toward speech. Generally, this research has focused on the extreme, unpopular and highly charged speech of its time. Samuel Stouffer conducted one of the first major investigations on public attitudes

[37] JULIE L. ANDSAGER, ROBERT O. WYATT AND ERNEST MARTIN, FREE EXPRESSION IN 5 DEMOCRATIC PUBLICS: SUPPORT FOR INDIVIDUAL AND MEDIA RIGHTS 78 (2004) (emphasis added).
[38] *Id*.
[39] *Id*. at 62.

toward speech in 1955. At that point in American legal history, the U.S. Supreme Court had begun relatively recently to expand the protections of the First Amendment to preclude state governments (not just the federal government) from passing laws that punished speech in the name of protecting law and order.[40] That era in American political history also witnessed the beginning of the Cold War and the rise of McCarthyism and the "Red Scare," which tested citizens' tolerance for the exercise of individual liberties to express communist messages.[41] Stouffer measured individuals' tolerance toward what was considered to be the most harmful type of speech of his time: communist speech.[42] He concluded that community elites and those with higher levels of education tended to be more tolerant of communist speech than those who were poorer or had lower levels of education.[43] However, Stouffer's work was later criticized for not adequately measuring individuals' perception of exactly how harmful communist speech was to society.[44]

The 1960s and 1970s saw a great expansion of First Amendment protections to many types of extreme speech, such as public display of curse words,[45] the KKK,[46] the American Nazi Party,[47] defamatory speech,[48] and publications potentially damaging to national security.[49] That era also saw the Supreme Court put some of the first major limits

[40] Near v. Minnesota, 283 U.S. 697 (1931); U.S. v. Carolene Products, 304 U.S. 144, FN 4 (1938)
[41] ANDSAGER, WYATT AND MARTIN, *supra* note 37, at 23.
[42] SAMUEL A. STOUFFER, COMMUNISM, CONFORMITY, AND CIVIL LIBERTIES (1955).
[43] *Id*.
[44] *See* JOHN L. SULLIVAN, JAMES PIERESON AND GEORGE E. MARCUS, POLITICAL TOLERANCE AND AMERICAN DEMOCRACY (1982).
[45] Cohen v. California, 403 U.S. 15 (1971).
[46] Brandenburg v. Ohio, 395 U.S. 444 (1969).
[47] National Socialist Party of America v. Skokie, 432 U.S. 43 (1977); Collin v. Smith, 439 U.S. 916 (1978).
[48] New York Times v. Sullivan, 376 U.S. 254 (1964).
[49] New York Times v. U.S., 403 U.S. 713 (1971).

on freedom of expression: obscenity or hardcore pornography;[50] incitement to imminent lawless action;[51] indecent speech on broadcast media;[52] and that burning a draft card was illegal because, despite its expressive qualities, it constituted destruction of government property.[53] This formidable period in First Amendment jurisprudence led scholars in the early 1980s to assess public opinion toward expression that fell within these protected and unprotected categories. In 1982, political scientists James Gibson and Richard Bingham published a study that measured support among ACLU and Common Cause members for the rights of controversial groups to engage in controversial speech acts.[54] The authors concluded that tolerance for speech varied depending on the group speaking and the message of that group.[55] That same year, philosophy professor John Immerwahr and legal researcher John Doble published a study that measured attitudes among both the public and journalists toward various controversial speech scenarios to determine the limits of public tolerance toward individuals or the press exercising their First Amendment rights.[56] They concluded that there is a great deal of disagreement among ordinary citizens and media elites over whether certain types of harmful speech should be allowed in society, with the former more likely to favor greater restriction of such types of speech than the latter.[57] Particularly, both of these studies show that public opinion was higher for speech with a strong political message (such as communist speech) than

[50] Miller v. California, 413 U.S. 15 (1973).
[51] Brandenburg v. Ohio, 395 U.S. 444 (1969); Hess v. Indiana, 414 U.S. 105 (1973).
[52] FCC v. Pacifica Foundation, 438 U.S. 726 (1978).
[53] U.S. v. O'Brien, 391 U.S. 367 (1968).
[54] James L. Gibson and Richard D. Bingham, *On the Conceptualization and Measurement of Political Tolerance*, 76 AMER. POL. SCI. REV. 603 (1982).
[55] *Id*. at 617.
[56] John Immerwahr and John Doble, *Public Attitudes toward Freedom of the Press*, 46 PUBLIC OPINION Q. 177 (1982).
[57] *Id*. at 185.

for speech with a more divisive and hateful (even if political) message (such as Nazis marching in Skokie).

Andsager and colleagues measured attitudes toward freedom of expression at the turn of the millennium, following a decade (the 1990s) that began with the Supreme Court upholding First Amendment protection for burning the American flag[58] and lampooning public officials with outrageous satire,[59] while striking down laws prohibiting cross burning[60] and punishing newspapers for publishing the name of a rape victim.[61] The decade also witnessed the growth of the so-called "PC" ("political correctness") movement, led by public universities' (failed) attempts to implement codes proscribing "hate speech,"[62] which later morphed into a movement that sought to socially condemn any forms of language used as a weapon "by the powerful to deny the interests of the oppressed."[63] Andsager and her colleagues compared attitudes toward freedom of expression among individuals in the United States, Russia, Hong Kong and Israel,[64] reaching several conclusions. First, somewhat obviously, "political, harmful, offensive, or routine," but the "exact components of these factors … vary from culture to culture."[65] Of particular interest among their U.S. study was their conclusion that individuals' "willingness to protect individual free speech rights seem[ed] paramount, with media

[58] Texas v. Johnson, 491 U.S. 397 (1989).
[59] Hustler Magazine, Inc. v. Falwell, 485 U.S. 46 (1988).
[60] R.A.V. v. St. Paul, 505 U.S. 377 (1992).
[61] Florida Star v. B.J.F., 491 U.S. 524 (1989).
[62] *See, e.g.*, Frank I. Michelman, *Universities, Racist Speech and Democracy in America: An Essay for the ACLU*, 27 Harv. C.R.-C-L. L. Rev 339 (1992); SMOLLA, *supra* note 4.
[63] Anthony Zurcher, *A Political Correctness War that Never Really Ended*, BBC NEWS ECHO CHAMBERS (Jan. 30, 2015). *See also* ANDSAGER, WYATT AND MARTIN, *supra* note 37, at 64.
[64] The researchers surveyed both Arabs and Jews in Israel, hence giving the book its title: "5 Democratic Publics."
[65] ANDSAGER, WYATT AND MARTIN, *supra* note 37, at 249.

rights scoring a somewhat distant second."[66] The authors attributed this gap to "the abstract nature of threats to personal security for most Americans."[67] Support for freedom of expression tended to be highest among relatively more affluent and educated males (and, in the United States, *white* males), leading the authors to propose the following theory: "people who are most secure within society are most likely to support expressive rights."[68]

Tolerance and Censorial Behavior

The studies cited above have indicated that support exists among certain groups of individuals to have certain types of unpopular or extreme speech legally proscribed under certain circumstances. However, Professor Jennifer Lambe has pointed out that there is a conceptual distinction between public opinion, tolerance and attitude toward certain types of speech and the willingness of people to censor that speech.[69] In other words, certain types of speech may be unpopular, but people still may believe that such speech should be allowed to exist in society. Thus, Lambe created the Willingness to Censor (WTC) scale to assess whether and how individuals would want certain types of speech to be regulated. The scale consists of a survey with 49 questions, each corresponding to a combination of one of seven types of extreme or harmful speech (hate speech, pornography, controversial political speech, abortion speech, commercial speech, defamation, speech that violates privacy) expressed via one of seven different media.[70]

---

[66] *Id*.
[67] *Id*. at 251. *See, e.g.*, Frederick Schauer, *The Exceptional First Amendment*, *in* AMERICAN EXCEPTIONALISM AND HUMAN RIGHTS (Michael Ignatieff ed., 2005).
[68] ANDSAGER, WYATT AND MARTIN, *supra* note 37, at 258.
[69] Jennifer L. Lambe, *Dimensions of Censorship: Reconceptualizing Public Willingness to Censor*, 7 COMM. L. & POL'Y 187 (2002).
[70] *Id*. at 222.

Participants are asked to what extent they would agree with the following responses by the government to harmful speech in a given medium: a law banning the speech; requiring that the speech be expressed only at certain times or in certain manners; no response; active support for the speech. Lambe then examined which potential demographic factors could be associated with willingness to censor in each of the 49 examples. In several studies published since, she concluded that multiple factors, such as age, political ideology and gender, predict willingness to censor particular messages in particular circumstances.[71]

Lambe's work is novel for its attempt at distinguishing attitude (public disdain toward speech) from behavior (actively taking steps to act on that disdain and censor the speech). It is also novel in its adoption of actual scenarios involving harmful speech and potential government responses to that speech into its instrument of measurement. However, one criticism of Lambe's work posed here is that it seems to assume that individuals are sufficiently familiar with the actual government responses to give accurate responses regarding their favorability toward each response. Thus, Lambe's participants may in fact be using her scale as a proxy simply to express their opinions toward the speech.

Tolerance, Censorial Behavior and Content Governance: A Research Agenda

Lambe's research involves individuals' support for censorship options available only to *government* actors. However, Lambe's work could prove valuable when applied to a context in which individuals have the ability to stifle unpopular speech with

[71] Jennifer L. Lambe, The Structure of Censorship Attitudes, 13 COMM. L. & POL'Y 485 (2008); Jennifer L. Lambe and Jason B. Reineke, *Public Attitudes about Government Involvement in Expressive Controversies*, 59 J. COMM. 225 (2009).

relatively greater ease and prospects for success than attempting to invoke the law to achieve the same goal. Namely, Lambe's WTC scale could be adapted to measure the likelihood that individuals would seek to have a digital intermediary remove unpopular speech from its platform, engage in a campaign to have advertisers boycott a platform that hosts the speech, engage in a campaign to publicly shame the speaker of the unpopular message, or rally an online crowd against the message. Such a study ultimately could get at the heart of all the preceding studies of attitudes toward speech, the notion that "[t]he force of public opinion may be as effective as laws, if not more so, in limiting expression."[72] This notion reflects the focus of the multi-method research agenda proposed in this chapter: the issue of how the definition and values of freedom of expression are changing in a world where more people have greater communicative agency than ever before, and where digital intermediaries have become the arbiters of individuals' communicative activity.

Building upon Lambe's work, I propose to measure individuals' support for the extralegal methods of managing extreme UGC discussed throughout this study. I also propose to measure individuals' willingness to engage in a censorial behavior (such as speaking out against the content offline or online, or flagging[73] the content to force the platform to remove it). This study, therefore, would seek to assess individuals' *agency* to censor and *deference* to content governance.

Based on the doctrinally grounded definitions of the limits and the social value of extreme speech discussed in this study, the research must now move into its second

[72] ANDSAGER, WYATT AND MARTIN, *supra* note 37, at 6.
[73] Crawford and Gillespie, *supra* note 6.

phase: developing an instrument to measure agency to censor and deference to content governance. Guided by the "exploratory model" of mixed-method research,[74] this phase of the research will begin by conducting focus groups. Ideally, the focus groups will facilitate complimentary and argumentative interactions among individuals in which a common and recognizable vocabulary can be developed for the subject of content governance.[75] Using the responses from the focus groups, instruments will be devised for two quantitative analyses that will be conducted in sequential stages. First, the researcher will use Q method to "unmask deeply held opinions in such a manner that people who respond … in specific ways can begin to be grouped into factors or types defined according to similarities and differences in the attitudes, motives, and wants they report."[76] The Q method study will involve giving participants roughly 30 statements that they will have to arrange from *most agreeable* to *least agreeable*. Examples of such statements may include:

- Individuals have a duty to police extreme speech on digital intermediaries.
- Individuals have a right to say whatever they want on digital intermediaries.
- Digital intermediaries have a duty to prevent harm caused by speech published on their platforms.
- Digital intermediaries have a duty to promote the freedom of expression of users on their platforms.

[74] JOHN W. CRESWELL AND VICKI L. PLANO CLARK, DESIGNING AND CONDUCTING MIXED METHODS RESEARCH 76 (2007) (arguing that the goal of the focus groups is to ensure strong reliability and construct validity when developing the instruments for the quantitative stages of the study).
[75] THOMAS R. LINDLOF AND BRYAN C. TAYLOR, QUALITATIVE COMMUNICATION RESEARCH METHODS 3RD ED., 183 (2011).
[76] Bryan H. Reber, Fritz Cropp and Glen T. Cameron, *Mythic Battles: Examining the Lawyer-Public Relations Counselor Dynamic*, 13 J. PUBLIC RELATIONS RES. 187, 192 (2001).

Second, I will use survey methodology to measure attitudes toward harmful UGC and assess willingness to engage in or endorse censorial behavior of such content, following Lambe's research on willingness to censor.[77] This method is effective because it translates complex legal abstractions into constructs that are comprehensible to a lay audience without also sacrificing the essential legal character of the constructs.[78] In that same spirit, questions will seek to probe how potential survey participants comprehend harmful UGC, with the goal being construction of survey questions that indeed reflect plausible real-life scenarios as closely as possible. Based on analysis of participant comprehension of key terms, the best scale to use to measure attitudes toward the stimuli also can be assessed (3-, 5-, and 7-point scales have all been used in previous studies).

The third phase of the research will be to test the survey on a convenience sample, most likely made up of undergraduate students. Once the reliability of the instrument has been assured through pilot study, the project will move to its fourth phase, which will be to administer the survey to a sample with demographic variability that is more representative of the general population. Possibilities for attaining such a sample include Amazon's Mechanical Turk or Knowledge Networks' KnowledgePanel.

The ultimate goal of this research is to give empirical structure to the abstract concepts of content governance discussed in this study. Now that individuals have the ability to act on what several big-name legal theorists long have called a natural

[77] Lambe, *supra* note 69.
[78] *Id*.

proclivity to censor,[79] it is imperative that social scientific research identifies whether and to what extent this proclivity exists. The very health of public discourse is at stake.

The Future of Content Governance: Concluding Thoughts

The analyses and discussions in this study are a slice in time. Whether a year or 10 years from now, Twitter, Facebook and YouTube almost certainly will be completely different than on the day this study was successfully defended. These intermediaries may no longer even exist, folding due to a MySpace-esque fate of lack of popularity. New intermediaries may take their place. As the communicative landscape changes, so too will the norms of communication in that landscape. Studying these changes means shooting at a moving target. However, the conclusions from this study and the research agenda that will grow from it ideally will provide a reference point from which to shoot.

Scholars of mass communication law have long had a duty to guide speakers and audiences through our chaotic world of extreme, offensive and potentially harmful speech. When angry crowds call for the censorship of such speech or the punishment of its speakers, it is our duty to facilitate a dialogue among all parties affected by the speech. We must have the courage to defend the right of the speaker to offend, the acumen to distinguish mere offense from greater harm, and the humanity to clearly explain our position to a lay audience. We also must have the sensitivity to understand the pain that speech can inflict upon certain social groups, and we must have the skill to ensure that their voice is fairly represented in the dialogue without allowing that voice to become a heckler's veto against the speech. Most importantly, we must encourage an active and

[79] *Supra* note 4.

mindful tolerance of extreme speech based on an understanding of why allowing such speech to have a place in public discourse is ultimately valuable for society.

In the end, the concern of this study is the health and robustness of public discourse. This is a subject at the core of research in both mass communication law and mass communication in general. It is a subject upon which rests the integrity of deliberative democracy. As law professor Robert Post writes,

> [M]ore is at stake in the regulation of public discourse than the simple question of laissez faire. Quite beyond values of individual human liberty and personal self-realization lies the significance of the *collective* virtue of self-government. Traditional First Amendment doctrine, with its quaint focus on autonomy and the indeterminacy of national identity, is one of the last remaining areas of constitutional law to engage seriously the project of self-determination. If we discard that project as childish myth, so do we also discard our commitment to democracy, at least as our constitutional tradition has so far understood democracy.[80]

[80] ROBERT POST, CONSTITUTIONAL DOMAINS 288-9 (1995).

Appendix 1: Content Code of Conduct (Corresponding to Chapter 5)

(May 24, 2007)
Source: Wayback Machine
http://web.archive.org/web/20070515003455/http://www.facebook.com/codeofconduct.php

Facebook is a social utility that connects people with friends and others who work, study and live around them. People use Facebook to keep up with friends, to share links, to share photos and videos of themselves and their friends, and to learn more about the people they meet. We want Facebook to be a place where people respect the rights and feelings of others, including third party intellectual property rights. Therefore, we have established certain rules for using Facebook and for posting messages, photos, video and other content ("Content") on Facebook, which rules are set forth in our Terms of Use and in this User Code of Conduct. WHEN YOU USE FACEBOOK, YOU ARE AGREEING TO ABIDE BY THE USER CODE OF CONDUCT AND THE OTHER RULES SET FORTH IN OUR TERMS OF USE. FAILURE TO ADHERE TO THIS CODE OF CONDUCT AND THE TERMS OF USE MAY RESULT, AMONG OTHER THINGS, IN TERMINATION OF YOUR ACCOUNT AND THE DELETION OF CONTENT THAT YOU HAVE POSTED ON FACEBOOK, WITH OR WITHOUT NOTICE, AS DETERMINED BY FACEBOOK IN ITS SOLE DISCRETION. Please refer to our Terms of Use for more information about the rules applicable to your use of Facebook and the other rights and remedies of Facebook.

Third-Party Content

[Pertains to copyright]

Inappropriate Content

While we believe users should be able to express themselves and their point of view, certain kinds of speech simply do not belong in a community like Facebook. Therefore, you may not post or share Content that:

- is obscene, pornographic or sexually explicit
- depicts graphic or gratuitous violence
- makes threats of any kind or that intimidates, harasses, or bullies anyone
- is derogatory, demeaning, malicious, defamatory, abusive, offensive or hateful

Unlawful or Harmful Content or Conduct

Although as an online service provider, we are not responsible for the conduct of our users, we want Facebook to be a safe place on the internet. Therefore, in using Facebook, you may not:

- violate any local, state, national or international law or post any Content that would encourage or provide instructions for a criminal offense
- impersonate any person or entity or otherwise misrepresent yourself, your age or your affiliation with any person or entity
- use Facebook to send or make available any unsolicited or unauthorized advertising, solicitations, promotional materials, "junk mail," "spam," "chain letters," "pyramid schemes," or any other form of solicitation
- post or share any personally identifiable or private information of any third party
- solicit passwords or personal information from anyone, including those under 18
- use information or content you obtained on the Facebook website or service in any manner not authorized by the Facebook Code of Conduct or Terms of Use
- post any material that contains software viruses or any other computer code, files or programs designed to interrupt, destroy or limit the functionality of any computer software or hardware or telecommunications equipment
- register for more than one account or use or attempt to use another's account, service or system without authorization or create a false identity on the Service or the Site
- engage in any predatory or stalking conduct

This Content Code of Conduct is subject to change at any time at Facebook's sole discretion.

Appendix 2: Facebook's Community Standards (Corresponding to Chapter 5)

Facebook Community Standards

(Feb. 9, 2011)
Source: Wayback Machine
http://web.archive.org/web/20110209013433/https://www.facebook.com/communitystandards//

Facebook is a global community where millions of people connect with each other. Each of these people represents unique opinions, ideals, and cultural values. Out of consideration for this diversity, we work to foster an environment where everyone can openly discuss issues and express their views, while respecting the rights of others. When millions of people get together to share things that are important to them, sometimes these discussions and posts include controversial topics and content. We believe this online dialog mirrors the exchange of ideas and opinions that happens throughout people's lives offline, in conversations at home, at work, in cafes, and in classrooms. As a trusted community of friends, family, coworkers, and classmates, Facebook is largely self-regulated. People who use Facebook can and do report content that they find questionable or offensive. To balance the needs and interests of a global community we ask everyone to respect the following content standards:

Threats We want our members to feel safe on the site. Any credible threats to harm others will be removed. We may also remove support for violent organizations.

Promoting Self-Harm Facebook is not a place for self-destructive behavior. To that end we don't allow the promotion of suicide, "cutting," eating disorders, or illegal drug use. We take threats of suicide very seriously and will contact the relevant authorities when we become aware of them.

Bullying & Harassment As a community, we place a high value on respecting each other, and take reports of harassment very seriously. We take action when private individuals are bullied or persistently contacted against their wishes. While we encourage you to make meaningful new connections, please keep in mind that contacting strangers or people you've never met in person can be a form of harassment.

Hate Speech Facebook does not tolerate hate speech. Please grant each other mutual respect when you communicate here. While we encourage the discussion of ideas, institutions, events, and practices, it is a serious violation of our terms to single out individuals based on race, ethnicity, national origin, religion, sex, gender, sexual orientation, disability, or disease.

Graphic Violence While we are a platform for sharing events that take place in your life and around the world, any inappropriately graphic content will be removed when found on the site. Sadistic displays of violence against people or animals, or depictions of sexual assault, are prohibited.

Sex & Nudity We have a strict "no nudity or pornography" policy. Any content that is inappropriately sexual will be removed. Before posting questionable content, be mindful of the consequences for you and your environment.

Theft, Vandalism, or Fraud We are trying to make the world a more open, connected, and ultimately better place. Organizing acts that harm others through theft, vandalism, or fraud is a violation of our terms.

Identity & Privacy Facebook is a community where real people connect and share using their real identities. When you represent yourself accurately on Facebook you are helping to build trust and safety for everyone. Claiming to be someone else, creating multiple accounts, or falsely representing an organization undermines this trust and violates our terms. Please also refrain from publishing other people's personal information.

Intellectual Property Before sharing content on Facebook, please be sure you have the right to do so. We ask that you respect copyrights, trademarks, and other legal rights.

Phishing & Spam We take the safety of our members seriously and work to prevent attempts to compromise their privacy or security. We also ask that you respect our members by not contacting them for commercial purposes without their consent.

Reporting Abuse

If you see something on Facebook that you believe violates our terms, you can report it to us. Please keep in mind that reporting a person, organization, or piece of content doesn't guarantee its removal from the site. Because of the diversity of our community, it's possible that something could be disagreeable or disturbing to you without meeting the criteria for being removed or blocked. For this reason, we also offer personal controls over what you see, such as the ability to hide or quietly cut ties with people, Pages, or applications that offend you. Content that does violate our terms may be removed from our site and (in some cases) subject to legal or other action.

Facebook Community Standards

(Dec. 15, 2012)
Source: Wayback Machine
http://web.archive.org/web/20121215024155/http://www.facebook.com/communitystandards

[NOTE: Red text denotes changes from original (Feb. 9, 2011) standards]**

Facebook gives people around the world the power to publish their own stories, see the world through the eyes of many other people, and connect and share wherever they go. The conversation that happens on Facebook – and the opinions expressed here – mirror the diversity of the people using Facebook.

To balance the needs and interests of a global population, Facebook protects expression that meets the community standards outlined on this page.

Please review these standards. They will help you understand what type of expression is acceptable, and what type of content may be reported and removed.

Violence and Threats Safety is Facebook's top priority. We remove content and may escalate to law enforcement when we perceive a genuine risk of physical harm, or a direct threat to public safety. You may not credibly threaten others, or organize acts of real-world violence. Organizations with a record of terrorist or violent criminal activity are not allowed to maintain a presence on our site. We also prohibit promoting, planning or celebrating any of your actions if they have, or could, result in financial harm to others, including theft and vandalism.

Self-Harm Facebook takes threats of self-harm very seriously. We remove any promotion or encouragement of self-mutilation, eating disorders or hard drug abuse. We also work with suicide prevention agencies around the world to provide assistance for people in distress.

Bullying and Harassment Facebook does not tolerate bullying or harassment. We allow users to speak freely on matters and people of public interest, but take action on all reports of abusive behavior directed at private individuals. Repeatedly targeting other users with unwanted friend requests or messages is a form of harassment.

Hate Speech Facebook does not permit hate speech, but distinguishes between serious and humorous speech. While we encourage you to challenge ideas, institutions, events, and practices, we do not permit individuals or groups to attack others based on their race, ethnicity, national origin, religion, sex, gender, sexual orientation, disability or medical condition.

Graphic Content People use Facebook to share events through photos and videos. We understand that graphic imagery is a regular component of current events, but must balance the needs of a diverse community. Sharing any graphic content for sadistic pleasure is prohibited.

Nudity and Pornography Facebook has a strict policy against the sharing of pornographic content and any explicitly sexual content where a minor is involved. We also impose limitations on the display of nudity. We aspire to respect people's right to share content of personal importance, whether those are photos of a sculpture like Michelangelo's David or family photos of a child breastfeeding.

Identity and Privacy On Facebook people connect using their real names and identities. We ask that you refrain from publishing the personal information of others without their consent. Claiming to be another person, creating a false presence for an organization, or creating multiple accounts undermines community and violates Facebook's terms.

Intellectual Property Before sharing content on Facebook, please be sure you have the right to do so. We ask that you respect copyrights, trademarks, and other legal rights.

Phishing and Spam We take the safety of our members seriously and work to prevent attempts to compromise their privacy or security. We also ask that you respect our members by not contacting them for commercial purposes without their consent.

Security We take the safety of our members seriously and work to prevent attempts to compromise their privacy or security, including those that use fraud or deception. Additionally, we ask that you respect our members by not contacting them for commercial purposes without their consent.

Reporting Abuse

If you see something on Facebook that you believe violates our terms, you should report it to us. Please keep in mind that reporting a piece of content does not guarantee that it will be removed from the site.

Because of the diversity of our community, it's possible that something could be disagreeable or disturbing to you without meeting the criteria for being removed or blocked. For this reason, we also offer personal controls over what you see, such as the ability to hide or quietly cut ties with people, Pages, or applications that offend you.

Facebook Community Standards

(Nov. 9, 2013)
Source: Wayback Machine
http://web.archive.org/web/20131109125448/https://www.facebook.com/communitystandards

[The Nov. 9, 2013, standards are the same as the Dec. 15, 2012, standards, except for the following update]:

Graphic Content Facebook has long been a place where people turn to share their experiences and raise awareness about issues important to them.

Sometimes, those experiences and issues involve graphic content that is of public interest or concern, such as human rights abuses or acts of terrorism. In many instances, when people share this type of content, it is to condemn it. However, graphic images shared for sadistic effect or to celebrate or glorify violence have no place on our site. When people share any content, we expect that they will share in a responsible manner. That includes choosing carefully the audience for the content. For graphic videos, people should warn their audience about the nature of the content in the video so that their audience can make an informed choice about whether to watch it.

[…]

Facebook Community Standards

(Feb. 8, 2015)
Source: Wayback Machine
http://web.archive.org/web/20150209145145/https://www.facebook.com/communitystandards/

[The Feb. 8, 2015, standards are the same as the Nov. 9, 2013, standards, except for the following update]:

Regulated Goods: It is not permitted to complete transactions involving regulated goods on our platform. If you post an offer involving firearms, alcohol, tobacco, or adult products, we expect you to make sure you're following all applicable laws and consider carefully the audience for that content. If you are using a Page to connect with your customers and other audiences, you need to abide by our Pages Terms.

[…]

Community Standards [completely revised]

(Mar. 15, 2015)
Source: Facebook
https://www.facebook.com/communitystandards

Our mission is to give people the power to share and make the world more open and connected. Every day, people come to Facebook to share their stories, see the world through the eyes of others and connect with friends and causes. The conversations that happen on Facebook reflect the diversity of a community of more than one billion people.

We want people to feel safe when using Facebook. For that reason, we've developed a set of Community Standards, outlined below. These policies will help you understand what type of sharing is allowed on Facebook, and what type of content may be reported to us and removed.

Because of the diversity of our global community, please keep in mind that something that may be disagreeable or disturbing to you may not violate our Community Standards.

Keeping You Safe

Overview: We remove content, disable accounts, and work with law enforcement when we believe there is a genuine risk of physical harm or direct threats to public safety. Learn more about how Facebook handles abusive content. [Links to categories below via sidebar]

Direct Threats: How we help people who feel threatened by others on Facebook. We carefully review reports of threatening language to identify serious threats of harm to public and personal safety. We remove credible threats of physical harm to individuals. We also remove specific threats of theft, vandalism, or other financial harm. We may consider things like a person's physical location or public visibility in determining whether a threat is credible. We may assume credibility of any threats to people living in violent and unstable regions.

Self-Injury: How we work to help prevent self-injury and suicide.

We don't allow the promotion of self-injury or suicide. We work with organizations around the world to provide assistance for people in distress. We prohibit content that promotes or encourages suicide or any other type of self-injury, including self-mutilation and eating disorders. We don't consider body modification to be self-injury. We also remove any content that identifies victims or survivors of self-injury or suicide and targets them for attack, either seriously or humorously. People can, however, share information about self-injury and suicide that does not promote these things.

Dangerous Organizations: What types of organizations we prohibit on Facebook.

We don't allow any organizations that are engaged in the following to have a presence on Facebook: Terrorist activity, or Organized criminal activity.

We also remove content that expresses support for groups that are involved in the violent or criminal behavior mentioned above. Supporting or praising leaders of those same organizations, or condoning their violent activities, is not allowed. We welcome broad discussion and social commentary on these general subjects, but ask that people show sensitivity towards victims of violence and discrimination.

Bullying and Harassment: How we respond to bullying and harassment.

We don't tolerate bullying or harassment. We allow you to speak freely on matters and people of public interest, but remove content that appears to purposefully target private individuals with the intention of degrading or shaming them. This content includes, but is not limited to:

- Pages that identify and shame private individuals,
- Images altered to degrade private individuals,
- Photos or videos of physical bullying posted to shame the victim,
- Sharing personal information to blackmail or harass people, and
- Repeatedly targeting other people with unwanted friend requests or messages.

We define private individuals as people who have neither gained news attention nor the interest of the public, by way of their actions or public profession.

Attacks on Public Figures: What protection public figures receive on Facebook.

We permit open and critical discussion of people who are featured in the news or have a large public audience based on their profession or chosen activities. We remove credible threats to public figures, as well as hate speech directed at them – just as we do for private individuals.

Criminal Activity: How we handle reports of criminal activity on Facebook.

We prohibit the use of Facebook to facilitate or organize criminal activity that causes physical harm to people, businesses or animals, or financial damage to people or businesses. We work with law enforcement when we believe there is a genuine risk of physical harm or direct threats to public safety.

We also prohibit you from celebrating any crimes you've committed. We do, however, allow people to debate or advocate for the legality of criminal activities, as well as address them in a humorous or satirical way.

Sexual Violence and Exploitation: How we fight sexual violence and exploitation on Facebook.

We remove content that threatens or promotes sexual violence or exploitation. This includes the sexual exploitation of minors, and sexual assault. To protect victims and survivors, we also remove photographs or videos depicting incidents of sexual violence and images shared in revenge or without permissions from the people in the images.

Our definition of sexual exploitation includes solicitation of sexual material, any sexual content involving minors, threats to share intimate images, and offers of sexual services. Where appropriate, we refer this content to law enforcement. Offers of sexual services include prostitution, escort services, sexual massages, and filmed sexual activity.

Regulated Goods

We prohibit any attempts by unauthorized dealers to purchase, sell, or trade prescription drugs and marijuana. If you post an offer to purchase or sell firearms, alcohol, tobacco, or adult products, we expect you to comply with all applicable laws and carefully consider the audience for that content. We do not allow you to use Facebook's payment tools to sell or purchase regulated goods on our platform.

Encouraging Respectful Behavior

Overview: People use Facebook to share their experiences and to raise awareness about issues that are important to them. This means that you may encounter opinions that are different from yours, which we believe can lead to important conversations about difficult topics. To help balance the needs, safety, and interests of a diverse community, however, we may remove certain kinds of sensitive content or limit the audience that sees it. Learn more about how we do that here. [Links to categories below via sidebar]

Nudity: People sometimes share content containing nudity for reasons like awareness campaigns or artistic projects. We restrict the display of nudity because some audiences within our global community may be sensitive to this type of content - particularly because of their cultural background or age. In order to treat people fairly and respond to reports quickly, it is essential that we have policies in place that our global teams can apply uniformly and easily when reviewing content. As a result, our policies can sometimes be more blunt than we would like and restrict content shared for legitimate purposes. We are always working to get better at evaluating this content and enforcing our standards.

We remove photographs of people displaying genitals or focusing in on fully exposed buttocks. We also restrict some images of female breasts if they include the nipple, but we always allow photos of women actively engaged in breastfeeding or showing breasts with post-mastectomy scarring. We also allow photographs of paintings, sculptures, and other art that depicts nude figures. Restrictions on the display of both nudity and sexual activity also apply to digitally created content unless the content is posted for educational, humorous, or satirical purposes. Explicit images of sexual intercourse are prohibited. Descriptions of sexual acts that go into vivid detail may also be removed.

Hate Speech: Facebook removes hate speech, which includes content that directly attacks people based on their: Race; Ethnicity; National origin; Religious affiliation; Sexual orientation; Sex, gender, or gender identity; or Serious disabilities or diseases.

Organizations and people dedicated to promoting hatred against these protected groups are not allowed a presence on Facebook. As with all of our standards, we rely on our community to report this content to us.

People can use Facebook to challenge ideas, institutions, and practices. Such discussion can promote debate and greater understanding. Sometimes people share content containing someone else's hate speech for the purpose of raising awareness or educating others about that hate speech. When this is the case, we expect people to clearly indicate their purpose, which helps us better understand why they shared that content.

We allow humor, satire, or social commentary related to these topics, and we believe that when people use their authentic identity, they are more responsible when they share this kind of commentary. For that reason, we ask that Page owners associate their name and Facebook Profile with any content that is insensitive, even if that content does not violate our policies. As always, we urge people to be conscious of their audience when sharing this type of content.

While we work hard to remove hate speech, we also give you tools to avoid distasteful or offensive content. Learn more about the tools we offer to control what you see. You can also use Facebook to speak up and educate the community around you. Counter-speech in the form of accurate information and alternative viewpoints can help create a safer and more respectful environment.

Violence and Graphic Content: Facebook has long been a place where people share their experiences and raise awareness about important issues. Sometimes, those experiences and issues involve violence and graphic images of public interest or concern, such as human rights abuses or acts of terrorism. In many instances, when people share this type of content, they are condemning it or raising awareness about it. We remove graphic images when they are shared for sadistic pleasure or to celebrate or glorify violence.

When people share anything on Facebook, we expect that they will share it responsibly, including carefully choosing who will see that content. We also ask that people warn their audience about what they are about to see if it includes graphic violence.

Keeping Your Account and Personal Information Secure

[…]*

Using Your Authentic Identity: How Facebook's real name requirement creates a safer environment.

People connect on Facebook using their authentic identities. When people stand behind their opinions and actions with their authentic name and reputation, our community is more accountable. If we discover that you have multiple personal profiles, we may ask you to close the additional profiles. We also remove any profiles that impersonate other people.

If you want to create a presence on Facebook for your pet, organization, favorite movie, games character, or another purpose, please create a Page instead of a Facebook Profile. Pages can help you conduct business, reach out to fans, or promote a cause you care about.

[…]*

*Denotes sections not pertinent to standards for users' speech

Reporting Abuse

Our global community is growing every day and we strive to welcome people to an environment free from abusive content. To do this, we rely on people like you. If you see something on Facebook that you believe violates our terms, please report it to us. We have dedicated teams working around the world to review things you report to help make sure Facebook remains safe.

Governments also sometimes ask us to remove content that violates local laws, but does not violate our Community Standards. If after careful legal review, we find that the content is illegal under local law, then we may make it unavailable only in the relevant country or territory.

Please keep the following in mind:

- We may take action any time something violates the Community Standards outlined here.
- We may ask Page owners to associate their name and Facebook Profile with a Page that contains cruel and insensitive content, even if that content does not violate our policies.
- Reporting something doesn't guarantee that it will be removed because it may not violate our policies.
- Our content reviewers will look to you for information about why a post may violate our policies. If you report content, please tell us why the content should be removed (e.g., is it nudity or hate speech?) so that we can send it to the right person for review.
- Our review decisions may occasionally change after receiving additional context about specific posts or after seeing new, violating content appearing on a Page or Facebook Profile.
- The number of reports does not impact whether something will be removed. We never remove content simply because it has been reported a number of times.
- The consequences for violating our Community Standards vary depending on the severity of the violation and the person's history on Facebook. For instance, we may warn someone for a first violation, but if we continue to see further violations we may restrict a person's ability to post on Facebook or ban the person from Facebook.

Not all disagreeable or disturbing content violates our Community Standards. For this reason, we offer you the ability to customize and control what you see by unfollowing, blocking, and hiding the posts, people, Pages, and applications you don’t want to see – and we encourage you to use these controls to better personalize your experience. Learn more. People also often resolve issues they have about a piece of content by simply reaching out to the person who posted it. We’ve created tools for you to communicate directly with other people when you’re unhappy with posts, photos, or other content you see on Facebook.