THE VIRTUES OF MODERATION

James Grimmelmann†

17 YALE J.L. & TECH. 42 (2015)

ABSTRACT

TL;DR—On a Friday in 2005, the Los Angeles Times launched an experiment: a "wikitorial" on the Iraq War that any of the paper's readers could edit. By Sunday, the experiment had ended in abject failure: vandals overran it with crude profanity and graphic pornography. The wikitorial took its inspiration and its technology from Wikipedia, but missed something essential about how the "the free encyclopedia that anyone can edit" staves off abuse while maintaining its core commitment to open participation.

The difference is moderation*: the governance mechanisms that structure participation in a community to facilitate cooperation and prevent abuse. Town meetings have moderators, and so do online communities. A community's moderators can promote posts or hide them, honor posters or shame them, recruit users or ban them. Their decisions influence what is seen, what is valued, what is said. They create the conditions under which cooperation is possible.*

This Article provides a novel taxonomy of moderation in online communities. It breaks down the basic verbs of moderation—exclusion, pricing, organizing, and norm-setting—and shows how they help communities walk the tightrope between the chaos of too much freedom and the sterility of too much control. Scholars studying the commons can learn from moderation, and so can policy-makers debating the regulation of online communities.

---

† Professor of Law, University of Maryland Francis King Carey School of Law. My thanks for their comments to Aislinn Black, BJ Ard, Jack Balkin, Shyam Balganesh, Nicholas Bramble, Danielle Citron, Anne Huang, Matt Haughey, Sarah Jeong, Amy Kapczynski, David Krinsky, Chris Riley, Henry Smith, Jessamyn West, Steven Wu, and participants in the 2007 Commons Theory Workshop for Young Scholars at the Max Planck Institute for the Study of Collective Goods, the 2007 Intellectual Property Scholars Conference, the 2007 Telecommunications Policy Research Conference, and the 2014 Free Expression Scholars Conference.

TABLE OF CONTENTS

> Building a community is pretty tough; it requires just the right combination of technology and rules and people. And while it's been clear that communities are at the core of many of the most interesting things on the Internet, we're still at the very early stages of understanding what it is that makes them work.
>
> –Aaron Swartz[1]

INTRODUCTION

If you've never seen the image known as "goatse," trust me—you don't want to.[2] But if you have, you understand why it was such a disaster when this notoriously disgusting photograph showed up on the website of the *Los Angeles Times* on June 19, 2005.[3] It wasn't a hack. The newspaper had invited its readers to post whatever they wanted. One of them posted a gaping anus.

It had started off innocently enough. Inspired by Wikipedia, the *Times* launched a "wikitorial," an editorial that any of the paper's readers could edit.[4] At first, readers fought over its position: should it be for or against the Iraq War?[5] Then one boiled the argument down to its essence—"Fuck USA"—touching off an edit war of increasingly rapid and radically incompatible changes.[6] By the second day, trolls were posting hardcore pornography, designed to shock and disgust.[7] The *Times* pulled the plug entirely in less than forty-eight hours.[8] What had started with "Rewrite the editorial yourself"[9] ended

---

1 Aaron Swartz, *Making More Wikipedias*, RAW THOUGHT (Sept. 14, 2006), http://www.aaronsw.com/weblog/morewikipedias [http://perma.cc/U2LR-CDTB].

2 The image, which has circulated on the Internet since 1999, depicts a man exposing himself to the camera in a particularly graphic and unpleasant way. In its heyday, goatse was most often used for its shock value: direct people to a website containing it, and revel in their horror. *See* Adrian Chen, *Finding Goatse: The Mystery Man Behind the Most Disturbing Internet Meme in History*, GAWKER, Apr. 10, 2012, http://gawker.com/finding-goatse-the-mystery-man-behind-the-most-disturb-5899787 [http://perma.cc/6RJ8-WVAW].

3 *See, e.g.*, Dan Glaister, *LA Times 'Wikitorial' Gives Editors Red Face*, THE GUARDIAN, June 21, 2005, http://www.theguardian.com/technology/2005/jun/22/media.pressandpublishing [http://perma.cc/NY5A-3A83].

4 *A Wiki for Your Thoughts*, L.A. TIMES, June 17, 2005, http://www.latimes.com/news/la-ed-wiki17jun17-story.html [http://perma.cc/4QW8-RH7C].

5 Glaister, *supra* note 3.

6 *Id.*

7 *Id.*

8 James Rainey, *'Wikitorial' Pulled Due to Vandalism*, L.A. TIMES, June 21, 2005, http://articles.latimes.com/2005/jun/21/nation/na-wiki21 [http://perma.cc/TJ2J-AD7S].

9 *A Wiki for Your Thoughts*, *supra* note 4.

with the admission that "a few readers were flooding the site with inappropriate material."[10]

The wikitorial debacle has the air of a parable: the *Los Angeles Times* hung a "KICK ME" sign on its website, and of course it got kicked. Open up an online community, and of course you'll bring out the spammers, the vandals, and the trolls. That's just how people act on the Internet. But consider this: the *Times*' model, Wikipedia, is going into its thirteenth year.[11] It is the sixth most-visited website on the Internet.[12] And despite being a website "that anyone can edit," it remains almost entirely goatse-free.[13] Anarchy on the Internet is not inevitable. Spaces can and do flourish where people collaborate and where all are welcome. What, then, separates the Wikipedias from the wikitorials? Why do some communities thrive while others become ghost towns?

The difference is moderation. Just as town meetings and debates have moderators who keep the discussion civil and productive,[14] healthy online communities have moderators who facilitate communication. A community's moderators can promote posts or hide them, honor posters or shame them, recruit users or ban them. Their decisions influence what is seen, what is valued, what is said. When they do their job right, they create the conditions under which cooperation is possible. Wikipedia, for all its faults, is moderated in a way that supports an active community of mostly productive editors. The *Los Angeles Times*, for all its good intentions, moderated the wikitorial in a way that provided few useful defenses against vandals. Wikipedia's moderation keeps its house in order; the *Times* gave arsonists the run of the place.

This Article is a guided tour of moderation for legal scholars. It synthesizes the accumulated insights of four groups of experts who have given the problem of moderation their careful and sustained attention. The first is moderators themselves—those who are entrusted with the care and feeding of online

---

10 Rainey, *supra* note 8.

11 *See generally* ANDREW LIH, THE WIKIPEDIA REVOLUTION (2009).

12 *See Top Sites*, ALEXA, http://www.alexa.com/topsites [http://perma.cc/36H3-9STW] (last visited Mar. 30, 2015); *see also Wikipedia: Statistics*, WIKIPEDIA, http://en.wikipedia.org/wiki/Wikipedia:Statistics#pageviews [http://perma.cc/HW25-U4WS] (last visited Jan. 20, 2015) (reporting 4,841,082 articles in the English-language version).

13 *But see goatse.cx*, WIKIPEDIA, http://en.wikipedia.org/wiki/Goatse.cx [http://perma.cc/7YQD-EBGH] (last visited Feb. 23, 2015) (telling rather than showing).

14 *See, e.g.*, ALEXANDER MEIKLEJOHN, FREE SPEECH AND ITS RELATION TO SELF-GOVERNMENT 25-26 (1948) ("[A]t a town meeting . . . [n]o competent moderator would tolerate . . . wasting . . . the time available for free discussion," but "no suggestion of policy shall be denied a hearing because it is on one side of the issue rather than another.").

communities. They have written at length about helpful interventions and harmful ones, giving guidelines and rules of thumb for nudging users towards collaborative engagement.[15] A second group, the software and interface designers who are responsible for the technical substrate on which online communities run, works closely with the first (indeed, they are often the same people). Their own professional literature offers a nuanced understanding of how the technical design of a social space influences the interactions that take place there.[16] The third group consists of academics from a wide variety of disciplines—psychology, communications, and computer science, to name just a few—who have turned a scholarly eye on the factors that make communities thrive or wither.[17] The fourth is

---

[15] *See generally* JONO BACON, THE ART OF COMMUNITY: BUILDING THE NEW AGE OF PARTICIPATION (2d ed. 2012); AMY JO KIM, COMMUNITY BUILDING ON THE WEB (2000); DEBORAH NG, ONLINE COMMUNITY MANAGEMENT FOR DUMMIES (2011); DEREK POWAZEK, DESIGN FOR COMMUNITY (2001); JENNY PREECE, ONLINE COMMUNITIES: DESIGNING USABILITY, SUPPORTING SOCIABILITY (2000).

[16] *See generally* GAVIN BELL, BUILDING SOCIAL WEB APPLICATIONS (2009); CHRISTIAN CRUMLISH & ERIN MALONE, DESIGNING SOCIAL INTERFACES (2009); F. RANDALL FARMER & BRYCE GLASS, BUILDING WEB REPUTATION SYSTEMS (2010); JENIFER TIDWELL, DESIGNING INTERFACES (2d ed. 2010). A particularly fruitful trend in this literature consists of *pattern languages*: interlocking networks of design elements that have repeatedly proven their worth. The idea of pattern languages comes from the work of the architectural theorist Christopher Alexander. *See, e.g.*, CHRISTOPHER ALEXANDER, THE TIMELESS WAY OF BUILDING (1979) (presenting a theory of patterns); CHRISTOPHER ALEXANDER ET AL., A PATTERN LANGUAGE: TOWNS, BUILDINGS, CONSTRUCTION (1977) (developing pattern language for architecture). Software designers took his idea of a pattern as "a careful description of a perennial solution to a recurring problem within a building context," *Aims & Goals*, PATTERNLANGUAGE.COM, http://www.patternlanguage.com/aims/intro.html [http://perma.cc/9BE6-BM4A], and generalized it to technical problems in computer system design. *See, e.g.*, ERICH GAMMA ET AL., DESIGN PATTERNS: ELEMENTS OF REUSABLE OBJECT-ORIENTED SOFTWARE (1994); RICHARD P. GABRIEL, PATTERNS OF SOFTWARE: TALES FROM THE SOFTWARE COMMUNITY (1996). From there, it was only a small step to develop patterns describing how people use software; indeed, these interaction patterns come closest to Alexander's goal of finding patterns that make "towns and buildings . . . able to come alive." ALEXANDER, A PATTERN LANGUAGE, *supra*, at x. Notable examples of pattern languages for social interactions using software include MEATBALLWIKI, http://meatballwiki.org/wiki [http://perma.cc/9RUZ-YZNK]; YAHOO DESIGN PATTERN LIBRARY, https://developer.yahoo.com/ypatterns/ [https://perma.cc/RAZ6-N4XM]; and ONLINE MODERATION STRATEGIES, https://web.archive.org/web/20070419071423/http://social.itp.nyu.edu/shirky/wiki [https://perma.cc/NWZ2-WM5L]. This Article uses a different analytical structure to describe moderation, but the themes of these pattern languages inform the thinking behind it.

[17] For an outstanding synthesis of the literature, see ROBERT E. KRAUT & PAUL RESNICK, BUILDING SUCCESSFUL ONLINE COMMUNITIES: EVIDENCE-BASED SOCIAL DESIGN (2012).

made up of journalists who cover the online beat by embedding themselves in communities (often in moments of high drama).[18]

The Article draws on these various sources to present a novel taxonomy of moderation. The taxonomy takes the form of a grammar—a set of nouns, verbs, adverbs, and adjectives suitable for describing the vast array of moderation techniques in common use on the Internet. The Article describes these techniques in terms of familiar jurisprudential categories such as *ex ante* versus *ex post* and norms versus architecture. This richer understanding of moderation should be useful to scholars and regulators in two ways. One is theoretical: well-moderated online communities catalyze human cooperation. Studying them can provide insights into the management of common-pool resources and the creation of information goods, two problems moderation must solve simultaneously. Studying online communities is thus like studying fisheries or fan fiction—a way to understand society. The other payoff is practical. Many laws either regulate the activities of online communities or exempt them from regulation. The wisdom of these choices depends on empirical facts about the value and power of moderation. Regulators cannot properly evaluate these laws without paying close attention to how moderation plays out on the ground.

Part I of the Article provides basic definitions and describes the dual commons problems that online communities confront. Part II supplies the detailed grammar of moderation, liberally annotated with examples. Part III presents four case studies of moderation in action: Wikipedia, the *Los Angeles Times* wikitorial, MetaFilter, and Reddit. Part IV offers some lessons for regulators by examining the two most important statutes that regulate moderation: § 230 of the Communications Decency Act, and § 512 of the Copyright Act. Part V concludes.

I. The Problem of Moderation

By "moderation," I mean *the governance mechanisms that structure participation in a community to facilitate cooperation and prevent abuse*. Part II will explain how moderation works; this Part lays the foundation by describing the problems it must solve. Section A supplies some basic definitions and details the motivations of community members; Section B describes the goals of good moderation; Section C explains why moderation must confront not one, but two commons problems;

---

[18] Examples will appear throughout the Article, but a good starting point would be Adrian Chen's work. *See, e.g.*, Adrian Chen, *The Laborers Who Keep Dick Pics and Beheadings out of Your Facebook Feed*, WIRED, Oct. 23, 2014, http://www.wired.com/2014/10/content-moderation [http://perma.cc/FJK6-B9SC].

and Section D provides a typology of the abuses against which moderation guards.

A. *Definitions*

Our object of study is an ***online community***.[19] A community can be as small as the handful of people on a private mailing list or as large as the Internet itself. Communities can overlap, as anyone on both Twitter and Facebook knows. Communities can also nest: the comments section at Instapundit is a meaningful community, and so is the conservative blogosphere. There is little point in being overly precise about any given community's boundaries, so long as we can identify three things: the community's ***members***, the ***content*** they share with each other, and the ***infrastructure*** they use to share it.[20] The Internet as a whole is both an agglomeration of numerous communities and a sprawling, loosely knit community in its own right. Its moderation includes both the moderation within its constituent communities and moderation that cannot easily be attributed to any of them. Thus, even though it is not particularly helpful to talk about Google as a community in its own right,[21] it and other search engines play an important role in the overall moderation of the Web.[22]

Members can wear different hats: there are ***owners*** of the infrastructure, ***moderators*** of the community, and ***authors*** and ***readers*** of content. For example, on YouTube, Google owns the infrastructure; video uploaders are authors; video viewers are readers; and the moderators include everyone who clicks to flag an inappropriate video,[23] the algorithms that collate user re-

---

[19] The defined terms that make up the vocabulary of moderation will be written in ***bolded italics*** when they make their first appearances in the Article.

[20] These are virtual communities, defined by a shared virtual place rather than by shared geography, meaning, or practice. *See generally* HOWARD RHEINGOLD, THE VIRTUAL COMMUNITY (1993); Quinn Warnick, What We Talk About When We Talk About Talking: Ethos at Work in an Online Community (2010) (unpublished Ph.D. dissertation, Iowa State University), http://lib.dr.iastate.edu/cgi/viewcontent.cgi?article=2480&context=etd [http://perma.cc/P9JK-HY2P]; *see also* PREECE, *supra* note 15, at 10-17.

[21] The problem is that there is not a close nexus between Google's users, the content it indexes, and the infrastructure in Google's server farms. Most of the websites whose content appears on Google are Google "users" only in a very loose sense, and they bring their own server infrastructure to the table. There is interesting moderation here, but "Google" is the wrong level of generality for identifying the community that the moderation affects.

[22] *See generally* James Grimmelmann, *Speech Engines*, 98 MINN. L. REV. 868, 893-96 (2014) (discussing the role of search engines in organizing the Internet).

[23] *See* Alistair Barr & Lisa Fleisher, *YouTube Enlists 'Trusted Flaggers' to Police Videos*, WALL ST. J., Mar. 17, 2014, http://blogs.wsj.com/digits/2014/03/17/youtube-enlists-trusted-flaggers-to-police-videos [http://perma.cc/Z6HY-RYKU].

ports, and the unlucky YouTube employees who manually review flagged videos.[24] Owners occupy a privileged position because their control over infrastructure gives them unappealable control over the community's software-based rules.[25] This control lets owners decide who can moderate and how. Moderators, in turn, shape the flow of content from authors to readers. Of course, members can wear multiple hats. "NO SPOILERS!" is both content and a gently chiding act of moderation.

Members have varied motivations.[26] Authors want their messages to be seen;[27] readers with diverse tastes seek content of interest to them.[28] Moderators, like authors, want to promote the spread of content they care about.[29] All of them can derive personal fulfillment and a sense of belonging from participation. On the other side of the ledger, these activities take time and effort. And where money changes hands, members would naturally prefer to be paid rather than to pay. YouTube is popular with video makers in part because it pays them a share of advertising revenue rather than charging them to host their videos.[30]

Because the same person could be an author, reader, moderator, and owner, these motivations interrelate. Thus, for example, users connect their computers to peer-to-peer networks to download files they want, but in the process they make files on their computers available to other users. [31] They are willing to act as owners supplying infrastructure because of the value they receive as readers receiving content. Similarly, participants on a discussion forum may shoulder some of the work of

---

[24] *See* Brad Stone, *Policing the Web's Lurid Precincts*, N.Y. TIMES, July 18, 2010, http://www.nytimes.com/2010/07/19/technology/19screen.html [http://perma.cc/Y493-7VKF].

[25] *See* James Grimmelmann, *Anarchy, Status Updates, and Utopia*, 35 PACE L. REV. (forthcoming 2015), http://papers.ssrn.com/sol3/papers.cfm?abstract_id=2358627 [http://perma.cc/4R29-AJC3] [hereinafter Grimmelmann, *Anarchy*].

[26] *See generally* FARMER & GLASS, *supra* note 16, at 111-20 (describing various user motivations).

[27] *See, e.g.*, *Using a Zip Code Puts You Under Military Rule According to Supreme Court*, YOUTUBE (June 5, 2014), https://www.youtube.com/watch?v=6TEOyp1ERVc [https://perma.cc/JC3J-A8AZ].

[28] *See, e.g.*, *[Search Results for] ASMR*, YOUTUBE, https://www.youtube.com/results?search_query=ASMR (last visited Mar. 19 2015) [https://perma.cc/9X65-RX5Y].

[29] *See, e.g.*, *What Does It Mean to "Like" Something?*, FACEBOOK, https://www.facebook.com/help/110920455663362 [https://perma.cc/FRY7-H27R] ("Clicking Like below a post on Facebook is an easy way to let people know that you enjoy it.") .

[30] *See, e.g.*, *What is the YouTube Partner Program?*, YOUTUBE, https://support.google.com/youtube/answer/72851 [https://perma.cc/3KCZ-QWHW].

[31] *See* Lior Jacob Strahilevitz, *Charismatic Code, Social Norms, and the Emergence of Cooperation on the File-Swapping Networks*, 89 VA. L. REV. 505 (2003).

moderation by flagging unwanted posts for deletion because they enjoy being part of a thriving community. Divergent motivations become important only when there is a clear separation of roles (e.g., paid professional moderators) or when a community is torn between participants with incompatible goals (e.g., amateur and professional photographers).

B. *Goals*

From these individual motivations, we can derive goals for moderation overall. Broadly speaking, moderation has three goals. First, a well-moderated community will be ***productive***: it will generate and distribute valuable information goods. Some of these information goods are valuable in themselves (*Welcome to Night Vale* fan fiction), others because they facilitate transactions (Freecycle listings), and others because they are part of socially important systems (political discussions). Productivity is the greatest common divisor of moderation goals, the one that everyone can agree on. Members share in the gains from productivity as authors and readers. Society gains, too, when valuable information spreads beyond the community—a classic example of a positive spillover.[32]

Second, moderation can increase access to online communities. ***Openness*** is partly about efficiency: more members can make the community more productive. But openness also has moral consequences: cutting people off from a community cuts them off from the knowledge the community produces.[33] Openness exists along a spectrum. A wiki usable by anyone on the Internet is more open than a wiki open to anyone on a school's network, which is in turn more open than a password-protected wiki open only to the graduate students of the geology department. An important aspect of openness is democracy—participation in moderation and in setting moderation policy. Again, part of the justification is instrumental: broad participation can help make moderation more effective.[34] But it can also be important in itself for members to have a voice in making moderation decisions. Democratic moderation is online self-governance.[35]

---

[32] *See generally* Brett M. Frischmann & Mark A. Lemley, *Spillovers*, 107 COLUM. L. REV. 257 (2007).

[33] *See generally* JOHN WILLINSKY, THE ACCESS PRINCIPLE: THE CASE FOR OPEN ACCESS TO RESEARCH AND SCHOLARSHIP (2005).

[34] *See*, *e.g.*, KRAUT & RESNICK, *supra* note 17, at 151-52.

[35] *See, e.g.*, David R. Johnson, David G. Post & Marc Rotenberg, *Governing Online Spaces: Virtual Representation*, VOLOKH CONSPIRACY (Jan. 3, 2013), http://www.volokh.com/2013/01/03/facebook-governance-and-virtual-representation [http://perma.cc/9DTM-FJTP] ("[A]ll users have a right to participate in the processes through which the rules by which they will be bound are made.").

Third, a well-moderated community will have low ***costs***: it will do its work while making as few demands as possible on the infrastructure and on participants. Costs here include the obvious computational ones—servers, hard drives, network connections, electricity, etc.—but also include the work required of participants, such as flagging a post for removal, removing a flagged post, or appealing an incorrectly removed post. Each individual decision may be small, but they add up quickly. Yahoo saved one million dollars per year in customer support costs by substantially automating its moderation system for Yahoo Answers.[36]

These virtues are incomparable. Different moderation techniques inevitably trade off among them. Excluding the heaviest users, for example, hurts productivity and openness while also reducing costs. Even productivity and cost, both efficiency concerns, have distributional components: two members may agree that a burden is worth bearing but disagree on who should bear it.

C. *Commons Problems*

One tension in particular animates the entire problem of moderation. Online communities have a commons problem.[37] In fact, they have two. On the one hand, they depend on shared infrastructure with limited capacity. Hard drives don't grow on trees. Members must collectively limit their use of infrastructure to keep this common-pool resource from collapsing.[38] On the other hand, online communities trade in information that can potentially be shared without limit, so members must collectively catalyze themselves into creating and sharing.[39] Solv-

---

[36] *See* FARMER & GLASS, *supra* note 16, at 243-77.

[37] This account draws heavily on James Grimmelmann, *The Internet Is a Semicommons*, 78 FORDHAM L. REV. 2799 (2010) [hereinafter Grimmelmann, *Semicommons*], which provides a more extensive exposition and literature review.

[38] *Id*. at 2806-10 (reviewing literature). A conventional understanding of commonly held resources, as captured in Garrett Hardin, *The Tragedy of the Commons*, 162 SCI. 1243 (1968), was that without external "mutual coercion, mutually agreed upon," *id.* at 1247, exhaustion through overuse was inevitable. Commons theorists, led by Elinor Ostrom, showed that under the right conditions a community of users could itself collectively moderate its use of a commonly held resource. *See* ELINOR OSTROM, GOVERNING THE COMMONS: THE EVOLUTION OF INSTITUTIONS FOR COLLECTIVE ACTION (1990).

[39] Grimmelmann, *Semicommons*, *supra* note 37, at 2810-15 (reviewing literature). Again, a conventional account emphasized the need for external restraints, such as intellectual property laws. *See, e.g.*, WILLIAM M. LANDES & RICHARD A. POSNER, THE ECONOMIC STRUCTURE OF INTELLECTUAL PROPERTY LAW 19-21 (2003) (providing a conventional account of intellectual property). Here, the counter-movement showed that some creative communities could self-regulate effectively and also that the absence of restraints could itself catalyze creativity. *See, e.g.*, KAL RAUSTIALA & CHRIS

ing both problems at once is particularly tricky because the most natural way to protect infrastructure is to discourage intensive use by limiting access, while the most natural way to promote the sharing of information is to encourage extensive use by opening up access.[40] Moderation is how online communities walk the tightrope between overuse and underuse.[41]

In previous work, I described the Internet as a semicommons—a resource that is owned and managed as private property at one level but as a commons at another, and in which "both common and private uses are important and impact significantly on each other."[42] The semicommons concept captures both the costs that authors and readers can impose on owners through overuse and the ways that owners can inhibit content-sharing uses by leveraging control of the infrastructure.[43] It also directs attention to moderation techniques that allow productive coexistence.[44] The emphasis there is on the Internet as a whole, but the same problems—and similar solutions—recur in smaller online communities.[45]

Brett Frischmann's theory of infrastructure also cleanly describes online communities. Indeed, I have borrowed the term because the fit is so precise.[46] To Frischmann, an infrastructural resource satisfies three criteria:

> (1) The resource may be consumed nonrivalrously for some appreciable range of demand.

SPRIGMAN, THE KNOCKOFF ECONOMY: HOW IMITATION SPARKS INNOVATION (2012); Yochai Benkler, *Coase's Penguin, or Linux and the Nature of the Firm*, 112 YALE L.J. 369 (2002); Jessica Litman, *The Public Domain*, 39 EMORY L.J. 965 (1990).

40 *See* Yochai Benkler, *Commons and Growth: The Essential Role of Open Commons in Market Economies*, 80 U. CHI. L. REV. 1499, 1505-06 (2013); Grimmelmann, *Semicommons*, *supra* note 37, at 2815.

41 *See* Mayo Fuster Morell, *Governance of Online Creation Communities for the Building of Digital Commons*, *in* GOVERNING KNOWLEDGE COMMONS 281 (Brett M. Frischmann, Michael J. Madison & Katherine J. Strandburg eds., 2014) (linking information production, infrastructure use, and community governance).

42 Grimmelmann, *Semicommons*, *supra* note 37, at 2816 (quoting Henry E. Smith, *Semicommon Property Rights and Scattering in the Open Fields*, 29 J. LEGAL STUD. 132 (2000)). *But see* Benkler, *supra* note 40, at 1522-23 (criticizing the application of semicommons theory).

43 Grimmelmann, *Semicommons*, *supra* note 37, at 2817.

44 *Id*. at 2816-18.

45 *Id*. at 2823-41 (giving case studies).

46 BRETT M. FRISCHMANN, INFRASTRUCTURE: THE SOCIAL VALUE OF SHARED RESOURCES (2012). Another early and sophisticated treatment of the nexus between tangible and intangible resources is Carol M. Rose, *The Comedy of the Commons: Custom, Commerce, and Inherently Public Property*, 53 U. CHI. L. REV. 711, 768 (1986). For a recent literature review, see Benkler, *supra* note 40.

> (2) Social demand for the resource is driven primarily by downstream productive activity that requires the resource as an input.
> (3) The resource may be used as an input into a wide range of goods and services . . . .[47]

This account captures the congestible but renewable nature of online infrastructure, the interdependence between infrastructure and content, and the diversity of content. Frischmann argues for managing infrastructure as a commons with nondiscriminatory access rules, subject to nondiscriminatory use restrictions for "securing the commons itself,"[48] an embrace of openness that recognizes the interplay of productivity and cost.

D. *Abuses*

The interface between infrastructure and information is vulnerable to some predictable forms of strategic behavior, including spam, harassment, and other famous pathologies of online life. These are the abuses against which moderation must guard. Moderation need not prevent them entirely—and probably cannot without killing the commons—but it must keep them within acceptable bounds, and without driving up the costs of moderation itself to unacceptable levels.[49] The abuses fall into four broad categories: congestion, cacophony, abuse, and manipulation.

The first pair of problems involves overuse. Each participant's contribution of content makes demands both on the infrastructure and on other participants. At the infrastructure level, overuse causes ***congestion***, which makes it harder for any information to get through and can cause the infrastructure to stagger and fall.[50] At the content level, overuse causes ***cacophony,*** which makes it harder for participants to find what they want. In trademark terms, they must incur search costs to sort through the information available to them.[51] Both congestion and cacophony are problems of prioritization: bad content crowds out good, to the private benefit of the content's promoters but at an overall cost to the community. The difference is that in congestion, the resource constraint is the infrastructure's capacity, whereas in cacophony, the constraint is partici-

---

[47] FRISCHMANN, *supra* note 46, at xiv.

[48] *Id.* at 92; *see also* Henry E. Smith, *Exclusion Versus Governance: Two Strategies for Delineating Property Rights*, 31 J. LEGAL STUD. S453, S454-55 (2002) (differentiating access and use restrictions as strategies for managing resource use).

[49] *See* Henry E. Smith, *Semicommon Property Rights and Scattering in the Open Fields*, 29 J. LEGAL STUD. 132, 141-42 (2000).

[50] *See* FRISCHMANN, *supra* note 46, at 136-58.

[51] *See* James Grimmelmann, *Information Policy for the Library of Babel*, 3 J. BUS. & TECH. L. 29 (2008).

pants' attention.[52] Spam is the classic example of overuse causing both congestion and cacophony.[53] A denial-of-service attack is an attempt to create congestion for its own sake.

Next, there is ***abuse***, in which the community generates negative-value content—information "bads" rather than information goods.[54] Abuse is distinctively a problem of information exchange. The harms it causes are the harms information causes as speech that is understood and acted on by humans. Harassment is the classic example of abuse directed at particular people, while trolling is the classic example of abuse directed at the community in general.[55] In its extreme form, abuse involves an entire community uniting to share content in a way that harms the rest of society, such as trading copyrighted movies pre-release, planning the assassination of doctors who perform abortions, or starting offensive hoaxes.[56]

Finally, there is ***manipulation***, in which ideologically motivated participants try to skew the information available through the community.[57] A forum moderator on a science discussion site who deletes posts from climate scientists while leaving posts from climate change deniers is engaging in manipulation, as is a retailer that games its way to the top of search rankings with link farms and hidden text. The classic pathological case of manipulation is the edit war, in which wiki

---

[52] For an early explanation of this distinction in terms of "exploitation" and "pollution," see Gian Maria Greco & Luciano Floridi, *The Tragedy of the Digital Commons*, 6 ETHICS & INFO. TECH. 73, 76 (2004).

[53] This general definition of overuse emphasizes that spam is a problem hardly confined to email. *See* Grimmelmann, *Semicommons*, *supra* note 37, at 2839 ("[A]ny sufficiently advanced technology is indistinguishable from a spam vector."); *see generally* FINN BRUNTON, SPAM: A SHADOW HISTORY OF THE INTERNET (2013) (cataloging other forms of spam).

[54] *See generally* Danielle Keats Citron & Helen Norton, *Intermediaries and Hate Speech: Fostering Digital Citizenship for Our Information Age*, 91 B.U. L. REV. 1435 (2011).

[55] *See* Lior Jacob Strahilevitz, *Wealth Without Markets?*, 116 YALE L.J. 1472, 1493-97 (2007). A troll "posts deliberately erroneous or antagonistic messages to a newsgroup or similar forum with the intention of eliciting a hostile or corrective response," *Troll*, n.1, OXFORD ENG. DICT. (Draft additions Mar. 2006), http://www.oed.com/view/Entry/206613 [http://perma.cc/4AHN-P3ZL], often with "willful, disingenuous provocation and malicious deceit." David Auerbach, *Anonymity as Culture: Treatise*, TRIPLE CANOPY, http://canopycanopycanopy.com/contents/anonymity_as_culture__treatise [http://perma.cc/3UNF-SJE4].

[56] *See, e.g.*, Caitlin Dewey, *4Chan's Latest, Terrible Prank: Convincing West Africans that Ebola Doctors Actually Worship the Disease*, WASH. POST, Sept. 22, 2014, http://www.washingtonpost.com/news/the-intersect/wp/2014/09/22/4chans-latest-terrible-prank-convincing-west-africans-that-ebola-doctors-actually-worship-the-disease [http://perma.cc/2WUX-DBFJ].

[57] *See* Christopher E. Peterson, User-Generated Censorship: Manipulating the Maps of Social Media (2013) (unpublished M.S. dissertation, Massachusetts Institute of Technology), http://cmsw.mit.edu/user-generated-censorship [http://perma.cc/8TJN-2CVB].

users with conflicting ideologies engage in a wasteful conflict to make a page reflect their point of view. Like abuse, manipulation is distinctively a problem of information exchange: it is possible whenever some information can be deleted entirely, or when participants can exploit each other's cognitive limits. The difference is that, in abuse, the content itself is the problem, while in manipulation, worthwhile content is handled in a way that harms the community. The dueling pro- and anti-war edits to the *Los Angeles Times* wikitorial were manipulation; the pornography that followed was abuse.

II. The Grammar of Moderation

Now that we have seen the problems that moderation faces, we can discuss how it solves them. We have already met the basic definitions: a *community* of *members* who use shared *infrastructure* to exchange *content*. The members have roles: as *owners* of infrastructure, as *authors* and *readers* of content, and as *moderators*. These are the nouns in the grammar of moderation.

Section A of this Part describes the verbs—the four principal techniques of moderation. Two are relatively simple. ***Exclusion*** keeps unwanted members out of the community entirely; ***pricing*** uses market forces to allocate participation. The other two are more complex. In ***organization***, moderators reshape the flow of content from authors to readers; in ***norm-setting***, they inculcate community-serving values in other members. Together, these are the basic tools of moderation.

Section B considers some important distinctions in how moderation is carried out. Each of these distinctions can be applied to any of the moderation techniques to give it different inflections. If the techniques are verbs, these distinctions are the adverbs. First, moderation can be carried out ***manually***, by human moderators making individualized decisions in specific cases, or ***automatically***, by algorithms making uniform decisions in every case matching a specified pattern. Second, moderation can be done ***transparently***, with each decision and its reasoning available for public review, or ***opaquely***, behind the electronic equivalent of closed doors. Third, there is the familiar distinction between regulation ***ex ante*** and regulation ***ex post***—deterrence versus punishment, protection versus repair. And fourth, moderation can be ***centralized*** and carried out by one powerful moderator making global decisions, or ***decentralized*** and carried out by many dispersed moderators making local decisions.

Section C then examines some underlying community characteristics that can significantly influence the success or failure of the different moderation techniques. These are adjectives: they modify the nouns (especially "community") in the gram-

mar of moderation. Sometimes they are set by the community at large, while at other times they are under the control of moderators or regulators. First, there is the ***capacity*** of the infrastructure. Greater capacity comes at a higher cost but is less prone to congestion. Second, there is the ***size*** of the community. Larger communities can engage in broader sharing but are less cohesive. Third, ownership of the infrastructure may be more or less ***concentrated***, which affects the distribution of power among members and hence their influence over moderation decisions. Fourth, members may be more or less ***identified*** within the community: rich identities enhance trust and cooperation but can also be a barrier to participation.

This is a rich, complicated taxonomy. Its subtleties are not to be grasped on this first, abbreviated glance. This is just the map, the outline whose broad contours we will now fill in.

A. *Techniques (Verbs)*

The real study of moderation begins with the verbs of moderation—the basic actions that moderators can take to affect the dynamics of a community. There are four: excluding, pricing, organizing, and norm-setting.[58]

1. Excluding

Exclusion is fundamental in property theory because of its simplicity.[59] Rather than attempt to calibrate specific uses, one simply excludes outsiders from all uses.[60] In an online community, exclusion deprives the community of the contributions that those who are excluded could have made. But that loss can be justified when exclusion inhibits strategic behavior. It can be used against any form of strategic behavior by targeting those users who engage in that behavior—for example, to reduce congestion by excluding known spammers.

The processes used to decide who will be excluded can fall anywhere along the spectrum from highly precise to absurdly crude. Mark Lemley individually vets each subscriber to the CyberProfs mailing list; for a time, Facebook was available only

---

[58] The account given here draws on several strands of legal theory. Foremost among them is Lessig's theory of four modalities of regulation: law, norms, markets, and architecture. *See* Lawrence Lessig, *The Law of the Horse: What Cyberlaw Might Teach*, 113 HARV. L. REV. 501, 507-11 (1999); James Grimmelmann, Note, *Regulation by Software*, 114 YALE L.J. 1719 (2005). Other strands include property theory and commons theory.

[59] Thomas W. Merrill, *Property and the Right to Exclude*, 77 NEB. L. REV. 730 (1998).

[60] *See, e.g.*, Smith, *supra* note 48. Exclusion is therefore an architectural constraint in Lessig's four-modalities taxonomy. Lessig, *supra* note 58. It acts automatically and immediately to prevent non-members from participating. Grimmelmann, *Regulation by Software*, *supra* note 58, at 1723.

to users with a .edu email address.[61] At any level of precision, a particularly important decision is whether the default is inclusion or exclusion. A default of inclusion gives everyone, well-intentioned or not, at least one bite at the apple.[62] Exclusion can also be applied independently to different roles. It is common, for example, to let anyone read a discussion board but to allow only registered users to post.[63]

2. Pricing

Pricing inhibits participation, both good and bad, by raising its costs.[64] Pricing is more information intensive than exclusion because one must set the level of prices.[65] Some prices are explicit, such as World of Warcraft's $14.99 per month subscription fee. Other prices are implicit: Twitter's abuse-reporting process is long and involved, so anyone who wants to report abuse must pay with their time.[66] Advertising is a prevalent form of implicit pricing: readers pay with their time and attention.

Any of the different roles can be priced separately. Authorship is the obvious target to be priced first because of its bandwidth demands.[67] Pricing can be applied at many levels of granularity, from flat-rate all-access passes to microtransactions for each action. At one extreme, prohibitively high prices collapse into *de facto* exclusion. At the other extreme, free is a price, too—one that sends a broadly welcoming signal to potential members.[68] Prices can even be negative, in which case they

---

61 *See* Janet Kornblum, *Facebook Will Soon Be Available to Everyone*, USA TODAY, Sept. 11, 2006, http://usatoday30.usatoday.com/tech/news/2006-09-11-facebook-everyone_x.htm [http://perma.cc/5CVL-WYKA].

62 For an example of why an inward-looking community might nonetheless choose inclusion by default, see Lauren Gelman, *Privacy, Free Speech, and "Blurry-Edged" Social Networks*, 50 B.U. L. REV. 1315 (2009).

63 *See, e.g.*, *Frequently Asked Questions*, METAFILTER, http://faq.metafilter.com/#38 [http://perma.cc/7ZJ4-S9GE].

64 *See* Lessig, *supra* note 58, at 507-08.

65 *See* Smith, *supra* note 48, at S471-72.

66 *See* Mary Anne Franks, *The Many Ways Twitter Is Bad at Responding to Abuse*, THE ATLANTIC, Aug. 14, 2014, http://www.theatlantic.com/technology/archive/2014/08/the-many-ways-twitter-is-bad-at-responding-to-abuse/376100 [http://perma.cc/X8EN-ADQM]. On implicit prices, see generally PREECE, *supra* note 15, at 133-43 (discussing usability factors in online communities).

67 Flickr, for example, offers unlimited access for viewing photos, requires a free account to post up to 1TB of photos, and sells a second TB of storage for $499 a year. The three bands consist of no price, an implicit price, and an explicit price. Overall, authorship is priced higher than readership. *See Free Accounts, Upgrading and Gifts*, FLICKR, https://www.flickr.com/help/limits [https://perma.cc/N9NN-BBW6].

68 *See generally* CHRIS ANDERSON, FREE: THE FUTURE OF A RADICAL PRICE (2009).

provide a subsidy. For example, when it launched, Epinions paid users to write reviews.[69]

As taxes on participation, prices have two basic purposes. One is to raise revenue for the community's use, typically by charging authors and readers to compensate owners (for supplying infrastructure) and professional moderators (for their work). The other is Pigouvian—to make members internalize some of the costs of their behavior.[70] Pricing is naturally well-suited to account for congestion.[71] A per-message email fee, for example, would reduce spam by forcing senders to account for some of the resources spam sucks up.[72] This type of pricing can also induce participants to signal their quality, ideally deterring those who have little to offer the community.[73] A $5, one-time registration fee, small as it is, can provide a substantial deterrent to casual malcontents.[74]

3. Organizing

Organization shapes the flow of content from authors to readers.[75] It is the verb of moderation that most takes advantage of the informational capabilities of computers.[76] Categorizing messages on a bulletin board by topic is organization. So is searching them by keyword, counting the number of messages, or deleting off-topic messages. These are all ways of remixing authors' contributions to give readers a more satisfying experience.

It is helpful to think of organizing techniques as being built up from several basic operations:

---

69 *See* Eric Goldman, *Epinions, The Path-Breaking Website, Is Dead. Some Lessons It Taught Us*, FORBES, Mar. 12, 2014, http://www.forbes.com/sites/ericgoldman/2014/03/12/epinions-the-path-breaking-website-is-dead-some-lessons-it-taught-us [http://perma.cc/M57J-RUVA]. Paying authors and moderators can backfire to the extent that payments crowd out other motivations. *See* YOCHAI BENKLER, THE WEALTH OF NETWORKS 96-97 (2006).

70 *See* KRAUT & RESNICK, *supra* note 17, at 157-58.

71 *See* FRISCHMANN, *supra* note 46, at 146-49.

72 *See, e.g.*, Cynthia Dwork & Moni Naor, *Pricing via Processing or Combatting Junk Mail*, *in* ADVANCES IN CRYPTOLOGY—CRYPTO '92 139 (1993) (implementing pricing by requiring "sender to compute some moderately expensive, but not intractable, function").

73 S*ee* FRISCHMANN, *supra* note 46, at 148.

74 KRAUT & RESNICK, *supra* note 17, at 200; rusty [Rusty Foster], *K5 Becomes "Gated Dysfunctional Community"*, KURO5HIN (Sept. 10, 2007), http://www.kuro5hin.org/story/2007/9/10/13920/3664 [http://perma.cc/X27E-BUUT].

75 In Lessig's taxonomy, organization is another application of architecture. Lessig, *supra* note 58, at 508-09.

76 For a thoughtful catalog of organizational interface patterns, see generally TIDWELL, *supra* note 16, especially chapters 3, 4, 5, and 7.

- *Deletion* is the removal of content. A bulletin board administrator who excises off-topic and profanity-laden posts is engaged in deletion.[77]
- *Editing* is the alteration of content. It ranges from correcting typos to changing the very essence of a post. At the limit, editing is deletion plus authorship: the moderator rejects an author's reality and substitutes her own.
- *Annotation* is the addition of information about content.[78] eBay's feedback system annotates buyers and sellers; Facebook's Likes annotate posts and comments; Amazon's user-written reviews and lists are annotations that have crossed the line and become content in their own right.
- *Synthesis* is the transformative combination of pieces of content. Wikipedia is the ultimate example of synthesis. There, small and heterogeneous changes by individual users are synthesized into entire encyclopedia entries. On a smaller scale, an online poll synthesizes individual votes into totals.
- *Filtering* is deletion's non-destructive cousin: the content is still there, but readers see a specialized subset of it. A search engine filters; so does a blog's list of the ten most recent comments. At the limit, filtering asymptotically approaches deletion: the ten-thousandth search result might as well not exist.
- *Formatting* is the styling of of content for presentation to readers. Good typography improves readability; sensible ordering and grouping of images makes it possible to scan through them quickly.

Like the other verbs, organization is itself costly but can reduce strategic behavior. Organization directly attacks cacophony by helping readers see only the content they want. At the

---

[77] Ephemerality is a species of deletion. Snapchat photos vanish within seconds to provide privacy and engagement. *See* danah boyd, *Why Snapchat Is Valuable: It's All About Attention*, APOPHENIA, Mar. 21, 2014, http://www.zephoria.org/thoughts/archives/2014/03/21/snapchat-attention.html [http://perma.cc/YA2Y-FLKF] ("The underlying message is simple: You've got 7 seconds. PAY ATTENTION. And when people do choose to open a Snap, they actually stop what they're doing and look."). *But see* Snapchat, Inc., F.T.C. 132 3078 (Dec. 31, 2014) (alleging a failure by Snapchat to secure privacy of photos). Ephemerality can be destructive of community. As Sarah Jeong says of Twitter, "Mix ephemerality, disconnectedness, and stable identities, and you get ever-lasting grudges filtered through a game of Telephone." @sarahjeong, TWITTER (Oct. 8, 2014), https://twitter.com/sarahjeong/status/519990219043389440 [https://perma.cc/8F8N-U9FE].

[78] *See* FARMER & GLASS, *supra* note 16, at 39-65, 131-61 (providing a rich taxonomy of annotation systems).

same time, organization indirectly reduces cacophony by reducing the incentives for authors to create low-value content that readers don't want and will never see.[79] Only deletion directly attacks congestion, but all forms of organization have the same indirect effect of reducing the incentive to spam.[80] On the other hand, organization can be a tool for manipulation in the hands of self-interested moderators. Think, for example, of a Judean People's Front sympathizer deleting every mention of the People's Front of Judea on a Roman forum.[81] Finally, depending on how it is used, organization can either greatly amplify or greatly inhibit abuse: compare a gossip site that deletes crude sexual comments with one that invites them.

The real complexity of organization comes when one uses multiple types of organization at once. An email list moderator who deletes some posts and marks others as "important" is simultaneously filtering and annotating. A user who flags an Amazon review as helpful is annotating it. Amazon then synthesizes the flags into totals and filters users' views based on those totals.[82] Wikipedia's Talk pages are annotation applied to the synthesis process.[83] Slashdot's moderation provides an annotative input into filtration (readers can choose to see only highly rated comments), in the process making the annotations *themselves* the subject of "meta-moderation."[84]

Finally, of course, organization interacts with the other verbs. Reddit gives its paid Gold users better filtering tools than regular users.[85] In many communities, those who are flagged by other participants for poor contributions may be

---

79 Filtration hides unwanted content; deletion removes it outright; annotation enables readers to evaluate the content's relevance to them without actually reading it; and synthesis turns several moderate-value contributions into one higher-value one.

80 On the other hand, readers facing lower search costs will increase their consumption, both encouraging them to greater creation of their own and also raising the incentives for authors to contribute. Thus, since organization can increase contribution through one mechanism and deter it through another, the overall impact of effective organization on infrastructure owners' private costs is indeterminate.

81 *See Monty Python's Life of Brian* (HandMade Films 1979).

82 For a detailed visual grammar for describing these multi-stage systems of organization, see FARMER & GLASS, *supra* note 16.

83 *See* DARIUSZ JEMIELNIAK, COMMON KNOWLEDGE? AN ETHNOGRAPHY OF WIKIPEDIA (2014).

84 See Clifford A. Lampe, Ratings Use in an Online Discussion System: The Slashdot Case (2006) (Ph.D. dissertation, University of Michigan), http://deepblue.lib.umich.edu/bitstream/handle/2027.42/39369/lampe_diss_revised.pdf [http://perma.cc/BN9G-REAK].

85 *See Reddit Gold*, REDDIT, http://www.reddit.com/gold/about [http://perma.cc/2X5J-QZU9].

banned—resulting in exclusion based on annotation based on social norms.[86]

4. Norm-Setting

Moderation's biggest challenge and most important mission is to create strong shared norms among participants. Norms can target every form of strategic behavior. For example, if every author refrains from personal attacks, there is no further personal-attack problem to be solved. Beneficial norms, however, cannot simply be set by fiat. By definition, they are an emergent property of social interactions. Moderators have limited power over group norms. Most of the levers they can pull will only nudge norms in one direction or another, possibly unpredictably. Good norm-setting is a classic example of know-how. There are heuristics, but knowing whether to chastise an uncivil user publicly or privately is not a decision that can be made in the abstract.[87] Blogger Jason Kottke summed up the challenges of norm-setting with characteristic verve:

> Punishing the offenders and erasing the graffiti is the easy part . . . [F]ostering "a culture that encourages both personal expression and constructive conversation" is much more difficult. Really fucking hard, in fact . . . it requires near-constant vigilance. If I opened up comments on everything on kottke.org, I could easily employ someone for 8-10 hours per week to keep things clean, facilitate constructive conversation, coaxing troublemakers into becoming productive members of the community, etc. Both MetaFilter and Flickr have dedicated staff to perform such duties . . . I imagine other community sites do as well. If you've been ignoring all of the uncivility on your site for the past 2 years, it's going to be difficult to clean it up. The social patterns of your community's participants, once set down, are difficult to modify in a significant way.[88]

---

[86] *See* Kate Crawford & Tarleton Gillespie, *What Is a Flag for? Social Media Reporting Tools and the Vocabulary of Complaint*, NEW MEDIA & SOC'Y (2014).

[87] *See, e.g.*, NG, *supra* note 15, at 77-83, 94-97; Warnick, *supra* note 20, at 113 ("[A] strong collective ethos—generated by individuals but bigger than any one person—is essential to maintaining a successful online community.").

[88] Jason Kottke, *The Blogger Code*, KOTTKE.ORG (Apr. 9, 2007), http://kottke.org/07/04/the-blogger-code [http://perma.cc/H9US-3V6N].

Some communities depend on shared norms. Discussion groups, for example, are acutely sensitive to group norms. It only takes a few determined spammers or trolls to bring a discussion to a screeching halt.[89] But other communities can prosper even when some norms are widely flouted. Spammers and trolls still abound on the Internet, but they have not yet managed to ruin it for everyone. Google may not be able to make spammers clean up their act, but it can hide their antics.[90] The difference illustrates the two roles that the other verbs of moderation can play. Sometimes, they keep order directly, in the face of bad behavior; at other times, they keep order indirectly, by encouraging good behavior. That is, the other three verbs are both substitutes for and sources of norms, and communities vary in the balance they strike between these two roles.

Moderators can influence norms directly by articulating them. They can do this either in general, with codes of conduct and other broad statements of rules, or in specific cases by praising good behavior and criticizing bad. The difference is the difference between "Don't post images containing nudity" and "This post has been deleted because it contained nudity." Note, however, that stating a norm does not automatically promote it. There is empirical evidence that, in some circumstances, expressing a norm about user behavior can induce exactly the opposite response.[91]

Moderators can also influence norms indirectly, through the other verbs.[92] A list of "new and noteworthy posts" doesn't just help users find good posts through organization, it also educates them in what makes a post good in the first place. Put another way, moderators can use the other three verbs not just to regulate but also to nudge. The flip side of this point, though, is that any time a moderator uses one of the other verbs, she

---

89 *See* Grimmelmann, *Semicommons*, *supra* note 37, at 2834-39 (discussing vulnerability of Usenet discussion groups in face of breakdown of shared norms).

90 *See* GOOGLE, *Fighting Spam*, http://www.google.com/insidesearch/howsearchworks/fighting-spam.html [http://perma.cc/8LDU-CTGK]. For an example of the controversies that search engine anti-spam efforts can generate, see Josh Constine, *Google Destroys Rap Genius' Search Rankings As Punishment For SEO Spam, But Resolution in Progress*, TECHCRUNCH, Dec. 25, 2013, http://techcrunch.com/2013/12/25/google-rap-genius [http://perma.cc/M465-5SYJ].

91 *See, e.g.*, Justin Cheng et al., *How Community Feedback Shapes User Behavior*, PROC. INT'L CONF. WEBLOGS & SOCIAL MEDIA (2014), http://cs.stanford.edu/people/jure/pubs/disqus-icwsm14.pdf [http://perma.cc/9WW3-A9CH] ("Instead, we find that community feedback is likely to perpetuate undesired behavior. In particular, punished authors actually write worse in subsequent posts, while rewarded authors do not improve significantly.").

92 The other verbs of moderation are, in this sense, secondary to norm-setting. Either they encourage users to comply with community norms, or they step in when norms have failed.

nudges participants' norms, whether she intends to or not. For example, excluding a well-known commenter can reduce participants' sense of trust in a moderator, even if the exclusion is justified. Experienced moderators evaluate every design decision in terms of its effects on community norms.

A few particularly important ways to promote good norms reflect the accumulated wisdom of community managers. By far the most significant is fostering a sense of shared identity that reinforces participants' sense of belonging and their commitment to the good of the community.[93] Another is the initiation of new participants, who must be taught the community's expectations at the same time as they are made to feel welcome.[94] Highlighting good behavior and hiding bad behavior reinforce participants' sense that good behavior is prevalent while also teaching them what to do.[95] As a result, designers frequently worry about how to balance competitive and cooperative impulses. Competition can spur users to individual effort at the cost of social cohesion, and different communities strike the balance differently.[96]

B. *Distinctions (Adverbs)*

Picking a verb of moderation does not end the process. Each verb can be used in quite different ways. There are four important distinctions that affect how a type of moderation operates: (1) humans vs. computers, (2) secret vs. transparent, (3) *ex ante* vs. *ex post*, and (4) centralized vs. decentralized. These are the "adverbs" of moderation. These four distinctions are independent: any Verb of moderation can be applied using any of the sixteen possible combinations. For example, spam filters are a secret, decentralized, automatic, *ex post* form of organization (specifically, deletion). A chat room facilitator is a centralized, human, transparent norm-setter who acts both *ex ante* and *ex post* and may have access to tools for exclusion and deletion.

1. Automatically / Manually

Moderation decisions can be made automatically by software or manually by people.[97] To take a simple example, a poli-

---

[93] *See, e.g.*, KRAUT & RESNICK, *supra* note 17, at 79-115.

[94] *See, e.g.*, NG, *supra* note 15, at 179-92.

[95] *See, e.g.*, KRAUT & RESNICK, *supra* note 17, at 140-150.

[96] *See, e.g.*, CRUMLISH & MALONE, *supra* note 16, at 155-59.

[97] In a narrow sense, all moderation decisions are applied by software because an online community is entirely a creature of software. Grimmelmann, *Anarchy*, *supra* note 25. And in a broad sense, all policy decisions are ultimately made by the people who control and program the software. *Id.* We are concerned here with the intermediate question of which actor is responsible for day-to-day, garden-variety moderation decisions. The

cy against foul language could be implemented either through a software filter that blocks the seven dirty words or by a censor who reads everything and decides what does and does not cross the line.[98] Humans have been setting norms, excluding, pricing, and organizing for millennia. Software, too, can do all four, albeit with varying aptitude. Software is effective at enforcing some exclusion decisions. Geotargeting, for example, limits access based on physical location.[99] On the other hand, some exclusion criteria remain easier to apply manually. Whisper employs a small army of moderators in the Philippines to screen images for "pornography, gore, minors, sexual solicitation, sexual body parts/images, [and] racism."[100] Software is even more effective at pricing. Offering standardized price terms to anyone in the world is the kind of low-granularity but universal application for which it is comparatively easy to write software.[101] Improved machine learning and data-mining technologies have led to stunning advances in software-abetted organization in the last decade, particularly in search (a hybrid of synthesis and filtration).[102] Software is least good at norm-setting, due to its lack of understanding of human subtleties, but it can still participate. Design features signal attitudes, can elicit empathetic reactions from participants, can mimic norm-affecting participation, and can shape what other participants see.[103]

Three characteristics of software I identified in *Regulation by Software* play out predictably when software is used for moderation.[104] First, software moderation has higher fixed costs but much lower marginal costs than human moderation. It takes more work to tell a computer what to do than to tell a

---

choice is whether humans themselves make specific decisions about particular content or whether they delegate those decisions to algorithms.

98 *But see* Declan McCullagh, *Google's Chastity Belt Too Tight*, CNET NEWS, Apr. 23, 2004, http://news.cnet.com/2100-1032_3-5198125.html [http://perma.cc/HM8L-ZWXA] (describing the "Scunthorpe problem" of overzealous software filters that find false positives of prohibited terms embedded in innocent phrases).

99 *See generally* Marketa Trimble, *The Future of Cybertravel: Legal Implications of the Evasion of Geolocation*, 22 FORDHAM INTELL. PROP. MEDIA & ENT. L.J. 567 (2012).

100 *See* Chen, *supra* note 18.

101 *See generally* Harry Surden, *Computable Contracts*, 46 U.C. DAVIS L. REV. 629 (2012); Grimmelmann, *Regulation by Software*, *supra* note 58, at 1746-49.

102 *See generally* Grimmelmann, *Speech Engines*, *supra* note 22*; see also* R. Stuart Geiger, *The Lives of Bots*, *in* CRITICAL POINT OF VIEW: A WIKIPEDIA READER 78-93 (Geert Lovink & Nathaniel Tkacz eds., 2011) (describing the use of bots on Wikipedia for mass organization).

103 *See, e.g.*, Ryan Calo, *People Can Be So Fake: A New Dimension to Privacy and Technology Scholarship*, 114 PENN ST. L. REV. 809 (2010); Neal Kumar Katyal, *Digital Architecture as Crime Control*, 112 YALE L.J. 2261 (2003).

104 Grimmelmann, *Regulation by Software*, *supra* note 58.

human, but once the programming work is done, it is cheap to use it in thousands or millions of cases.[105] No human could possibly carry out the millions of sorting decisions Reddit makes on a daily basis. Software is also more rule-bound than humans are. Thus, software is comparatively more effective at making decisions that can be reduced to "hard" facts and figures, such as how many messages a user has sent or how widely a given message has been distributed.[106] And third, software fails differently than humans: it can fail all at once and is vulnerable to hacking. This is not to say that software is always less reliable or secure—humans make inexplicable errors and are vulnerable to social manipulation. But most of the time, human decision-making is more robust than software decision making.[107]

The tradeoff between cost and quality is characteristic of the choice between human and automated moderation. More human attention generally means better but costlier decisions. One of the reasons that user-generated moderation is so attractive to Internet companies is that it allows for human moderation's greater responsiveness while pushing the associated costs off onto users. Companies are also now increasingly using outsourced labor to drive down the cost of human review.[108] Paradoxically, by turning human moderation into assembly-line piecework, these companies make it more and more like automated moderation—cheap, but also rule-bound and inflexible.

2. Transparently / Secretly

Every moderation decision has some observable consequences, but some are more observable than others. Transparent moderation makes explicit and public what the moderators have done and why, revealing what the overall moderation policies are and how they apply in each specific case. Secret moderation hides the details. This distinction is really a spectrum: moderation could be transparent about the what but not the why, or transparent only some of the time. Generally speaking, transparency takes additional work to implement, just as having judges give reasoned explanations of their decisions increases the judicial workload.

---

[105] *Id.* at 1729.

[106] *Id.* at 1732-34.

[107] *Id.* at 1742-45.

[108] *See* Chen, *supra* note 18; *see also* Tarleton Gillespie, *The Dirty Job of Keeping Facebook Clean*, CULTURE DIGITALLY, Feb. 22, 2012, http://culturedigitally.org/2012/02/the-dirty-job-of-keeping-facebook-clean [http://perma.cc/A4ED-B3G8] (discussing Facebook's detailed guidelines for outsourced moderators).

It is easier to be secretive about some kinds of moderation than others. Someone who is excluded from a community will generally be able to tell that they are being denied access, although it is sometimes possible to disguise the fact that it is deliberate.[109] Prices, for the most part, also need to be known to be effective in shaping choices, although implicit prices and micropayments can create some wiggle room.[110] Organization has the most room for secrecy. Search users don't know what pages Google hides from them; Facebook users may not realize that the News Feed is only a partial list of posts from friends.[111] Conversely, secret norms are close to an oxymoron: norms must be known to be effective.

The choice between transparency and secrecy in exclusion, pricing, and organization can have indirect effects on norms. On the one hand, transparency enhances legitimacy, providing community support for moderation, while secrecy raises fears of censorship and oppression.[112] On the other, the "Streisand Effect" can undermine the effectiveness of exclusion or deletion: censorship attempts call attention to the censored material.[113] Indeed, censorship can undermine norms by suggesting that the unwanted behavior is prevalent and can even draw trolls seeking attention.[114] One clever technique for splitting the difference is disemvoweling—leaving only the consonants in an inappropriate comment.[115]

The choice between secrecy and transparency also interacts with the choice between software and humans. The more com-

---

[109] *See, e.g.*, KRAUT & RESNICK, *supra* note 17, at 137-38.

[110] "Micropayments are systems that make it easy to pay small amounts of money." Michael Kinsley, *You Can't Sell News by the Slice*, N.Y. TIMES, Feb. 9, 2009, at A27. *See generally* Clay Shirky, *The Case Against Micropayments*, O'REILLY P2P (Dec. 19, 2000), http://www.openp2p.com/pub/a/p2p/2000/12/19/micropayments.html [http://perma.cc/ZU9F-JK7F] (critizing micropayment systems).

[111] *See* J. Nathan Matias, *Uncovering Algorithms: Looking Inside the Facebook News Feed*, MIT CTR. FOR CIVIC MEDIA (July 22, 2014), https://civic.mit.edu/blog/natematias/uncovering-algorithms-looking-inside-the-facebook-news-feed [https://perma.cc/9GUD-87YT].

[112] *See* KRAUT & RESNICK, *supra* note 17, at 138; NG, *supra* note 15, at 104-07.

[113] The canonical example of the Streisand effect is the trope namer: Barbra Streisand's failed attempt to suppress distribution of an aerial photograph of her house, which led hundreds of thousands of people to seek out the photograph. *See* Paul Rogers, *Streisand's Home Becomes Hit on Web*, SAN JOSE MERCURY NEWS, June 24, 2003, http://www.californiacoastline.org/news/sjmerc5.html [http://perma.cc/4JE4-VUSR].

[114] *See* KRAUT & RESNICK, *supra* note 17, at 145.

[115] *See* Cory Doctorow, *How to Keep Hostile Jerks from Taking over Your Online Community*, INFORMATION WEEK (May 14, 2007), http://www.informationweek.com/how-to-keep-hostile-jerks-from-taking-over-your-online-community/d/d-id/1055100 [http://perma.cc/7ZS5-DWCS] (arguing that disemvoweling "takes the sting out of" abusive comments without censoring them, and also signals community norms).

plex an algorithm, the harder to explain *why* it does what it does in a way that is intelligible to humans and the greater the risk that it will act unaccountably.[116] Yet secrecy may be necessary: transparency is riskier with software than with people because there is a danger of unchecked loopholes.[117] Anti-spam email filtering, for example, depends in part for its success on the fact that spammers are unaware of the exact details of filtering and so cannot send messages guaranteed to sneak past filters. The costs of secrecy do not just fall on abusive users, though. Google's secretive ways, adopted as a defense against search engine optimizers, make it hard for innocent websites to understand why their search rankings have fallen.

3. Ex Ante / Ex Post

Moderators can act *ex ante*—using their power over the infrastructure to allow some actions and prohibit others—or they can act *ex post*—using their powers to punish evildoers and set right that which has gone wrong. Acting *ex ante* takes advantage of software's architectural features; acting *ex post* is a more traditionally law-like technique.[118] *Ex ante* moderation can produce consistency by applying the same rules to all content. *Ex post* moderation can conserve resources by directing moderators' attention only where it is needed.

The distinction plays out differently for different verbs. *Ex ante* exclusion can work in three ways. First, it can ration access to limit congestion and cacophony. A chat room with ten participants is easier to follow than one with a hundred all going full-speed.[119] Second, it can be a crude filter that uses a member's identity as a proxy for the value of her contributions: a company might reasonably assume that non-employees have little to add to the discussion on its legal department's email list.[120] Finally, it can limit community size: smaller communities may *eo ipso* be better able to cooperate because they have stronger norms.[121] *Ex post* exclusion is a punishment for mis-

---

[116] See Matias, *supra* note 111 ("Is there any person at Facebook who knows how the algorithm works?"); *see generally* Danielle Keats Citron, *Technological Due Process*, 85 WASH. U. L. REV. 1249 (2007).

[117] *See* FARMER & GLASS, *supra* note 16, at 91-93.

[118] *See* Grimmelmann, *Regulation by Software*, *supra* note 58, at 1729-30. The distinction has been regularly rediscovered. *See, e.g.*, Michael L. Rich, *Should We Make Crime Impossible?*, 36 HARV. J.L. & PUB. POL'Y 795 (2013); Danny Rosenthal, *Assessing Digital Preemption (and the Future of Law Enforcement?)*, 14 NEW CRIM. L. REV. 576 (2011).

[119] *See* FRISCHMANN, *supra* note 46, at 144-46.

[120] PREECE, *supra* note 15, at 273 (discussing special-purpose communities that limit registration); Smith, *supra* note 48, at S468-71 (discussing exclusion as a crude filter).

[121] *See generally* MANCUR OLSON, THE LOGIC OF COLLECTIVE ACTION (rev. ed. 1971).

behavior that puts teeth in community rules.[122] If *ex post* exclusion is coupled with some transparency about the reasons for excluding participants, its existence gives members an incentive to behave.[123]

Ex ante pricing implements the usual understanding of a market—pay to play. *Ex post* pricing could be one of three things. First, the moderators may simply be extending credit to members or offering a free trial: those who do not pay up when billed will then be excluded. Second, it could be an honor-system—pricing backed up by norms—in which participants who appreciate content or the community are encouraged to chip in to support it.[124] Third, an *ex post* price could be a sanction for misbehavior, a punishment short of exclusion.[125]

The choice between *ex ante* and *ex post* organization is tied to the choice of actor and thus to the incidence of implicit costs. Authors act *ex ante*; readers act *ex post*; moderators can do either. If authors must pick *ex ante* a topic on a discussion board in which to post, the cost of posting is higher by the amount of effort involved in picking the right one. If moderators come along *ex post* and assign topics, authors bear less of a burden, and moderators bear more of one. *Ex post* organization is widespread, as everyone who has ever liked a photo on Facebook or flagged an abusive YouTube comment can confirm.

Regardless of who performs it, *ex ante* organization imposes a time cost on distribution because readers do not receive content until it has been organized. It may be more convenient to get your email from a mailing list in a daily digest, but you will miss out on fast-breaking conversations. *Ex post* organization can be faster-moving,[126] but if the goal is to edit out unwanted content and to inculcate norms against it, leaving it in place for too long can be dangerous.[127] Further, because *ex post* organiza-

---

[122] *See* KRAUT & RESNICK, *supra* note 17, at 158-60; NG, *supra* note 15, at 97.

[123] For example, MetaFilter bans users who use posts for self-promotion. *See Self Link*, MEFI WIKI, http://mefiwiki.com/wiki/Self_Link [http://perma.cc/V9GW-N7M4]. Note that this point holds only when users have enough invested in their community identities to make exclusion a meaningful sanction.

[124] *But see* Matthew Ingram, *Why Online "Tip Jar"-Style Payment Systems Don't Work*, GIGAOM, May 11, 2011, http://gigaom.com/2011/05/11/why-online-tip-jar-style-payment-systems-dont-work [http://perma.cc/APR9-M4U9].

[125] *See* Robert Cooter, *Prices and Sanctions*, 84 COLUM. L. REV. 1523 (1984). The difference between sanctions and prices is that sanctions attempt to defend a known line from transgression, rather than to measure the precise amount of harm. The temporal asymmetry provides a different way to distinguish sanctions and prices. Prices can be applied *ex ante* or *ex post*, but sanctions can only be applied *ex post*.

[126] *See* BELL, *supra* note 16, at 59 (praising *The Guardian* for its use of *ex post* moderation).

[127] *See* KRAUT & RESNICK, *supra* note 17, at 132.

tion can alter content or change its attributes, it can paradoxically impose search costs on participants. If you have ever found a web page through a search engine, only to lose it later when the page is no longer a prominent result for your original search term, you are a victim of a change in *ex post* filtration.

An especially important aspect of *ex post* organization—specifically of *ex post* deletion—is the spectrum from ephemerality to permanence.[128] By lowering the stakes, ephemerality promotes experimentation, risk-taking, and contingency.[129] It also inhibits the formation of recognizable individual identities, which can ironically promote the development of a shared collective ethos.[130] More persistent content allows for norm-enhancing community memory and for enduring individual reputations. The good that community members do lives on; so does the bad.

Finally, effective social norms have both *ex ante* and *ex post* aspects. *Ex post*, the community expresses its approval or disapproval after a member has acted. Once someone has hit reply-all for a personal aside, others can only glower and make pointed remarks. *Ex ante* social norms are those that members have internalized. With enough pointed remarks, members will learn to check themselves before they hit reply-all.

4. Centrally / Distributedly

Moderation decisions can be made either centrally by a single moderator whose decision affects the entire community, or by multiple distributed moderators whose individual decisions affect only part of the community.[131] For the most part, the consequences are as one would expect, and track the usual legal debates about hierarchy and federalism. Centralized moderation provides consistency: there is only one domain-name system. Distributed moderation promotes diversity: TMZ and PatientsLikeMe have different moderation policies, and should. Centralized moderation aggregates information: Google's PageRank algorithm draws on the entire structure of the Web. Distributed moderation relies on local knowledge: mailing list moderators have experience with their members' sense of

---

[128] I am indebted to Sarah Jeong for pointing out this distinction.

[129] *See* Lee Knuttila, *User Unknown: 4chan, Anonymity, and Contingency*, FIRST MONDAY, Oct. 3, 2011, http://firstmonday.org/article/view/3665/3055 [http://perma.cc/VP3A-57CS] (explaining how ephemerality of posts on 4chan is responsible for "a unique, virtual ontological experience" and "fortuitous encounter[s]").

[130] *See* Auerbach, *supra* note 55; Jay Allen, *How Chan-Style Anonymous Culture Shapes #gamergate*, STORIFY (Dec. 3, 2014), https://storify.com/a_man_in_black/how-chan-style-anonymous-culture-shapes-gamergate [https://perma.cc/AFN4-PW9S].

[131] *Cf.* Raaj Kumar Sah & Joseph E. Stiglitz, *The Architecture of Economic Systems: Hierarchies and Polyarchies*, 76 AM. ECON. REV. 716 (1986).

which messages are off-topic. Centralized moderation offers the ability to stop unwanted content and participants by creating a single checkpoint through which all must pass: a spammer kicked off of Facebook will not bother anyone else on Facebook. But chokepoints are also single points of failure: a spammer who gets through on Facebook can bother a lot of people. In comparison, distributed moderation offers more robustness and defense in depth. Centralized moderation offers a clear focal point for policy-making. If you don't like my post, you know where to complain. Distributed moderation permits those with ideological differences to agree to disagree: if you don't want to read my weblog, no one is putting it in front of you.

In a sense, the choice between centralized and distributed exclusion is the choice between a single community and many. Similarly, the choice between centralized and distributed pricing is the choice between a big-box retailer and a bazaar of many small merchants. It is in organization that the dichotomy between centralized and distributed moderation is the sharpest and the richest. A search engine is powerfully centralized; a social network devolves many organizational decisions to members who decide which friends to share and converse with. Taxonomies are centralized annotation; folksonomies of user-assigned tags are distributed annotation.[132] A top-ten list is a centralized filter; user-created playlists are distributed filters. But norms, by their nature, cannot be fully centralized. The power to adopt, shape, or reject them is always in the hands of members. The larger a community, the more competing voices and normative focal points it is likely to have.

C. *Community Characteristics (Adjectives)*

Just as one size does not fit all forms of moderation, one size does not fit all communities. Communities differ along many axes: the email system has different properties than Wikipedia, which has different properties than the comments section of a blog. Four characteristics of a community are particularly important in affecting the kinds of strategic behavior threatening it and the effectiveness of various types of moderation: (1) the capacity of the infrastructure, (2) the size of the user community, (3) the distribution of ownership, and (4) the identifiability of participants. As with the adverbs above, these characteristics are mostly independent of each other.

---

[132] *See* Adam Mathes, *Folksonomies—Cooperative Classification and Communication Through Shared Metadata* 3-5 (2004), http://adammathes.com/academic/computer-mediated-communication/folksonomies.pdf [http://perma.cc/Y5KE-RMCY].

1. Infrastructure Capacity

Infrastructure's capacity—hard drive space, bandwidth, processing power, electric power, etc.—affects its ability to support members' use. Where there is too much use for a given capacity, congestion results. Members find the system unpleasant, unreliable, or unusable. In theory, it is almost always possible to add infrastructure, increase capacity, and reduce congestion. But in practice, limited capacity affects the community in two ways. First, infrastructure costs money, so paying for it often requires pricing. Second, adding capacity takes time, which means congestion can be a major short-run problem, particularly in growing communities, even when long-run upgrades are feasible.[133] Friendster stumbled over technical issues as it grew and was surpassed by MySpace,[134] which stumbled in turn and was surpassed by Facebook.[135]

A community in which capacity is a significant bottleneck looks very different from one in which it is not. With little capacity, the common-pool resource problem at the infrastructure level dominates, favoring moderation that closely regulates usage: exclusion, pricing, and deletion. As capacity increases, infrastructure recedes and content comes to the fore. There may still be cacophony, abuse, and manipulation, but these problems are more amenable to additive solutions, as captured in the slogan that the best remedy for bad speech is more speech. Annotation, filtration, synthesis, and norm-setting become comparatively more attractive. Where the balance between capacity constraints and cognitive constraints falls will vary by community. Ones in which members share rich multimedia content will experience congestion sooner and more painfully than ones in which members share short textual content.[136]

The minimum practical unit of infrastructure is often sufficient to enable a great deal of use, making it an important spe-

---

[133] Infrastructure can be lumpy. When a community has outgrown its first server, it cannot easily add one-tenth of a second server. You cannot make a terabyte database simply by connecting a thousand gigabyte databases to each other. Cloud computing, however, is smoothing out infrastructure capacity by making it much easier to throw more computing resources at a problem, quickly and scalably.

[134] *See* Gary Rivlin, *Wallflower at the Web Party*, N.Y. TIMES, Oct. 15, 2006, http://www.nytimes.com/2006/10/15/business/yourmoney/15friend.html [http://perma.cc/E8EE-K254].

[135] *See* JULIA ANGWIN: STEALING MYSPACE: THE BATTLE TO CONTROL THE MOST POPULAR WEBSITE IN AMERICA 246-53 (2009).

[136] *Compare, e.g.*, TWITCH, http://www.twitch.tv [http://perma.cc/5PG9-PKLB] (users share gaming videos), *with* YO, http://www.justyo.co [http://perma.cc/ZUC2-VSWM] (users share the word "Yo").

cial case of abundance.[137] For less than $250, you can buy a computer capable of carrying out two billion operations per second and with a hard drive capable of storing half a million full-text novels.[138] Accordingly, a typical blog could receive thousands of comments a month without increasing its owner's costs in the slightest. When participants bring their own infrastructure—as in peer-to-peer systems—they may be able to support a substantial community without substantial effort. Growth out of this range creates an important scale transition: capacity becomes something the community must worry about, pay for, and safeguard.

2. Community Size

Closely related to infrastructure capacity is the number of members in a community. One important issue, discussed above, plays out at the infrastructure layer: more members means greater use and thus greater congestion.[139] The more interesting consequences of increasing community size play out at the information layer. There are two offsetting effects for readers and authors. On the one hand, greater size catalyzes informational network effects in this two-sided market for attention: readers would rather join a fan fiction community with ten thousand stories than one with ten, while authors would rather post to a fan fiction community with ten thousand readers than one with ten. These effects are critical when a community starts. Like airplanes, communities need forward momentum to take off.[140] On the other hand, a large community of authors will generate cacophony, making moderation increasingly essential if readers are to find anything of value. Once a fan fiction community has ten thousand stories, it needs tags or a search function to separate the Harry/Draco slash[141] from the

---

[137] *See* Yochai Benkler, *Sharing Nicely: On Shareable Goods and the Emergence of Sharing as a Modality of Economic Production*, 114 YALE L.J. 273, 301-04 (2004).

[138] As of March 2015, a Dell Inspiron 14-inch laptop with a 500-gigabyte hard drive and a 2.16-gigahertz CPU cost $229.99 on sale from Dell.com. *See New Inspiron 14 3000 Series Laptop*, DELL, http://www.dell.com/us/p/inspiron-14-3451-laptop/pd [http://perma.cc/8DWA-JXK6]. Also as of March 2015, Amazon Web Services offers 750 hours of computing time per month and tens of gigabytes of storage *for free*. *See AWS Free Tier*, AMAZON, http://aws.amazon.com/free [http://perma.cc/X272-92C9].

[139] This is the classic common-pool resource problem, and conventional wisdom recommends restricting community membership for just this reason. *See, e.g.*, OSTROM, *supra* note 38, at 91-92.

[140] The importance of catalyzing the initial roll-out of a community is a recurring focus of books on community management. *See, e.g.*, NG, *supra* note 15, at 113-23.

[141] *See, e.g.*, Blackie & Yoyo, *[Tag: Harry/Draco]*, FUCK YEAH HP SLASH, http://fuckyeahhpslash.tumblr.com/tagged/draco [http://perma.cc/D5V5-4FYR].

Ronbledore,[142] and stars or favorites or votes to filter the cream from the chaff. To summarize, the larger a community is, the better it is at competing with external alternatives, but the more internal moderation it requires.

Moreover, community size interacts with the effectiveness of the forms of moderation. Growth is often notably unkind to social norms. It is easier to maintain any given norm in a smaller community than a larger one. As a community grows, it becomes easier for individuals and groups to resist a norm. This breakdown makes it harder to use social norms to moderate large communities. A group of twenty can operate by unspoken consensus in a way that a group of twenty thousand cannot. Thus, decentralized moderation becomes increasingly attractive as the community grows because it fragments the community into smaller subcommunities that can maintain their own norms. Reddit's subreddits, described below, are a superlative example. Exclusion can also be more difficult in a large community because it is easier for the unwanted to sneak in (for example, by stealing a password or giving a false name) and avoid immediate detection.

On the other hand, pricing and organization can benefit from community size. Pricing at scale benefits from the salami-slicing effect: a great many small payments can add up to a surprisingly large number. This is the key, for example, to advertising. Each pageview is good only for a small fraction of a penny, but those fractions add up fast. Organization can take advantage of the law of large numbers. Any individual moderator's assessment of an action's value may or may not accurately reflect the community's sense of value, but the average of a thousand moderators' assessments is likely to express it fairly well. Thus, some techniques of synthesis become increasingly reliable as the community grows. Google's assessment of Web pages' importance, for example, synthesizes the individual decisions of many millions of Web authors. The same is true of Amazon's averages of user reviews, of Reddit's upvoting algorithms,[143] and even of American Idol.

Finally, community size shapes the way in which moderation can best be executed. All of the verbs have costs that increase with volume, and moderation requires greater and

---

[142] *See* estel, *Weasley is Dumbledore Theory, If You Have Time to Spare,* HARRY POTTER'S PAGE DISCUSSION BOARDS (MAY 21, 2004, 11:31 AM), http://www.harrypotterspage.com/forums/index.php?s=&showtopic=393&view=findpost&p=17361 [http://perma.cc/Y8JG-WST8]; *see generally* Mallory Ortberg, *Ronbledore Archive*, THE TOAST, http://the-toast.net/tag/ronbledore [http://perma.cc/MWF4-BR9N].

[143] Indeed, the Reddit algorithm has been tweaked to reflect the fact that assessments of content become more reliable as more people provide them. *See infra* note 273.

greater investments as a community grows. I have mentioned the scale transition that occurs when a community becomes big enough that it must start worrying about congestion. There is a second common transition, which occurs when a community becomes too big for one person to moderate without help. Professional moderation becomes attractive, but paying those professionals requires pricing. A third scale transition occurs when the community becomes too big for any reasonably sized group of humans to moderate entirely on their own. YouTube, for example, would need three shifts of six thousand employees each, working around the clock, to prescreen all the videos uploaded to the site.[144] Either decentralized or automatic moderation—and quite possibly both—becomes a necessity.

(1) Ownership Concentration

Just as moderation can be centralized or distributed, so can ownership. The two questions are distinct. Wikipedia has centralized ownership (the Wikimedia Foundation) but decentralized moderation. Bitcoin has centralized moderation (there is only one blockchain) but decentralized ownership (many different people and organizations run computers that participate in the Bitcoin network).[145] For the sake of clarity, I will refer to centralized ownership as *concentrated*, and decentralized ownership as *dispersed*.

Concentrated ownership has one substantial advantage: there is only one owner whose account books must balance, rather than many. With dispersed ownership, if one of the many owners finds that she is absorbing a disproportionate share of the costs, she may simply withdraw. Put another way, concentrated owners can afford to be indifferent to the distribution of costs, since one part of the infrastructure may subsidize another. The *New York Times* does not have to worry about separately accounting for the profit and loss of comments on each individual article on its website. Distributed owners have no choice but to worry about the balance of payments. Such a system is far more likely to be stable when the owners are also participants, so that they subsidize themselves. Peer-to-peer file sharing is the classic example of a case in which the rewards of participation induce users to contribute their computing resources to the infrastructure. This self-organizing, self-provisioning as-

---

[144] *See Statistics*, YOUTUBE, https://www.youtube.com/yt/press/statistics.html (last visited Jan. 20, 2015) [https://perma.cc/ZG7H-CL5J] ("100 hours of video are uploaded to YouTube every minute."). Assuming each moderator can watch one video at a time, watching every video would therefore require 6,000 moderators working simultaneously, all the time.

[145] *See generally* ANDREAS M. ANTONOPOULOS, MASTERING BITCOIN: UNLOCKING DIGITAL CURRENCIES (2014) (describing Bitcoin).

pect of dispersed ownership can be particularly robust in communities that do not rely heavily on exclusion or pricing.

Both forms of ownership can be useful in resisting attacks. On the one hand, dispersed ownership can align incentives by allocating the costs of heavy use to those users' own portions of the infrastructure.[146] For example, in peer-to-peer file-sharing networks, uers who are only willing to pay their ISPs for low-bandwitdh connections can download less than users who are willing to pay more for faster connections. On the other hand, a concentrated owner can mount a coordinated defense against denial of service attacks. The other participants have no infrastructure of their own at risk. This point is more important than it may seem at first because popularity can be an unintentional denial-of-service attack.[147] If George Takei tweets about your website, your server might crash.[148] But if your Facebook Page goes viral, Facebook will take care of it without blinking.

The choice to concentrate or disperse ownership also affects the political economy of the choice among moderation techniques. Owners can use their power over the infrastructure layer to make policy at the content layer, for good and for ill. This is why Facebook had years of controversy over banning breastfeeding photos: as infrastructure owner, it made and applied a broad anti-nudity moderation policy.[149] Manipulation to favor the owner's interests is the constant fear.[150] Distributed ownership gives community members more power to force democratic moderation decisions. To take a simple example, compare the openness of the web with the walled garden that is Facebook. For another, compare Bitcoin with Paypal.

---

[146] This is the pattern described by Smith in his study of the open-field semi-commons, where farming ownership was divided into strips. *See* Smith, *supra* note 48. He explains that the boundaries of privately held portions can be set to prevent strategic behavior—in the case of the open fields, to make it hard for shepherds to concentrate grazing harms on particular owners.

[147] *See Slashdot Effect*, KNOW YOUR MEME, http://knowyourmeme.com/memes/slashdot-effect [http://perma.cc/U3N4-ZEPK].

[148] *See* Anna Leach, *Mr. Sulu Causes DDoS Panic After Posting Link on Facebook*, THE REGISTER, June 8, 2012, http://www.theregister.co.uk/2012/06/08/takei_ddos_facebook_fans [http://perma.cc/M5KW-NJXA].

[149] *See* Soraya Chemaly, *#FreeTheNipple: Facebook Changes Breastfeeding Mothers Photo Policy*, HUFFINGTON POST, June 9, 2014, http://www.huffingtonpost.com/soraya-chemaly/freethenipple-facebook- changes_b_5473467.html [http://perma.cc/8976-4D37].

[150] *See, e.g.*, Christian Sandvig, *Corrupt Personalization*, SOCIAL MEDIA COLLECTIVE, June 26, 2014, http://socialmediacollective.org/2014/06/26/corrupt-personalization [http://perma.cc/AJ7M-Q4ZP].

3. Identity

The final community characteristic is the distinction between identity and anonymity. At one extreme, participants in an online community could be completely identified, bringing with them a complete biography of their online and offline lives. At the other, they could be completely anonymous.[151] Compare Google+, which launched with a strict, stringently enforced, and much-criticized "real names" policy,[152] with 4chan, where "most posts . . . are disconnected from *any* identity."[153] There are many gradations in between.[154] Participants could have identities that mostly match their offline lives, but in which the details are potentially questionable, as in an online dating service where participants sometimes lie about their height.[155] They could have rich and persistent but avowedly fictitious identities, as in virtual worlds where they play the same avatar thirty hours a week for years. They could have stable but thin identities, as on a discussion board that uses pseudonyms and keeps users' real names and email addresses secret. They could have thin identities purely as a matter of convention, as in some blogs' comment sections, where a commenter can pick a fresh display name with each comment. Participants could even have one level of identifiability at the infrastructure level (supply a valid email address to sign up) but a different level at the content layer (that email address is hidden from other participants). Whatever its nature, the most important role of identity is creating stable reputations through time so that others can link past behavior to a present identity.[156]

All four verbs of moderation can tap into identity. Exclusion absolutely depends on it; without identity, the distinction between "outsiders" and "insiders" collapses. You can identify the

---

[151] *See, e.g.*, E. Gabriella Coleman, *Our Weirdness Is Free*, TRIPLE CANOPY (Jan. 2012), http://www.canopycanopycanopy.com/contents/our_weirdness_is_free [http://perma.cc/DV5P-AYX3] (discussing "the sublimation of identity" in hacker collective Anonymous).

[152] *See* Jillian York, *A Case for Pseudonyms*, DEEPLINKS, July 29, 2011, https://www.eff.org/deeplinks/2011/07/case-pseudonyms [https://perma.cc/RU9U-DCPM] (last visited Mar. 16, 2015).

[153] Michael S. Bernstein et al., *4chan and /b/: An Analysis of Anonymity and Ephemerality in a Large Online Community*, PROC. INT'L CONF. ON WEBLOGS AND SOCIAL MEDIA 3 (2011) (emphasis added); *see also* Auerbach, *supra* note 55.

[154] *See generally* FARMER & GLASS, *supra* note 16, at 21-36 (presenting graphical grammar of reputation).

[155] *See* Christian Rudder, The Big Lies People Tell in Online Dating, OKTRENDS (July 7, 2010), http://blog.okcupid.com/index.php/the-biggest-lies-in-online-dating [http://perma.cc/7W4R-YVQ9].

[156] *See, e.g.*, FARMER & GLASS, *supra* note 16, at 17 (describing "The Reputation Virtuous Cycle").

unwanted outsiders and blacklist them or identify the wanted insiders and whitelist them, but both versions require some notion of identity.[157] As anyone who has moderated a blog's comments can testify, this is a huge problem for communities that are open to new and unknown members from the Internet at large. There is often no way to tell that a "new" commenter is actually an old and well-known miscreant, back from a ban for another round of malice.[158]

In theory, pricing could be identity-free—transactional and transitory. But in practice, any explicit pricing system will depend on some non-trivial identity infrastructure, such as a credit card processor. Having a credit card number is not necessarily a guarantee of anything, but it is a significant identity hurdle, and one that many online businesses, for example, use to create some minimal level of accountability among users. More complex pricing builds on persistent identity. "For $10, you can post as often as you want for a year" requires a notion of "you" that will be stable for a year.

Organization can both piggyback on and produce identity. Filtering and deletion both often treat the identity of an author as a significant data point. Most anti-spam systems, for example, use whitelisting to allow trusted senders' emails to bypass the spam check altogether. In reputation systems, participants provide annotations on each other's actions, and those annotations become part of one's community identity.[159] eBay's feedback system, in which buyers and sellers use feedback left by others to decide whom to trust, is an example of reputational annotation. Slashdot's multi-level moderation system has several variables that use others' ratings of one's actions to decide how much power one will have to moderate in the future.[160]

It is well known that identifiability plays a significant role in setting social norms.[161] Persistent reputations make it possible for participants to build credibility as respected elders within the community.[162] They make it possible to hold participants accountable for their actions, enabling the effective monitoring and graduated sanctions beloved by commons scholars.[163] By contrast, anonymity enables consequence-free norm violation

---

[157] *See generally* PREECE, *supra* note 15, at 96-97.

[158] KRAUT & RESNICK, *supra* note 17, at 138.

[159] *See generally* FARMER & GLASS, *supra* note 16.

[160] *See* Lampe, *supra* note 84.

[161] Lessig famously described the difference in tone between two class newsgroups, one anonymous and one with stronger identity; the anonymous one was hijacked by a malicious flamer. LAWRENCE LESSIG, CODE 2.0 102-06 (2006). *See generally* Bryan H. Choi, *The Anonymous Internet*, 72 MD. L. REV. 501 (2013).

[162] *See, e.g.*, Warnick, *supra* note 20, at 103-04.

[163] *See* KRAUT & RESNICK, *supra* note 17, at 155-57; OSTROM, *supra* note 38, at 94-100; *see generally* FARMER & GLASS, *supra* note 16.

and can undermine the appearance of reciprocity among real human beings. But stronger identity is not always better. Sometimes it creates a badge for misbehavior: a leaderboard is an invitation to fame-seeking cheaters. Making participants more anonymous (for example, by resetting a server) can drive trolls away because it deprives them of the opportunity to make a (bad) name for themselves.

Paradoxically, both identity and its opposite—anonymity—can be expensive to establish. Externally produced identity requires participants to prove facts about themselves, which can cost both time and money. It also requires owners and moderators to be prepared to check these assertions, which too is costly. Internally produced reputation systems require participants to take the time to learn about, comment on, and rate each other.[164] An important question for online communities is who controls these socially constructed identities: users themselves, the community, or infrastructure owners.[165] Anonymity might seem cheaper, but genuinely effacing participants' identities requires some significant effort—deleting log files, stripping out snooping software, and taking action against participants who "out" one another's offline identities.

Finally, identity can be the enemy of privacy, for good and for bad. Divulging information about oneself is itself a cost. Privacy is virtually a precondition for some kinds of speech. Some conversations simply cannot take place in public. This phenomenon can be good: think of therapeutic conversations on a discussion board for adult victims of childhood abuse. It can also be bad: think of virulently misogynistic and racist conversations on a law student board.

Two forms of abuse are characteristically tied to the misuse of identity. Impersonation—the hijacking of another's identity—requires that participants have recognizable identities to hijack.[166] And sock puppetry—creating fake personas to create the false appearance of support for a position—requires that the community recognize personas as distinct participants in the first place.[167] Both become possible when a community ac-

---

[164] FARMER & GLASS, *supra* note 16, at 223-41.

[165] *See, e.g.*, Beth Simone Noveck, *Trademark Law and the Social Construction of Trust: Creating the Legal Framework for Online Identity*, 83 WASH. U.L.Q. 1733 (2005); Omer Tene, *Me, Myself, and I: Aggregated and Disaggregated Identities on Social Networking Service*, 8 J. INT'L COM. L. & TECH. 118 (2013).

[166] *See, e.g.*, Dylan Loeb McClain, *Chess Group Officials Accused of Using Internet to Hurt Rivals*, N.Y. TIMES, Oct. 8, 2007.

[167] *See, e.g.*, Simon Owens, *The Battle to Destroy Wikipedia's Biggest Sockpuppet Army*, THE DAILY DOT, Oct. 8, 2013, http://www.dailydot.com/lifestyle/wikipedia-sockpuppet-investigation-largest-network-history-wiki-pr [http://perma.cc/2E4D-BEZ2].

cepts claims of identity that it is not capable of properly validating.

III. Case Studies

This Part discusses four case studies to give a feel for how moderation can play out in practice.[168] Two (Wikipedia and MetaFilter) are hard-won successes. One (the *Los Angeles Times* wikitorial) was an abject failure. The fourth (Reddit) is deeply ambivalent—an immensely popular site with an immensely loyal user base that is nonetheless also responsible for some notoriously destructive episodes.

A. *Wikipedia*

Other than the Internet itself, Wikipedia is the preeminent example of successful online collaboration.[169] It started as an offshoot of Nupedia, one of several attempts in the 1990s and early 2000s to create an online encyclopedia through volunteer contributions.[170] Nupedia relied on peer-reviewed contributions from experts—centralized, transparent, *ex ante* human exclusion and organization—but its founders, Jimmy Wales and Larry Sanger, were frustrated at the slow pace of contributions. Sanger's friend Ben Kovitz suggested that Nupedia use a wiki for initial collaboration on articles that would then go through the full editorial review.[171] And thus, on January 15, 2001, Wikipedia was born.[172] It took off so rapidly that "when the server hosting Nupedia crashed in September 2003 (with little more than twenty-four complete articles and seventy-four more in progress) it was never restored."[173] Today the English-language Wikipedia alone has over four and a half million articles. Twenty-three million registered users and countless anonymous ones have made more than seven hundred million edits.[174] One meta-analysis concluded that Wikipedia has "a valuation in the tens of billions of dollars, a one-time replace-

---

[168] Other case studies illustrate the principles as well. *See, e.g.*, FARMER & GLASS, *supra* note 16, at 243-77 (Yahoo! Answers); Grimmelmann, *Semicommons*, *supra* note 37, at 2831-39 (USENET).

[169] *See generally* PHOEBE AYERS, CHARLES MATTHEWS & BEN YATES, HOW WIKIPEDIA WORKS: AND HOW YOU CAN BE A PART OF IT (2008); ANDREW DALBY, THE WORLD AND WIKIPEDIA: HOW WE ARE EDITING REALITY (2009); JEMIELNIAK, *supra* note 83; LIH, *supra* note 11; JOSEPH REAGLE, GOOD FAITH COLLABORATION: THE CULTURE OF WIKIPEDIA (2010).

[170] *See* LIH, *supra* note 11, at 32-41.

[171] REAGLE, *supra* note 169, at 39. *See generally* BO LEUF & WARD CUNNINGHAM, THE WIKI WAY: ONLINE COLLABORATION ON THE WEB (2001) (describing wiki technology).

[172] *See* LIH, *supra* note 11, at 60-67.

[173] REAGLE, *supra* note 169, at 40; *see also* JEMIELNIAK, *supra* note 83, at 11 (explaining that twenty thousand articles were created in the first year).

[174] *See Wikipedia:Statistics*, WIKIPEDIA, http://en.wikipedia.org/wiki/Wikipedia:Statistics [http://perma.cc/7LC2-GYLC] (last visited Jan. 20, 2015).

ment cost of $6.6 billion with an annual updating cost of $630 million, and consumer benefit in the hundreds of billions of dollars."[175]

A wiki is merely a tool: switching from Microsoft Word to MediaWiki will not make you a master encyclopedist, just as Diderot and d'Alembert's success in creating the *Encyclopédie* was not a matter of having better pens than other *philosophes*. Rather, the technical switch from Nupedia to Wikipedia mattered because it enabled a social shift—dropping the exclusion and switching from *ex ante* to *ex post* organization. Editors could now draw on a much larger pool of potential contributors and improve each other's work incrementally, iteratively, and interactively.[176] These changes dramatically increased the community size and dramatically reduced the implicit price of participation. As new authors added more articles and improved existing ones, they quickly established strong, positive norms. The initial success served as an advertisement for further participants and participation in a virtuous cycle of growth.

Wikipedia's system of moderation is sophisticated and intricate, but its two basic commitments have remained distributed organization and strong social norms. The two are in significant tension.[177] Most of Wikipedia's moderation choices can be understood in terms of the difficult task of sustaining norm-based "soft security," which works through "group dynamics rather than hard-coded limits" in a massive community with millions of members.[178] "But Wikipedia's openness isn't a mistake; it's the source of its success. A community solves problems that official leaders wouldn't even know were there."[179]

---

[175] Jonathan Band & Jonathan Gerafi, Wikipedia's Economic Value (Oct. 7, 2013), http://papers.ssrn.com/sol3/papers.cfm?abstract_id=2338563 [http://perma.cc/58SJ-VZV7].

[176] These are classic features of successful open source collaboration. *See generally* STEVEN WEBER, THE SUCCESS OF OPEN SOURCE (2004). BENKLER, *supra* note 69, describes their broader applicability.

[177] *See* REAGLE, *supra* note 169, at 83-88; Eric Goldman, *Wikipedia's Labor Squeeze and Its Consequences*, 8 J. TELECOMM. & HIGH TECH. L. 157 (2010); Andrew George, *Avoiding Tragedy in the Wiki-Commons*, 12 VA. J.L. & TECH. no. 8 (2007), http://www.vjolt.net/vol12/issue4/v12i4_a2-George.pdf [http://perma.cc/HNV9-9GXU]; Aaron Halfaker et al., *The Rise and Decline of an Open Collaboration System: How Wikipedia's Reaction to Popularity Is Causing Its Decline*, 57 AM. BEHAV. SCIENTIST 664 (2012).

[178] AYERS ET AL., *supra* note 169, at 45; *see also Soft Security*, MEATBALL WIKI, http://meatballwiki.org/wiki/SoftSecurity [http://perma.cc/G78W-9G3T]; Jimmy Wales, *Foreword to* LIH, *supra* note 11, at xvii-xviii ("[T]rying to make sure that nobody can hurt anyone else actually eliminates all the opportunities for trust."); JONATHAN ZITTRAIN, THE FUTURE OF THE INTERNET—AND HOW TO STOP IT 127-48 (2008).

[179] Aaron Swartz, *Who Runs Wikipedia?*, RAW THOUGHT (Sept. 7, 2006), http://www.aaronsw.com/weblog/whorunswikipedia [http://perma.cc/6MNH-2YYZ]. Swartz's six-part series on Wikipedia's self-moderation is

The power of Wikipedia's first commitment, to distributed organization, is by now well established.[180] Wikipedia uses it with remarkable flexibility. The best Wikipedia articles are synthesized from the contributions of thousands of authors, and the hyperlinks between articles are a beautiful use of annotation.[181] Organizationally, Wikipedia uses pages, subpages, lists, lists of lists, categories, categories of categories, sidebars, standardized templates, and even a special-purpose markup and programming language, all tools that enable the richly multimedia and complexly interlinked web of knowledge.[182] These are *ex ante* filtering carried out by authors and moderators; they split the flood of information into manageable streams. Further, Wikipedia offers readers *ex post* filtering though its search engine.[183] It also enjoys additional filtering simply as a consequence of being openly available and searchable on the Web: if you want to learn about widgets, you need only Google "wikipedia widget." [184]

Wikipedia's pricing strategy similarly supports large-scale participation. Just as it is open to anyone, it is also free to read and to edit. Wikipedia's socially beneficial mission allows it to function as a charitable organization. The Wikimedia Foundation, which subsists on donations, keeps the site free for authors, readers, and moderators. Implicitly, the use of a wiki makes it easy, at least in theory, for anyone to dive in and make edits. Indeed, Wikipedia now prohibits undisclosed paid editing because it is worried about incentives for manipulation.[185]

Wikipedia's second commitment, to positive social norms, is even richer and more complex. The basic attitude of epistemic humility is summed up in the two mottoes "[adopt a] neutral point of view" and "assume [that others are acting in] good

---

well worth reading, and holds up quite well eight years later. *See also* James Grimmelmann, *Seven Wikipedia Fallacies*, THE LABORATORIUM (Aug. 27, 2006), http://laboratorium.net/archive/2006/08/27/seven_wikipedia_fallacies_1 [http://perma.cc/C3E8-BRWY].

180 *See generally* WEBER, *supra* note 176; Benkler, *supra* note 39.

181 For example, as of November 3, 2014, the readable and informative article on West Point was roughly 13,000 words long and had been edited 3,815 times by 1,252 editors.

182 *See generally* AYERS ET AL., *supra* note 169, at 68-298 (describing Wikipedia's technical organization).

183 *See id.* at 60-65. Another useful reader filtering technique is the "watchlist," which provides an editor with a chronological list of edits to whichever pages she wishes to track. *See Help:Watching pages*, WIKIPEDIA, http://en.wikipedia.org/wiki/Help:Watching_pages [http://perma.cc/T3B2-V8CS].

184 *See* AYERS ET AL., *supra* note 169, at 65-76.

185 *See Terms of use/FAQ on paid contributions without disclosure*, WIKIPEDIA, https://meta.wikimedia.org/wiki/Terms_of_use/FAQ_on_paid_contributions_without_disclosure [https://perma.cc/5H24-FQSZ].

faith."[186] Editors are expected to adhere to these attitudinal norms while editing pages in order to advance an extensive list of substantive standards for articles (they are expected, for example, to make entries verifiable by citing reliable sources[187] and to craft entries of appropriate length[188]) and are expected to follow extensive procedural rules[189] (for example, the procedures for deleting a controversial entry[190]). Reproducing these norms requires an immense amount of work.[191] In fact, simply learning the social ropes of Wikipedia can be notoriously discouraging to new members.[192] The endless restatement of Wikipedian norms—often in the process of accusing others of violating them—testifies to just what a big job this is in a community the size of Wikipedia.[193] Indeed, Wikipedia has an extensive parallel architecture of talk pages devoted to conversations about Wikipedia and its norms.[194] The norms of discourse here are rich. There is even a tradition of Wikipedia humor.[195]

Wikipedia does not run on exhortation alone. Beneath the surface, its moderators use the other verbs of moderation extensively to sustain positive norms. The most important decision is structural, and so deeply embedded in the idea of a wiki that it can be invisible: Wikipedia is highly modular, and its editorial work factors into loosely coupled subunits.[196] Wikipedia would not work—it could not work—if it consisted of a single massive webpage that only one person at a time could

---

[186] *See* REAGLE, *supra* note 169, at 45-71 (describing Wikipedia's "collaborative culture"); *see generally Wikipedia: Civility*, WIKIPEDIA, http://en.wikipedia.org/wiki/Wikipedia:Civility [http://perma.cc/FS37-ZP5H]; *Wikipedia: Neutral Point of View*, WIKIPEDIA, http://en.wikipedia.org/wiki/Wikipedia:Neutral_point_of_view [http://perma.cc/JMA4-9Z59]. Dariusz Jemielniak argues that Wikipedia's policy against personal attacks is central to its culture. *See* JEMIELNIAK, *supra* note 83, at 17-18.

[187] *Wikipedia: Verifiability*, WIKIPEDIA, http://en.wikipedia.org/wiki/Wikipedia:Verifiability [http://perma.cc/L425-K4DG].

[188] *Wikipedia: Article Size*, WIKIPEDIA, http://en.wikipedia.org/wiki/Wikipedia:Article_size [http://perma.cc/CY68-3V94].

[189] *See* AYERS ET AL., *supra* note 169, at 363-81 (summarizing policies).

[190] *Wikipedia: Deletion Process*, WIKIPEDIA, http://en.wikipedia.org/wiki/Wikipedia:Deletion_process [http://perma.cc/8P7A-NGXJ].

[191] *See generally* JEMIELNIAK, *supra* note 83; REAGLE, *supra* note 169.

[192] *See generally* E. GABRIELLA COLEMAN, CODING FREEDOM: THE ETHICS AND AESTHETICS OF HACKING 123-60 (2013) (describing the process of norm transmission in a community of open-source hackers).

[193] *See, e.g.*, JEMIELNIAK, *supra* note 83.

[194] *See, e.g.*, *id.* at 92-96.

[195] *See* AYERS ET AL., *supra* note 169, at 350-53; REAGLE, *supra* note 169, at 68-70.

[196] On modularity in general, see HERBERT A. SIMON, THE SCIENCES OF THE ARTIFICIAL (1969). For a powerful application of modularity to open collaborative projects, see Carliss Y. Baldwin & Kim B. Clark, *The Architecture of Participation: Does Code Architecture Mitigate Free Riding in the Open Source Development Model?*, 52 MGMT. SCI. 1116 (2006).

edit. Instead, it is split into different linguistic versions,[197] into WikiProjects for specific topics,[198] and into individual pages (each with its own associated Talk page for discussion).[199] For one thing, this factoring allows different editors to work in parallel, making independent decisions. For another, it allows different *groups* of editors to work in parallel, creating smaller and more cohesive subcommunities with a more localized sense of purpose and stronger shared norms.[200]

In sustaining its collaborative norms, Wikipedia also makes subtle and judicious compromises on openness, using deletion and exclusion in controversial but probably necessary ways. Take a simple act of vandalism akin to the one that brought down the *Los Angeles Times* wikitorial: changing the Wikipedia page on the Iraq War to say "FUCK USA." Wikipedia has entire projects devoted to fighting vandalism.[201] Some users sign up for a "recent changes patrol" or "counter-vandalism unit" and watch for suspicious changes to attractive targets, like politically controversial pages.[202] When they see an obvious act of vandalism, they revert the edit and restore the page to its previous state. This is *ex post*, distributed, transparent, human deletion. Other users run bots that watch for recent changes and revert changes that are especially likely to be vandalism, as when a page goes from thousands of words to two.[203] This is *ex post*, distributed, transparent, automatic deletion. These anti-vandalism efforts are why, despite the large number of bogus edits daily, most Wikipedia articles are in good shape most of the time. Vandals who don't succeed quickly tend to give up quickly.[204]

---

197 *See* AYERS ET AL., *supra* note 169, at 407-18; LIH, *supra* note 11, at 133-67.

198 *See* AYERS ET AL., *supra* note 169, at 212-16.

199 *See generally id.* at 99-117; Almila Akdag Shah et al., *Generating Ambiguities: Mapping Category Names of Wikipedia to UDC Class Numbers*, *in* CRITICAL POINT OF VIEW: A WIKIPEDIA READER 63-77 (Geert Lovink & Nathaniel Tkacz eds., 2011) (describing complexities in Wikipedia's classification systems).

200 This is an example of an effect described by Howard Rheingold: online, small and dispersed communities of interest can find each other for collaborative purposes. *See generally* RHEINGOLD, *supra* note 20.

201 *See Wikipedia: Cleaning up Vandalism*, WIKIPEDIA, http://en.wikipedia.org/wiki/Wikipedia:Cleaning_up_vandalism [http://perma.cc/H8WP-5NKS].

202 *Wikipedia: Recent Changes Patrol*, WIKIPEDIA, http://en.wikipedia.org/wiki/Wikipedia:Recent_changes_patrol [http://perma.cc/4DKH-UJCB]; *Wikipedia: Counter-Vandalism Unit*, WIKIPEDIA, http://en.wikipedia.org/wiki/Wikipedia:Counter-Vandalism_Unit [http://perma.cc/UFZ3-LTET].

203 *See* Geiger, *supra* note 102; Jesse Hicks, *This Machine Kills Trolls*, THE VERGE, Feb. 18, 2014, http://www.theverge.com/2014/2/18/5412636/this-machine-kills-trolls-how-wikipedia-robots-snuff-out-vandalism [http://perma.cc/5HGX-DB2P]. Bots are also useful for repetitive tasks such spell-checking and filling basic articles with standardized information (e.g., county demographics). *See* LIH, *supra* note 11, at 99-106.

204 *See* ZITTRAIN, *supra* note 178, at 138-39.

Wikipedia has sterner stuff in store when reversion isn't enough. Some articles are "semi-protected": only logged-in users can edit them, thus adding enough of a cost barrier to deter casual vandals.[205] Highly controversial articles may be "fully protected": only the much smaller group of administrators can make changes to them.[206] Protection switches from *ex post* to *ex ante* deletion, trading off openness for better protection against cacophony, abuse, and manipulation. Protection is regularly used not merely to prevent norm-defying users from making changes, but also to reassert norms without aliening users by establishing "cooling off" periods.[207]

When protection doesn't work, Wikipedia can also act against users themselves. Those who engage in large-scale vandalism or serious abuse, or who flout other important community policies, can be banned. Their accounts are prevented from making any edits at all, either to a few specific pages, or in severe cases, to Wikipedia as a whole.[208] Since there are also anonymous users and banned users who return with sock-puppet accounts, Wikipedia also blocks anonymous edits from some IP addresses entirely.[209] Banning and blocking are centralized, *ex post*, human, transparent exclusion. These methods are imposed by administrators through a review process that includes appeals.

As this discussion suggests, Wikipedia has a complicated relationship to identity. On the one hand, the fundamental "anyone can edit" policy acts as a strong check on pressures to prevent all anonymous edits.[210] Indeed, it gives Wikipedia a

---

[205] *Wikipedia: Protection Policy*, WIKIPEDIA, http://en.wikipedia.org/wiki/Wikipedia:Protection_policy [http://perma.cc/BV9Q-DT2A].

[206] *Id.* Currently, there are about 1,400 administrators on the English-language Wikipedia, who are selected through discussion and voting by other Wikipedians. *See Wikipedia: Administrators*, WIKIPEDIA, http://en.wikipedia.org/wiki/Wikipedia:Administrators [http://perma.cc/X7VB-7F95].

[207] *Wikipedia: Banning Policy*, WIKIPEDIA, http://en.wikipedia.org/wiki/Wikipedia:Banning_policy [http://perma.cc/FM6L-SN33].

[208] *See* R. Stuart Geiger & David Ribes, *The Work of Sustaining Order in Wikipedia: The Banning of a Vandal*, PROC. ACM CONF. ON COMPUTER SUPPORTED COOPERATIVE WORK (2010), http://www.stuartgeiger.com/papers/cscw-sustaining-order-wikipedia.pdf [http://perma.cc/7Y46-AGEP].

[209] *Wikipedia: Blocking IP Addresses*, WIKIPEDIA, http://en.wikipedia.org/wiki/Wikipedia:Blocking_IP_addresses [http://perma.cc/D37W-2RJM]. In 2014, Wikipedia blocked an IP address associated with the House of Representatives from anonymous edits because of "disruptive editing." *See* Abby Phillip, *Wikipedia Blocks Anonymous Edits (and Trolling) from a Congressional IP Address*, WASH. POST SWITCH BLOG (July 24, 2014), http://www.washingtonpost.com/blogs/the-switch/wp/2014/07/24/wikipedia-blocks-anonymous-edits-and-trolling-from-a-congressional-ip-address [http://perma.cc/LA5W-U6F4].

[210] *See Wikipedia: Introduction*, WIKIPEDIA, http://en.wikipedia.org/wiki/Wikipedia:Introduction [http://perma.cc/2Q2J-DEEU] ("Don't be afraid to ed-

strong (though much-criticized[211]) anti-expert ethos, in which offline credentials are nominally considered irrelevant to one's authority as an editor.[212] On the other hand, for registered users, Wikipedia is a surveillance society: user pages track one's complete editing history.[213] Reputation plays a major role in the Wikipedia community. One is expected to have a substantial history of numerous productive edits to be accepted as a trusted voice.[214] Editors also celebrate each other's work, individually with "barnstars" (images awarded for feats of hard but valuable work),[215] and collectively by making especially good articles "featured" on the Wikipedia homepage,[216] thereby using reputation to fuel positive norms.

Relatedly, transparency is a key aspirational virtue. Because every edit is logged, Wikipedians are expected to explain and if necessary defend their actions in sometimes excruciating detail. The process of being given administrator privileges can involve a harrowing examination of one's editing history, often by other editors with an axe to grind.[217] Decisions on everything from whether to rename a page to whether to ban a user are also debated publicly, often ad nauseam. Opacity is anathema. A persistent, if overblown, criticism of administrators is that they have access to private mailing lists.[218] Wikipedia's

---

it—*anyone* can edit almost every page, and we are encouraged to be bold.").

[211] *See, e.g.*, DALBY, *supra* note 169, at 50-81 (collecting criticisms); *Criticism of Wikipedia*, WIKIPEDIA, http://en.wikipedia.org/wiki/Criticism_of_Wikipedia [http://perma.cc/5UE4-T2PR]. *But see* Clay Shirky, *Larry Sanger, Citizendium, and the Problem of Expertise*, MANY 2 MANY (Sept. 18, 2006) (arguing that that Wikipedia succeded where Nupedia failed because it avoided the institutional overhead costs created by deference to experts).

[212] *See* JEMIELNIAK, *supra* note 83, at 106-124 (arguing that Wikipedia trusts procedures rather than people). Famously, Philip Roth was considered not to be a reliable source when changing an entry to describe the origins of his novel, *The Human Stain*. *See* Phillip Roth, *An Open Letter to Wikipedia*, THE NEW YORKER, Sept. 6, 2012, http://www.newyorker.com/books/page-turner/an-open-letter-to-wikipedia [http://perma.cc/TN3T-KB7G].

[213] *See* JEMIELNIAK, *supra* note 83, at 87-99 (discussing control through tracking); *see also* Geiger, *supra* note 102, at 83-92 (describing extensive controversy over a bot that added signatures to users' comments on talk pages).

[214] *See* JEMIELNIAK, *supra* note 83, at 39-41.

[215] *See* AYERS ET AL., *supra* note 169, at 333-34. *Wikipedia: Barnstars*, WIKIPEDIA, http://en.wikipedia.org/wiki/Wikipedia:Barnstars [http://perma.cc/7LQU-L3RV]. For the history and nomenclature of Barnstars, see *BarnStar*, MEATBALLWIKI, http://meatballwiki.org/wiki/BarnStar [http://perma.cc/AM8X-YEP3].

[216] *See* AYERS ET AL., *supra* note 169, at 227-28; *Wikipedia: Featured Articles,* WIKIPEDIA, http://en.wikipedia.org/wiki/Wikipedia:Featured_articles [http://perma.cc/H775-JTP5].

[217] *See* JEMIELNIAK, *supra* note 83, at 37-50.

[218] *See* JEMIELNIAK, *supra* note 83, at 50-58; Ayelet Oz, "Move Along Now, Nothing to See Here": The Private Discussion Spheres of Wikipedia (Aug.

use of open-source software and freely licensed contributions also mean that forking is always a possibility.[219]

Wikipedia's dispute resolution system is complex and multi-tiered, as might be expected from a project as capacious and contentious as creating a global encyclopedia. Officially, at least, it tries to operate by consensus.[220] Initially, many differences of opinion are simply argued over until one side or the other is either persuaded or gives up.[221] Other questions, such as whether to delete an article or how best to describe a political issue neutrally, may be put to a vote of all interested Wikipedians. The votes themselves are usually non-binding; they serve instead as a tool for measuring consensus. When that doesn't suffice, both a Mediation Committee and an Arbitration Committee exist to hear disputes through a relatively formal multi-level process.[222] Beyond that, the nonprofit Wikimedia Foundation, which oversees Wikipedia in the role of owner,[223] can ultimately step in, although it generally tries to avoid be-

---

29, 2009), http://papers.ssrn.com/sol3/papers.cfm?abstract_id=1726450 [http://perma.cc/U2MZ-GADA].

219 In a fork, a group of participants makes a copy of another shared, freely-licensed informational resource and work on the newly independent copy rather than on the original. *See* Andrew Famigletti, *The Right to Fork: A Historical Survey of De/Centralization in Wikipedia*, *in* CRITICAL POINT OF VIEW: A WIKIPEDIA READER 296-308 (Geert Lovink & Nathaniel Tkacz eds., 2011); Nathaniel Tkacz, *The Politics of Forking Paths*, *in* CRITICAL POINT OF VIEW: A WIKIPEDIA READER 94-107 (Geert Lovink & Nathaniel Tkacz eds., 2011). The most famous fork of Wikipedia is Citizendium, created by Wikipedia co-founder Larry Sanger to give more deference to credentialed experts. *See* Timothy B. Lee, *Citzendium Turns Five, But the Wikipedia Fork is Dead in the Water*, ARS TECHNICA. Oct. 27, 2011, http://arstechnica.com/tech-policy/2011/10/five-year-old-wikipedia-fork-is-dead-in-the-water [http://perma.cc/5ZPU-UNYV] (discussing history of Citizendium).

220 *See* REAGLE, *supra* note 169, at 97-115.

221 *See* JEMIELNIAK, *supra* note 83, at 76-81 (describing different "trajectories" that conflicts on Wikipedia can take).

222 *See* JEMIELNIAK, *supra* note 83, at 81-84; David Hoffman & Salil Mehra, *Wikitruth Through Wikiorder*, 59 EMORY L.J. 151 (2010) (arguing that Wikipedia's dispute resolution procedures work mainly by weeding out problematic users who will not adhere to Wikipedia's discourse norms); Sara Gwendolyn Ross, *Your Day in 'Wiki-Court': ADR, Fairness, and Justice in Wikipedia's Global Community* (Osgoode Legal Studies, Research Paper No. 56, 2014).

223 *See* AYERS ET AL., *supra* note 169, at 447-53; Mayo Fuster Morell, *The Wikimedia Foundation and the Governance of Wikipedia's Infrastructure*, *in* CRITICAL POINT OF VIEW: A WIKIPEDIA READER 325-41 (Geert Lovink & Nathaniel Tkacz eds., 2011); *see generally* Jyh-An Lee, *Organizing the Unorganized: The Role of Nonprofit Organizations in the Commons Communities*, 50 JURIMETRICS 275 (2010) (arguing that the nonprofit organizational form presents distinctive trust advantages for stewards of commons-based communities).

coming involved in specific issues.[224] Wikipedia takes its community democracy as seriously as it can.

Wikipedia is a sprawling, messy, and often bureaucratic organization. It combines norm-setting, exclusion, pricing, and organization. Its moderation is human, automatic, transparent, opaque, *ex ante*, *ex post*, centralized, and distributed. Its participants fight about everything—from the Shakespeare authorship question to the choice between "Gdańsk" and "Danzig"[225]—at great length. These endless wrangles are not simply wasted breath, or signs of a community about to crack apart. They are part and parcel of why Wikipedia works. The free encyclopedia is not free: its participants create it at great expense of time and effort. Not all of that effort goes into research and writing. The greater part of it is spent on the community-oriented work that actually holds Wikipedia together—the work of moderation.

B. *The* Los Angeles Times *Wikitorial*

The *Los Angeles Times* ignored all of this. Like Wikipedia, it was open to the world, but it had none of Wikipedia's devices for helping the well intentioned collaborate while keeping the ne'er-do-wells at bay. Unlike Wikipedia, the *Times* had no way to block persistently harmful users—not even a mechanism to track and identify the worst abusers. Unlike Wikipedia, it had no back channel for users to converse and develop community norms[226] or dispute-resolution mechanisms to contain conflict, and the experiment failed long before they could evolve. The *Times* forced users with strongly divergent beliefs on a controversial topic together, exacerbating normative conflict.[227] It brought them together for a one-off project, with no long-term reputations to recognize trustworthy members of the community. It had no dedicated cadre of administrators cleaning up destructive edits. Vandals who saw the broken windows decided to storm the front door.

---

[224] *See* JEMIELNIAK, *supra* note 83, at 125-44 (describing the sometimes fraught relationship between the Foundation and the Wikipedia community); Shun-Ling Chen, *The Wikimedia Foundation and the Self-Governing Community: A Dynamic Relationship Under Constant Negotiation*, *in* CRITICAL POINT OF VIEW: A WIKIPEDIA READER 351-69 (Geert Lovink & Nathaniel Tkacz eds., 2011). Previously, Jimmy Wales acted as "benevolent dictator" before voluntarily sidelining himself to reduce controversy. *See* REAGLE, *supra* note 169, at 117-35 (theorizing the concept of leadership via benevolent dictatorship)

[225] *See* JEMIELNIAK, *supra* note 83, at 65-76 (describing four-year edit war); LIH, *supra* note 11, at 122-32.

[226] *See* OSTROM, *supra* note 38, at 100-01 (emphasizing the importance of such fora).

[227] Well into the experiment, Wikipedia's Jimmy Wales tried to split the editorial into pro-war and anti-war versions to separate the warring camps, but by then it was too late. *See* Rainey, *supra* note 8.

The *Los Angeles Times* embraced Wikipedia's technology and its commitment to distributed organization, but neglected its commitment to positive social norms. The wikitorial had neither hard security nor soft. There were two kinds of naiveté at work. First, there was the assumption that communities operate without moderation, that broad participation by itself suffices. Second, there was the assumption that even if moderation were needed, it would develop spontaneously. The *Los Angeles Times* neither moderated the wikitorial effectively nor created the conditions under which the community of participants could develop its own effective moderation. The wikitorial was cargo cult collaboration.[228]

C. *MetaFilter*

Not many online communities can say that they came together to save people from human trafficking, but MetaFilter can.[229] Two young Russian women had come to the United States in May 2010 to work as lifeguards on Virginia Beach, but when they arrived, their contact instead told them to come to a bar on Long Island, the Lux Lounge, for some unspecified hostess work. Annoyed at their unreliable employer but enjoying their American adventure, they called Dan Reetz, an American they had befriended in Russia two years before. Reetz, who recognized the telltale signs of a sex-trafficking ring, was immediately alarmed. But he "was in his car on a highway in Wyoming with all his earthly belongings on his way to start a new job," and couldn't convince his friends of the danger they were in.[230] Instead, he took the situation to MetaFilter.

MetaFilter, founded in 1999 by programmer and blogger Matthew Haughey, calls itself a "community weblog."[231] It hosts discussions on user-submitted topics on a text-heavy

---

[228] *Cf.* Richard Feynman, *Cargo Cult Science*, *in* "SURELY YOU'RE JOKING, MR. FEYNMAN!": ADVENTURES OF A CURIOUS CHARACTER 338, 342 (1997) ("[T]hey follow all the apparent precepts and forms . . . but they're missing something essential."); *see also* Aaron Swartz, *Making More Wikipedias*, RAW THOUGHT (Sep. 14, 2006), http://www.aaronsw.com/weblog/morewikipedias ("For the most part, people have simply assumed that Wikipedia is as simple as the name suggests: install some wiki software, say that it's for writing an encyclopedia, and *voila*!—problem solved.").

[229] *See* Stephen Thomas, *The Internet's First Family*, HAZLITT (Oct. 31, 2014), http://penguinrandomhouse.ca/hazlitt/longreads/internets-first-family [http://perma.cc/R8WT-AMSZ]. The story unfolded in real time on the Metafilter thread *Help me help my friend in DC*, at http://ask.metafilter.com/154334/Help-me-help-my-friend-in-DC [http://perma.cc/6P36-WWPA], and sparked extensive discussion on a related thread, *The kindness of strangers*, http://metatalk.metafilter.com/19304/The-kindness-of-strangers [http://perma.cc/67LP-PZY2].

[230] Thomas, *supra* note 229.

[231] METAFILTER, http://www.metafilter.com [http://perma.cc/4HSR-RDYS].

website with a laid-back feel and a simple interface.[232] Each discussion—initiated when a logged-in user clicks on an unobtrusive text link to create a new post—starts with a link to a page somewhere else on the web, along with a description to provide some context. Each new post then appears on the front page of the site, where the posts are sorted in reverse chronological order like a blog. Click on a post, and you go to a dedicated page for that post, where you can read previous readers' comments and add your own. Users who aren't logged in can read posts and comments, but not add their own. The Metafilter community's interests are diverse. As I write this, the first three links feature rare concert footage of the Velvet Underground,[233] an effort to stop public urination in India,[234] and pre-WWII African-American science fiction.[235]

Reetz posted to Ask MetaFilter, a sister site for questions to the community. His post went up at 5:09 PM, and increasingly concerned community members started exchanging information about the bar, the girls' situation, and human trafficking resources. The next morning, another MetaFilter user, an unemployed nanny named Katherine Gutierrez, operating on almost no sleep, thought she might be able to divert the girls from the very bad idea of going to the Lux Lounge at midnight. She got their phone number from Reetz, and gave them a call. To avoid alarming them about her own intentions, she "presented herself as Just Another Fun-loving Young Gal In The Big City, Much Like Yourselves, and told the girls she'd gotten their numbers from a mutual friend and would be happy to hang out and show them around."[236] It worked. With the assistance of some plainclothes police and many other MetaFilter users, she convinced the girls not to go to their meeting with the mysterious and menacing "George." Instead, "they ultimately stayed with her for a full month, during which time MeFites [MetaFilter users] in New York and around the country sent the out-of-work nanny money to help feed the girls, and helped also in other ways, such as taking the girls out on

---

[232] *See generally Frequently Asked Questions*, METATALK, http://faq.metafilter.com [http://perma.cc/N49A-URA6] [hereinafter *MetaFilter FAQ*]; METAFILTER WIKI, http://mefiwiki.com/wiki/Main_Page [http://perma.cc/4AP9-GPSV].

[233] item, *Velvet Underground / Exploding Plastic Inevitable Live in Boston 1967*, METAFILTER (May 26, 2014), http://www.metafilter.com/139386 [http://perma.cc/WKR8-4UE3].

[234] KokoRyu, *"How can India stop people urinating in public?"*, METAFILTER (May 26, 2014), http://www.metafilter.com/139385 [http://perma.cc/T57P-BMF8].

[235] Martin Wisse, *Before Delany, before Butler*, METAFILTER (May 26, 2014), http://www.metafilter.com/139384 [http://perma.cc/9DE6-NUZ3].

[236] Thomas, *supra* note 229.

the town and putting them in touch with immigration lawyers and employment agencies."[237]

That's MetaFilter in a nutshell (albeit its very best nutshell). Let's dig in to how and why it works so well. In moderation terms, *ex post,* centralized, human norm-setting dominates, with editing and exclusion (and a tiny bit of pricing) in supporting roles. The central technique of moderation is simple: Haughey and a small group of paid moderators[238] read posts and comments and take action against inappropriate ones.[239] Some receive a gentle chiding, in the form of a comment or email; others are deleted.[240] Deleted posts are visible on the site (since people may have left comments on them or saved the URL), but they carry a short notice of why they were removed—for example, because there is already an active discussion of the same story or issue on the site in another post.[241] For particularly controversial or important actions, the moderators or concerned users will create a discussion post on MetaTalk, another sister site for conversations about Metafilter itself.[242]

The overriding goal is to maintain positive community norms. In its initial days, Haughey was the primary author of posts and was active in all discussion threads to set a good example. The moderators' policy of hiding inappropriate material quickly reinforces positive norms by making good behavior far more visible than bad. The explanations treat users who make mistakes as well-intentioned, and indicate that they are still welcomed members of the community. These discussions invite broad participation in articulating and shaping the community's norms. This is deletion in service of social norms as much as it is deletion for its own sake. The moderators enjoy substantial credibility on the site not just by virtue of their author-

---

[237] *Id.*

[238] *See Mods*, MEFI WIKI, http://mefiwiki.com/wiki/Moderators [http://perma.cc/Y8CP-R3SC]. While this Article was in press, Haughey announced his retirement from MetaFilter to take up a day job, handing over the moderation reins to the other members of the team. *See* mathowie [Matthew Haughey], *Sixteen Years*, METATALK (Mar. 4, 2015), http://metatalk.metafilter.com/23626/Sixteen-Years [http://perma.cc/6B6R-LQKS].

[239] *See* Matt Haughey, Real World Moderation: Lessons from 11 Years of Community, Presentation at SXSW Interactive (Mar. 12, 2011), *available at* https://vimeo.com/21043675 [https://perma.cc/J3GP-SGKH].

[240] *See* Warnick, *supra* note 20, at 120-23.

[241] *E.g.*, tofu_crouton [Sara Gore], *We Must Not Call Him Sister*, METAFILTER (July 28, 2014), http://www.metafilter.com/141401/We-Must-Not-Call-Him-Sister [http://perma.cc/JXV2-G4QN] ("This post was deleted for the following reason: This kinda feels like a big fight in the making for no particular good reason. –cortex."); *see generally MetaFilter FAQ*, *supra* note 232; METAFILTER DELETED POSTS, http://mefideleted.blogspot.com [http://perma.cc/UK6C-XKGE].

[242] METATALK, http://metatalk.metafilter.com [http://perma.cc/CC7E-ZBRW].

ity, but because "they've proven over and over again that they understand how communities work and deal with most issues patiently and courteously."[243] Humility is a key virtue. Even a recent visual makeover was an occasion for consultation rather than simply being imposed from the top down.[244]

At the same time, the moderators and the community call out particularly noteworthy posts and comments for praise. Haughey maintains a small sideblog[245] and a Twitter feed[246] that he uses to link to high-quality posts. There is a small, unobtrusive button on Metafilter to mark any post or comment as a "favorite". The counts appear next to the post or comment and on the user's profile page, functioning as a visible symbol of community praise. On Ask Metafilter, the question-asker can flag replies as "best answers," again a visible symbol of praise.[247] Users participate extensively in the explicit norm-setting, too. They post comments rebuking and praising each other,[248] take their debates to MetaTalk,[249] and occasionally flag inappropriate posts and comments to bring them to the moderators' attention.[250] A lightweight message system, MeFi Mail, gives members a private back channel.[251]

The other verbs of moderation appear almost entirely in secondary, supporting roles. There is a smattering of organization: posts can be tagged and searched. In a form of *ex ante* deletion, users are limited to one post per twenty-four hours (though very few come anywhere near that pace). Commenting, however, is unlimited. There is $5 signup fee for new members,[252] which looks like pricing but functions more as a speed bump to exclude participants not really interested in the com-

---

[243] *See* Warnick, *supra* note 20, at 101 (quoting MetaFilter user Rhaomi); *see also* Paul Lawton, Capital and Stratification Within Virtual Community: A Case Study of Metafilter.com 87-91 (2003) (unpublished B.A. dissertation, University of Lethbridge), https://www.uleth.ca/dspace/bitstream/handle/10133/267/MR17405.pdf [https://perma.cc/9AU8-E52N].

[244] *See* mathowie [Matt Haughey], *A new theme for MeFi: Modern*, METATALK (Sept. 24, 2014), http://metatalk.metafilter.com/23445/A-new-theme-for-MeFi-Modern [http://perma.cc/BQ5D-MRRA].

[245] BEST OF METAFILTER, http://bestof.metafilter.com [http://perma.cc/WHM6-264S].

[246] *@MetaFilter*, TWITTER, https://twitter.com/metafilter [https://perma.cc/5X4V-G2DC].

[247] ASK METAFILTER, http://ask.metafilter.com [http://perma.cc/2J27-EWKB].

[248] *See* Leiser Silva, Lakshmi Goel & Elham Mousavidin, *Exploring the Dynamics of Blog Communities: The Case of MetaFilter*, 19 INFO. SYS. J. 55, 67-73 (2008) (describing debates shaping norms of "good" and "bad" posts).

[249] *See* Lawton, *supra* note 243, at 70-85; Warnick, *supra* note 20, at 89-91 (breaking down rhetorical functions of MetaTalk posts).

[250] *See MetaFilter FAQ*, *supra* note 232.

[251] *See id.*

[252] *Create a New User*, METAFILTER, http://www.metafilter.com/newuser.mefi [http://perma.cc/BQ5D-MRRA].

munity.[253] Those who misbehave have their accounts deactivated, no refunds offered. The real pricing consists of some lightweight advertising.[254] Most ads are hidden for members;[255] the site sustains itself primarily on the revenue from outsiders who come across particular pages on web searches.[256]

MetaFilter's treatment of identity is carefully modulated. It is easy to browse a user's history of posts and comments, and some users choose to decorate their profiles with detailed information about themselves,[257] but the default is persistent pseudonymity.[258] Members are known primarily by their usernames, such that dedicated discussants can build up extensive reputations on MetaFilter without revealing their real-life identities. In one part of the site, however, these rules are suspended: members can post anonymous questions on AskMetaFilter—just the thing for seeking advice on an abusive relationship or a difficult workplace issue.[259] Because of the decreased norm-based constraints on abusive questions, anonymous questions go through *ex ante* human moderator review

---

[253] *See* Hannah Pileggi, Brianna Morrison & Amy Bruckman, *Deliberate Barriers to User Participation on MetaFilter* (Georgia Institute of Technology School of Interactive Computing, Technical Report No. GT-IC-14-01 2014), https://smartech.gatech.edu/handle/1853/50776 [https://perma.cc/E55N-CBUR].

[254] *See How Does Advertising on MetaFilter Work?*, METAFILTER FAQ, http://faq.metafilter.com/130/how-does-advertising-on-metafilter-work [http://perma.cc/AB92-6HSF].

[255] *Id*.

[256] MetaFilter's biggest threat currently comes not from internal community dynamics but from a shift in the online advertising ecosystem. In November 2012, Google changed its ranking algorithms in a way that appears to have significantly demoted MetaFilter, instantly slashing the site's traffic from non-members and with it, the site's advertising revenue. *See* Matt Haughey, *On the Future of MetaFilter*, MEDIUM (May 21, 2014), https://medium.com/technology-musings/on-the-future-of-metafilter-941d15ec96f0 [https://perma.cc/G28F-8USR]. Haughey was forced to lay off three members of the already small moderation staff as a result. *See* mathowie [Matt Haughey], *The State of MetaFilter*, METATALK (May 19, 2014), http://metatalk.metafilter.com/23245/State-of-MetaFilter [http://perma.cc/JA6V-R5YH]. The exact reason for the decline in MetaFitler's Google rankings remains unclear. The change seems to have been unintentional on Google's part, but Google has not acted decisively to fix it. *See* Danny Sullivan, *On MetaFilter Being Penalized By Google: An Explainer*, SEARCH ENGINE LAND (May 22, 2014), http://searchengineland.com/metafilter-penalized-google-192168 [http://perma.cc/2RNQ-WPJ9].

[257] *See* Noor Ali-Hasan, *MetaFilter: Analysis of a Community Weblog* (2005), http://www.nooratwork.com/pdf/ali-hasan_metafilter.pdf [http://perma.cc/7SRF-PHDE].

[258] *See MetaFilter FAQ*, *supra* note 232.

[259] *Id.*; akomom, *confused about how to successfully ask anonymously*, ASK METAFILTER (Mar. 27, 2012), http://metatalk.metafilter.com/21588/confused-about-how-to-successfully-ask-anonymously [http://perma.cc/5YJ6-CZZ2].

rather than *ex post*.[260] "Anonymous questions are for basic privacy, not for hiding from Interpol."[261]

Overall, MetaFilter is a moderation success story. In 2013, it had about 40,000 active users, who created about 11,000 posts and 600,000 comments.[262] The flow of posts is consistently interesting but not overwhelming. Although the site is occasionally challenged by vandals or infiltrated by marketers, it is for the most part an authentic conversation among engaged participants. Users have been known to spend weeks crafting massive posts that do deep dives on a topic, collecting hundreds of related links into a perfectly curated collection.[263] In 2001, the MetaFilter community collaboratively exposed an Internet hoax—a fictional teenager with equally fictional terminal leukemia.[264] The site has been self-sustaining for years, and there are strong feelings of belonging and community among active posters. There are face-to-face meet-ups in major (and minor) cities,[265] a rich vocabulary of references and in-jokes,[266] a holiday gift exchange called Secret Quonsar,[267] and at least one

---

[260] *See MetaFilter FAQ*, *supra* note 232.

[261] *Id.*

[262] *See MetaFilter Stats 2013*, MEFI LABS, http://labs.metafilter.com/mefi-stats-2013 [http://perma.cc/GG7T-4JYB].

[263] For a particularly outstanding example, see Miko, *Alice's Restaurant*, METAFILTER (Nov. 25, 2010), http://www.metafilter.com/97904/Alices-Restaurant [http://perma.cc/5FKC-EZ6X], a post celebrating the Arlo Guthrie song "Alice's Restaurant" by extensively hyperlinking the song's lyrics; *see also What Is a Good Post*, MEFI WIKI, http://mefiwiki.com/wiki/What_Is_A_Good_Post [http://perma.cc/D5WB-QGDN]; *Posting Guidelines*, METAFILTER, http://www.metafilter.com/guidelines.mefi [http://perma.cc/H52N-T8S5].

[264] *See* acidrabbit, *Is it possible that Kaycee did not exist?*, METAFILTER (May 19, 2001), http://www.metafilter.com/7819/Is-it-possible-that-Kaycee-did-not-exist [http://perma.cc/A74A-J9SM]; Katie Hafner, *A Beautiful Life, an Early Death, a Fraud Exposed*, N.Y. TIMES, May 31, 2001, http://www.nytimes.com/2001/05/31/technology/a-beautiful-life-an-early-death-a-fraud-exposed.html [http://perma.cc/K3MV-QMJS].

[265] Indeed, MetaFilter has an entire subsite devoted to meetups. MEFIIRL, http://irl.metafilter.com [http://perma.cc/BS3W-ENYL]; see also Lauren F. Sessions, *How Offline Gatherings Affect Online Communities*, 13 INFO., COMM., & SOC. 375-95 (2010) (analyzing the effect of meetups on the MetaFilter community).

[266] *See In Jokes*, MEFI WIKI, http://mefiwiki.com/wiki/In_Jokes [http://perma.cc/9KBM-C8K8]. For example, "Pepsi Blue" is "a sort of catch-all cat-call for something that is a possible shill posting on MetaFilter—that is, an ad or product endorsement for reasons other than just overall consumer joy." *See Pepsi Blue*, MEFI WIKI, http://mefiwiki.com/wiki/Pepsi_Blue [http://perma.cc/F7NT-QWNM]. More seriously, it is common to leave comments consisting solely of a period, as a moment of silent mourning. *See The Period*, MEFI WIKI, http://mefiwiki.com/wiki/The_Period [http://perma.cc/7LYK-H9BW].

[267] *See Secret Quonsar*, MEFI WIKI, http://mefiwiki.com/wiki/Secret_Quonsar [http://perma.cc/U4TV-2KQF]. Quonsar was the username of a prolific, if problematic, MetaFilter user. *See* Lawton, *supra* note 243, at 100.

marriage of users who met through MetaFilter.[268] It is a strikingly different kind of community than Wikipedia with strikingly different moderation, but it also works.

D. *Reddit*

If the Portland-based MetaFilter is artisanal small-batch news, then the San Francisco-based Reddit is crowd-sourced post-industrial news.[269] Instead of MetaFilter's loving, centralized, *ex post*, human moderation with a strong emphasis on norm-setting, Reddit depends on finely machined, distributed, *ex post*, algorithmic moderation with a strong emphasis on annotation and filtering.[270] The contrast between them shows both the diversity of moderation and some of its recurring challenges.

Reddit's users moderate primarily by voting on content.[271] Each post and each comment are accompanied by two arrows. Click the up arrow, and the item gains an "upvote"; click the down arrow, and the item gains a "downvote." Reddit uses the upvotes and downvotes to determine the order in which posts and comments are displayed.[272] Well-liked posts bubble to the top and are seen by more users; disliked ones are rapidly driven down to invisibility. This is *ex post*, distributed, human annotation, used as an input to centralize automatic filtration. The algorithm that weights upvotes and downvotes has been carefully tuned both to maintain a fresh flow of new content

---

[268] *See* MrMoonPie, *the wedding of NortonDC and onlyconnect, who met at a meetup*, METATALK (Sept. 24, 2005), http://metatalk.metafilter.com/10231 [http://perma.cc/8EZL-C7EP].

[269] REDDIT, http://www.reddit.com [http://perma.cc/7QL4-JCJK].

[270] *See generally* Tom Lamont, *Reddit: How to Win the Internet*, THE GUARDIAN, Feb. 7, 2014, http://www.theguardian.com/technology/2014/feb/07/reddit-how-to-win-the-internet [http://perma.cc/M2MN-JYZV].

[271] *See Frequently Asked Questions*, REDDIT, http://www.reddit.com/wiki/faq [http://perma.cc/4UDT-7JV8] [hereinafter *Reddit FAQ*].

[272] *Id.* This core idea of upvotes and downvotes has been in widespread use for years. Slashdot, a social news site, pioneered the extensive reliance on algorithms to sort comments. *See generally* Lampe, *supra* note 84. Slashdot's system grew particularly baroque over time: it developed a system of "meta-moderation," in which users would examine each others' moderation decisions and then vote on whether those decisions were correct or incorrect. Users whose decisions were frequently voted correct would receive "karma" points, which in turn allowed them to moderate and meta-moderate more frequently. *Id.* For further discussion of karma systems, see FARMER & GLASS, *supra* note 16, at 75-82. Reddit passes posts themselves through the voting algorithm, not just comments. Thus, unlike on Slashdot or MetaFilter, where every post has been vetted or even edited by moderators affiliated with the site, the choice of which Reddit posts are prominent is in the hands of the voting algorithm. This is not unique to Reddit—the social news site Digg.com had it first—but it is characteristic of Reddit.

and to keep early votes from disproportionately influencing a post's or comment's fate.[273]

Reddit also relies on a layer of distributed organization. Any user can create a "subreddit" devoted to discussion on a particular topic.[274] Within the subreddit, the usual upvote and downvote mechanics apply, but moderators also enjoy substantial editorial discretion: they can remove "objectionable or off topic" comments and ban abusive users from the subreddit.[275] This, in effect, splits Reddit into a large number of smaller communities, each combining automated filtration with human deletion.[276] It also makes exiting an appealing option within Reddit: users who dislike a subreddit can easily avoid it.[277]

There are a few other techniques in use on Reddit, but they occupy subsidiary roles. The site has a custom advertising platform that allows either generic site-wide advertising or advertising targeted at the users of particular subreddits.[278] It also

---

[273] For explanations of the algorithmic details, see Michael Billard, *Reddit's Empire No Longer Founded on a Flawed Algorithm*, OUT OF SCOPE, Feb. 16, 2014, http://www.outofscope.com/reddits-empire-no-longer-founded-on-a-flawed-algorithm [http://perma.cc/2SNA-LYN3]; Randall Munroe, *Reddit's New Comment Sorting System*, REDDIT BLOG (Oct. 15, 2009), http://www.redditblog.com/2009/10/reddits-new-comment-sorting-system.html [http://perma.cc/2MZR-KYWX]; Jonathan Rochkind, *Reddit's Actual? (Or a Variation?) Story Ranking Algorithm Explained (Significant Typos in Previously Published Version (Or Not))*, BIBLIOGRAPHIC WILDERNESS (May 8, 2012), http://bibwild.wordpress.com/2012/05/08/reddit-story-ranking-algorithm [http://perma.cc/8MFX-CHFS]; and Amir Salihefendic, *How Reddit Ranking Algorithms Work*, HACKING AND GONZO (Nov. 23, 2010), http://amix.dk/blog/post/19588 [http://perma.cc/ECF5-HTAS].

[274] *See Reddit FAQ*, *supra* note 271.

[275] *Id.*

[276] *See Moderation*, REDDIT, http://www.reddit.com/wiki/moderation [http://perma.cc/9R3Q-74ND]. Users, in turn, can combine up to 100 subreddits into a personal "front page" that brings together posts from all of the subreddits they follow—another form of filtration. *See Reddit FAQ*, *supra* note 271. Users who are not logged in see a default front page combining posts from a curated list of fifty subreddits. *See* cupcake1713 [Alex Angel], *What's That, Lassie? The Old Defaults Fell Down a Well?*, REDDIT BLOG (May 7, 2014), http://www.redditblog.com/2014/05/whats-that-lassie-old-defaults-fell.html [http://perma.cc/X87N-ETZU]. There are also quotas on the number of front-page posts from each subreddit to prevent the largest and most popular subreddits from dominating the front page. *See* Todd W. Schneider, *The Reddit Front Page Is Not a Meritocracy*, TODD W. SCHNEIDER (Nov. 6, 2014), http://toddwschneider.com/posts/the-reddit-front-page-is-not-a-meritocracy [http://perma.cc/8LSB-47FS].

[277] *See* Adrian Chen, Reddit CEO Speaks Out On Violentacrez In Leaked Memo: 'We Stand for Free Speech', GAWKER, Oct. 16, 2012, http://gawker.com/5952349/reddit-ceo-speaks-out-on-violentacrez-in-leaked-memo-we-stand-for-free-speech [http://perma.cc/XH2A-KTA9].

[278] *See Advertise*, REDDIT, http://www.reddit.com/advertising [http://perma.cc/QU5Q-HCET]; Mike Isaac, *Can Reddit Grow Up?*, N.Y. TIMES, July 27, 2004, http://www.nytimes.com/2014/07/28/technology/can-reddit-grow-up.html [http://perma.cc/E2JQ-K3MD].

implements pricing through a "Gold" membership tier for $3.99 per month that hides ads and gives users a few more sophisticated filtration choices.[279] The site uses a spam filter to delete automated posts, and it fights hard against voting manipulation (such as using multiple accounts to upvote a post).[280] User accounts are banned for abuse (i.e., excluded),[281] and if a website is caught repeatedly trying to manipulate its way onto Reddit, the entire domain may be banned: all links to it are deleted *ex ante* on sight.[282] Norm-building meetups take place, but given the sheer size of the Reddit community, they reach only a small portion of the user population.[283]

In many ways, Reddit is transparent. Its source code, for example, is made publicly available.[284] But there is a strong undercurrent of opacity. A few operational details are shrouded in secrecy to protect the voting system from being gamed. Thus, actual upvote and downvote totals are "fuzzed" so that users cannot tell exactly which tactics are successfully getting past the vote-cheating detectors.[285] The site sometimes shadowbans spammers, letting them think their accounts are active and working, while quietly deleting their posts and ignoring their votes.[286] Individual subreddit moderators frequently push their personal political agendas by using their power to secretly delete content. Moderators have even been known to take bribes in exchange for promoting particular content.[287]

Something about the combination works.[288] After a few years of steady growth following its 2005 founding, Reddit took

---

[279] *See Gold*, REDDIT, http://www.reddit.com/gold/about [http://perma.cc/Z9H4-AKY5].

[280] *See Reddit FAQ*, *supra* note 271.

[281] *See id.*

[282] *See, e.g.*, Peter Bright, *Year-Long E-Sports Site Ban Shows the Dangers of Gaming Reddit*, ARS TECHNICA, July 3, 2014, http://arstechnica.com/gaming/2014/07/year-long-e-sports-site-ban-shows-the-dangers-of-gaming-reddit [http://perma.cc/F933-VKYA].

[283] *See, e.g.*, Matthew Shaer, *Reddit in the Flesh*, N.Y. MAG., July 8, 2012, http://nymag.com/news/features/reddit-2012-7 [http://perma.cc/N9TT-ZB2M].

[284] *See Reddit—Reddit*, GITHUB, https://github.com/reddit/reddit [https://perma.cc/Z4D7-2LHD].

[285] *See Reddit FAQ*, *supra* note 271.

[286] *See User Specific FAQs*, REDDIT, http://www.reddit.com/r/help/wiki/faq [http://perma.cc/M8KG-72LG]. *See also* cojoco, *An Unofficial Guide on How to Avoid Being Shadowbanned*, REDDIT, http://www.reddit.com/r/ShadowBan/comments/1x92jy/an_unofficial_guide_on_how_to_avoid_being [http://perma.cc/SF3M-2YX5].

[287] *See, e.g.*, David Auerbach, *Does Reddit Have a Transparency Problem?*, SLATE, Oct. 9, 2014, http://www.slate.com/articles/technology/technology/2014/10/reddit_scandals_does_the_site_have_a_transparency_problem.html [http://perma.cc/EHL8-2KSN].

[288] For an example of Reddit at its best, see Kevin Morris, *The Greatest Story Reddit Ever Told*, THE KERNEL, Nov. 2, 2014, http://kernelmag.dailydot

off like a rocket in the early 2010s.[289] Where MetaFilter is a water fountain, Reddit is a firehose. In 2013, Reddit had over 2 million users who made 41 million posts, 400 million comments, and 6.7 billion votes.[290] Reddit's "Ask Me Anything" crowd-sourced interviews have featured everyone from President Obama[291] to Jerry Seinfeld,[292] and one literally epic Reddit thread—who would win in a fight between a U.S. Marine Expeditionary Unit and the Roman Empire?—was optioned by Warner Brothers.[293]

The contrast between MetaFilter and Reddit is striking. Even though they have broadly similar missions—threaded discussions about things from around the web—the two sites have succeeded as communities for very different reasons. Metafilter relies on its core team of administrators to set consistent rules and norms across the site. Reddit, on the other hand, is built to scale. Its site-wide administrators tweak the algorithms occasionally but avoid almost all individual moderation decisions. All of those decisions are delegated either to the ranking algorithms or to the moderators of subreddits. MetaFilter works because almost all of its users want it to work, because its moderators are personally attuned to its users' interests, and because it offers a single coherent community. Reddit works because many of its users want it to work, because its algorithms are well-tuned to reflect its users' overall preferences, and because its subreddits are compartmentalized from each others' failures.

Every community has to deal with abuse. Reddit's responses show both the power and the limits of its moderation tech-

---

.com/issue-sections/headline-story/10727/dante-orpilla-youngluck-reddit gifts/ [http://perma.cc/MH7A-KHAQ].

289 *See* Farhad Manjoo, *The Great and Powerful Reddit*, SLATE, Jan. 19, 2012, http://www.slate.com/articles/technology/technology/2012/01/reddit_how_the_site_went_from_a_second_tier_aggregator_to_the_web_s_unstoppable_force_.html [http://perma.cc/5BP9-TWUP].

290 *See* hueypriest [Erik Martin], *Top Posts of 2013, Stats, and Snoo Year's Resolutions*, REDDIT BLOG (Dec. 31, 2013), http://www.redditblog.com/2013/12/top-posts-of-2013-stats-and-snoo-years.html [http://perma.cc/Q5CX-GTQH].

291 *See I am Barack Obama, President of the United States—AMA*, REDDIT, http://www.reddit.com/r/IAmA/comments/z1c9z/i_am_barack_obama_president_of_the_united_states [http://perma.cc/4FEF-6B8P].

292 *See Jerry Seinfeld here. I will give you an answer.*, REDDIT, http://www.reddit.com/r/IAmA/comments/1ujvrg/jerry_seinfeld_here_i_will_give_you_an_answer [http://perma.cc/CS2E-Y7XM]; *see generally* Ryan Holiday, *Inside the Reddit AMA: The Interview Revolution That Has Everyone Talking*, FORBES, May 1, 2012, http://www.forbes.com/sites/ryanholiday/2012/05/01/inside-the-reddit-ama-the-interview-revolution-that-has-everyone-talking [http://perma.cc/EVE4-V8CU].

293 *See* Jason Fagone, *How One Response to a Reddit Query Became a Big-Budget Flick*, WIRED, Mar. 20, 2012, http://www.wired.com/2012/03/ff_reddit/all [http://perma.cc/EVE4-V8CU].

niques. On the one hand, the combination of decentralization and filtering is often effective in enabling users to avoid content they dislike. While sometimes users choose to stay and argue over the direction of a subreddit, on the whole, exit dominates over voice.[294] Reddit invites dissatisfied users to "consider making a new subreddit and shaping it the way you'd like rather than performing a sit-in and/or witch hunt."[295]

Reddit therefore adopts a strongly libertarian official attitude toward free speech.[296] The administrators will not intervene to remove content. Users who dislike something are expected to avoid it rather than seek to have it removed. The only exception is when there is a legal requirement to remove content, and even then, the administrators make a show of acting only when compelled to.[297] Relatedly, Reddit users are encouraged to protect their privacy with pseudonymity. "It is thought bad form on Reddit to reveal your real name,"[298] and there is a strong norm against "doxxing"—revealing personal information about members without their consent.[299]

But if these features—strong subreddit communities, tolerance of differing views, and pseudonymous speech—make Reddit effective at defusing internal conflicts and catalyzing internal cooperation, they can make it downright dangerous to outsiders. Reddit is passionate about creating strong communities but completely indifferent as to whether those communities collaborate for good or for ill. After the Aurora shooting, Reddit was a leading source for sorting through the chaos of conflict-

---

[294] *See* Grimmelmann, *Anarchy*, *supra* note 25 (describing a controversy within /r/politics subreddit after moderators banned links from the left-leaning *Mother Jones*).

[295] *Id.*

[296] *See, e.g.*, yishan [Yishan Wong], *Fundraising for Reddit*, REDDIT BLOG (Sept. 30, 2014), http://www.redditblog.com/2014/09/fundraising-for-reddit.html [http://perma.cc/FA9E-B2NS] ("We believe in free speech, self-governing communities, and the power of voting."); Morris, *supra* note 288 ("The site's founders . . . instilled an institutional devotion to ideals of free speech, turning Reddit into an online petri dish for experiments in stretching the First Amendment to its breaking point.").

[297] This ethos made Reddit a crucial nexus in the online protests that halted the copyright-filtering bills SOPA and PIPA in 2012. Reddit was the first major site to announce a blackout for the day of protest. *See Stopped they must be; on this all depends*, REDDIT BLOG (Jan. 10, 2012), http://www.redditblog.com/2012/01/stopped-they-must-be-on-this-all.html [http://perma.cc/Z9FD-NLR7]; *see also* Tom Cheredar, *Reddit Goes Black Jan. 18 to Protest SOPA & PIPA—Who else will join?*, VENTUREBEAT, Jan. 10, 2012, http://venturebeat.com/2012/01/10/reddit-blackout-sopa-pipa [http://perma.cc/2JT6-3ZFH].

[298] Lamont, *supra* note 270.

[299] *See* C. S.-W., *What Doxxing Is, and Why It Matters*, THE ECONOMIST, Mar. 10, 2014, http://www.economist.com/blogs/economist-explains/2014/03/economist-explains-9 [http://perma.cc/T9X3-XBJJ].

ing reports.[300] But after the Boston Marathon bombing, an *ad hoc* community of Reddit users misidentified a missing Brown undergraduate, Sunil Tripathi, as one of the bombers, touching off a media firestorm and causing his family great distress.[301]

In some cases, entire subreddits are devoted to illegal and immoral purposes. Take /r/jailbait, which featured "sexually suggestive pictures of teenage girls, most of whom appear[ed] to be under the age of 18."[302] After CNN's Anderson Cooper ran an expose on /r/jailbait in 2011, its traffic spiked. Eventually Reddit staff shut it down amid allegations that users were trading actual child pornography.[303] A year later, the creator of /r/jailbait, a user with the name violentacrez, became involved in a similar controversy over /r/creepshots, "where users posted covert photos they had taken of women in public . . . for a voyeuristic sexual thrill."[304] This time, journalist Adrian Chen identified the person behind the violentacrez account, a programmer from Texas named Michael Brutsch.[305] The initial response from many subreddit moderators was defensive: they banned links to Gawker on the grounds that the story violated violentacrez's privacy.[306] Indeed, for a while, Reddit itself banned links to Gawker because of the unmasking.[307] In the end, the bad publicity was too much to withstand. Brutsch was fired from his job at a financial services company, Gawker was unbanned, and /r/creepshots was deleted.[308] But the story

---

[300] *See* Jay Caspian Kang, *Should Reddit Be Blamed for the Spreading of a Smear?*, N.Y. TIMES, July 25, 2013, http://www.nytimes.com/2013/07/28/magazine/should-reddit-be-blamed-for-the-spreading-of-a-smear.html [http://perma.cc/RU9Y-SJZS].

[301] *Id.*

[302] *See* Kevin Morris, *Anderson Cooper Addresses Reddit's Teen Pics Section*, THE DAILY DOT, Sept. 30, 2011, http://www.dailydot.com/news/anderson-cooper-jailbait-reddit [http://perma.cc/V59S-X4BL].

[303] *See* Kevin Morris, *Reddit Shuts Down Teen Pics Section*, THE DAILY DOT, Oct. 11, 2011, http://www.dailydot.com/society/reddit-r-jailbait-shutdown-controversy [http://perma.cc/DY2M-XAET].

[304] Adrian Chen, *Unmasking Reddit's Violentacrez, The Biggest Troll on the Web*, GAWKER, Oct. 12, 2012, http://gawker.com/5950981/unmasking-reddits-violentacrez-the-biggest-troll-on-the-web [http://perma.cc/Z457-87XA].

[305] *Id.*

[306] *See* Kevin Morris, *Clearing up Rumors and Hearsay as the Internet Eagerly Awaits the Gawker Reddit Story*, THE DAILY DOT, Oct. 12, 2012, http://www.dailydot.com/news/reddit-adrian-chen-violentacrez-gawker-rumors [http://perma.cc/2JUL-NAQL].

[307] *See* Katie Notopoulos, *Leaked Reddit Chat Logs Reveal Moderators' Real Concern*, BUZZFEEDNEWS (Oct. 13, 2012), http://www.buzzfeed.com/katienotopoulos/leaked-chat-logs-between-reddit-moderators-and-sta [http://perma.cc/M5VQ-UXLY].

[308] *See* Fernando Alfonso III, *Reddit's Most Notorious Troll Loses Job After Gawker Profile*, THE DAILY DOT, Oct. 15, 2012, http://www.dailydot.com/news/violentacrez-reddit-troll-fired-gawker-profile [http://perma.cc/J5KW-WWZG]. *But see* Fernando Alfonso III, *Creepshots Never Went Away—We Just Stopped Talking About Them*, THE DAILY DOT, Feb. 7, 2014, http://

shows how toxic Reddit's combination of tolerance and pseudonymity can be.

More recently, but just as alarmingly, Reddit played a prominent role in the 2014 release of nude photographs of celebrities such as Jennifer Lawrence and Kirsten Dunst. The photos were initially stolen by a loose-knit coalition of hackers who scour the Internet looking for enough personal information to gain access to the victims' online accounts.[309] The photos might have stayed hidden within the hackers' semi-private networks had it not been for the Reddit user johnsmcjohn, who created the /r/TheFappening subreddit to share them.[310] Within a day, the subreddit had tens of thousands of members.[311] Reddit's official response was a masterpiece of muddled messaging. CEO Yishan Wong wrote a blog post, portentously titled "Every Man Is Responsible For His Own Soul," that doubled down on Reddit's commitment to free speech, explaining "why we will *not* ban questionable subreddits, of which /r/TheFappening is one of them."[312] Almost simultaneously, and supposedly by complete coincidence, Reddit banned /r/TheFappening.[313] The stated reason for the ban was not that trading links to stolen nude photographs was wrong, or that trading links to stolen nude photographs was illegal, but that the burden of responding to DMCA takedown requests had become unsustainable in light of users' continual attempts to repost the photographs after each takedown.[314] Wong's explanation of Reddit's sense of itself is telling, and deserves to be quoted at length:

---

www.dailydot.com/lifestyle/reddit-creepshots-candidfashionpolice-photos [http://perma.cc/4B8U-CJEA].

309 *See* Nik Cubrilovic, *Notes on the Celebrity Data Theft*, NEW WEB ORDER (Sept. 2, 2014), https://www.nikcub.com/posts/notes-on-the-celebrity-data-theft [https://perma.cc/5NMQ-MJH5].

310 *See* Caitlin Dewey, *Meet the Unashamed 33-Year-Old Who Brought the Stolen Celebrity Nudes to the Masses*, WASH. POST., Sept. 5, 2014, http://www.washingtonpost.com/news/the-intersect/wp/2014/09/05/meet-the-unashamed-33-year-old-who-brought-the-stolen-celebrity-nudes-to-the-masses [http://perma.cc/PK3J-6FA2]. "'Fap' is an onomatopoeic Internet slang term for the act of masturbation." *Fap*, KNOW YOUR MEME, http://knowyourmeme.com/memes/fap [http://perma.cc/QJ6Z-G6E6].

311 *See* Rob Price, *Reddit's Privacy Rules Fail as Celebrity Nudes Spread Like Wildfire*, THE DAILY DOT, Sept. 1, 2014, http://www.dailydot.com/business/reddit-jennifer-lawrence-kate-upton-nude-photos-leak-privacy-dox-ban [http://perma.cc/4XFQ-NTFY].

312 *See* yishan [Yishan Wong], *Every Man Is Responsible For His Own Soul*, REDDIT BLOG (Sept. 6, 2014), http://www.redditblog.com/2014/09/every-man-is-responsible-for-his-own.html [http://perma.cc/9HYP-T2N7].

313 *Id.*

314 *See* alienth, *Time to Talk*, REDDIT ANNOUNCEMENTS, https://www.reddit.com/r/announcements/comments/2fpdax/time_to_talk [https://perma.cc/Z68C-Z5RD].

> The reason is because we consider ourselves not just a company running a website where one can post links and discuss them, but the government of a new type of community. The role and responsibility of a government differs from that of a private corporation, in that it exercises restraint in the usage of its powers.
>
> . . .
>
> The philosophy behind this stems from the idea that each individual is responsible for his or her moral actions.
>
> We uphold the ideal of free speech on reddit as much as possible not because we are legally bound to, but because we believe that you—the user—has the right to choose between right and wrong, good and evil, and that it is *your responsibility* to do so. When you know something is right, you should choose to do it. But as much as possible, we will not force you to do it.[315]

If Reddit is like a nation, it has all of the best and worst features of real ones. On the one hand, it is thriving and pluralistic, capable both of fostering diverse communities and binding them together in a common collective project. On the other, it turns a blind eye to terrorist training camps on its soil.[316] Filibusters regularly set forth from Reddit in search of adventure and infamy as they destabilize the rest of the Internet.

IV. Lessons for Law

The moderation techniques presented here may seem to defy easy generalization. But it is possible to highlight a few things that tie them together:

- *Moderation is complex.* The Reddit algorithm has been finely tuned over years. So has Matt Haughey's situational sense of how to respond to intemperate comments. Wikipedia's fractious norms have been the subject of multiple books. The grammar of moderation in Parts II and III is complicated because moderation itself is complicated.
- *Moderation is diverse.* Reddit, MetaFilter, Google News, and *The New York Times* solve the same problem in four radically different ways.

---

[315] yishan, *supra* note 312 (emphasis in original).

[316] *See* T.C. Sottek, *Reddit Is a Failed State*, THE VERGE, Sept. 8, 2014, http://www.theverge.com/2014/9/8/6121363/reddit-is-a-failed-state [http://perma.cc/J7QM-NANP].

> Wikipedia alone uses dozens of different moderation techniques. There is no one formula for success. Moderation contains multitudes.
>
> • *Moderation is necessary*. It may be possible for an online community to go without centralized moderation, or without *ex ante* moderation, or without human moderation, or without exclusion, or without explicit pricing, or without opening any particular drawer in the moderation tool chest. But it cannot go without moderation entirely, as the *Los Angeles Times* discovered. If a successful community appears to be unmoderated, look more closely, until its implicit technical constraints or shared norms come into focus.
>
> • *Moderation is messy*. Moderation depends on a site's technological affordances: they shape how members can communicate and interact. But a community's fate is not determined by its technology. Redditors' passion does not come from the voting algorithm; not all wikis are Wikipedia. Moderation depends just as much on social norms, and thus it is always emergent, contingent, and contestable.
>
> • *Moderation is both top-down and bottom-up*. Moderation takes place at the interface between infrastructure and interaction. Both owners and authors can influence a community's course, but neither can control it. MetaFilter would not exist without Matt Haughey's long stewardship. But if he were to say, "jump!" the community would mostly give him side-eye. A well-moderated community is like a wildflower garden. A good moderator can create the conditions for life, but not life itself.

To see how these insights play out in a regulatory context, consider the two principal legal regimes that govern online communities: § 230 of the Communications Decency Act, which provides broad immunity for interactive computer services,[317] and § 512 of the Copyright Act, which provides a copyright safe harbor.[318] Both of them seek to give communities substantial breathing room to set their own moderation policies, but they adopt very different attitudes and expectations towards moderators.

---

[317] 47 U.S.C. § 230 (2012).

[318] 17 U.S.C. § 512 (2012).

A. *Communications Decency Act § 230*

Section 230 of the Communications Decency Act (CDA) famously immunizes any "provider or user of an interactive computer service" from being treated as "the publisher or speaker" of user-generated content.[319] At the same time, it provides equally broad immunity for "any action voluntarily taken in good faith to restrict access to" objectionable material.[320] This is a double-pronged protection for moderation: it gives moderators immunity both for the content they moderate and the content they miss. The overall effect, despite the "good faith" qualifier in the second prong, is that moderators have blanket immunity: no moderation decision can lead to liability under defamation law,[321] securities law,[322] civil rights law,[323] consumer-protection law,[324] or almost-anything law. A wide variety of moderation techniques are protected, including paying authors,[325] selective deletion,[326] and extensive organization.[327]

The underlying policy is to encourage moderation by taking away the threat of liability for mismoderation.[328] A pre-CDA decision held that the pre-Web service Prodigy could be held liable for a defamatory post by a user. The court reasoned that Prodigy was "making decisions as to content" and was therefore a publisher of the defamatory material.[329] A similar service, CompuServe, had escaped liability because it exercised "no more editorial control . . . than does a public library, book store, or newsstand."[330] Taken together, the decisions created a perverse disincentive to moderate.

The CDA's solution thus embodies three assumptions about moderation. First, it views moderation as desirable—better to

---

[319] 47 U.S.C. § 230(c)(1) (2012). My discussion focuses on § 230(c)(1), which gives providers immunity when they *fail* to remove objectionable content. For a useful discussion of § 230(c)(2), which gives providers immunity when they act in good faith to remove objectionable content, see Eric Goldman, *Online User Account Termination and 47 U.S.C. § 230(c)(2)*, 2 U.C. IRVINE L. REV. 659 (2012) (describing and defending this immunity as applied to exclusion).

[320] *Id.* § 230(c)(2)(A).

[321] *E.g.*, Zeran v. America Online, 129 F.3d 327 (4th Cir. 1997).

[322] *E.g.*, Universal Commc'n Systems v. Lycos, Inc., 478 F.3d 413 (1st Cir. 2007).

[323] *E.g.*, Fair Housing Council v. Roommates.com, LLC, 521 F.3d 1157 (9th Cir. 2008) (en banc).

[324] *E.g.*, Goddard v. Google, Inc., 640 F. Supp. 2d 1193 (N.D. Cal. 2009).

[325] *E.g.*, Blumenthal v. Drudge, 992 F. Supp. 44 (D.D.C. 1998).

[326] *E.g.*, Batzel v. Smith, 333 F.3d 1018 (9th Cir. 2003).

[327] *E.g.*, Carafano v. Metrosplash.com, 339 F.3d 1119 (9th Cir. 2003).

[328] Zeran v. America Online, 129 F.3d 327, 331 (4th Cir. 1997).

[329] Stratton Oakmont, Inc. v. Prodigy Servs. Co., 1995 WL 323710, at *4 (N.Y. Sup. Ct. May 24, 1995).

[330] Cubby, Inc. v. CompuServe, Inc., 776 F. Supp. 135, 140 (S.D.N.Y. 1991).

have Prodigies and not just CompuServes. Second, it treats moderation as fallible. Even if Prodigy can find and remove some offensive posts, it is all but certain to miss others. Third, it believes that looking over moderators' shoulders is ill advised. Rather than trying to enforce a reasonable-moderator standard of conduct, § 230 says that any moderation, even do-nothing moderation, is good enough.

In light of the discussion above, the first two assumptions are eminently justified: moderation is necessary, and moderation is messy. These are close to universal laws of moderation. A regulatory scheme that does not take them into account verges on madness. But the third assumption is more contestable. On the one hand, moderation's complexity and diversity counsel regulatory caution. A judicially enforced standard of conduct risks flattening out distinctions among communities and moderators, thereby stomping on valuable experiments in self-governance.[331] On the other hand, moderators really do have some power. Some gardeners grow lilies; others grow nightshade.

Take the late and little-lamented Is Anyone Up, a website dedicated to revenge porn.[332] Like/r/jailbait and /r/TheFappening, it was a toxic community built on criminal conduct. Moderation made it better for participants and worse for the rest of the world. This is not a kind of collaboration society should encourage. Is Anyone Up's operator, Hunter Moore, pleaded guilty to federal hacking and identity theft charges for paying a co-conspirator to steal nude photos and post them to Is Anyone Up.[333] But should his liability turn on his personal involvement? The site itself was illegitimate, the privacy equivalent of a service "good for nothing else but infringement."[334] Moore moderated it with malice aforethought. There is a pragmatic question of whether it is feasible for the judicial system to distinguish the Is Anyone Ups of the world from the Prodigies, but the question is really one about how much we wish to subject moderation to judicial review.[335] It seems un-

---

[331] *See* H. Brian Holland, *In Defense of Online Intermediary Immunity: Facilitating Communities of Modified Exceptionalism*, 56 U. KAN. L. REV. 369 (2008).

[332] *See* Camille Dodero, *Hunter Moore Makes a Living Screwing You*, VILLAGE VOICE, Apr. 4, 2012, http://www.villagevoice.com/2012-04-04/news/revenge-porn-hunter-moore-is-anyone-up [http://perma.cc/QXD5-CXAL].

[333] *See* Plea Agreement for Defendant Hunter Moore, U.S. v. Moore, No. 2:13-cr-00917-DMG (Feb. 18, 2015), *available at* http://cdn.arstechnica.net/wp-content/uploads/2015/02/revenge-porn-Hunter-Moore-plea-agreement.pdf [http://perma.cc/ESP6-3QFS].

[334] Metro-Goldwyn-Mayer Studios Inc. v. Grokster, Ltd., 545 U.S. 913, 932 (2005).

[335] *Cf.* Jennifer L. Mnookin, *Virtual(ly) Law: The Emergence of Law in LambdaMOO*, 2 J. COMP.-MEDIATED COMM. (June 1996).

likely that § 230's across-the-board immunity is ideal. Even if in many cases judges are poorly situated to second-guess moderation decisions, they should not be writing blank checks to moderators like Hunter Moore, violentacrez, and johnsmcjohn.[336] Moderation to please a community's insiders is not the same as moderation to protect outsiders, and we need not treat them the same.

Another way in which § 230 currently exhibits too much deference to bad-faith moderation is illustrated by *Jones v. Dirty World*.[337] The Dirty is a website that features user-provided "Dirty Army intel, opinions, gossip, satire, and celebrities." It is edited by Nik Lamas-Richie.[338] A user submitted photographs of the plaintiff, Sarah Jones, along with a note reading, "Nik, this is Sara J, Cincinnati Bengal Cheerleader. She's been spotted around town lately with the infamous Shayne Graham. She has also slept with every other Bengal Football player."[339] Richie added his own comments: "Everyone in Cincinnati knows this kicker is a Sex Addict."[340] Another user submitted a photograph of Jones with the comment, "Her ex Nate.. cheated on her with over 50 girls in 4 yrs.. in that time he tested positive for Chlamydia Infection and Gonorrhea.. so im sure Sarah also has both.. whats worse is he brags about doing sarah in the gym.. football field.. her class room at the school she teaches at DIXIE Heights."[341] Richie posted these along with his own comment: "Why are all high school teachers freaks in the sack?nik."[342] There was more, but you get the picture.

Jones repeatedly complained to Richie, who refused to remove the posts. Under the prevailing judicial interpretation of § 230, he was clearly in the right, as the Sixth Circuit confirmed when Jones sued for defamation, false light, and intentional infliction of emotional distress.[343] All of the offending posts were written by users. It did not matter that Richie added his own comments, because his own comments were not defamatory. If moderators were forbidden to post to their own sites on pain of losing their § 230 immunity, they would lose a powerful tool for norm-setting. It did not matter that Richie

---

[336] For proposals to limit Section 230, see, for example, DANIELLE KEATS CITRON, HATE CRIMES IN CYBERSPACE 177-81 (2014); and Nancy S. Kim, *Website Design and Liability*, 52 JURIMETRICS 383 (2012).

[337] Jones v. Dirty World Entertainment Recordings, LLC, 755 F.3d 398 (6th Cir. 2014), *rev'g* 965 F. Supp. 2d 818 (E.D. Ky. 2013).

[338] THE DIRTY, http://thedirty.com [http://perma.cc/9SU9-GAFB].

[339] *Jones*, 755 F.3d at 403. At the time, Graham was a placekicker for the Bengals.

[340] *Id.*

[341] *Id.*

[342] *Id.* at 404.

[343] *Id.*

refused to remove the posts after being notified because § 230 makes no distinction based on notice.[344] If notice took away immunity, high-volume sites would be easy targets for a heckler's veto. Moderators who were unable to review notices carefully would simply remove any content that was the subject of a notice.[345]

These two rationales for blanket immunity are individually well taken, but considered together, they contradict each other. Richie's defense rests on the simultaneous claims that human moderation is desirable and that human moderation is impossible. Both cannot be true at the same time. By individually commenting on the posts about Jones, Richie provided active, human moderation. In so doing, he showed that he is not the kind of moderator for whom immunity even after notice was designed. We know that individual review of the user-submitted posts about Jones is feasible because Richie himself *had already engaged in just such a review when he posted and commented on them*.

Section 230, then, should perhaps apply differently to automated moderation and human moderation. For automated moderation, immunity even after notice can potentially be justified. If a website does not already use human moderation, goes the argument, it should not be forced to. At the scale of a YouTube or a Reddit, such a mandate could be debilitating. But where a specific post has already been the subject of human moderation, the argument for immunity after notice is weaker. The website's own actions show that it is capable of providing substantive human review. This is not an argument for going back to the world before § 230, in which Prodigy could be held liable as a publisher (liable even without notice) because it used human moderation. The point is narrower: websites like The Dirty that rely extensively on content-specific human curation could be treated as distributors (liable after notice) without undercutting the core rationales of § 230.[346] Thus, even if Richie should not have been liable for the initial postings, there is a stronger case that he should have been liable for failing to remove them.[347]

---

[344] *See* Zeran v. America Online, 129 F.3d 327, 331-34 (4th Cir. 1997) (rejecting the argument that notice defeats immunity).

[345] *Id.* at 333.

[346] For a detailed and thoughtful discussion of those rationales, see Felix T. Wu, *Collateral Censorship and the Limits of Intermediary Immunity*, 87 NOTRE DAME L. REV. 293 (2011).

[347] Assuming, that is, that the posts were defamatory—an assumption we must make when considering the threshold question of Section 230 immunity.

Not all moderation is the same. Insights like these are possible only when we delve into the details of how different communities are moderated differently.

B. *Copyright Act § 512*

The Online Copyright Infringement Liability Limitation Act, codified at § 512 of the Copyright Act, takes a rather different approach. It offers online "service provider[s]" immunity from copyright infringement for user-uploaded content.[348] But unlike § 230's blanket immunity, § 512's is shot through with exceptions. The provider must "respond[] expeditiously to remove" material that is the subject of a notice of infringement.[349] It must also have a policy for terminating the accounts of "repeat infringer[s]."[350] The safe harbor is suspended when the provider has knowledge of specific infringing activity and does nothing,[351] or when it has both a "financial benefit directly attributable to" the infringement and the "right and ability to control" it.[352] At the same time, § 512 limits moderators' duties to these specified ones. They need not "affirmatively seek[] facts indicating infringing activity."[353]

In moderation terms, § 512 specifies particular moderation strategies that a provider must use. It must use *ex post* deletion when it receives notices or knowledge of infringement, and it must use *ex post* exclusion against repeat infringers. The financial benefit test is a restriction on pricing: it rules out moderation models with prices that can be too "directly" linked to infringing material. The underlying assumption, in common with § 230, is that infringement-screening moderation is both desirable and necessarily imperfect. But § 512 more realistically recognizes that not all moderators will voluntarily adopt the law's goals—some users, and some moderators, are all for infringement.

Again, the survey of moderation shows some substantial wisdom in § 512's approach. In particular, by specifying particular moderation techniques rather than requiring perfect compliance or setting forth a vague standard for judicial elaboration, § 512 gives moderators clear and realistic orders. Takedown notices and repeat-infringer suspensions are susceptible to straightforward automated enforcement. The choice of *ex post* over *ex ante* moderation also simplifies the moderation task: required activity is triggered only by specific events. At the same time, § 512's design also pushes moderation in some

---

348 17 U.S.C. § 512(c) (2012).
349 *Id.* § 512(c)(1)(C).
350 *Id.* § 512(i)(1)(A).
351 *Id.* § 512(c)(1)(A).
352 *Id.* § 512(c)(1)(B).
353 *Id.* § 512(m)(1).

perhaps unintended directions. The red-flag knowledge test discourages moderators from looking too closely at content on their sites lest they become liable for knowing of infringement and failing to remove it—thereby discouraging hands-on moderation for other considerations, such as cultivating positive community norms.

Copyright owners have been frustrated by the rule that intermediaries have no duty to search for infringing content,[354] even though such a duty would create potentially crippling legal uncertainty.[355] Interestingly, YouTube's ContentID system now blocks uploaded videos that match an extensive list of copyrighted works.[356] YouTube has, in effect, invented around § 512's *ex ante*/*ex post* distinction, a move that was feasible because of advances in computing power and content-matching algorithms.[357] This is a good example of a moderation technique that would be hard to mandate directly. Any court looking at ContentID would be hard-pressed to explain whether its matching algorithms were too aggressive or not aggressive enough in general, let alone in any specific case.[358] But even where specific commands are unworkable, the general principle is reasonable. A good legal regime for moderation should find ways to encourage both the development of better moderation techniques and the deployment of ones that already exist.

V. Conclusion

The patterns of moderation, at once enabling and constraining, are like the basic steps of a dance. They can be combined in an infinite number of ways, and a skilled dancer can always find new and surprising variations, but the audience member who knows the steps can recognize how the dance brings them together. This Article, then, is an initial dance lesson for legal scholars. The grammar of moderation provides a convenient way to reason about online communities: it directs attention to significant features and makes tentative predictions about

---

[354] *See, e.g.*, UMG Recordings, Inc. v. Shelter Capital Partners, 667 F.3d 1022, 1041-43 (9th Cir. 2011), *withdrawn*, 718 F.3d 1006 (9th Cir. 2011).

[355] *See* John Blevins, *Uncertainty As Enforcement Mechanism: The New Expansion of Secondary Copyright Liability to Internet Platforms*, 34 CARDOZO L. REV. 1821 (2013).

[356] *See generally How Content ID Works*, YOUTUBE.COM, https://support.google.com/youtube/answer/2797370 [https://perma.cc/2M4M-UKGU] (describing ContentID's automated *ex ante* filtering)

[357] *But see* Rebecca Tushnet, *All of This Has Happened Before and All of This Will Happen Again: Innovation in Copyright Licensing*, 29 BERK. TECH. L.J. 1447, 1457-67 (criticizing ContentID).

[358] *See* Sonia K. Katyal & Jason M. Schultz, *The Unending Search for the Optimal Infringement Filter*, 112 COLUM. L. REV. SIDEBAR 83, 83-84 (2012) (critizing proposal to "incentiviz[e] webhosts to screen materials prior to posting them" by "offering immunity only for webhosts that employ the best available method for filtering content prior to publication").

what certain forms of moderation are likely to do. The theory of moderation should be applicable to such matters as network neutrality, regulation of social software, intermediary liability, online privacy, and media policy. It may also be useful in describing the institutional options open to communities dealing with the management of offline resources.

We should not expect so subtle a term as "moderation" to have only one meaning. This Article has dealt with "moderation" as practiced by *moderators*, those who bring order to a discussion. But "moderation" is also a matter of being *moderate*, "avoidance of excess or extremes in behaviour."[359] Online communities are caught between freedom and control, openness and closure, abundance and scarcity. The theory of moderation presented in this Article emphasizes that none of these oppositions is ever absolute. No community is ever perfectly open or perfectly closed; moderation always takes place somewhere in between.

[359] *Moderation*, OXFORD ENGLISH DICTIONARY (3d ed. 2002).