JOURNAL OF LAW, TECHNOLOGY & THE INTERNET • VOLUME 10 • ISSUE 1 • 2019

– ESSAY –

CONTENT MODERATION IN AN AGE OF EXTREMES

Rebecca Tushnet[1]

CONTENTS

INTRODUCTION

Every time I sat down to revise this talk for publication, there was a new story of some mis- or malfeasance by a major social media platform: Twitter declined to ban a man who was subsequently charged with sending bombs to the kind of people he harassed online;[2] Facebook repeatedly allowed discriminatory

[1] Rebecca Tushnet was the keynote speaker at the 2018 Journal of Law, Technology & the Internet Symposium. She received her A.B. from Harvard College and J.D. from Yale Law School. She previously clerked for Chief Judge Edward R. Becker on the U.S. Court of Appeals for the Third Circuit and for Associate Justice David H. Souter on the U.S. Supreme Court. She is presently a professor of Law at Harvard Law School.

[2] *E.g.*, Andrew Liptak, *Twitter Says It 'Made a Mistake' For Not Removing Threatening Tweets from Florida Bomb Suspect*, VERGE (Oct. 27, 2018, 10:32 AM) https://www.theverge.com/2018/10/27/18031888/twitter-alleged-florida-bomber-threats-rochelle-ritchie [https://perma.cc/8FCV-TYG9].

advertising;[3] and YouTube's algorithm encouraged radicalization in order to keep people on the site.[4] As James Grimmelmann depressingly concludes in his article on the same topic, everything is broken because platforms are broken.[5]

I would put it a bit differently: everything is broken because people are broken. But that doesn't mean we need to give up; rather, it means that we can't. There is no alternative to being broken. There is only trying to make things better, imperfectly and unevenly. This process is inevitably ugly. As Tarleton Gillespie has written, content moderation is so difficult that "all things considered, it's amazing that it works at all, and as well as it does…. Given the sheer enormity of the undertaking, most platforms' definition of success includes failing users on a regular basis."[6]

---

[3] *See* Julia Angwin, Madeleine Varner and Ariana Tobin, *Facebook Enabled Advertisers to Reach 'Jew Haters'*, PROPUBLICA (Sept. 14, 2017, 4:00 PM), https://www.propublica.org/article/facebook-enabled-advertisers-to-reach-jew-haters [https://perma.cc/3M26-WV64] (enabling marketers to direct their advertisements to profiles that expressed anti-Semitic interests); Charles V. Bagli, *Facebook Vowed to End Discriminatory Housing Ads. Suit Says It Didn't*, N.Y. TIMES (Mar. 27, 2018), https://www.nytimes.com/2018/03/27/nyregion/facebook-housing-ads-discrimination-lawsuit.html [https://perma.cc/U9TH-7HFU] (alleging the website allowed housing discrimination); Jacob Kastrenakes, *ACLU Says Facebook Allowed Discriminatory Job Ads That Didn't Appear to Women*, VERGE (Sept. 18, 2018, 9:19 AM), https://www.theverge.com/2018/9/18/17873448/facebook-job-ads-discrimination-women-aclu [https://perma.cc/NYK2-8PTQ] (allegedly allowed employers to post discriminatory job advertisements); Joe Miller, *Facebook Sorry for 'White Supremacist Ad'*, BBC NEWS (Nov. 3, 2018), https://www.bbc.com/news/technology-46083026 [https://perma.cc/7783-Y9HH] (allowing an advertisement that spread a conspiracy theory advocated for by white supremacists political ads).

[4] *See* Zeynep Tufekci, *YouTube, the Great Radicalizer*, N.Y. TIMES: OPINION (Mar. 10, 2018), https://www.nytimes.com/2018/03/10/opinion/sunday/youtube-politics-radical.html. [https://perma.cc/66Z6-RNAF] (arguing YouTube has programmed its algorithms to suggest more extreme content to improve the platform's profits); *see also* David Streitfeld, *'The Internet Is Broken': @ev Is Trying to Salvage It*, N.Y. TIMES (May 20, 2017) https://www.nytimes.com/2017/05/20/technology/evan-williams-medium-twitter-internet.html [https://perma.cc/AP6M-WF3K] ("The trouble with the internet, Mr. Williams says, is that it rewards extremes. Say you're driving down the road and see a car crash. Of course you look. Everyone looks. The internet interprets behavior like this to mean everyone is asking for car crashes, so it tries to supply them.").

[5] James Grimmelmann, *The Platform Is the Message*, 2 GEO. L. TECH. REV. 217, 233 (2018).

[6] TARLETON GILLESPIE, CUSTODIANS OF THE INTERNET: PLATFORMS, CONTENT MODERATION, AND THE HIDDEN DECISIONS THAT SHAPE SOCIAL MEDIA, 197 (2018); *see also id.* at 77 ("[T]here are no hypothetical situations, and there are no cases that are different or really edgy . . . There's just more and less frequent cases, all of which happen all of the time.").

Where everything, including bad behavior, can be found in endless proliferation, what if anything can be done? My experience leads me to suggest a few lessons. I helped found a nonprofit, the Organization for Transformative Works ("OTW"), which focuses on protecting and preserving noncommercial transformative works, specifically fanfiction and other fanworks—that is, new stories and art based on existing, often in-copyright, works such as the Harry Potter series and Marvel's Avengers movies. Fandom-specific nonprofit archives have existed for a long time and some are maintained by hardy individuals, but the OTW wanted to create a multi-fandom archive and to protect its continued existence with a legal structure that would be separate from the individual situation of any given archive maintainer.

Our all-volunteer team—the largest majority-female open source coding project on the web, as far as we are aware—created software that allows fan creators to post and tag their works so that they can reach the right audiences. Our terms of service are not influenced by advertisers and our mission is to host all legal, noncommercial fanworks, even if that content is controversial, offensive, or badly spelled. The Archive of Our Own ("AO3"), our main website, hosts over four million works, supports over a million users, and garners over 115 million page views per week,[7] making the AO3 the second most popular writing-focused website in existence.[8]

It turns out, however, that while most fans are lovely people, some of them do really bad things, because fans are, first and foremost, *people*. Our success therefore hasn't come without costs; the AO3's all-volunteer content and abuse team faces some serious challenges. Not only do people sometimes abuse our platform by harassing or spamming other users, but they also sometimes abuse our content moderation process as a means of harassment.

If you had asked me ten years ago, I would have been skeptical that a person would pretend to be another person's parent, a police officer, or a lawyer representing a copyright claimant in order to get another person's account closed. I no longer need to believe in these things—I've seen them. Indeed, according to Eugene Volokh, who's extensively researched these falsehoods, fake court orders

[7] Archive of Our Own, Fanfore.com, https://fanlore.org/wiki/Archive_of_Our_Own (last visited Feb. 17, 2019).

[8] Archive of Our Own, Similarweb.com, https://www.similarweb.com/website/archiveofourown.org (last visited Feb. 17, 2019).

to get content removed are popping up in other contexts, such as when someone wants to silence a blogger on Medium or remove negative content from search results so it will be unfindable.[9] We aren't alone in experiencing our own, intermediary version of "fake news."

While the OTW believes in fair use, we have also encountered situations where our users' freely available works have been copied wholesale and sold without authorization on platforms like Amazon, violating our users' copyright rights.[10] We have also encountered cases where we have to act against confusing uses of our own trademarks. For example, a typosquatter used our website's domain name, plus one letter, and mimicked our interface, including asking people for their login information—which looked like classic phishing for login credentials.[11] We've experienced a rightsholder's frustration with intermediaries that don't seem to care until you spend a lot of money on a formal process, whether that's the Uniform Domain-Name Dispute-Resolution Policy ("UDRP") for domain names or a potential lawsuit. We've had problems with apps on the Google and Apple app stores using our name and logos, at which point we receive complaints from frustrated users who paid for performance they aren't getting and blame us—whereas our site is and will remain free and ad-free.

So, as the song says, I've seen the world from both sides now. We've sent cease and desist letters and received them. The legal process, in general, has worked very well for us, in substantial part because we are fortunate to have a highly qualified group of volunteer lawyers. Most sites, even commercial sites of our size, won't have four or five experienced lawyers on call at any hour of the day. Nonetheless, despite our abundance of volunteer talent, we can't hire ten thousand people to moderate content. With a limited annual budget, most of which goes to keeping the website running, we can't even hire ten people. We rely on a small, hard-working, long-suffering team of volunteers to investigate reports of abuse, and they are doing their best, but they will never be able to catch everything.

---

[9] *See* Revised Amicus Curiae Brief of Eugene Volokh in Support of Appellant, Hassell v. Bird, 274 Cal. App. 4th 1336 (2016), https://digitalcommons.law.scu.edu/cgi/viewcontent.cgi?article=2484&context=historical.

[10] To preserve confidentiality, I will only discuss these types of incidents broadly, rather than identifying specific affected people.

[11] Organization for Transformative Works v. Binkley, Case No. D2018-0132, http://www.wipo.int/amc/en/domains/search/text.jsp?case=D2018-0132.

This is the personal background with which I come to what's commonly understood as the most serious problem in the online ecosystem and, perhaps, in democracy as a whole: the presence of bad content. Copyright owners decry massive amounts of copyright infringement. Then there is the harassment and revenge porn mostly used to humiliate and blackmail women and massive amounts of fake news. What is to be done?

While some people argue for increased government intervention in the form of legal duties to remove various types of unwanted content, others maintain that the best solution is to reconstruct some sort of democratic process within a service's "polity" itself, a procedural solution to knotty problems of substance. I want to complicate the debate by discussing the multiple types of actors in the intermediary space; some entities, like the OTW, don't resemble the profit-seeking model at which most regulatory and governance proposals are directed. Other online entities, such as those that participate in the domain name system, have very different functions and abilities than the websites and apps most people think of as "the internet." If we don't keep these variances in mind, we are unlikely to get the results we seek. It's very hard to generalize beyond those cautions because things are changing so fast in terms of both content moderation policies and government action (such as the recent preliminary approval of a copyright filtering requirement for intermediaries in Europe,[12] and the even more recent "embedding" of French officials into Facebook to see how it regulates hate speech[13]).

I. THE DIGITAL MILLENNIUM LASTED TEN YEARS: INTELLECTUAL PROPERTY

I will start with intellectual property issues, because that's where I started my career and my work at the OTW. One thing that non-specialists are rarely aware of is the extensive regulation of domain names, done by the Internet Consortium for Assigned Names and Numbers ("ICANN") and the providers to which it has delegated authority to administer various top-level domains, like .com, .uk, or

[12] James Vincent & Russell Brandom, *Everything You Need To Know About Europe's New Copyright Directive*, VERGE, (Sept. 13, 2018, 2:14 PM), https://www.theverge.com/2018/9/13/17854158/eu-copyright-directive-article-13-11-internet-censorship-google.

[13] Tony Romm & James McAuley, *Facebook Will Let French Regulators Study Its Efforts to Fight Hate Speech*, WASH. POST, (Nov. 12, 2018), https://www.washingtonpost.com/technology/2018/11/12/facebook-will-let-french-regulators-study-its-efforts-fight-hate-speech/?utm_term=.b2e3f9953d54.

newer entrants like .bike and .london. In 1999, trademark owners succeeded in procuring the UDRP, which is mandatory arbitration for trademark claims based on allegedly infringing domain names. Like other arbitration proceedings, the UDRP is generally cheaper and faster than litigation.

Trademark owners secured this extra protection, which no other category of objections to individual domain names had, by virtue of their persistence (not to mention some decent arguments) in the multi-stakeholder, consensus-based process through which ICANN makes decisions. ICANN is not a government, and it's not a democracy, but what it *is* can be hard to define, which is why its role in overseeing intellectual property issues in the domain name ecosystem is apparently still up for grabs.

More recently, ICANN introduced a large number of new global top level domains—.london and .bike and hundreds of others. In order for trademark owners to accede to this expansion, which would also expand the potential number of domain names that could use their marks, trademark owners secured the Uniform Rapid Suspension ("URS") process—which is supposed to be faster and cheaper than the UDRP, for no-brainer cases of infringing use. Complainants and respondents get fewer words and less time for their submissions than they would for a UDRP claim.

The URS has now been in place for several years. Roughly 900 claims have been adjudicated, but it's still hard to say whether the game is worth the candle. Most of the domain name registrants defaulted; those cases don't seem to be decided noticeably faster than UDRP default cases. Though trademark owners were opposed to looking at individual cases to see if they truly were no-brainers, I asked my research assistants to code the cases to look at some of the more obvious questions. It seems that the URS is used by a relatively small number of trademark owners, even compared to the universe of trademark owners who have resort to the UDRP. Most cases do seem to involve simple infringing or at least non-bona fide uses, but a significant minority don't provide enough information to figure out what the basis for the arbitrator's decision was—which means that no one else can be sure what the rationale was, or evaluate whether it made sense.

Trademark owners' representatives are currently proposing to shorten response times and eliminate the ability for a domain registrant to cure a default, as well as to make the URS a "consensus policy," exposing millions of domain names registered in "legacy" domains such as .com to potential challenge under the URS, with its lesser procedural protections compared to those of the UDRP. Likewise,

they are proposing to expand something known as the "claims" system to allow trademark owners broader rights to block or threaten potential registrants—even though the current system has been extensively used to make claims on common words like "cloud" and "hotel."

The broader point is that rightsholders don't stop when they lose a public battle. Big trademark owners, like big copyright owners, can afford to be persistent, and they often believe they can't afford *not* to be. As a result, when they don't get remedies in national law, they often try to get more through international treaties, or in this case, through private policymaking at the chokepoints of internet connectivity. And, because the entities administering the system—domain name registries and registrars—don't have much skin in the game for any given domain name, they may not fully take into account individual domain name registrants' interests in their practices. For example, some registries and registrars are now selling extra services to trademark owners to allow them to block any registrations using their names, regardless of whether the use would be barred by the URS, UDRP, or trademark law—so Apple could block apple.farm for an apple farming operation, if .farm is operated by one of these registries.[14] The carefully negotiated balance between registrant and free speech interests versus trademark interests that was supposed to result from the ICANN policies has been replaced by these private mechanisms, just as Content ID has replaced copyright and fair use rules on YouTube.

This is also the future of content moderation, if we don't make content moderation a public policy issue. And it's not just trademark owners who want to use the domain name system to shut down allegedly bad actors—copyright owners are trying to get in on the game. They are arguing that registries should shut down domain names where the underlying domain is used to infringe copyrights. Annemarie Bridy has done important work documenting what people are willing to admit on the record about these new policies.[15] Copyright owners claim that they

[14] *See* Donuts, Brand Protection, https://donuts.domains/what-we-do/brand-protection (firm that controls a large number of top-level domains explaining the private control it sells to trademark owners) (last visited Feb. 27, 2019).

[15] *See generally* Annemarie Bridy, *Notice and Takedown in the Domain Name System: ICANN's Ambivalent Drift into Online Content Regulation*, 74 WASH. & LEE L. REV. 1345 (2017).

are only going after sites devoted to infringement, but without public guidelines and other transparency mechanisms we just have to take their word for it.[16]

Some copyright owners argue that United States law isn't good enough to protect their interests.[17] After the failure of a proposed law that would have fundamentally changed platform liability for infringing content due to a massive public outcry, copyright owners are touting their policy preferences in subtler ways, both by the aforementioned private agreements with online platforms and through arguing that the current notice and takedown system for copyright infringement online is broken. As in Europe, certain copyright owners argue for mandatory filtering of all content, brilliantly renamed "notice and staydown" to sound only slightly different from the familiar notice and takedown.[18]

Proponents of filtering argue that filtering isn't much of a challenge for websites, and that the costs to speakers would be minimal. They take YouTube as their model: YouTube implements filters, so why not the internet as a whole? A reality check can be found in Jennifer M. Urban, Joe Karaganis, and Brianna L. Schofield's important empirical work on the functioning of the copyright notice and takedown system in the U.S., which includes in-depth interviews with multiple intermediaries, including websites that host millions of pieces of user-generated content.[19]

Although the OTW wasn't part of their research, the results corresponded to our experience. What they found was that there are distinct models of copyright takedowns online. The vast majority of services, like the OTW, are what they call "DMCA[20] Classic": we receive relatively few notices of claimed infringement, few of them are automated, and we subject them to individual review for validity. We

[16] *Id*. at 1347.
[17] Numerous examples can be found in the comments submitted in response to the Copyright Office's inquiry on the functioning of the notice and takedown system, as well as in statements made at various roundtables the Office held on the subject. See Section 512 Study, https://www.copyright.gov/policy/section512/ (collecting comments and transcripts) (last visited May 1, 2019).
[18] At the most recent Copyright Office roundtable, numerous industry participants applauded European developments and argued that the U.S. should follow Europe's lead. See Transcript, Library of Congress, United States Copyright Office, Section 512 Study Roundtable, Monday, Apr. 8, 2019 (on file with journal).
[19] *See generally* Jennifer M. Urban, Joe Karaganis, & Brianna Schofield, *Notice and Takedown: Online Service Provider and Rightsholder Accounts of Everyday Practice*, 64 J. COPYRIGHT SOC'Y OF THE U.S.A 371 (2017).
[20] Digital Millennium Copyright Act (DMCA), Pub. L. No. 105-304, 112 Stat. 2860 (1998).

reject notices that ignore free speech and fair use or are trying to use the DMCA process to achieve non-copyright aims, such as trying to protect a claimed trademark in a title, and remove content that seems to be correctly identified as infringement.[21] In essence, our process is artisanal rather than mass market.[22] But we couldn't continue to do that if we started to receive a flood of automated notices.

Other models are adopted by services that *do* receive a large volume of automated notices, and they have responded by automating in return and sometimes also by going further than the DMCA requires. This "DMCA Plus" model includes some types of filtering and sometimes includes cutting deals with content owners to monetize user uploads of content claimed by someone else, as with YouTube's Content ID.[23]

There are plenty of problems with the Content ID model, but its deficiencies are not my focus here. A key objection to a filtering requirement for intermediaries that host user-generated content in general, as the European Community has decided to impose, is that it is fundamentally mistargeted. It takes a solution that has benefits for a few big copyright owners and big internet services and demands its imposition on other intermediaries—most of which don't have a big infringement problem in the first place and many of which couldn't continue to operate if they had to bear the costs of developing and constantly updating a filtering system. Ironically, because Europe is hostile to Facebook and YouTube, it has adopted a solution that ensures that Facebook and YouTube will continue to dominate, since they are the ones most likely to survive filtering and licensing requirements.[24]

It's not even a matter of size—some very big sites, like ours, Wikipedia, and Medium, are DMCA Classic sites receiving relatively few copyright claims

[21] Urban et al., *supra* note 19, at 385.
[22] *Cf.* Robyn Caplan, Content or Context Moderation? Artisanal, Community-Reliant, and Industrial Approaches, https://datasociety.net/wp-content/uploads/2018/11/DS_Content_or_Context_Moderation.pdf, at 17 (noting that "Alex Feerst, head of legal for the blogging platform Medium, referred to their approach as 'artisanal,' or (being tongue-in-cheek) as 'small-batch,' to note that despite their more than 80 million users, their moderation approach is still done manually, 'by human beings.'").
[23] Urban et al., supra note 19, at 382-383.
[24] The newly adopted rules do exclude sites run by nonprofits, such as Wikipedia and the OTW's archive, and they purport to allow small startups a few years to become compliant with filtering and licensing requirements, but given how quickly websites can expand the minimal exceptions for startups are unlikely to help. See Rachel Wolbers, Transcript, supra note 18, at 329-30.

and a relatively large proportion of invalid claims out of those few. Even Amazon's large Kindle Direct program, which allows self-publishing, turns out to be better suited to DMCA Classic treatment. At hearings on the continued viability of the DMCA's notice and takedown system, Amazon disclosed that half of the takedown requests it receives for Kindle Direct are from competitors trying to suppress another person's book. These are not valid claims.[25] "Notice and staydown" would worsen this problem rather than ameliorate it, harming legitimate creators trying to reach audiences.

This isn't to say that DMCA Classic sites are the only ones who experience problems. As Urban and her coauthors explained, "[n]early every OSP recounted stories of deliberate gaming of the DMCA takedown process, including to harass competitors, to resolve personal disputes, to silence critics, or to threaten the OSP or damage its relationship with its users. And although the proportion of problematic requests varied by type of OSP, every OSP also told stories of takedowns that ignored fair use defenses or that targeted non-infringing material."[26] Attempts to impose more obligations are likely to fail to decrease pure, substitutionary copying, because users dedicated to piracy excel at finding ways to route around obstacles. By contrast, ordinary good-faith users are more likely to get accidentally caught in a trap. Given this dynamic, where automated filtering and blocking deters the wrong people, each failed measure is likely to lead to demands for ever more restrictive actions, suppressing more and more legitimate speech in the process—if only by driving websites with fewer resources out of business.[27]

---

[25] Stephen Worth, Assoc. Gen. Couns. of Amazon.com, Inc., Testimony at the U.S. Copyright Off. Section 512 Study, Public Roundtable (May 13, 2016) (transcript available in the U.S. Copyright Office website). ("[W]ith Kindle Direct publishing, authors routinely try to climb to the top spot in their category . . . by issuing bogus notices against higher ranking titles. And this for us actually accounts for more than half of the takedown notices that we receive.").

[26] Urban et al., *supra* note 16, at 389.

[27] *Id.* at 397 (one "respondent described his company's history of experimentation with 'reasonable and effective' measures beyond notice and takedown, but viewed most of the strategies proposed by rightsholders as ineffective, and therefore certain to be followed by more demands if adopted."); Id. at 399 ("In some striking cases, it appears that the vulnerability of smaller OSPs to the costs of implementing large-scale notice and takedown systems and adopting expensive DMCA Plus practices can police market entry, success, and competition. Those without sufficient resources to build or license automated systems described being in precarious positions, at risk of being priced out of the market by better-resourced competition if floods of notices or DMCA Plus requirements were to arrive."); Id. at 400 (noting estimated cost of filtering: $60 million to develop Google's Content ID, and $10,000–12,000 per month for Audible Magic for

Ultimately, the DMCA's notice and takedown system, like democracy, might be the worst possible system except for all the others that have been tried. While market pressures might drive large sites like YouTube to more filter-based systems, it is important not to treat YouTube as a model for the internet at large—unless all we want from the internet is YouTube.

II. STRANGE APPS HANDING OUT BADGES IS NO BASIS FOR A SYSTEM OF GOVERNMENT

A. The Need for Eyeballs and the Impulse to Regulate

Outside of intellectual property, content moderation faces even more varied and unpredictable instances of wrongdoing, including false news and harassment. Unlike a situation of claimed copyright infringement, there's no underlying work with which to compare the allegedly violative content, so some other metric for removing a user's post has to be devised. But what might that be? As James Grimmelmann has explained, there's no answer to this question in the abstract. A post that decries eating Tide Pods and one that encourages eating Tide Pods, for example, can be indistinguishable from the outside, given the ways in which the social meaning of an individual piece of communication is constructed: "The difficulty of distinguishing between a practice, a parody of the practice, and a commentary on the practice is bad news for any legal doctrines that try to distinguish among them, and for any moderation guidelines or ethical principles that try to draw similar distinctions."28 Even worse, this can equally be true of arguably racist content, especially when moderators are from other countries and don't have the full background required to distinguish someone who is reporting a racial slur they experienced from someone who is using that racial slur to cause more harm.

These intractable problems are generated and worsened by commercial interests. Because major social media platforms are ad-driven, they need to keep their users on their sites as long as possible. Therefore, they prioritize showing users content that will keep them watching or reading and they are indifferent to whether

---

one medium-sized service and $25,000 per month for another service, plus additional, ongoing implementation, negotiation, and review costs).

[28] Grimmelmann, *supra* note 4, at 221-22.

that's doing overall social harm to society, or to users themselves.[29] Their responses to "bad speech" depend on their profit motives, which both drive some speech protections (when most users like the speech) and provide a reason (or an excuse) for other speech suppression. For example, pressure from advertisers led YouTube to crack down on "pro-terrorism" speech on the platform, while addressing white supremacist speech has proved more challenging, in significant part because of the right-wing elected officials who might be caught up in any white supremacist purge.[30] Klonick argues that "platforms are economically responsive to the expectations and norms of their users," which leads them "to both take down content their users don't want to see and keep up as much content as possible," including by pushing back against government takedown requests.[31] But this formulation equivocates about who the relevant "users" are—participants on the platform or advertisers. Content that advertisers or large copyright owners don't want to see is far more vulnerable than content that individual participants don't want to see.[32]

Here again it's worth comparing the AO3 to commercial platforms. The AO3 is a volunteer, voluntarily-funded space whose commitments are directed towards satisfying users' preferences without the need to generate ad revenue.[33] Although it lacks Facebook's massive engineering staff, the AO3 also doesn't seem to need complex algorithms to present users with the content that the AO3 "thinks"

---

[29] *Id.* at 227.

[30] Joseph Cox & Jason Koebler, *Why Won't Twitter Treat White Supremacy Like ISIS? Because It Would Mean Banning Some Republican Politicians Too*, MOTHERBOARD, Apr. 25, 2019, https://motherboard.vice.com/en_us/article/a3xgq5/why-wont-twitter-treat-white-supremacy-like-isis-because-it-would-mean-banning-some-republican-politicians-too.

[31] Kate Klonick, *The New Governors: The People, Rules, and Processes Governing Online Speech*, 131 HARV. L. REV. 1598, 1664 (2018).

[32] On YouTube, for example, shifting policies (or lack of policies) on what content can be monetized have repeatedly angered the platform's content creators, because YouTube is trying to placate advertisers. *See* Julia Alexander, *The Yellow $: A Comprehensive History of Demonetization and Youtube's War With Creators*, POLYGON, May 10, 2018, https://www.polygon.com/2018/5/10/17268102/youtube-demonetization-pewdiepie-logan-paul-casey-neistat-philip-defranco ("Ariel Bardin, YouTube's vice president of product management, addressed the changes that Schindler spoke about in her own creator's blog post. 'There's a difference between the free expression that lives on YouTube and the content that brands have told us they want to advertise against,' Bardin said ….").

[33] *See* Elizabeth Minkel, *The Online Free Speech Debate is Raging in Fan Fiction, Too*, The Verge(Nov. 8, 2018), https://www.theverge.com/2018/11/8/18072622/fanfic-ao3-free-speech-censorship-fandom. (showing that these rules are hotly debated within fandom, because what is welcoming to some users can be exclusionary to others).

those users want to see. Instead, it has a user-centered set of tools for displaying works of interest. Users can subscribe to the specific tags creators use to describe their works, or they can assemble their own search terms. Search results are shown in order of newest to oldest by default, though a user can tweak how search results are displayed in various ways—sorting by length, number of views, or kudos (the AO3 equivalent of "likes"), and so on. Users complain about the difficulty of finding what they want—users always complain about the difficulty of finding what they want—but they don't need to worry about the AO3 manipulating what they see to make them happier or sadder, or to make them more or less likely to vote, or to make them more (never less) likely to buy advertisers' products.

B. Facebook as Government, Facebook as Fiduciary?

What might we do about those platforms that do try to shape users' experiences to keep them on-site as long as possible, in ways that seem to cause harmful externalities? As with so many issues of free expression in the modern world, Jack Balkin got there first.34 Balkin argues that many of the dominant online players, like Facebook and Google, should be treated as "information fiduciaries" toward their end-users.[35] Under the fiduciary model, they would have duties to act in good faith and to avoid manipulation of those users.[36] Concepts of due process, transparency in decision-making, and equal treatment are also consistent with the fiduciary model, as well as with an idea of the social media platform as a little republic: a system that is in need of governance and thus of government.[37]

---

[34] Jack M. Balkin, *Free Speech in the Algorithmic Society: Big Data, Private Governance, and New School Speech Regulation*, 51 U.C.D. L. REV.. 1149 (2017-2018). *See also* Jack M. Balkin, *Fixing Social Media's Grand Bargain*, HOOVER WORKING GROUP ON NAT'L SECURITY, TECH., AND L. (HOOVER INSTITUTION, Aegis Series Paper No. 1814), Oct. 16, 2018 at 11-15.

[35] Jack M. Balkin, *The First Amendment in the Second Gilded Age*, BUFF. L. REV. (forthcoming 2019) (manuscript at 28-29) (https://ssrn.com/abstract=3253939) [https://perma.cc/H8QV-CRR6] (explaining how the relationship between social media companies and their users meet the four standards of a fiduciary relationship—"(1) the company provides special services based on special expertise; (2) there is a great asymmetry in knowledge between the company and its clients; (3) clients are especially vulnerable to the company because of the company's knowledge about them; and (4) the need for clients to trust the company to receive the benefit of the service.").

[36] Id. at 29.

[37] *See* Balkin, *supra* note 24 at 1162, 1196-97.

What we have in the online ecospace is a version of ontogeny recapitulating phylogeny. As online services become more like places where we live most of our lives, we want them to behave more like governments, or branches of governments. For example, Facebook's policy team has a biweekly content meeting, in which different teams across the company—engineering, legal, content reviewers, and external partners like nonprofit groups—provide recommendations to the content moderation team for inclusion in the policy guidebook.[38] Tellingly, the team leader called this meeting a "mini legislative session."[39] Facebook has also adopted and adapted concepts familiar from First Amendment doctrine to identify when objectionable content should be removed, specifically newsworthiness and the concept of a public figure.[40]

Indeed, Kate Klonick's study of major platforms finds that Facebook is not unique: everyone's policies have "marked similarities to legal or governance systems. This includes the creation of a detailed list of rules, trained human decision-making to apply those rules, and reliance on a system of external influence to update and amend those rules."[41] The rules are usually similar for users in and outside the U.S. Regardless of what the underlying legal structure is or whether an institution is essentially inventing a structure from scratch, it turns out that speech regulations pose standard issues of definition (defamation and hate speech are endlessly flexible), enforcement (who will catch the violators?), and equity/fairness (who will watch the watchmen?).

These details of implementation are far more than trivial. Among other things, as we have seen here and in other countries, governments quickly learn how

[38] Alexis C. Madrigal, *Inside Facebook's Fast-Growing Content-Moderation Effort*, THE ATLANTIC, (Feb. 7, 2018) https://www.theatlantic.com/technology/archive/2018/02/what-facebook-told-insiders-about-how-it-moderates-posts/552632.

39 *Id*. S*ee also* Heather Whitney, *Emerging Threats: Search Engines, Social Media, and the Editorial Analogy*, KNIGHT FIRST AMEND. INST. 27-28 (Feb. 2018), https://knightcolumbia.org/content/search-engines-social-media-and-editorial-analogy ("The government-like character of the leading tech companies has been acknowledged by the companies themselves. Almost a decade ago, Zuckerberg opined, "In a lot of ways Facebook is more like a government than a traditional company. We have this large community of people, and more than other technology companies we're really setting policies.").

[40] Kate Klonick, *Facebook v. Sullivan*, KNIGHT FIRST AMEND. INST. (2018), https://knightcolumbia.org/sites/default/files/content/Kate_Klonick_Emerging_Threats.pdf [https://perma.cc/3XNJ-PSL3].

[41] Klonick, *supra* note 22, at 1602.

to use, and misuse, platform mechanisms for their own benefit.[42] Reacting to government manipulation of an abuse team by coordinated reporting of dissidents for policy violations can be difficult—and platforms may even decide not to resist that manipulation. Some of these abusive techniques, moreover, resist handling by an abuse team even when identified. When government-backed teams overwhelm social media with trivialities in order to distract from a potentially important political event, as is apparently common in China, what policies or algorithms could identify the pattern, much less sort the wheat from the chaff?

Fiduciary treatment of the kind Balkin has advocated for could require some sort of attempt to fix these problems – to protect users against epistemic fraud that destabilizes the foundations of knowledge and social trust.[43] However, paradoxically or ironically, proposals for regulation tend to lock in the idea that large online spaces where people engage with one another will be privatized and profit-seeking, and thus will have the cash on hand to hire the moderators and build the automated tools that the regulators think are likely to diminish whatever harms they're concerned about. But this presumption has significant negative consequences not just for the startups who can't necessarily comply with complex regulatory schemes, but also for non-governmental, nonprofit actors who don't want to build something advertising-driven.

For example, Balkin's proposal—in part to deal with First Amendment objections that would arise if the government tried to impose new fiduciary content moderation duties on platforms—suggests incentivizing platforms to opt in to fiduciary obligations by conditioning the platform's immunity for harmful content posted by users on its assumption of such obligations.[44] However, this solution requires changing the baseline in the US to remove that immunity, which presently protects the AO3, Wikipedia, Medium, YouTube, Facebook, Twitter, and hundreds

[42] ZEYNEP TUFCEKI, TWITTER AND TEAR GAS: THE POWER AND FRAGILITY OF NETWORKED PROTEST 230-31 (2017) (surveying the ways in which governments have used online platforms for their own ends, including generating public hatred and uncertainty about truth).

[43] *But see* Lina Khan & David Pozen, *A Skeptical View of Information Fiduciaries*, 133 Harvard Law Review (forthcoming 2019), https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3341661, draft at 13 (noting that "it is unclear how a digital fiduciary is supposed to fulfill its duty of loyalty to users under conditions of 'perpetual' conflict" between the interests of advertisers and the interests of ordinary platorm users).

[44] *See also* GILLESPIE, *supra* note 5, at 43-44 (suggesting that legal immunities could be conditioned on having obligations such as meeting minimum standards for effective content moderation, some degree of transparency or specific structures for due process and appeal of decisions).

of other smaller platforms—many of which will lack the resources to look like a fiduciary if pervasive content moderation is the desired model. That change could expose many smaller platforms to potentially devastating liability and encourage sites to take down any challenged content no matter its truth value. Maybe we could at least exempt the AO3 and Wikipedia by being more generous to nonprofits or to sites with non-Facebook-sized userbases, but the problem suggests the complexity of the issues.

Another question that deserves more attention is this: if what we are talking about are truly governance structures, why should citizens of a democracy accept anything other than actual democracy, either representative or otherwise, in the regulation of these spaces? Shouldn't you have to win an election to be "mayor" of a place on Foursquare? Right now, this criticism is mainly raised in the context of the fact that the employees and corporate officers of the major platforms "aren't nearly as diverse as their user bases."[45] A more democratic form of governance would seem to necessitate product design involving actual representatives of the user base, not just the advertisers.[46] Anti-gerrymandering activists say that voters should pick their politicians, rather than politicians picking their voters; perhaps we need a similar design principle for dominant platforms. Any such solution would have to involve representation, not direct democracy. When Facebook said it would allow users to vote on its policies, the experiment was a massive failure—because it wanted them to vote on policies they had no reason to read, and more generally because it tried to use direct voting on a population of a size requiring representative government.[47]

Given the lack of transparency and democratic input into the rules, Facebook's content moderation systems—and those of many other platforms—resemble, as David Pozen argues, "a system of authoritarian or absolutist constitutionalism. Authoritarian constitutionalism … accepts many governance features of constitutional democracy 'with the noteworthy exception of …

[45] Tonya Riley, *Who's Afraid of Online Speech? A Future Tense Event Recap*, SLATE (Feb. 08, 2018), https://slate.com/technology/2018/02/whos-afraid-of-online-speech-a-future-tense-event-recap.html.

[46] *See also* GILLESPIE, *supra* note 5, at 12 (pointing to an overrepresentation of libertarian white males in Silicon Valley).

[47] Eric Goldman, *Facebook Isn't—and Shouldn't Be—A Democracy*, Dec 17, 2012, https://www.forbes.com/sites/ericgoldman/2012/12/17/facebook-isnt-and-shouldnt-be-a-democracy.

democracy itself.' … [A]bsolutist constitutionalism … occurs when 'a single decisionmaker motivated by an interest in the nation's well-being consults widely and protects civil liberties generally, but in the end, decides on a course of action in the decisionmaker's sole discretion, unchecked by any other institutions.'"[48] Even if we are highly confident that the decisionmaker is good—and very few platforms have earned our trust in this regard—there remains the question of whether it is legitimate for decisions to be made in this way. Is users' presumed consent enough, especially when it is so very hard to avoid Facebook or its subsidiaries?

C. But Can a Government Moderate Content? Possible American Analogies

An obvious problem of treating Facebook like a real government is that, in the United States, Facebook-as-government could not regulate much of the speech that makes parts of the web a hateful cesspool, given the commands of the First Amendment—nor could the government regulate for Facebook.[49] This blunt legal fact seems to be driving the somewhat clumsy "fiduciary" workaround for improving governance from U.S. theorists.

Many people who think the First Amendment is generally a good idea might still be willing to accept more interventionist models of content moderation because institutions like Facebook and Google are more like schools than like full governments. These platforms are institutions that have near-total effects on some people some of the time, but that still leave a broader world outside with which users have significant interactions. Similarly, schools needn't be representative democracies as long as they have sufficient connection to the overall democratic polity, such that policies are in the end influenced both by area experts and by the people's elected representatives. There's a hell full of devils in the details. Still, democratic oversight might be sufficient to establish the kinds of policies we want, while leaving implementation largely to unelected people within given platforms.

The school model may also provide some guidance for handling the inevitability of terrible individual decisions by platforms. For both schools and platforms, failing an individual is simultaneously a tragedy worth investing heavily

[48] David Pozen, *Authoritarian Constitutionalism in Facebookland*, KNIGHT FIRST AMEND. INST. (2017), https://knightcolumbia.org/content/authoritarian-constitutionalism-facebookland**.**
[49] Whitney, *supra* note 29, at 29.

in preventing and also guaranteed to happen some percentage of the time, given the size of the task—and even a small percentage is a large absolute number. Failures should neither be excused nor used to condemn the entire enterprise. The spread of viral content ensures that there will always be something else shocking to worry about, some new violation of the rules or violation of norms that requires a new rule we didn't know we needed before. Children are equally inventive, and constantly new ones appear, in need of proper socialization. Schools will therefore always be in crisis, always making thousands of small-scale content moderation decisions. We generally think of handling this ongoing chaos as a matter of training and some outside supervision, rather than individualized judicial review except in the very worst cases of abuse. The system can be functional even if it predictably fails in small ways all the time, as long as we neither think we have the best system possible nor give up on it as too rotten to improve. That might be the best feasible model for the larger information environment, as well.

We could also look beyond schools to find other models of messy content regulation within democratically responsive governance structures. For example, what if platforms were run the way public libraries are? Libraries are the real "sharing" economies, and in the U.S., they have resisted government surveillance and content filtering as a matter of mission. The AO3 has much more in common with libraries than it does with Facebook. Just as libraries struggle to welcome everyone and to arrange their space so that the people looking at adult content don't drive out the kids and the kids don't preclude adults from reading what they choose to read, the AO3's design prioritizes the ability to host multiple overlapping and sometimes conflicting communities.

Instead of managing physical space, the emphasis online is on "tagging" content so that only those who want to see, or don't mind seeing, particular types of content will be exposed to it. Along with requiring creators to identify the fandom of a work and encouraging them to identify the characters who appear and the romantic or sexual relationships that will be featured, the AO3 has several major content warnings—for rape, major character death, and underage sex—that users repeatedly indicated were important to them. Creators can also add whatever additional "freeform" tags they want to warn or entice readers. Further, the AO3 allows creators a "Choose Not To Warn" option, which itself puts readers on notice that any of the major content warnings may or may not be present. The theory is that readers can decide for themselves what risks they are willing to take. A violation of the tagging rules in general does not lead a user's work to be removed,

but, in appropriate cases, the abuse team may add "Choose Not To Warn" if the creator fails to do so herself. Again, the nonprofit structure and ethos matters to the sustainability of this model: creators and users have generally embraced tagging because they want to find the right audiences and creators, but tagging can also be cumbersome and time-consuming, so most commercial platforms require far less than the AO3 does in order to not deter potentially valuable content.[50]

Together, the school and the library provide additional models for dealing with disputes over content and behavior. While they certainly aren't perfect, they have the advantage of being established parts of public infrastructure, with histories of imagining the world differently than do private profit-seeking corporations.

III. LOVE IN THE TIME OF CONTENT MODERATION

James Grimmelmann memorably argued that "responsible content moderation is necessary, and … responsible content moderation is impossibly hard."[51] Paradox, however, need not mean defeat. It could mean instead that we must continue to fight bad behavior, and fail, and continue again. If there is any hope for the future of democratic governance (which there may not be) it is that contradiction is where democracies, if not algorithms, regularly live, trying not to be overwhelmed by the paradox of tolerance. In this chaos, it would be helpful to remember that Facebook and YouTube aren't the world, and they aren't even the internet. We don't need to accept them as they are, and we also shouldn't accept them as our models for internet governance, whether internally generated or externally imposed. There are other possibilities, and some of them already exist; as lawmakers increasingly experiment with new forms of internet regulation, they should take care not to crush those alternatives in the name of bringing Facebook and Google to heel.

[50] *See* GILLESPIE, *supra* note 5, at 199-200 (discussing platforms' hesitancies to embrace flagging, but expanding upon the possibility of tagging as a mechanism to filter out material and content).
[51] Grimmelmann, *supra* note 4, at 217.