Yale Information Society Project

House of Representatives Committee on Energy and Commerce
Subcommittee on Communications and Technology
"Algorithms: How Companies' Decisions About Data and Content Impact Consumers"
November 29, 2017

Written Remarks of Kate Klonick
Yale Law School Information Society Project*

Chairman Blackburn, Ranking Member Doyle, and Members of the Subcommittee:
Every day millions of people around the world post pictures, videos, and text to online speech platforms, but not everything that is posted remains there. Sites like Facebook, Twitter, and YouTube actively curate the content that is posted by their users through a mix of algorithmic and human processes, broadly termed content moderation. Until recently, how and why these platforms made these decisions on user speech was largely opaque. For two years I have interviewed over three-dozen former and current executives and content moderation workers at these companies in an effort to better understand how and why these platforms regulate content.

This written testimony borrows heavily from my Article summarizing those findings[1] and attempts to clarify a few major points about content moderation, including:

- The vast majority of content moderation of user content (roughly estimated at over 90%) is done by trained human content moderators who review content only after it has been flagged by platform users and **not by algorithms**, contrary to this hearing's title.
- While users at sites like Facebook are given a public set of "Community Standards" guiding what kind of content is posted on the site, a separate much more detailed, and much more regularly updated set of internal rules is used by human moderators in making their decisions. These internal rules, at least at Facebook, are not currently known to the public.[2]
- That Facebook, and most platforms, use one global set of rules (with exceptions to comply with Nation-State laws) to curate content. This means, for example, that definitions of "inappropriate sexual activity" are the same for users in Canada, as they

* Resident Fellow at the Information Society at Yale Law School; Ph.D Candidate in Law, Yale University; J.D. Georgetown University Law Center; A.B. Brown University. I'm testifying on own behalf, not on behalf of my employer or anyone else.
Email: kate.klonick@yale.edu Website: www.kateklonick.com

[1] Kate Klonick, *The New Governors: The People, Rules, and Processes Governing Online Speech, forthcoming* HARV. L. REV. (2018). Available for download at https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2937985

[2] In May 2017, *The Guardian* published a series of documents claiming to be the "leaked rules" of Facebook. In fact, these were not the precise rules, but rather slides used to train human content moderators on Facebook's internal rules. Nick Hopkins, *Revealed: Facebook's internal rulebook on sex, terrorism and violence*, GUARDIAN (May 21, 2017) https://www.theguardian.com/news/2017/may/21/revealed-facebook-internal-rulebook-sex-terrorism-violence.

are for users in India, as they are for users in France—irrespective of the norms of each country.

- These platforms intricate systems of governance to regulate content are a response to the Communications Decency Act Section 230, which incentivized sites to remove offensive content with immunity from intermediary liability.[3] In many ways, these platforms' self-regulation has met the goals of Section 230, but as access to online speech platforms has increasingly become an essential public right[4] new concerns about the expansive immunity granted under Section 230 are being raised. While these and other concerns are undoubtedly present, changes to Section 230 or new regulation that might affect it, should be considered with extreme caution and with a full appreciation of the potential damage that could be caused to consumer rights.
- While there have long been worries about internet service providers favoring access to some content over others, there has been less concern about companies further along the pipeline holding an internet on/off switch. In large part, this is because at other points in the pipeline, users have choice. But the fewer choices you have for the infrastructure you need to stay online, the more serious the consequences when companies refuse service. This is one important reason net neutrality is so important. As Section 230 reveals, we generally agree that it's appropriate for social media companies to take down certain kinds of content — that's how they ensure our newsfeeds aren't full of pornography or violence. But that doesn't mean we don't want that type of content to be able to exist *somewhere* on the Internet. Ensuring that ISPs remain neutral is necessary to guaranteeing the continuation of a free and open Internet.

How Platforms Moderate Content

Content moderation happens at many levels. It can happen before content is actually published on the site as with *ex ante* moderation, or after content is published, as in *ex post* moderation. These methods can be either *reactive*, in which moderators passively assess content

[3] The ability of private platforms to moderate content comes from § 230 of the Communications Decency Act, which gives online intermediaries broad immunity from liability for user generated content posted on its site. 47 U.S.C. § 230. The purpose of this grant of immunity was both to encourage platforms to be "Good Samaritans" and take an active role in removing offensive content, and also to avoid free speech problems of collateral censorship. *See* Zeran v. Am. Online, Inc. 129 F.3d 327, 330 (4th Cir. 1997) (discussing the purposes of intermediary immunity § 230 were not only to incentivize platforms to remove indecent content, but to protect the free speech of platform users). *See also* Eric Goldman, *Ten Worst Section 230 Rulings of 2016 (Plus the Five Best*), (Jan. 4, 2017) at http://blog.ericgoldman.org/archives/2017/01/ten-worst-section-230-rulings-of-2016-plus-the-five-best.htm. For a comprehensive and complete cataloging of § 230 cases with context and commentary, see Professor Eric Goldman's blog, http://blog.ericgoldman.org/.

[4] Packingham v. North Carolina, 137 S.Ct. 1730 (2017) (holding that a state statute barring registered sex offenders from using online social media platforms was unconstitutional under the First Amendment). In his opinion, Justice Kennedy wrote that "[w]hile in the past there may have been difficult in identify the most important places (in a spatial sense) for the exchange of views, today the answer is clear. It is cyberspace¬¬–the 'vast democratic forums of the Internet' in general, and social media in particular." *Id.* at 1735(quoting Reno v. ACLU, 521 U.S. 844, 868 (1977)).

and update software only after other users bring the content to their attention, and *proactive* moderation, in which teams of moderators actively seek out published content for removal. Additionally, these processes can be *automatically* made by software or algorithms, or *manually* made by humans. [5]

1. *Ex Ante* Content Moderation[6]

When a user uploads a video to Facebook, a message appears: "Upload Completed: The video in your post is being processed. We'll send you a notification when it's done and your post is ready to view."[7] *Ex ante* content moderation is the process that happens in this moment between upload and publication. The vast majority of *ex ante* content moderation is an automatic process largely run through algorithmic screening without the active use of human decision-making.

An example of such content is child pornography, which can reliably be identified on upload to a site through a picture recognition algorithm called PhotoDNA.[8] Under federal law, production, distribution, reception, and possession of an image of child pornography is illegal, and as such, sites are obligated to remove it.[9] A known universe of child pornography—around

[5] *Cf.* James Grimmelmann, *The Virtues of Moderation*, 17 YALE J.L. & TECH. 42, 63-70 (2015) (describing how a moderation system operates through distinctions between automatic, manual, transparent, secret, *ex ante*, *ex post*, centralized, and decentralized features). Grimmelmann's taxonomy, while foundational, speaks more generally to all of Internet moderation rather than content publishing platforms, specifically. In the context of speech, the distinction between *ex ante* and *ex post* is especially important to determine if moderation is happening before or after publication. Of secondary concern is whether content is being moderated through reaction or through proactive measures. Finally, for the purposes of this hearing, the distinction between automatic or algorithmic moderation and human manual moderation is of central importance.

[6] Because it happens before publication takes place, *ex ante* content moderation is the type of prior restraint that scholars like Professor Jack Balkin are concerned with. *See* Jack M. Balkin, *Old-School/New-School Speech Regulation*, 127 HARV. L. REV. 2296,2299 (2014). Of the two automatic means of reviewing and censoring content—algorithm or geo-blocking—geo-blocking is of more concern for the purposes of collateral censorship and prior restraint. In contrast, algorithm take down is currently used to remove illegal content like child pornography or copyright violations. *But see* Rebecca Tushnet, *Power Without Responsibility: Intermediaries and the First Amendment*, 76 GEO.WASH. L. REV. 986, 1003-05 (2008) (noting that the DMCA notice-takedown provisions give platforms no incentive to investigate and therefore "suppress critical speech as well as copyright infringement.").

[7] FACEBOOK, UPLOADING & VIEWING VIDEOS (accessed Mar. 1, 2017) https://www.facebook.com/help/154271141375595/?helpref=hc_fnav

[8] Tracy Ith, *Microsoft's PhotoDNA: Protecting children and businesses in the cloud*, MICROSOFT NEWS (accessed Mar. 1, 2017) https://news.microsoft.com/features/microsofts-photodna-protecting-children-and-businesses-in-the-cloud/#sm.001eom8zb14bad5htm11ixrkpzssa.

[9] *See* 18 U.S.C. § 2251; 18 U.S.C. § 2252; 18 U.S.C. § 2252A. It is important to remember that § 230 expressly states that no Internet entity has immunity from federal criminal law, intellectual property law or communications privacy law. This means that every Internet service provider, search engine, social networking platform and website is subject to thousands of laws, including child pornography laws, obscenity laws, stalking laws and copyright laws. 47 U.S.C. § 230 (e).

720,000 illegal images–exists online.[10] By converting each of these images to gray scale formatting, overlaying a grid, and assigning a numerical value to each square, researchers were able to create a "hash" or signature that remained even if the images were altered. As a result, platforms can determine within micro-seconds between upload and publication if an image contains child pornography.[11] Geo-blocking is another form of automatic *ex ante* moderation. Unlike PhotoDNA, which prevents the publication of illegal content, geo-blocking prevents both the publication and viewing of certain content based on a user's location. As happened in the controversy over the *Innocence of Muslim* video, geo-blocking usually comes at the request of a government notifying a platform that a certain type of posted content violates its local laws.

It is important to note that, of course, algorithms do not decide for themselves which kind of content they should block from being posted. Content screened automatically is typically content that can reliably be identified by software and is illegal or otherwise prohibited on the platform. This universe of automatically moderated *ex ante* content is regularly evaluated and updated through iterative software updates and machine learning. For example, in a similar fashion to PhotoDNA, potential copyright violations can be moderated proactively through software like ContentID. Developed by YouTube, ContentID allows creators to give their content a "digital fingerprint" so it can be compared against other uploaded content. Copyright holders can also flag already published copyright violations through notice and takedown.[12] These two systems work together, with user-flagged copyrighted material eventually added to ContentID databases for future proactive review. This mix of proactive, manual moderation, informed and automatic *ex ante* moderation is also evident in the control of spam. All three platforms (and most Internet companies, generally) struggle to control spam postings on their sites. Today, spam is mostly blocked automatically from publication through software. Facebook, Twitter, and YouTube, however, all feature mechanisms for users to report spam manually.[13] *Ex ante* screen software is iteratively updated to reflect these flagged spam sources.

2. *Ex Post* Proactive Manual Content Moderation

Recently, a form of content moderation that harkens to the earlier era of AOL chat rooms has re-emerged: platforms proactively using their own moderators, instead of relying on flagging by users to seek out and remove published content. Currently, this method is largely confined to the moderation of extremist and terrorist speech. As of February 2016, dedicated teams at

---

[10] This "known universe" of child pornography is maintained and updated by the International Centre for Missing and Exploited Children and the U.S. Department of Homeland Security in a program known as Project Vic. Mark Ward, *Cloud-based archive tool to help catch child abusers*, BBC NEWS (Mar. 24, 2014) http://www.bbc.com/news/technology-26612059.

[11] Ith, *supra* note 8.

[12] *See e.g.*, YOUTUBE, *YouTube Help: Submit a copyright takedown notice*, https://support.google.com/youtube/answer/2807622 (last visited Aug. 15, 2016).

[13] *See e.g.* Panda Security, *How Twitter aims to prevent your timeline from filling up with spam* (Sept. 12, 2014) http://www.pandasecurity.com/mediacenter/social-media/twitter-spam/; James Parsons, *Facebook's War Continues Against Fake Profiles and Bots*, HUFF. POST (May 22, 2015) http://www.huffingtonpost.com/james-parsons/facebooks-war-continues-against-fake-profiles-and-bots_b_6914282.html.

Facebook proactively removed all posts or profiles with links to terrorist activity.[14] Such efforts were doubled in the wake of terrorist attacks and the events in Charlottesville.[15] This is an important new development affecting content moderation with an ever-evolving balance between ensuring national security yet maintaining individual liberty and freedom of expression, but it still only comprises a small amount of the total moderation that happens on these sites.

3. *Ex Post* Reactive Manual Content Moderation

As previously mentioned, with the exception of proactive moderation for terrorism described above, almost all user-generated content that is published is reviewed *reactively*, that is, through *ex post* flagging by other users and reviewed by human content moderators against internal guidelines. Flagging—alternatively called reporting—is the mechanism provided by platforms to allow users to express concerns about potentially offensive content.[16] The adoption by social media platforms of a flagging system serves two main functions: (1) it is a "practical" means of reviewing huge volumes of content, and (2) its utilization of users serves to legitimize the system when platforms are questioned for censoring or banning content.[17]

Facebook users flag over one million pieces of content worldwide every day.[18] Content can be flagged for a variety of reasons and the vast majority of items flagged do not violate the Community Standards of Facebook. Instead they often reflect internal group conflicts or disagreements of opinion. To resolve the issue, Facebook created a new reporting "flow"—the industry term to describe the sequence of screens users would experience as they made selections—that would encourage users to resolve issues themselves rather than report them for review to Facebook.[19] Facebook has also designed its reporting flow to triage flagged content for review. This makes it possible for Facebook to immediately prioritize certain content for review, and when necessary, notify authorities of emergency situations like suicide, imminent threats of violence, terrorism, or self-harm. Other content, like possible hate speech or harassment, can be queued into less urgent databases for general review. [20]

When content is flagged or reported it is sent to a server where it awaits review by a human content moderator. At Facebook, there are three basic tiers of content moderators: "Tier 3" moderators, who do the majority of the day-to-day reviewing of content; "Tier 2" moderators,

---

[14] Natalie Andrews & Deepa Seetharaman, *Facebook Steps Up Efforts Against Terrorism*, WALL ST. J. (Feb. 11, 2016), http://www.wsj.com/articles/facebook-steps-up-efforts-against-terrorism-1455237595.

[15] *Id.*

[16] Kate Crawford & Tarleton Gillespie, *What is a flag for? Social media reporting tools and the vocabulary of complaint*, NEW MEDIA & SOC. (2014), at 2.

[17] *Id*. at 3.

[18] *See* Catherine Buni & Soraya Chemaly, The Secret Rules of the Internet, THE VERGE (Mar. 13, 2014), www.theverge.com/2016/4/13/11387934/internet-moderator-history-youtube-facebook-reddit-censorship-free-speech.

[19] *Radiolab: The Trust Engineers*, WNYC (Feb. 9, 2015) (downloaded using iTunes).

[20] *Facebook Reporting Guide: What Happens When You Report Something?*, uploaded to Scribd June 19, 2012 by Facebook Washington DC. https://www.scribd.com/doc/97568769/Facebook-Reporting-Guide. After content has been flagged to a platform for review, the precise mechanics of the decision-making process become murky. Platforms do not publish details of their internal content moderation guidelines; no major platform has made such guidelines public. Buni & Chemaly, *supra* note 18.

who supervise Tier 3 moderators and review prioritized or escalated content; and "Tier 1" moderators, who are typically lawyers or policy makers based at company headquarters.

In the early days—before 2008 to 2009—recent college graduates based in the San Francisco Bay Area did much of the Tier 3 content moderation.[21] Today, most platforms, including Facebook, either directly employ content moderation teams or outsource much of their content moderation work to companies based in the Philippines, Ireland, Singapore, India, or Eastern Europe.[22] Today, Tier 3 moderators typically work in "call-centers" in the Philippines, Ireland, Singapore, India, or Eastern Europe. Within Facebook, these workers are called "community support" or "user support teams."[23]

Tier 2 moderators are typically supervisors of Tier 3 moderators or specialized moderators with experience judging content. They work both remotely (many live in the United States and supervise groups that are internationally based) and locally at call-centers.[24] Tier 2 moderators review content that has been prioritized, like imminent threats of violence, self-harm, terrorism, or suicide that arrive to Tier 2 directly through the reporting flow or are identified and escalated to Tier 2 by Tier 3 moderators. Tier 1 moderation is predominantly performed by the legal or policy headquarters of a platform. At Facebook, for example, a Tier 3 worker could be based in Hyderabad, the Tier 2 supervisor could be based in Hyderabad, or remotely in a place like Dublin, but a Tier 1 contact would be based in Austin, Texas or the San Francisco Bay Area.

At Facebook, Tier 3 moderators have three decision-making options regarding content: they can "confirm" the content violates the Community Standards and remove it, "unconfirm" that the content violates Standards and leave it up, or escalate review of the content to a Tier 2 moderator or supervisor. The internal rules describe certain types of content requiring mandatory escalations. For example in 2012 at Facebook: child nudity or pornography, promotion or encouragement of bestiality, credible threats, bullying, self-harm content, poaching of endangered animals, Holocaust denial, all attacks on Ataturk, maps of Kurdistan and Burning Turkish Flags.[25] If a moderator has decided to ban content, a Facebook user's content is taken down, and she is automatically signed off of Facebook. When the user next attempts to sign in, she will be given the following message explaining without detail that an offensive post was removed in violation of community standards. At Facebook, users who repeatedly have content

---

[21] Buni & Chemaly, *supra* note 18.

[22] Adrian Chen, *The Laborers Who Keep Dick Pics and Beheadings Out of Your Facebook Feed*, Wired, (Oct. 23, 2014) https://www.wired.com/2014/10/content-moderation/);Adrian Chen, *Inside Facebook's Outsources Anti-Porn and Gore Brigade, Where 'Camel Toes' are More Offensive Than 'Crushed Heads*,' GAWKER (Feb. 16, 2012) http://gawker.com/5885714/inside-facebooks-outsourced-anti-porn-and-gore-brigade-where-camel-toes-are-more-offensive-than-crushed-heads. Within Facebook, these workers are called "community support" or "user support teams."

[23] *Id.*

[24] *Id.*; Telephone Interview with Dave and Charlotte Willner (Mar. 23, 2016).

[25] Abuse Standards (AS) 6.1 *available at* https://www.scribd.com/doc/81863464/oDeskStandards; Abuse Standards (AS) 6.2 *available* at https://www.scribd.com/doc/81877124/Abuse-Standards-6-2-Operation-Manual *hereinafter collectively* "Abuse Standards." These are copies of documents that were leaked from a content moderator working at oDesk (now UpWork) doing content moderation for Facebook. They are not the actual internal rules of Facebook, but they were oDesk's approximation of Facebook's rules in 2012.

removed are gradually escalated in punishment: two removed posts in a certain amount of time, for example, might mean your account is suspended for 24-hours.

Normative Implications of Platform Governance on Potential Regulation

These details about how and why platforms are governing user speech have direct implications on potential regulation and our understanding of online speech.

1. Any reform to Section 230 should be approached with caution

When CDA Section 230 was put into place in 1996, the Internet was a very different place. Spam and pornography were threatening to dominate platforms, but courts were beginning to hold platforms civilly liable if they acted to remove such content.[26] Section 230 lifted the "specter of tort liability" that might "deter service providers from blocking and screening offensive material" and also result in platforms removing too much use speech resulting in an "obvious chilling effect."[27] "Faced with potential liability for each message republished by their services, interactive computer service providers might choose to severely restrict the number and type of messages posted."[28]

In many ways, these major social media platforms' self-regulation has met the goals of Section 230—removing content that users find normatively unpalatable, while keeping up as much content as possible.[29] But in the 21 years since Section 230 was passed, access to online speech platforms has increasingly become an essential public right new and concerns about the expansive immunity granted under Section 230 are being raised. While these and other concerns are undoubtedly present, changes to Section 230 or new regulation that might affect it, should be considered with extreme caution and with a full appreciation of the potential damage that could be caused to consumer rights and free speech online.

2. Speech platforms' ability to self-regulate content has little to no direct applicability to broadband ISPs ability to self-regulate

Generally speaking, there are two kinds of corporate players on the internet: companies that build infrastructure through which content flows, and companies that seek to curate content and create a community. Internet service providers like Verizon and Comcast, domain name servers, web hosts and security services providers are all the former — or the "pipe." They typically don't look at the content their clients and customers are putting up, they just give them the means to do it and let it flow. Social media platforms like Facebook are the latter. They

---

[26] *See* Cubby v. Compuserve, 776 F. Supp. 135, 138 (S.D.N.Y. 1991) (holding CompuServe could not be held liable for the defamatory content because the intermediary did not review any of the content posted to the forum) *and* Stratton Oakmont, Inc. v. Prodigy Servs. Co., 1995 WL 323710 (N.Y. Sup. Ct. 1995) (holding intermediary Prodigy was liable as a publisher for all posts made on its site, because it voluntarily deleted some forum postings).

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

[28] *Id*. The quote continues: "Congress considered the weight of the speech interests implicated and chose to immunize service providers to avoid any such restrictive effect."

[29] Eric Goldman, *The Ten Most Important Section 230 Rulings*, 20 TULANE J. TECH& I.P. __ (2017),https://ssrn.com/abstract=3025943.

encourage their users to create, share and engage with content — so they look at content all the time and decide whether they want to allow hateful material like that of neo-Nazis to stay up.

While there have long been worries about internet service providers favoring access to some content over others, there has been less concern about companies further along the pipeline holding an Internet on/off switch. In large part, this is because at other points in the pipeline, users have choice. Private companies can make their own rules, and consumers can choose among them. If GoDaddy won't register your domain, you can go to Bluehost or thousands of other companies. And while there may only be one Facebook, there are billions of other platforms and places online to post speech.

But the fewer choices you have for the infrastructure you need to stay online, the more serious the consequences when companies refuse or throttle service. This is one important reason net neutrality is so important. As Section 230 reveals, we all generally agree that it's appropriate for social media companies to take down certain kinds of content — that's how they ensure our newsfeeds aren't full of pornography or violence. But that doesn't mean we don't want that type of content to be able to exist *somewhere* on the Internet. Ensuring that ISPs remain content-neutral is necessary to guarantee that.[30]

[30] This section borrows heavily from my article about the potential dangers in allowing internet infrastructure to regulate content. Kate Klonick, *The Terrifying Power of Internet Censors*, N.Y. TIMES (Sept. 13, 2017) https://www.nytimes.com/2017/09/13/opinion/cloudflare-daily-stormer-charlottesville.html?.