IMPROVING MODERATION

Tarleton Gillespie

This is an expanded version of part of Chapter 8 of my book, Custodians of the Internet: Platforms, Content Moderation, and the Hidden Decisions that Shape Social Media *(Yale, 2018). http://custodiansoftheinternet.org/*
The trimmer version can be found on pages 198-202.
The section titled "Amend the Section 230 safe harbor" previously appeared in "How Social Networks Set the Limits of What We can Say Online," Wired*, June 26, 2018. https://www.wired.com/story/how-social-networks-set-the-limits-of-what-we-can-say-online/ and in "Platforms are not Intermediaries,"* Georgetown Law and Technology Review*, July 2018.*

There are many things social media companies could do to improve their content moderation: More human moderators. More expert human moderators. More diverse human moderators. More transparency in the process. Better tools for users to block bad actors. Better detection software. More empathetic engagement with victims. Consulting experts with training on hatred and sexual violence. Externally imposed monitors, public liaisons, auditors, and standards. And we could imagine how we might compel those changes: Social pressure. Premium fees for a more protected experience. Stronger legal obligations.

But these are all are just tweaks—more of the same, just more of it. And some of them are likely to happen, in the short term, as the pressure and scrutiny social media platforms face increase, and they look for steps to take that moderately address the concerns while preserving their ability to conduct business as usual. But it is clearer now than ever that the fundamental arrangement itself is flawed.

If social media platforms wanted to do more, to come at the problem in a fundamentally different way, I have suggestions that more substantively rethink the not only their approach but how platforms conceive of themselves and their users. This is not an exhaustive list of what must be done, and some of the suggestions have also been made by others. I fully acknowledge that some are politically untenable and economically outlandish, and are almost certain never to happen.

Design for Deliberate and Actionable Transparency

Calls for greater transparency in the critique of social media are so common as to be nearly vacant. But the workings of content moderation at most social media platforms are shockingly opaque, and not by accident.[1] The labor, the criteria, and the outcomes are almost entirely kept from the public. On some platforms, content disappears without explanation and rules change with notification; when platforms do respond publicly regarding controversial decisions, their statements are often short on detail and rarely articulate a larger philosophy.

Platform moderation should be much more transparent. Full stop. I am certainly not the first to say so. A superb example is the 2018 "Santa Clara Principles," which argues for what categories of moderation data platforms should report and what notification and appeals procedures they should offer users.[2] Important efforts to induce more transparency from platforms include the "Corporate Accountability Index" from Rebecca McKinnon and Ranking Digital Rights,[3] the efforts of Jillian York and Onlinecensorship.org to gather user reports about their tangles with the moderation process, and the 2018 report from David Kaye, Special Rapporteur for the U.N. Human Right Council, arguing that a commitment to transparency (among others) is essential if platform moderation is to better align with basic human rights.[4]

But transparency is not merely the absence of opacity. It requires designing new ways to make processual information visible but unobtrusive. For instance, if one of my tweets is receiving lots of responses from "egg" accounts—often the ones dedicated to trolling—that I have already blocked, I may be left unaware of a growing threat and unable to document that threat for the benefit of law enforcement, and I may radically misunderstand the impact of my own statements. How could the fact of these harassing tweets, and their number and velocity, still be made visible to me?[5] Tiny eggs, swarming like angry bees below my tweet? A pop-up histogram that indicates the emotional intensity of the responses, algorithmically estimated? The imperative for platforms to smooth and sanitize the user experience must be tempered with an obligation to make the moderation visible.

This runs counter to the efforts of most platforms, that have aimed to make moderation not just unobtrusive but undetectable. Theirs is a fundamentally paternalistic approach, reserving for the platform

[1] Mikkel Flyverbom, 2016. "Digital Age Transparency: Mediation and the Management of Visibilities." *International Journal of Communication* 10: 13.

[2] "The Santa Clara Principles," May 2018. https://cdt.org/files/2018/05/Santa_Clara_Principles.pdf

[3] 2018 Corporate Accountability Index, Ranking Digital Rights, https://rankingdigitalrights.org/index2018/

[4] David Kaye, "Report of the Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression," April 6, 2018. https://freedex.org/a-human-rights-approach-to-platform-content-regulation/

[5] Twitter has since changed the "egg" icon that used to represent accounts that had not added a profile photo, because it had become associated with trolling. Sorry, it's still an egg. See Kaitlyn Tiffany, "Twitter Wants You to Stop Saying 'Twitter Eggs,'" *Verge*, March 31, 2017, https://www.theverge.com/tldr/2017/3/31/15139464/twitter-egg-replaced-harassment-terms-trolls-language.

the right to decide not only what to hide, from whom, and why, but also when to override even the stated wants of the user. In Chapter 7 of the book, I raised my concerns about "moderation by design": putting objectionable content in front of some users and hiding it from others. The central appeal of this tactic is its opacity. Of course, the platforms have compelling reasons for this approach: by rendering not just the content but even its removal invisible, they clean up and smooth the user experience. Rendering harassers invisible to their targets not only protects the targets from further attack, it undercuts the sense of impact the harasser gets to feel, which may diminish motivation to do it in the first place.

On balance, I find this approach troubling. There is something dishonest about being able to hide what elsewhere is made available. Even when a video store puts the adult movies in the back room, the fact that they offer such videos is still visible—that visibility forces the store owner to stand behind that decision, and forces the primmest customer to acknowledge that there must be demand for such films.

What does get removed should be signaled as such. Except when the platform can specifically demonstrate that some harm would follow, making their intervention visible and leaving some trace evidence of what has been removed is always preferable. This gives the user the certainty and the actionable knowledge of what the platform is doing on her behalf. Twitter accomplished this with country-specific legal takedowns, indicating that a tweet is gone, noting who requested its removal, and providing links to resources like chillingeffects.org.[6] This should also be true for content removed by the platform, users it has banned, comments it has hidden from view. A user should have an attenuated ability to access content directed at her, even if it has been removed: make it reviewable with a click, or offer a way to request that the platform preserve it on her behalf as a record, or make it downloadable as an archive. Deleted content could be tagged with the date of removal and rule invoked, as well as the number of users who had flagged it over what period of time, as well as whether it was identified by an algorithmic process, by users, or by the content moderation team itself.

Platforms could also be much more transparent not just about what has been policed away but about what they are still interacting with, so that users can make savvier decisions about them. Platforms collect and analyze an immense amount of data, and much of this is already packaged and made available to their content reviewers, as a handy dashboard to help them make moderation judgments. They owe users that data too, all of it, in a form that helps them make smarter navigation judgments.

[6] See, for example, Twitter's transparency reports, at https://transparency.twitter.com/. See also Loretta Chao and Amir Efrati, "Twitter Can Censor by Country" *Wall Street Journal*, January 28, 2012. https://www.wsj.com/articles/SB10001424052970204573704577185873204078142; Zeynep Tufekci, "Why Twitter's New Policy is Helpful for Free-speech Advocates," Technosociology, January 27, 2012. http://technosociology.org/?p=678

For example, the concern about "fake news" made clear how easy it is for a user to forward a link without having thought about its source or validity. Cascades of gossip and propaganda could flow before anyone noted that the headline being passed around was false. Facebook's solution to this problem was to mark disputed articles, and provide users a way to flag content they were suspicious of. But why not provide more information about every forwarded headline or link, and every user doing the forwarding. If I hover over a link in my news feed, I would like to see a dashboard that reports: how long the source of that link has been online, a graph of how much and how quickly that headline is being forwarded, the headlines before and after this one from the same source, and how often other articles from that source have been disputed by fact checkers. I would like to see how long the person who posted the link has been on Facebook, the last five articles he forwarded before this one, whether he has ever been reprimanded or suspended by the platform, and whether he read the article before forwarding it.

Some platforms do the tiniest bit of this, such as indicating how many followers someone has or whether a profile is verified. This is an anemic commitment to transparency, shaped more by the economic and design imperatives of the site. Moving forward, platforms should make a radical commitment to turning the data they already have back to me in a legible and actionable form, everything they could tell me contextually about why a post is there and how I should assess it. We have already paid for this transparency, with our data.

Honest Ads, for all Ads

The need for radical transparency goes double for advertising, as well as paid or otherwise boosted content, on social media platforms. The debate about content moderation has focused almost exclusively on what users say in the spaces designed for them: posts, photos, videos, comments, groups. Advertising has been considered a separate element, that has until recently been left out of the public debates.

This turned out to be an unfortunate omission, one that misunderstood how advertising has been changed by social media platforms. The cost of advertising has dropped precipitously, because small buyers can narrowly target just a few users; this means advertising on social media platforms as almost as open and available to users as the platforms themselves. Advertising no longer comes only from well-funded companies and campaigns, it includes anyone motivated enough to drop a couple hundred dollars to make sure they're heard. And on social media platforms, advertising can circulate almost as fluidly as user content, in ways that television and print ads do not. This became abundantly clear in 2016, when interested parties from around the world, including Russia, Macedonia, as well as inside the U.S., found ways to manipulate the political discourse surrounding the 2016 U.S. presidential election – using tactics that took advantage of the circulation of ads and the circulation of content, as well as the blurry lines between them.[7]

In response to these concerns, Congress introduced the Honest Ads Act.[8] If enacted, the law would require that "political ads" on social media disclose who paid for them; this would extend the existing Federal Elections Commission rules that already apply to television and print advertising. Though initially reluctant to support the bill,[9] Facebook and Twitter have now instituted similar policies themselves – though they have also faced criticism since for how they discern and label "political" content.[10]

Passing the Honest Ads Act would be an excellent start. There's no reason why the sensible obligations around political campaign ads on television shouldn't also extend to social media. But the bill is fundamentally insufficient, given the way social media platforms were exploited in 2016 and since. The bill applies only to "political ads," but very few of the ads purchased by foreign interests during the 2016 presidential race were for or against a particular candidate. Most didn't name a candidate at all, so they would not qualify as traditional "electioneering communications."[11] The Honest Ads Act does try to address this, by applying also to messages relating to "a national legislative issue of public importance." Still, it is not clear that an ad arguing that "all lives matter" qualifies. Most of the 2016 ads aimed to stir up discord by slamming a political position – or by speaking as if in support of that position, but presenting it in its most extreme form.[12] Few of these would have to honor the requirements that would be imposed by the Honest Ads Act.

What's worse, recent evidence suggests that there are particular problems with platform advertising as a whole. A 2017 ProPublica exposé revealed some of the problematic mechanisms that

[7] Nicholas Thompson, "Exclusive: Facebook Opens Up About False News" Wired May 23, 2018. https://www.wired.com/story/exclusive-facebook-opens-up-about-false-news

[8] S. 1989, Honest Ads Act. https://www.congress.gov/bill/115th-congress/senate-bill/1989. See for updates, https://www.warner.senate.gov/public/index.cfm/the-honest-ads-act

[9] Heather Timmons and Hanna Kozlowska, "Facebook's quiet battle to kill the first transparency law for online political ads" *Quartz*, March 22, 2018. https://qz.com/1235363/mark-zuckerberg-and-facebooks-battle-to-kill-the-honest-ads-act/

[10] Russell Brandom, "Facebook's new political ad policy is already sweeping up non-campaign posts," *The* Verge, May 29, 2018. https://www.theverge.com/2018/5/29/17406658/facebook-political-ad-speech-block-russia-troll

[11] Federal Election Commission, "Making electioneering communication", https://www.fec.gov/help-candidates-and-committees/making-disbursements-ssf-or-connected-organization/making-electioneering-communications/

[12] Bill Allison, "Most of Divisive Facebook Ads Were Paid for by 'Suspicious' Groups," Bloomberg, April 16, 2018. https://www.bloomberg.com/news/articles/2018-04-16/most-of-divisive-facebook-ads-paid-for-by-suspicious-groups; Tony Romm, "'Pro-Beyoncé' vs. 'Anti-Beyoncé': 3,500 Facebook ads show the scale of Russian manipulation" *Washington Post*, May 10, 2018. https://www.washingtonpost.com/news/the-switch/wp/2018/05/10/here-are-the-3400-facebook-ads-purchased-by-russias-online-trolls-during-the-2016-election/

Facebook offers to advertisers.[13] Advertisers can target users not only by traditional demographics like age, gender, or region, but highly specific and personalized information like friends, patterns of likes or other activity, or expressed interests. ProPublica was able to buy ads targeting users who had put "Jew hater" in their Facebook profile. Buzzfeed noticed a similar problem with Google: the tool allowed advertisers to use hateful slurs to target their ads, and even proposed others: "Type 'Why do Jews ruin everything,' and Google will suggest you also run ads next to searches including 'the evil jew' or 'jewish control of banks.'"[14] Platforms should prevent such egregious techniques. But I bring them up to highlight a broader problem: online advertising can be targeted along a nearly infinite array of vectors, including problematic ones, all invisible to the user.

Microtargeting tactics have become commonplace. But some platforms have developed sophisticated tools to allow advertisers to target even more precisely. Facebook allows "dark ads": advertisers can run different versions of an ad, targeted at different micro-audiences; this lets them A/B test different ads to see how much traction they get. The way dark ads currently work, it is difficult for critics and lawmakers to see which ads went where. And on some platforms, ads can quickly become organic content, further obscuring their origins. Users on Facebook can "like" an ad – at which point it will be delivered to their friends' news feeds, just as if they had liked an article or video. Further circulation of that ad loosens it from any sign that it was ever an ad at all. Advertisers take advantage of this, by designing ads to look like news articles, or to pay advertising rates to circulate their news articles.

All of this means that, whether it's a campaign ad, a politically divisive piece of content paid for through the advertising tools, or a traditional commercial, users are structurally inhibited from knowing much about who bought that ad, how it ended up in front of them, or whether it is even an ad at all. This must change.[15] Thankfully, while Facebook, Google, and other social media platforms exacerbate the problem of transparency and honesty in ads, they also offer a solution. I believe the answer is not to drop this effort to make political ads transparent, but to double down. I think platforms should label not just political ads, but paid ones, all of them.

[13] Julie Angwin, Madeleine Varner, and Ariana Tobin, "Facebook Enabled Advertisers to Reach 'Jew Haters'," *ProPublica*, September 14, 2017. https://www.propublica.org/article/facebook-enabled-advertisers-to-reach-jew-haters; Robinson Meyer, "Could Facebook Have Caught Its 'Jew Hater' Ad Targeting?" *The Atlantic*, September 15, 2017. https://www.theatlantic.com/technology/archive/2017/09/on-facebook-advertisers-can-show-their-ads-only-to-jew-haters/539964/

[14] Alex Kantrowitz, "Google Allowed Advertisers To Target People Searching Racist Phrases," *Buzzfeed*, September 15, 2017. https://www.buzzfeed.com/alexkantrowitz/google-allowed-advertisers-to-target-jewish-parasite-black

[15] Sara M. Watson, "Russia's Facebook ads show how Internet microtargeting can be weaponized" *Washington Post*, October 12, 2017. https://www.washingtonpost.com/news/posteverything/wp/2017/10/12/russias-facebook-ads-show-how-internet-microtargeting-can-be-weaponized/

Social media platforms, search engines, and advertising networks have an immense amount of data about their ads. They know who purchased them (or could demand to know), they know how they were targeted, they know when they ran, they know who saw them, they know how they circulated further through their network. They have all of this data, or could have it, and they deliver much of it back to the advertisers themselves in the form of analytics. So why couldn't they offer that data, all of that data, to users?[16]

Knowing who has paid for an ad is not as easy as it sounds. But, especially given Facebook's history, to offer this as a reason not to make payment transparent is astounding in its bravado. Facebook has for years dug in its heels, standing by the requirement that users only have one account, and be represented by their real names.[17] They asserted that requiring users to present their real identity makes for richer and kinder social interactions, diminishes the kind of harassment and fraud that was possible under the cloak of anonymity, an alias, or multiple accounts. They have suspended users who wanted to use a second name, or an alias, including whistleblowers and drag queens.[18] And yet they have not extended the same expectation to advertisers.

Call it the Honest Ads Act on steroids. Such a law might require the following of all platforms and search engines that accept advertising or payment of any kind:

* every advertiser, at whatever dollar amount, must have a public profile on the platform on which they purchased the ad; payments to a platform cannot be made anonymously

* their real identity must be made known to the platform and published on the profile, before paid content can be posted – including an institutional name, geographic location, and how long they have been on that platform

* every ad they post to that platform must be publicly visible on that profile, and remain there

* with every ad on that profile the platform should provide, at minimum, the following information:

- the days on which that ad was delivered to any users

[16] Abby K. Wood, Ann M Ravel, and Irina Dykhne. 2018. "Fool Me Once: Regulating 'Fake News' and Other Online Advertising." *Southern California Law Review* 6: 91+.

[17] danah boyd, 2012. "The Politics of 'Real Names.'" *Communications of the ACM* 55 (8): 29. https://doi.org/10.1145/2240236.2240247; Haimson, Oliver L., and Anna Lauren Hoffmann. 2016. "Constructing and Enforcing 'Authentic' Identity Online: Facebook, Real Names, and Non-Normative Identities." *First Monday* 21 (6). http://ojs-prod-lib.cc.uic.edu/ojs/index.php/fm/article/view/6791.

[18] Jessa Lingel and Tarleton Gillespie. 2014. "One Name to Rule Them All: Facebook's Identity Problem." *The Atlantic*, October 2, 2014. http://www.theatlantic.com/technology/archive/2014/10/one-name-to-rule-them-all-facebook-still-insists-on-a-single-identity/381039/.

- the targeting criteria by which it was delivered, including demographics, psychographics, target phrases, similar users, etc.
- the number of users the ad (1) was delivered to directly, and (2) reached indirectly, by being forwarded, liked, reposted, etc.
- how much the advertiser paid for that ad to circulate

* every ad should include a link back to the advertiser's profile page, where this information is available, both as a publicly readable listing for that specific ad and a machine-readable database of every ad transaction by that advertiser

* paid ads should also be held in a separate, public archive that is open and searchable, in ways suited to both qualitative and data scientific research techniques

If this sounds like a great deal more than what U.S. law asks of advertisers and publishers in broadcasting or print, it is. But there are (at least) four good reasons why we should demand that platforms provide this information. First, they have the data, or can get it, and it would not be a high burden to provide it. Second, such obligations could help to address the kind of divisive political advertising that became so visible in 2016. Third, social media platforms gain an immense economic value based on the data contributed by their users: their posts, their activity, their attention, and the digital traces they leave. This value is realized in the sophisticated and precise advertising they can offer. Platforms owe their users something for all that extracted value, more than just a smooth interface and some entertaining videos. They owe users the same commitment to transparency, real identity, and accountability that they expect of us. And fourth, this is an opportunity to demand not just the same commitment we asked of broadcasters, but more: that advertisers stand by their ads and how they target them, and make them available for scrutiny by journalists, activists, and regulators. For advertisers making fair and reasonable appeals to consumers or voters, this should be no burden at all.

Distribute the Agency of Moderation, not just the Work

Platforms not only distribute content, they also distribute responsibility for that content. Far from washing their hands of the task of moderation, most platforms have made it a central part of what they do. But to do so, they have distributed the work of and the responsibility for that moderation: across different parts of the company, across fluctuating teams of independent contractors, to users, and to third-party advisory organizations. So, if we think of platforms as distributors of responsibility, we should then ask whether that responsibility is being distributed responsibly, or how it might be distributed differently.

When social media platforms task users with the work of moderation, overwhelmingly it is as individuals. Flagging is individual, rating content is individual, muting and blocking are by an individual and of an individual. This is not true in all cases. Some collaborative projects include collective forms of moderation, like the talk pages on Wikipedia, or Reddit groups moderated by hundreds of volunteers. And a few sites have experimented with more democratic efforts at moderation: LiveJournal involved users in setting site policies, and the game League of Legends adjudicated some of its cases in what it called The Tribunal, enlisting players to serve as an ad hoc jury. But beyond these few exceptions, there is little support for users to moderate collectively, or even to have their individual efforts accumulate into something of collective value, not just for the platform but for other users.

Platforms should also let flagging accumulate into actionable data for users. Social media platforms don't generally want flags to trigger consequences automatically, since the data is too noisy: some flaggers misunderstand the content or the rules, and some flag content they don't agree with in order to get it removed. But the data is not so noisy that it can't be returned to users as a signal. Heavily flagged content, especially if those flags are coming from unconnected users, could be labeled as such, or put behind a clickthrough warning, even before it is reviewed.

There are other ways to accumulate the work users already do into a collective resource valuable to all. The simplest is shared blocklists.[19] If I have identified a dozen harassers, and you trust me and are experiencing similar problems, you might just want to block my dirty dozen from the start. This makes intuitive sense, not only because it takes advantage of many users' incremental efforts (the fundamental premise of crowdsourcing) but also because harassment is often itself a social phenomenon: communities of users, often in conversation with one another, find themselves harassed in similar ways, and brigades of trolls often seek out and torment users together.

Platforms have been slow to embrace even the simplest version of blocklists. Inexplicably, even as it faced intense public criticism for harassment, Twitter actively resisted shared blocklists for a long time; it finally made them accessible in 2015, but Twitter's blocklists are severely limited in how they can be shared or updated, compared with third-party alternatives already available but unsupported by the platforms.[13] Easily shared blocklists would aggregate the work of blocking and quickly propagate it across users, protecting those who want protection and taking some of the thrill out of trolling. Danielle Citron and Ben Wittes also proposed an easy but ingenious addition, that users be able to share lists of those they

[19] Stuart Geiger, 2016. "Bot-Based Collective Blocklists in Twitter: The Counterpublic Moderation of Harassment in a Networked Public Space." *Information, Communication, and Society* 19 (6): 787–803.

follow as well, allowing users to generate lists of bad actors and lists of recommended sources, which other users could then subscribe to.[20]

This approach could be taken much farther, to what I will call *collective lenses*. Flagging a video on YouTube brings down a complex menu for categorizing the offense, to streamline their review process. But what if the site offered a similar tool for tagging videos as sexual, violent, spammy, false, or obscene? These would not be complaints per se, nor would they be taken as requests for their removal (as the flag currently is), though they would help YouTube find the truly reprehensible and illegal. Instead, these tags would produce aggregate data by which users could filter their viewing experience. I could subscribe to an array of these collective lenses: I don't want to see videos that more than x users have categorized as violent.[21] Platforms could make clear, to the poster and to all users, that a particular piece of content is among those being filtered out by a particular lens, and could provide an appeals process to declassify it. They could host a debate page for every lens, where interested users could discuss what should and should not fall into that category.

Better yet, trusted organizations could develop and manage their own collective lenses: imagine a lens run by the Southern Poverty Law Center to avoid content that allied users have marked as "racist," or one from Factcheck.org filtering out "disputed" news articles. Building on Citron and Wittes' suggestion, collective lenses could double as a recommendation system: I could subscribe to a lens run by a prestigious university that brings up content that their allied users had marked as "educational" or "scientific" or "fascinating." This would not help the "filter bubble" problem; it might in fact exacerbate it. Then again, users would be choosing what not to see, rather than having it deleted on their behalf.

With collective lenses and the crowdsourced tags it would generate, platforms would benefit from an immense amount of additional data about their own content, improving their recommendations as well as helping them identify illegal content or troublesome users. Groups of users and independent organizations could partner to help curate the platform landscape, in areas and around topics they are most invested and expert in, at a granularity and precision greater than the platform could by itself. Some forms of harassment would be quickly deflated: while their speech would not be silenced, harassers would know that targets of their abuse were easily and collectively ignoring them.

[20] Danielle Citron and Ben Wittes, "Follow Buddies and Block Buddies: A Simple Proposal to Improve Civility, Control, and Privacy on Twitter," *Lawfareblog*, January 4, 2017, https://lawfareblog.com/follow-buddies-and-block-buddies-simple-proposal-improve-civility-control-and-privacy-twitter.

[21] It is worth noting that, if for some reason the platform wanted to make it possible, users could even seek out content that others had marked "violent."

This approach would be far preferable to having a platform "choreograph" content away, because users would get to choose what they filtered and why; they would have more agency in the moderation of their platforms experience, not less. Collective lenses would prioritize those who want a curated experience over those who take advantage of an uncurated one. This would, of course, not solve the problem of harassment altogether—nothing can, since the problem to be "solved" is how to make a common life with other human beings. But it would grant users more individual and collective agency in the project of moderation, not just solicit their labor.

Platforms would still have to moderate—some things are simply too harmful to keep, and some harms cannot be collectively identified. But they could focus on the most reprehensible content, and enjoy more legitimacy for removing it. Platforms could even opt to be more permissive, especially regarding content that might offend, knowing that lenses are available for avoiding it. And if these lenses were publicly made and publicly sponsored, there would be more opportunity to debate why we moderate, according to what values, and with what consequences—instead of just dinging Facebook or Apple for their specific and opaque interventions.

Protect Users across Platforms

Little of what a user does to curate and defend her experience on one platform can easily be exported to others. Given that, in reality, most users use several platforms, all of these preferences should be portable. If I have flagged a video as offensive on Tumblr, I presumably don't want to see it on YouTube either; if I have marked an advertisement as misleading on Google, I presumably don't want it delivered to me again on Facebook; if I have been harassed by someone on Twitter, I presumably don't want him also harassing me on Snapchat.[22] Given that users are already being asked to rate, flag, and block, and that this labor is almost exclusively for the benefit of the platform, it is reasonable to suggest that users should also enjoy the fruits of that labor, in their interactions on this platform and elsewhere.

This need not require that platforms share user data, and it shouldn't. But it would require developing a way to format content metadata so that users could carry it with them, from platform to platform, in a form that would be recognizable. (As the technology improves, it might not merely be my selections and preferences that could be made portable: platforms could gradually "learn" from my flags and blocks, set up an AI sentry on my behalf—and make that portable too.) If the harms on social media

[22] According to the WAM report, "Reporters were asked if the harassment was occurring on multiple platforms. 54 reports (17%) mention harassment taking place on multiple platforms" (15). Nathan J. Matias, Amy Johnson, Whitney Erin Boesel, Brian Keegan, Jaclyn Friedman, Charlie Detar. 2015. "Reporting, Reviewing, and Responding to Harassment on Twitter." http://womenactionmedia.org/twitter-report/

come from the people and their behavior (as much as or more than the content they share), there would need to be a way to know that the person behind that Twitter handle is the same person behind that Tumblr comment and that Instagram follow request. That is a more involved undertaking, I admit, but it's one I expect we will soon find worth considering.

Social media platforms have been resistant to making users profiles and preferences interoperable. At least publicly, platform managers say doing so would make it too easy for users to decamp to a hot, new service, if their interest in this one cooled. Managers at successful platforms tell ghost stories about the hasty deaths of MySpace, Friendster, and Digg, and regularly predict the death of Twitter and Facebook, to justify any means for retaining their users. But the accumulated history users have with a platform—established social networks, a legacy of interactions, an archive of photos, an accumulated matrix of preferences—does in fact discourage them from abandoning it, even when they are dissatisfied with how it governs their use, even when they are fed up, even when they must endure harassment to stay. This means that making it difficult to export preferences, though it may seem to make economic sense (at least in the short term), keeps some people in unsatisfactory and even abusive conditions. Portability could make all that uncompensated effort into an asset. The fact that this is not yet possible is a cold reminder that what users need from platforms is constrained by what platforms need in the market.[23]

Amend the Section 230 safe harbor

I am sympathetic to the desire to defend some of the safe harbor offered by Section 230. But the typical defense of Section 230 in the face of these compelling concerns tends to adopt an all-or-nothing rhetoric. Some claim that any conditional liability opens the door to proactive moderation and could prove a slippery slope to full liability; some imply that platforms do not moderate already. This rigorous defense of "expressive immunity," as Julie Cohen calls it, requires a "carefully tended hysteria about censorship and injured protestations of First Amendment virtue."[24] And, it is worth noting, this perspective lines up well with the way platforms themselves defend Section 230's protections. Realistically, there is a great deal of room between complete legal immunity offered by a robust Section 230 without exceptions and total liability for platforms as Section 230 crumbles away.[25]

However Section 230 grows to meet today's challenges, we must redress the opportunity that was missed when Section 230 was first drafted. Safe harbor, including the right to moderate in good faith and

[23] Many thanks to Sharif Mowlabocus for this insight.

[24] Julie Cohen, "Law for the Platform Economy," 51 *University of California Davis Law Review* 133 (2017).

[25] In this, I am in agreement with Danielle Citron and Ben Wittes, in their forthcoming essay in the *Georgetown Law and Technology Review*.

the permission not to moderate at all, was an enormous gift to the young Internet industry. Over the history of U.S. regulation of the media and telecommunication industries, gifts of this enormity were always fitted with a matching obligation to serve the public a monopoly granted to a telephone company comes with the obligation to serve all users; a broadcasting license comes with obligations about providing news or weather alerts or educational programming.

The gift of safe harbor could finally be fitted with public obligations – not external standards for what to remove, but parameters for how moderation should be conducted fairly, publicly, and humanely. Such matching obligations might include:

> * transparency obligations – platforms could be required to report data on the process of moderation to the public or to a regulatory agency. Several of the major platforms already voluntarily report takedown requests, but these have typically focused on government requests. Until recently none systematically report data on flagging, policy changes, or removals made on their own accord. Facebook and YouTube began to do so in 2018, and should be encouraged to continue.[26]
> * minimum standards for moderation – without requiring moderation be handled in a particular way, minimum standards for the worst content, minimum response times, or obligatory mechanisms for redress or appeal could assure a base level of responsibility and parity across platforms.
> * shared best practices – a regulatory agency could provide a means for platforms to share best practices in content moderation without running into questions of antitrust rules. Outside experts could be enlisted to develop best practices in consultation with industry representatives.
> * public ombudsman – most major platforms address the public only through their corporate blogs, when announcing major changes in policy or responding to public controversies. But this is on their own initiative and offers little room for public response. Platforms could each be required to have a public ombudsman who responds to public concerns and translates those concerns to policy managers internally, or a single "social media council"[27] could field public complaints and demand accountability from platforms.

[26] Tarleton Gillespie, "Facebook and YouTube Just Got More Transparent. What Do We See?" Neiman Lab, May 3, 2018. http://www.niemanlab.org/2018/05/facebook-and-youtube-just-got-more-transparent-what-do-we-see/

[27] Article 19, "Regulating Social Media: We Need a New Model That Protects Free Expression," April 25, 2018 https://www.article19.org/resources/regulating-social-media-need-new-model-protects-free-expression/

* financial contributions to support organizations and digital literacy programs – major platforms like Twitter have leaned on non-profit organizations to advise and even handle some moderation, as well as to mitigate the socio-emotional costs of the harms some users encounter.[28] Digital literacy programs could expand to better address online harassment, hate speech, and misinformation. Enjoying safe harbor protections of Section 230 might platforms to help fund these non-profit efforts.

* consultations with an expert advisory panel – without assuming regulatory oversight of a government body, a blue ribbon panel of regulators, experts, academics, and activists could be given access to platforms and their data to oversee content moderation, without revealing platforms' inner workings to the public.

* advisory oversight from regulators – a government regulatory agency could consult on and review content moderation procedures at major platforms. By focusing on reviewing *procedures*, such oversight could avoid the appearance of imposing a political viewpoint; review could be sensitized to the more systemic problems of content moderation.

* labor protections for moderators – content moderation at large platforms depends on crowdworkers, either internal to the company or contracted through third party temporary services. Guidelines could ensure these workers basic labor protections like health insurance, assurances against employer exploitation, and resources to address the potential psychological harm that can be involved.

* obligation to share moderation data with qualified researchers – the right to safe harbor could come with an obligation to set up reasonable mechanisms for qualified academics to access platform moderation data, in order to allow for the investigation of questions that platforms might not think to, or want to, answer. The research partnership between Facebook and the Social Science Research Council.[29] Announced in 2018, has yet to work out the details, but some version of this model could be extended to all platforms.

* data portability - Social media platforms have been resistant to making users' profiles and preferences interoperable across platforms. But moderation elements like blocked users and flagged content could be made portable so the work users have put into moderating their information space carries across the platforms they frequent.

[28] Nathan J. Matias, Amy Johnson, Whitney Erin Boesel, Brian Keegan, Jaclyn Friedman, Charlie Detar. 2015. "Reporting, Reviewing, and Responding to Harassment on Twitter." http://womenactionmedia.org/twitter-report/

[29] Social Science Research Council, April 2018. https://www.ssrc.org/programs/view/social-data-initiative/

* build their systems for regular audits – without requiring complete transparency in the moderation process, platforms could build a mechanism for researchers, journalists, and even users to conduct their own audits of the moderation process, to understand better the rules in practice.

* regular legislative review of Section 230 – the Digital Millennium Copyright Act stipulated that the Library of Congress revisit the list of exceptions every three years to account for changing technologies and emergent needs. Section 230, and whatever matching obligations that might be fitted to it, could similarly come up for reexamination to account for the rapidly changing workings of social media platforms and the even more rapidly changing nature of harassment, hate, misinformation, and other harms.

Reject the Economics of Popularity

Many of the harms that have proliferated are hard to moderate because they emulate the shape of the participation that platforms try to encourage. In the dreams of their founders, these platforms were intended to allow everyone to speak their minds, to connect with others around issues that matter to them, to be findable on the network, to present themselves as they choose, and to form bonds through conversation untrammeled by status or location. For the harasser, harassment embodies all of those benefits. Harassment is not a perversion of the social media dream; it is one logical version of it—just not the one designers had in mind.

For platforms, popularity is one of the most fundamental metrics, often serving as proxy to every other: relevance, merit, newsworthiness.[30] Platforms amplify the popular by returning it to users in the form of recommendations, cued-up videos, trends, and feeds. Harassment and hate take advantage of this: cruel insults that classmates will pass around, slurs aimed at women that fellow misogynists will applaud, nonconsensual porn that appeals to prurient interests. These are not just attacks; they generate likes, views, comments, and retweets, making it hard for platforms to discern their toxicity or pass up their popularity.

From a business perspective, at least in a short term, these posts are just as valuable to the platform as other forms of participation.[19] If it's advertising revenue that platforms seek, these are eyeballs to be sold like any other. If it's data, these are traces to be collected like any other. Would Reddit make more money if it cleaned up its act, or has it dragged its feet because all those trolls and men's rights harassers and alt-right blowhards and pornographers bring an awful lot of activity and energy to the

[30] José van Dijck, 2013. *The Culture of Connectivity*, Oxford: Oxford UP: 13.

platform? Don't mess with success, as they say. With business models that use popularity as the core proxy for engagement, too often platforms err on the side of encouraging as many people to stay as possible, imposing rules with the least consequences, keeping troublesome users if they can, and bringing them back quickly if they can't.

Under a different business model, platforms might be more willing to uphold a higher standard of compassionate and just participation, and forgo users who prove unwilling to consent to the new rules of the game. Where is the platform that prioritizes the longer-term goal of encouraging people to stay and helping them thrive, and sells that priority to us for a fee? Where are the platforms that gain value when fewer users produce a richer collaboration? Until those platforms appear and thrive, general-use platforms are unlikely to pursue an affirmative aspiration (what are we here to accomplish?) rather than a negative one (what shouldn't we do while we're here?).

Put Real Diversity behind the Platform

In the end, it will remain difficult for managers of these platforms to radically rethink moderation, because these platforms have been built and are governed by a tiny tribe of people with a specific and limited perspective on the world. Silicon Valley engineers, managers, and entrepreneurs are by and large a privileged lot, who tend to see society as fair and meritocratic; to them, communication just needs to be more open and information more free. They tend to build tools "for all" that sustain, extend, and reify the inequities they overlook. But harassment and hatred are not problems specific to social media; they are endemic to a culture in which the powerful maintain their position over the less powerful through tactics of intimidation, marginalization, and cruelty, all under cover of a nominally open society. Silicon Valley engineers and entrepreneurs are not the community most likely to really get this, in their bones. It turns out that what they are good at is building communication spaces designed as unforgiving economic markets, where it is necessary and even celebrated that users shout each other down to be heard; where some feel entitled to toy with others as an end in itself, rather than accomplishing something together; where the notion of structural inequity is alien, and silencing tactics take cover behind a false faith in meritocracy. Platform management cannot see what the world, or even its platforms, look like from the perspective of someone who has endured structural inequity or blatant hatred.

As a straight, white, and privileged man myself, I am not suggesting that we are incapable of compassion, unwilling to make progressive changes that largely benefit others, or necessarily oblivious to ingenious solutions. But the straight, well-off, white and Asian men of Silicon Valley have proven, convincingly and repeatedly, that they cannot resolve these problems alone. There's a reason why social media seem to make room not only for abuse and hatred per se but for the very same abuse and hatred that plagued society long before social media: against women, against racial and ethnic minorities, against gay, queer, and transgender people.

Slight improvements in workplace diversity aren't going to make the difference. We've seen what corrosive environments some of these companies are for those who do show up.[31] What would happen if social media platforms promised that for the next decade, all of their new hires, 100 percent, would be women, queer people, or people of color? Sounds like an outrageous exercise in affirmative action and social engineering? It sure is. Slight improvements in workplace diversity aren't going to make the difference, we've seen what corrosive environments some of these companies can be for those who do show up. But I suggest this not only for the benefit of the new employees, but for the benefit of the platform and its users. It is not that women and queer people and people of color necessarily know how to solve the problems of harassment, revenge porn, or fake news—or that the job of solving these problems should fall on their shoulders. But to truly diverse teams, the landscape will look different, the problems will surface differently, the goals will sound different. Teams that are truly diverse might be able to better stand for their diverse users, might recognize how platforms are being turned against users in ways that are antithetical to the aims and spirit of the platform, and might have the political nerve to intervene.

* * *

Would these suggestions solve the problem? No. When I began writing this book, the pressing questions about content moderation seemed to be whether Silicon Valley could figure out which photos to delete and which to leave, which users to suspend and which to stop suspending. The years 2014 and 2015 helped reveal just how many people were suffering while platforms tried to figure this out. But 2016 and 2017 fundamentally transformed the nature of the problem. It turns out the issue is much bigger than it first seemed...

[31] Liza Mundy, "Why Is Silicon Valley So Awful to Women?" *The Atlantic*, April 2017. https://www.theatlantic.com/magazine/archive/2017/04/why-is-silicon-valley-so-awful-to-women/517788/