8

what platforms are, and what they should be

> Facebook is not just technology or media, but a community of people. That means we need Community Standards that reflect our collective values for what should and should not be allowed. In the last year, the complexity of the issues we've seen has outstripped our existing processes for governing the community.
>
> *Mark Zuckerberg, chairman and CEO of Facebook,*
> *"Building Global Community," February 2017*

Content moderation is such a complex sociotechnical undertaking that, all things considered, it's amazing that it works at all, and as well as it does. Even so, as a society we have once again handed over to private companies the power to set and enforce the boundaries of appropriate public speech for us. That is enormous cultural power held by a few deeply invested stakeholders, and it is being done behind closed doors, making it difficult for anyone else to inspect or challenge. Platforms frequently, and conspicuously, fail to live up to our expectations; in fact, given the sheer enormity of the undertaking, most platforms' definition of success includes failing users on a regular basis.

And while we sometimes decry the intrusion of platform moderation, at other moments we decry its absence. We are partly to blame for having put platforms in this untenable situation, by asking way too much of them. Users cannot continue to expect platforms to be hands-off *and* expect them to solve problems perfectly *and* expect them to get with the times *and* expect them to be impartial and automatic.

We must recognize that moderation is hard work, that we are asking platforms to intervene, and that they have responded by enlisting us in the labor. What is important, then, is that we understand the ways in which platforms are moderated, by whom, and to what ends. But more than that, the discussion about content moderation needs to shift, away from a focus on the harms users face and the missteps platforms sometimes make in response, to a more expansive examination of the responsibilities of platforms, that moves beyond their legal liability to consider their greater obligations to the public.

IMPROVING MODERATION

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 not only their approach but how platforms conceive of themselves and their users. I fully acknowledge that some are politically untenable and economically outlandish, and are almost certain never to happen. I spell them out in more detail in a separate document, but in brief:

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. But transparency is not merely the absence of opacity. It requires designing new ways to make processual information visible but unobtrusive. If one of my tweets is receiving lots of responses from “egg” accounts—often the ones dedicated to trolling—that I have already blocked, how could this fact, and their number and velocity, still be made visible to me?[2] Tiny eggs, swarming like angry bees below my tweet? A pop-up histogram that indicates the 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. 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.

Distribute the Agency of Moderation, Not Just the Work

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. With few exceptions, there is little support for any of this work to accumulate into something of collective value, not just for the platform but for other users. As Danielle Citron and Ben Wittes have noted, platforms have been slow to embrace even the simplest version of this, shared block lists. They also propose an easy but ingenious addition, that users be able to share lists of those they follow as well.[3]

Platforms should also let flagging accumulate into actionable data for users. Heavily flagged content, especially if by multiple, unconnected users, could be labeled as such, or put behind a clickthrough warning, even before it is reviewed. But this could be taken farther, to what I’ll call *collective lenses.* Flagging a video on YouTube brings down a menu for categorizing the offense, to streamline the 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.[4] 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. 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. It would prioritize those who want a curated experience over those who take advantage of an uncurated one.

Protect Users as They Move 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.[5] 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.

Social media platforms have been resistant to making users profiles and preferences interoperable. At least publicly, platform managers say that doing so would make it too easy for users to decamp to a hot, new service, if their interest in this one cooled. 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 your preferences keeps some people in unsatisfactory and even abusive conditions. 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.[6]

Reject the Economics of Popularity

For platforms, popularity is one of the most fundamental metrics, often serving as proxy to every other: relevance, merit, newsworthiness. 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. 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

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. 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. They tend to build tools "for all" that continue, extend, and reify the inequities they overlook.

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.

THE LONG HANGOVER OF WEB 2.0

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.

In the run-up to the 2016 presidential election in the United States, and in elections that followed on its heels in France, Germany, the Netherlands, Myanmar, and Kenya, it became clear that misinformation was spreading too readily on social media platforms like Facebook and Twitter, and in many cases, its spread was deliberate. The clamor about "fake news" may tend to erase important distinctions between propaganda, overstatement, partisan punditry, conspiracy theories, sensationalism, clickbait, and downright lies. But even so, it has made clear that social media platforms facilitated the circulation of falsehoods, and algorithmically rewarded the most popular and outlandish over more substantive journalism. The politically motivated, including Russian operatives keen on disrupting democratic political processes, and the economically motivated, including Macedonian teens keen on turning a quick profit on the clicks of the gullible, seeded online discussions with false content masquerading as legitimate news and political debate. Though it may be impossible to prove, some worried that this flood of propaganda may have been a factor in the U.S. election, handing the presidency to Donald Trump.[7] Finding the very mechanisms of democracy in peril has dramatically raised the stakes for what platforms allow and what they can prevent.

Evidence that Russian operatives also bought advertisements on Facebook, Twitter, and Google, targeted at users in battleground U.S. states, and designed to fuel racial and economic tensions on both sides of the political spectrum, expanded the issue further. Advertising on Facebook was already moderated; it is not surprising that the platform's response to this revelation was to promise better moderation. But the Russian operatives' use of social media advertising also offers a powerful reminder that, all this time, we may have been thinking about platforms in the wrong way. Facebook is not a content platform supported by advertising. It is two intertwined networks, content and advertising, both open to all. Given that small ad buys are relatively cheap, advertisers are no longer just corporate brands and established institutional actors; they can be anyone. Persuading someone through an ad is as available to almost every user as persuading him through a post. The two networks may work according to different rules: posts go to friends and friends of friends, ads go to those targeted through Facebook's algorithmically generated microdemographics. But the two also bleed into each other, as "liking" an ad will forward it to friends, and ads can be designed to look like posts. So while platforms moderate content that users circulate and content that advertisers place, the problem of policing propaganda—or harassment, hate speech, or revenge porn, for that matter—must now trace persuasive tactics that take advantage of both networks together.

The concerns around political discourse and manipulation on social media platforms feel like the crest of a larger wave that has been breaking for a few years now, a broader reconsideration, on so many fronts, of social media platforms and their power in society: concerns about privacy and surveillance, topped by the Snowden revelations of the links between Silicon Valley companies and the National Security Administration (NSA);

vulnerability to hackers, criminal or political, made plain by high-profile attacks on retailers, credit agencies, and political parties; their impact on the economics of journalism, particularly Facebook's oversized footprint, which changes as often as Facebook's priorities do; concerns about the racial and gender biases baked into their algorithms; research conducted on users without their consent; their growing influence over policy, in places like Washington, D.C., Brussels, and Davos; the inequities in their workplaces, and the precarious labor dynamics they foster as part of the "gig economy"; their impact on San Francisco, on manufacturing zones around the world, and on the environment.

Perhaps we are now experiencing the long hangover after the ebullient high of web 2.0, the birth of social media, and the rise of a global, commercial, advertising-supported Internet culture—the bursting of a cultural bubble, if not a financial one.[8] It's possible that we've simply asked too much, or expected too much, from social media. As Virginia Heffernan quipped, about Twitter, "I think we've asked way, way too much of a little microblogging platform that was meant to talk about where to get a beer in Austin, Texas, and now is moving mountains, and is a centerpiece of geopolitics. It's like asking nodes of Rubik's Cubes to manage world history."[9] Or perhaps these are growing pains. In a 2013 interview, Ken Auletta took the mounting criticism of social media as evidence that "Silicon Valley is in the equivalent of its adolescence. And, in adolescence, is suddenly a time when you become aware of things beyond yourself, become aware the world. . . . So suddenly the people at Google, the people at Twitter, the people at Amazon, awake to the fact that, oh my God, we have to learn how the rest of the world operates and lives and what they expect of us."[10] Or maybe platforms simply are vulnerable to both unpredictable and tactical misuse, because they're designed to be. As Tom Malaby put it, thinking specifically about virtual game worlds, "Like few other products we can identify—early telephone service is one, Internet search engines may be another— . . . [a platform] depends on unanticipated uses by its consumers . . . meant to make itself." Striking a new balance between control and contingency means platforms must assure an open-endedness sufficient to produce "socially legitimate spaces for the unexpected."[11]

The dreams of the open web did not fail, exactly, nor were they empty promises to begin with. Many people put in a great deal of effort, time, resources, and dollars to pursue these dreams, and to build infrastructures to support them. But when you build a system that aspires to make possible a certain kind of social activity, even if envisioned in the most positive terms, you also make possible its inverse—activity that adopts the same shape in order to accomplish the opposite end. In embracing the Internet, the web, and especially social media platforms for public discourse and sociality, we made a Faustian bargain, or a series of them, and we are now facing the sober realities they produced. If we dreamed of free information, we found we also got free advertising. If we dreamed of participation, we also got harassment. If we dreamed of sharing, we also got piracy. If we dreamed of political activism online, we also got clicktivism, political pandering, and tweetstorms. If we dreamed of forming global, decentralized communities of interest, we also got ISIS recruitment. If we dreamed of new forms of public visibility, we also got NSA surveillance. If we dreamed of free content and crowdsourced knowledge, we also got the exploitation of free labor. If we dreamed of easy interconnection between complex technical resources, we also got hacked passwords, data breaches, and cyberwar.

The companies that have profited most from our commitment to platforms did so by selling the promises of participatory culture. As those promises have begun to sour and the reality of their impact on public life has become more obvious and more complicated, these companies are now grappling with how best to be stewards of public culture, a responsibility that was not evident to them at the start. The debates about content moderation over the past half-decade can be read as social media's slow and bumpy maturation, its gathering recognition that it is a powerful infrastructure for knowledge, participation, and public expression. The adjustments that platforms have already made have not sufficiently answered the now relentless scrutiny being paid to them by policy makers, the changing expectations articulated by the press, and the deep ambivalence now felt by users. Social media platforms have, in many ways, reached an untenable point. This does not mean they cannot function—clearly they can—but that the challenges they face are now so deep as to be nearly paradoxical.

Social media platforms have, to a remarkable degree, displaced traditional media, and they continue to enlarge their footprint. They have indeed given all comers the opportunity to speak in public and semipublic ways, and at an unprecedented, global scale. While they are not used by all, and many parts of the world are still excluded by limited resources, infrastructure, the constraints of language, or political censorship, those who do find their way to these platforms find the tools to speak, engage, and persuade. The general-purpose platforms, especially, aspire to host all public and semipublic communication of every form, aim, and import. At the same time, they are nearly all private companies, nearly all commercially funded, nearly all built on the economic imperatives of advertising and the data collection that targeted advertising demands. Their obligations are, like those of traditional commercial media, pulled between users, content producers, and advertisers—with only the one twist: that users and producers are one and the same.

These platforms now function at a scale and under a set of expectations that increasingly demands automation. Yet the kinds of decisions that platforms must make, especially in content moderation, are precisely the kinds of decisions that should not be automated, and perhaps cannot be. They are judgments of value, meaning, importance, and offense. They depend both on a human revulsion to the horrific and a human sensitivity to contested cultural values. There is, in many cases, no right answer for whether to allow or disallow, except in relation to specific individuals, communities, or nations that have debated and regulated standards of propriety and legality. And even then, the edges of what is considered appropriate are constantly recontested, and the values they represent are always shifting.

WHAT IT TAKES TO BE A CUSTODIAN

We desperately need a thorough, public discussion about the social responsibility of platforms. This conversation has begun, but too often it revolves around specific controversies, making it hard to ask the broader question: what would it mean for social media platforms to take on some responsibility for their role in organizing, curating, and profiting from the activity of their users? For more than a decade, social media platforms have presented themselves as mere conduits, obscuring and disavowing the content moderation they do. Their instinct has been to dodge, dissemble, or deny every time it becomes clear that they produce specific kinds of public discourse in specific ways. While we cannot hold platforms responsible for the fact that some people want to post pornography, or mislead, or be hateful to others, we are now painfully aware of the ways in which platforms can invite, facilitate, amplify, and exacerbate those tendencies: weaponized and coordinated harassment; misrepresentation and propaganda buoyed by its quantified popularity; polarization as a side effect of algorithmic personalization; bots speaking as humans, humans speaking as bots; public participation emphatically figured as individual self-promotion; the tactical gaming of algorithms in order to simulate genuine cultural value. In all of these ways, and others, platforms invoke and amplify particular forms of discourse, and they moderate away others, all under the guise of being impartial conduits of open participation. As such, platforms constitute a fundamentally new information configuration, materially, institutionally, financially, and socially. While they echo and extend traditional forms of communication and exchange, they do so by being, like computers themselves, “universal machines” for many different kinds of information exchange.

Our thinking about platforms must change. It is not just, as I hope I have shown, that all platforms moderate, or that all platforms have to moderate, or that most tend to disavow it while doing so. It is that moderation, far from being occasional or ancillary, is in fact an essential, constant, and definitional part of what platforms do. I mean this literally: moderation is the essence of platforms, it is the commodity they offer. By this point in the book, this should be plain. First, moderation is a surprisingly large part of what they do, in a practical, day-to-day sense, and in terms of the time, resources, and number of employees they devote to it. Moreover, moderation shapes how platforms conceive of their users—and not just the ones who break rules or seek help. By shifting some of the labor of moderation, through flagging, platforms deputize users as amateur editors and police. From that moment, platform managers must in part think of, address, and manage users as such. This adds another layer to how users are conceived of, along with seeing them as customers, producers, free labor, and commodity. And it would not be this way if moderation were handled differently.

But in an even more fundamental way, content moderation is precisely what platforms offer. Anyone could make a website on which any user could post anything he pleased, without rules or guidelines. Such a website would, in all likelihood, quickly become a cesspool and then be discarded. But it would not be difficult, nor would it require skill or financial backing. To produce and sustain an appealing platform requires, among other things, moderation of some form. Moderation is hiding inside every promise social media platforms make to their users, from the earliest invitations to “join a thriving community” or “broadcast yourself,” to Mark Zuckerberg’s 2017 manifesto quoted at the start of this chapter. Every platform promises to offer something in contrast to something else—and as such, every platform promises moderation.[12]

Content moderation is a key part of what social media platforms do that is different, that distinguishes them from the open web: they moderate (removal, filtering, suspension), they recommend (news feeds, trending lists, personalized suggestions), and they curate (featured content, front-page offerings). Platforms use these three levers together to, actively and dynamically, *tune* the unexpected participation of users, produce the “right” feed for each user, the “right” social exchanges, the “right” kind of community. (“Right” here may mean ethical, legal, and healthy; but it also means whatever will promote engagement, increase ad revenue, and facilitate data collection. And given the immense pushback from users, legislators, and the press, these platforms appear to be deeply out of tune.

If content moderation is the commodity, if it is the essence of what platforms do, then it no longer makes sense to treat it as a bandage to be applied or a mess to be swept up. Too often, social media platforms treat content moderation as a problem to be solved, and solved privately and reactively. Platform managers understand their responsibility primarily as protecting users from the offense and harm they are experiencing. But now platforms find they must answer also to users who find themselves troubled by and implicated in a system that facilitates the reprehensible—even if they never see it. Removing content is no longer enough: the offense and harm in question is not just to individuals, but to the public itself, and to the institutions on which it depends. This is, according to John Dewey, the very nature of a public: “The public consists of all those who are affected by the indirect consequences of transactions to such an extent that it is deemed necessary to have those consequences systematically cared for.”[13] What makes something of concern to the public is the potential need for its inhibition.

Despite the safe harbor provided by the law and the indemnity enshrined in their terms of service contracts as private actors, social media platforms inhabit a position of responsibility—not only to individual users but to the public they powerfully affect. When an intermediary grows this large, this entwined with the institutions of public discourse, this crucial, it has an *implicit contract* with the public that, whether platform management likes it or not, can differ from the contract it required users to click through. The primary and secondary effects these platforms have on essential aspects of public life, as they become apparent, now lie at their doorstep.

Fake news is a useful example. Facebook and Twitter never promised to deliver only reliable information, nor are they legally obligated to spot and remove fraud. But the implicit contract is now such that they are held accountable for some of the harms of fake news, and must find ways to intervene. This is not a contract that will ever bind the platforms in court, but it is certain to be upheld in the court of public opinion. Even as moderation grows more complicated and costly, the expectations of users have grown not more forgiving but more demanding. That is the collective enforcement of the implicit contract, and right now it is pushing platforms away from the safe harbors they have enjoyed.[14]

Rethinking content moderation might begin with this recognition, that content moderation is the essential offer platforms make, and part of how they tune the public discourse they purport to host. Platforms could be held responsible, at least partially, for how they tend to that public discourse, and to what ends. The easy version of such an obligation would be to require platforms to moderate more, or more quickly, or more aggressively, or more thoughtfully. We have already begun to see public and legal calls for such changes. But I am suggesting something else: that their shared responsibility for the public requires that they share that responsibility *with* the public—not just the labor, but the judgment.

To their credit, the major social media platforms have been startlingly innovative in how they present, organize, recommend, and facilitate the participation of users. But that innovation has focused almost exclusively on how users participate at the level of content: how to say more, see more, find more, like more, friend more. Little innovation, by comparison, has supported users' participation at the level of governance, shared decision making, collaborative design, or the orchestration of collective values. The promise of the web was not only that everyone could speak but that everyone would have the tools to form new communities on their own terms, design new models of democratic collectivity. This second half of the promise has been largely discarded by social media platforms.

In 2012, Facebook held a vote. Starting in 2009, Facebook users could actually "veto" a policy, but it required 30 percent of users to participate to make it binding. The 2012 vote received just 0.038 percent participation.[15] Facebook went ahead and amended the policy, even though a clear majority of those who did vote were against it. And it got rid of the veto policy itself.[16] The vote was, in the eyes of the press and many users, considered a failure. But instead of considering it a failure, what if it had been treated as a clumsy first step? Hundreds of thousands of users voted on an obscure data policy, after all. What if the real failure was that Facebook was discouraged from trying again? The idea of voting on site policies could have been improved, with an eye toward expanding participation, earning the necessary legitimacy, developing more sophisticated forms of voting, and making a more open process. And mechanisms of collective governance could mean much more than voting. Platforms should be developing structures for soliciting the opinions and judgment of users in the governance of groups, in the design of new features, and in the moderation of content.[17]

Facebook has followed a well-worn path, from involving its users as skilled participants with agency to treating them as customers who prioritize ease and efficiency. Much the same happened with commercial radio and the telephone, even electricity itself.[18] It would take a miracle to imagine social media platforms voluntarily reversing course. But I am not imagining some overbuilt exercise in deliberative democracy, nor do I mean to make every user accept or reject a bunch of gruesome flagged images every time she logs on. But given how effective commercial platforms have become at gleaning from users their preferences, just to more effectively advertise to them, I can only imagine what would be possible if that same innovative engineering went to gleaning from users their civic commitments—not what they like as consumers, but what they value as citizens.

It seems reasonable to think that, given everything users already do on these platforms, the data traces they leave should already make these civic values legible. Social media platforms are not just structures filled with content. Each contribution is also a tiny value assertion from each user: this is what we should be able to say, out loud, in this place. These claims attesting to what should be acceptable are implicit in the very act of posting. These are not grand claims, typically. When someone tweets what he had for breakfast, that is a tiny claim for what should be acceptable, what platforms should be for: the mundane, the intimate, the quotidian. When a critic of Twitter moans, on Twitter, that all people do is tweet what they had for breakfast, that is a claim as well: Twitter should be of more significance, should amount to something.

Some claims, from the beginning, require consideration: does "the way we do things" include this? Sometimes we debate the question explicitly. Should Twitter be more social, or more journalistic? What does it make possible, and where does it fail? But more often, these considerations accumulate slowly, over time, in the soup of a billion assertions—each one tiny, but together, legion. The accumulated claims and responses slowly form the platform.

Content moderation is fundamental to this collective assertion of value. As it currently functions, it is where, in response to some claims of what should be acceptable, platforms will sometimes refuse. Every post is a test claim for what platforms should include, and from the start, some receive the answer “no.” No, this platform is not for porn. No, you can’t use this platform to threaten people. No, you mustn’t mislead people for a quick buck. Because every post is a “yes” claim that something does belong, it shows in relief exactly when platform managers must, or feel they must, act counter to the wants of their users.

But what if social media platforms, instead of policing content, could glean the assertions of civic value that they represent? What if they could display that back to users in innovative ways? Given the immense amount of data they collect, platforms could use that data to make more visible the lines of contestation in public discourse and offer spaces in which they can be debated, informed by the everyday traces of billions of users and the value systems they imply. Could their AI research efforts currently under way, to improve machine-learning techniques to automate the identification and removal of pornography, instead identify what we think of pornography and where it belongs? From our activity across platforms, artificial intelligence techniques could identify clusters of civic commitments—not to then impose one value system on everyone, as they do now, but to make visible to users the range of commitments, where they overlap and collide, and to help users anticipate how their contributions fit amid the varied expectations of their audience.

This would be a very different understanding of the role of “custodian”—not where platforms quietly clean up our mess, but where they take up guardianship of the unresolvable tensions of public discourse, hand back with care the agency for addressing those tensions to users, and responsibly support that process with the necessary tools, data, and insights. I do not pretend to know how to do this. But I am struck by the absence of any efforts to do so on behalf of major social media platforms.

Platforms can no longer duck the responsibility of being custodians to the massive, heterogeneous, and contested public realm they have brought into being. But neither can we duck this responsibility. As Roger Silverstone noted, “The media are too important to be left to the media.”[19] But then, to what authority can we even turn? The biggest platforms are more vast, dispersed, and technologically sophisticated than the institutions that could possibly regulate them. Who has sufficient authority to compel Facebook to be a good Facebook?

As users, we demand that they moderate, and that they not moderate too much. But as citizens, perhaps we must begin to be that authority, be the custodians of the custodians. Participation comes with its own form of responsibility. We must demand that platforms take on an expanded sense of responsibility, and that they share the tools to govern collectively.

So far, we have largely failed to accept this responsibility, too easily convinced, perhaps, that the structure of the digital network would somehow manufacture consensus for us. When users threaten and harass, when they game the system, when they log on just to break the fragile accomplishments of others for kicks, this hardly demonstrates a custodial concern for what participation is, and what it should be. But simply crying foul when you don't agree with someone, or when you don't share his normative sense of propriety, or you don't like a platform's attempt to impose some rules isn't a custodial response either. And in the current arrangement, platforms in fact urge us to go no farther: "If you don't like it, flag it, and we'll handle it from there."

If platforms circulate information publicly, bring people into closer contact, and grant some the visibility they could not have otherwise—then with that comes sex and violence, deception and manipulation, cruelty and hate. Questions about the responsibility of platforms are really just part of long-standing debates about the content and character of public discourse. It is not surprising that our dilemmas about terrorism and Islamic fundamentalism, about gay sexuality, about misogyny and violence against women, each so contentious over the past decade, should erupt here too. Just as it was not surprising that the *Terror of War* photograph was such a lightning rod when it first appeared in U.S. newspapers, in the midst of a heated debate about the morality of the Vietnam War. The hard cases that platforms grapple with become a barometer of our society's pressing concerns about public discourse itself: Which representations of sexuality are empowering and which are obscene, and according to whose judgment? What is newsworthy and what is gruesome, and who draws the line? Do words cause harm and exclude people from discussion, or must those who take part in public debate endure even caustic contributions? Can a plurality of people reach consensus, or is any consensus always an artifact of the powerful? How do we balance freedom to participate with the values of the community, with the safety of individuals, with the aspirations of art, and with the wants of commerce?

The truth is, we wish platforms could moderate away the offensive and the cruel. We wish they could answer these hard questions for us and let us get on with the fun of sharing jokes, talking politics, and keeping up with those we care about. But these are the fundamental and, perhaps, unresolvable tensions of social and public life. Platforms, along with users, should take on this greater responsibility. But it is a responsibility that requires attending to these unresolvable tensions, acknowledging and staying with them—not just trying to sweep them away.

This page intentionally left blank