1

all platforms moderate

> In the ideal world, I think that our job in terms of a moderating function would be really to be able to just turn the lights on and off and sweep the floors . . . but there are always the edge cases, that are gray.
>
> —*personal interview, member of content policy team, YouTube*

Titled *The Terror of War* but more commonly known as "Napalm Girl," the 1972 Pulitzer Prize–winning photo by Associated Press photographer Nick Ut is perhaps the most indelible depiction of the horrors of the Vietnam War. You've seen it. Several children run down a barren street fleeing a napalm attack, their faces in agony, followed in the distance by Vietnamese soldiers. The most prominent among them, Kim Phuc, naked, suffers from napalm burns over her back, neck, and arm. The photo's status as an iconic image of war is why Norwegian journalist Tom Egeland included it in a September 2016 article reflecting on photos that changed the history of warfare. And it was undoubtedly some combination of that graphic suffering and the underage nudity that led Facebook moderators to delete Egeland's post.

After reposting the image and criticizing Facebook's decision, Egeland was suspended twice, first for twenty-four hours, then for three additional days.[1] Norway's daily newspaper *Aftenposten* then reported on his suspensions and included the photo; Facebook moderators subsequently instructed the newspaper to remove or pixelate the photo, then went ahead and deleted it anyway.[2] The editor in chief of *Aftenposten* took to the newspaper's front page to express his outrage at Facebook's decision, again publishing the photo along with a statement directed at Facebook CEO Mark Zuckerberg.

Front page of the Norwegian newspaper *Aftenposten,* September 8, 2016, including the "Terror of War" photograph (by Nick Ut / Associated Press) and editor in chief Espen Egil Hansen's open letter to Mark Zuckerberg, critical of Facebook's removal of Ut's photo. Newspaper used with permission from *Aftenposten;* photo used with permission from Associated Press.

In it he criticized both the decision and Facebook's undue influence on news, calling Facebook "the world's most powerful editor."[3] Many Norwegian readers, even the prime minister of Norway herself, reposted the photo to Facebook, only to have it quickly removed.[4]

More than a week after the image was first removed, after a great deal of global news coverage critical of the decision, Facebook reinstated the photo. Responding to the controversy, Facebook Vice President Justin Osofsky explained:

> These decisions aren't easy. In many cases, there's no clear line between an image of nudity or violence that carries global and historic significance and one that doesn't. Some images may be offensive in one part of the world and acceptable in another, and even with a clear standard, it's hard to screen millions of posts on a case-by-case basis every week. Still, we can do better. In this case, we tried to strike a difficult balance between enabling expression and protecting our community and ended up making a mistake. But one of the most important things about Facebook is our ability to listen to our community and evolve, and I appreciate everyone who has helped us make things right. We'll keep working to make Facebook an open platform for all ideas.[5]

It is easy to argue, and many did, that Facebook made the wrong decision. Not only is Ut's photo of great historical and emotional import, but it also has been "vetted" by Western culture for decades. And Facebook certainly could have handled the removals differently. At the same time, what a hard call to make! This is an immensely challenging image: a vital document of history, so troubling an indictment of humanity that many feel it must be seen—and a graphic and profoundly upsetting image of a fully naked child screaming in pain. Cultural and legal prohibitions against underage nudity are firm across nearly all societies, with little room for debate. And the suffering of these children is palpable and gruesome. It is important precisely because of how Kim Phuc's pain, and her nakedness, make plain the horror of chemical warfare. Its power is its violation: "the photo violates one set of norms in order to activate another; propriety is set aside for a moral purpose. It is a picture that shouldn't be shown of an event that shouldn't have happened."[6] There is no question that this image is obscenity. The question is whether it is the kind of obscenity of representation that should be kept from view, no matter how relevant, or the kind of obscenity of history that must be shown, no matter how devastating.

Much of the press coverage treated Facebook's decision to remove the image as a thoughtless one, even an automatic one. But Egeland's post was almost certainly removed by a paid, human reviewer, though it may have been flagged by automatic software or by a user.[7] Nor was it an error: in fact, Facebook had a specific policy on this specific image, which it had encountered before, many times. It was later reported by Reuters that the famous photo "had previously been used in training sessions as an example of a post that should be removed. . . . Trainers told content-monitoring staffers that the photo violated Facebook policy, despite its historical significance, because it depicted a naked child, in distress, photographed without her consent."[8] Ut's photo is a test image, literally and figuratively, a proxy with which an industry and a society draws parameters of what is acceptable.[9]

It is important to remember, however, that traditional media outlets also debated whether to publish this image, long before Facebook. In 1972 the Associated Press struggled with whether even to release it. As Barbie Zelizer tells it, "Ut took the film back to his bureau, where he and another photographer selected eight prints to be sent over the wires, among them the shot of the napalmed children. The photo at first met internal resistance at the AP, where one editor rejected it because of the girl's frontal nudity. A subsequent argument ensued in the bureau, at which point photo department head Horst Faas argued by telex with the New York office that an exception needed to be made; the two offices agreed to a compromise display by which there would be no close-up of the girl alone. Titled 'Accidental Napalm Attack,' the image went over the wires."[10] The first version of the photo the Associated Press released was lightly airbrushed to minimize the hint of Kim Phuc's pubic hair—though the untouched photo was also made available, and was what most newspapers ran the next day.[11] The *New York Times* was the first to publish the photo, and it too had an internal debate as to whether it could do so.[12] Though many U.S. and European newspapers published the photo, many did so after much debate, and some did not. And some readers were offended by the photo, enough to send their complaints to the newspapers: "Readers' letters labeled the display 'nauseating,' 'obscene,' and in 'poor taste,' on the one hand, and urged the photo's widespread display so as to end the war, on the other."[13]

Since the moment it was taken, this photo has been an especially hard case for Western print media—and it continues to be so for social media.[14] It was always both a document of war and a troubling object of concern itself: "Kim's suffering was captured and published movingly in a still photograph—and the still is what became the iconic image—but the photograph

also immediately became a *story*."[15] At the time, commentators like Susan Sontag wrote extensively about it; U.S. President Richard Nixon mused on his secret White House recordings whether it had been faked;[16] many others have acknowledged its troubling power ever since—in articles like Egeland's.[17]

PLATFORMS MUST MODERATE, WHILE ALSO DISAVOWING IT

Social media platforms arose out of the exquisite chaos of the web. Many were designed by people who were inspired by (or at least hoping to profit from) the freedom the web promised, to host and extend all that participation, expression, and social connection.[18] But as these platforms grew, that chaos and contention quickly found its way back onto them, and for obvious reasons: if I want to say something, whether it is inspiring or reprehensible, I want to say it where others will hear me.[19] Social media platforms put more people in direct contact with one another, afford them new opportunities to speak and interact with a wider range of people, and organize them into networked publics.[20] Though the benefits of this may be obvious, and even seem utopian at times, the perils are also painfully apparent, more so every day: the pornographic, the obscene, the violent, the illegal, the abusive, and the hateful.

The fantasy of a truly "open" platform is powerful, resonating with deep, utopian notions of community and democracy—but it is just that, a fantasy.[21] There is no platform that does not impose rules, to some degree. Not to do so would simply be untenable.[22]

Platforms must, in some form or another, moderate: both to protect one user from another, or one group from its antagonists, and to remove the offensive, vile, or illegal—as well as to present their best face to new users, to their advertisers and partners, and to the public at large.

This project, content moderation, is one that the operators of these platforms take on reluctantly. Most would prefer if either the community could police itself or, even better, users never posted objectionable content in the first place. But whether they want to or not, platforms find that they must serve as setters of norms, interpreters of laws, arbiters of taste, adjudicators of disputes, and enforcers of whatever rules they choose to establish. Having in many ways taken custody of the web, they now find themselves its custodians.

The challenge for platforms, then, is exactly when, how, and why to intervene. Where they draw the line between the acceptable and the prohibited rehearses centuries-old debates about the proper boundaries of public

expression, while also introducing new ones. The rules imposed by social media platforms today respond to contemporary fears (for example, sexual predation, terrorism), and they revisit traditional concerns around media and public culture (sex, obscenity, graphic violence).[23] They also revive a perennial challenge, particularly for Western media: do private information providers, having achieved a place of prominence, have an obligation to shape and sometimes restrict content? Do such obligations accompany, or run counter to, the legal promise of free speech?[24]

Moreover, the particular ways in which these platforms enforce their policies have their own consequences. Regardless of the particular rule, it matters whether the enforcement comes in the form of a warning or a removal, whether action comes before or only after someone complains, whether the platform segregates the offending content behind an age barrier or removes it completely. And, however it is enforced, moderation requires a great deal of labor and resources: complaints must be fielded, questionable content or behavior must be judged, consequences must be imposed, and appeals must be considered. For most platforms, this is now a significant portion of what they do.

The very fact of moderation shapes social media platforms as tools, as institutions, and as cultural phenomena. Across the prominent social media platforms, these rules and procedures have coalesced into functioning technical and institutional systems—sometimes fading into the background, sometimes becoming a vexing point of contention between users and platform. Users, whether they sense it or not, are swarming within, around, and sometimes against the parameters that platforms set.

The ways that platforms moderate today are slowly settling in as the familiar and accepted ways to handle user-generated content, mundane features of the digital culture landscape. Approaches battle-tested over time by many platforms are shared among them as "best practices." They are picked up by new sites looking for "what works." To the extent that regulators see such industry "self-regulation" as effective, they tend to then craft policy to complement it or give it legal teeth.

As more and more of our public discourse, cultural production, and social interactions move online, and this handful of massive, privately owned digital intermediaries continues to grow in economic and cultural power, it is crucial that we examine the choices moderators make.

Moderation is hard to examine, because it is easy to overlook—and that is intentional. Social media platforms are vocal about how much content they

make available, but quiet about how much they remove. Content-sharing platforms typically present themselves as cornucopias: thousands of apps, millions of videos, billions of search results, more than you could ever consume. With so much available, it can start to seem as if nothing is unavailable. These sites also emphasize that they are merely hosting all this content, while playing down the ways in which they intervene—not only how they moderate, delete, and suspend, but how they sort content in particular ways, algorithmically highlight some posts over others, and grant their financial partners privileged real estate on the site.

This requires regularly disavowing all the ways in which platforms are much more than mere conduits: in 2016, in the face of mounting criticism, Mark Zuckerberg made a pointed statement that Facebook was not a "media company."[25] Phil Napoli and Robyn Caplan convincingly argue that this claim is both inaccurate and strategic: Zuckerberg and his colleagues do not want to be saddled with the social and legal obligations that apply to media companies.[26] Platforms offer to merely host: positioned front and center are your friends and those you follow, and all the content they share. The platform would like to fall away, become invisible beneath the rewarding social contact, the exciting content, the palpable sense of community.

When they acknowledge moderation at all, platforms generally frame themselves as open, impartial, and noninterventionist—in part because their founders fundamentally believe them to be so, and in part to avoid obligation or liability.[27] Twitter, for example, begins its posted community guidelines with: "We believe that everyone should have the power to create and share ideas and information instantly, without barriers. In order to protect the experience and safety of people who use Twitter, there are some limitations on the type of content and behavior that we allow."[28] These companies prefer to emphasize their wide-open fields of content, and then their impartial handling of it.[29]

It's also not surprising that so few users are aware of how platforms moderate, given that few users ever encounter these rules, or feel the force of their imposition. For many, using these sites as intended, there is little reason to bump up against these restrictions. It is easy to imagine these platforms as open and unregulated, if there appears to be no evidence to the contrary. Since users tend to engage with those like them and use the platform in similar ways, the lack of any sign of rules or their enforcement can be self-confirming.[30] Even some of those suffering harassment, or regularly offended by the content they're encountering, are unaware that the platforms have rules against it or remedies on offer.

On the other hand, many users—more and more every day—are all too aware of how social media platforms moderate. Believing that these platforms are wide open to all users, and that all users experience them that way, reveals some subtle cultural privilege at work. For more and more users, recurring abuse has led them to look to the platforms for some remedy. Others know the rules because they're determined to break them.[31] And others know about platform moderation because they are regularly and unfairly subjected to it. Social media platforms may present themselves as universal services suited to everyone, but when rules of propriety are crafted by small teams of people that share a particular worldview, they aren't always well suited to those with different experiences, cultures, or value systems. Put another way, I am not a pornographer or a terrorist, but I am also not a whistleblower, a drag queen, a Muslim, a lactation specialist, a sex educator, or a black antiviolence activist. So while I may experience social media platforms as wide open, international human rights activists don't; they experience them as censored, unreliable, and inhospitable to their efforts.[32] While I have never had a post deleted or my account suspended, other users with just as much legitimacy to participate as I regularly run up against the rules the platform imposes. Moderation is meant to disappear, but it does so for some more than others.

In the press, there have been growing attention to and debate about how and why platforms moderate. In the earliest days of social media, there was sporadic coverage of the moderation decisions of platforms. Most was little more than "gotcha journalism," typically criticizing a platform for a specific decision that seemed either boneheaded or hypocritical. But in recent years the press has raised deeper concerns: about the implications of platforms intervening too much, the rampant harms for which some platforms do too little, or the punishing labor that this moderation requires. In 2010, Apple was roundly criticized for removing more than five thousand apps from its App Store, because of their sexual nature. The technology press raised a collective eyebrow when Steve Jobs said the then new iPad should offer its users "freedom from porn," but the issue did help show moderation to be systemic and values-laden.[33] In 2012, for the first time but not the last, some of Facebook's moderator training documents were leaked, giving a rare, unvarnished glimpse of what Facebook does and does not want taken down.[34] Two years later the national press finally took notice of the pervasive misogyny online in the wake of #Gamergate, a dispute in the gaming community that blew up into a venomous campaign of harassment and threats

targeting women in gaming, feminist critics, and just about anyone who came to their defense.[35] When private nude photos of celebrities stolen by hackers began circulating on the news aggregation platform Reddit, the site was lambasted for allowing the groups responsible to persist.[36] Late-night talk show host Jimmy Kimmel enjoyed viral success with "Mean Tweets," a recurring feature in which celebrities read aloud to the camera hateful tweets they had received. Journalists began to examine the hidden labor behind content moderation, most notably a 2014 *Wired* report by Adrian Chen documenting the experiences of Filipino workers who scrubbed U.S. social media platforms for dollars a day.[37] Cover stories about trolling and harassment moved from the online technology press to the major newspapers, weekly magazines, and national radio programs and podcasts, especially when celebrities like Ashley Judd, Zelda Williams, Leslie Jones, and Megyn Kelly were targeted.[38] Many were troubled when terrorist organization ISIS circulated gruesome videos of civilians and journalists being beheaded, and some called on YouTube, Twitter, and Facebook to remove them.[39] And in the run-up to and the aftermath of the 2016 U.S. presidential election, many drew attention to the efforts of the "alt-right" to shut down outspoken commentators and journalists through coordinated tactics of online harassment, and to the surge of "fake news," deliberately false news stories meant to mislead voters and/or make a tidy profit from the clicks of curious readers.[40]

In 2016 *Wired,* long a source of unbridled optimism about all things digital, published an open letter to the Internet, decrying not only harassment but the failure of platforms to handle this whole array of problems: "Things aren't great, Internet. Actually, scratch that: they're awful. You were supposed to be the blossoming of a million voices. We were all going to democratize access to information together. But some of your users have taken that freedom as a license to victimize others. This is not fine."[41]

MODERATION IS HARD

But before beginning to challenge platforms for their moderation policies and their responsibility for the Internet's many troubles, let's start with a simple reminder. Content moderation is hard. This should be obvious, but it is easily forgotten. Moderation is hard because it is resource intensive and relentless; because it requires making difficult and often untenable distinctions; because it is wholly unclear what the standards should be; and because one failure can incur enough public outrage to overshadow a million quiet successes.

It would be too simple to say that platforms are oblivious to the problems or too self-interested to do enough about them. Moderators do in fact want to exclude the worst atrocities and champion some basic rules of decency, while allowing everything else to flow undisrupted. Their efforts may be driven by a genuine desire to foster a hospitable community, or by a purely economic imperative not to lose users driven away by explicit content or relentless abuse, or by a fear of legal intervention if they are unable to protect their users themselves. But in speaking with representatives of the content policy teams at some of the major platforms, I found them genuinely committed to their work, and well aware of the difficulty of the task they have taken on. In some cases, it is even their job to press the seriousness of these issues back onto their own engineers, who often fail to imagine the ways their tools can be misused: as one policy manager from Flickr observed, "There have been so many different times that you think, 'Haven't you guys thought about how people are going to abuse this?' "[42]

Given the true atrocities that regularly appear on social media platforms, the question of *whether* to intervene is, for most, settled. But figuring out where and why to intervene means wading into some thorny questions: not just determining what is unacceptable, but balancing offense and importance; reconciling competing value systems; mediating when people harm one another, intentionally or otherwise; honoring the contours of political discourse and cultural taste; grappling with inequities of gender, sexuality, race, and class; extending ethical obligations across national, cultural, and linguistic boundaries; and doing all that around the hottest hot-button issues of the day.

Another way to think about it is that every well-intentioned rule has equally important exceptions. It is more complicated than simply "How bad is bad?" The blurry edges of bright line rules involve important and long-contested cultural questions: What is the difference between sexually explicit and pornographic? When is an image of the human body artistic, educational, or salacious? Are representations of fictional violence merely entertaining or psychologically harmful? Does discussing a dangerous behavior help those who suffer, or tempt them to act? Or, as in the *Terror of War* photo, does the fact that something is newsworthy supersede the fact that it is also graphic? These questions plague efforts to moderate questionable content, and they hinge not only on different values and ideologies but also on contested theories of psychological impact and competing politics of culture.

Too often the public debate about platform moderation happens at one of two extremes. Those looking to criticize social media platforms for being too permissive point to the most extreme material that can be found there: child pornography, graphic obscenities, rape threats, animal torture, racial and ethnic hatred, self-mutilation, suicide. Those looking to criticize platforms for intervening too much or for the wrong reasons point to arguably legitimate material that was nevertheless removed: the mildest of racy content, material that is frank or explicit but socially valuable, or material simply removed in error.

One of the biggest challenges platforms face is establishing and enforcing a content moderation regime that can address *both* extremes simultaneously. The rules must account for the most egregious atrocities as well as material that is questionable but defensible. Those in charge of content policy are often motivated by, and called to task for, the worst offenses, but must be careful not to ban culturally valuable material in the process. Users troubled by the most offensive content condemn it with the same passion as those who defend the material that rides the very edge of the rule. The reviewers enforcing those rules must maintain sensitive judgment about what does or does not cross a line while also being regularly exposed to—traumatized by—the worst humanity has to offer.

A second question immediately follows: according to whose criteria? We will see in later chapters how the major platforms work this out in practice, but it is a fundamentally difficult, perhaps intractable, problem. Even an online community that is self-governed faces the challenge of who should set the rules that will apply to everyone.[43] Users within that community will have competing values; the challenge only grows as the community does.

For a platform with commercial aims, run by a small team, this tends to turn into a question about either "our values" or the "values of our users." A platform is a product of the company that runs it, so there is a certain logic that it should be the company's values and interests that determine what is acceptable and what should be removed. But these values do not exist in a vacuum. Nearly all social media platforms are commercial enterprises, and must find a way to make a profit, reassure advertisers, and honor an international spectrum of laws. For social media platforms, what ends up standing as "our values" is not some moral core that exists beneath these many competing pressures. It is whatever solution can resolve those pressures—perhaps presented in a language of "the right thing to do," but

already accounting for the competing economic and institutional demands these platforms face.

On the other hand, if these platforms were imagined to be "for" their users, perhaps the values of the users should be preeminent. But how can a content policy team know the "values of our users"? Platform operators have limited ways of knowing their users. Listening to those they hear from most regularly can lead to either attending too much to those who complain, or too easily dismissing them as a noisy minority. It can also lend credence to the assumption that those who do not complain represent a "silent majority" who have no complaints—which may or may not be the case.

In the face of all this uncertainty, or sometimes in total disregard of it, designers and managers often assume their users are "just like us."[44] But platform operators are hardly a cross-section of their user base. Currently, the full-time employees of most social media platforms are overwhelmingly white, overwhelmingly male, overwhelmingly educated, overwhelmingly liberal or libertarian, and overwhelmingly technological in skill and worldview.[45] This can lead these teams to overlook minority perspectives, and only worsens as a user base grows and diversifies. And it explodes when those platforms born in northern California open up to international user communities.[46] As soon as these sites expand beyond the United States, platforms "face a world where the First Amendment is merely a local ordinance."[47] Their distinctly American assumptions about free speech, civil discourse, and healthy community are being subtly (and sometimes knowingly) imposed on an international user base with very different values.

All this means that platforms simply cannot "get it right," in some simple or universal sense. Moderation policies are, at best, reasonable compromises—between users with different values and expectations, as well as between the demands of users and the demands of profit. I am not suggesting that platforms are beyond reproach or that their efforts should not be criticized. They aren't, and they should. Users who feel wronged by the interventions, even when made on their behalf, have every right to challenge a particular decision, policy, or platform. But I'm convinced that most of the challenges are structural, that even the missteps are endemic to how the problem is approached.

The hard questions being asked now, about freedom of expression and virulent misogyny and trolling and breastfeeding and pro-anorexia and terrorism and fake news, are all part of a fundamental reconsideration of social media platforms. To move this reconsideration forward, we need to

examine the moderation apparatus that has been built over the past decade: the policies of content moderation, the sociotechnical mechanisms for its enforcement, the business expectations it must serve, the justifications articulated to support it. We need to look into why this apparatus is straining under the messiness of real uses, note the harms that have been made apparent, and document the growing pressure to address them. And finally, we must ask: If moderation should not be conducted the way it has, what should take its place?

But the reason to study moderation on social media platforms goes beyond preventing harm or improving enforcement. Moderation is a prism for understanding what platforms *are,* and the ways they subtly torque public life. Our understanding of platforms, both specific ones and as a conceptual category, has largely accepted the terms in which they are sold and celebrated by their own managers: open, impartial, connective, progressive, transformative. This view of platforms has limited our ability to ask questions about their impact, even as their impact has grown and/or concern about them has expanded.

In this celebratory vision of platforms, content moderation is treated as peripheral to what they do—a custodial task, like turning the lights on and off and sweeping the floors. It is occasionally championed in response to criticism, but otherwise it is obscured, minimized, and disavowed. I propose turning this understanding of platforms on its head. What if moderation is central to what platforms do, not peripheral? Moderation is an enormous part of the work of running a platform, in terms of people, time, and cost. And the work of policing all this caustic content and abuse haunts what they think their platforms are and what they must accomplish.

And moderation is, in many ways, *the* commodity that platforms offer. Though part of the web, social media platforms promise to rise above it, by offering a better experience of all this information and sociality: curated, organized, archived, and moderated. Consider two details, the first from Julia Angwin's history of the now nearly forgotten social media platform MySpace. Tila Tequila, since disgraced for her association with white supremacists, in 2003 was becoming one of the first online celebrities for the flirty, revealing photos she posted on Friendster. But Friendster had repeatedly deleted her profile for violating their policy, and after each deletion she had to re-create her list of followers. After a fifth deletion, she decided to move to the fledgling MySpace—and brought her forty thousand followers

with her. Traffic spiked the day she arrived. Her choice, between two regimes of moderation, helped buoy the new site in its transition from spyware provider to social media giant, at a moment when its future was far from certain.[48] Second: in 2016, Twitter was in negotiations with the likes of Google, Salesforce, and Disney to be acquired, but all three passed on the deal. Some in the financial press wondered whether Twitter could not be sold, might not even survive, because it had become a toxic environment marred by harassment and misogyny.[49] In one instance, the promise of lenient moderation may have saved that platform for another week, or month, or year; in another, insufficient moderation may have rendered a billion-dollar company toxic to potential buyers.

By understanding moderation not just as an occasional act platforms must engage in but as a fundamental aspect of their service and a fundamental part of their place in public discourse, we can reconsider what platforms are, and ask new questions about their power in society. A focus on moderation slices through the myth that they are neutral conduits, to reveal their inner workings, their animating logics, their economic imperatives, and the actual footprint they leave on the dynamics of sociality and public discourse. It allows us to question their claim of deserving certain legal rights and obligations, and of being free of others. It helps reveal the real and often hidden investments platforms require, including the human, technical, and financial resources necessary, and it helps make sense of their responses to growing criticism from users and the press. It highlights the solutions platform managers prefer in the face of intractable social problems, like how to reconcile competing value systems within the same community, or how to uphold consistent policies in the face of competing societal expectations. And it can help us understand our commitment to platforms, and the ramifications of that cultural shift.

WHAT IS A PLATFORM?

We talk so much about platforms these days, it is easy to forget that they are still surrounded by the "world wide web" of home pages, personal blogs, news sites, oddball discussion spaces, corporate sites, games, porn, 404 error pages, file listings, and forgotten ephemera.[50] Over the course of more than a decade, the kinds of encounters with information and people that were once scattered across the web have been largely gathered up by a small set of companies onto a handful of social media platforms. Today we are, by and large, speaking from platforms. In fact, when these platforms are

compared with their less regulated counterparts, it is to Reddit or 4chan—big platforms compared with smaller ones, mainstream platforms compared with marginal ones—not with the open web, not any more.

The dream of the open web emphasized new, expanded, and untrammeled opportunities for knowledge and sociality. Access to the public would no longer be mediated by the publishers and broadcasters that played such powerful gatekeeper roles in the previous century. The power to speak would be more widely distributed, with more opportunity to respond and deliberate and critique and mock and contribute.[51] This participatory culture, many hoped, would be more egalitarian, more global, more creative, and more inclusive. Communities could be based not on shared kinship or location but on shared interest, and those communities could set their own rules and priorities, by any manner of democratic consensus. The web itself was to be the "platform."[52] It would finally provide an unmediated public sphere, a natural gathering of the wisdom of the crowd, and a limitless cultural landscape.

Soon, new services began offering to facilitate, host, and profit from this participation. This began with the commercial provision of space for hosting web pages, offered first by Internet service providers (ISPs) themselves, and increasingly by web-hosting services like Tripod, Angelfire, and Geocities. Yet these still required knowledge of web design, HTML programming, and file management. The earliest content platforms—MP3.com, SixDegrees, Livejournal, Blogger, Cyworld, Friendster, LinkedIn, MySpace, Delicious, Orkut, Flickr, Dodgeball, YouTube—often began by trying to facilitate one element of being on the web (write without needing to know HTML; keep a list of friends or fans; make content easier to find through directed search).[53] These services were meant to "solve" some of the challenges of navigating the open web. They substantially simplified the tools needed for posting, distributing, sharing, commenting; they linked users to a larger, even global audience; and they did so at an appealing price. They also had acute network effects: if you want to share and participate, you want to do so where there are people to share and participate with.

These platforms were, of course, nearly all for-profit operations. This made them quite interested in not just facilitating but also incorporating the kinds of participation that the web itself made possible.[54] Platform companies developed new ways to keep users navigating and posting, coax them into revealing their preferences and proclivities, and save all of it as personalized data, to sell to advertisers and content partners. Some pushed to become all-in-one services—combining storage, organization, connection,

canvas, delivery, archive; some looked to partner with traditional media and news providers, to draw the circulation of their content onto the platforms. Some extended their services infrastructurally, building identity architectures (profiles, login mechanisms) that extend to other sites and computationally linking themselves to the rest of the web.[55] These were all strategic attempts by platforms to counter the economic risks of being mere intermediaries, by turning themselves into ecosystems that keep users using their services and make data collection more comprehensive and more valuable.

In other words, to be free of intermediaries, we accepted new intermediaries. Platforms answer the question of distribution differently from the early web or traditional media. But they do offer the same basic deal: we'll handle distribution for you—but terms and conditions will apply. These terms may be fewer and less imposing, though you may be asked to do more of the labor of posting, removing, maintaining, tagging, and so on. But the platform still acts as a provider.[56]

Many content moderators and site managers came to these roles because they themselves were active users of the early Internet. Some of today's platforms, particularly the ones that began as startups rather than as projects of large corporations, grew out of that participatory culture. The fundamental mythos of the open web was extended to the earliest platforms: they often characterize themselves as open to all; in their promotion they often suggest that they merely facilitate public expression, that they are impartial and hands-off hosts with an "information will be free" ethos, and that being so is central to their mission.[57] Unfettered speech and participation on one's own terms, they believed, meant that rough consensus would emerge and democratic values would flourish.

On the other hand, early adopters of the web were also enamored with the possibility of "virtual community": like-minded individuals, joined by interest rather than geography or social obligation, building meritocratic and progressive social structures from scratch, and achieving the kind of communitarianism that had eluded Western society thus far.[58] These champions of online communities quickly discovered that communities need care: they had to address the challenges of harm and offense, and develop forms of governance that protected their community and embodied democratic procedures that matched their values and the values of their users.[59]

Both of these were, in important ways, myths. Nevertheless, they were familiar and meaningful to many of the people who found themselves

in charge of moderating early social media platforms, and remain part of the corporate culture of many of the platforms today. This has had two consequences. First, many of social media platform designers were initially caught off guard by the proliferation of obscenity and cruelty on their sites. As one content policy manager at Dreamwidth put it, "Everybody wants their site to be a place where only Good Things happen, and when someone is starting up a new user-generated content site, they have a lot of enthusiasm and, usually, a lot of naïveté. . . . They think of their own usage of social media and their friends' usage, and design their policies on the presumption that the site will be used by people in good faith who have the same definitions that they do as to what's unacceptable. That works for a while."[60]

Second, even as it became clear that content moderation was necessary, these two animating principles were in many ways at odds when it came to deciding how to intervene. Social media platform moderators often invoke one or even both principles when framing the values by which they moderate. But a platform committed to free speech, and comfortable with the wild and woolly Internet that early web participants were accustomed to, might install a very different form of moderation from that of a platform conceived as the protector of community, its moderators attuned to all the forces that can tear such community apart.

Still, from an economic perspective, all this talk of protecting speech and community glosses over what in the end matters to platforms more: keeping as many people on the site spending as much time as possible, interacting as much as possible. But even in this sense, platforms face a double-edged sword: too little curation, and users may leave to avoid the toxic environment that has taken hold; too much moderation, and users may still go, rejecting the platform as either too intrusive or too antiseptic. This is especially true as platforms expand their user base: platforms typically begin with users who are more homogenous, who share the goal of protecting and nurturing the platform, and who may be able to solve some tensions through informal means.[61] As their user base broadens it tends also to diversify, and platforms find themselves hosting users and whole communities with very different value systems, and who look to the platform to police content and resolve disputes.

Today, there are many social media platforms vying for our attention, but only a handful in each domain seem to enjoy the bulk of users and of the public's interest. Here is a representative but not exhaustive list of the social

media platforms I think about, and that will be central to my concern in this book: social network sites like Facebook, LinkedIn, Google+, Hi5, Ning, NextDoor, and Foursquare; blogging and microblogging providers like Twitter, Tumblr, Blogger, Wordpress, and Livejournal; photo- and image-sharing sites like Instagram, Flickr, Pinterest, Photobucket, DeviantArt, and Snapchat; video-sharing sites like YouTube, Vimeo, and Dailymotion; discussion, opinion, and gossip tools like Reddit, Digg, Secret, and Whisper; dating and hookup apps like OK Cupid, Tinder, and Grindr; collaborative knowledge tools like Wikipedia, Ask, and Quora; app stores like iTunes and Google Play; live broadcasting apps like Facebook Live and Periscope.[62]

To those I would add a second set that, while they do not neatly fit the definition of platform, grapple with many of the same challenges of content moderation in platformlike ways: recommendation and rating sites like Yelp and TripAdvisor; exchange platforms that help share goods, services, funds, or labor, like Etsy, Kickstarter, Craigslist, Airbnb, and Uber; video game worlds like League of Legends, Second Life, and Minecraft; search engines like Google, Bing, and Yahoo.

At this point I should define the term that I have already relied on a great deal. *Platform* is a slippery term, in part because its meaning has changed over time, in part because it equates things that nevertheless differ in important and sometimes striking ways, and in part because it gets deployed strategically, by both stakeholders and critics.[63] As a shorthand, "platform" too easily equates a site with the company that offers it, it implies that social media companies act with one mind, and it downplays the people involved. Platforms are sociotechnical assemblages and complex institutions; they're not even all commercial, and the commercial ones are commercial in different ways. At the same time, "platform" is a widely used term, including by the companies themselves. And when assigning responsibility and liability (legal and otherwise) we often refer to institutions as singular entities, and for good reason.

For my purposes, platforms are: online sites and services that

a) host, organize, and circulate users' shared content or social interactions for them,
b) without having produced or commissioned (the bulk of) that content,
c) built on an infrastructure, beneath that circulation of information, for processing data for customer service, advertising, and profit.

For the most part, platforms don't make the content; but they do make important choices about it.[64] While the early platforms merely made user contributions available and searchable, increasingly they determine what users can distribute and to whom, how they will connect users and broker their interactions, and what they will refuse. This means that a platform must negotiate any tensions between the aims of independent content providers who want their work to appear on a public forum and the platform's own economic and political imperative to survive and flourish.[65] And it must do so without having produced or commissioned the content. This means that platform managers by and large cannot oversee content through more traditional media industry relations such as salary, contract, or professional norms. For traditional media, employment arrangements and shared norms were key means of prohibiting illicit content. Platforms must find other ways.

Most platforms still depend on ad revenue, extending the monetization strategy common to amateur home pages, online magazines, and web portals. Advertising still powerfully drives their design and policy decisions. But most social media companies have discovered that there is more revenue to be had by gathering and mining user data—the content users post, the profiles they build, the search queries they enter, the traces of their activity through the site and beyond, the preferences they indicate along the way, and the "social graph" they build through their participation with others. This data can be used to better target all that advertising, and can be sold to customers and data brokers. This means platforms are oriented toward data collection and retention; toward eliciting more data, and more kinds of data, from its users; and toward finding new ways to draw users to the platform, and to follow users off the platform wherever they may go.[66]

And now, for the fine print. Some would argue that I am using the term *platform* incorrectly. It has a more specific computational meaning, where it means a programmable infrastructure upon which other software can be built and run, like the operating systems in our computers and gaming consoles, or information services that allow developers to design additional layers of functionality.[67] Some have suggested that the term should be constrained to this meaning—that Facebook, for example, is not a platform because it hosts our updates and photos, it is a platform only in that it provides an application programming interface (API) for software developers to design extensions and games atop it.[68] The distinction is convincing,

but at this point, it's simply too late: *platform* has been widely embraced in its new sense—by users, by the press, by regulators, and by the platform providers themselves.

I may also be using the term too broadly. Platforms vary, in ways that matter both for the influence they can assert over users and for how they should be governed. It is deceptively easy in public debates, and in scholarship, to simply point in the direction of Facebook and move on, without acknowledging the variety of purpose, scope, membership, economics, and design across different platforms. For instance, YouTube has developed a program for paying some of its users, which changes the dynamics between platform and those users significantly. A live-streaming platform like Periscope faces different challenges moderating content in real time.

I may also be using the term too narrowly. First, my location and limited proficiency in languages limits my analysis to platforms based in the West and functioning largely in English. This overlooks massive platforms in other countries and languages, like Sina Weibo in China, VK in Russia, and, until 2014, Google's Orkut in South America. However, many of the platforms I consider have a global reach and influence, mattering a great deal across many parts of the world. While this does not make my analysis universal, it does extend it beyond the specific platforms I focus on.

The platforms I spend the most time discussing are the largest, the most widely used, the best known. These are, of course, all good reasons to pay particular attention to them. It matters how Facebook sets and enforces rules, even if you're not on Facebook. And it is harder and harder to not be on Facebook, even if you are uncomfortable with its oversight. But there are dozens of other platforms competing with these to be national or global services, and there are many thousands of smaller sites, with no such ambitions, more focused on specific regions or interests. All face many of the same moderation challenges, though on a smaller scale and with substantially less public scrutiny and criticism. Smaller sites may even be breeding grounds for innovative approaches and solutions to the challenges all platforms face. And there are also plenty of social media sites that are long dead, or nearly so—Friendster, MySpace, Orkut, Revver, Veoh, Chatroulette, Ping, Delicious, Xanga, Airtime, Diaspora, Vine, Yik Yak—that also faced the challenges of moderation, and can still be illuminating examples.

I did not include messaging services, which are hugely popular competitors to the platforms mentioned above. Millions regularly use WhatsApp, Facebook Messenger, QZone, WeChat, Kik, Line, Google Hangout, and

Skype to communicate and congregate online. Because they are generally person-to-person or group-to-group, and overwhelmingly between known contacts, they sidestep many of the problems that plague platforms that offer public visibility and contact with strangers. But they too engage in their own forms of moderation.

Finally, there is a broader set of information sites and services that, while I would not lump them into this category, face similar questions about user activity and their responsibility for it: online discussion forums, unmoderated social spaces online, gaming worlds that allow for player-to-player interaction, amateur porn platforms, comment threads on blogs, news sites, and inside e-commerce sites.

PLATFORMS ARE NOT PLATFORMS WITHOUT MODERATION

To the definition of platforms, I would like this book to add a fourth element:

d) platforms do, and must, moderate the content and activity of users, using some logistics of detection, review, and enforcement.

Moderation is not an ancillary aspect of what platforms do. It is essential, constitutional, definitional. Not only can platforms not survive without moderation, they are not platforms without it. Moderation is there from the beginning, and always; yet it must be largely disavowed, hidden, in part to maintain the illusion of an open platform and in part to avoid legal and cultural responsibility. Platforms face what may be an irreconcilable contradiction: they are represented as mere conduits *and* they are premised on making choices for what users see and say.

Looking at moderation in this way should shift our view of what social media platforms really do: from transmitting what we post, to constituting what we see. There is no position of impartiality. Platform moderators pick and choose all the time, in all sorts of ways. Excluding porn or threats or violence or terrorism is just one way platforms constitute the social media product they are generating for the audience.

The persistent belief that platforms are open, impartial, and unregulated is an odd one, considering that *everything* on a platform is designed and orchestrated. Economists know this: like with any "multisided market," a platform company is a broker, profiting by bringing together sellers and buyers, producers and audiences, or those in charge of tasks and those with the necessary skills to accomplish them.[69] So, if Uber profits by bringing

independent drivers to interested passengers, coordinating and insuring their interaction, and taking a fee from the exchange, Twitter does much the same: it brings together independent speakers with interested listeners, coordinates their interaction, and takes a fee from the exchange—in the form of valuable user data.

It is a position that can be, for a few, extremely lucrative: as John Herrman notes, "If successful, a platform creates its own marketplace; if extremely successful, it ends up controlling something closer to an entire economy."[70] And it depends on platforms not only bringing independent parties together but completely structuring every aspect of the exchange. YouTube connects videomakers with viewers, but also sets the terms: the required technical standards, what counts as a commodity, what is measured as value, how long content is kept, and the depth and duration of the relationship. YouTube can offer established videomakers a share of the advertising revenue or not, and it gets to decide how much, to whom, and under what conditions. And like any market, game world, or information exchange that invites users to participate according to their own interests, this requires excluding some to serve others: those who provide unwanted goods, those who game the system, those who disrupt the entire arrangement.

How platforms are designed and governed not only makes possible social activity, it calls it into being, gives it shape, and affirms its basic legitimacy as a public contribution. Platforms don't just mediate public discourse, they constitute it.[71] As José van Dijck observes, "Sociality is not simply 'rendered technological' by moving to an online space; rather, coded structures are profoundly altering the nature of our connections, creations, and interactions."[72] They are designed so as to invite and shape our participation toward particular ends. This includes how profiles and interactions are structured; how social exchanges are preserved; how access is priced or paid for; and how information is organized algorithmically, privileging some content over others, in opaque ways. These "social media logics" are the repertoires of expression and action that social media platforms trade in.[73] Put simply, if Twitter were designed and managed in fundamentally different ways, that would have some effect on what users could and would do with it. This includes what is prohibited, and how that prohibition is enforced.[74]

On the other hand, it is also easy to overstate the influence platforms have as straightforward and muscular—either facilitating participation in powerful ways or constraining and exploiting it in powerful ways. Users

don't simply walk the paths laid out by social media platforms. They push against them, swarm over them, commandeer them, and imbue them with new meanings. The instant a social media platform offers its service publicly, it is forever lashed to a ceaseless flow of information and activity that it cannot quite contain. So yes, Facebook tweaks its newsfeed algorithm and suspends users for breaking the rules. But it also hosts the regular participation of more than a billion people, who use the platform for countless different activities. This torrent of participation never stops, and is shaped much more by its own needs and tactics.[75] Whatever structure a platform attempts to introduce may cause its own little eddies and detours, but they are minor compared to the massive perturbations endemic to public discourse: shifting sentiments, political flare-ups, communal and national rhythms, and the recursive loops of how forms of participation emerge and propagate, then are superseded. Platform managers may want to support and expand this ceaseless flow, but they also remain in constant fear of it turning sour or criminal, or simply drying up. While platforms structure user activity, users also have power over platforms—maybe less so as mere individuals or groups, but more in the aggregate, the slow, unrelenting shifts in what people seem to want to do.

This is not to say that platforms are of no consequence. I simply mean that we must examine their role, without painting them as either all-powerful or merely instrumental. We must recognize their attenuated influence over the public participation they host and the complex dynamics of that influence, while not overstating their ability to control it. Examining moderation and how it works slices these questions open for scrutiny.

Platforms may not shape public discourse by themselves, but they do shape the shape of public discourse. And they know it.