*Article*

SM+S
social media + society

Social Media + Society
July-September 2022: 1–13

DOI: 10.1177/20563051221117552
journals.sagepub.com/home/sms
SAGE

# Do Not Recommend? Reduction as a Form of Content Moderation

**Tarleton Gillespie**[1,2]

**Abstract**

Public debate about content moderation has overwhelmingly focused on removal: social media platforms deleting content and suspending users, or opting not to do so. However, removal is not the only available remedy. Reducing the visibility of problematic content is becoming a commonplace element of platform governance. Platforms use machine learning classifiers to identify content they judge misleading enough, risky enough, or offensive enough that, while it does not warrant removal according to the site guidelines, warrants demoting them in algorithmic rankings and recommendations. In this essay, I document this shift and explain how reduction works. I then raise questions about what it means to use recommendation as a means of content moderation.

**Keywords**

platforms, content moderation, algorithms, shadowbanning, borderline content

It is easy to assume that, in the current debates about how social media platforms moderate problematic content, content moderation is content *removal*: deleting content and suspending users. Even the journalists, critics, policymakers, industry stakeholders, and academics who know better spend an inordinate amount of time focused on removal. Debating whether a platform should permanently ban the sitting president of the United States is a potent way to interrogate its influence. The First Amendment implications of removal lure journalists, pundits, law scholars, even social scientists. Accusations of "censorship" by critics have the most traction when content has been completely deleted or a user has been permanently banned.

While removal may be the most visible response, it is by no means the only remedy available. Besides being able to (1) remove content and users, platforms can also (2) implement age barriers, geo-blocking, or temporary holds, to keep problematic content away from some users some of the time; (3) append fact-checking labels and interstitial warnings, to alert users to problematic content before or as they encounter it; (4) impose demonetization and punitive strike systems, as disincentives for producing problematic content; and (5) provide counter-speech and model preferred norms, to raise the overall level of discourse (Goldman, 2021). Some of these strategies are not as visible or controversial as removal; others may go unnoticed because they are imposed by different product teams, intervene at different points in the platform, are more difficult for users to identify in practice, or get justified in different ways.

This essay focuses on yet another type of remedy: when platforms (6) reduce the visibility or reach of problematic content. Many social media platforms have quietly begun to identify content that they deem not quite bad enough to remove. The offending content remains on the site, still available if a user can find it directly. However, the platform limits the conditions under which it circulates: whether or how it is offered up as a recommendation or search result, in an algorithmically generated feed, or "up next" in users' queues. This is not a new policy or technique, exactly, though it is being deployed to new ends, for an expanding list of reasons. Reducing the visibility of risky, misleading, or salacious content is now used by many platforms to curb any content they deem as nearly violating their rules, across nearly all the traditional Trust & Safety concerns.

YouTube and Facebook, which have most clearly acknowledged these techniques, call them "borderline content" policies. I am reluctant to adopt that term: using "borderline" as a

[1]Microsoft Research, USA
[2]Cornell University, USA

**Corresponding Author:**
Tarleton Gillespie, Microsoft Research, 1 Memorial Drive, Cambridge, MA 02142, USA.
Email: tarleton@microsoft.com

**Correction (May 2023):** Article updated online to include Figure 2 on page 5.

pejorative has quiet resonances with assumptions too often made about both geographic borders and "borderline" mental health conditions, in ways I do not want to reify. Critics have called this "shadowbanning"[1] or "suppression"[2]—I find these terms more compelling, though I worry, along with Cotter (2021), that they may inadvertently help platforms dodge the very criticism being leveled at them. I will use *reduction* to encompass all of these.

That there is not yet a settled industry term is telling. Platforms are, understandably, wary of being scrutinized for these policies—either for being interventionist and biased, or opaque and unaccountable. Some platforms have not acknowledged them publicly at all. Those that have are circumspect about it. It is not that reduction techniques are hidden entirely, but platforms benefit from letting them linger quietly in the shadow of removal policies. So, despite their widespread use, reduction policies remain largely absent from news coverage, debate, policymaking, and even much of the scholarly conversations about content moderation and platform governance.

My aim is not simply to ask whether these techniques are good or bad, or suggest how they could be improved or regulated. I want to argue that reduction techniques should be included within a broadened definition of "content moderation," not adjacent to it. Content moderation and algorithmic recommendation are mostly treated as different platform functions, handled by different product teams, and scrutinized by different critics according to different measures and concerns (Caplan, 2019). However, if we remain analytically agnostic about these distinctions, we might instead see both content moderation and algorithmic recommendation as governance—by different means, deployed in different ways, with different justifications.

First, I will define reduction techniques and identify how they are implemented by several of the major social media platforms. Then, I will examine how the platforms explain and justify these interventions. I will explain how reduction works and offer a typology of reduction techniques. I will then make a case for why reduction must be understood as part of a broader content moderation project, and how doing so changes how we think about content moderation and platform governance.

## A Methodological Note

For this research, I examined company policies, corporate communications, and some of the technical literature coming from both academia and industry. However, I also lean on conversations I have had with platform employees in Trust & Safety and adjacent areas—a dozen or more representatives from five major social media platforms. Let me make a quick note on my access to and use of these conversations.

Many Silicon Valley technology companies are reluctant to grant researchers access, especially for qualitative inquiry into the company's decision-making. As Bonini and Gandini (2020) note, more than the technology, it may be the industry itself that is "black boxed." This reluctance protects powerful institutions from scrutiny (Monahan & Fisher, 2015, p. 710). Nondisclosure agreements (NDA), once a way to protect trade secrets, are now a ubiquitous and routine part of employee contracts and encounters with interested outsiders, and help manage scrutiny and suppress criticism (Starr, 2020). These agreements are shaping how technologists think about those who come asking interested questions, as a matter of secrecy (Seaver, 2016).

However, this guarded industry also regularly trades on informal intra- and inter-company consultations. Managers seek out informal advice from their industry peers; consult with nonprofit, advocacy, and academic groups; attend conferences on pressing issues; and orchestrate one-off conversations with people elsewhere in their own company, or at adjacent ones. When these interactions cross the outer membrane of the corporation, sometimes they are governed by an NDA as a matter of course. However, often they are not, depending instead on informal relations and mutual trust.

I find myself with a bit of this "insider access," where different questions can be asked, and are more likely to be answered. However, the instant I would identify a conversation as a research interview, no matter how informally I did so, that performative gesture often triggered a nervousness that any "interview" would need to be under NDA. Formality seemed to invoke a matching formality: mine an Institutional Review Board (IRB) consent process, theirs an NDA. I felt an NDA would be a substantive barrier to talking frankly and publicly about these policies later on. So instead, I promised informants that our conversations would be treated as background. I fully acknowledge that there is a privilege to my position—not so much from being employed by a technology company myself, ironically, which in some ways complicates my access to other companies—more from the working relationships developed over the course of previous research, aided by presumptions my informants may make about me as "working in industry" like them. These are not conversations that another researcher could necessarily have, and I recognize the methodological problem that raises.

Any details I learned, I tried to verify using that company's published policies and statements. In the article, I use published evidence to present details I may have first learned in confidence. For anything else, I have not disclosed specific details, though they inform my argument. Even so, I remain concerned that these routinized logics of secrecy and informality, especially in Silicon Valley, are a problematic barrier to qualitative inquiry, and that our ethical commitments may at times hamper us from investigating and challenging the strategies of the powerful (Souleles, 2021). I am keenly aware that my efforts to know these companies are entangled in their use of me. In these conversations, am I finding out things from them, or are my questions a form of

free consulting? Both, I suspect. And like other researchers, I function within the "contact zones that media industries stage carefully around us as we study them" (Caldwell, 2013, p. 165, cited in Vonderau, 2014). Still, I am unwilling to let that prevent me entirely from revealing and explaining policies these platforms continue to publicly obscure.

## "Borderline Content"

YouTube announced its "borderline content" policy in a January 2019 post, though the practice had already been in place for a few months or more. This followed on the heels of nearly 2 years of blistering criticism, not only for hosting problematic content but also for amplifying it with its recommendation system. Ex-YouTube designer Guillaume Chaillot specifically blamed the recommendation algorithm for the glut of misinformation, warning that "fiction is outperforming reality";[3] Zeynep Tufekci, commenting on a *Wall Street Journal* exposé of conspiracy videos, called YouTube "the great radicalizer" for its tendency to recommend increasingly extreme videos, regardless of the topic.[4]

It is worth lingering for a moment on how YouTube framed this new intervention. The title of the post, "Continuing our work to improve recommendations on YouTube,"[5] gives no indication that this is in any way a Trust & Safety concern. The post begins with a lineage of adjustments YouTube had already made, including limiting "clickbaity videos with misleading titles and descriptions," and "getting too many similar recommendations," part of the "hundreds of changes to improve the quality of recommendations for users" YouTube had made just in the past year. Then, the new policy is introduced as if it isn't new at all:

> We'll continue that work this year, including taking a closer look at how we can reduce the spread of content that comes close to—but doesn't quite cross the line of—violating our Community Guidelines. To that end, we'll begin reducing recommendations of borderline content and content that could misinform users in harmful ways—such as videos promoting a phony miracle cure for a serious illness, claiming the earth is flat, or making blatantly false claims about historic events like 9/11.

Harms are presented as falling along a spectrum: "content that comes close to—but doesn't quite cross the line of—violating our Community Guidelines." In this spatial understanding of harm lies a "borderline," just left of the existing prohibitions (Maddox & Malson, 2020).

Notice that the problems identified are traditionally moderation issues—misinformation and conspiracy—but the tactic is drawn from recommendation quality: clickbait, repetitive suggestions, and so on. While discussions of online harm tend to highlight victims and perpetrators, quality interventions tend to treat users as consumers, and are more concerned about customer satisfaction than the public interest.

YouTube then quickly assures the reader that:

> To be clear, this will only affect recommendations of what videos to watch, not whether a video is available on YouTube. As always, people can still access all videos that comply with our Community Guidelines and, when relevant, these videos may appear in recommendations for channel subscribers and in search results. We think this change strikes a balance between maintaining a platform for free speech and living up to our responsibility to users.

Reduction techniques depend on this kind of demarcation, between what is hosted and what is recommended, between archive and algorithm. Reduction only pertains to where content is offered up, suggested: not where it lives in the archive (Gillespie, 2016).[6]

YouTube later incorporated this strategy into a broader approach to governance that the company called "The Four Rs": remove, raise, reward, and reduce. By late 2019, YouTube PR could assert that the effort to "reduce" borderline content and harmful misinformation was working: "The result is a 70% average drop in watch time of this content coming from non-subscribed recommendations in the U.S."[7]—though as Lewis noted, such claims are nearly impossible to confirm, especially given that the category itself is defined by YouTube's efforts to identify it.[8]

Facebook and Instagram had already announced their reduction policy in May 2018,[9] after also enduring 2 years of criticism for amplifying misinformation.[10] The terminology and justifications are similar to YouTube's:

> There are other types of problematic content that, although they don't violate our policies, are still misleading or harmful and that our community has told us they don't want to see on Facebook—things like clickbait or sensationalism. When we find examples of this kind of content, we reduce its spread in News Feed using ranking . . .[11]

Facebook later published a detailed list of what pages, groups, or events they will not recommend[12] and an exhaustive "Content Distribution Guidelines" indicating what they will "demote" in the News Feed.[13] Facebook's list includes three categories of concern, two of which echo YouTube's: borderline content and harmful misinformation, but also low-quality junk. Facebook will not recommend "content that users broadly tell us they dislike," meaning clickbait, "engagement bait," contest giveaways, and links to deceptive or malicious sites; or "content that is associated with low-quality publishing," including unoriginal or repurposed content, news that's unclear about its provenance, or, content enjoying a surge of engagement on Facebook that's unmatched on the wider web. Including this in their reduction strategy, just like when YouTube frames theirs in the legacy of clickbait, further confirms that this is the aim of content moderation using the tools developed in the pursuit of recommendation quality.

In July 2021, Instagram introduced "sensitive content control," which allowed the user to adjust the degree to which sensitive content should be filtered out of the "explore" recommendations the platform offers.[14] While the announcement emphasized the agency users were being given, the fact that users could now "allow," "limit (default)," or "limit even more" sensitive content revealed that such content already was, by default, being reduced. The company did not immediately specify what counted as sensitive; Instagram head Adam Mosseri later indicated that the intervention focuses on "sexually suggestive, firearm, and drug-related content"[15] and was separate from parallel efforts to reduce misinformation and self-harm.

Twitter, LinkedIn, and TikTok have similar reduction strategies already in place, though they have been less vocal about them. In a July 2020 promise to address COVID-19 vaccine misinformation, Twitter asserted, "Tweets that are labeled under this expanded guidance will have reduced visibility across the service . . . however, anyone following the account will still be able to see the Tweet and Retweet."[16] They used similar language to describe their approach to election misinformation in 2020 and 2021.[17] LinkedIn has acknowledged that, in response to community guidelines violations, they may "limit the visibility of certain content, or remove it entirely."[18] That language first appeared in LinkedIn's "Professional Community Policies" when they were substantially rewritten in December 2019.[19] TikTok has been even coyer, perhaps because their widely praised recommendation engine is so central to their service. However, in the community guidelines, the company acknowledges that

> for some content—such as spam, videos under review, or videos that could be considered upsetting or depict things that may be shocking to a general audience—we may reduce discoverability, including by redirecting search results or limiting distribution in the For You feed.[20]

Depending on how we broaden the definition, we can find other platforms engaged in reduction strategies that share at least a family resemblance. Tumblr, which once took a more permissive approach to sexual content (Tiidenberg et al., 2021), used to limit the circulation of explicit content by refusing to serve up search results to queries it judged explicit. Users could post pornographic images, could even tag them as "#porn"—but if a user searched for "#porn" no results would be returned. Hashtag blocking has its problems (Gerrard, 2018; Pilipets & Paasonen, 2022; Sybert, 2021). However, like reduction, it similarly demarcates between hosting content and offering it up in search results or recommendations.

Reddit's "quarantine" policy, though structured differently (Chandrasekharan et al., 2022; Copland, 2020; DeCook, 2022), belongs as well. Reduction is just one part of the Reddit quarantine, but the effect is similar. For users who don't already subscribe to a quarantined subreddit, no posts from within that subreddit will appear on the Reddit front page or be returned in search results or other site-wide recommendations. In other words, it's there if you go looking for it, but it won't be circulated broadly. Reddit's reasoning will sound familiar: "to prevent its content from being accidentally viewed by those who do not knowingly wish to do so, or viewed without appropriate context."[21] It is worth noting that Reddit has been more explicit and transparent about their quarantines than YouTube and Facebook have about reduction. Still, Reddit users outside the quarantined subreddit may not know why some things aren't bubbling up on their front page anymore.

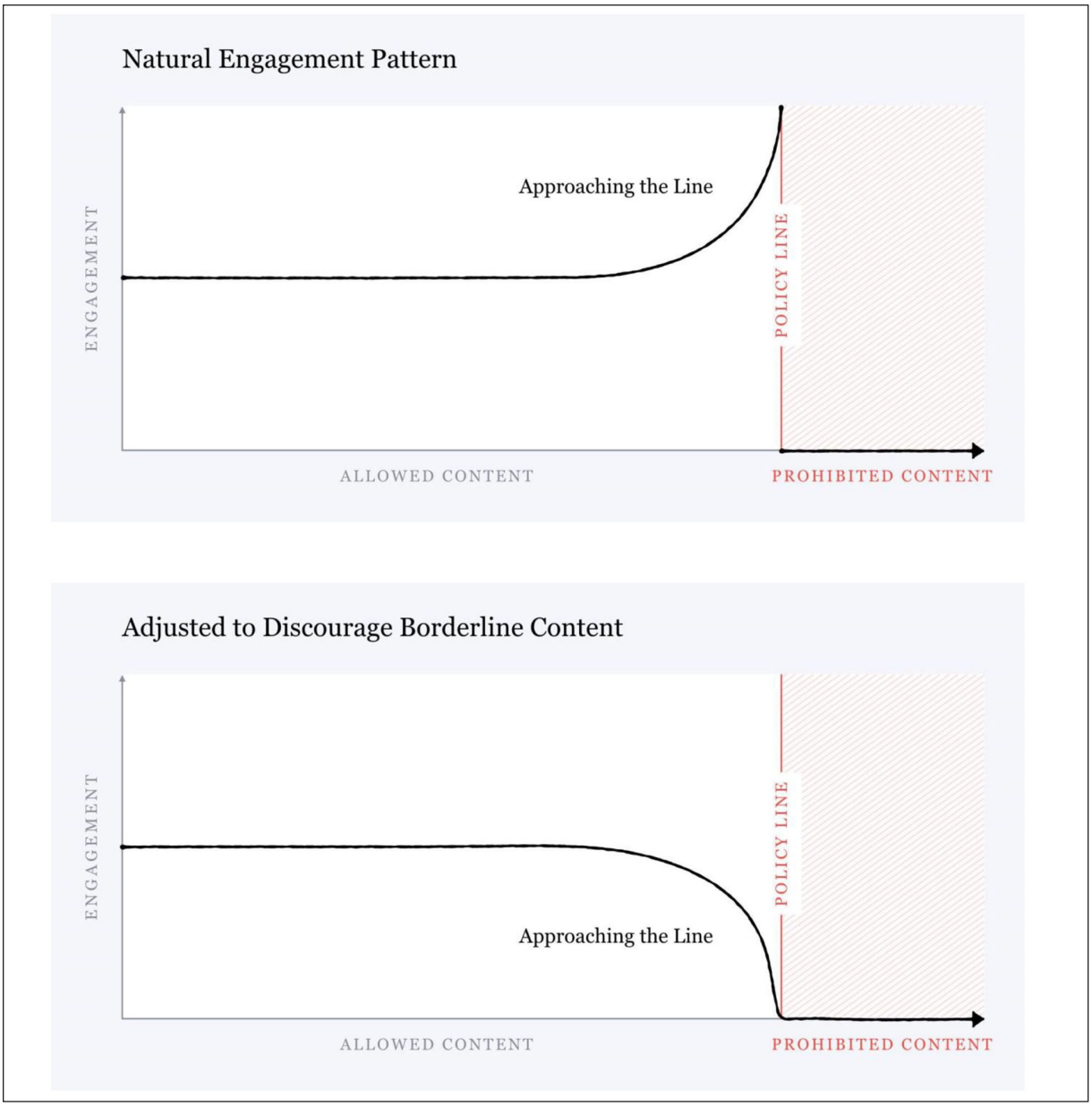
## Why Not Just Remove It

If a platform deems content problematic enough to justify reducing its circulation, why not simply remove it? In other words, if the problem is at the borderline, why not just move the border? After all, Trust & Safety teams write these rules themselves, change them frequently, and reserve the right to interpret them as they see fit; why not adjust the rule and remove the content? The platforms that have acknowledged implementing these approaches have offered several kinds of reasons; I will suggest a few more that, unsurprisingly, tend to go unsaid publicly.

Facebook has offered a jumble of justifications for their reduction policy. First, they were responding to their users' wishes—what "our community has told us they don't want to see"—though without indicating how those wishes were articulated. Six months later, Mark Zuckerberg and his PR team offered a second, somewhat different justification:

> Our research suggests that no matter where we draw the lines for what is allowed, as a piece of content gets close to that line, people will engage with it more on average—even when they tell us afterwards they don't like the content. This is a basic incentive problem that we can address by penalizing borderline content so it gets less distribution and engagement.[22]

According to Zuckerberg, users are (invariably) drawn to "sensationalist and provocative content." Wherever a line is drawn, interest balloons alongside it. This incentivizes creators to be daring, saucy, outrageous, and dangerous. By this logic, Facebook is not implicated, it is simply dealing with the inevitable outcome of policing itself: a "natural engagement pattern" it is not directly responsible for (Hallinan, 2021). This is akin to concerns about "gaming the algorithm" (Petre et al., 2019), that clever content creators will do precisely as much as they can get away with—except that, in this characterization, demand comes first. Zuckerberg included Figures 1 and 2 (that look mathematical, but are not) to explain both the problem and his solution (Intriguingly, after the reduction policy was in place, Zuckerberg not only hoped to reduce demand and supply down to an equivalent level, but to mute it further).




**Figures 1 and 2.** Mark Zuckerberg, Facebook, 15 November 2018. "Blueprint for content governance and enforcement."

In Facebook's view, moving the line would only shift the problem: to the left of every line, no matter where it is drawn, is a bubble of demand for the sensational and illicit that someone will fulfill. It is worth noting that this is a very specific theory of culture and taste, and I am not convinced that it is correct. Still, naturalizing this as human and inevitable allows Facebook to act responsibly without admitting responsibility.

A year later, Facebook offered a third justification:

> Since recommended content doesn't come from accounts you choose to follow, it's important that we have certain standards for what we recommend. This helps ensure we don't recommend potentially sensitive content to those who don't explicitly indicate that they wish to see it.[23]

Consent, which Facebook infers when you like a page or friend another user, justifies a lesser standard of responsibility than a recommendation coming "from" Facebook.[24]

YouTube management has generally framed their reduction policies as an acknowledgment of a growing public responsibility. In a long-form piece on the new policy, *Wired* reported that the company's incessant drive to grow had pushed the YouTube recommendation team to redesign the

recommendation system in whatever way would increase watch time; by 2016, "recommendations had become the thrumming engine of YouTube, responsible for an astonishing 70 percent of all its watch time."[25] The sense of being implicated in the problems of misinformation and conspiracy that followed—whether this was an ethical awakening from within YouTube, or a turgid response to public outcry—is now presented as YouTube's justification for reducing conspiratorial, misleading, and otherwise "borderline" content.

In my informal conversations, I was also offered other, less noble reasons for not simply removing this problematic content. First, reduction is less politically risky than removal. Given the recent climate, platforms fear reprisals from conservative critics who air their outrage whenever their posts are removed.[26] Demoting reprehensible content lets platforms avoid "censoring" it, or facing charges of bias that are difficult to refute. Flip this around, and it is not difficult to imagine that reducing problematic content allows platforms to continue to benefit financially from the users who do seek it out, in the form of advertising revenue and/or data collection, while still answering public concerns by reducing its reach (Matamoros-Fernández, 2017; Siapera & Viejo-Otero, 2021).

Reduction strategies may also be preferable when the types of problematic content are difficult to identify, in flux, or difficult to police. Reducing without removing allows platforms the flexibility to intervene around quickly emerging phenomena, to go after content designed to elude prohibitions, and to curtail content they "know" is bad but have a hard time articulating why. Seen in the best light, this flexibility makes it easier to respond quickly to changing problems—from the unpredictable outbursts of White nationalism, to the evasive tactics of pro-ana groups, to the constantly evolving QAnon conspiracy. In a less flattering light, reduction also avoids public accountability, as the interventions themselves are hard to spot, and are not—yet—reported as part of the platform's transparency obligations.

## How Reduction Works

Reduction is not the exclusive purview of the Trust & Safety teams who traditionally handle content moderation. Like the conceptual distinction being made between different zones of responsibility on the platform, reduction lives at a similar institutional distinction between the teams that handle algorithm design and those that handle policing. This is not an insignificant detail: where work done within a company can reveal a great deal about how a problem is perceived, what approaches are likely to be pursued, who has to approve and justify the resources required, and how the solution will be understood and valued by the company's senior leadership (Caplan & boyd, 2018).

Recommendation is a central component of social media platforms, but it is driven by a different set of concerns and priorities: while Trust & Safety teams *select out* what is deemed least appealing, the teams that manage recommender systems and newsfeeds *select for* what is deemed most appealing. Their north star is engagement, usually measured by the amount of time users spend on the platform, the number and types of actions taken, and other proxy measures of satisfaction (Bucher, 2018; McKelvey & Hunt, 2019; S. Singh, 2019). Their primary technique is to collect signals about the specific user, about users like them, and about all the available content in the corpus, to produce a personalized feed of content that will be maximally appealing.

Generally these signals measure some aspect of the content understood to be positive—that is, reasons why the content should be shown: Is this video recent? Is this link recommended by this user's friends or network? Is this post often liked by users who share a similar matrix of interests? Content with a higher score on these measures is more likely to be recommended. Sophisticated recommender systems can calculate hundreds of these signals, measure each signal against its own specific threshold, weigh them differently, and adjust dynamically depending on the user, region, situation, genre, or moment. Reduction techniques take advantage of this, but flip it: the calculation of what to recommend now includes a negative signal, indicating that a particular piece of content should not be recommended to this particular user.

There are several points along the recommendation process where reduction can take place. In its own technical literature, for example, YouTube engineers distinguish two key steps in their recommendation algorithm: inventory and ranking.[27] YouTube does not rank millions of videos every time it needs to recommend content to a user. Inventory, or what in Figure 3 is called "candidate generation," represents a first pass at gathering videos that might be relevant to this user. YouTube's software quickly identifies a small set of videos to even consider recommending, numbering in the hundreds, based generally on their similarity to videos the user watched before, their recent popularity, and so on. "Ranking" then orders these videos to determine which to recommend next, and in what order.

This means reduction can happen in several ways, with different implications for how widely "reduced" content would circulate, and to whom:

1. *Do not recommend at all*: problematic content can be left out of the inventory altogether; a video excluded from the inventory will never be recommended, no matter how far down the user scrolled.
2. *Do not recommend as much*: problematic content can be included in the inventory, but ranked lower than other content. A video included in the inventory, but whose "borderline" status meant that other videos are likely to be recommended ahead of it, might still be recommended eventually; it could even be recommended highly, if it so perfectly matched that user's query and viewing history.[28] How often it will be recommended, then, will depend on how strongly it is downweighted, how many other criteria are also weighed strongly, and what other content it's competing against.

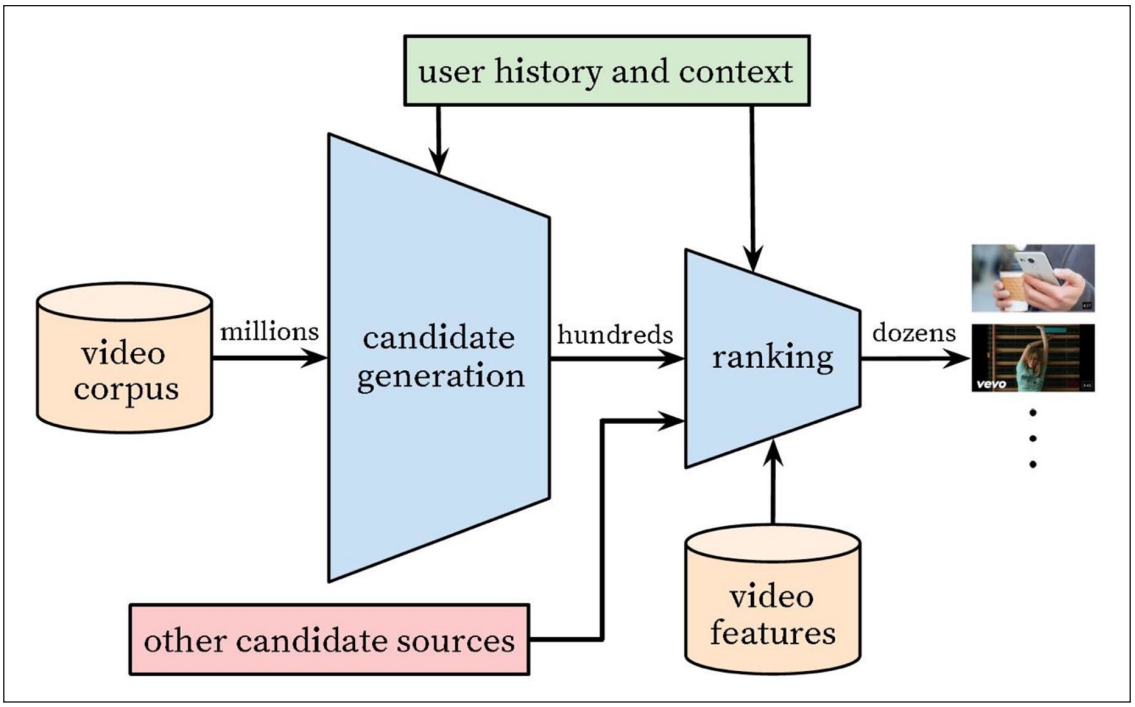



**Figure 3.** Deep neural networks for YouTube recommendations.
*Source.* Covington et al. (2016).[27]

3. *Do not recommend to some*: either of the above could be applied only to some users, based on their age, region, time of day, or distinguished by their indicated preference or search query, by past traces, or by whether the user is known or unknown.

In whatever form, this is an example of what Ananny (2020) calls a "probabilistic" approach to content moderation: these platforms want "less" of the content they have deemed problematic, and use mathematical tweaks to achieve it.

The other half of the challenge, common to all content moderation efforts, is how to identify the problematic content in the first place—and do so quickly, accurately, based on limited information, and at scale. It should surprise no one that, to accomplish this, most platforms have turned to a now well-worn Silicon Valley technique: develop a machine learning classifier that can estimate what content is "problematic," by training that classifier on a heap of data that's already been evaluated by human raters (Ekbia & Nardi, 2014; Gorwa et al., 2020; Gray & Suri, 2019). Ideally, the judgments made by the human raters will be approximated by the machine learning classifier, which can then make the same judgment over and over on millions of pieces of content.

Facebook seems to use the same machine learning classifiers already used to identify content they may remove.[29] Posts that score highly are removed, and posts that score close to that may now be algorithmically reduced. YouTube opted to develop a separate classifier, using the same machine learning techniques, specifically to identify the content they consider borderline or harmful.[30] This meant they needed data rated by humans. The labor of rating videos as "borderline" is farmed out as piecework to thousands of independent workers, much like other forms of commercial content moderation labor (Roberts, 2019; Ruckenstein & Turunen, 2019; Steiger et al., 2021). But YouTube asserts that, here, the work is structured differently. Evaluators are given more time per evaluation and are encouraged to do research—that is, Google searches—to help assess which videos are problematic: How fishy is this video and how false are its claims?[31] (Tellingly, the training manual YouTube gives its raters to spot conspiracy videos, misinformation, and other borderline content, is the same document Google uses to train evaluators of search result quality.[32] The tools of quality, turned to the concerns of harm). Using the scores from these human evaluators, the machine learning classifier is trained to spot borderline content and harmful misinformation.

## Making Visible What's Been Made Less Visible

Given how they have conducted content moderation up to this point, reduction policies may be the most mature step platforms have yet taken (DiResta, 2018). While it may not sit well with idealized notions of the marketplace of ideas, it may nevertheless be true that healthy public spheres actually require their information landscapes be curated. Whether platform managers are genuinely motivated by a sense of responsibility, or it is

just a good PR move, this may finally recognize the complex ways in which recommendation algorithms do seem to amplify some kinds of contributions over others, do seem to exacerbate the more chaotic elements of civic discourse. The technical field of recommender systems is beginning to think about what "healthy" recommendations might look like (Beutel et al., 2020; A. Singh et al., 2020), as part of a broader commitment to fairness, accountability, and transparency.

However, reduction avoids none of the legal or societal problems that already haunt content removal: the presumptive power of the arbiter, the possible biases, the inequitable impact on different user communities, and the implications for free speech (Keller, 2021). In fact, reduction techniques concentrate even more of that curatorial power in the hands of these elite, private, profit-oriented intermediaries, by embedding and obscuring their judgment and worldviews within these nearly imperceptible adjustments.

One common rejoinder to such unchecked power is to demand that platforms be transparent about and accountable to their own policies. For years, critics have pushed social media companies to be more transparent about the moderation decisions they make,[33] and the major platforms have slowly acquiesced; many now publish regular transparency reports, articulate their policies in greater detail, and offer more robust mechanisms for appealing moderation decisions (Ananny & Crawford, 2018; Suzor et al., 2019). However, reduction techniques, at least so far, remain completely invisible to these apparatuses of observability or accountability (Rieder & Hofmann, 2020). None of these interventions are reported or documented, and data about what is reduced and to what degree are not made publicly available. Except for Facebook, even the policies are not spelled out.

This is especially problematic because reduction is so hard to detect. When content is removed or a user suspended, there are at least traces left behind: the video is missing, and there is a mark where the deleted post used to be. However, there is no trace left when a post, tweet, or video simply has not circulated as far as it might have otherwise. The content remains, it can be found, commented on, and forwarded, yet it seems to not have the audience or traction that it might have. This uncertainty leaves users grasping for explanations, and is part of why so many users are suspicious that murky machinations are at work under the hood of these platforms (Caplan & Gillespie, 2020; de Keulenaar et al., 2021; Nicholas, 2022; West, 2018). Although critics may not have all the details right, or an easy way to confirm their suspicions, many have long suspected that platforms are gaming their own systems without making it known to their users. YouTube creators raise alarm bells when their work gets demonetized without explanation, forced to read the tea leaves on why revenue is down on one video and not another (Caplan & Gillespie, 2020; Cotter, 2021). Sex workers accuse platforms of shadowbanning, but must gather the evidence themselves (Are, 2021; Blunt et al., 2020, 2021; Smith et al., 2021). Conservative pundits accuse the ranking mechanisms of social media algorithms of being tipped against them, in ways that seem nefarious but cannot be pinpointed. While not all of these accusations will prove to be true, and some are politically self-serving for those making them, they tap into an underlying truth: platforms *are* in fact using a wider array of tools to shape the flow of information, in ways that are not particularly transparent. Frustrated users grapple with that uncertainty, asserting that they are being thwarted—yet any assertion "falls short as actionable knowledge" (Savolainen, 2022, p. 7). It is disheartening, even reprehensible, that our best evidence of these techniques continues to come from surveys of users who *suspect* that they have been shadowbanned (Nicholas, 2022).

At the same time, it is hard to imagine how platform companies would be transparent about reduction policies. How does one measure or document reduction? What should the reduced visibility of a piece of content be compared to? Because the circulation of a piece of content tomorrow depends in part on who happens to see it today, both amplification and reduction have a cumulative effect. For the same reason, we cannot know how that content would have traveled had it not been reduced. There is no "normal" reach of content to measure against. How something might have performed—versus how it did—depends on its quality, who saw and liked it early, whether it got traction and how much, what it was up against on the platform, what news was breaking at the same time, how the machine learning algorithms rated it, what criteria and thresholds they were tuned to at that moment, and on and on (Magalhães & Yu, 2022).

Algorithms in practice are frustratingly and impossibly opaque (Burrell, 2016). And that impossibility of being precise about what impact a reduction policy may have had leaves users uncertain about when and to what degree they have been subject to an intervention, allows platforms to both overstate and oversimplify the effects of their algorithms, and may hinder what the law could do to oversee platforms and prevent abuses of this power (Cobbe & Singh, 2019; Helberger, 2020; Heldt, 2020; Keller, 2021).

## Conclusion: Moderation by Other Means

Major platforms, including Facebook, YouTube, Instagram, Twitter, Tumblr, TikTok, LinkedIn, and Reddit have added reduction to their content moderation techniques. They use machine learning classifiers to identify content that is misleading enough, risky enough, and problematic enough to warrant reducing its visibility, by demoting or excluding it from the algorithmic rankings and recommendations—while not going so far as to remove it. They distinguish different aspects of their platforms to ascribe different logics of responsibility and justify different interventions: passive hosting versus active recommending, what users asked for versus what they were presented with. Reduction policies are explained as either responsible oversight of algorithmic systems, or unavoidable

responses to the human impulse toward the sensational: Or, they are not explained publicly at all.

The way YouTube, Facebook, and others have developed and legitimated their reduction policies reveals a great deal about how content moderation now works. This is content moderation by other means, conducted sometimes by other parts of the companies, deployed in different parts of the platform. To address growing concerns about *harm* (and user speech), these platforms are turning to mechanisms designed traditionally to manage *quality* (and consumer satisfaction). But, because platforms have until recently been circumspect about these strategies, and because the public debate has focused so strongly on removal, reduction is rarely included in discussions of content moderation, or of the power of platforms over public discourse. These new strategies should remind us that platforms have long managed content they want less of by making adjustments to their recommendations, rankings, or search results: duplicates, spam, clickbait, engagement bait (Hallinan, 2021), not-safe-for-work (NSFW) content, bots, coordinated inauthentic content (McGregor, 2020). Yet, we tend not to think of any of these as content moderation—even though someone's speech is certainly being constrained, so as to prevent users from encountering what they either don't or shouldn't want.

Whether it is *selecting out* and *selecting for*, through policy or through design, with whatever justification—all of it "moderates" not only what any one user is likely to see but also what society is likely to attend to, take seriously, struggle with, and value. Reducing news content so as to improve the "organic reach" of posts from your friends and family is a form of moderation (Cobbe & Singh, 2019). When Mark Zuckerberg, after the 6 January insurrection at the U.S. Capitol, announced that Facebook would begin testing ways to show less political content in the newsfeed,[34] that is moderation too. Whether a platform intervenes at a single post, or all posts that include a single term, or a machine learning classifier's best guess of which content falls on the wrong side of a rule—or the reduction of an entire category, so as to decrease the likelihood of polarizing, hateful, or misleading content—that's all moderation (Carmi, 2021). Platforms intervene in the circulation of information, culture, and political expression by removing, reducing, personalizing, rewarding, and elevating; these are overlapping and cumulative strategies, both in practice and in effect, and they must be examined together. As Carmi (2020) put it,

> the separation between signal and noise in this context is complicated, as what constitutes a disturbance is decided by multiple actors, and is not restricted to those who create the medium. What needs to be filtered constantly changes because what is considered to be an interference to the business model is also constantly in flux. (p. 186)

If reduction is a form of content moderation, then it must be included in the ongoing debates about platform responsibility. Does it benefit the public, or undermine it, when platforms regularly and quietly reduce what they deem to be misinformation, conspiracy, and "borderline content" violations? We do not know the impact of reduction techniques. We do not know whether that impact differs when what's being reduced is White nationalism, junk news links, explicit sex work, or users struggling with the impulse to harm themselves (Gerrard, 2020). We do not know whether we can trust platforms to engage in these reduction practices thoughtfully, in ways that produce a robust but fairer public sphere. We have little access to who is making these policies and distinctions, and according to what criteria.

In fact, I suspect that reduction policies will have the effect of further normalizing the underlying logics by which social media select and circulate information. Reduction is a corrective mechanism that presumes and reifies the logic of the system it is designed to correct. It prefigures a system that takes *expansion* to be the "natural" impulse, to which constraints must occasionally be applied.[35] When we are fighting about the particular dynamics of virality (Nahon & Hemsley, 2013), we are not asking whether there are other logics of circulation that we should prefer.

If platforms are unavoidably curators, always selecting for and selecting out; if asking them not to be is nonsensical; and if that power to curate might be a more mature, progressive approach than anything they've done until now, then what we must grapple with is a very old problem: private intermediaries with power over the public speech that circulates through them (Gillespie, 2018). What we know, not just from two decades of studying social media but from studying media for many more, is that the power to select is important and unavoidable, it can be exerted in both progressive and destructive ways—but that, again and again, the impact of these selections are felt inequitably. Whatever combination of algorithmic reductions, moderation removals, and design adjustments platforms employ, the questions driving analyses should be: Who is excluded and who is included? Who enjoys the largesse and who bears the constraints? Which groups are further marginalized and which are given center stage? These inquiries must go beyond mere calculations of profit, for sure, but also beyond the First Amendment, beyond unspecified notions of public interest, beyond reassuring notions of openness and meritocracy. We know, and should now take as axiomatic, that universalized principles of fairness cloak and reaffirm systems of inequity (Díaz & Hecht-Felella, 2021; Gerrard & Thornham, 2020; Gray & Stein, 2021; Haimson et al., 2021; Marshall, 2021; Noble, 2018; Southerton et al., 2021; Zolides, 2021). The study of platforms and their interventions, then, should put questions of social inequity front and center, and explain why they always seem to come second to other priorities.

There is a world in which reduction techniques could, at least in theory, be a salve to structural inequities of visibility: if they were used to counter how marginalized communities regularly lose ground amid the proclaimed evenhandedness of intermediaries and in the face of the bad faith tactics of their antagonists. Platforms willing to reduce content that is harmful, abusive, or in bad faith might also reduce the swollen voices of the already amplified, to make more room for the kind of pluralist civic sphere that social media has sometimes promised but never achieved. However, that is probably not what will happen. Unless reduction techniques are premised on a different theory of participation and equity, driven by an ethic of care, and fitted with accountable oversight and public debate, they could further scuttle marginalized communities—as early evidence suggests they are already doing. Marginalized communities have long been "reduced" by the centrism and conservatism of traditional media, their content dismissed as "low quality" because it doesn't look like it is "supposed to." Reduction on social media platforms, for different reasons, could easily do the same.

## Acknowledgements

The author thanks the anonymous reviewers, his excellent colleagues in the Social Media Collective at MSR, and to those who helped him with these ideas or read drafts of this paper, especially Mike Ananny, Robyn Caplan, Zoe Darmé, Fernando Diaz, Stephan Dreyer, Ysabel Gerrard, Amélie Heldt, Kate Klonick, Oliver Haimson, Ashley Mears, and Dylan Mulvin. The author thanks Wendy Chun and the Digital Democracies Institute at Simon Fraser University, Oscar Westlund and the Digital Journalism Research Group at Oslo Metropolitan University, Sally Wyatt and the Dutch Digital Society 2021 conference, Ana Viseu and the Universidade Nova de Lisboa, and Joan Feigenbaum and DIMACS for the opportunities to present this as a work in progress. Finally, the author thanks Rachel Bergmann and Elizabeth Fetterolf for their superb assistance with this research.

## Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

## Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

## ORCID iD

Tarleton Gillespie https://orcid.org/0000-0002-2601-6073

## Notes

1. Samantha Cole, 31 July 2018. "Where did the concept of 'shadow banning' come from?" *Vice*. https://www.vice.com/en_us/article/a3q744/where-did-shadow-banning-comefrom-trump-republicans-shadowbanned
2. Margaux MacColl, 16 September 2020. "TikTok creators say the Creator Fund is killing their views. Some are leaving." *Digital Trends*. https://www.digitaltrends.com/social-media/tiktok-creators-say-creator-fund-is-killing-their-views/
3. Paul Lewis, 2 February 2019. "'Fiction is outperforming reality': How YouTube's algorithm distorts truth." *Guardian*. https://www.theguardian.com/technology/2018/feb/02/how-youtubes-algorithm-distorts-truth
4. Zeynep Tufekci, 10 March 2018. "YouTube, the great radicalizer." *New York Times*. https://www.nytimes.com/2018/03/10/opinion/sunday/youtube-politics-radical.html
5. YouTube, 25 January 2019. "Continuing our work to improve recommendations on YouTube." https://youtube.googleblog.com/2019/01/continuing-our-work-to-improve.html
6. Facebook makes a similar distinction: most newsfeed items are there because of that user's attachments, likes, and clicks that make up their social graph; Facebook sees it as the user's responsibility when problematic content appears there. Facebook bears more responsibility for the "unconnected content," the groups and pages Facebook recommends, separate from the user's social graph. Their reduction efforts began here, with the recommendation of groups, though they have since expanded. Facebook, August 11, 2020. "Our commitment to safety." https://www.facebook.com/business/news/our-commitment-to-safety
7. YouTube, 3 December 2019. "The Four Rs of responsibility, part 2: Raising authoritative content and reducing borderline content and harmful misinformation." https://blog.youtube/inside-youtube/the-four-rs-of-responsibility-raise-and-reduce/
8. Becca Lewis, 27 April 2021. https://twitter.com/beccalew/status/1387125473553960963
9. Facebook had also made passing references to these techniques as far back as maybe 2015, certainly 2017: Erich Owens and Udi Weinsberg, Facebook, 20 January 2015. "Showing fewer hoaxes." https://about.fb.com/news/2015/01/news-feed-fyi-showing-fewer-hoaxes/; Adam Mosseri, Facebook, 6 April 2017. "Working to stop misinformation and false news." https://about.fb.com/news/2017/04/working-to-stop-misinformation-and-false-news/. See also Caplan et al. (2018).
10. Nicholas Thompson, Fred Vogelstein, 12 February 2018. "Inside the two years that shook Facebook—and the World." *Wired*. https://www.wired.com/story/inside-facebook-mark-zuckerberg-2-years-of-hell/
11. Tessa Lyons, Product Manager, Facebook, 22 May 2018. "The Three-part recipe for cleaning up your news feed." https://about.fb.com/news/2018/05/inside-feed-reduce-remove-inform/; Reporting by *TechCrunch* at the time made clear that Instagram had imposed similar policies: Josh Constine, 10 April 2019. "Instagram now demotes vaguely 'inappropriate' content." *TechCrunch*. https://techcrunch.com/2019/04/10/instagram-borderline/
12. Facebook, first published 31 August 2020. "What are recommendations on Facebook?" https://www.facebook.com/help/1257205004624246; Instagram has a matching policy, "What are recommendations on Instagram?" https://www.facebook.com/help/instagram/313829416281232
13. Facebook, "Content distribution guidelines." https://transparency.fb.com/en-gb/features/approach-to-ranking/types-of-content-we-demote/

14. Instagram, 20 July 2021. "Introducing sensitive content control." https://about.fb.com/news/2021/07/introducing-sensitive-content-control/
15. Adam Mosseri, 21 July 2021. https://twitter.com/mosseri/status/1417672062110507008
16. Twitter, 14 July 2020, "Clarifying how we assess misleading information." https://blog.twitter.com/en_us/topics/company/2020/covid-19.html#protecting
17. Twitter, 10 September 2020. "Expanding our policies to further protect the civic conversation." https://blog.twitter.com/en_us/topics/company/2020/civic-integrity-policy-update; Twitter, January 2021, "Civic integrity policy." https://help.twitter.com/en/rules-and-policies/election-integrity-policy
18. LinkedIn, "LinkedIn professional community policies." https://www.linkedin.com/legal/professional-community-policies
19. LinkedIn, "LinkedIn professional community policies" (dated: 8 May 2018; collected by Wayback Archive: 11 November 2019). https://web.archive.org/web/20191111185459/https://www.linkedin.com/help/linkedin/answer/34593?lang=en&trk=homepage-basic_footer-community-guide
20. TikTok, "Community guidelines." https://www.tiktok.com/community-guidelines. This language appears to have been added in December 2020, and has been revised since.
21. Reddit, "Quarantined communities" (dated: February 2021). https://reddit.zendesk.com/hc/en-us/articles/360043069012-Quarantined-Subreddits
22. Mark Zuckerberg, Facebook, 15 November 2018. "Blueprint for content governance and enforcement." https://www.facebook.com/notes/751449002072082/
23. Facebook, 31 August 2020. "Recommendation guidelines." https://about.fb.com/news/2020/08/recommendation-guidelines/
24. This logic reappears in the explanation for Instagram's "sensitive content control" feature. See Instagram, "Introducing sensitive content control."
25. Clive Thompson, 18 September 2020. "YouTube's plot to silence conspiracy theories." *Wired*. https://www.wired.com/story/youtube-algorithm-silence-conspiracy-theories/
26. Jeff Horwitz, 13 September 2021. "Facebook says its rules apply to all. Company documents reveal a secret elite that's exempt." *Wall Street Journal*. https://www.wsj.com/articles/facebook-files-xcheck-zuckerberg-elite-rules-11631541353
27. Covington, P., Adams, J., & Sargin, E. (2016). Deep Neural Networks for YouTube Recommendations. *Proceedings of the 10th ACM Conference on Recommender Systems*. (pp 191-198). doi:10.1145/2959100.2959190
28. This comment, made by Facebook VP Nick Clegg in 2021, suggests that Facebook handles reduction not in the inventory, but the ranking: "there are types of content that might not violate Facebook's Community Standards but are still problematic because users say they don't like them. For these, Facebook reduces their distribution . . . In other words, how likely a post is to be relevant and meaningful to you acts as a positive in the ranking process, and indicators that the post may be problematic (but nonviolating) act as a negative. The posts with the highest scores after that are placed closest to the top of your Feed." However, this was said in of a document that was otherwise largely laying the blame for misinformation at the feet of users, so this may not be a particular careful, thorough, or even good faith explanation. Nick Clegg, March 31, 2021. "You and the algorithm: It takes two to tango." *Medium*. https://nickclegg.medium.com/you-and-the-algorithm-it-takes-two-to-tango-7722b19aa1c2
29. Neil Potts, VP of Public Policy, Facebook, direct communication in Q&A with the author; High Tech Law Institute, Santa Clara School of Law, October 8, 2020. https://law.scu.edu/event/content-policy-and-enforcement-people-policy-and-product/
30. Thompson, "YouTube's plot."
31. Thompson, "YouTube's plot."
32. YouTube, "The Four Rs of Responsibility, Part 2"; Google, "Search Quality Evaluator Guidelines" last updated: 14 October 2020. https://static.googleusercontent.com/media/guidelines.raterhub.com/en//searchqualityevaluatorguidelines.pdf
33. "The Santa Clara Principles: On transparency and accountability in content moderation." 2018 https://santaclaraprinciples.org/
34. Aastha Gupta, Facebook, 10 February 2021. "Reducing political content in news feed." https://about.fb.com/news/2021/02/reducing-political-content-in-news-feed/
35. W. Chun, personal communication, 8 September 2021.

## References


Ananny, M. (2020). Making up political people: How social media create the ideals, definitions, and probabilities of political speech. *Georgetown Law and Technology Review*, *4*(2), 351-366. https://doi.org/10.31219/osf.io/7pd62

Ananny, M., & Crawford, K. (2018). Seeing without knowing: Limitations of the transparency ideal and its application to algorithmic accountability. *New Media & Society*, *20*(3), 973-989. https://doi.org/10.1177/1461444816676645

Are, C. (2021). The shadowban cycle: An autoethnography of pole dancing, nudity and censorship on Instagram. *Feminist Media Studies*. Advance online publication. https://doi.org/10.1080/14680777.2021.1928259

Beutel, A., Chi, E. H., Diaz, F., & Burke, R. (2020). *Responsible recommendation and search systems*. The Web Conference. http://alexbeutel.com/www2020_resprecs/tutorial.pdf

Blunt, D., Duguay, S., Gillespie, T., Love, S., & Smith, C. (2021). Deplatforming sex: A roundtable. *Porn Studies*, *8*(4), 420-438. https://doi.org/10.1080/23268743.2021.2005907

Blunt, D., Wolf, A., Coombes, E., & Mullin, S. (2020). *Posting into the void: Studying the impact of shadowbanning on sex workers and activists*. Hacking/Hustling. https://hackinghustling.org/posting-into-the-void-content-moderation/

Bonini, T., & Gandini, A. (2020). The field as a black box: Ethnographic research in the age of platforms. *Social Media + Society*, *6*(4), 205630512098447. https://doi.org/10.1177/2056305120984477

Bucher, T. (2018). *If . . . then: Algorithmic power and politics*. Oxford University Press.

Burrell, J. (2016). How the machine "thinks": Understanding opacity in machine learning Algorithms. *Big Data & Society*, *3*(1), 205395171562251. https://doi.org/10.1177/2053951715622512

Caldwell, J. T. (2013). Para-industry: Researching Hollywood's blackwaters. *Cinema Journal*, *52*(3), 157-165. https://doi.org/10.1353/cj.2013.0014

Caplan, R. (2019). *Content or context moderation? Artisanal, community-reliant, and industrial approaches*. Data & Society

Research Institute. https://datasociety.net/library/content-or-context-moderation/

Caplan, R., & boyd, d. (2018). Isomorphism through algorithms: Institutional dependencies in the case of Facebook. *Big Data & Society*, *5*(1), 1-12. https://doi.org/10.1177/2053951718757253

Caplan, R., & Gillespie, T. (2020). Tiered governance and demonetization: The shifting terms of labor and compensation in the platform economy. *Social Media + Society*, *6*(2), 205630512093663. https://doi.org/10.1177/2056305120936636

Caplan, R., Hanson, L., & Donovan, J. (2018). *Dead reckoning: Navigating content moderation after "fake news."* Data & Society Research Institute. https://datasociety.net/library/dead-reckoning/

Carmi, E. (2020). *Media distortions: Understanding the power behind spam, noise, and other deviant media*. Peter Lang.

Carmi, E. (2021). "It's not you, Juan, it's us": How Facebook takes over our experience. *Tech Policy Press*. https://techpolicy.press/its-not-you-juan-its-us-how-facebook-takes-over-our-experience/

Chandrasekharan, E., Jhaver, S., Bruckman, A., & Gilbert, E. (2022). Quarantined! Examining the effects of a community-wide moderation intervention on Reddit. *ACM Transactions on Computer-Human Interaction*, *29*(4). https://doi.org/10.1145/3490499

Cobbe, J., & Singh, J. (2019). Regulating recommending: Motivations, considerations, and principles. *European Journal of Law and Technology*, *10*(3). https://doi.org/10.2139/ssrn.3371830

Copland, S. (2020). Reddit quarantined: Can changing platform affordances reduce hateful material Online? *Internet Policy Review*, *9*(4). https://doi.org/10.14763/2020.4.1516

Cotter, K. (2021). "Shadowbanning Is not a thing": Black box gaslighting and the power to independently know and credibly critique algorithms. *Information, Communication & Society*. Advance online publication. https://doi.org/10.1080/1369118X.2021.1994624

DeCook, J. R. (2022). r/WatchRedditDie and the politics of Reddit's bans and quarantines. *Internet Histories*, *6*(1-2), 206-222. https://doi.org/10.1080/24701475.2021.1997179

de Keulenaar, E., Burton, A. G., & Kisjes, I. (2021). Deplatforming, demotion and folk theories of big tech persecution. *Fronteiras—Estudos Midiáticos*, *23*(2), 118-139. https://doi.org/10.4013/fem.2021.232.09

Díaz, Á., & Hecht-Felella, L. (2021). *Double standards in social media content moderation*. Brennan Center for Justice. https://www.brennancenter.org/our-work/research-reports/double-standards-social-media-content-moderation

DiResta, R. (2018, April 11). Up next: A better recommendation system. *Wired*. https://www.wired.com/story/creating-ethical-recommendation-engines/

Ekbia, H., & Nardi, B. (2014). Heteromation and its (dis)contents: The invisible division of labor between humans and machines. *First Monday*, *19*(6). https://firstmonday.org/article/view/5331/4090

Gerrard, Y. (2018). Beyond the hashtag: Circumventing content moderation on social media. *New Media & Society*, *20*(12), 4492-4511. https://doi.org/10.1177/1461444818776611

Gerrard, Y. (2020). The COVID-19 mental health content moderation conundrum. *Social Media + Society*, *6*(3). https://doi.org/10.1177/2056305120948186

Gerrard, Y., & Thornham, H. (2020). Content moderation: Social media's sexist assemblages. *New Media & Society*, *22*(7), 1266-1286. https://doi.org/10.1177/1461444820912540

Gillespie, T. (2016). #trendingistrending: When algorithms become culture. In R. Seyfert & J. Roberge (Eds.), *Algorithmic cultures: Essays on meaning, performance and new technologies* (pp. 52-75). Routledge.

Gillespie, T. (2018). *Custodians of the internet: Platforms, content moderation, and the hidden decisions that shape social media*. Yale University Press.

Goldman, E. (2021). Content moderation remedies. *Michigan Technology Law Review*, *28*(1), 1-59. https://doi.org/10.36645/mtlr.28.1.content

Gorwa, R., Binns, R., & Katzenbach, C. (2020). Algorithmic content moderation: Technical and political challenges in the automation of platform governance. *Big Data & Society*, *7*(1), 205395171989794. https://doi.org/10.1177/2053951719897945

Gray, K. L., & Stein, K. (2021). "We 'said her name' and got Zucked": Black women calling-out the carceral logics of digital platforms. *Gender & Society*, *35*(4), 538-545. https://doi.org/10.1177/08912432211029393

Gray, M. L., & Suri, S. (2019). *Ghost work: How to stop Silicon Valley from building a new global underclass*. Houghton Mifflin Harcourt.

Haimson, O. L., Delmonaco, D., Nie, P., & Wegner, A. (2021). Disproportionate removals and differing content moderation experiences for conservative, transgender, and black social media users: Marginalization and moderation gray areas. *Proceedings of the ACM on Human-Computer Interaction*, *5*, 1-35. https://doi.org/10.1145/3479610

Hallinan, B. (2021). Civilizing infrastructure. *Cultural Studies*, *35*(4-5), 707-727. https://doi.org/10.1080/09502386.2021.1895245

Helberger, N. (2020). The political power of platforms: How current attempts to regulate misinformation amplify opinion power. *Digital Journalism*, *8*(6), 842-854. https://doi.org/10.1080/21670811.2020.1773888

Heldt, A. (2020). Borderline speech: Caught in a free speech limbo? *Internet Policy Review*. https://policyreview.info/articles/news/borderline-speech-caught-free-speech-limbo/1510#:~:text=Why%20call%20it%20a%20free,content%20to%20be%20fully%20permissible

Keller, D. (2021, June 8). *Amplification and its discontents*. Knight First Amendment Institute.

Maddox, J., & Malson, J. (2020). Guidelines without lines, communities without borders: The marketplace of ideas and digital manifest destiny in social media platform policies. *Social Media + Society*, *6*(2), 205630512092662. https://doi.org/10.1177/2056305120926622

Magalhães, J. C., & Yu, J. (2022). Mediated visibility and recognition: A taxonomy. In A. Brighenti (Ed.), *The new politics of visibility: Spaces, actors, practices, and technologies in the visible* (pp. 72-99). Intellect.

Marshall, B. (2021). *Algorithmic misogynoir in content moderation practice*. Heinrich Böll Stiftung. https://www.boell.de/en/2021/06/21/algorithmic-misogynoir-content-moderation-practice

Matamoros-Fernández, A. (2017). Platformed racism: The mediation and circulation of an Australian race-based controversy on Twitter, Facebook and YouTube. *Information, Communication & Society*, *20*(6), 930-946. https://doi.org/10.1080/1369118X.2017.1293130

McGregor, S. (2020, September 17). What even is "coordinated inauthentic behavior" on platforms? *Wired*. https://www.wired.com/story/what-even-is-coordinated-inauthentic-behavior-on-platforms/

McKelvey, F., & Hunt, R. (2019). Discoverability: Toward a definition of content discovery through platforms. *Social Media + Society*, *5*(1), 205630511881918. https://doi.org/10.1177/2056305118819188

Monahan, T., & Fisher, J. (2015). Strategies for obtaining access to secretive or guarded organizations. *Journal of Contemporary Ethnography*, *44*(6), 709-736. https://doi.org/10.1177/0891241614549834

Nahon, K., & Hemsley, J. (2013). *Going viral*. Polity.

Nicholas, G. (2022). *Shedding light on shadowbanning*. Center for Democracy and Technology. https://cdt.org/insights/shedding-light-on-shadowbanning/

Noble, S. U. (2018). *Algorithms of oppression: How search engines reinforce racism*. NYU Press.

Petre, C., Duffy, B. E., & Hund, E. (2019). "Gaming the system": Platform paternalism and the politics of algorithmic visibility. *Social Media + Society*, *5*(4), 205630511987999. https://doi.org/10.1177/2056305119879995

Pilipets, E., & Paasonen, S. (2022). Nipples, memes, and algorithmic failure: NSFW critique of Tumblr censorship. *New Media & Society*, *24*(6), 1459-1480. https://doi.org/10.1177/1461444820979280

Rieder, B., & Hofmann, J. (2020). Towards platform observability. *Internet Policy Review*, *9*(4). https://doi.org/10.14763/2020.4.1535

Roberts, S. T. (2019). *Behind the screen: Content moderation in the shadows of social media*. Yale University Press.

Ruckenstein, M., & Turunen, L. (2019). Re-humanizing the platform: Content moderators and the logic of care. *New Media & Society*, *22*(6), 1026-1042. https://doi.org/10.1177/1461444819875990

Savolainen, L. (2022). The shadow banning controversy: Perceived governance and algorithmic folklore. *Media, Culture & Society*, *44*(6), 1091-1109. https://doi.org/10.1177/01634437221077174

Seaver, N. (2016). *After secrecy*. Society for Ethnomusicology.

Siapera, E., & Viejo-Otero, P. (2021). Governing hate: Facebook and digital racism. *Television & New Media*, *22*(2), 112-130. https://doi.org/10.1177/1527476420982232

Singh, A., Halpern, Y., Thain, N., Christakopoulou, K., Chi, E., Chen, J., & Beutel, A. (2020). *Building healthy recommendation sequences for everyone: A safe reinforcement learning approach. Facctrec Workshop, ACM Recommender Systems (Recsys)*. http://www.ashudeepsingh.com/publications/facctrec2020_singh_et_al.pdf

Singh, S. (2019). *Rising through the ranks: How algorithms rank and curate content in search results and on news feeds*. New American Foundation.

Smith, S., Haimson, O. L., Fitzsimmons, C., & Brown, N. E. (2021). *Censorship of marginalized communities on Instagram*. Salty Algorithmic Bias Collective. https://saltyworld.net/exclusive-report-censorship-of-marginalized-communities-on-instagram-2021-pdf-download/

Souleles, D. (2021). How to think about people who don't want to be studied: Further reflections on studying up. *Critique of Anthropology*, *41*(3), 206-226. https://doi.org/10.1177/0308275X211038045

Southerton, C., Marshall, D., Aggleton, P., Rasmussen, M. L., & Cover, R. (2021). Restricted modes: Social media, content classification and LGBTQ sexual citizenship. *New Media & Society*, *23*(5), 920-938. https://doi.org/10.1177/1461444820904362

Starr, P. (2020, January 22). How money now tries to bury the truth. *The American Prospect*. https://prospect.org/power/how-money-now-tries-to-bury-the-truth/

Steiger, M., Bharucha, T. J., Venkatagiri, S., Riedl, M. J., & Lease, M. (2021). The psychological well-being of content moderators: The emotional labor of commercial moderation and avenues for improving support. In Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems (pp. 1-14). Association for Computing Machinery.

Suzor, N. P., West, S. M., Quodlin, A., & York, J. (2019). What do we mean when we talk about transparency? Toward meaningful transparency in commercial content moderation. *International Journal of Communication*, *13*, 1526-1543.

Sybert, J. (2021). The demise of #NSFW: Contested platform governance and Tumblr's 2018 adult content ban. *New Media & Society*. Advance online publication. https://doi.org/10.1177/1461444821996715

Tiidenberg, K., Hendry, N. A., & Abidin, C. (2021). *Tumblr*. Polity Press.

Vonderau, P. (2014). Industry proximity. *Media Industries*, *1*(1), 69-74.

West, S. M. (2018). Censored, suspended, shadowbanned: User interpretations of content moderation on social media platforms. *New Media & Society*, *20*(11), 4366-4383. https://doi.org/10.1177/1461444818773059

Zolides, A. (2021). Gender moderation and moderating gender: Sexual content policies in Twitch's community guidelines. *New Media & Society*, *23*(10), 2999-3015. https://doi.org/10.1177/1461444820942483

## Author Biography

Tarleton Gillespie (PhD, University of California, San Diego) is a senior principal researcher at Microsoft Research New England and an affiliated associate professor in the Department of Communication and Department of Information Science at Cornell University. He is the author of *Wired Shut: Copyright and the Shape of Digital Culture* (MIT Press, 2007), co-editor of *Media Technologies: Essays on Communication, Materiality, and Society* (MIT, 2014), and author of *Custodians of the Internet: Platforms, Content Moderation, and the Hidden Decision that Shape Social Media* (Yale, 2018).