Reduction / Borderline content / Shadowbanning

Tarleton Gillespie

introduction

The broader public debate about content moderation has overwhelmingly focused on *removal*: social media platforms deleting content and suspending users—or opting not to. This is not surprising. Debating whether a platform should ban the sitting president of the United States is delicious, and a potent way to ask about the platform's influence. The First Amendment aspect lures journalists, pundits, policymakers, and researchers alike. And cries of "censorship" have the most traction when the content has been completely deleted or a user has been permanently banned.

But while removal may be the most visible response, it is by no means the only remedy available.[1] Many platforms, including Facebook, YouTube, Instagram, Twitter, Tumblr, TikTok, LinkedIn, and Reddit, also identify content that they deem not quite bad enough to remove, but bad enough. As an example, In July 2021, after U.S. President Biden accused Facebook of "killing people"[2] by allowing misinformation about COVID vaccines to proliferate, Facebook pointedly countered:

> …when we see misinformation about COVID-19 vaccines, we take action against it. Since the beginning of the pandemic we have removed over 18 million instances of COVID-19

[1] *See* Eric Goldman, *Content Moderation Remedies*, 28 MICH. TECH. L. REV. 1 (2021).

[2] Eugene Scott & Rachel Lerman, *Biden Clarifies Comments About Facebook "Killing People" with Vaccine Misinformation*, WASH. POST (July 19, 2021), https://www.washingtonpost.com/politics/2021/07/19/biden-facebook-misinformation.

> misinformation. We have also labeled and reduced the visibility (emphasis mine) of more than 167 million pieces of COVID-19 content debunked by our network of fact-checking partners so fewer people see it...[3]

Reducing the visibility of risky, misleading, or salacious content is becoming a commonplace and large-scale part of platform governance. Using machine learning classifiers, platforms identify content that is misleading enough, risky enough, problematic enough to warrant reducing its visibility by demoting or excluding it from the algorithmic rankings and recommendations. The offending content remains on the site, still available to the user who can find it directly; but the platform limits the conditions under which it circulates: how it is offered up as a recommendation, search result, part of an algorithmically-generated feed, or "up next" in users' queues.

In this essay, I will call these "reduction policies." There are several emergent terms for this practice, as I will discuss, and they are all problematic. But the fact that there is not yet a settled industry term is itself revealing. Understandably, platforms are wary of being scrutinized for these reduction policies. Some platforms have not publicly acknowledged them; those that have are circumspect. It is not that they are hidden entirely, but the major platforms are only just beginning to acknowledge these techniques as a significant element of how they now manage problematic content. Consequently, reduction policies remain largely absent from public, policy, and scholarly conversations about content moderation and platform governance.[4]

Borderline Content

Let me start with YouTube. The company announced what it calls its "borderline content" policy in January 2019, though the practice had already been in place for a few months or more:

> 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

[3] Guy Rosen, *Moving Past the Finger Pointing*, FACEBOOK (July 17, 2021), https://about.fb.com/news/2021/07/support-for-covid-19-vaccines-is-high-on-facebook-and-growing.

[4] TARLETON GILLESPIE, CUSTODIANS OF THE INTERNET: PLATFORMS, CONTENT MODERATION, AND THE HIDDEN DECISIONS THAT SHAPE SOCIAL MEDIA (2018).

> 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.[5]

Notice that harms are being characterized as on a spectrum: "content that comes close to—but doesn't quite cross the line of—violating our Community Guidelines." YouTube's spatial understanding of harm stakes out a "borderline" just shy of the existing prohibitions.[6]

When Mark Zuckerberg announced a similar policy for Facebook in May 2018,[7] his terminology and justifications were 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…"[8] The announcement included two diagrams (that look mathematical but are not) to explain both the problem and the proposed solution.

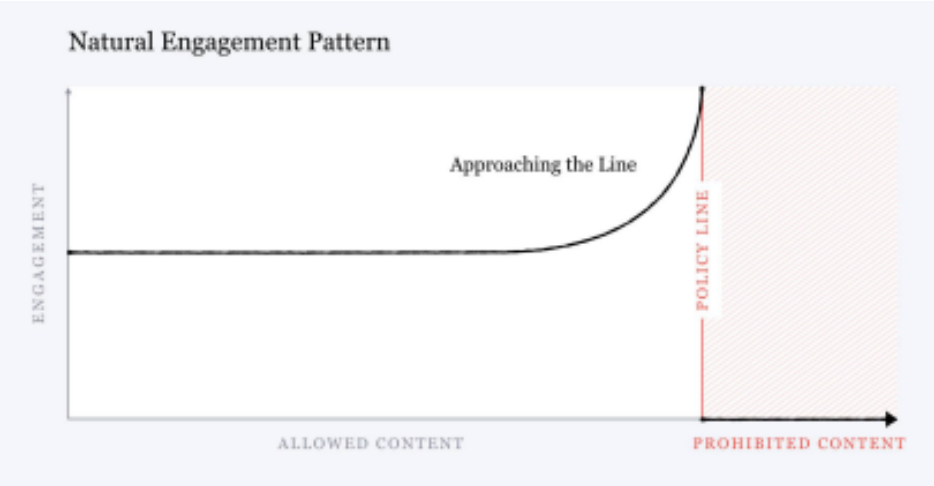
---

[5] The YouTube Team, *Continuing Our Work to Improve Recommendations on YouTube*, YOUTUBE (Jan. 25, 2019), https://blog.youtube/news-and-events/continuing-our-work-to-improve.

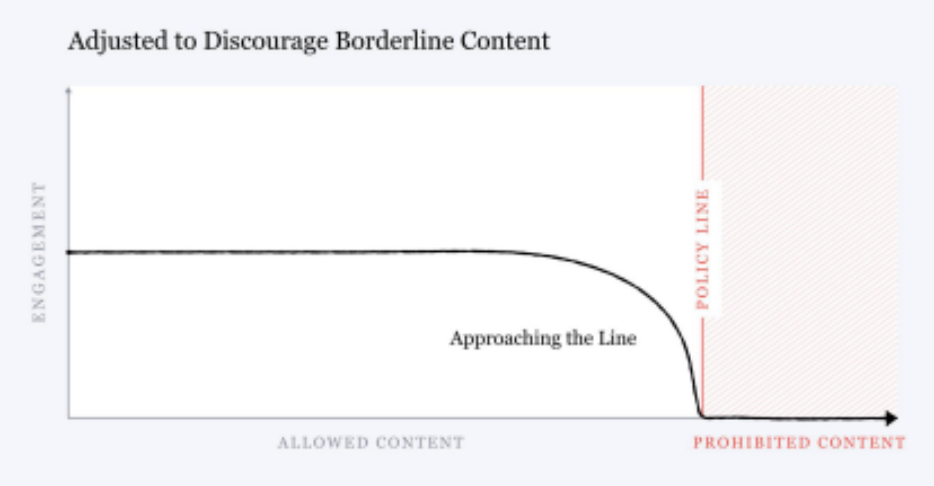
[6] *See* Jessica Maddox & Jennifer Malson, *Guidelines Without Lines, Communities Without Borders: The Marketplace of Ideas and Digital Manifest Destiny in Social Media Platform Policies*, SOC. MEDIA + SOC'Y, Apr.-June 2020, at 1, https://journals.sagepub.com/doi/pdf/10.1177/2056305120926622.

[7] Mark Zuckerberg, *Blueprint for Content Governance and Enforcement*, FACEBOOK (Nov. 15, 2018), https://www.facebook.com/notes/751449002072082. Facebook had also made passing references to these techniques as far back as maybe 2015, certainly 2017. *See* Erich Owens & Udi Weinsberg, *Showing Fewer Hoaxes*, FACEBOOK (Jan. 20, 2015), https://about.fb.com/news/2015/01/news-feed-fyi-showing-fewer-hoaxes; Adam Mosseri, *Working to Stop Misinformation and False News*, FACEBOOK (Apr. 6, 2017), https://about.fb.com/news/2017/04/working-to-stop misinformation-and-false-news; *see also* Robyn Caplan, Lauren Hanson & Joan Donovan, *Dead Reckoning: Navigating Content Moderation After "Fake News,"* DATA & SOC'Y RSCH. INST. (Feb. 2018), https://datasociety.net/pubs/oh/DataAndSociety_Dead_Reckoning_2018.pdf.

[8] Tessa Lyons, *The Three-Part Recipe for Cleaning Up Your News Feed*, FACEBOOK (May 22, 2018), 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, *Instagram Now Demotes Vaguely "Inappropriate" Content,* TECHCRUNCH (Apr. 10, 2019), https://techcrunch.com/2019/04/10/instagram-borderline.



Zuckerberg hoped the reduction policy would not just level out that surge of demand near the borderline, but to reduce it further, to make demand for questionable content approach zero.[9]



[9] Blake Hallinan, *Civilizing Infrastructure*, 35 CULTURAL STUD. 707 (2021).

Facebook later published an exhaustive "Content Distribution Guidelines," indicating what it now "demotes" in the News Feed.[10] The list reveals how broad this "borderline content" technique has become: Facebook will not recommend "what [users] do and don't like seeing on Facebook" (meaning clickbait, "engagement bait," contest giveaways, and links to deceptive or malicious sites); low-quality and inaccurate content (including misinformation, unoriginal or repurposed content, and news that's unclear about its provenance); and borderline content near to violating any of Facebook's community standards.[11]

Twitter,[12] LinkedIn,[13] and TikTok[14] have similar reduction strategies already in place, though they have been less vocal about them. And depending on how we broaden the definition, other platforms are engaged in strategies that have at least a family resemblance, including Tumblr's hashtag blocking,[15] Instagram's "sensitive content control,"[16] and

---

[10] *Content Distribution Guidelines*, FACEBOOK, https://transparency.fb.com/en-gb/features/approach-to ranking/types-of-content-we-demote (last visited Mar. 24, 2022).

[11] *Id.*

[12] *Clarifying how we assess misleading information*, TWITTER (July 14, 2020), https://blog.twitter.com/en_us/topics/company/2020/covid-19.html#protecting.

[13] *LinkedIn Professional Community Policies*, LINKEDIN, https://www.linkedin.com/legal/professional-community-policies (last visited Nov. 23, 2021).

[14] *Community Guidelines*, TIKTOK, https://www.tiktok.com/community-guidelines (last visited Mar. 24, 2022). This language appears to have been added in December 2020.

[15] Before it restricted sexual content in 2018, Tumblr used to limit the circulation of explicit content by refusing to serve up search results to explicit queries. Users could post pornographic images and could even tag them with a term like "#porn"—but if a user searched for "#porn" no results would be returned. *See* Tarleton Gillespie, *Tumblr, NSFW Porn Blogging, and the Challenge of Checkpoints*, CULTURE DIGITALLY (July 26, 2013), https://culturedigitally.org/2013/07/tumblr-nsfw-porn-blogging-and-the-challenge-of-checkpoints. This hashtag-blocking approach, like the borderline content techniques, demarcate between hosting content and offering it up in search results or recommendations, allowing one while restricting the other. *See also* KATRIN TIIDENBERG, ET.AL. TUMBLR (2021).

[16] In July 2021, Instagram introduced "sensitive content control," giving individual users the ability to adjust how much sensitive content should be filtered from the "explore" recommendations the platform offers. While the announcement emphasized the agency users are being offered, the very fact that users can now "allow," "limit (default)," or "limit even more" how much sensitive content is recommended revealed that such content already was being reduced already. *Introducing sensitive content control*, INSTAGRAM (July 20, 2021), https://about.instagram.com/blog/announcements/introducing-sensitive-content-control. The company did not immediately specify what counts as sensitive; Instagram head Adam Mosseri later indicated that the intervention focuses on "sexually suggestive, firearm, and drug-related content"

Reddit's quarantine policy.[17]

YouTube and Facebook call these their "borderline content" policies, and the term has begun to circulate more widely in the discussion of platform policies. But I am reluctant to adopt this term. First, the spatial metaphor it implies is misleading; it treats complex sociocultural behavior as if it can be mapped on a single line, from acceptable to unacceptable, and imagines a rule as a singular line cleanly intersecting it. It also plays up the similarity between limiting the visibility of content and removing it, as if what they are doing is "almost" content removal. That is misleading. These policies have much more in common with the array of decisions platforms make about what to emphasize and what to disregard—the "sorting for" rather than the "sorting out."

More importantly, the term "borderline" has connotations I do not particularly want to reify by affirming its use as a pejorative, to describe disputably problematic online content, as deemed so by social media platforms fumbling for their new sense of responsibility amid the misinformation and conspiracy that courses through their systems. First, decades of scholarship examining the complex political and cultural dynamics of state borders remind us that borders are not just territorial edges, or walls to be policed.[18] They are sites at which sovereign power is exercised, and are by no means stable, singular, or clear: "Borders are relational in the sense that they are produced, reproduced, and transformed in diverging

---

and was separate from other efforts to reduce misinformation or self-harm. Adam Mosseri (@mosseri), TWITTER (July 21, 2021, 10:25 PM) https://twitter.com/mosseri/status/1417672062110507008.

[17] Reduction is just one part of Reddit's quarantine policy, but the effect is similar. Users who seek out a quarantined subreddit first encounter a warning page, requiring they opt in if they want to continue. The quarantined subreddit can generate no revenue. But also, no posts from within that subreddit will appear on the front page of Reddit unless the user is already a subscriber, and they will not be returned among search or recommendation results. It's worth noting that Reddit is more explicit and transparent about their quarantines, meaning this reduction is much more overt than what YouTube and Facebook are doing – though Reddit users outside the quarantined subreddit may not know why some things aren't bubbling up on their front page anymore. Still, the reasoning behind quarantining a problematic subreddit sounds a lot like Facebook's and YouTube's: "to prevent its content from being accidentally viewed by those who do not knowingly wish to do so or viewed without appropriate context." *Quarantined Communities*, REDDIT, https://reddit.zendesk.com/hc/en-us/articles/360043069012-Quarantined-Subreddits (last accessed Nov. 23, 2021); see also Simon Copland, *Reddit Quarantined: Can Changing Platform Affordances Reduce Hateful Material Online?*, 9 INTERNET POL'Y REV. (2020), https://policyreview.info/articles/analysis/reddit-quarantined-can-changing-platform-affordances-reduce-hateful-material.

[18] Corey Johnson et al., *Interventions on rethinking "the border" in border studies*, 30 POL. GEOGRAPHY 61 (2011).

social relations and networks."[19] The "borderlands" that emerge around borders are spaces of flow permeable to people, resources, and ideas—places of historical and personal pain[20] and cultural spaces meaningful to the people that struggle to inhabit them.[21] Perhaps the use of the term "borderline content" is in fact more accurate than was intended, by the sovereign powers that are YouTube and Facebook. The casual shorthand use of the term by Facebook and YouTube risks making the same missteps: treating an artificial line as natural, treating the restriction of those near it as politically obvious and unproblematic, and failing to see that these are borderlands in which practices meaningful to those who inhabit it are taking place, despite the exercise of power being imposed.

The term also echoes "borderline" mental health conditions, a characterization that the mental health community has been attempting to move away from. The fields of psychology and psychiatry have begun discarding the term "borderline personality disorder," concerned that its original meaning (displaying aspects of both neurosis and psychosis) is inaccurate, and that the term has a pejorative connotation implying a flaw in the subject's personality ;[22] feminist critics have noted how the vague diagnosis is applied substantially more often to women than men, reminiscent of other discarded diagnoses like "hysteria";[23] critics of its use in common parlance and media representations condemn the stigma it places on neurodiversity.[24] Some worry that the way it implies that someone is "almost" mentally ill may seed doubt about those who suffer from it, and research funding less forthcoming.[25] But the term lingers in popular discourse,

---

[19] Anssi Paasi, Commentary, *Border Studies Reanimated: Going Beyond the Territorial/Relational Divide*, 44 ENV'T & PLAN. A 2303 (2012).

[20] JASON DE LEON, THE LAND OF OPEN GRAVES: LIVING AND DYING ON THE MIGRANT TRAIL (2015).

[21] JOSÉ DAVID SALDÍVAR, BORDER MATTERS: REMAPPING AMERICAN CULTURAL STUDIES (1997).

[22] *See* Peter Tyrer, *Why Borderline Personality Disorder Is Neither Borderline nor a Personality Disorder*, 3 PERSONALITY & MENTAL HEALTH 86 (2009).

[23] *See* JANET WIRTH-CAUCHON, WOMEN AND BORDERLINE PERSONALITY DISORDER: SYMPTOMS AND STORIES (2000).

[24] *See* Valéry Brousseau, *Why We Need Better Representation of Borderline Personality Disorder*, NAT'L ALL. ON MENTAL ILLNESS (June 2021), https://www.nami.org/Blogs/NAMI-Blog/June-2021/Why-We-Need-Better-Representation-of-Borderline-Personality-Disorder; Johnson et al., *supra* note 18.

[25] Jayashri Kulkarni, *Borderline Personality Disorder Is a Hurtful Label for Real Suffering –Time We Changed It*, THE CONVERSATION (July 20, 2015, 4:11 PM), https://theconversation.com/borderline-personality-disorder-is-a-hurtful-label-for-real-suffering-time-we-changed-it-41760.

retaining both the pejorative that clings to so many mental health terms, and the quiet suggestion that it is not a "real" condition.

"Shadowbanning"

When a video is removed or a user suspended, traces of that removal remain: the user is alerted; the video is missing; there is a terse explanation where the deleted tweet used to be. For all the concerns about censorship that attend to the removal of content, at least the intervention can be seen, and potentially held accountable. But 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, forwarded, yet it seems to not have earned the audience or reach that it might have otherwise. This uncertainty leaves users grasping for explanations, and it is part of why many users are so suspicious that murky machinations are at work under the hood at these platforms.[26] Some frustrated users, suspicious that platforms are taking invisible action against them, have begun to take note, developing folk theories as to what may be happening and why, implementing homegrown techniques for proving that their content is suppressed, and in some cases attempting to document these interventions. The term critics have most often adopted for such interventions is "shadowbanning." Most vocally, some in the sex work community have accused platforms of shadowbanning.[27] Similar critiques have emerged from other marginalized communities.[28]

---

[26] *See* Robyn Caplan & Tarleton Gillespie, *Tiered Governance and Demonetization: The Shifting Terms of Labor and Compensation in the Platform Economy*, SOC. MEDIA + SOC'Y, April-June 2020, at 1, https://journals.sagepub.com/doi/pdf/10.1177/2056305120936636; Emillie de Keulenaar, Anthony Glyn Burton & Ivan Kisjes, *Deplatforming, Demotion and Folk Theories of Big Tech Persecution*, 23 FRONTEIRAS – ESTUDOS MIDIÁTICOS 118 (2021); Sarah Myers West, *Censored, Suspended, Shadowbanned: User Interpretations of Content Moderation on Social Media Platforms*, 20 NEW MEDIA & SOC'Y 4366 (2018).

[27] *See* Carolina Are, *The Shadowban Cycle: An Autoethnography of Pole Dancing, Nudity and Censorship on Instagram*, FEMINIST MEDIA STUD. (forthcoming 2022), https://www.tandfonline.com/doi/full/10.1080/14680777.2021.1928259; Danielle Blunt et al., *Deplatforming Sex: A Roundtable*, 8 PORN STUD. 420 (2021); Danielle Blunt et al., *Posting into the Void: Studying the Impact of Shadowbanning on Sex Workers and Activists*, HACKING//HUSTLING (2020), https://hackinghustling.org/posting-into-the-void-content-moderation/.

[28] *See* Carolina Are, *How Instagram's Algorithm is Censoring Women and Vulnerable Users but Helping Online Abusers*, 20 FEMINIST MEDIA STUD. 741 (2020); de Keulenaar et al., *supra* note 22; Oliver L. Haimson et al., *Disproportionate Removals and Different Content Moderation Experiences for Conservative, Transgender, and Black Social Media Users: Marginalization and Moderation Gray Areas*, 5 PROC. OF THE ACM ON HUMAN-COMPUT. INTERACTION,

Reports like *Posting into the Void* collect evidence from members of the community that their posts are being constrained in some way.[29] But how to prove it? Some evidence offered is anecdotal—users suspicious that their content is not traveling as far as it used to. Sometimes users will point to financial evidence, that revenue from ad-sharing programs like YouTube's has diminished.[30] Some will confirm that their post is being suppressed by asking users to find it and comment on it. All of these are workarounds, attempting to address the fact that it is extremely difficult to document or measure reduction: what is the reduced visibility of a piece of content measured against? Because the circulation and visibility of a piece of content tomorrow depends on who happens to see it today, reduction has a cumulative effect; for the same reason, it is difficult for a user to know how that content would have travelled had it not been reduced. There is no "normal" reach of content; how it *might* have performed, and how it did, depends on its quality, who saw or forwarded 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, and on and on.

It is clear that these critics are right: whether their particular post was affected, platforms are suppressing the circulation of some content in ways that are difficult to identify. But here too, I have concerns about the term. I agree with Cotter[31] that in using the term "shadowbanning," critics may be inadvertently offering platforms ways of avoiding accountability. The word originated in early online communities and bulletin board systems.[32] It referred to a very specific technique where a moderator could make an offending user invisible to every other user, while to the offender it appeared as everything was functioning normally. The only way the offender might discover they were shadowbanned is if they noticed that they were getting no reactions from anyone else.

The range of "borderline content" interventions on today's platforms

---

No. 466, (2021), https://dl.acm.org/doi/pdf/10.1145/3479610; Shakira Smith et al., *Censorship of Marginalized Communities on Instagram*, SALTY ALGORITHMIC BIAS COLLECTIVE (Oct. 2021), https://saltyworld.net/exclusive-report-censorship-of-marginalized-communities-on-instagram-2021-pdf-download/.

[29] Blunt et al., *supra* note 27.

[30] *See* Caplan & Gillespie, *supra* note 26.

[31] Kelley Cotter, *"Shadowbanning Is Not a Thing": Black Box Gaslighting and the Power to Independently Know and Credibly Critique Algorithms*, INFO., COMMC'N & SOC'Y (forthcoming 2022), https://www.tandfonline.com/doi/full/10.1080/1369118X.2021.1994624.

[32] Samantha Cole, *Where Did the Concept of "Shadow Banning" Come From?*, VICE (July 30, 2018), https://www.vice.com/en_us/article/a3q744/where-did-shadow-banning-comefrom-trump-republicans shadowbanned.

are much subtler: a post could be throttled so as to be recommended less; or only to certain kinds of users, such as to followers only; or for a shorter time. When critics lump these under "shadowbanning," it makes semantic sense: they are interventions surreptitiously made by the platform, leaving the user thinking they are participating normally, while other users see less of them. However, because these interventions do not all match the more specific, original meaning of shadowbanning, platforms can answer critics by saying that "we do not shadowban"—as some platforms have[33]—even as they elsewhere admit to having similar policies in place.

"Reduction"

I prefer the term "reduction" policies. Reduction is a term some platforms user, so it is a good candidate as a term of art. It avoids the baggage of the word "borderline" and includes a wider array of techniques than "shadowbanning." But most importantly, it better captures the fundamental orientation beneath what is being done, how it is being done, and what concerns it is responding to. Reduction policies are the flipside to charges that platforms algorithmically "amplify" problematic content and should bear some responsibility for doing so.[34] The two are conceptual twins; if platforms amplify, perhaps they can or should also reduce. And amplification and reduction work in similar ways, just towards opposite ends.

Recommendation is a central component of social media platforms, but it is driven by a very different set of concerns and priorities than content moderation: while trust and safety teams *select out* what is least appealing, the teams that manage recommender systems and newsfeeds *select for* what is most appealing.[35] Their north star is engagement, usually measured by the time users spend on the platform, the number and types of actions taken, and other measures of satisfaction.[36] Their primary

---

[33] *See* Are, *supra* note 27; Cotter, *supra* note 31.

[34] Joe Whittaker et al., *Recommender Systems and the Amplification of Extremist Content*, 10 INTERNET POL'Y REV. (2021), https://policyreview.info/pdf/policyreview-2021-2-1565.pdf.

[35] *See* Tarleton Gillespie, *The Relevance of Algorithms*, *in* MEDIA TECHNOLOGIES: ESSAYS ON COMMUNICATION, MATERIALITY, AND SOCIETY 167 (Tarleton Gillespie, Pablo J. Boczkowski & Kirsten A. Foot eds., 2014); Tarleton Gillespie, *#trendingistrending: When Algorithms Become Culture*, *in* ALGORITHMIC CULTURES: ESSAYS ON MEANING, PERFORMANCE AND NEW TECHNOLOGIES 52 (Robert Seyfert & Jonathan Roberge eds., 2016).

[36] *See* TAINA BUCHER, IF…THEN: ALGORITHMIC POWER AND POLITICS (2018); Sandana Singh, *Rising Through the Ranks: How Algorithms Rank and Curate Content in Search Results and on News Feeds*, NEW AM. (Oct. 21, 2019), https://www.newamerica.org/oti/reports/rising-through-ranks/.

technique is to collect signals, both about the user and about all the available content in the corpus, to produce a personalized feed of content that will be maximally appealing. Generally, these signals indicate some aspect of the content understood to be positive, or valuable: 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. With reduction techniques, the calculation of what to recommend now includes a negative signal, indicating that a particular piece of content should not be considered relevant to this particular user.

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 to discover that, to accomplish this, most platforms 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 has already been evaluated as problematic by human raters.[37] Ideally, the judgments made by the human raters will be approximated by the machine learning classifier, which can then make the same value judgment over and over on millions of pieces of content. YouTube management has generally framed their reduction policies as an acknowledgement of a growing responsibility.[38] Facebook suggests that it is unavoidable: to the left of every line is a bubble of demand for the sensational and illicit that someone will fulfill.[39] But it is easy to imagine other, less noble reasons for not simply removing this problematic content. First, reduction is less politically risky than removal. Given the recent political climate, platforms fear reprisals from conservative critics who air their outrage whenever their posts are

[37] *See* Hamid Ekbia & Bonnie Nardi, *Heteromation and Its (Dis) Contents: The Invisible Division of Labor between Humans and Machines*, 19 FIRST MONDAY (2014), https://firstmonday.org/article/view/5331/4090; Robert Gorwa, Reuben Binns & Christian Katzenbach, *Algorithmic Content Moderation: Technical and Political Challenges in the Automation of Platform Governance*, BIG DATA + SOC'Y, Jan.-June, at 1 (2020), https://journals.sagepub.com/doi/pdf/10.1177/2053951719897945; MARY L. GRAY & SIDDHARTH SURI, GHOST WORK: HOW TO STOP SILICON VALLEY FROM BUILDING A NEW GLOBAL UNDERCLASS (2019); SARAH T. ROBERTS, BEHIND THE SCREEN: CONTENT MODERATION IN THE SHADOWS OF SOCIAL MEDIA (2019).

[38] Neal Mohan, *Perspective: Tackling Misinformation on YouTube*, YOUTUBE (August 25, 2021), https://blog.youtube/inside-youtube/tackling-misinfo; *see also* Clive Thompson, *YouTube's plot to silence conspiracy theories*," WIRED (Sept. 18, 2020), https://www.wired.com/story/youtube-algorithm-silence-conspiracy-theories.

[39] Zuckerberg, *supra* note 7.

removed.[40] 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 seek it out, whether in the form of advertising revenue or data collection, while still answering public concerns by reducing its reach.[41]

Reduction strategies may also be preferable when the types of problematic content platforms face are difficult to identify, in flux, or difficult to police. Reducing without removing means not having to articulate an explicit policy; this gives platforms the flexibility to intervene around quickly emerging phenomena, go after content designed to elude prohibitions, and 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 to changing problems, from the many faces of white nationalism to the evasive tactics of pro-ana users, to the constantly evolving QAnon conspiracy. In a less flattering light, reduction also avoids accountability, as the interventions themselves are hard to spot, and are not—yet—reported as part of the platform's transparency obligations.

If the judgment of what is most worthwhile can serve as a means to reduce what is least worthwhile, then reduction policies are content moderation by other means. And we probably need to expand our definition even further. Reducing news content so as to improve the "organic reach" of posts from your friends and family is a form of moderation.[42] When Mark Zuckerberg, after the January 6 insurrection at the U.S. Capitol, announced that Facebook would begin testing ways to show less political content in the newsfeed,[43] 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

---

[40] Jeff Horwitz, *Facebook Says Its Rules Apply to All. Company Documents Reveal a Secret Elite That's*

Exempt., WALL ST. J. (Sept. 13, 2021, 10:21 AM), https://www.wsj.com/articles/facebook-files-xcheck-zuckerberg-elite-rules 11631541353.

[41] *See* Ariadna Matamoros-Fernández, *Platformed Racism: The Mediation and Circulation of an Australian Race-Based Controversy on Twitter, Facebook and YouTube*, 20 INFO. COMMC'N & SOC'Y 930 (2017); Eugenia Siapera & Paloma Viejo-Otero, *Governing Hate: Facebook and Digital Racism*, 22 TELEVISION & NEW MEDIA 112 (2021).

[42] Jennifer Cobbe & Jatinder Singh, *Regulating Recommending: Motivations, Considerations, and Principles*, 10 EUR. J.L. & TECH. (2019), https://ssrn.com/abstract=3371830.

[43] Aastha Gupta, *Reducing Political Content in News Feed*, FACEBOOK (Feb. 10, 2021), https://about.fb.com/news/2021/02/reducing-political-content-in-news-feed.

the likelihood of polarizing, hateful, or misleading content—that is moderation, too.[44] Selecting out and selecting for, through policy and through design, with whatever justification, all of it "moderates" not only what any one user is likely to see, but what society is likely to attend to, take seriously, struggle with, and value.[45]

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. If reduction is a form of content moderation, then it must be included in the ongoing debates about platform responsibility. If platforms are responsible for amplifying problematic content, then reduction may be the most mature response.[46] But 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? What is the impact of reduction techniques, and does that impact differ when what is being reduced is white nationalism, junk news links, explicit sex work, or users struggling with the impulse to harm themselves?[47] Can we trust platforms to engage in these reduction practices thoughtfully, in ways that produce a robust but fairer public sphere? Who is making these policies and distinctions, and according to what criteria?

To begin to answer these questions, platforms must be more transparent and when and where they reduce content, and specifically how it works in their recommendation algorithms. Independent researchers need access to the platform to study the downstream effects for content that is reduced in various ways. Policymakers need to take into consideration not only what platforms remove, but what they reduce. And both platforms and critics need to be humbler about admitting that, at this point in the concern about online harms, society may be clamoring for gatekeepers

---

[44] Elinor Carmi, "*It's Not You, Juan, It's Us": How Facebook Takes Over Our Experience*, TECH POL'Y PRESS (Jan. 28, 2021), https://techpolicy.press/its-not-you-juan-its-us-how-facebook-takes-over-our-experience.

[45] Caitlin Petre, Brooke Erin Duffy & Emily Hund, *"Gaming the System": Platform Paternalism and the Politics of Algorithmic Visibility*, Soc. Media + Soc'y, Oct.-Dec. 2019, at 1, https://journals.sagepub.com/doi/pdf/10.1177/2056305119879995.

[46] Daphne Keller, *Amplification and Its Discontents*, KNIGHT FIRST AMEND. INST. (June 8, 2021), https://knightcolumbia.org/content/amplification-and-its-discontents.

[47] Ysabel Gerrard, *The COVID-19 Mental Health Content Moderation Conundrum*, SOC. MEDIA + SOC'Y, July-Sept. 2020, at 1, https://journals.sagepub.com/doi/pdf/10.1177/2056305120948186.

again – and, that any new gatekeepers must be interrogated to avoid the sins of the old: who is being excluded and included, who enjoys the largesse and who bears the constraints, which groups are given center stage, and which are further marginalized.