Content Moderation Tools to Stop Extremism

By Daniel Byman[*]

September 2022

Content moderation, with a few exceptions like copyright and child sexual abuse material, remains largely voluntary: U.S.-based companies have considerable leeway as to what is allowed on their platforms. However, government and civil society pressure, both legal and political, is growing. In the United States, a range of laws and proposals floating around would require companies to change how they do content moderation. If companies decide to act more aggressively, what are their options?

Technology companies are more active than ever in trying to stop terrorists, white supremacists, conspiracy theorists, and other hateful individuals, organizations, and movements from exploiting their platforms. Analyses often focus on the danger of allowing extremists to leverage the internet, the types of content that should be stopped, the difficult technical issues, and the harm to civil liberties and other problems that arise from over-removal. However, the different content moderation techniques available to most companies receive less attention, even though success depends on employing a robust range of tools.

Content moderation, with a few exceptions like copyright and child sexual abuse material, remains largely voluntary: U.S.-based companies have considerable leeway as to what is allowed on their platforms. However, government and civil society pressure, both legal and political, is growing. Germany's *Netzwerkdurchsetzungsgesetz* law (NetzDG, also referred to as the "Facebook Act"), enacted in 2017, requires services to remove or block access to various types of illegal content in a

[*] Daniel Byman is foreign policy editor of *Lawfare.* He is a senior fellow at the Center for Middle East Policy at the Brookings Institution, where he focuses on counterterrorism and Middle East security. He is also a professor at Georgetown University's School of Foreign Service. His latest book is *Spreading Hate: The Global Rise of White Supremacist Terrorism*. He also consults for Google. He would like to thank Marley Carroll, Chris Meserole, and Alan Rozenshtein for their comments on previous versions of this draft.

short period of time or face large fines.[1] In May 2021, the European Union strengthened its Code of Practice on Disinformation, claiming that fake news and similar problems are "putting people's [lives] in danger."[2] In the United States, a range of laws and proposals floating around would require companies to change how they do content moderation.[3]

If companies decide to act more aggressively, what are their options? Much of the debate centers around whether to remove offensive content or leave it up, ignoring the many options in between. As Tarleton Gillespie notes, "Removal is a blunt instrument, an all-or-nothing determination."[4] Depending in part on the services they offer, platforms often enjoy considerable choice that goes beyond the binary "take down/leave up" approach presented by defenders of free speech principles or those who criticize the platforms for not doing enough.

This paper presents a range of options for technology companies, discussing how they work in practice, their advantages, and their limits and risks. It offers a primer on the many options available and then discusses the numerous trade-offs and limits that affect the various approaches.

The first section of the paper offers a brief background on issues related to content moderation, noting its necessity and providing an overall context. The second section, the bulk of the paper, reviews a range of content moderation options, detailing three approaches—removing content, reducing its distribution, and shaping dialogue—and discussing variants within each approach. The third section discusses tensions with regard to the various options as well as implications for the companies and for content moderation in general.

[1] See text at https://perma.cc/7UCW-AA3A.
[2] European Commission, "Commission Presents Guidance to Strengthen the Code of Practice on Disinformation," May 26, 2021, https://ec.europa.eu/commission/presscorner/detail/en/IP_21_2585.
[3] For a list of efforts, see Chris Riley and David Morar, "Legislative Efforts and Policy Frameworks Within the Section 230 Debate," Brookings TechStream, September 21, 2021, https://www.brookings.edu/techstream/legislative-efforts-and-policy-frameworks-within-the-section-230-debate/.
[4] Tarleton Gillespie, *Custodians of the Internet: Platforms, Content Moderation, and the Hidden Decisions That Shape Social Media* (New Haven, CT: Yale University Press, 2018), 176.

BRIEF BACKGROUND: THE NECESSITY AND DIFFICULTIES OF CONTENT MODERATION

Harmful Speech

Hate speech, incitement, conspiracy theories, and similar content can cause many real-world harms. Dylann Roof, who massacred nine Black worshippers at a church in Charleston, South Carolina, claims that it was Google searches that shaped his views on race.[5] For several years, the FBI has warned that the QAnon conspiracy theory may lead to violence, and indeed it has, with adherents arrested for numerous crimes, particularly violence against those in their families or other intimate circles.[6] False election claims, coronavirus mis- and disinformation, and other dubious content is decreasing faith in democracy, increasing polarization, leading millions to eschew lifesaving vaccinations, and at times inspiring real-world violence.

Even beyond the risk of violence and death, hate speech that remains online can harm the mental health of its victims.[7] Inaction by the platforms can "send a powerful message that targeted group members are second-class citizens," as Danielle Keats Citron and Helen Norton put it.[8] As women and members of minority groups are often the target of harassment, inaction can disproportionately affect certain categories of users who already suffer a range of disadvantages.

---

[5] Rebecca Hersher, "What Happened When Dylann Roof Asked Google for Information About Race?" National Public Radio, January 10, 2017, https://www.npr.org/sections/thetwo-way/2017/01/10/508363607/what-happened-when-dylann-roof-asked-google-for-information-about-race.

[6] Michael A. Jensen and Sheehan Kane, "QAnon-Inspired Violence in the United States, *Behavioral Sciences of Terrorism and Political Aggression* (December 2021), https://doi.org/10.1080/19434472.2021.2013292; Ben Collins, "Local FBI Field Office Warns of 'Conspiracy Theory-Driven Domestic Extremists,'" NBC News, August 1, 2019, https://www.nbcnews.com/tech/tech-news/local-fbi-field-office-warns-conspiracy-theory-driven-domestic-extremists-n1038441.

[7] Koustuv Saha, Eshwar Chandrasekharan, and Munmun De Choudhury, "Prevalence and Psychological Effects of Hateful Speech in Online College Communities," *Proceedings of the 11th ACM Conference on Web Science* (June 2019): 255–64, https://doi.org/10.1145/3292522.3326032.

[8] Danielle Keats Citron and Helen Norton, "Intermediaries and Hate Speech," *Boston University Law Review* 91, no. 4 (2011): 1435–84, at 1441, https://www.bu.edu/law/journals-archive/bulr/volume91n4/documents/CITRONANDNORTON.pdf.

The problem can also easily spiral. Left unchecked, hate speech by one user can increase the likelihood that others on a platform will adopt hate speech.[9] The platforms often risk becoming toxic, driving away potential users and decreasing engagement from existing ones.

The Necessity of Content Moderation

To ensure the reputation and openness of their platforms, all social media companies moderate content, assessing user-generated content to ensure that it complies with the platform's terms of service and community guidelines.[10] Indeed, as Gillespie contends, content moderation is at the heart of the business: Platforms do not usually produce content themselves, but they distinguish their platforms for both users and advertisers by how they rank or display user-generated content.[11] What this means in practice varies by platform. For Google Search, for example, it may be the search ranking system or auto-fill words, while for Twitter it might be "who to follow" or the display of trending subjects. The Facebook news feed displays constantly curated content selected by the platform's algorithm, ranging from status updates and "likes" from people you follow to videos, links, and other activity.

One size, of course, does not fit all with regard to content moderation. Robyn Caplan points out that large platforms like Google must strive for consistency given their geographic reach and global scale, and content moderation decisions are often made outside the language and cultural context in which the comments or other content were made. In contrast, more decentralized platforms like Reddit, which rely on volunteers to run their communities, and more "artisanal" platforms like Discord or Patreon can prioritize local context. Some companies are supported via advertisements and thus have more incentives to host offensive and divisive content that might attract users and drive engagement, while others, such as Vimeo, use a subscription-based approach that attracts more professional users. Companies like Facebook directly or indirectly employ tens of thousands of people to work on content moderation, while smaller platforms like Discord and Patreon have

---

[9] Bing He, Caleb Ziems, Sandeep Soni, Naren Ramakrishnan, Diyi Yang, and Srijan Kumar, "Racism Is a Virus: Anti-Asian Hate and Counterspeech in Social Media During the COVID-19 Crisis," *Proceedings of the 2021 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining* (November 2021): 90–94, https://doi.org/10.1145/3487351.3488324.

[10] Gillespie, *Custodians of the Internet,* 5; see also Miriah Steiger, Timir J. Bharucha, Sukrit Venkatagiri, Martin J. Riedl, and Matthew Lease, "The Psychological Well-Being of Content Moderators: The Emotional Labor of Commercial Moderation and Avenues for Improving Support," *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* no. 341 (May 2021): 1–14, https://doi.org/10.1145/3411764.3445092.

[11] Gillespie, *Custodians of the Internet,* 13.

content moderation staff in the single digits or dozens and have little non–English language capacity.[12]

Content moderation can also happen at different levels of the internet stack, including not only social media platforms like Facebook and YouTube but also cloud service providers, domain registrars, app stores, and others.[13] Actors that are deeper in the internet stack tend not to look at the content they help keep up—content moderation is not at the core of their business—but at times they are compelled to act by the same pressures facing more public-facing platforms. After the 2019 mass shootings in Christchurch, New Zealand, internet service providers there and in Australia blocked access to 8chan and other sites hosting the footage of the killings, and after the 2019 El Paso shooting, the content distribution network provider Cloudflare also denied protection to 8chan (as did a major domain registrar).[14]

Inherent Difficulties

Content moderation is difficult, and some would say impossible, to do well at scale. Facebook makes far more content moderation decisions in a day than the U.S. justice system makes decisions in a year.[15] When content moderation decisions must be scaled to involve millions or hundreds of millions of decisions a day, a small error rate still means large numbers of mistakes in absolute numbers; there will always be an example of a mistake, even if the vast majority of content moderation decisions are sensible.[16] Even if technology companies were to abandon artificial intelligence (AI), which is impossible for many given the scale of what they must do, individual

[12] Robyn Caplan, "Content or Context Moderation? Artisanal, Community-Reliant, and Industrial Approaches," Data and Society Research Institute, Nov. 14, 2018, https://apo.org.au/node/203666.
[13] Joan Donovan, "Navigating the Tech Stack: When, Where and How Should We Moderate Content," Centre for International Governance Innovation, Oct. 28, 2019, https://www.cigionline.org/articles/navigating-tech-stack-when-where-and-how-should-we-moderate-content; Jenna Ruddock and Justin Sherman, "Widening the Lens on Content Moderation," *Joint PIJIP/TLS Research Paper Series* no. 69 (July 2021), https://digitalcommons.wcl.american.edu/research/69.
[14] Ruddock and Sherman, "Widening the Lens on Content Moderation."
[15] Evelyn Douek, "Verified Accountability," The Hoover Institution, Sept. 17, 2019, https://www.hoover.org/research/verified-accountability, 8.
[16] Mike Masnick, "Masnick's Impossibility Theorem: Content Moderation at Scale Is Impossible to Do Well," Techdirt, Nov. 20, 2019, https://www.techdirt.com/2019/11/20/masnicks-impossibility-theorem-content-moderation-scale-is-impossible-to-do-well/.

workers will apply the rules inconsistently, with some removing too much content and others too little.[17]

The scale problems are compounded by the need for speed. If some content is left up, it can rapidly go viral: Taking down a terrorist's video of his attack does less good once it has spread throughout the internet's four corners. Because of this concern, some regulatory efforts, such as the German NetzDG, demand rapid removal. This need for speed, in turn, requires AI or simple decision-making processes that can be applied quickly across a vast range of content.

Another tension is between consistency, especially at a global scale, and local context.[18]A noose, for example, might suggest an imminent threat but can also be used to condemn lynching.[19] In addition, the meaning of a post will vary by its cultural and subcultural context. Thus, keeping up the same content in one part of the world or part of a country while taking it down in another is logical given this variance, though in practice—and especially at scale—it is nearly impossible to do.

Vulnerable people and communities are at risk both from a lack of content moderation and from aggressive content moderation. As Quinta Jurecic points out, women, minorities, and other vulnerable groups are more likely to be harassed and targeted online and are thus in need of protection from the platforms. At the same time, automated content moderation is disproportionately likely to take down their content: Black users, for example, are more likely than white users to find their posts incorrectly labeled as hate speech.[20]

Legitimacy is a key question for content moderation. Although companies have the legal right to remove or shape the content on their sites, most company decisions lack transparency and thus can seem arbitrary or politicized. The idea that a small, unrepresentative, and unelected group of people, usually based in the United States and motivated by profit, controls the huge megaphones of social media sits poorly with many people. This problem is particularly acute for automated moderation. As one analysis argues, algorithmic moderation is necessary, but "these systems remain opaque, unaccountable and poorly understood."[21] In part because of a lack of legitimacy, social media

---

[17] Caplan, "Content or Context Moderation?"
[18] Caplan, "Content or Context Moderation?"
[19] See, for example, tweets such as https://twitter.com/Smokeahontis111/status/1387774190024019990 and https://twitter.com/BlueWave215/status/1377993372581892103.
[20] Quinta Jurecic, "The Politics of Section 230 Reform: Learning From FOSTA's Mistakes," The Brookings Institution, March 1, 2022, https://www.brookings.edu/research/the-politics-of-section-230-reform-learning-from-fostas-mistakes/.
[21] Robert Gorwa, Reuben Binns, and Christian Katzenbach, "Algorithmic Content Moderation: Technical and Political Challenges in the Automation of Platform Governance," *Big Data & Society* 7, no. 1 (January–June 2020): 1–15, https://doi.org/10.1177%2F2053951719897945.

companies are under fire from both sides of the political spectrum: In the United States, conservatives think the companies moderate far too much, while liberals believe they are moderating far too little.

The government role in content moderation is limited, at least in the United States. Moderating the sheer volume of content is itself a massive task, requiring a huge expansion for any would-be agency to oversee it. More fundamentally, the First Amendment would stop the U.S. government from requiring companies to block or demote most forms of hateful speech, though other governments with less restrictive legal frameworks are playing a greater role.[22] For now, however, companies are acting on their own. Facebook CEO Mark Zuckerberg claimed he did not want his and other social media companies to be "arbiters of truth," but they have ended up in this role by default.[23]

These problems are exacerbated when platforms adopt a strategy of aggressively taking down harmful content. The most obvious outcome is that important ideas are not shared, with the overall debate suffering as a result. Human rights activists in Tunisia, Syria, and other countries claim that Facebook has removed their posts documenting regime human rights abuses.[24] Another harm is a sense of anger and randomness on the part of users. As Mike Masnick notes, "By definition, content moderation is always going to rely on judgment calls, and many of the judgment calls will end up in gray areas where lots of people's opinions may differ greatly."[25]

Making this all more complex—and more consequential—is the lack of competition that many platforms enjoy; when they allow or take down content, it may have a far bigger effect than similar efforts in the past, when a speaker could more easily open a rival newspaper or go to another public space to speak. Although it is not difficult for skilled programmers to create a social networking site, many of the big ones have network effects advantages: Their size dissuades users from trying alternatives, and with more users they have access to better data to further improve their services.

---

[22] Jen Patja Howell, "The Lawfare Podcast: Content Moderation and the First Amendment for Dummies," *Lawfare*, March 11, 2021, https://www.lawfareblog.com/lawfare-podcast-content-moderation-and-first-amendment-dummies.

[23] Tom McCarthy, "Zuckerberg Says Facebook Won't Be 'Arbiters of Truth' After Trump Threat," *The Guardian,* May 28, 2020, https://www.theguardian.com/technology/2020/may/28/zuckerberg-facebook-police-online-speech-trump.

[24] Olivia Solon, "'Facebook Doesn't Care: Activists Say Accounts Removed Despite Zuckerberg's Free Speech Stance," NBC News, June 15, 2020, https://www.nbcnews.com/tech/tech-news/facebook-doesn-t-care-activists-say-accounts-removed-despite-zuckerberg-n1231110.

[25] Masnick, "Masnick's Impossibility Theorem."

They are, as Kate Klonick has written, "New Governors" who shape speech worldwide, an intermediary between governments and traditional speakers and publishers.[26]

THREE APPROACHES TO CONTENT MODERATION

Companies can remove individual posts they deem offensive, but they have other options as well. This section assesses three broad approaches and the specific options within them: removing content, reducing distribution, and shaping dialogue.

Approach One: Removing Content

The most straightforward approach is simply to remove offending material or, if it is particularly offensive or dangerous or if there are repeated infractions, remove the user altogether. In rarer cases, this might be done for entire collections of users and even platforms, denying them the services and protection necessary for them to continue.

Removing Individual Posts

One of the most common approaches is to block or remove posts, videos, or other hateful content, an approach most major platforms take (or try to take) with content that egregiously violates their terms of service.

Such removals, however, face many problems. One difficulty is avoiding over-removal while ensuring that hateful content is banished. As Evelyn Douek argues, rapid decisions are necessary, which means there is little time for deliberation (or for the review of decisions made by AI). Because they need to act quickly, platforms often choose to use too heavy a filter, as they are more likely to face criticism if they err on the side of leaving up posts.[27] As a result, content condemning violence, postings from human rights activists highlighting hatred, and at times just random content gets taken down without explanation.

---

[26] Kate Klonick, "The New Governors: The People, Rules, and Processes Governing Online Speech," *Harvard Law Review* 131, no. 6 (April 2018): 1598–1670, https://harvardlawreview.org/2018/04/the-new-governors-the-people-rules-and-processes-governing-online-speech/.

[27] Douek, "Verified Accountability," 10; Evelyn Douek, "Governing Online Speech: From 'Posts-as-Trumps' to Proportionality and Probability," *Columbia Law Review* 121, no. 3 (April 2021): 759–834, https://columbialawreview.org/wp-content/uploads/2021/04/Douek-Governing_Online_Speech-from_Posts_As-Trumps_To_Proportionality_And_Probability.pdf.

Removals often have collateral damage. Eric Goldman points out that when content is deleted, the comments on the offending posts are either deleted as well or become orphaned and lose their context and meaning, effectively degrading the content of other users regardless of whether their content is offensive.[28]

It is often difficult to train algorithms to remove troubling content, as much of it concerns sexuality, gender, race, and other issues where context and tone can dramatically change the meaning of text or an image. A group of YouTubers tried to reverse engineer the company's algorithm and found that it disproportionately removed posts with LGBTQ-related vocabulary, which in turn meant that some creators who made videos on LGBTQ issues faced demonetization and other penalties. At times it seemed like the algorithm targeted gay people: Simply replacing a word like *lesbian* with *friend* would allow a video to stay up. Their research also found that the YouTube algorithm seemed to take down content related to several odd terms, such as *healing* and *Oklahoma*, among others, as well as names like *Josh*.[29]

YouTube and other companies are constantly tinkering with their algorithms, and of course the AI systems are constantly learning, but as a result of the latter, engineers often do not know what, exactly, shapes what is removed.[30] Indeed, AI "explainability" or "interpretability" is a tremendous challenge, and the lack thereof leads to results that are not on the surface intuitive, appear inconsistent, and are difficult to explain to users. It makes it harder for individuals to challenge content takedowns and decreases confidence in AI accuracy and safety.[31]

At times, it is vital to preserve—or even exploit—offensive content, but only for a subset of users. In cases where there are links related to atrocities or other grave crimes, the social

---

[28] Eric Goldman, "Content Moderation Remedies," *Michigan Technology Law Review* 28, no. 1 (2021), https://doi.org/10.36645/mtlr.28.1.content.

[29] Aja Romano, "A Group of YouTubers Is Trying to Prove the Site Systematically Demonitizes Queer Content," *Vox,* Oct. 10, 2019, https://www.vox.com/culture/2019/10/10/20893258/youtube-lgbtq-censorship-demonetization-nerd-city-algorithm-report.

[30] Romano, "A Group of YouTubers Is Trying to Prove the Site Systematically Demonitizes Queer Content."

[31] Tim Rudner and Helen Toner, "Key Concepts in AI Safety: An Overview," Center for Security and Emerging Technology (March 2021), https://cset.georgetown.edu/publication/key-concepts-in-ai-safety-an-overview/.

media record is important evidence. YouTube faced criticism for its algorithm's inability to distinguish between actual hate videos and those that documented hate groups for educational purposes, and some reporters who cover these issues have found their accounts demonetized. Reporters, historians, and others may also want to learn about a political leader or otherwise inform a broader audience about the leader's views. Access to the leader's social media record can be critical even, or especially, if the leader's remarks are bigoted. Removing them thus can decrease public awareness of a leader's troubling record.[32]

When removing content, companies must develop archives for offensive content and effective carve-outs for journalists, legitimate investigators, academics, and civil society organizations. This enables them to better comment on the impact of social media on politics and society, find evidence for crimes, and learn more about how to combat hate.

Deplatforming

Deplatforming goes a step beyond removing posts and consists of the temporary or permanent banning of figures who deviate from a platform's terms of service and community guidelines. This can be done to an individual account; an individual user (to prevent them from simply rejoining the platform with a new account); or an entire community of users, such as removing a channel, group, or subreddit.

Deplatforming can be especially powerful because a small number of users often drive a massive amount of bad content. Facebook found that in some of its most vaccine-hostile communities, only 0.016 percent of accounts drove half of the anti-vaccine content.[33] As a result, deplatforming a few users has the potential to make entire platforms less toxic.

An important and influential study of the banning of two offensive subreddits on Reddit, r/fatpeoplehate and r/CoonTown (an anti-Black racist forum), found that the ban worked well at both the user and community levels, though not necessarily for the internet as a whole. Many users who posted offensive content stopped using Reddit, and hate speech by

[32] Julia Alexander, "YouTube's New Policies Are Catching Educators, Journalists, and Activists in the Crossfire," *The Verge,* June 7, 2019, https://www.theverge.com/2019/6/7/18657112/youtube-hate-policies-educators-journalists-activists-crossfire-takedown-demonetization.
[33] David Klepper and Amanda Seitz, "Facebook Froze as Anti-Vaccine Comments Swarmed Users," Associated Press, Oct. 26, 2021, https://apnews.com/article/the-facebook-papers-covid-vaccine-misinformation-c8bbc569be7cc2ca583dadb4236a0613.

those who remained declined by at least 80 percent. After the ban, some of those active in the two subreddits migrated to other subreddits, but they did not significantly worsen the discourse on those subreddits.[34]

A study of Twitter's deplatforming of several high-profile figures showed similar positive results. When Twitter removed the influencers and prevented them from opening new accounts, posts referencing the influencers fell by over 90 percent on average, new users were far less likely to tweet about them, and, overall, their supporters tweeted almost 13 percent less. These results suggest that many of the influencers' most enthusiastic supporters left the platform or became less active; with this change, the platform as a whole became less toxic (though some individual users did become more toxic), with the proportion of tweets coded as such falling by almost 6 percent. Nor did a "Streisand effect" occur: Users were not newly attracted to the deplatformed toxic figures as a result of the high degree of publicity.[35]

One criticism of Twitter's and Reddit's approaches is that users simply migrate to other social media sites in darker corners of the internet; as the Reddit researchers argued, "Reddit has made these users (from banned subreddits) *someone else's problem*" and thus did not make the internet safer overall even if Reddit improved the quality of its own platform.[36] Indeed, a study of those deplatformed on Twitter and Reddit found that many accounts went to Gab after being suspended from mainstream platforms. Those who went to Gab were more active and, especially for those who migrated from Reddit, more toxic.[37]

---

[34] Eshwar Chandrasekharan, Umashanthi Pavalanathan, Anirudh Srinivasan, Adam Glynn, Jacob Eisenstein, and Eric Gilbert, "You Can't Stay Here: The Efficacy of Reddit's 2015 Ban Examined Through Hate Speech," *Proceedings of the ACM on Human-Computer Interaction* 1, no. CSCW: 1–22, https://doi.org/10.1145/3134666.

[35] Shagun Jhaver, Christian Boylston, Diyi Yang, and Amy Bruckman, "Evaluating the Effectiveness of Deplatforming as a Moderation Strategy on Twitter," *Proceedings of the ACM on Human-Computer Interaction* 5, no. CSCW2 (October 2021): 1–30, https://doi.org/10.1145/3479525.

[36] Chandrasekharan et al., "You Can't Stay Here," 18. Italics in the original.

[37] Shiza Ali, Mohammad Hammas Saeed, Esraa Aldreabi, Jeremy Blackburn, Emiliano De Cristofaro, Savvas Zannettou, and Gianluca Stringhini, "Understanding the Effect of Deplatforming on Social Networks," *13th ACM Web Science Conference 2021* (June 2021): 187–95, https://doi.org/10.1145/3447535.3462637. See also Richard Rogers, "Deplatforming: Following Extreme Internet Celebrities to Telegram and Alternative Social Media," *European Journal of Communication* 35, no. 3 (2020): 213–29, https://doi.org/10.1177%2F0267323120922066.

Although migration to other platforms certainly occurs, this criticism of deplatforming appears overblown. A study of those deplatformed from Twitter and Reddit also showed that the reach of those who went to Gab decreased as they lost substantial numbers of followers.[38] Another study of deplatformed YouTube channels showed that the vast majority of deplatformed channels did not move to a new site. Some did migrate to the alternative video platform BitChute—which claims it is dedicated to promoting free speech—and a few prominent users, such as Martin Sellner, who leads the far-right Austrian Identitarian Movement, flourished there. Indeed, some users knew they might be deplatformed, and they and their fans prepared for their content to move to alternative platforms. However, the extreme nature of discourse in alternative platforms alienates some users, and even extreme platforms at times bow to hosting services or others that demand that they remove the worst of the content on their sites. In addition, the reach of Gab, BitChute, and similar services is far more limited than their mainstream equivalents, making it harder for extremists to spread their message to new users. Users like conspiracy theorist Alex Jones and white supremacist James Allsup suffered tremendous losses of audience when YouTube banned them.[39] Platforms have not completely stopped the proliferation of extremist content with deplatforming, but what matters most is the degree to which this content reaches and harms others, and deplatforming helps to mitigate that threat.

Deplatforming need not be permanent: Platforms can suspend accounts briefly (as warnings) or for longer periods of time. Snapchat, for example, as a warning will temporarily lock user accounts that engage in prohibited activities, suspending them only if they continue to engage in such activities.[40] Twitter permanently suspended Rep. Marjorie Taylor Greene's account in January 2022, but only after she accumulated five "strikes" for repeatedly posting misinformation on coronavirus vaccines and other issues.[41] YouTube puts a weeklong freeze on new uploads for accounts that receive a second warning for posting offensive materials, while Twitter, Facebook, and other platforms limit posting, retweeting, and so on for set

[38] Ali et al., "Understanding the Effect of Deplatforming on Social Networks."
[39] Adrian Rauchfleisch and Jonas Kaiser, "Deplatforming the Far-Right: An Analysis of YouTube and BitChute," *SSRN* (2021), https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3867818.
[40] Goldman, "Content Moderation Remedies," 37.
[41] Davey Alba, "Twitter Permanently Suspends Marjorie Taylor Greene's Account," *New York Times,* Jan. 2, 2022, https://www.nytimes.com/2022/01/02/technology/marjorie-taylor-greene-twitter.html.

time periods without removing accounts completely.[42] Such a temporary block serves as a warning and, ideally, educates a user on the type of behavior that is permissible.

As with removing individual posts, deplatforming has costs and limits. In some cases, a banned user can simply create a new profile and evade restrictions on the old one.[43] However, this is usually true only for less important users; prominent individuals are easily spotted. Those with significant followings often find it difficult to recreate them under a new name, and aggressive deplatforming efforts can again reverse what limited gains these users make under a new alias.

Deplatforming, however, also increases the sense of victimhood and political bias, even when companies are careful (some would say too careful) to focus on the specifics of the violation.[44] Thus, many mainstream conservatives took umbrage when President Donald Trump or other prominent figures were deplatformed, believing it reflects political bias by liberal Silicon Valley leaders rather than bad behavior by the prominent figures.

In addition, for many platforms, the financial cost of deplatforming can be quite real. Because deplatforming may involve removing important users and then seeing far less activity among their supporters, it can lead to less use of the platform and less engagement.[45] If platforms alienate powerful politicians, they are also at risk of punitive regulation.

Denying Service

A variant of deplatforming can also occur via internet-related services farther down the internet stack, denying websites or entire platforms the assistance companies need to reach large numbers of users. App stores, like the Google Play Store, might deny access to apps linked to platforms or individuals deemed to support hate. Browsers like Google Chrome can stop access from their browser to certain websites based on offensive content, and domain registrars like GoDaddy can deny service, making offensive websites inaccessible. After Charlottesville, the content delivery network (CDN) Cloudflare withdrew its protection from 8chan, leaving it vulnerable to distributed denial of service (DDoS) attacks and other

[42] Goldman, "Content Moderation Remedies," 37–38.
[43] Bente Kalsnes and Karoline Andrea Ihlebæk, "Hiding Hate Speech: Political Moderation on Facebook," *Media, Culture & Society* 43, no. 2 (2021): 326–42, https://doi.org/10.1177%2F0163443720957562.
[44] Rauchfleisch and Kaiser, "Deplatforming the Far-Right."
[45] Jhaver et al., "Evaluating the Effectiveness of Deplatforming as a Moderation Strategy on Twitter."

cyber threats.[46] Similarly, internet service providers like AT&T could deny support for websites like 8chan, as they did after the New Zealand attacks. They might also use deep packet inspection, filtering for certain words or images in the traffic that goes through their pipelines.[47]

Relying on domain registrars, CDNs, and others farther down the internet stack is difficult, however. These companies do not see content moderation as part of their mission: They usually consider themselves to be basic infrastructure, having no role in what is done with their services. As a result, they often lack a significant content moderation staff, do not have any expertise on different cultures, and are otherwise ill suited for this mission. In addition, techniques like deep packet inspection or preventing browsers from accessing certain websites can easily be used to bolster repression under the pretext of stopping extremist content.

Approach Two: Reducing Distribution

Platforms do not just host content: They often amplify it, increasing its exposure beyond what would occur due to the platform's basic hosting and transmissions functions alone. And while they may rightly evade legal or even ethical responsibility for the original content, the question of what is promoted, to whom, and how much is very much at the center of their services. Instead of removal, platforms could choose to demote content, distribute it less widely, or otherwise make it less available: the reverse of amplification.[48] As Renée DiResta famously put it, "Free speech is not the same as free reach."[49]

Reducing Visibility

Platforms can hinder casual encounters with potentially offensive content but still allow access to those who know its precise location. For example, a platform can tag content with "noindex" to ensure that Google and other search engines do not bring users there, even

[46] Donovan, "Navigating the Tech Stack."
[47] Ruddock and Sherman, "Widening the Lens on Content Moderation."
[48] Daphne Keller, "Amplification and Its Discontents: Why Regulating the Reach of Online Content Is Hard," *Journal of Free Speech Law* 1, no. 1 (2021): 227–72, https://www.journaloffreespeechlaw.org/keller.pdf.
[49] Renée DiResta, "Free Speech Is Not the Same as Free Reach," *Wired*, Aug. 30, 2018, https://www.wired.com/story/free-speech-is-not-the-same-as-free-reach/.

though the content remains present on the platform itself. A variant is to do this for internal searches, with platforms removing select content from on-platform search results as a way of making it harder to access. Platforms may also relocate content, moving it to a new URL and, in so doing, removing comments and breaking links to the original posting. YouTube, for example, does this for videos it identifies as being promoted by spam.[50] Search engines may also remove auto-suggest content for offensive terms, making it less likely that ordinary users will use them.[51]

Some platforms quarantine content, allowing the offending posts to stay up and the users behind them to remain on the platform but trying to limit access to and promotion of the community of which they are part. Reddit attempted to quarantine the pro-Trump r/The_Donald, which contained considerable racist and anti-Muslim content, and the misogynistic "men's rights" subreddit r/TheRedPill. The quarantined subreddits could still be accessed directly through their URLs, but they did not appear in search results. In addition, the warning "shocking or highly offensive content" appeared before users could enter the subreddit. Researchers found that quarantining greatly decreased the number of new members—a fall of over 50 percent for both subreddits. Nor did the users simply move to another community and submit toxic posts there; toxicity even decreased in some associated communities. As with deplatforming, quarantining did not attract users hitherto unaware of the toxic site, thus failing to produce a "Streisand effect."[52]

One hope was that the quarantine would serve as a warning to the community and its moderator to rein in their offensive behavior. However, researchers found that there was not a significant change in hateful speech within the quarantined community itself.[53]

Part of the goal of quarantining was simply to introduce "friction" into the process in order to discourage behavior online that, with a bit of reflection, the offending users themselves might not favor. One way to do this is by making it harder to go to the toxic site, requiring

[50] Goldman, "Content Moderation Remedies," 33.
[51] Goldman, "Content Moderation Remedies," 42–44.
[52] Eshwar Chandrasekharan, Shagun Jhaver, Amy Bruckman, and Eric Gilbert, "Quarantined! Examining the Effects of a Community-Wide Moderation Intervention on Reddit," *ACM Transactions on Computer-Human Interaction* 29, no. 4 (August 2022): 1–26, https://doi.org/10.1145/3490499.
[53] Chandrasekharan et al., "Quarantined!"

extra clicks or other additional efforts. Sometimes, it may also require the user to acknowledge that the content is toxic before going forward.[54]

Downranking and Demoting

Companies try to rank content to ensure the most user engagement. For several years, Facebook actively promoted divisive content, tuning its promotion algorithm to give more weight to "anger" emojis and similar responses despite knowing that such responses often came in response to toxic content and misinformation.[55] Facebook even allowed pro-vaccine posts to be swarmed by negative commentary and elevated vaccine misinformation because it drove high user engagement. Ranking posts by the quality of information or chronologically would have reduced this problem but driven down overall engagement.[56]

Instead, platforms could downrank certain content, making it less visible on the overall platform, or, at the very least, not use promotion tools such as recommendations for materials that are not permitted or are at the borderline between what is permitted and what is not acceptable. YouTube decided in 2019, for example, not to recommend borderline videos even though it allows them to remain on the platform. Companies may also take steps to prevent certain content from going viral, such as Twitter's limits on retweeting for tweets that it found violative but that it kept up for public interest.[57] Such measures, of course, would have financial consequences for companies always seeking more ways to engage new users.

Some disturbing content, however, should still be actively amplified, or at least not downranked. The video that documented the slow suffocation of George Floyd, for example, was vital to fueling outrage over deep-seated societal issues and sparking a broader social movement. Content about the 2022 war in Ukraine also might be upsetting, but awareness of the war and its horrors is vital for broad public understanding.

---

[54] Thomas Mejtoft, Sarah Hale, and Ulrik Söderström, "Design Friction," *Proceedings of the 31st European Conference on Cognitive Ergonomics* (September 2019): 41–44, https://doi.org/10.1145/3335082.3335106; Chandrasekharan et al., "Quarantined!"

[55] Jeremy B. Merrill and Will Oremus, "Five Points for Anger, One for a 'Like': How Facebook's Formula Fostered Rage and Misinformation," *Washington Post*, Oct. 26, 2021, https://www.washingtonpost.com/technology/2021/10/26/facebook-angry-emoji-algorithm/.

[56] Klepper and Seitz, "Facebook Froze as Anti-Vaccine Comments Swarmed Users."

[57] Goldman, "Content Moderation Remedies," 45–46.

Using Warning Labels and Other Forms of Friction

Some companies have begun using warning labels to flag dangerous content or misinformation. Facebook, YouTube, Twitter, and other companies have used warnings for graphic or hateful content, at times simply placing it before the viewer and in other cases forcing the viewer to click an additional time before seeing the content.[58] However, many warnings appear to have little or no impact on behavior, regardless of framing.[59]

Applying Geographic Limits

What is offensive (or illegal) can vary by country, with different cultures and countries taking different approaches to the same material. "Geoblocking" is one way to adjust content to local sensitivities, enabling content in some areas but blocking it in others. This is often done for copyright purposes, where website administrators use a computer's IP address to determine where the user is and, from there, to permit only content that is allowed in that country. Geoblocking has also been used to accommodate German law, which makes Holocaust denial illegal, but allow the same content to stay up in a more permissive country.[60]

Geoblocking is at times a legal necessity, but it harms several categories of users, as Peter Yu points out. Geoblocking can be an easy excuse for censorship, with governments using copyright claims as a way to deny access to critical news coverage, for example. Those who travel cannot access the content they legitimately enjoyed in another country. Geoblocking can also be challenging technically, with a constant back-and-forth between workarounds such as virtual private networks, or VPNs, and blocking methods, requiring considerable

---

[58] Goldman, "Content Moderation Remedies," 35.

[59] Ciara M. Greene and Gillian Murphy, "Quantifying the Effects of Fake News on Behavior: Evidence From a Study of COVID-19 Misinformation," *Journal of Experimental Psychology: Applied* 27, no. 4 (2021): 773–84, https://doi.org/10.1037/xap0000371.

[60] Eline Jeanné, "Geoblocking: What Is It, and How Effective Is It in Practice?" Getthetrollsout.org, June 26, 2019, https://getthetrollsout.org/articles/geoblocking-what-is-it-and-how-effective-is-it-in-practice; Peter K. Yu, "A Hater's Guide to Geoblocking," *Boston University Journal of Science and Technology Law* 25, no. 2 (2019): 503–29, https://scholarship.law.tamu.edu/facscholar/1339/.

resources to keep up. Finally, the legal landscape is confusing, reflecting a lack of international consensus on this issue.[61]

Restricting the Audience

Age restrictions are common on many platforms, allowing platforms to permit a broader array of content—including content that might offend audiences—on the grounds that the most vulnerable can be shielded. YouTube, for example, has a "Restricted Mode" that, when activated, prevents a user from accessing content the company deems unsuitable for younger audiences.[62]

In practice, however, such restrictions are often not used. On average, only 1.5–2% of YouTube viewers use Restricted Mode.[63] Many parents are not aware of these options or otherwise do not use them, and many young viewers, of course, do not turn on Restricted Mode or otherwise evade this possible restriction.

Approach Three: Shaping Dialogue

Technology companies can try to reshape dialogue on their platforms to make it less toxic and more positive without the more extreme measures of banning users and removing content. There are a variety of ways to do this, some of which focus on particular content and others on incentives and disincentives for the users themselves.

Moderating the Debate

On many platforms, discussions are moderated by humans, and the moderator plays an important role in the quality of the discourse. Among other functions, moderators can set the goals for a discussion, publicize it, and establish rules for the participants.[64] On some

[61] Yu, "A Hater's Guide to Geoblocking."
[62] *Prager University v. Google LLC,* No. 18-15712 (9th Cir. 2020), 7.
[63] *Prager University v. Google LLC*, 7.
[64] Arthur R. Edwards, "The Moderator as an Emerging Democratic Intermediary: The Role of the Moderator in Internet Discussions About Public Issues," *Information Polity* 7, no. 1 (2002): 3–20, at 5, https://dl.acm.org/doi/10.5555/1412444.1412446.

platforms, moderators can also exclude participants, remove a post, or otherwise determine who speaks and what content is acceptable.[65]

Interactions often occur by liking, sharing, or commenting on the post of an influential account, and these comments in turn become grist for further discussion. Moderators can elevate some discussions while decreasing the audience for others.[66] Platforms can facilitate this process with technological changes to further empower moderators. One set of political forum moderators modified Facebook's profanity filter, which can identify racist terms as well as swear words, to add specific words or phrases that they found offensive relevant to their group discussions.[67]

Much depends, of course, on the intentions of the moderator. Because moderators are usually unpaid, they take on this role because of their commitment to the community. In many cases that is admirable, but in some they are committed to misogyny, racism, or other ugly causes. Facebook, for example, relied on moderators to ensure content quality in Facebook Groups, but it found that moderators often led efforts to push conspiratorial content such as election fraud in the 2020 election.[68] At the very least, dependence on moderators produces uneven results, with some seeking a balanced and productive discussion while others do little or even make the problem worse.

Elevating Trustworthy and Positive Content and Counterspeech

Another approach is to elevate "good" speech beyond what a company's algorithms would do on their own. Often the hate and counterspeech communities are well connected, engaging each other rather than living in echo chambers. Many of the most extreme thrive in combative environments, but for those not fully committed, exposure to counterspeech can deter the spread of hate speech by raising social inhibitions and thus the "cost" of hate speech.[69] Platforms could give groups like the NAACP or an immigration rights organization free or cheaper advertisements to enable their content to be seen alongside searches or

[65] Kalsnes and Ihlebæk, "Hiding Hate Speech."
[66] Kalsnes and Ihlebæk, "Hiding Hate Speech."
[67] Kalsnes and Ihlebæk, "Hiding Hate Speech," 336–37.
[68] Sheera Frenkel, "The Rise and Fall of the 'Stop the Steal' Facebook Group," *New York Times*, Nov. 5, 2020, https://www.nytimes.com/2020/11/05/technology/stop-the-steal-facebook-group.html.
[69] He et al., "Racism Is a Virus."

promoted with various comments.[70] Similarly, platforms could boost content from reputable news sources or individuals deemed to be more objective (the latter being a harder and more labor-intensive approach).

A variant of counterspeech is to redirect potentially radicalized individuals to different information and real-world support organizations. Instagram, for example, tries to disrupt groups that promote eating disorders by offering a content warning and displaying information about sites where those with eating disorders can receive help.[71] Similarly, Facebook and Twitter highlight material that offers users a chance to fact check false stories.[72]

In the extremism space, Facebook initiated a "Redirect" program to steer users searching for hateful content to organizations that move people away from radicalism, such as Life After Hate. Drawing on a list of known keywords linked to hate groups, it then activated a safety module that encouraged searchers of these keywords to go to the website of an anti-hate group. An evaluation of the program found that it achieved modest success, with a few users becoming engaged with anti-hate groups. Another goal of the program, though harder to evaluate, was to create friction for searches for neo-Nazi and similar material, slowing down the process and otherwise making it more frustrating.[73]

Such programs have several problems and limits. Most obviously, they work on only a fraction of potential radicals: The evaluation of Facebook's program found that of the almost 60,000 searches for hateful terms in the five-month period studied, 4 percent of those searches led to users clicking on the safety module, and, of those, only 25 individuals eventually clicked "get help" on the website of an anti-hate organization.[74]

Even if more individuals respond positively, some of these anti-hate organizations are not set up for an influx of potential users from Facebook and other platforms. In addition, given

[70] Citron and Norton, "Intermediaries and Hate Speech," 1481.
[71] Chandrasekharan et al., "Quarantined!"
[72] Goldman, "Content Moderation Remedies," 36.
[73] Moonshot, *From Passive Search to Active Conversation: An Evaluation of the Facebook Redirect Programme* (London: Moonshot, 2020), https://counterspeech.fb.com/en/wp-content/uploads/sites/2/2020/11/Facebook-Redirect-Evaluation_Final-Report_Moonshot-1.pdf.
[74] Moonshot, *From Passive Search to Active Conversation*.

the disparity between the number of potential radicals and those who seek help, the cost of the program and its scalability are both open to question.

Counterspeech in general faces many challenges. Bad actors can exploit or game counterspeech attempts. For example, YouTube highlighted the #MoreThanARefugee video in 2017, but far-right communities reframed, mocked, negatively reviewed, and otherwise ensured a steady stream of negative comments associated with the video.[75] Good counterspeech is also difficult to produce. One analysis called for "sexy" counterspeech that is bold and funny; however, "bold" and "funny" are context dependent and can cause reputational damage if they are perceived as on the wrong side of the line of respectability—the lack of examples for such an approach indicates how difficult it is.[76] Even basic fact checking often fails and at times backfires. Some people double down on their false beliefs as a backlash against the fact check.[77]

Providing Reminders and Warnings

Many content violations occur due to ignorance of company rules or involve only small amounts of dangerous content. In such cases, users might be warned about their behavior, both to educate them on the rules and to let them know that continued violations will be punished more severely. YouTube, for example, has a strike system in which initial violations receive a warning and the penalties are more severe for subsequent strikes, ultimately resulting in suspension. [78]

Simple warnings can help. Nextdoor offers a "Kindness Reminder" to users, asking them to reconsider posts that the site identifies as negative, drawing on previously flagged comments. Nextdoor found that 20 percent of users hit "edit" in response to the prompt,

---

[75] Julia Ebner, "Counter-Creativity: Innovative Ways to Counter Far-Right Communication Tactics," in *Post-Digital Cultures of the Far Right: Online Actions and Offline Consequences in Europe and the US*, ed. Maik Fielitz and Nick Thurston (Bielefeld, Germany: Transcript Verlag, 2019), 176–77.
[76] Ebner, "Counter-Creativity," 179.
[77] Alice E. Marwick, "Why Do People Share Fake News? A Sociotechnical Model of Media Effects," *Georgetown Law Technology Review* 2, no. 2 (July 2018): 474–512, https://georgetownlawtechreview.org/why-do-people-share-fake-news-a-sociotechnical-model-of-media-effects/GLTR-07-2018/.
[78] Rauchfleisch and Kaiser, "Deplatforming the Far-Right."

resulting in far fewer negative comments.[79] Twitter also found that when it nudges users to read a story before retweeting it or to reconsider a reply that might offend others, at least some users reconsider—around a third in the case of retweeting unread articles.[80]

Such reminders and warnings do little for users who knowingly violate terms of service or otherwise do not care about being on the right side of the rules. In addition, companies face financial consequences and possible reputational concerns when they use strikes or other measures against popular users who attract engagement to the platform. Not surprisingly, platforms like Facebook have tried to excuse, or limit the strikes given to, high-profile influencers who violate their rules (like Charlie Kirk), while YouTube simply refused to investigate many problems, fearing the financial consequences of what it would find.[81]

Turning Off and Hiding Comments

Another approach is to simply turn off comments on a piece of content in order to limit the amount of toxic content and prevent engagement from spiraling into virality. For instance, YouTube disables comments when a video is identified as containing white supremacist content.[82] At times, though, such an action will come from the user rather than the platform. In one case, the Canadian Broadcasting Corporation (CBC) found that its Facebook pages had "an inordinate amount of hate, abuse, misogyny and threats in the comments under [its] stories." The subjects of the stories were often attacked in the comments—as were the journalists themselves—and the overall user experience suffered. This stood in contrast to comments on the news site itself, which were curated to allow genuine criticism but not

[79] "Announcing Our New Feature to Promote Kindness in Neighborhoods," Nextdoor.com, Sept. 18, 2019, https://blog.nextdoor.com/2019/09/18/announcing-our-new-feature-to-promote-kindness-in-neighborhoods/.
[80] James Vincent, "Twitter Is Bringing Its 'Read Before You Retweet' Prompt to All Users," *The Verge,* Sept. 25, 2020, https://www.theverge.com/2020/9/25/21455635/twitter-read-before-you-tweet-article-prompt-rolling-out-globally-soon.
[81] Olivia Solon, "Sensitive to Claims of Bias, Facebook Relaxed Misinformation Rules for Conservative Pages," NBC News, Aug. 7, 2020, https://www.nbcnews.com/tech/tech-news/sensitive-claims-bias-facebook-relaxed-misinformation-rules-conservative-pages-n1236182; Mark Bergen, "YouTube Executives Ignored Warnings, Let Toxic Videos Run Rampant," *Bloomberg*, April 2, 2019, https://www.bloomberg.com/news/features/2019-04-02/youtube-executives-ignored-warnings-letting-toxic-videos-run-rampant.
[82] Kent Walker, "Four Ways Google Will Help to Tackle Extremism," *Financial Times*, June 18, 2017, https://www.ft.com/content/ac7ef18c-52bb-11e7-a1f2-db19572361bb.

"vile abuse." After comments were removed from the Facebook pages, the CBC staff's well-being improved, and they felt freer to run more diverse stories.[83]

A study on how political party officials in Norway moderate Facebook shows how moderators might shape dialogue. The study found that moderators hid comments they deemed offensive or that otherwise hindered constructive dialogue, even deleting entire conversations. The Norwegian users did not know their comments were hidden, a fact that the moderators found allowed users to be visible to their friends but prevented them from shaping the overall debate.[84] Similarly, Instagram, Reddit, and Twitter have "shadowbanned" users, giving them the impression they are still contributing to the overall discussion but, in reality, preventing other users from seeing their comments.[85]

Reducing comments reduces engagement, a key metric that social media companies use as part of their business model. In addition, the lack of transparency inherent in shadowbanning and other hidden means of controlling content is frustrating for online commentators. Such efforts, if not undertaken for clearly justified and consistent reasons, create a sense that it is a political agenda and censorship, rather than a failure to follow impartial rules, that causes a post or user to be removed.[86] This increases political support for antitrust regulation and other restrictions that companies oppose.

Demonetizing

Some platforms rely on users as creators and reward them with a share of ad revenue or other income. Such payments, in turn, can be manipulated to try to incentivize better user behavior. YouTube has claimed that the company has a "higher standard" for creators who make money from their videos than it does for ordinary users. It withheld payment to right-

[83] Brodie Fenlon, "CBC Is Keeping Facebook Comments Closed on News Posts," Canadian Broadcasting Corporation, Nov. 1, 2021, https://www.cbc.ca/news/editorsblog/facebook-comments-ed-blog-1.6230921.
[84] Kalsnes and Ihlebæk, "Hiding Hate Speech," 337.
[85] Rauchfleisch and Kaiser, "Deplatforming the Far-Right."
[86] Kalsnes and Ihlebæk, "Hiding Hate Speech."

wing commentator Steven Crowder, who used his channel to harass LGBTQ creators, even though it did not take down the content.[87]

Demonetizing can also be done farther down the internet stack. Stores like Google Play can prevent sales of an app, while payment providers like Stripe, Square, and Amazon Pay can deny services to a platform or user.[88] In 2017, after the "Unite the Right" rally and associated criticism of technology companies for helping facilitate the event, PayPal cut off its services to groups run by right-wing extremists like Jason Kessler and Richard Spencer.

Demonetizing, however, has many limits. Individuals like Crowder can make money in other ways, such as by selling offensive T-shirts via the channel.[89] Demonetizing also raises concerns about censorship and political bias, as it tilts the political playing field against those being punished. More importantly, many of the most prominent individuals, groups, and states that spread bad content are not motivated by immediate financial gain; thus, these penalties mean little. Demonetizing means little for the many toxic users whose posts do not reach large audiences and thus are not benefiting financially from their activity. Finally, demonetizing does not stop the content itself, so other users might see it and become radicalized or suffer emotionally.

Ending Anonymity and Shaming Users

Another way to shape dialogue is to force users to take more personal responsibility for the content they produce. Some platforms, such as Facebook, require users to verify their identity and otherwise know a considerable amount about their users, giving them the potential to identify extremists and, if they choose, require them to use their real identities. Other platforms, however, allow users to join with just an email address or phone number (or sometimes even less) or otherwise may not know users' true identities. For ordinary users, anonymity allows them to voice more controversial opinions and otherwise express themselves more freely. However, this anonymity is often a license for abuse: When users are not directly associated with their posts, they can avoid real-world retaliation or shame.

---

[87] Evelyn Douek, "YouTube's Bad Week and the Limitations of Laboratories of Online Governance," *Lawfare,* June 11, 2019, https://www.lawfareblog.com/youtubes-bad-week-and-limitations-laboratories-online-governance.
[88] Ruddock and Sherman, "Widening the Lens on Content Moderation."
[89] Douek, "YouTube's Bad Week and the Limitations of Laboratories of Online Governance."

Therefore, platforms could consider revoking anonymity when users violate their terms of service.

Anonymity, however, is vital for a small but important subset of users: political dissidents. In authoritarian countries or other states where there is a risk of repression and political violence against regime or social critics and their families, anonymity is a way to promote important content and alternative perspectives with less danger of intimidation. Putin's Russia, Erdoğan's Turkey, and other dictatorial regimes monitor social media and arrest those who voice any dissent. Ending anonymity would make it far easier for such dictators to silence opponents.

Another approach is to shame the account or individual in a visible way, both to deter others and to encourage that user not to repeat offensive behavior. Some video games have turned players who violate their rules into virtual toads as a way to embarrass them, for example.[90] This requires, of course, a technical option for such an embarrassing label and it may not work for truly incorrigible users who may even welcome such "toading." More serious platforms may also see "toading" as damaging their brand, though they could change the "toading" to a less whimsical transformation such as a red stripe through a username.

Enabling User Control and Encouraging Reporting

Companies could give users more control over what they see in their accounts, letting them set their content filters with more fidelity rather than having the company do it (or not do it) on behalf of all its users. Truly illegal content would still be removed, but the user would set the border for borderline content. Many platforms already have some degree of user control, at times linked to age restrictions and in other cases just to appease critics or prevent legal action. For example, Twitter allows its users to block certain terms to better curate their own experiences on the platform.[91]

Given that many users may not fully understand the implications of their choices, they could also tie their preferences to those of organizations or individuals they trust. As Daphne

[90] Goldman, "Content Moderation Remedies," 40.
[91] See Twitter, "How to Use Advanced Muting Options," https://help.twitter.com/en/using-twitter/advanced-twitter-mute-options.

Keller notes, "Users could opt for their church's ranking preferences, or Vox's, or Fox News's—or even just Facebook's, Google's, or Twitter's."[92]

Some platforms also allow users to bulk-delete harassing comments, giving users a way to more effectively curate their own accounts.[93] In 2021, Facebook allowed all of its users to control who can see comments on a post, and Twitter provided an option to limit who could reply to tweets. These decisions followed a 2019 legal ruling from Australia that held companies like Facebook liable for defamatory comments that users posted on public pages.[94]

Some sites have encouraged users to report harassment, hate speech, and other content that may violate a platform's terms of service. This gives users a sense of empowerment, allowing them to feel some sense of agency despite the companies' complex and opaque rules and at times seemingly random content moderation decisions. At times it may even lead to the removal of bad content that the platform otherwise might not have noticed.

At the very least, this approach's success relies on the audience understanding the harms related to hate speech and the basics of the platform's terms of service.[95] As this is difficult for many paid content moderators, it is not surprising that many users do not understand it well. Indeed, users often simply report content they find objectionable (for example, pro- or anti-Trump content) rather than content that violates the platform's terms of service. As a result, platforms must often examine a wide range of reported content even when it clearly does not violate their terms of service. Indeed, there is a problem of a "heckler's veto," where individuals, groups, or even countries submit false allegations to platforms in order to get critical, but legitimate, content taken down.[96]

Given these problems, some users might be trusted more than others. Some platforms allow government agencies, civil society organizations, and trusted individual experts to report bad content and have their information receive expedited attention (though their reports are

---

[92] Keller, "Amplification and Its Discontents."
[93] Evelyn Douek, "More Content Moderation Is Not Always Better," *Wired,* June 2, 2021, https://www.wired.com/story/more-content-moderation-not-always-better/.
[94] Josh Taylor, "Facebook Now Lets Users and Pages Turn Off Comments on Their Posts," *The Guardian*, March 31, 2021, https://www.theguardian.com/media/2021/mar/31/facebook-turn-off-comments-on-post-limit-restrict-disable-comment-posts-moderation-control-tool.
[95] Citron and Norton, "Intermediaries and Hate Speech," 1478.
[96] Keller, "Amplification and Its Discontents."

still subject to review).[97] This effectively outsources content moderation, creating its own risks. Companies must develop a process for identifying and incorporating the views of outsiders. Even more important, companies must recognize the agendas, biases, and limits (especially for governments) of those doing the reviewing.

TENSIONS AND IMPLICATIONS

As the above discussion makes clear, there is no single option that works best when seeking to reduce toxic content. When issuing penalties, companies must consider a range of factors. The most obvious is the harm involved with the rule violation: Some are far more dangerous than others. Companies also want to be able to scale their penalties and treat users around the world equally, both for ethical reasons and because the AI they use depends on consistency. That said, they must consider the collateral damage of penalties, as penalties often affect other users who were not guilty of the violation and disproportionately harm already-disadvantaged communities. Further, several tensions between different objectives such as rapid action and the proper level of content removal inevitably arise. Finally, some users and communities may have more ability to self-correct; thus, certain approaches that involve more warnings and user learning are better for them than for others.[98] These factors shape the advantages and disadvantages of different approaches; combinations are necessary, as is tailoring the approach to the needs and interests of specific platforms and their stakeholders.

The Free Speech Question

One tension, seemingly endlessly debated, is between content moderation and free speech. Under recent interpretations of the First Amendment, companies have considerable latitude to determine their own policies free of government regulation, with narrow exceptions for universally agreed evils like child sexual abuse material.[99] Yet despite this legal freedom, the companies—the majority based in the United States and often influenced by U.S.-trained lawyers—value the First Amendment and are seen by American audiences in particular as

[97] See one such approach by YouTube at https://support.google.com/youtube/answer/7554338?hl=en.
[98] For a discussion, see Goldman, "Content Moderation Remedies," 53–63.
[99] Howell, "The Lawfare Podcast"; Alan Z. Rozenshtein, "Silicon Valley's Speech: Technology Giants and the Deregulatory First Amendment," *Journal of Free Speech Law* 1, no. 1 (2021): 337—76. Available at SSRN, https://ssrn.com/abstract=3911460.

needing to respect it. However, years of criticism about fake news, Nazis online, and other harms have led to less emphasis on free speech and more on creating safe environments online.[100]

Over- vs. Under-Removal

Another tension is between approaches that leave up too much bad content versus those that take down too much acceptable content: There is no true Goldilocks solution. A company may reason, correctly, that it will face more criticism for leaving up a neo-Nazi post than for inadvertently taking down posts that criticize Nazism but are wrongly identified by the company's algorithms. Smaller platforms, which have fewer lawyers and less ability to survive a negative financial judgment, are likely to be particularly risk averse.[101]

Even uncontroversial and well-meaning efforts like stopping child sex trafficking online produced unwanted costs due to over-removal, with platforms like Craigslist eliminating all personal ads and other platforms denying services to all sex workers (not just, as intended, those exploiting underage ones) because they feared the liability inherent in underenforcement. Cloudflare terminated service to the website Switter, created as an online refuge for sex workers, because it felt that the law intended to stop child sex trafficking was confusing and left the company vulnerable. These measures made the lives of sex workers more dangerous.

Newsworthiness and evidentiary issues also increase the costs of over-removal. Even horrific content should be available for a small group of people, and removing it can hinder investigations or prevent political figures from receiving the scrutiny and criticism they deserve.

At the same time, there are incentives for under-removal. Some platforms may abdicate responsibility for content, fearing censure or legal risk for mistakes on over-enforcement.[102] Others may simply lack the staff or AI capabilities to do content removal at scale. As

---

[100] Tim Wu, "Will Artificial Intelligence Eat the Law? The Rise of Hybrid Social-Organizing Systems," *Columbia University Law Review* 119, no. 7 (2019), https://columbialawreview.org/content/will-artificial-intelligence-eat-the-law-the-rise-of-hybrid-social-ordering-systems/.
[101] Keller, "Amplification and Its Discontents"; Jurecic, "The Politics of Section 230 Reform."
[102] Jurecic, "The Politics of Section 230 Reform."

Goldman notes, however, it is also important for companies to assess the confidence that a violation actually occurred.[103] Especially as most decisions are done by AI, they are inherently probabilistic, and a higher degree of certainty is necessary to justify more severe punishments to minimize the harm of false positives.

Under-removal is particularly likely in non-English-speaking areas. Companies like Facebook overwhelmingly prioritize English-language content in their moderation, even though non-English content can be as bad or far worse. One internal Facebook memo on vaccine hesitancy noted that Facebook's ability to detect disinformation in comments "is bad in English—and basically non-existent elsewhere."[104] And for all the criticism Facebook has (often rightly) received, other companies have the same problem or are worse.

Fortunately, lawmakers do seem to be learning from past mistakes. Legislative proposals concerning content moderation are now more focused on particular harms rather than calling for sweeping changes that are technically difficult and lead to broad over-removal out of an abundance of caution. The proposals are also more nuanced with regard to how they treat companies that have different market capitalizations and different positions on the technology stack, recognizing that different platforms have different capacities and roles to play.[105]

Yet politics still intrudes. As Jurecic contends, "Where Republican politicians were often irate that platforms had taken down content or otherwise limited access to it, Democratic politicians voiced frustration that platforms were leaving too much content *up*—misinformation around the coronavirus, far-right extremism, or lies posted by Trump and his associates."[106] Grandstanding by both parties will remain a limit to serious reform.

Process Transparency and Oversight

To improve the legitimacy of removal and moderation policies, some degree of external involvement is necessary. Douek has pointed out that companies like YouTube have a lack

[103] Goldman, "Content Moderation Remedies," 42-44.
[104] Klepper and Seitz, "Facebook Froze as Anti-Vaccine Comments Swarmed Users."
[105] Jurecic, "The Politics of Section 230 Reform."
[106] Jurecic, "The Politics of Section 230 Reform." Italics in the original.

of transparency and accountability regarding the content they take down or allow.[107] Such opacity can allow companies to claim their processes are objective when, in reality, they are fraught.[108] Leading companies like Facebook have released more information on their content moderation practices in recent years, but key information on policies and on the scope of the problem remains hidden.

Given the growing role that algorithms play in content moderation for larger platforms, transparency is exceptionally difficult to achieve. The interpretability challenge for AI remains significant, and progress is moving far more slowly than the adoption of AI for content moderation. A lack of transparency on parameters compounds the problem. For example, the definitions used to bound fuzzy terms like *hate speech* and the data and labels used to train the algorithm are usually not shared.[109]

Douek calls on government regulation to emphasize transparency regarding the process of how platforms make speech decisions rather than on the outcomes of the decisions themselves. Examples might include imposing disclosure obligations and requiring internal oversight mechanisms. Companies can, and perhaps should, end up with different approaches, but there would be greater clarity and thus greater legitimacy as a result.[110] Transparency should involve a dialogue with users about what hate speech is most troubling and the potential harms of various platform moderation efforts.[111] Some users, of course, will game the rules that are publicized, calibrating harassment and other bad behavior to fall just below the threshold for moderation.[112]

Another approach is to outsource some decisions to an external oversight body, as Meta has done by creating its Oversight Board, which provides input into content on Facebook and Instagram. In theory, such an oversight board uses a small number of cases as examples to illustrate deeper principles, which are then applied by the platform more widely. This

---

[107] Douek, "Verified Accountability," 2; Douek, "YouTube's Bad Week and the Limitations of Laboratories of Online Governance."
[108] Sarah T. Roberts, "Digital Detritus: 'Error' and the Logic of Opacity in Social Media Content Moderation," *First Monday* (2018), https://journals.uic.edu/ojs/index.php/fm/article/download/8283/6649.
[109] Gorwa et al., "Algorithmic Content Moderation."
[110] Douek, "Verified Accountability," 7.
[111] Citron and Norton, "Intermediaries and Hate Speech," 1441.
[112] Caplan, "Content or Context Moderation?"

approach offers greater public visibility, an understanding of legitimate areas of disagreement when there is no clear answer, as well as an outside check on a company. Even if a company overrules its decisions or simply ignores them, the body still highlights important issues. Company officials who know there may be outside review may also be more careful and consistent in their decisions.[113]

Counter Reactions

A final concern is that the extreme right and other toxic communities develop their own platforms and information ecosystem, relying on "alt-tech" platforms to help them recruit and organize. Many hateful voices already use an array of platforms, leading to considerable redundancy. The 2017 "Unite the Right" rally, for example, was organized in part on major platforms and gaming sites like Discord, but also via 8chan, altright.com, and podcasts such as the Daily Shoah. Deplatforming and similar steps can also help extremist platforms like Gab find a greater audience.[114] This can create small but powerful echo chambers, where users become more extreme as they are egged on by those also using the platform.

Yet there are limits to the alt-tech world. Such platforms often have little funding, as most companies are not willing to advertise there. Their mobile apps are also taken down by Apple and Google's mobile stores. In part as a result of this, the technology and user experience are often poor: "glitchy and unstable" is how one analysis described Gab.[115] And, as discussed above, when users switch to alternative platforms, their audiences are far smaller.

CONCLUSION

This paper argues that companies have many options for content moderation, but all of them are flawed. Moreover, platforms face inherent tensions and trade-offs when they try to shape discourse on their platforms. Some of these involve tensions balancing free speech versus other rights, while others require confronting whether to err on the side of removal

[113] Douek, "Verified Accountability," 14–16.
[114] Joan Donovan, Becca Lewis, and Brian Friedberg, "Parallel Ports: Sociotechnical Change From the Alt-Right to Alt-Tech," in *Post-Digital Cultures of the Far Right: Online Actions and Offline Consequences in Europe and the US*, ed. Maik Fielitz and Nick Thurston (Bielefeld, Germany: Transcript Verlag, 2019), 50–55.
[115] Donovan et al., "Parallel Ports," 58–61.

or permissibility. Together, however, the options for content moderation presented in this paper offer a helpful menu that companies can use to tailor their approaches and offer users a more vibrant and less toxic user experience.

The Digital Social Contract paper series is supported by funding from the John S. and James L. Knight Foundation and Meta, which played no role in the selection of the specific topics or authors and which played no editorial role in the individual papers.