BRENNAN CENTER FOR JUSTICE

Double Standards in Social Media Content Moderation

By Ángel Díaz and Laura Hecht-Felella PUBLISHED AUGUST 4, 2021

Brennan Center for Justice at New York University School of Law

Table of Contents

Introduction 3

I. Content Moderation Policies: Discretion Enabling Inequity 4

A. Terrorism and Violent Extremism Policies 5

B. Hate Speech 7

C. Harassment 8

II. Enforcement of Content Moderation Policies: Who Is Affected and How? 10

A. Latent Biases in How Offending Content Is Identified and Assessed . . . 10

B. Intermediate Enforcement and Public Interest Exceptions — Protections for the Powerful 12

III. Appeals Processes and Transparency Reports: A Start, Not a Solution 18

A. User Appeals 18

B. Transparency Reports 19

IV. Recommendations 20

A. Legislative Recommendations 20

B. Platform Recommendations 23

Conclusion 27

Endnotes 28

ABOUT THE BRENNAN CENTER FOR JUSTICE

The Brennan Center for Justice at NYU School of Law is a nonpartisan law and policy institute that works to reform, revitalize — and when necessary defend — our country's systems of democracy and justice. The Brennan Center is dedicated to protecting the rule of law and the values of constitutional democracy. We focus on voting rights, campaign finance reform, ending mass incarceration, and preserving our liberties while also maintaining our national security. Part think tank, part advocacy group, part cutting-edge communications hub, we start with rigorous research. We craft innovative policies. And we fight for them — in Congress and the states, in the courts, and in the court of public opinion.

STAY CONNECTED TO THE BRENNAN CENTER

Visit our website at
www.brennancenter.org

Introduction

Social media plays an important role in building community and connecting people with the wider world. At the same time, the private rules that govern access to this service can result in divergent experiences across different populations. While social media companies dress their content moderation policies in the language of human rights, their actions are largely driven by business priorities, the threat of government regulation, and outside pressure from the public and the mainstream media.[1] As a result, the veneer of a rule-based system actually conceals a cascade of discretionary decisions. Where platforms are looking to drive growth or facilitate a favorable regulatory environment, content moderation policy is often either an afterthought or a tool employed to curry favor.[2] All too often, the viewpoints of communities of color, women, LGBTQ+ communities, and religious minorities are at risk of over-enforcement, while harms targeting them often remain unaddressed.

This report demonstrates the impact of content moderation by analyzing the policies and practices of three platforms: Facebook, YouTube, and Twitter.[3] We selected these platforms because they are the largest and the focus of most regulatory efforts and because they tend to influence the practices adopted by other platforms. Our evaluation compares platform policies regarding terrorist content (which often constrict Muslims' speech) to those on hate speech and harassment (which can affect the speech of powerful constituencies), along with publicly available information about enforcement of those policies.[4]

In section I, we analyze the policies themselves, showing that despite their ever-increasing detail, they are drafted in a manner that leaves marginalized groups under constant threat of removal for everything from discussing current events to calling out attacks against their communities. At the same time, the rules are crafted narrowly to protect powerful groups and influential accounts that can be the main drivers of online and offline harms.

Section II assesses the effects of enforcement. Although publicly available information is limited, we show that content moderation at times results in mass takedowns of speech from marginalized groups, while more dominant individuals and groups benefit from more nuanced approaches like warning labels or temporary demonetization. Section II also discusses the current regimes for ranking and recommendation engines, user appeals, and transparency reports. These regimes are largely opaque and often deployed by platforms in self-serving ways that can conceal the harmful effects of their policies and practices on marginalized communities. In evaluating impact, our report relies primarily on user reports, civil society research, and investigative journalism because the platforms' tight grip on information veils answers to systemic questions about the practical ramifications of platform policies and practices.

Section III concludes with a series of recommendations. We propose two legislative reforms, each focused on breaking the black box of content moderation that renders almost everything we know a product of the information that the companies choose to share. First, we propose a framework for legally mandated transparency requirements, expanded beyond statistics on the amount of content removed to include more information on the targets of hate speech and harassment, on government involvement in content moderation, and on the application of intermediate penalties such as demonetization. Second, we recommend that Congress establish a commission to consider a privacy-protective framework for facilitating independent research using platform data, as well as protections for the journalists and whistleblowers who play an essential role in exposing how platforms use their power over speech. In turn, these frameworks will enable evidence-based regulation and remedies.

Finally, we propose a number of improvements to platform policies and practices themselves. We urge platforms to reorient their moderation approach to center the protection of marginalized communities. Achieving this goal will require a reassessment of the connection between speech, power, and marginalization. For example, we recommend addressing the increased potential of public figures to drive online and offline harms. We also recommend further disclosures regarding the government's role in removals, data sharing through public-private partnerships, and the identities of groups covered under the rules relating to "terrorist" speech.

I. Content Moderation Policies: Discretion Enabling Inequity

This section compares platform rules governing terrorist content, hate speech, and harassment. These policies showcase how content moderation rules are designed to give the platforms broad discretion, which can translate into inequitable enforcement practices that leave marginalized groups at risk while protecting dominant groups and their leaders. In addition, we find that platforms seem to choose to interpret their rules in ways that allow them to avoid politically charged removals and to deflect questions about unequal enforcement.

A Brief Overview of Content Moderation Policy

>> **All platforms incorporate** content moderation in some form. Without taking steps to remove some content, such as spam, they would quickly become unusable. At the same time, platforms have historically been reluctant to invest the time and resources necessary to develop comprehensive content moderation practices until forced to do so by scandal or public pressure.[5] This reluctance largely reflects a belief that content moderation is a resource drain rather than an essential part of product development.

Eventually, most companies publish a set of "rules" or "community standards" that aim to explain to their users the kinds of content that is acceptable on the platform. These policies typically contain rules against hate speech, terrorist or extremist content, nudity, and harassment (among other categories). Although the specific wording differs from platform to platform, these guidelines are frequently more similar than dissimilar. In practice, high-profile removals by one platform tend to be mirrored by other major platforms, regardless of differences in the specific language in their respective rules.[6]

The last decade has seen a clear trend across the largest platforms toward greater complexity in their content moderation policies. For example, Twitter's initial rules, published in January 2009, were less than 600 words long.[7] YouTube's first policy simply instructed users not to "cross the line."[8] Today, content moderation policies are much more comprehensive, containing specific provisions on various content categories spanning multiple web pages. Facebook's community standards evolved from a few sentences in 2011 to a multipage, encyclopedic collection of rules and blog posts by 2021. The greater diversity of speech and content available on the platforms, their increasingly global reach, and the fact that they have faced mounting public and government pressure to both moderate content and justify their decisions may explain this expansion.[9]

The major platforms rely on their policies to buttress their assertion that they operate within a rule-based system. But when confronting high-profile or difficult decisions, they often work backward to reach the desired result. High-profile removals are frequently made on an ad hoc basis, with action explained via new rules and announced in company statements spread across multiple locations ranging from company blogs to corporate Twitter accounts to third-party websites.[10] The resulting policy changes are often disjointed, unclear, or limited to addressing the narrow issue of controversy rather than encompassing a systematic reimagining.[11] As the Facebook Oversight Board pointed out in an early decision in January 2021, Facebook's practice of updating its content moderation rules through blog posts and failing to memorialize them in its community standards makes it difficult both for users to understand or adhere to company policies and for external groups to hold the platform accountable for its enforcement decisions.[12] Often, situations would have been foreseeable had teams more thoroughly addressed potential abuses of their products and services at the development and implementation stages.

A. Terrorism and Violent Extremism Policies

Despite detailed rules, blog posts, and other announcements, platform rules against terrorist and violent extremist content remain opaque, failing to provide clarity on which groups have been designated as terrorist organizations and granting the platforms immense discretion in enforcement. Largely shaped by government calls to launch an "offensive" against Islamic State of Iraq and Syria (ISIS) propaganda, these policies disproportionately target speech from Muslim and Arabic-speaking communities.[13] Imprecise rules combined with overbroad tools for mass removals have resulted in what researchers describe as "mistakes at scale that are decimating human rights content."[14]

After years of claiming that they could not reliably identify and remove "terrorist" speech, the major social media platforms responded to pressure from the U.S. and European governments to move aggressively to remove content deemed as supporting terrorism.[15] Each platform has a slightly different approach to removing terrorist and violent extremist content: Facebook removes it under its policy on dangerous individuals and organizations, Twitter under its violent organizations policy, and YouTube under its violent criminal organizations policy.[16] All these policies rely on the content being tied to a specific group, but no platform has published a list of the groups that they target.[17] In statements, the platforms have all indicated that they rely on national and international terrorism designations, such as the U.S. Department of State's list of foreign terrorist organizations or the United Nations Security Council's consolidated sanctions list, in addition to their own designations.[18] Reliance on these lists demonstrates the political nature of the removals, as well as the risk of disparate impact on particular communities. Many U.S. and international sanctions lists specifically target al-Qaeda, the Taliban, and ISIS, making it likely that over-removals will disproportionately affect Muslim and Middle Eastern communities. Moreover, because the platforms' full lists (which might well extend beyond the sanctions lists) are not public, users do not necessarily know that the organization they are posting about is on a list of prohibited organizations — a failing highlighted by the Oversight Board constituted by Facebook in late 2018 to review its content moderation decisions.[19]

Once an organization falls under the terrorism policies' umbrella, the platforms typically remove content that "promotes" or expresses "support or praise for groups, leaders, or individuals involved in these activities."[20] These imprecise terms are inevitably applied in a broad-brush manner that captures general sympathy or understanding for political viewpoints as well as news reporting and research projects. For example, in multiple reported instances, Facebook has erroneously deleted news articles and suspended accounts of journalists and human rights activists, including at least 35 accounts of Syrian journalists in the spring of 2020 and 52 Palestinian activists in a single day in May 2020.[21] In a 2018 letter to Facebook, Fionnuala Ní Aoláin, the UN special rapporteur on the promotion and protection of human rights and fundamental freedoms while countering terrorism, similarly warned against "the use of overly broad and imprecise definitions as the basis for regulating access to and the use of Facebook's platform as these may lead to indiscriminate implementation, over-censoring and arbitrary denial of access to and use of Facebook's services."[22] While Facebook claims that it allows some breathing room for discussion, the policy requires individuals to "clearly indicate their intent" in order to avoid removal;[23] this appears ill-suited to accommodate the reality of political discussion, which is rarely black and white.

None of the platforms' enforcement reports distinguish between terrorist content itself and so-called glorification of terrorism content, instead providing overall numbers on removals under their broader policies.[24] Since all three platforms are founding members of the Global Internet Forum to Counter Terrorism (GIFCT), a collaborative effort to combat online terrorist content through information sharing and automated removals, their enforcement probably aligns largely with the content flagged in the database — more than 70 percent of which is removed as glorification of terrorism.[25] As with the platforms' policies, GIFCT's approach is vague: glorification is defined as content that "glorifies, praises, condones or celebrates attacks after the fact."[26] However, unlike the secrecy that shrouds platform designations, GIFCT claims that it uses only the UN Security Council's consolidated sanctions list (although it did make ad hoc designations for attacks in Christchurch, New Zealand, and Halle, Germany, in 2019, and Glendale, Arizona, in 2020).[27]

By choosing to rely on prohibitions of expansive categories like "support" and "glorification," platforms have established a regime in which a wide range of political speech and human rights documentation is inevitably swept up in a removal dragnet. Overall, platform policy regarding "terrorist" content pays little heed to nuance and context, willingly accepting errors that affect communities with little political power.

Case Study: Comparing Facebook Responses to ISIS and White Supremacist Content

>> **Historically, Facebook's** dangerous individuals and organizations policy has primarily focused on targeting content from terrorist groups like ISIS and al-Qaeda, with seemingly little concern for how these removals potentially limit ordinary Muslim users whose content can be too readily misinterpreted as glorification, support, or praise of these groups.[28] In a report analyzing the effects of "extremist" removals, researchers documented how this approach resulted in the removal of a Facebook group supporting independence for the Chechen Republic of Ichkeria, groups advocating for an independent Kurdistan, and the account of a prominent Emirati journalist who posted satirical commentary criticizing a Hezbollah leader.[29] In each of these cases, errors were prompted by an inability to account for context and intent, reflecting Facebook's broad approach to removals that accepted mistakes as acceptable trade-offs in exchange for rapid and sweeping enforcement. These mistakes were mostly kept from public view until 2019, when Facebook finally started publishing information about errors and appeals regarding its dangerous individuals and organizations policy.[30]

After a white supremacist livestreamed his attacks on mosques in Christchurch, New Zealand, Facebook announced that it was beginning to enact additional measures to combat white supremacy.[31] However, this rollout struck a different tone than was applied for combating groups like ISIS and al-Qaeda.[32] Facebook clarified that it was not attempting to ban "American pride" or limit people's ability to "demonstrate pride in their ethnic heritage."[33] Instead, the company said it was banning the "praise, support and representation of white nationalism and white separatism."[34] While the company's initial announcement said that it was banning more than 200 organizations under the policy, later posts significantly narrowed the scope of the removals to around 12.[35]

In practice, Facebook's new rule on white supremacy was narrow: only posts explicitly calling for white nationalism or white separatism were subject to removal. Facebook's 2020 Civil Rights Audit confirmed this approach, noting that the policy did not capture content that "explicitly espouses the very same ideology without using those exact phrases."[36] Despite multiple blog posts and an analysis by Facebook's civil rights auditor Laura Murphy, white supremacy as a concept was nowhere to be found within Facebook's community standards until an update in June 2021.[37] Now, white supremacy is listed as an example of a "hateful ideology," alongside Nazism, white separatism, and white nationalism, but it remains undefined — making it difficult to assess the rule's scope or impact.

Unlike the bans against posts praising ISIS, taking a meaningful stand against white supremacy would require Facebook to remove content from users with powerful political support or with significant followings within the United States. In one instance, Facebook's own policy team recommended adding Alex Jones, who regularly targeted Muslim and transgender communities, to its list of dangerous individuals and organizations.[38] According to *BuzzFeed News*, Facebook CEO Mark Zuckerberg rejected the proposal, claiming that he did not view Jones as a hate figure.[39] Instead, he called for a more "nuanced" strategy, according to a Facebook spokesperson, resulting in Jones's suspension but not the more systemic removal of his account or content from people who "praise" or "support" him.[40]

As part of its latest update, Facebook disclosed that it uses a three-tiered system for its dangerous organizations and individuals policy. Tier 1 covers "terrorism, organized hate, large-scale criminal activity, mass and multiple murderers, and violating violent events." Facebook bans all praise, support, and representation of these groups and activities. Tier 2 covers "violent non-state actors" that do not generally target civilians. For these groups, Facebook removes support and representation, but only removes praise that is specific to an entity's violence. Finally, Tier 3 covers "militarized social movements," "violence-inducing conspiracy networks," and "hate banned entities."[41] Tier 3 entities are only prohibited from representation; praise and support of these organizations does not violate the policy. It would appear that individuals like Alex Jones and organizations like QAnon fall into Tier 3, meaning they are subjected to a narrower set of restrictions than what is reserved for organizations like ISIS.

These attempts to meaningfully target narrow and individualized enforcement stand in stark contrast to sweeping removals that are typically reserved for "dangerous organizations" that come from marginalized communities. Moreover, they demonstrate the extent to which detailed rules around enforcement mean little when they can be adjusted on the fly to accommodate powerful individuals and their followers.

B. Hate Speech

According to the three platforms, hate speech is one of the most difficult areas of content moderation because of the complexities inherent in developing a definition of hate speech that can be applied globally and at scale.[42] The companies struggle to address regional and cultural nuances, dog whistles, and terms that have been reclaimed by marginalized groups. These complexities are aggravated by an eagerness to expand into parts of the world where they have no local moderator expertise and may not even offer translations of their community rules into the local language.[43] Over time, the companies have developed elaborate systems for assessing hate speech based on defined protected characteristics, but these systems also allow for immense discretion in enforcement,[44] and they often result in over- and under-removals of hate speech. As described in the case study below, this constrains the ability of marginalized groups to use the platforms.

Facebook's policy on hate speech is the most complex of the three platforms, involving a three-tiered system of prohibited content that distinguishes among protected characteristics (i.e., race, ethnicity, national origin, religious affiliation, sexual orientation, caste, sex, gender, gender identity, and serious disease or disability) and quasi-protected characteristics (i.e., age and immigration status).[45] Both Twitter's and YouTube's protected characteristics lists overlap significantly with Facebook's.[46] Twitter's list considers age on par with the other protected attributes, and attacks based on immigration status are not protected.[47] YouTube's list includes victims of a major violent event as well as veteran status.[48]

Case Study: Twitter's Failure to Protect Black Twitter Users and Its Light-Touch Approach to QAnon

>> **The reasoning behind** Twitter's hateful conduct policy is among the most well-articulated of the three companies, explicitly acknowledging the effects of hate speech on marginalized communities, particularly those who identify with multiple underrepresented groups.[49] Nevertheless, like the other platforms, its enforcement efforts have failed to protect people of color, leaving up content from influential accounts that have repeatedly been connected with everything from online harassment to offline violence.

In 2014, for example, a group of Black women uncovered and fought a coordinated harassment campaign targeting them.[50] A group of 4chan users had created a number of fake Twitter accounts impersonating Black feminists, attempted to use Black slang, and promoted a fake Twitter campaign to #EndFathersDay. When Twitter failed to provide adequate support, these women fought back against the campaign with their own hashtag #YourSlipIsShowing.[51] The women organized and shared block lists and helped prevent other Black women from being driven off the platform due to the tireless racist and misogynistic attacks.[52]

Twitter also took an incremental approach to the conspiracy theory QAnon, despite a long history of its followers posting incendiary, threatening, racist, and abusive speech that seemed to blatantly violate the company's rules.[53] Other platforms acted in 2018 (Reddit) and mid-2019 (YouTube) to close QAnon-related forums and remove the group's content.[54]

When Twitter finally acted in July of 2020, it did not enforce its existing prohibitions on hate speech or harassment. Instead, the company suspended several QAnon accounts under a new "coordinated harmful activity" designation added to Twitter's rules in January 2021.[55] Under this new rule, Twitter said that it would now act when it found "both evidence that individuals associated with a group, movement, or campaign are engaged in some form of coordination and that the results of that coordination cause harm to others."[56] It would not typically remove content or suspend accounts but instead limit the conspiracy theory's appearance in Twitter's trends and recommendations features, as well as its appearance in search results.[57] It was only after the January 6 attack on the U.S. Capitol that Twitter "began permanently suspending thousands of accounts that were primarily dedicated to sharing QAnon content."[58]

One Twitter employee told the *Washington Post*, "whenever we introduce a change to our policies, we can look back and wish that we'd introduced it earlier."[59] At the same time, Twitter's hesitance to act could also be explained by the growing affinity for QAnon expressed by several prominent American politicians, from former President Trump to various members of Congress.[60]

In designing their response to hate speech, platforms do not fully incorporate power dynamics into their rule construction, which can lead to bizarre and illogical outcomes. For example, an internal Facebook training document from 2017 revealed that out of three groups — female drivers, Black children, and white men — only white men would be protected under the company's hate speech policy.[61] The rationale was that both race (white) and sex (male) are protected characteristics, whereas the other examples included quasi- or nonprotected characteristics, namely age (in the Black children example) and driving (in the female drivers example).[62] Implementation of Facebook's hate speech policy in this manner resulted in the platform reinforcing protections for white men, a dominant group, while failing to address speech targeting more vulnerable communities (women and Black people). In response to media fallout following the release of the training documents, Facebook announced that it had reformed its hate speech enforcement systems to deprioritize comments about "Whites," "men," and "Americans."[63] However, it did not change its underlying policies, nor did it fully show how it implemented these updates or how they were evaluated for success.

C. Harassment

Harassment is one of the most omnipresent elements of online life. According to the Pew Research Center, 41 percent of U.S. adults have experienced online harassment; women are more than twice as likely to report that their most recent experience was very or extremely upsetting, and roughly half of Black and Hispanic individuals subjected to online harassment reported it being tied to their race or ethnicity, compared with 17 percent of white targets.[64] Amnesty International has tracked Twitter's efforts to address online abuse against women since 2017, time and again reporting that the platform has failed to adequately protect female users — particularly those with intersectional identities — noting LGBTQ+ women, women with disabilities, and women from ethnic or religious minorities are disproportionately harmed by abuse.[65]

Online harassment can cause offline harms ranging from doxxing (publicly revealing private information about an individual with malicious intent) to violence, but it can also lead to online harms such as causing people to withdraw from social media or to self-censor around certain topics.[66] As a report from PEN America notes, individual harms stemming from harassment "have systemic consequences: undermining the advancement of equity and inclusion, constraining press freedom, and chilling free expression."[67] Offline, the law acknowledges the collateral repercussions of unchecked harassment. For example, people are not only protected from a lunch counter denying them service on the basis of race but also protected against hecklers looking to interfere with their equal access to the restaurant.[68] The state of comparable protections online remains underdeveloped.

Platform rules against harassment acknowledge the impact of online harassment but balance it — in different ways — against freedom of expression. For example, Facebook says that harassment "prevents people from feeling safe" but that it also wants to ensure that people can share "critical commentary of people who are featured in the news or who have a large public audience," positing the two values as oppositional.[69] YouTube's rules do not provide a policy rationale, but its exception for "debates related to high-profile officials or leaders" suggests that it too attempts to balance open debate and user safety.[70] Twitter, on the other hand, says that freedom of expression "means little as an underlying philosophy if voices are silenced because people are afraid to speak up"; it weighs that value against the platform's interest in "direct access to government and elected officials, and maintaining a robust public record [that] provides benefits to accountability."[71] According to Twitter, "insults may be present in tweets related to heated debate over matters of public policy," but the company is more likely to remove a tweet targeting a "private individual" without relevant political context.[72] Platforms rightfully provide more breathing room for criticism of public figures, but only Twitter specifies who qualifies as a public figure and addresses how these same individuals can have a stronger ability to indirectly prompt their users to harass individuals. In June 2021, Facebook agreed to implement an Oversight Board recommendation to clearly define its differing approaches to moderating harassment against public figures and private individuals, as well as to provide additional information on how the company defines these user categories.[73]

While the companies' harassment policies incorporate a wide range of prohibitions, they provide considerable discretion for platforms to choose when and how they will act. This indeterminacy can leave users unsure of both what is actually prohibited in practice and who must report a violation in order for platforms to act. For example, Facebook and YouTube prohibit not only direct threats and incitement but also "repeatedly" making unwanted advances, abusive behavior, or "inciting hostility."[74] While Twitter does not have a comparable prohibition, it does prohibit "excessively aggressive insults."[75] These terms are not well defined, and none of the platforms specify what level of repeated or aggressive attacks merit action (or the enforcement it will trigger). Similarly, while platforms acknowledge the importance of context and assessing patterns of behavior, it remains unclear how platform moderators acquire the necessary context to make accurate determinations. Instead, Twitter and Facebook's policies say that in "certain" circumstances, they "may" need

to hear from the person being targeted to "understand context" or "understand that the person targeted feels bullied or harassed."[76] But they do not specify the circumstances that require self-reporting, and their policies do not disclose other ways that platforms may receive relevant context. At the same time, in response to a recommendation from the Oversight Board, Facebook indicated that it would assess the feasibility of providing moderators with more social and political context when applying the harassment policy against private individuals. While the company acknowledged the need to balance "speed, accuracy, consistency, and non-arbitrary content moderation," it remains unclear how it will weigh these principles.[77]

The platforms' rules against harassment come closest to articulating the competing interests of freedom of expression and user safety, but they do not explain how the companies assess these competing interests. Instead, the rules contain considerable leeway for the platforms to decide at what point repeated or excessive harassment merits immediate enforcement. As with hate speech and terrorist content, platforms regularly use this discretion to delay action against powerful figures, whereas they seem to apply a trigger-happy approach to speech from marginalized groups using aggravated language to speak out against injustice.

The ever-expanding list of content rules can obscure the reality that platforms retain tremendous discretion. Moreover, content policy is often imprecise and broad when applied against marginalized communities, yet narrowly drafted and interpreted when it concerns dominant groups. This disparity sets the groundwork for an online ecosystem that reinforces existing power dynamics, leaving marginalized communities simultaneously at risk of removal and over-exposed to a number of harms.

Case Study: YouTube's Checkered Approach to Harassment by Far-Right Personalities

>> **In late May 2019,** Carlos Maza, a former *Vox* journalist, posted a Twitter thread describing consistent racist and homophobic harassment he endured at the hands of popular alt-right YouTuber Steven Crowder and his followers.[78] Crowder repeatedly called Maza a "lispy sprite," a "little queer," "Mr. Gay Vox," and "gay Mexican" in videos that were watched millions of times.[79] Maza reported that these attacks led to doxxing, death threats, and targeted harassment from Crowder's followers,[80] who also harassed Maza on Instagram and Twitter and via a torrent of phone calls and texts.[81] In response to Maza's Twitter thread, Crowder posted a response on YouTube defending his show as political humor and disavowing harassment by his followers.[82]

In June 2019, YouTube published an initial response on Twitter asserting that "while we found language that [Crowder used] was clearly hurtful, the videos as posted don't violate our policies."[83] The platform found it convincing that Crowder had not explicitly instructed his viewers to harass Maza — even though that was not a necessary element of YouTube's harassment policy, which at the time banned "content or behavior intended to maliciously harass, threaten, or bully others."[84]

Only after LGBTQ+ organizations, YouTube employees, politicians, industry executives, and others criticized its decision did YouTube choose to demonetize Crowder's account — not for the harassment of Maza but for the sale of "Socialism is for F*gs" T-shirts on his channel.[85] This decision meant that Crowder would no longer be able to make money off advertising from his channel. Subsequently, in response to continued public outrage, the platform followed up by saying that Crowder would have to address other unspecified "issues" with his account. YouTube said it came to this decision after "further investigation" of the account revealed a "pattern of egregious actions" that harmed an unspecified "broader community."[86]

However, in August 2020, YouTube announced that it was lifting the penalty because of Crowder's improved behavior, seemingly overlooking the fact that Crowder's videos over the previous year included one calling the Black Lives Matter movement a "domestic terrorist organization" and another titled "WHEN TRANSGENDERS ATTACK!"[87]

The public outcry around Maza's experience triggered a new and expanded harassment policy, which YouTube unveiled in December 2019.[88] This new policy added a new "creator-on-creator" harassment policy that prohibited "demeaning language that goes too far."[89]

YouTube's creation of a new rule to address harmful content already falling within the scope of its existing harassment policy exemplifies the ways in which platforms attempt to deflect criticisms of their disparate enforcement practices by announcing new policies. YouTube had rules and enforcement mechanisms that it could use against Crowder, but it chose to exercise its discretion to allow him to continue violating the rules.

II. Enforcement of Content Moderation Policies: Who Is Affected and How?

As with the specific language of their content policies, the companies' enforcement mechanisms operate under the veneer of a rule-based system but leave platforms with immense flexibility. From how violating content is identified to how it is sanctioned, enforcement decisions reflect judgments that often take a deferential approach for the powerful and appear to allow more room for error when removing the speech of the marginalized voices.

For example, Black activists have long complained that their attempts to post examples of attacks targeting them have resulted in enforcement actions. In one instance in 2019, Facebook removed a post by Black Lives Matter organizer Tanya Faison that read "Dear white people, it is not my job to educate you or to donate my emotional labor to make sure you are informed [. . .]"; in another, the company suspended the Black Lives Matter Minneapolis Facebook page when organizers spoke out against a police officer who suggested that drivers run over protestors to keep traffic flowing.[90] Some users have succeeded in getting enforcement errors overturned — especially if they were able to elicit press attention — but others, such as the activist collective Hood Communist, were unsuccessful in getting either their Twitter accounts reinstated or adequate explanations as to why their accounts were actioned.[91] Similarly, in the midst of a collective reckoning around sexual harassment and abuse in the workplace in 2017, several women saw their accounts suspended for posting about their experiences and saying variations of the phrase "men are scum."[92]

The problem is not limited to the United States. Earlier this year, for example, Palestinians and their allies faced widespread removals, account suspensions, and hashtag and livestream blocks as they sought to document abuses by Israeli government forces stemming from evictions in Jerusalem's Sheikh Jarrah neighborhood.[93] In one instance, Instagram removed or blocked posts using hashtags related to al-Aqsa Mosque as Israeli police stormed the complex while Palestinians were congregating there to observe the last Friday of the holy month of Ramadan.[94] In another, Twitter suspended the account of a journalist documenting protests in East Jerusalem.[95] Platforms said that these removals were due to glitches or technical errors, but their collective effect was to suppress Palestinian speech.

This section analyzes platforms' enforcement mechanisms and how companies' decisions can harm marginalized communities, downplaying the threats facing them while employing tools and methods that can limit their ability to organize, comment, and document threats to their safety. Because of the difficulties in obtaining detailed, empirical data from the platforms themselves beyond the general statistics in their transparency reports, much of the following analysis relies on user reports, independent audits, civil society research, and investigative journalism.

A. Latent Biases in How Offending Content Is Identified and Assessed

Facebook, Twitter, and YouTube rely on a combination of user reports, trusted flaggers, human reviewers, and automated tools to identify content that violates their policies.[96] Whether content is assessed by a human or by a machine, bias can creep in in numerous ways.

Platforms employ automated processes at various stages of policy enforcement, although no one outside of company walls knows the extent to which each platform relies on these systems for removals. Each piece of content uploaded to a platform is typically prescreened against a database to determine if it matches content previously removed as terrorist content or pertaining to child exploitation, among other categories.[97] The platforms continuously use additional automated systems in a number of contexts to try and identify content that violates community standards, but the complete universe of policies enforced via algorithm is unknown.[98]

Of the three platforms, Facebook's methods of identifying violating content are the most accessible, in part due to a Data Transparency Advisory Group report that the company commissioned in May 2018.[99] If Facebook's automated systems flag a piece of content as "clearly" in violation of the rules, it may be removed without a human reviewer. If the algorithm is uncertain, the content is forwarded to a human for review.[100] When a user reports content, it is typically routed through an automated system that determines whether it should be taken down or sent to a human reviewer based on the same standards.[101] It is unclear if content identified by trusted flaggers is put through automated review or if it

gets escalated for higher-level review.

Whether any of the platforms have subjected their automated tools to peer review, third-party audits, or any form of independent review for accuracy or efficacy is unclear. For example, despite Facebook's assurances that only "clear" violations are automatically removed, in practice, there is little understanding of the confidence thresholds that trigger automatic removal, how often human moderators override algorithmic determinations, or whether automated tools are being employed globally despite limited training data across languages and cultures. Having often touted their automated tools, when the Covid-19 pandemic decreased the role of human reviewers, the three platforms acknowledged that increased reliance on automated tools would lead to more errors.[102]

The automated processes for reviewing and removing content reflect a series of choices that can put speech from marginalized communities at risk of over-removal. For example, upload filters are typically applied to content related to copyright, nudity, and suspected terrorism, reflecting a decision on the part of the platforms to accept errors in the name of rapid enforcement. Automated filtering largely relies on a hashing system: a piece of offending content is "fingerprinted" so that duplicates of the content can be more rapidly removed.[103] Of course, this system may miss situations in which the image or video is being shared for purposes such as criticism, artistic expression, satire, or news reporting. As explained by one Twitter employee who works on machine learning issues, the application of content filters requires accepting that innocent users will be inevitably swept into the dragnet.[104] Speaking in an individual capacity, he explained that while the company viewed possibly restricting ordinary Arabic-speaking users and journalists as an acceptable trade-off in the fight against ISIS, deploying a similar tactic to fight white supremacy in a manner that would constrain American users and Republican politicians was not.[105]

Platforms have also expanded their use of automated tools to incorporate natural language processing. At a high level, this process relies on a series of computer science techniques for analyzing text and making predictions about it.[106] For purposes of content moderation, natural language processing tools are trained to guess whether a piece of content is, say, hate speech or harassment. However, studies into the use of automated tools for identifying hate speech have found that these systems can amplify racial bias. In one study, researchers found that models for automatic hate speech detection were 1.5 times more likely to flag tweets written by self-identified Black people as offensive or hateful, with tweets written using African American English "more than twice as likely to be labeled as 'offensive' or 'abusive.'"[107] Another study analyzed five widely used data sets for studying hate speech and found "consistent, systemic and substantial racial biases in classifiers trained on all five data sets."[108] That study reported that "in almost every case, black-aligned tweets are classified as sexism, hate speech, harassment, and abuse at higher rates than white-aligned tweets."[109] Although these analyses were not conducted on the proprietary tools employed by Facebook, Twitter, or YouTube, they illustrate issues common across the academic approaches to using automated tools to identify hate speech.

One reason that automated tools for speech removals can be biased against people of color is that they are ill-equipped to understand nuances in context. For example, the systems may not know that a term considered a slur when directed toward a marginalized group can be used neutrally or positively among members of the targeted group. For the automated tools to make their assessments, they must be trained on large amounts of data labeled by humans as either belonging or not belonging in a given category. In some circumstances, the human labelers are not given the context necessary to make informed determinations of whether a piece of content is likely to be offensive. One set of researchers found that simply providing human annotators with additional context about the identity of the poster or the dialect they were writing in could decrease the likelihood that tweets written in African American English would be labeled as offensive.[110] In other circumstances, humans helping code content are pressured by their employers to label content in a manner that falls in line with a majority view or else risk sacrificing their wages.[111] As a result, these individuals may choose not to flag instances where, counter to the status quo, they believe an algorithm is making incorrect assessments about a piece of speech, allowing the system to encode incorrect lessons.

Bias can also work its way into automated systems because many natural language processing tools developed in one language may not perform as well when applied to other dialects or languages. This risk is amplified when the languages have a smaller online footprint in terms of available training data, such as Bengali, Indonesian, Punjabi, Cebuano, and Swahili.[112] Unaddressed, this paucity of data for the algorithms to learn from can lead to the development of automated tools with higher error rates, raising the risk of over-removals of content from the people who speak these languages.[113] Unfortunately, minimal information is available about the systems that Facebook, Twitter, and YouTube employ, making it difficult to understand questions such as what languages their algorithms are trained in or how they address latent biases against users speaking in dialects or less common languages.

Human content moderation can also be biased. For example, human moderators face tremendous pressure to rapidly finish their queues and may not have adequate time or context to arrive at correct assessments, making it likely that they will have to let their own biases or intuitions guide them — consciously or unconsciously.[114] This pressure can put content from marginalized communities

at risk, given the extent to which these communities can use language in novel ways or attempt to reclaim slurs used against them. In addition, despite a global footprint and dispersed workforce, platforms have entered into markets where they lack moderators with local expertise or even appropriate language skills to make accurate determinations about content.[115] Here again, human moderation can be prone to error and bias, as moderators are forced to make determinations without the necessary knowledge or context to make accurate decisions.

Whether content is assessed by human or machine, platforms' enforcement mechanisms are often ill-equipped to understand political developments, shifts in language, or other emerging trends that are essential to proper content moderation, especially when moderating speech from marginalized groups.[116] Given these areas where bias can creep in, the need for independent assessments of platform systems being used live and at scale is urgent and overdue.

B. Intermediate Enforcement and Public Interest Exceptions — Protections for the Powerful

Historically, platform approaches to content moderation were binary: take it down or leave it up. However, Facebook, Twitter, and YouTube have all begun expanding their repertoires to encompass a broader range of enforcement actions. Although the rules are not applied equally and mistakes are common, platforms can comfortably point to the scale of their operations and say that they will inevitably make mistakes. This posture becomes more complicated when influential figures with large followings regularly break the rules with impunity. This section analyzes intermediate enforcement actions and public interest exemptions, finding that they can be used to avoid making politically charged decisions, ultimately allowing the platforms to take a permissive posture that can amplify harms to marginalized communities.

1. Intermediate Enforcement

Intermediate enforcement options vary by platform but can range from disqualifying a person from making money off their content (demonetization) to decreased distribution of their content (downranking) to adding restrictions or warnings to their posts (warning labels).[117] Some measures are more readily visible than others. For example, a user is likely to be notified when they are disqualified from running ads on their YouTube channel, but they may never know that their content is receiving less distribution or is disqualified from a recommendation engine because it repeatedly came "close" to breaking existing rules. Even less transparency is available at a systemic level. Currently, platform transparency reports do not account for intermediate enforcement measures or adequately define how platforms make determinations for what qualifies as "borderline content," making it difficult to understand how often these enforcement options are applied or if they are having a disparate effect on certain communities. At best, platforms issue statements about the effectiveness of their measures without allowing independent verification.[118]

a. Warning Labels

One increasingly popular intermediate enforcement measure is the discretionary use of labels and click-through screens (called interstitials). Particularly in cases where the harms of leaving a post up are more attenuated or where a platform wants to slow the spread of content, labels and interstitials give platforms a way to still moderate content while prioritizing freedom of expression. However, they often deploy these measures as a means of sidestepping politically fraught content moderation, which amounts to a protection for the powerful.

For example, Facebook adopted labels to avoid controversial moderation choices in the lead-up to the 2020 U.S. presidential election. Facing vigorous criticism from both parties for its failure to moderate election content — with Democrats arguing that it was not doing enough to stop the spread of misinformation and Republicans arguing that it was censoring conservative viewpoints — Facebook reportedly added warning labels to more than 180 million posts between March and November 2020, while removing 265,000 for violating its policies on voter suppression.[119] Unfortunately, these labels seemed designed to fail. Oftentimes they contained generic warnings or simply directed users to election results.[120] As several civil society groups pointed out, in choosing to label all election-related posts uniformly, regardless of their potential to cause harm, Facebook essentially made the labels meaningless.[121] News reports alleged that Facebook knew the labels did little to stop the spread of misinformation.[122]

Platforms are missing an opportunity in not harnessing the potential of labels and interstitials to serve as proportionate interventions. While intermediate enforcement tools might remediate some of content moderation's over- and under-inclusivity issues, they appear to be used merely as tools for tempering controversial decisions.

b. Feed Ranking and Recommendation Engines

Facebook, Twitter, and YouTube all use automated systems to rank, organize, and recommend content. This curation is an essential aspect of what makes their platforms usable

and engaging. These invisible decisions highlight posts from particular accounts, recommend groups to join or content to share, and facilitate targeted advertising. These systems dictate which posts get wide distribution and which ones remain in obscurity. Receiving favorable treatment by platform algorithms is essential for gaining notoriety. These systems thus play a decisive role in shaping what kinds of content people are incentivized to post. They are also a form of content moderation.

Whether a piece of content will be engaging and encourage people to spend more time on the platform is one important factor for ranking and recommendation engines. However, people's inclination to interact with sensationalism and outrage can result in these systems learning to prioritize inflammatory content, at least up until the point that it becomes too toxic and turns away advertisers. Savvy marketers and bad-faith actors alike exploit this algorithmic preference for sensational and incendiary content.[123]

Case Study: Facebook's News Feed

>> **Facebook's news feed** feature is a collection of content posts from a user's friends, pages they follow, and groups they belong to — along with paid ads labeled as "sponsored" peppered in. The news feed also contains a number of recommendations, such as groups to join or past posts in the form of "memories" to share.

Although Facebook does not make the full details of how it decides which posts to show a user publicly available, the company has published blog posts that provide an overview of the four elements that inform its algorithmic ranking:[124]

- the inventory of available posts, which can include content a person's friends have posted or that have appeared in groups or on pages they follow;
- the signals (i.e., data points) that inform ranking decisions, derived from the account a given post was shared by, the virality of a post, the speed of a user's internet connection, or whether a piece of content has been rated as objectionable in some way;
- predictions such as how likely a user is to comment or share a post; and
- unspecified "integrity processes" to evaluate objectionability, diversity of content, etc.

Based on these elements, Facebook assigns each piece of content a relevancy score in order to rank how they appear in a user's news feed.

These explanations provide limited insight.[125] Facebook does not explain what role (if any) the various methods it employs to track users as they use the platform — and browse the internet more broadly — play. For example, it does not reference the "traits" that machine learning models can estimate, such as race, political and religious leanings, socioeconomic class, and level of education.[126] The company also conceals what role advertisements play, and how it deploys automated interface features like memories to prompt engagement with the platform. These data points are ripe for incorrect or biased assessments.

In the face of ongoing public criticism of its ranking system's role in helping spread everything from misinformation to hate speech, Facebook announced a series of changes to its ranking algorithms. In 2018, Mark Zuckerberg said the company was changing its news feed algorithm to prioritize posts from friends and family in an attempt to facilitate "meaningful social interactions."[127] However, news reports indicate that Facebook may have scrapped efforts to adjust its ranking system to create "the best possible experiences for people" because promoting civility interfered with the company's interest in growth and having people keep visiting the platform.[128] The status of the initiative is unclear.

The company recently claimed that it is also experimenting with approaches such as surveying its users and offering a more easily accessible ability to opt out of its ranking algorithm, instead allowing users to simply view posts in reverse chronological order.[129] This move appears to be part of a concerted effort to shift the focus away from its lack of meaningful transparency around algorithmic ranking by making the argument that people are choosing how to populate their feeds.[130]

Ultimately, the fact that Facebook now makes some of its algorithmic methodology public and allows users some control over what they want to see does not change the reality that the company controls how it designs its systems, which it does behind closed doors. The company can walk away from any reforms at will and is under no obligation to allow independent research to assess the impact of its decisions. Additionally, while users can tinker with some controls and features, the possibility remains that Facebook's systems will make incorrect, biased, or even discriminatory assessments about what to show its users based on inputs that users cannot control.

Case Study: YouTube's Recommendation Engine

>> **YouTube's recommendation engine** is the automated process by which the platform generates a queue of videos that its systems guess will keep a user on the platform. According to current and former YouTube engineers, the overarching purpose behind the company's algorithm is to make the platform more "sticky" — that is, a place where people spend more time.[131] In essence, the system analyzes various inputs to pull a set of videos that a given user may want to watch next and then ranks them.[132] The full scope of the inputs is unknown, but YouTube says that it analyzes a number of metrics, including the following:

- click-through rates;
- watch time and engagement metrics such as likes, dislikes, and shares;
- user history, including search history and previously watched videos;
- feedback from other users;
- users' demographic and location information; and
- how new or "fresh" a video is.[133]

The recommendation engine plays an outsize role in how people find videos. According to a 2018 statement from YouTube's chief product officer, the recommendation engine accounts for more than 70 percent of the videos watched on the platform.[134] At the same time, some investigations into the prioritization of watch time and engagement found that the recommendation engine promotes inflammatory and extreme content to users.[135]

Attempting to downplay and discredit criticism, YouTube initially said that its recommendation system had "changed substantially over time and no longer works the way it did five years ago," but there is no way to verify this assertion. The company claims that it no longer optimizes for watch time but instead focuses on "satisfaction," as measured by likes, dislikes, shares, and user surveys.[136] In a 2019 white paper detailing how it fights disinformation, the company claimed that YouTube prioritizes content by "authoritative sources" to avoid misleading users about developing news stories.[137]

Later that year, YouTube announced a series of new measures that seem to give people greater control in influencing the videos that its recommendation engine promotes.[138] These measures included the ability to remove suggestions from channels and the addition of contextual explanations as to why a particular video is being recommended.[139] According to YouTube, its measures to reduce "borderline content" (i.e., content that "comes close to" but does not violate its community guidelines) achieved "a 70% average drop in watch time of this content coming from non-subscribed recommendations in the U.S."[140] This obscure figure does not suggest a change in the recommendation of borderline content, just a change in how long certain people watch the videos that YouTube recommends. The number tells us nothing about the overall prevalence of content that the company considers "borderline," which guidelines were *almost* violated, or how the platform makes those determinations. Nor does YouTube identify the people tasked with training or evaluating its recommendation engine.

We cannot fully ascertain the effectiveness of these changes, but there is reason to be concerned. For example, YouTube says that it elevates "authoritative voices" in its search results and "watch next" panels for topics related to "news, science, and historical events," but one of the named authoritative voices is Fox News, which researchers have identified as a locus for spreading misinformation.[141] And algorithms are bound to make mistakes. Case in point: in creating an algorithmically generated knowledge panel with contextual information about the 2019 fire at the Notre-Dame cathedral in Paris, YouTube accidentally treated the incident as an act of terrorism and provided information about the 9/11 attacks.[142] In 2021, the Mozilla Foundation published a report that relied on 37,380 YouTube users who had installed a browser extension to understand how the platform's recommendation engine was operating in practice.[143] The study found that 71 percent of the content that users identified as "regrettable" (a term that allowed users to define what they believed caused harm, and contained everything from hate speech and harassment to scientific misinformation and what YouTube might consider "borderline content") came from YouTube's recommendation algorithm. The study also found that the rate of "regrets" was 60 percent higher in countries where English is not the primary language.[144]

Despite the reported success of its new measures, YouTube has not released any underlying data, such as how much time viewers still spend watching the borderline videos or how it populated its sample size. Similarly, existing independent studies operate with limited information, as YouTube does not facilitate research at even the limited scale offered by Twitter and Facebook.

As with most of content moderation's inner workings, little public information about how these systems work and how they may be affecting different groups is available. Social media platforms zealously guard their algorithms and constantly tweak them, so public understanding of how they work is largely based on platforms' own explanations, leaks from internal whistleblowers, and academic studies based on limited data.[145] After sustained criticism and pressure, platforms have slowly rolled out changes to the underlying formulas that constitute their algorithms and provided some information about them.[146] But these piecemeal disclosures and back-end updates are not subjected to independent review, making them of limited use as a measure of platform transparency and accountability — particularly in analyzing how they may be disproportionately affecting marginalized communities.

Platforms increasingly rely on intermediate enforcement measures to moderate some of the most high-profile issues facing social media platforms. At the same time, these methods appear to be employed selectively as restrained measures when traditional enforcement mechanisms would impact dominant groups. Despite its growing prominence, transparency around platforms' intermediate enforcement lags far behind their disclosures around removals, making it difficult to fully assess the scope and impact of these moderation tools.

2. Public Interest and Newsworthiness Policies

Notwithstanding their voluminous policies, platforms maintain enormous and often invisible discretion over how they employ content moderation by applying newsworthiness or public figure exceptions. The lack of specificity and transparency regarding how they deploy these exceptions creates a large gap in our understanding of how platforms approach moderating the content of their most influential users. The actions of influential accounts can be among the most consequential drivers of harassment and offline violence. We have seen this reality play out across the globe, with the social media posts of politicians, religious figures, and cultural leaders, among other high-reach accounts, linked to violent incitement in countries ranging from Myanmar to India to the United States. (The latter is discussed in a case study below.)[147]

Platforms take a lenient approach when moderating content from public figures because of the public interest in seeing and interacting with it. But the details of these policies — and how they are applied — are notoriously difficult to track, which can give the impression that they are fallbacks that platforms use to allow public figures to break the rules with impunity. Although Facebook has had an ad hoc public figure/newsworthiness policy since 2009, it took almost 10 years for the company to publicly disclose its policy in two blog posts from 2016 and 2019.[148] Under its policy, Facebook may allow content to remain on the platform even if it violates the Facebook community standards when the company believes that the public interest in seeing it outweighs the risk of harm.[149] We know little about how that determination is made, how often the policy is applied, or whether the default is to leave content alone. Users have almost no way to know the policy has been applied absent press inquiries that are answered by company representatives. Moreover, the policy is not codified in the company's community standards, which forces the public to wade through Facebook's various public statements and blog posts and third-party analyses to estimate the policy's parameters and application use cases.

This policy's opacity is illustrated by the fact that the world outside of Facebook assumed that the company had applied the public figure exception to allow content from former President Trump to remain online even when what he was posting seemed to violate its policies. In May 2021, however, Facebook told the Oversight Board that it did not exempt Trump's posts based on this policy; in fact, it told the board that it had "never applied the newsworthiness allowance to content by the Trump Facebook page or Instagram account.[150] (The company subsequently stated that it had accidentally forgotten to disclose one instance in which it did employ the policy, when Trump insulted someone at a political rally.)[151] Instead of applying a newsworthiness exemption, Facebook says it employs a newly disclosed "cross-check system" for certain accounts considered high-profile to minimize enforcement errors.[152] While it is understandable that the company would involve its senior executives in decisions about the president's account, as the Oversight Board noted in May 2021, "the lack of transparency regarding these decision-making processes appears to contribute to perceptions that the company may be unduly influenced by political or commercial considerations."[153]

In response to the Oversight Board's recommendation for more clarity about its cross-check system, Facebook merely stated that it "may employ additional reviews for high-visibility content that may violate our policies."[154] Absent from its response are essential details such as who qualifies for this cross-check system, whether every post from popular accounts undergoes this secondary review, and which individuals or teams are responsible for the additional layer of review. Also in response to a board recommendation, Facebook published a blog post on its Transparency Center blog that sought to explain its newsworthiness policy.[155] Most notably, it announced that it was "removing the presumption we announced in 2019 that speech from politicians is inherently of public interest."[156] It said that the company would "begin providing regular updates about when we apply our newsworthy allowance" but did not specify how or when. Among the considerations for evaluating newsworthiness, Facebook lists "country-specific circumstances" (for example, when

an election is underway or a country is at war)," the "nature of the speech," and the "political structure of the country, including whether it has a free press."[157] These disclosures do not connect newsworthiness with other processes, such as the cross-check system or how this policy ties to its efforts to label or downrank certain types of content.[158]

Twitter's public interest exception is accessible through its rules and policies hub.[159] It takes a more detailed approach, explaining which elected and government officials or candidates qualify and how the public will know when an exception is applied.[160]According to its policy, Twitter is supposed to apply a label that discloses when it has applied a public interest exception and specify which rule(s) the public figure violated, and users are supposed to be prompted to click or tap through before they can see the tweet.[161] Twitter claims that it rarely applies this policy, noting that it was employed fewer than five times in all of 2018.[162] Comparatively, in 2020, the company used it numerous times to mark tweets by former President Trump during the 2020 presidential election before ultimately suspending his account.[163]

Unlike the other platforms, YouTube claims that it does not have a public figure exception but does have an exception for content that is "educational, scientific, newsworthy, or a documentary."[164] However, in two competing appearances, YouTube CEO Susan Wojcicki made conflicting representations about the company's policies. In 2019, she said that politicians are exempt from some of YouTube's rules because "it's important for other people to see."[165] In 2021, she said that YouTube has no exceptions for politicians and that the rules are the same for everyone.[166] Regardless of what its actual policy is, YouTube is typically among the last of the major platforms to act against public figures, generally waiting until other platforms have taken enforcement actions.[167]

Platform approaches to public figures often seem to give these influential accounts wide latitude to break the rules with impunity. Even where companies have taken steps to write a policy for moderating public figures, a timid posture precludes their enforcement actions even having a chance to succeed. As observed by Jillian York, the director of freedom of expression at the Electronic Frontier Foundation, "When the citizen's speech is deemed less valuable than that of the politician, when the activist is silenced by state or corporate powers (or both acting in concert), the same structures that enable offline repression are being replicated online."[168]

Case Study: The Deplatforming of President Trump

>> **In the wake of the** January 6, 2021, insurrection at the U.S. Capitol, Twitter and other platforms began suspending Donald Trump's accounts for his role in inciting the violence.[169] These decisions came after years of apprehension. Among the offending posts that pushed the platforms over the edge, the former president told the insurrectionists "we love you, you're very special," called them "true patriots," and encouraged them to "remember this day forever."[170]

While there was significant posturing by the major platforms that Trump's behavior was extraordinary or unprecedented, there was in fact ample evidence going back years that Trump was violating their policies. There were numerous times where his statements seemed to clearly violate platform policies against inciting violence. For example, in May 2020, as nationwide protests were erupting in the aftermath of the police murder of George Floyd, Trump published incendiary comments across multiple platforms, describing protesters as "THUGS" and saying that "when the looting starts, the shooting starts."[171] Trump's tweet to that effect quoted a statement from a white police chief in the 1960s about cracking down against civil rights protestors and could be seen as a call for others to commit acts of violence.[172] In another case, his ongoing calls to "LIBERATE MICHIGAN" and "LIBERATE MINNESOTA" acted as a call to arms within extremist circles, perhaps even contributing to a plot to kidnap Michigan Governor Gretchen Whitmer.[173] Some members of Facebook's Oversight Board, which was charged with reviewing Trump's suspension from the platform, also noted that Trump's multiple posts regarding the "CHINA VIRUS" may have advocated racial or national hatred constituting incitement to hostility, discrimination, or violence.[174]

To be sure, Trump's racist rhetoric, lies, and bullying were important for people to see. When Trump regularly crossed the line throughout the years, it would have made sense for the platforms to consistently label his posts as breaking specific rules but note that a public interest exemption was being applied and that limits had been placed on their distribution. But the platforms put off these actions for years, and it was only in the immediate lead-up to the 2020 presidential election that they publicly considered moderating Trump's account and actually began doing so.[175]

Even in their belated enforcement, the platforms' decisions did not come from a principled standpoint and

continued on next page

continued from previous page

did not entail transparent application of their policies. Instead, they took a reactive approach that was opaque and reliant on new designations. For example, YouTube's decision to suspend Trump did not result from his support of the January 6 insurrection but from a video that was posted nearly a week after the event, in which Trump defended his rally before the riot and said the national anger was due to congressional efforts to impeach and convict him. Without specifying which rule this video broke or even identifying the video itself, YouTube announced on its Twitter account that it was removing content and applying a strike and a one-week suspension.[176] The company subsequently extended the suspension without explaining the original violation or how it would approach reinstatement, simply saying that its teams were "closely monitoring for any new developments."[177] In a panel appearance before the Atlantic Council, CEO Susan Wojcicki provided slightly more information, saying that YouTube assesses the risk of violence by looking at "signals" such as government statements and warnings, law enforcement presence, and violent rhetoric on YouTube.[178] In contrast, Twitter followed up its suspension of Trump's account with an announcement that it was permanently suspending his account "due to the risk of further incitement of violence."[179]

Reviewing Facebook's decision to suspend Trump indefinitely, the Oversight Board found that the threat of ongoing violence justified a suspension but that an "indefinite" one — a remedy not contemplated in the company's rules — was unacceptable.[180] The board demanded that the company instead "apply and justify a defined penalty."[181] It also made a number of recommendations regarding the moderation of influential accounts, directing the platform to provide appropriate resources and personnel to apply principled rules to address these individuals' unique ability to incite online and offline harms.[182] In its response to the Oversight Board's decision, Facebook set Trump's suspension at two years but reserved the right to reassess whether "the risk to public safety has receded" or if another suspension might be warranted.[183] Facebook also noted that in "extreme cases," it can permanently disable an account if the account "persistently posts any violating content, despite repeated warnings and restrictions" (among other reasons).[184]

Hateful speech from high-profile individuals has a greater potential to cause significant harm. The Brookings Institution followed the impact of three tweets the former president wrote attacking political rivals in Michigan and Ohio and federal employees in Virginia. The study found that in the aftermath of the posts, levels of severe toxicity and threats against the targeted parties spiked.[185] Furthermore, one of the Capitol insurrectionists was quoted in the *Washington Post* as saying "I thought I was following my president. . . . I thought I was following what we were called to do."[186] Others apparently hoped for a pardon before Trump left office, believing that he would protect them because they were following his "orders."[187]

Of course, this problem is not limited to social media or even to Donald Trump himself. Numerous other elected officials, such as Sens. Ted Cruz (R-TX) and Josh Hawley (R-MO), also played a role in casting doubt on the results of the 2020 election, as did traditional media outlets like Fox News, which reportedly raised doubts as to the 2020 election outcome nearly 800 times in the two weeks after the election was called for Joe Biden.[188]

III. Appeals Processes and Transparency Reports: A Start, Not a Solution

Over the last several years, social media platforms have started providing users with greater access to appeals and provided more transparency on their enforcement practices.[189] However, these efforts have fallen short when it comes to protecting users who come from marginalized communities. The appeals processes are limited to certain types of content,[190] and users often lack the information they need to meaningfully utilize them: information on specific rules violated, for example, or whether content was removed due to an algorithm or human reviewer — which could affect the likelihood that a removal was made in error.[191] These limitations have stymied the effectiveness of the platforms' appeals processes.

A. User Appeals

Despite greater access to appeals, the processes remain inconsistent and hidden from public view. Moreover, as companies shifted to utilizing fewer human reviewers during the Covid-19 pandemic, momentum around greater appeals accessibility and transparency faced a significant setback.[192] These developments placed users from marginalized communities at greater risk for adverse content moderation decisions due to the higher error rates in algorithmic versus manual takedowns — particularly when non-English speech is involved. Those users also found themselves without meaningful ways to appeal such decisions.[193] For example, in 2020, Facebook demonstrated its inability to account for local context in Nigeria when its automated tools removed posts using the #EndSARS hashtag to call attention to police attacks against protesters.[194] According to the company, its automated systems mistook this hashtag organizing against police violence for misinformation about Covid-19.[195]

YouTube was an early pioneer in the appeals space, allowing users to appeal strikes for community guidelines violations as early as 2010.[196] Twitter and Facebook lagged behind, creating opportunities for user appeals only in the last several years.[197] Facebook began allowing appeals for some takedowns only in 2018 and has slowly expanded the scope of appealable content since then.[198] It recently took the lead in creating novel appeals procedures,[199] and it is the only platform that created an independent body to hear appeals of its enforcement decisions (as well as decisions that the company itself refers to the board). The impact of this quasi-independent body remains to be seen;[200] the Facebook Oversight Board began hearing its first cases in the fall of 2020 and issued its first decisions and policy recommendations in early 2021.[201] The high overturn rate in the board's initial set of decisions indicates that Facebook's decision-making is far from sound and that appeals are necessary to correct mistakes.[202] Unfortunately, because of the board's limited jurisdiction and its ability to hear only a small number of cases each year, the venture is likely to be more impactful from a policy standpoint than as a true means of providing due process for users.[203]

Despite the push in recent years toward expanding appeals and notice to users, central elements of the process remain vague, and appeal options remain limited. YouTube has traditionally relied on human reviewers, but for Twitter and Facebook, it is often unclear when appeals are conducted by human reviewers as opposed to AI systems.[204] This ambiguity was particularly true in the midst of the Covid-19 pandemic, during which all of the platforms acknowledged having limited access to human moderators, leading to slower and reduced user appeals.[205] Facebook did start giving users the option to "Disagree with Decision Instead of Request Review" in some circumstances.[206] Essentially, though, that process just allows people to provide feedback to the company, not to receive another review of a content moderation decision.

Furthermore, not all types of rule violations can be appealed. For example, until early 2019, Twitter only allowed appeals for account suspensions or locked accounts, not for enforcement actions against specific pieces of content.[207] Currently, none of the platforms allow appeals of intermediate enforcement actions like downranking or interstitials or even notify users of such actions. Until recently, users could not appeal in cases where the platforms decided to leave up content that they reported.[208] The reality is that users are left without any recourse in a wide range of situations. Even when appeals are allowed, users rarely receive adequate notice about the rule they allegedly violated or the reasoning behind the platforms' enforcement actions. Oftentimes the platforms will identify the content as violating a broad, general policy, such as "hate speech," without providing information on what part of the rule was violated. As the

Facebook Oversight Board highlighted in early 2021, this opacity leaves users lacking the information they need to conduct a meaningful appeal.[209] It also constrains their ability to explain why they believe a decision was incorrect.

Robust appeals are an essential mechanism to safeguard the speech of all users, but they are particularly important for those from marginalized communities. Given the platforms' inconsistent, quasi-arbitrary, and obfuscated enforcement of their content moderation rules, their appeals processes are ripe for overhaul.

B. Transparency Reports

Following years of civil society advocacy, Facebook, Twitter, and YouTube now regularly issue transparency reports covering matters such as the volume of their content removals and the percentage of offending content that was identified using automated tools.[210] However, these reports largely reflect how the companies want their efforts to be evaluated: on the quantity of removals (often without providing information regarding the prevalence of various types of content) instead of their quality. As a result, transparency reports in their current form fail to provide the information necessary to evaluate who is most affected by enforcement decisions, whether some communities are disproportionately targeted by hate speech or harassment, and whether automated tools are making more mistakes when assessing certain categories of content.

One way to assess the impact of community standards enforcement is to understand who is most affected by policies that remove broader swaths of content, such as rules against terrorist content or violent extremism. Yet none of the platforms disclose the individuals or organizations covered by those policies or whether content, when removed, was pulled due to imminent threats of violence or under vague prohibitions such as "praise" or "support" of banned organizations.[211] Assessing whether these removals are unilaterally targeting a particular type of content or subset of users — and whether groups targeting marginalized populations are being ignored — is thus extremely difficult to ascertain.

We know that all three platforms rely on a combination of users, trusted flaggers, and automated tools to identify offending content, but transparency reports do not disclose enough information to assess the effectiveness or biases contained within flagging systems. Of the three platforms, only YouTube provides some insight into the identity of the trusted flaggers it relies on for a subset of its rules, and even then, the company only offers a representative sampling.[212] YouTube notes that it works with "academics, government partners, and NGOs," but it only discloses the identities of a few NGOs and academics, omitting any information about government agencies.[213] Similarly, its disclosures around the use of automated tools do not specify if those tools are only used for specific rules, how often their removals are overturned compared to human flags, and whether the tools have been subjected to independent assessments. Instead, the company touts its system's growing effectiveness with little proof beyond the volume of removals.[214]

Another way to assess the effects of removals would be to better understand the targets of hate speech and harassment. Despite detailed lists of protected categories (at times even broken down into tiers), transparency reports bundle removals by category.[215] While Twitter reports data such as the percentage increase in harassment, and Facebook recently started reporting information on the prevalence of hate speech, the reports do not indicate whether this information reflects an increase in attacks against particular groups, such as women or LGBTQ+ communities.[216] This lack of disclosure makes it impossible to assess enforcement or to understand the extent to which marginalized communities are being subjected to attacks.

Similarly, only Facebook discloses how often it restores content removed under these rules with or without an appeal.[217] By comparison, YouTube bundles appeals across all of its policies, grouping categories like spam with more complex and potentially overbroad removals under its violent extremism or hate speech policies.[218] Without more information, transparency reports offer only minimal insight into how specific rules are written or interpreted in a manner that may be causing higher error rates. Finally, information about intermediate removals, such as demonetization, labeling, or downranking, is not currently disclosed in transparency reports, making it impossible to fully assess the scope or effectiveness of these enforcement actions.

Without significant expansion of platform disclosures, transparency reports fail to provide the insights necessary to assess the effectiveness of content moderation policies and practices, and whether certain communities are bearing the burden of an unequal system.

IV. Recommendations

The social media status quo often fails marginalized communities. While this must end, we must also take care to ensure that attempts to provide solutions do not aggravate existing harms or introduce new ones. This section contains several recommendations for legislative and industry reforms. These recommendations are aimed at protecting marginalized communities by empowering government and the public with the information necessary to hold platforms accountable. They are not an exhaustive inventory of what needs to be done to address the myriad harms wrought by social media companies. But they are important steps for ensuring that society understands the disparate influence of content moderation on particular communities, which in turn will lead to better, evidence-based policymaking.

A. Legislative Recommendations

Section 230 of the Communications Decency Act of 1996 immunizes online platforms from most liability for their users' content and for their voluntary efforts to remove content that they consider objectionable.[219] It is a political lightning rod that spurred bipartisan energy around holding Big Tech accountable.[220] Reform of Section 230 is one potential path to effecting serious changes to content moderation if those reforms can wrest more transparency and accountability from platforms, as we discuss below. Alternatively, our proposed reforms could be achieved via stand-alone regulation.[221] However, some current proposals run the risk of aggravating existing inequities and further harming marginalized communities.[222]

Who Is Covered (and Who Is Not)

>> **Our proposed regulatory updates** target social media platforms that host user-generated content — the companies whose largely inscrutable decisions can have the biggest impact in shaping "the availability and discoverability of information."[223] But those companies represent only a subset of the information service providers covered by Section 230 and of those that make content moderation decisions. We recommend that the transparency and data access requirements outlined in this report should not apply to the following services or entities:

- **Online service providers at layers of the internet infrastructure other than the application layer, such as cloud services, domain name system (DNS) providers, content delivery networks, domain name registrars, and internet service providers (e.g., Cloudflare's protection against distributed denial of service attacks).** We believe that these companies should largely act as neutral conduits and avoid making content moderation decisions. These services are essential in providing the infrastructure necessary for a functioning internet, but the companies maintain a more attenuated relationship with users and can be more ill-equipped to target individual pieces of content. The concern is that greater moderation by these companies could limit the ability of individuals from marginalized communities to freely communicate using the internet.[224]

- **Traditional media outlets that offer journalistic or editorial content (not including editorial decisions by online platforms to rank and organize third-party content).** These companies operate at a smaller scale, have traditional regulatory controls in place, and are ill-equipped to make the types of disclosure determinations expected of social media platforms.

- **Applications and functionalities that enable private communications, such as email, direct messages (e.g., private Twitter accounts or Facebook posts shared with certain friends), videoconferencing services, and encrypted communication services (e.g., iMessage and Signal).** It is essential for individuals to remain free to make private communications using the internet, just as phone companies or the postal service should not deny people access to speak or correspond with others. The Covid-19 pandemic laid bare how vital the ability to communicate online is for participation in modern society.

Recommendation 1: Require Clear Content Moderation and Appeals Policies for All Platforms

Regardless of their size, it is essential that all platforms adequately invest in developing clear, consistent, and publicly accessible content moderation policies instead of relying on ad hoc decisions that are reactive to scandal and treat the harms imposed on marginalized communities as an acceptable trade-off for growth. We see this cycle repeating itself with new platforms ranging from TikTok to Substack to Clubhouse.[225] To this end, we recommend that the following requirements apply to all social media companies:

- **Public and easily accessible community standards and enforcement protocols.** First and foremost, all platforms must publish community standards that explain their rules for acceptable content and explain the steps taken in response to rule violations. These standards and protocols should be presented in one location instead of the current practice of scattering content policies across numerous blog posts, press releases, and even third-party websites.[226]

- **Moderator disclosures.** Platforms should disclose the variety of methods used to identify offending content and provide visibility into which methods are employed in which situations. This increased transparency is an essential step in allowing users and watchdog organizations alike to identify issues ranging from lack of language expertise to biased algorithms. At a minimum, transparency reports should disclose the following:

  - The identities of organizational trusted flaggers and the specific content policies for which they have special flagging privileges.[227]

  - The types of automated tools deployed to identify and remove offending content, such as the use of hashing systems and natural language processing systems. Platforms should also disclose their participation in cross-industry initiatives such as GIFCT, as well as how they respond to content flagged by the consortium (i.e., automatic removal or flagging for human review) or other similar arrangements.

  - The types of human moderation employed, including whether in-house employees or contractors perform this role and whether the platform leverages community moderators, as well as the identities of third-party vendors that assist with moderation, the number of languages in which content moderation is offered, and the extent of local expertise.

- **Right to appeal and disclosure of appeals policies.** All platforms should provide users with the right to appeal all content moderation decisions, from down-ranking to removal, as well as publicly accessible information on their appeals processes. The platforms should clearly delineate to users the reason for an enforcement decision (including the specific rule violated), how to file an appeal, what enforcement decisions can be appealed, who will review the appeal, how the decision will be delivered, and if there is any avenue for further appeal.

- **Disclosure of government requests for content removals and user information.** All platforms should disclose the number of orders and requests received from government agencies to remove content or suspend accounts and whether action was taken on the basis of law or platform community standards. These disclosures should be broken down by the agencies requesting information and should include the specific law cited in the request. Additionally, as a check against closed-door information exchange, platforms should be required to disclose any partnerships with government entities (including supra- and quasi-governmental entities) to provide or exchange information about the identities of users without following a legal process. These disclosures are essential to protect against government suppression of marginalized communities, who are often the target of state oppression. While many platforms already do this, the information they disclose is not granular enough to identify harms against marginalized groups.

- **Error rates.** Platforms should publish information about their error rates for enforcing each content policy as well as error rates for human and algorithmic moderation. This information is imperative to understanding the trade-offs that platforms accept regarding over-removal, along with which communities are disproportionately affected.

Recommendation 2: Require Mandatory Transparency Reporting for Larger Platforms

While many of the larger platforms already publish transparency reports, current self-reporting is insufficient to allow policymakers and the public to evaluate who is most affected by their decisions. To correct this imbalance, we recommend that platforms meeting the criteria outlined below be required to publish transparency reports that disclose, at a minimum, the following information in an openly licensed, cross-referenceable, and machine-readable format. We believe that the disclosures described

below represent the base level of transparency that should be required of all companies. However, we recognize potential resource limitations and for now recommend that these regulatory requirements be reserved for larger companies that meet the following threshold: platforms that have more than 30 million active monthly users in the United States or more than $500 million in global annual revenue during the most recent 12-month period. This threshold is intended to target the largest and most influential companies while giving emerging companies an opportunity to build up their content moderation practices at a pace proportionate to their scale and influence. Platforms that surpass this threshold should be required to make the following disclosures:

- **Flagging violating content.** For each specific area of their content moderation rules, platforms should disclose information on the number of posts that have been flagged as potentially violating and by whom — a user, automated tool, human moderator, or trusted flagger (and if they were an individual, government agency, or NGO). In addition, for each type of moderator, platforms should disclose the percentage of the flagged content that was determined to have violated content moderation rules.

- **Enforcement of content policies.** Currently, some platforms release data on the numbers of posts removed and accounts suspended for violation of their content policies. This requirement should be expanded to all platforms meeting the above threshold and supplemented with information on other enforcement mechanisms, such as interstitials, demonetization, and other intermediate restrictions applied. Each of these disclosures should be broken down by the specific part of the rule violated.

  - **Impact of hate speech.** To facilitate a better understanding of content moderation's reach, we recommend that platforms track and disclose their enforcement of hate speech policies by the group targeted (when one is identified). Instead of tracking users by race, gender, or other protected class — which are open to abuse across multiple contexts — tracking the targets of attacks would only require creating a record of an assessment already made by platform moderators.[228] For example, when assessing hate speech, platforms already must assess whether a post is promoting violence against individuals or groups based on characteristics such as religion, national origin, ethnicity, sexual orientation, disability, gender identity, immigration status, or race.[229] Thus, tracking these decisions should not be overly burdensome. Platforms should also disclose the type of victim targeted (e.g., whether they are a group, individual, or organization). For content enforcement related to terrorism or violent extremism, platforms should disclose actions based on the organization or individual designated as "dangerous" or "violent" that the post related to. This step will help assess whether these removals are focusing on certain users or groups.

  - **Impact of algorithmic removals.** Platforms should publish the number of removals and suspensions that occurred without human review and distinguish the type of algorithm used. This measure will help assess whether platforms are over-relying on tools that may be ill-suited for the types of assessments required.

  - **Impact of enforcement.** In addition to reporting on the amount of violating content removed, platforms should disclose the number of users who have had their accounts suspended or terminated and the reasons why. For each specific area of their content moderation rules, platforms should also report on intermediate enforcement actions, including downranking, demonetization, interstitials, and exceptions made under their newsworthiness or similar policies.

- **Appeals disclosures.** Platforms should disclose the number of user appeals per type of enforcement action (from account suspension to demonetization), as well as information about the rate of overturned decisions by type of violation. These checks are necessary to ensure equitable enforcement and to assess whether appeals are adequately robust.

Together, these proposals set a floor, not a ceiling, for transparency reporting. Platforms should consider these recommendations as setting forth a minimum amount of information that they should report, with the understanding that the particular details of their reporting may vary depending on the types of services they offer. Moreover, while these recommendations are currently proposed as mandatory for only those platforms meeting the above threshold requirements, they should inform the transparency reporting of all platforms regardless of their size.

Recommendation 3: Establish a Federal Commission to Evaluate Academic Research, Third-Party Auditing, and Independent Investigations

For too long, it has fallen to journalists and academics to uncover discrepancies in platforms' explanations of how their systems work. Platforms regularly dispute research

findings and news reports, claiming that they reflect an incomplete understanding of how their systems work, but the same companies refuse to provide the data necessary to allow informed assessments, at times even taking steps to impede ongoing research.[230] The unilateral control of the data necessary to assess social media's positive and negative effects on society is untenable.

We recommend that Congress establish a federal commission to investigate how to best facilitate the disclosure of platform data to enable independent research and auditing, protect privacy, and establish whistleblower protection for company employees and contractors who expose unlawful practices. The commission should convene a broad set of stakeholders, including representatives of impacted communities, activists, journalists, members of civil society organizations and NGOs, researchers, privacy experts, academics, platform representatives, and content moderators.

The commission should consider the following issues, among others:

- how to arrange for researcher access to data held by private companies, including whether an intermediary is needed;[231]
- how to protect the privacy and security of user data;[232]
- how to ensure that platform users are appropriately notified that their data may be shared with independent researchers for public interest research projects;[233]
- how to protect companies' proprietary information;[234]
- how to set up clear walls separating academic research from law enforcement and intelligence agencies;[235]
- the appropriate liability protections, both for the companies and the researchers, required to enable public interest research; and
- whether, if a regime is proposed or created by the commission to facilitate the disclosure of platform data for independent research and auditing, there should be whistleblower protections to allow employees and contractors at platform companies to share information about unlawful actions pertaining to the privacy and security of user data.[236]

Clear legal obligations and protections are necessary first steps to ensure that external research and evaluation can play an informed role in improving online discourse and protecting groups that are disproportionately harmed by opaque and inconsistent content moderation enforcement. As noted in one research institute's analysis of how social media companies responded to the first 100 days of the Covid-19 pandemic: "Without better access to data and insight on companies' decision-making systems, both human- and machine-led, we cannot determine with certainty why some areas of policy appear more effective or better enforced than others. . . . [A]ny conclusions drawn must rely on some element of extrapolation and inference."[237]

B. Platform Recommendations

This section contains recommendations for platforms to do more to protect marginalized communities. First, we make a deceptively simple proposal: changing content moderation policies and practices to meaningfully integrate free expression, user privacy, and equal protection. This change will require a paradigm shift in everything from hiring to product design to policy drafting; at times, it will even require difficult pushback against government actors. Second, we propose designing a series of consistent controls for public figures who pose the largest threat of inciting online and offline harms. Finally, we outline the need to overhaul transparency mechanisms to allow the public to better understand the groups and communities targeted by terrorism and violent extremism removals, and the role that governments may be playing in content moderation.

Recommendation 1: Shift Policy to Equitably Protect Marginalized Communities

Content moderation policy must be revamped to account for power dynamics among different groups and structural inequalities that may diminish opportunity for equal access to platforms. Currently, the amorphous nature of social media platforms' content moderation policies gives companies enormous discretion in their enforcement. Platforms must begin exercising this discretion in a way that truly balances free expression with equity and prioritizes user safety and due process.

This approach will require more limited enforcement in some circumstances and more expansive enforcement in others. When it comes to terrorist content, policy and enforcement lean too heavily toward over-removal, censoring everything from important political debate to timely and relevant journalistic content. This tendency can severely limit the ability of marginalized groups to speak out and engage with the political developments of the day, effectively creating a second-tier social media platform where anything that does not clearly denounce terrorism could be misinterpreted as praise or support,

leaving it subject to removal with limited rights of appeal. On the other hand, hate speech policies attempt to take a more measured approach with narrower restrictions, which can create a convoluted order of operations that ends up protecting powerful groups or allowing all but the most explicit attacks based on protected characteristics. Platforms should recalibrate this system to ensure that all users are treated equally. The first step is to collect and publish data on who bears the brunt of platform enforcement policies and priorities.

Moving forward, platforms must acknowledge and document the unique ways in which minority communities are most susceptible to harassment, violence, and hate speech and the ways in which such content can result in both offline and online harms. Too often, platforms rely on the risk of offline harm, such as "real-world violence" or doxxing, as a critical tipping point for identifying content that violates their policies. In fact, online harms ranging from reputational attacks to someone's voice being silenced are significant injuries that merit action. To date, it has fallen on affected communities to raise these issues, usually with limited (if any) response from the companies. An appropriate response will not always require removals but rather a more robust build-out into intermediate enforcement practices as well as tools for users to protect themselves.

Meaningful consideration of marginalized communities also requires analysis of the harms of various methods for enforcing content policy. Automated tools for content removals, whether they rely on hashing or natural language processing, must be evaluated for their ability to accurately assess different dialects, slang, and related variations of context.[238] Additionally, platforms must ensure that human moderation teams have the necessary linguistic and cultural competence to achieve effective content moderation, and that those employees receive adequate compensation and health care to accomplish their challenging but essential work. These basic first steps should be carried out *before* tools and moderation practices are employed, and they must be continually reevaluated to protect against disparate impact.

Recommendation 2: Invert the Status Quo — Moderation of Public Figures

Social media platforms' rules for public figures and newsworthiness need an overhaul. Accounts with the largest reach merit close scrutiny because they have the most potential to cause harm. As explained by Susan Benesch, director of the Dangerous Speech Project, this subset of users can range from elected officials to religious leaders to celebrities and political pundits; identifying these individuals requires analyzing the speaker, audience, context, and medium.[239] In the Trump case, Facebook's Oversight Board noted that influential users such as heads of state and high-ranking government officials "can carry a heightened risk of encouraging, legitimizing, or inciting violence — either because their high position of trust imbues their words with greater force and credibility or because their followers may infer they can act with impunity."[240] Instead of addressing this issue, companies largely take a hands-off approach, looking the other way when influential users regularly break rules. Facebook claims that it is fully implementing the Oversight Board's recommendation to "prioritize safety over expression when taking action on a threat from influential accounts," but it also claims to already take this approach — an assertion that seems contrary to the company's long-standing policy (which was only changed after the decision in the Trump case) that politicians' speech is inherently newsworthy.

The application of public figure rules is mostly invisible, with platforms usually disclosing it only when pressured by journalists. In June 2021, Facebook committed to greater transparency over how it moderates public figures, particularly during times of civil unrest. The company also said that it would begin providing regular updates about when it applies newsworthiness exceptions. How these commitments will play out in practice remains to be seen. Similarly, whereas Twitter has committed to disclosing when it applies its public interest exception, it also said that in 2018, it applied that exception fewer than five times.[241]

Hearing what public figures have to say is an important public interest, and robust freedom of expression requires space for offensive and hurtful views. At the same time, as the Facebook Oversight Board noted in May 2021, international human rights standards also expect state actors to condemn violence and provide accurate information on matters in the public interest.[242] At a minimum, social media companies must ensure that public figures do not leverage platforms to incite violence or other types of offline harm. Protecting the ability of ordinary people to use their speech to challenge the powerful requires a different calculus than the protections necessary for influential figures who often have multiple avenues for disseminating their message.

Platforms must write public figure policies that recognize the connection between reach, authority, and influence. They must also enforce their policies in a clear, consistent, and transparent manner. The following recommendations represent a baseline:

- Platforms should establish pilot programs to identify and more rapidly moderate posts from users who have a high probability of causing imminent harm. As part of this process, platforms will need to undertake an inventory of the accounts that necessitate escalated moder-

ation. This group may include elected officials but also may include influential people such as religious leaders, celebrities, and political pundits. The process for identifying these accounts will vary by platform. In a 2019 report titled *Alternative Influence: Broadcasting the Reactionary Right on YouTube*, researcher Rebecca Lewis suggests junctures at which YouTube might assess: the instances where it awards "silver," "gold," and "diamond" awards for reaching 100,000, 1 million, or 10 million subscribers.[243] Similarly, in the context of election integrity, the nonprofit Accountable Tech recommends focusing on accounts with more than 250,000 followers and applying strikes against those accounts for violations of applicable election integrity policies.[244]

- Platforms should fund and train content moderation staff who have appropriate linguistic, cultural, and political fluency and authority to moderate these influential accounts. According to Facebook, it already ensures that its content reviewers have "regional and linguistic expertise." Nonetheless, instances such as the platform's confusing the Nigerian #EndSARS campaign ("SARS" in that instance denoting a tactical unit of the Nigerian police force accused of brutality against civilians) with health misinformation clearly demonstrate the need for improvements in the company's screening and training processes.[245] Furthermore, as the Facebook Oversight Board noted in May 2021, moderators must be insulated from political and economic interference and undue influence.[246] This safeguard may necessitate decision-making by individuals other than the heads of social media platforms. When moderating posts from influential accounts, messages must be assessed by how people are likely to understand them, regardless of superficial attempts to couch them in language that skirts responsibility.[247]

- Platforms should publish policies for moderating public figures as part of their community guidelines. These descriptions should explain how they determine who is covered by an exemption, how the exemption applies to the remaining community standards, and whether there are different restrictions for how ordinary users can interact with public figures.

- Platforms should explain and disclose intermediate controls for public figures, specify when and how downranking and warning labels are applied, and explicate the strike policies that are applied to public figures who regularly break platform rules.

- Platforms should specify the protocols for moderating public figures during volatile situations such as mass protests, elections, or other times where the risk of offline harms is high. Facebook has indicated that it assembles a special team to escalate more rapidly during high-risk events; neither YouTube nor Twitter have indicated whether they follow a similar practice.[248]

- Whenever they apply a public interest exception, platforms should specify the rule that the post would normally break and clearly indicate that the post is not being removed because it is protected by the exemption. When an individual is suspended or banned, users visiting that public figure's profile should be notified that the person was suspended for violating a specific rule — and told what that rule was.

- Whenever platforms take steps to ban a public figure's account, they should take steps to appropriately archive the account. The steps taken to collect, preserve, and share information for the investigation and prosecution of human rights or other violations must be public and transparent, including how content that was removed can be made available to researchers in compliance with applicable laws.

- Platforms should disclose public figure censures in transparency reports by region and country, detailing the measures taken (e.g., removal, warning labels, or downranking).

Focusing on the biggest drivers of harm provides an effective way to allocate resources. It also protects the free expression and user safety of a majority of people, but especially marginalized communities that are the more frequently targeted by the powerful.

Recommendation 3: Ensure Consistent and Transparent Actions Regarding Terrorist and Violent Extremism Content

Removals of terrorist content are plagued by opacity and overbreadth. As a preliminary matter, platforms should publish a public list of the individuals and organizations covered by their terrorism, dangerous organizations and individuals, and related policies. To the extent that they simply rely on sanctions lists from the United States or United Nations, that should be disclosed. These disclosures will help assess whether policy rules such as those addressing white supremacy are written in a manner that does not miss the organizations that are driving violence, and whether these policies remain predominantly focused on ISIS and al-Qaeda.

Policies that broadly target content based on "praise," "support," or "glorification" should not be used, regardless of the type of violent extremism being targeted. These

imprecise terms will inevitably capture expressions of general sympathy for or understanding of certain viewpoints, not to mention news reporting. Relying on vague labels makes it more likely that content will be misinterpreted or inaccurately flagged by context-blind hashing algorithms. These terms also introduce opportunities for policy misuse, as praise or glorification provide catchall categorizations that become the easiest way to justify removal.[249] Facebook's decision to allow praise and support of certain conspiracy networks and hate-banned entities reflects a decision to avoid overburdening speech from users with more powerful political support; the same calculations should be extended to users from marginalized communities.

Finally, government involvement in terrorist or violent extremism content removals raises serious free expression and disparate impact concerns. In many jurisdictions, overbroad laws around terrorism and national security are already used to target dissent. For example, in 2017, the government of India pressured Twitter to block the accounts of activists, journalists, and academics critical of the government's military actions in Kashmir, relying on an Indian law prohibiting incitement that threatens national security.[250] This pattern is playing out again in 2021, as the Indian government pressured social media companies to block accounts connected to protests over new agricultural reforms as well as posts criticizing the government's response to the second wave of Covid-19, at times even threatening to jail the companies' employees for not complying with removal requests.[251]

In other countries, the embedding of law enforcement officials within platforms — commonly referred to as internet referral units — raises serious concerns that governments are directing platforms to target disfavored groups and seeking to remove content that they would be legally prohibited from removing themselves. In 2017, Facebook reached an agreement with the Israeli government to address "incitement," prompting the platform to remove content from Palestinian journalists and civil society.[252] This long-standing practice is reaching new levels: in May 2021, Justice Minister Benny Gantz responded to violence between Israelis and Palestinians by attempting to pressure platform executives to remove Palestinian content "that incites to violence or spreads disinformation."[253] The Israeli attorney general's office reported that its cyber unit had submitted 1,010 requests over a 10-day period for social media companies to remove content or reduce its exposure; the office claimed that Twitter complied with 82 percent of its requests and Facebook with 46 percent, and that it was awaiting removals from YouTube.[254]

In order to empower a global response to protect these marginalized groups from government censorship, transparency reports should disclose how often government agencies are flagging content for removal, specify the rule, and identify the agency. To the furthest extent that is legally permissible, platforms should also notify the affected user when content is removed due to government flagging. Additionally, when removals are accompanied by voluntary information-sharing arrangements with government agencies, these relationships must be publicly disclosed. Removals around "coordinated inauthentic behavior" (i.e., platform manipulation removals typically related to online influence operations) most clearly implicate such arrangements, but there is an ongoing risk that this private-public partnership is occurring outside of public view.[255]

Conclusion

We are at a crucial juncture in reevaluating the role that social media plays in world discourse. For too long, content moderation policies and practices have often failed to appropriately mitigate the numerous harms facing women, communities of color, LGBTQ+ communities, and ethnic and religious minorities. The time has come for a systematic overhaul of how platforms carry out their policies to address these harms and ensure that they actually provide products that facilitate free expression. At the same time, legislators must step in and end the information vacuum that prevents public interest evaluations of platform systems and their impact on society. These dual reforms are vital for ensuring that modern threats to the civil rights and liberties of marginalized groups are no longer treated as acceptable trade-offs.

Endnotes

1 Chinmayi Arun, "Facebook's Faces," *Harvard Law Review* 135 (forthcoming), https://ssrn.com/abstract=3805210.

2 See, e.g., Newley Purnell and Jeff Horwitz, "Facebook's Hate-Speech Rules Collide with Indian Politics," *Wall Street Journal*, August 14, 2020, https://www.wsj.com/articles/facebook-hate-speech-india-politics-muslim-hindu-modi-zuckerberg-11597423346.

3 Facebook's Community Guidelines also cover Instagram, as the Instagram Community Guidelines regularly link to and incorporate Facebook's rules regarding hate speech, bullying and harassment, violence and incitement, and dangerous organizations and individuals (among others). See Instagram Community Guidelines, accessed July 6, 2021, https://help.instagram.com/477434105621119?ref=igtos. This report does not address alternative content moderation models such as community moderation, which have had comparative success at a smaller scale on platforms like Reddit.

4 See, e.g., See Joseph Cox and Jason Koebler, "Why Won't Twitter Treat White Supremacy Like ISIS? Because It Would Mean Banning Some Republican Politicians Too," *Vice*, August 25, 2019, https://www.vice.com/en/article/a3xgq5/why-wont-twitter-treat-white-supremacy-like-isis-because-it-would-mean-banning-some-republican-politicians-too (Describing the statement of a technical employee at Twitter who works on machine learning and artificial intelligence (AI) issues noting at an all-hands meeting on March 22, 2019: "With every sort of content filter, there is a tradeoff, he explained. When a platform aggressively enforces against ISIS content, for instance, it can also flag innocent accounts as well, such as Arabic language broadcasters. Society, in general, accepts the benefit of banning ISIS for inconveniencing some others, he said. In separate discussions verified by Motherboard, that employee said Twitter has not taken the same aggressive approach to white supremacist content because the collateral accounts that are impacted can, in some instances, be Republican politicians.").

5 Compare Josh Halliday, "Twitter's Tony Wang: 'We Are the Free Speech Wing of the Free Speech Party,'" *Guardian*, March 22, 2012, https://amp.theguardian.com/media/2012/mar/22/twitter-tony-wang-free-speech (quoting Tony Wang, then-general manager of Twitter in the United Kingdom, as saying that "generally, we remain neutral as to the content because our general council and CEO like to say that we are the free speech wing of the free speech party.") with Nitasha Tiku and Casey Newton, "Twitter CEO: 'We Suck at Dealing with Abuse,'" *Verge*, February 4, 2015, https://www.theverge.com/2015/2/4/7982099/twitter-ceo-sent-memo-taking-personal-responsibility-for-the (quoting an internal company memo written by Twitter chief executive Dick Costolo in the wake of a radio piece criticizing the company's approach to harassment: "We suck at dealing with abuse and trolls on the platform and we've sucked at it for years. . . . It's no secret and the rest of the world talks about it every day. We lose core user after core user by not addressing simple trolling issues that they face every day."). See also Kate Klonick, "The New Governors: The People, Rules, and Processes Governing Online Speech," *Harvard Law Review,* 131 (2018): 1631 https://harvardlawreview.org/2018/04/the-new-governors-the-people-rules-and-processes-governing-online-speech (describing the "ethos of the pre-2008 moderation guidelines" for Facebook as "if it makes you feel bad in your gut, then go ahead and take it down" and noting that "it was not until November 2009, five years after the site was founded, that Facebook created a team of about twelve people to specialize in content moderation." Klonick also describes a similar situation at YouTube in the early days.).

6 See, e.g., Sara Fischer and Ashley Gold, "All the Platforms That Have Banned or Restricted Trump So Far," *Axios*, January 11, 2021, https://www.axios.com/platforms-social-media-ban-restrict-trump-d9e44f3c-8366-4ba9-a8a1-7f3114f920f1.html.

7 Twitter Support, "The Twitter Rules," Terms of Service and Rules Policies, January 14, 2009, https://web.archive.org/web/20090118211301/http:/twitter.zendesk.com/forums/26257/entries/18311.

8 YouTube, "YouTube Community Guidelines," October 24, 2006, https://web.archive.org/web/20061024061946/http://www.youtube.com/t/community_guidelines.

9 It was this pressure that led companies like Facebook to finally publish details of their internal content moderation policies a few years ago. See, e.g., Monika Bickert, "Publishing Our Internal Enforcement Guidelines and Expanding Our Appeals Process," Facebook Newsroom, April 24, 2018, https://about.fb.com/news/2018/04/comprehensive-community-standards.

10 See, e.g., Facebook Newsroom, "Enforcing Our Community Standards," August 6, 2018, https://about.fb.com/news/2018/08/enforcing-our-community-standards ("We believe in giving people a voice, but we also want everyone using Facebook to feel safe. It's why we have Community Standards and remove anything that violates them, including hate speech that attacks or dehumanizes others. Earlier today, we removed four Pages belonging to Alex Jones for repeatedly posting content over the past several days that breaks those Community Standards."); Twitter Safety (@TwitterSafety), "Updating Our Rules Against Hateful Conduct," Twitter Blog, December 2, 2020, https://blog.twitter.com/en_us/topics/company/2019/hatefulconductupdate.html; Facebook Newsroom, "Mark Zuckerberg Stands for Voice and Free Expression," October 17, 2019, https://about.fb.com/news/2019/10/mark-zuckerberg-stands-for-voice-and-free-expression; Matt Halprin, "An Update to Our Harassment Policy," YouTube Official Blog, December 11, 2019, https://blog.youtube/news-and-events/an-update-to-our-harassment-policy; and Kate Klonick, "Facebook v. Sullivan," Knight First Amendment Institute at Columbia University, October 1, 2018, https://knightcolumbia.org/content/facebook-v-sullivan.

11 See, e.g., Alex Hern, "Facebook and YouTube Defend Response to Christchurch Videos," *Guardian*, March 19, 2019, https://www.theguardian.com/world/2019/mar/19/facebook-and-youtube-defend-response-to-christchurch-videos.

12 Facebook Oversight Board, Case Decision 2020-006-FB-FBR, January 28, 2021, https://www.oversightboard.com/decision/FB-XWJQBU9A ("The Board also found Facebook's misinformation and imminent harm rule, which this post is said to have violated, to be inappropriately vague and inconsistent with international human rights standards. A patchwork of policies found on different parts of Facebook's website make it difficult for users to understand what content is prohibited. Changes to Facebook's COVID-19 policies announced in the company's Newsroom have not always been reflected in its Community Standards, while some of these changes even appear to contradict them.").

13 See European Commission, "European Commission and IT Companies Announce Code of Conduct on Illegal Online Hate Speech," press release, May 31, 2016, https://ec.europa.eu/commission/presscorner/detail/en/IP_16_1937. See also Amar Toor, "France Wants Facebook and Twitter to Launch an 'Offensive' Against ISIS Propaganda," *Verge*, December 3, 2015, https://www.theverge.com/2015/12/3/9842258/paris-attacks-facebook-twitter-google-isis-propaganda.

14 See Abdul Rahman Al Jaloud et al., *Caught in the Net: The Impact of 'Extremist' Speech Regulations on Human Rights Content*, Electronic Frontier Foundation, May 30, 2019, 3, https://www.eff.org/files/2019/05/30/caught_in_the_net_whitepaper_2019.pdf ("Social media companies have long struggled with what to do about extremist content on their platforms. While most companies include provisions about 'extremist' content in their community standards, until recently, such content was often vaguely defined, providing policymakers and content moderators a wide berth in determining

what to remove, and what to allow. Unfortunately, companies have responded with overbroad and vague policies and practices that have led to mistakes at scale that are decimating human rights content.").

15 See, e.g., European Parliament Think Tank, "Addressing the Dissemination of Terrorist Content Online," April 9, 2021, https://www.europarl.europa.eu/thinktank/en/document.html?reference=EPRS_BRI(2020)649326.

16 See Twitter Help Center, "Violent Organizations Policy," General Guidelines and Policies, October 2020, https://help.twitter.com/en/rules-and-policies/violent-groups; Facebook Community Standards, "Dangerous Individuals and Organizations Policy," accessed June 22, 2021, https://www.facebook.com/communitystandards/dangerous_individuals_organizations; and YouTube Help, "Violent Criminal Organizations Policy," YouTube Policies, accessed June 22, 2021, https://support.google.com/youtube/answer/9229472?hl=en&ref_topic=9282436.

17 Facebook removes terrorist content under its policy on dangerous individuals and organizations, which prohibits "any organizations or individuals that proclaim a violent mission or are engaged in violence to have a presence on Facebook." Facebook Community Standards, "Dangerous Individuals and Organizations Policy." This policy covers a wide range of content, including terrorist activity, organized hate, mass murder, human trafficking, organized violence, and "militarized social movements." Twitter's violent organizations policy applies to terrorist organizations and violent extremist groups — generally defined as groups that participate in or promote violence against civilians. Twitter Help Center, "Violent Organizations Policy" ("You may not threaten or promote terrorism or violent extremism. Violent extremist groups are those that meet all of the below criteria: identify through their stated purpose, publications, or actions as an extremist group; have engaged in, or currently engage in, violence and/or the promotion of violence as a means to further their cause; and target civilians in their acts and/or promotion of violence. . . . Other violent organizations are those that meet all of the below criteria: a collection of individuals with a shared purpose; and have systematically targeted civilians with violence."). YouTube does not even attempt to define terrorist organizations, although it removes terrorist content under its violent criminal organizations policy. YouTube Help, "Violent Criminal Organizations Policy" ("Don't post content on YouTube if it fits any of the descriptions noted below. · Content produced by violent criminal or terrorist organizations · Content praising or memorializing prominent terrorist or criminal figures in order to encourage others to carry out acts of violence · Content praising or justifying violent acts carried out by violent criminal or terrorist organizations · Content aimed at recruiting new members to violent criminal or terrorist organizations · Content depicting hostages or posted with the intent to solicit, threaten, or intimidate on behalf of a violent criminal or terrorist organization · Content that depicts the insignia, logos, or symbols of violent criminal or terrorist organizations in order to praise or promote them.").

18 Monika Bickert and Erin Saltman, "An Update on Our Efforts to Combat Terrorism Online," Facebook Newsroom, December 20, 2019, https://about.fb.com/news/2019/12/counterterrorism-efforts-update; Facebook Newsroom, "Next Steps for the Global Internet Forum to Counter Terrorism," September 23, 2019, https://about.fb.com/news/2019/09/next-steps-for-gifct; Anna Meier, "Why Do Facebook and Twitter's Anti-extremist Guidelines Allow Right-Wingers More Freedom than Islamists?" *Washington Post*, August 1, 2019, https://www.washingtonpost.com/politics/2019/08/01/why-do-facebook-twitters-anti-extremist-guidelines-allow-right-wingers-more-freedom-than-islamists; Organisation for Economic Co-operation and Development, *Current Approaches to Terrorist and Violent Extremist Content Among the Global Top 50 Online Content-Sharing Services*, Directorate for Science, Technology, and Innovation Committee on Digital Economic Policy, August 14, 2020, www.oecd.org/officialdocuments/publicdisplaydocumentpdf/?cote=DSTI/CDEP(2019)15/FINAL&docLanguage=En; and Twitter Help Center, "Violent Organizations Policy."

19 In a case appealed to the Facebook Oversight Board, the panel overturned a decision to remove a post that Facebook claimed violated its dangerous individuals and organizations policy, finding that the rules are not clearly explained to users. Among its policy recommendations, the panel called on Facebook to "provide a public list of the organizations and individuals designated as 'dangerous' under the Dangerous Individuals and Organizations Community Standard or, at the very least, a list of examples." Facebook Oversight Board, "Oversight Board Overturns Facebook Decision: Case 2020-005-FB-UA," January 2021, https://oversightboard.com/news/141077647749726-oversight-board-overturns-facebook-decision-case-2020-005-fb-ua. Facebook has responded that it is committed to increasing transparency around its definitions of "praise," "support," and "representation" by adding definitions within the next few months. Facebook, "Facebook's Response to the Oversight Board's First Decisions," February 2021, https://about.fb.com/wp-content/uploads/2021/02/OB_First-Decision_Detailed_.pdf.

20 Twitter Help Center, "Violent Organizations Policy"; YouTube Help, "Violent Criminal Organizations Policy"; and Facebook Community Standards, "Dangerous Individuals and Organizations Policy."

21 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 ("Over the last two months Facebook has deleted at least 35 accounts of Syrian journalists and activists, according to the Syrian Archive, a database of documentary evidence of human rights violations and other crimes committed by all sides of the conflict in Syria sourced mostly from social media.").

22 See Fionnuala Ní Aoláin, "Mandate of the Special Rapporteur on the Promotion and Protection of Human Rights and Fundamental Freedoms While Countering Terrorism," letter to Mark Zuckerberg, July 24, 2018, 6, https://www.ohchr.org/Documents/Issues/Terrorism/OL_OTH_46_2018.pdf. While a June 2021 update to Facebook's dangerous organizations and individuals policy provided additional information on what the company considers praise (e.g., "speaking positively about a designated entity or event" or giving it "a sense of achievement") and support (e.g., an "act which provides material aid to a designated entity or event" or puts out "a call to action" on behalf of it), the threat of overbroad application remains. See Facebook recent updates to Community Standards, available https://www.facebook.com/communitystandards/recentupdates/dangerous_individuals_organizations/.

23 See Facebook recent updates to Community Standards, available https://www.facebook.com/communitystandards/recentupdates/dangerous_individuals_organizations/.

24 Facebook Transparency Center, "Dangerous Organizations: Terrorism and Organized Hate," *Community Standards Enforcement Report*, February 2021, https://transparency.facebook.com/community-standards-enforcement#dangerous-organizations; Google, "YouTube Community Guidelines Enforcement," *Google Transparency Report*, accessed June 22, 2021, https://transparencyreport.google.com/youtube-policy/removals?hl=en; and Twitter, "Rules Enforcement," *Twitter Transparency Report*, January 11, 2021, https://transparency.twitter.com/en/reports/rules-enforcement.html#2020-jan-jun.

25 Global Internet Forum to Counter Terrorism, *GIFCT Transparency Report July 2020*, July 2020, https://gifct.org/wp-content/uploads/2020/10/GIFCT-Transparency-Report-July-2020-Final.pdf.

26 Global Internet Forum to Counter Terrorism, *GIFCT Transparency Report*; and Ángel Díaz, "Global Internet Forum to Counter Terrorism Transparency Report Raises More Questions than Answers," Brennan Center for Justice, September 25, 2019, https://www.brennancenter.org/our-work/analysis-opinion/global-internet-forum-counter-terrorism-transparency-report-raises-more.

27 See Nicholas J. Rasmussen, "GIFCT Launches Multi-Stake-

holder Effort to Develop an Expanded Taxonomy Framework for the Hash-Sharing Database," letter from GIFCT Executive Director Nicholas J. Rasmussen to the GIFCT stakeholder community, February 24, 2021, https://gifct.org/2021/02/24/gifct-launches-taxonomy-framework-rfp.

28 In a 2017 blog post describing the platform's capabilities, Facebook explained that its automated removals, ranging from image matching to natural language processing, were focused on targeting "ISIS, al-Qaeda, and their affiliates." Monika Bickert and Brian Fishman, "Hard Questions: How We Counter Terrorism," Facebook Newsroom, June 15, 2017, https://about.fb.com/news/2017/06/how-we-counter-terrorism. In the second quarter of 2018, Facebook removed 9.4 million pieces of terrorist content, all of which were related to ISIS, al-Qaeda, and their affiliates. Monika Bickert and Brian Fishman, "Hard Questions: How Effective Is Technology in Keeping Terrorists off Facebook?" Facebook Newsroom, April 23, 2018, https://about.fb.com/news/2018/04/keeping-terrorists-off-facebook.

29 Al Jaloud et al., *Caught in the Net*, 6–7.

30 Facebook Transparency Center, "Dangerous Organizations."

31 Facebook Newsroom, "Standing Against Hate," March 27, 2019, https://about.fb.com/news/2019/03/standing-against-hate.

32 Compare Facebook Newsroom, "Standing Against Hate" with Monika Bickert and Brian Fishman, "Hard Questions: What Are We Doing to Stay Ahead of Terrorists?" Facebook Newsroom, November 8, 2018, https://about.fb.com/news/2018/11/staying-ahead-of-terrorists.

33 Facebook Newsroom, "Standing Against Hate."

34 Facebook Newsroom, "Standing Against Hate."

35 Compare Facebook Newsroom, "Combating Hate and Extremism," September 17, 2019, https://about.fb.com/news/2019/09/combating-hate-and-extremism ("We've banned more than 200 white supremacist organizations from our platform") with Adam Mosseri, "An Update on Our Equity Work," Facebook Newsroom, September 9, 2020, https://about.fb.com/news/2020/09/an-update-on-our-equity-work ("This includes removing 23 different banned organizations, over half of which supported white supremacy.").

36 Laura W. Murphy et al., *Facebook's Civil Rights Audit — Final Report*, July 8, 2020, 50, https://about.fb.com/wp-content/uploads/2020/07/Civil-Rights-Audit-Final-Report.pdf.

37 See Facebook Community Standards Update, https://www.facebook.com/communitystandards/recentupdates/dangerous_individuals_organizations/.

38 See Ryan Mac and Craig Silverman, "'Mark Changed the Rules': How Facebook Went Easy on Alex Jones and Other Right-Wing Figures," *BuzzFeed News*, last updated February 22, 2021, https://www.buzzfeednews.com/article/ryanmac/mark-zuckerberg-joel-kaplan-facebook-alex-jones.

39 Mac and Silverman, "'Mark Changed the Rules.'"

40 Facebook Newsroom, "Enforcing Our Community Standards."

41 See Facebook Community Standards Update.

42 Arcadiy Kantor, "Measuring Our Progress Combating Hate Speech," Facebook Newsroom, November 19, 2020, https://about.fb.com/news/2020/11/measuring-progress-combating-hate-speech; and Twitter Safety (@TwitterSafety), "Updating Our Rules Against Hateful Conduct."

43 Timothy McLaughlin, "How Facebook's Rise Fueled Chaos and Confusion in Myanmar," *Wired*, July 6, 2018, https://www.wired.com/story/how-facebooks-rise-fueled-chaos-and-confusion-in-myanmar (discussing Facebook's failures in Myanmar, in part due to a failure to understand local context and language).

44 This is a common way to define hate speech. See United Nations, "United Nations Strategy and Plan of Action on Hate Speech," May 2019, https://www.un.org/en/genocideprevention/documents/UN%20Strategy%20and%20Plan%20of%20Action%20on%20Hate%20Speech%2018%20June%20SYNOPSIS.pdf.

45 Within the first tier is violent or dehumanizing speech that targets a person based on a protected characteristic or their immigration status. The second tier prohibits statements of inferiority targeting a person based on a protected characteristic. The third tier includes calls for segregation, exclusion, or slurs targeting a person based on a protected characteristic. Facebook also protects against attacks on the basis of age, but only when age is paired with another protected characteristic. Facebook Community Standards, "Hate Speech," accessed June 22, 2021, https://www.facebook.com/communitystandards/hate_speech.

46 The Twitter rules now state: "You may not promote violence against, threaten, or harass other people on the basis of race, ethnicity, national origin, caste, sexual orientation, gender, gender identity, religious affiliation, age, disability, or serious disease." Twitter Help Center, "The Twitter Rules," Twitter Rules and Policies, accessed June 22, 2021, https://help.twitter.com/en/rules-and-policies/twitter-rules. YouTube's Community Guidelines prohibit content that either alleges that a group is superior to justify discrimination or promotes violence or hatred based on someone's protected attributes — age, caste, disability, ethnicity, gender identity and expression, nationality, race, immigration status, religion, sex/gender sexual orientation, victims of a major violent event and their kin, and veteran status. YouTube Help, "Hate Speech Policy," YouTube Policies, accessed June 22, 2021, https://support.google.com/youtube/answer/2801939?hl=en&ref_topic=9282436.

47 Twitter Help Center, "Hateful Conduct Policy," Twitter Rules and Policies, accessed June 22, 2021, https://help.twitter.com/en/rules-and-policies/hateful-conduct-policy.

48 YouTube Team, "Our Ongoing Work to Tackle Hate," YouTube Official Blog, June 5, 2019, https://blog.youtube/news-and-events/our-ongoing-work-to-tackle-hate.

49 Twitter is the only platform to explicitly justify its decision to single out particular characteristics as deserving of greater protection on the basis of research showing that certain groups are disproportionately subject to harmful speech online, and to acknowledge that for those who identity with multiple underrepresented groups, abuse may be more common and more severe and cause greater harm. Twitter Help Center, "Twitter Rules."

50 Rachelle Hampton, "The Black Feminists Who Saw the Alt-Right Threat Coming," *Slate*, April 23, 2019, https://slate.com/technology/2019/04/black-feminists-alt-right-twitter-gamergate.html.

51 See Hampton, "Black Feminists."

52 See Hampton, "Black Feminists" ("Within days of the creation of #YourSlipIsShowing, Crockett, Hudson, and others had documented a small army of fake accounts numbering in the hundreds — accounts that users could not only cross-reference with their followers but also mass-report to Twitter. But despite the evidence that harassment campaigns fueled by a noxious mixture of misogyny and racism spelled out a threat to users from vulnerable groups, Hudson and Crockett felt that Twitter basically did nothing. At most, the company suspended a few of the mass-reported accounts tagged under #YourSlipIsShowing.").

53 See Craig Timberg and Elizabeth Dwoskin, "As QAnon Grew, Facebook and Twitter Missed Years of Warning Signs About the Conspiracy Theory's Violent Nature," *Washington Post*, October 3, 2020, https://www.washingtonpost.com/technology/2020/10/01/facebook-qanon-conspiracies-trump.

54 Timberg and Dwoskin, "As QAnon Grew."

55 Ben Collins and Brandy Zadrozny, "Twitter Bans 7,000 QAnon Accounts, Limits 150,000 Others as Part of Broad Crackdown," NBC News, July 21, 2020, https://www.nbcnews.com/tech/tech-news/twitter-bans-7-000-qanon-accounts-limits-150-000-others-n1234541. See also Twitter Help Center, "Coordinated Harmful

Activity," General Guidelines and Policies, January 2021, https://help.twitter.com/en/rules-and-policies/coordinated-harmful-activity.

56 Twitter Help Center, "Coordinated Harmful Activity."

57 See Twitter Safety (@TwitterSafety), "We've been clear that we will take strong enforcement action on behavior that has the potential to lead to offline harm. In line with this approach, this week we are taking further action on so-called 'QAnon' activity across the service," Twitter, July 21, 2020, 5:00 p.m., https://twitter.com/TwitterSafety/status/1285726277719199746.

58 Twitter Safety (@TwitterSafety), "An Update Following the Riots in Washington, DC," Twitter Blog, January 12, 2021, https://blog.twitter.com/en_us/topics/company/2021/protecting--the-conversation-following-the-riots-in-washington--.html.

59 Timberg and Dwoskin, "As QAnon Grew."

60 See, e.g., Manu Raju and Sam Fossum, "Trump Praised QAnon During Meeting About Keeping the Senate," CNN, December 3, 2020, https://www.cnn.com/2020/12/03/politics/donald-trump-qanon/index.html; and Alex Kaplan, "Here Are the QAnon Supporters Running for Congress in 2020," Media Matters for America, January 7, 2020, https://www.mediamatters.org/qanon-conspiracy-theory/here-are-qanon-supporters-running-congress-2020.

61 Julia Angwin and Hannes Grassegger, "Facebook's Secret Censorship Rules Protect White Men from Hate Speech but Not Black Children," ProPublica, June 28, 2017, https://www.propublica.org/article/facebook-hate-speech-censorship-internal-documents-algorithms.

62 Angwin and Grassegger, "Facebook's Secret Censorship Rules."

63 Adam Smith, "Facebook Comments Like 'White Men Are Stupid' Were Algorithmically Rated as Bad as Antisemitic or Racist Slurs, According to Internal Documents," *Independent*, December 4, 2020, https://www.independent.co.uk/life-style/gadgets-and-tech/facebook-comments-algorithm-racism-b1766209.html.

64 Emily A. Vogels, "The State of Online Harassment," Pew Research Center, January 13, 2021, https://www.pewresearch.org/internet/2021/01/13/the-state-of-online-harassment.

65 Amnesty International's investigation found that 55 percent of respondents were less able to focus on everyday tasks, 54 percent experienced panic attacks, anxiety, or stress, and 57 percent reported feeling apprehension when thinking about using the internet or social media. See Amnesty International, "Twitter Still Failing Women over Online Violence and Abuse," September 22, 2020, https://www.amnesty.org/en/latest/news/2020/09/twitter-failing-women-over-online-violence-and-abuse.

66 See Jillian C. York, *Silicon Values: The Future of Free Speech Under Surveillance Capitalism* (New York: Verso, 2021), 196.

67 Viktorya Vilk, Elodie Vialle, and Matt Bailey, *No Excuse for Abuse: What Social Media Companies Can Do Now to Combat Online Harassment and Empower Users*, PEN America, March 2021, https://pen.org/report/no-excuse-for-abuse.

68 See, e.g., David Brody and Sean Bickford, *Discriminatory Denial of Service: Applying State Public Accommodations Laws to Online Commerce*, Lawyers' Committee for Civil Rights Under Law, January 2020, 6, https://lawyerscommittee.org/wp-content/uploads/2019/12/Online-Public-Accommodations-Report.pdf ("Online threats, harassment, and intimidation . . . interfere with users' right to equal enjoyment of online services, chill speech and civic engagement, and cause serious harm. When a user self-censors or quits an online platform after experiencing hateful harassment, that user is deprived of their equal right to enjoy the services offered by that business.").

69 See Facebook Community Standards, "Bullying and Harassment," accessed June 22, 2021, https://www.facebook.com/communitystandards/bullying.

70 See YouTube Help, "Harassment and Cyberbullying Policies," YouTube Policies, accessed June 22, 2021, https://support.google.com/youtube/answer/2802268.

71 See Twitter Help Center, "Abusive Behavior," General Guidelines and Policies, accessed June 22, 2021, https://help.twitter.com/en/rules-and-policies/abusive-behavior. See also Twitter Help Center, "About Public-Interest Exceptions on Twitter," General Guidelines and Policies, accessed June 22, 2021, https://help.twitter.com/en/rules-and-policies/public-interest.

72 Twitter Help Center, "About Public-Interest Exceptions."

73 Facebook, "Case on a Comment Related to the January 2021 Protests in Russia," Facebook Transparency Center, June 25, 2021, https://transparency.fb.com/oversight/oversight-board-cases/comment-related-to-january-2021-protests-in-russia/.

74 For example, Facebook prohibits repeatedly contacting someone in a manner that is "unwanted," "sexually harassing," or "directed at a large number of individuals with no prior solicitation." YouTube prohibits repeatedly encouraging "abusive audience behavior," or "targets, insults and abuses an identifiable individual." See Facebook Community Standards, "Bullying and Harassment"; and YouTube Help, "Harassment and Cyberbullying Policies."

75 Twitter Help Center, "Abusive Behavior."

76 See, e.g. Facebook Community Standards, "Bullying and Harassment" ("In certain instances, we require self-reporting because it helps us understand that the person targeted feels bullied or harassed."); and Twitter Help Center, "Abusive Behavior" ("To help our teams understand the context of a conversation, we may need to hear directly from the person being targeted, to ensure that we have the information needed prior to taking any enforcement action.").

77 Facebook, "Case on a Comment Related to the January 2021 Protests in Russia," Facebook Transparency Center, June 25, 2021, https://transparency.fb.com/oversight/oversight-board-cases/comment-related-to-january-2021-protests-in-russia/.

78 See Julia Alexander, "YouTube Investigating Right-Wing Pundit Steven Crowder for Harassing Vox.com Host," *Verge*, May 31, 2019, https://www.theverge.com/2019/5/31/18647621/youtube-steven-crowder-bullying-harassment-twitter-vox-carlos-maza; and Rebecca Lewis, *Alternative Influence: Broadcasting the Reactionary Right on YouTube*, Data & Society, September 2019, 1, https://datasociety.net/wp-content/uploads/2018/09/DS_Alternative_Influence.pdf.

79 Mahita Gajanan, "YouTube Says Homophobic Harassment Targeting a Popular Host Doesn't Violate Its Policies," *Time*, June 5, 2019, https://time.com/5601302/youtube-vox-carlos-maza-steven-crowder-homophobia.

80 Benjamin Goggin, "YouTube's Week from Hell: How the Debate over Free Speech Online Exploded After a Conservative Star with Millions of Subscribers Was Accused of Homophobic Harassment," *Business Insider*, June 9, 2019, https://www.businessinsider.com/steven-crowder-youtube-speech-carlos-maza-explained-youtube-2019-6.

81 See Alexander, "YouTube Investigating Right-Wing Pundit."

82 See Alexander, "YouTube Investigating Right-Wing Pundit."

83 Nick Statt, "YouTube Decides That Homophobic Harassment Does Not Violate Its Policies," *Verge*, June 4, 2019, https://www.theverge.com/2019/6/4/18653088/youtube-steven-crowder-carlos-maza-harassment-bullying-enforcement-verdict.

84 Goggin, "YouTube's Week from Hell."

85 See TeamYouTube (@TeamYouTube), "Sorry for the confusion, we were responding to your tweets about the T-shirts. Again, this channel is demonetized due to continued egregious actions that have harmed the broader community. To be reinstated, he will need to address all of the issues with his channel," Twitter, June 5, 2019, 1:06 p.m., https://twitter.com/TeamYouTube/status/1136363701882064896.

86 See Julia Alexander, "YouTube Revokes Ads from Steven Crowder Until He Stops Linking to His Homophobic T-Shirts," *Verge*, June 5, 2019, https://www.theverge.com/2019/6/5/18654196/steven-crowder-demonetized-carlos-maza-youtube-homophobic-language-ads.

87 Julia Alexander, "YouTube Will Let Steven Crowder Run Ads After Year-Long Suspension for Harassment," *Verge*, August 12, 2020, https://www.theverge.com/2020/8/12/21365601/youtube-steven-crowder-monetization-reinstated-harassment-carlos-maza.

88 Halprin, "An Update to Our Harassment Policy."

89 The company also claimed that it was tightening its YouTube Partner Program to punish channels that "repeatedly brush up" against its harassment policy by suspending them from the program, thereby eliminating their ability to make money on YouTube. Halprin, "An Update to Our Harassment Policy."

90 See Jessica Guynn, "Facebook While Black: Users Call It Getting 'Zucked,' Say Talking About Racism Is Censored as Hate Speech," *USA Today*, April 24, 2019, https://www.usatoday.com/story/news/2019/04/24/facebook-while-black-zucked-users-say-they-get-blocked-racism-discussion/2859593002; Erin Golden, "Black Lives Matter Minneapolis Says Facebook Suspended Its Accounts over Critical Posts, *Star Tribune*, January 26, 2016, https://www.startribune.com/black-lives-matter-minneapolis-says-facebook-suspended-its-accounts-over-critical-posts/366588641; and Chao Xiong, "St. Paul Officer on Leave After Allegedly Telling Drivers to Run Over Marchers," *Star Tribune*, January 19, 2016, https://www.startribune.com/st-paul-police-officer-on-leave-after-allegedly-telling-drivers-to-run-over-black-lives-matter-marchers/365697691.

91 Ready for Revolution, "The Lullaby Aint Loud as It Used to Be," *Hood Communist*, March 29, 2021, https://hoodcommunist.org/2021/03/29/the-lullaby-aint-loud-as-it-used-to-be/?amp=1.

92 Samuel Gibbs, "Facebook Bans Women for Posting 'Men Are Scum' After Harassment Scandals," *Guardian*, December 5, 2017, https://www.theguardian.com/technology/2017/dec/05/facebook-bans-women-posting-men-are-scum-harassment-scandals-comedian-marcia-belsky-abuse.

93 See Jeremy Bowen, "Israel-Palestinians: Old Grievances Fuel New Fighting," BBC News, May 11, 2021, https://www.bbc.com/news/world-middle-east-57074460; Kassem Mnejja and Marwa Fatafta, "Sheikh Jarrah: Facebook and Twitter Systematically Silencing Protests, Deleting Evidence," Access Now, May 7, 2021, https://www.accessnow.org/sheikh-jarrah-facebook-and-twitter-systematically-silencing-protests-deleting-evidence; Access Now (@AccessNow), "Palestinian activists and citizens are primarily using social media to draw global attention. People have taken to social media to document and denounce Israel police brutality, violent attacks, occupation and apartheid, and forced dispossession from their homes," Twitter, May 17, 2021, 12:26 p.m., https://twitter.com/accessnow/status/1394373888109338627; and Matthew Ingram, "Social Networks Accused of Censoring Palestinian Content," *Columbia Journalism Review*, May 19, 2021, https://www.cjr.org/the_media_today/social-networks-accused-of-censoring-palestinian-content.php.

94 See Ryan Mac, "Instagram Censored Posts About One of Islam's Holiest Mosques, Drawing Employee Ire," *BuzzFeed News*, May 12, 2021, https://www.buzzfeednews.com/article/ryanmac/instagram-facebook-censored-al-aqsa-mosque.

95 Emanuel Maiberg and Joseph Cox, "Twitter Said It Restricted Palestinian Writer's Account by Accident," *Vice*, May 11, 2021, https://www.vice.com/en/article/qj8b4x/twitter-said-it-restricted-palestinian-writers-account-by-accident.

96 Trusted flaggers are recruited from civil society organizations and government agencies, and some are individual users. See, e.g., YouTube Help, "YouTube Trusted Flagger Program," Reporting and Enforcement, accessed June 22, 2021, https://support.google.com/youtube/answer/7554338?hl=en ("Individual users, government agencies, and NGOs are eligible for participation in the YouTube Trusted Flagger program. Ideal candidates have identified expertise in at least one policy vertical (listed here), flag content frequently with a high rate of accuracy, and are open to ongoing discussion and feedback with YouTube about various content areas."); and Google, "YouTube Community Guidelines Enforcement."

97 Casey Newton, "Facebook Open-Sources Algorithms for Detecting Child Exploitation and Terrorism Imagery," *Verge*, August 1, 2019, https://www.theverge.com/2019/8/1/20750752/facebook-child-exploitation-terrorism-open-source-algorithm-pdq-tmk; Google, "Featured Policies: Child Safety," *Google Transparency Report*, accessed June 22, 2021, https://transparencyreport.google.com/youtube-policy/featured-policies/child-safety?hl=enl ("Community Guidelines and enforcement details: How YouTube uses technology to detect violative content"); Twitter, "Rules Enforcement"; and Global Internet Forum to Counter Terrorism, "GIFCT Transparency Report July 2020," accessed July 6, 2021, https://gifct.org/wp-content/uploads/2020/10/GIFCT-Transparency-Report-July-2020-Final.pdf.

98 Guy Rosen, "An Update on Our Work to Keep People Informed and Limit Misinformation About COVID-19," Facebook Newsroom, April 16, 2020, https://about.fb.com/news/2020/04/covid-19-misinfo-update ("Once a piece of content is rated false by fact-checkers, we reduce its distribution and show warning labels with more context. Based on one fact-check, we're able to kick off similarity detection methods that identify duplicates of debunked stories."); and Twitter Inc., "Coronavirus: Staying Safe and Informed on Twitter — An Update on Our Proactive Enforcement and Spam Detection," Twitter Blog, May 4, 2020, https://blog.twitter.com/en_us/topics/company/2020/covid-19.html#proactiveenforcement ("Since introducing our updated policies on March 18, we have removed more than 4,074 Tweets containing misleading and potentially harmful content from Twitter. Additionally, our automated systems have challenged more than 3.4 million accounts which were targeting discussions around COVID-19 with spammy or manipulative behaviors. We will continue to use both technology and our teams to help us identify and stop spammy behavior and accounts.").

99 Ben Bradford et al., *Report of the Facebook Data Transparency Advisory Group*, Yale Law School Justice Collaboratory, April 2019, https://law.yale.edu/sites/default/files/area/center/justice/document/dtag_report_5.22.2019.pdf.

100 Bradford et al., *Report of the Facebook Data Transparency Advisory Group*.

101 Bradford et al., *Report of the Facebook Data Transparency Advisory Group*.

102 See Paresh Dave, "Social Media Giants Warn of AI Moderation Errors as Coronavirus Empties Offices," Reuters, March 16, 2020, https://www.reuters.com/article/us-health-coronavirus-google/social-media-giants-warn-of-ai-moderation-errors-as-coronavirus-empties-offices-idUSKBN2133BM.

103 See Evan Engstrom and Nick Feamster, *The Limits of Filtering: A Look at the Functionality and Shortcomings of Content Detection Tools,* Engine, March 2017, https://static1.squarespace.com/static/571681753c44d835a440c8b5/t/58d058712994ca536bbfa47a/1490049138881/FilteringPaperWebsite.pdf.

104 See Cox and Koebler, "Why Won't Twitter Treat White Supremacy Like ISIS?"

105 See Cox and Koebler, "Why Won't Twitter Treat White Supremacy Like ISIS?"

106 See Natasha Duarte, Emma Llansó, and Anna Loup, *Mixed Messages? The Limits of Automated Social Media Content Analysis*, Center for Democracy & Technology, November 2017, 9, https://cdt.org/wp-content/uploads/2017/11/Mixed-Messages-Paper.pdf. See also Carey Shenkman, Dhanaraj Thakur, and Emma Llansó, *Do You See What I See? Capabilities and Limits of Automated Multimedia Content Analysis*, Center for Democracy & Technology, May 20, 2021, https://cdt.org/wp-content/uploads/2021/05/2021-05-18-Do-You-See-What-I-See-Capabilities-Limits-of-Automated-Multimedia-Content-Analysis-Full-Report-2033-FINAL.pdf.

107 See Maarten Sap et al., "The Risk of Racial Bias in Hate Speech Detection," *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, Florence, Italy, July 28–August 2, 2019,

1671, https://www.aclweb.org/anthology/P19-1163.pdf.

108 Thomas Davidson et al., "Racial Bias in Hate Speech and Abusive Language Detection Datasets," *Proceedings of the Third Abusive Language Workshop at the Annual Meeting for the Association for Computational Linguistics*, Florence, Italy, August 1–2, 2019, 6, https://arxiv.org/pdf/1905.12516.pdf.

109 Davidson et al., "Racial Bias in Hate Speech," 6.

110 See Sap et al., "Risk of Racial Bias," 1672.

111 See Aliide Naylor, "Underpaid Workers Are Being Forced to Train Biased AI on Mechanical Turk," *Vice*, March 8, 2021, https://www.vice.com/en/article/88apnv/underpaid-workers-are-being-forced-to-train-biased-ai-on-mechanical-turk.

112 See Duarte et al., *Mixed Messages?*, 14.

113 See Duarte et al., *Mixed Messages?*, 15.

114 These moderators are frequently underpaid, under supported, and largely hidden from public view. See, e.g., Olivia Solon, "Underpaid and Overburdened: The Life of a Facebook Moderator," *Guardian*, May 25, 2017, https://www.theguardian.com/news/2017/may/25/facebook-moderator-underpaid-overburdened-extreme-content; Davey Alba, "Google Drops Firm Reviewing YouTube Videos," *Wired*, August 4, 2017, https://www.wired.com/story/google-drops-zerochaos-for-youtube-videos; and Adrian Chen, "The Laborers Who Keep Dick Picks and Beheadings Out of Your Facebook Feed," *Wired*, October 23, 2014, https://www.wired.com/2014/10/content-moderation.

115 See Sarah T. Roberts, *Behind the Screen: Content Moderation in the Shadows of Social Media* (New Haven, CT: Yale University Press, 2019), 41–43.

116 See Roberts, *Behind the Screen*, 41–43.

117 For example, content that does not violate YouTube's policies but is close to the removal line and could be offensive to some viewers may have some features disabled, such as comments, suggested videos, or likes, or it may be placed behind a warning label. YouTube Help, "Limited Features for Certain Videos," Community Guidelines Enforcement, last accessed June 22, 2021, https://support.google.com/youtube/answer/7458465?hl=en. Twitter employs various intermediate enforcement mechanisms, including limiting tweet visibility, placing a tweet behind a notice, and placing accounts in read-only mode for a set time period. See Twitter Help Center, "Our Range of Enforcement Options," General Guidelines and Policies, last accessed June 22, 2021, https://help.twitter.com/en/rules-and-policies/enforcement-options. Facebook is the least transparent about what intermediate techniques it employs. The company says that the consequences for violating its Community Standards "vary depending on the severity of the violation and the person's history on the platform." Facebook Community Standards, last accessed June 22, 2021, https://www.facebook.com/communitystandards.

118 For example, Twitter claimed that its efforts to slow the spread of election misinformation ahead of the 2020 U.S. presidential election worked, reporting that it saw a 29 percent reduction in the sharing of tweets after being prompted that they might be spreading misleading information. Without more transparency, the danger exists that these intermediate controls could be employed haphazardly and inconsistently. See Shannon Boyd, "Twitter Says Steps To Curb Election Misinformation Worked," NPR, November 12, 2020, https://www.npr.org/sections/live-updates-2020-election-results/2020/11/12/934267731/twitter-says-steps-to-curb-election-misinformation-worked.

119 Danielle Abril, "Facebook Reveals That Massive Amounts of Misinformation Flooded Its Service during the Election," *Fortune*, November 19, 2020, https://fortune.com/2020/11/19/facebook-misinformation-labeled-180-million-posts-2020-election-hate-speech-prevalence.

120 Adi Robertson, "Facebook Users Rarely Saw Voting Misinformation Labeled 'False,' Says Study — Especially If It Came from Trump," *Verge*, February 16, 2021, https://www.theverge.com/2021/2/16/22285553/facebook-the-markup-citizen-browser-data-election-labels-trump.

121 See, e.g., Spandana Singh and Margerite Blase, "Facebook/Instagram," *Protecting the Vote: How Internet Platforms Are Addressing Election and Voter Suppression-Related Misinformation and Disinformation*, Open Technology Institute, New America, September 30, 2020, https://www.newamerica.org/oti/reports/protecting-vote/facebookinstagram.

122 Craig Silverman and Ryan Mac, "Facebook Knows That Adding Labels to Trump's False Claims Does Little to Stop Their Spread," *BuzzFeed News*, November 16, 2020, https://www.buzzfeednews.com/article/craigsilverman/facebook-labels-trump-lies-do-not-stop-spread; and Tyler Sonnemaker, "Facebook Failed to Put Fact-Check Labels on 60% of the Most Viral Posts Containing Georgia Election Misinformation That Its Own Fact-Checkers Had Debunked, a New Report Says," *Business Insider*, December 4, 2020, https://www.businessinsider.com/facebook-mislabeled-60-of-georgia-election-misinfo-posts-report-2020-12.

123 See Cade Metz, "Feeding Hate with Video: A Former Alt-Right YouTuber Explains His Methods," *New York Times*, April 15, 2021, https://www.nytimes.com/2021/04/15/technology/alt-right-youtube-algorithm.html.

124 Adam Mosseri, "News Feed Ranking in Three Minutes Flat," Facebook Newsroom, May 22, 2018, https://newsroom.fb.com/news/2018/05/inside-feed-news-feed-ranking. See also Akos Lada, Meihong Wang, and Tak Yan, "How Machine Learning Powers Facebook's News Feed Ranking Algorithm," Facebook Engineering, January 26, 2021, https://engineering.fb.com/2021/01/26/ml-applications/news-feed-ranking.

125 See Elinor Carmi, "'It's Not You, Juan, It's Us': How Facebook Takes over Our Experience," *Tech Policy Press*, January 28, 2021, https://techpolicy.press/its-not-you-juan-its-us-how-facebook-takes-over-our-experience.

126 Karen Hao, "How Facebook Got Addicted to Spreading Misinformation," *MIT Technology Review*, March 11, 2021, https://www.technologyreview.com/2021/03/11/1020600/facebook-responsible-ai-misinformation ("All Facebook users have some 200 'traits' attached to their profile. These include various dimensions submitted by users or estimated by machine-learning models, such as race, political and religious leanings, socioeconomic class, and level of education.").

127 Mark Zuckerberg, "One of our big focus areas for 2018 is making sure the time we all spend on Facebook is time well spent . . . ," Facebook, January 11, 2018, https://www.facebook.com/zuck/posts/one-of-our-big-focus-areas-for-2018-is-making-sure-the-time-we-all-spend-on-face/10104413015393571 ("Based on this, we're making a major change to how we build Facebook. I'm changing the goal I give our product teams from focusing on helping you find relevant content to helping you have more meaningful social interactions.").

128 See Kevin Roose, Mike Isaac, and Sheera Frenkel, "Facebook Struggles to Balance Civility and Growth," *New York Times*, November 24, 2020, https://www.nytimes.com/2020/11/24/technology/facebook-election-misinformation.html.

129 See Nick Clegg, "You and the Algorithm: It Takes Two to Tango," Medium, March 31, 2021, https://nickclegg.medium.com/you-and-the-algorithm-it-takes-two-to-tango-7722b19aa1c2.

130 Clegg, "You and the Algorithm."

131 See Jack Nicas, "How YouTube Drives People to the Internet's Darkest Corners," *Wall Street Journal*, February 7, 2018, https://www.wsj.com/articles/how-youtube-drives-viewers-to-the-internets-darkest-corners-1518020478. Recommendations find their way into various elements of YouTube, from search results to the home page to an "Up Next" queue of videos that are set up to play automatically once the previous one ends.

132 See Paul Covington, Jay Adams, and Emre Sargin, "Deep

Neural Networks for YouTube Recommendations," *RecSys '16: Proceedings of the 10th Association for Computing Machinery Conference on Recommender Systems*, Boston, MA, September 15–19, 2016, https://dl.acm.org/doi/pdf/10.1145/2959100.2959190.

133 See Covington et al., "Deep Neural Networks."

134 See Joan E. Solsman, "YouTube's AI Is the Puppet Master over Most of What You Watch," *CNET*, January 10, 2018, https://www.cnet.com/news/youtube-ces-2018-neal-mohan/.

135 Nicas, "How YouTube Drives People." See also Lewis and McCormick, "How an Ex-YouTube Insider Investigated." On the other hand, at least one recent study suggests that YouTube's recommendation engine promotes extremist content for users already predisposed to engage with it, but that its promotion across all users may not be uniform. See Annie Y. Chen et al., "Exposure to Alternative and Extremist Content on YouTube," ADL Center for Technology & Society, 2021, https://www.adl.org/resources/reports/exposure-to-alternative-extremist-content-on-youtube#results.

136 Ben Popken, "As Algorithms Take Over, YouTube's Recommendations Highlight a Human Problem," NBC News, April 19, 2018, https://www.nbcnews.com/tech/social-media/algorithms-take-over-youtube-s-recommendations-highlight-human-problem-n867596.

137 Google, *How Google Fights Disinformation*, February 2019, https://storage.googleapis.com/gweb-uniblog-publish-prod/documents/How_Google_Fights_Disinformation.pdf.

138 Essam El-Dardiry, "Giving You More Control over Your Homepage and Up Next Videos," YouTube Official Blog, June 26, 2019, https://blog.youtube/news-and-events/giving-you-more-control-over-homepage.

139 El-Dardiry, "Giving You More Control."

140 See YouTube Team, "The Four Rs of Responsibility, Part 2: Raising Authoritative Content and Reducing Borderline Content and Harmful Misinformation," YouTube Official Blog, December 3, 2019, https://blog.youtube/inside-youtube/the-four-rs-of-responsibility-raise-and-reduce. See also YouTube Help, "External Evaluators and Recommendations," Help Center, accessed June 22, 2021, https://support.google.com/youtube/answer/9230586.

141 Leslie Miller, "Our Approach to Election Day on YouTube," YouTube Official Blog, October 27, 2020, https://blog.youtube/news-and-events/our-approach-to-election-day-on-YouTube. See also Yochai Benkler, Robert Faris, and Hal Roberts, *Network Propaganda: Manipulation, Disinformation, and Radicalization in American Politics* (New York: Oxford University Press, 2018).

142 See Niraj Chokshi, "YouTube Fact-Checks the Fire at Notre-Dame with Facts About . . . 9/11," *New York Times*, April 16, 2019, https://www.nytimes.com/2019/04/16/technology/youtube-notre-dame-fire.html.

143 See Jesse McCrosky and Brandi Geurkink, "YouTube Regrets: A crowdsourced investigation into YouTube's recommendation algorithm," Mozilla Foundation, July 2021, https://assets.mofoprod.net/network/documents/Mozilla_YouTube_Regrets_Report.pdf.

144 McCrosky and Geurkink, "YouTube Regrets," 3–4.

145 See, e.g., Corin Faife, "In Georgia, Facebook's Changes Brought Back a Partisan News Feed," *Markup*, January 5, 2021, https://themarkup.org/citizen-browser/2021/01/05/in-georgia-facebooks-changes-brought-back-a-partisan-news-feed; and Paul Lewis and Erin McCormick, "How an Ex-YouTube Insider Investigated Its Secret Algorithm," *Guardian*, February 2, 2018, https://www.theguardian.com/technology/2018/feb/02/youtube-algorithm-election-clinton-trump-guillaume-chaslot.

146 See, e.g., Casey Newton, "How Extremism Came to Thrive on YouTube," *Verge*, April 3, 2019, https://www.theverge.com/interface/2019/4/3/18293293/youtube-extremism-criticism-bloomberg.

147 See, e.g., Alexandra Stevenson, "Facebook Admits It Was Used to Incite Violence in Myanmar," *New York Times*, November 6, 2018, https://www.nytimes.com/2018/11/06/technology/myanmar-facebook.html; Purnell and Horwitz, "Facebook's Hate-Speech Rules"; and Ben Collins and Brandy Zadrozny, "In Trump's 'LIBERATE' Tweets, Extremists See a Call to Arms," NBC News, April 17, 2020, https://www.nbcnews.com/tech/security/trump-s-liberate-tweets-extremists-see-call-arms-n1186561.

148 Klonick, "Facebook v. Sullivan." See also Joel Kaplan and Justin Osofsky, "Input from Community and Partners on Our Community Standards," Facebook Newsroom, October 21, 2016, https://about.fb.com/news/2016/10/input-from-community-and-partners-on-our-community-standards/; and Nick Clegg, "Facebook, Elections and Political Speech," Facebook Newsroom, September 24, 2019, https://about.fb.com/news/2019/09/elections-and-political-speech.

149 Facebook employees evaluate newsworthiness on a case-by-case basis. Facebook officials purportedly weigh the value of "voice" against the risk of harm and consider whether the content was posted by a public figure — meaning they are a politician, have a certain number of followers, or are part of the media. Facebook Newsroom, "Mark Zuckerberg Stands for Voice and Free Expression."

150 See Facebook Oversight Board, Case Decision 2021-001-FB-FBR, May 5, 2021, 12, https://www.oversightboard.com/sr/decision/2021/001/pdf-english.

151 See Facebook, "Facebook Responses to Oversight Board Recommendations," accessed June 22, 2021, https://about.fb.com/wp-content/uploads/2021/06/Facebook-Responses-to-Oversight-Board-Recommendations-in-Trump-Case.pdf ("We applied the newsworthiness exception to an August 15th, 2019 video on Mr. Trump's Page. During the video from a New Hampshire rally, Mr. Trump says: 'That guy's got a serious weight problem. Go home. Start exercising.'").

152 See Facebook, "Facebook Responses to Oversight Board Recommendations."

153 Facebook Oversight Board, Case Decision 2021-001-FB-FBR, 24.

154 Facebook Transparency Center, "Reviewing High-Visibility Content Accurately," last updated June 11, 2021, https://transparency.fb.com/enforcement/detecting-violations/reviewing-high-visibility-content-accurately.

155 See Facebook Transparency Center, "Our Approach to Newsworthy Content," last updated June 16, 2021, https://transparency.fb.com/features/approach-to-newsworthy-content.

156 Facebook, "Facebook Responses to Oversight Board Recommendations," 11.

157 Facebook Transparency Center, "Our Approach to Newsworthy Content."

158 In the company's responses to the Oversight Board's recommendations, Facebook claims that it will "begin providing regular updates about when we apply our newsworthiness allowance." Facebook, "Facebook Responses to Oversight Board Recommendations," 11.

159 Twitter Help Center, "Twitter Rules."

160 Twitter Help Center, "About Public-Interest Exceptions" ("At present, we limit exceptions to one critical type of public-interest content — Tweets from elected and government officials — given the significant public interest in knowing and being able to discuss their actions and statements. . . . Criteria for exceptions . . . The Tweet violates one or more Twitter Rules; . . . The Tweet was posted by a verified account; . . . The account has more than 100,000 followers; and . . . The account represents a current or potential member of a local, state, national, or supra-national governmental or legislative body: 1) Current holders of an elected or appointed leadership position in a governmental or legislative body, OR 2) Candidates or nominees for political office, OR 3) Registered political parties").

161 Twitter Help Center, "About Public-Interest Exceptions." See also Twitter Safety (@TwitterSafety), "Defining Public Interest on

Twitter," Twitter Blog, June 27, 2019, https://blog.twitter.com/en_us/topics/company/2019/publicinterest.html.

162 Twitter Help Center, "About Public-Interest Exceptions." Given the extent to which public figures break Twitter's rules, it remains unclear whether the policy is actually applied sparingly, or if there are other ways that Twitter decides not to apply its rules to certain public figures.

163 Compare Twitter Help Center, "Our Range of Enforcement Options" (noting that the public interest exception is applied in "rare cases") with Tara Law, "In a First, Twitter Adds 'Unsubstantiated' Warning to 2 of President Trump's Tweets," *Time*, May 26, 2020, https://time.com/5842896/trump-warning-twitter-tweets-misleading. See also Elizabeth Dwoskin and Tony Romm, "Twitter Adds Labels for Tweets That Break Its Rules — A Move with Potentially Stark Implications for Trump's Account," *Washington Post*, June 27, 2019, https://www.washingtonpost.com/technology/2019/06/27/twitter-adds-labels-tweets-that-break-its-rules-putting-president-trump-companys-crosshairs ("The new policy applies to political candidates and government officials who have more than 100,000 followers, Twitter said, and will be used in rare occasions.").

164 YouTube Help, "The Importance of Context," YouTube Policies, accessed June 22, 2021, https://support.google.com/youtube/answer/6345162?hl=en.

165 Theodore Schleifer, "Facebook and YouTube Will Keep Letting Politicians Say What They Want If It's 'Newsworthy,'" *Vox*, September 26, 2019, https://www.vox.com/recode/2019/9/26/20885783/facebook-twitter-youtube-policies-political-content.

166 See Susan Wojcicki, "A Conversation with YouTube CEO Susan Wojcicki," discussion hosted virtually by the Atlantic Council, March 4, 2021, https://www.atlanticcouncil.org/event/youtubes-wojcicki.

167 See Rory Cellan-Jones, "YouTube Suspends Donald Trump's Channel," BBC News, January 13, 2021, https://www.bbc.com/news/technology-55643774.

168 York, *Silicon Values*, 41.

169 See Fischer and Gold, "All the Platforms."

170 See Associated Press, "The Latest: Trump Promises 'Orderly Transition' on Jan. 20," January 7, 2021 https://apnews.com/article/ap-electoral-college-congress-7af85d3c702e070464d7713c-42cf254a; and Mark Moore, "Trump Calls His Supporters 'Great Patriot,' Says They Will 'Remember This Day Forever,'" *New York Post*, January 6, 2021, https://nypost.com/2021/01/06/trump-calls-his-supporters-great-patriots/.

171 See Donald J. Trump (@DonaldTrump), "I can't stand back & watch this happen to a great American City, Minneapolis. A total lack of leadership . . . ," Facebook, May 28, 2020, https://www.facebook.com/153080620724/posts/10164767134275725 ("... These THUGS are dishonoring the memory of George Floyd, and I won't let that happen. Just spoke to Governor Tim Walz and told him that the Military is with him all the way. Any difficulty and we will assume control but, when the looting starts, the shooting starts. . . .").

172 Barbara Sprunt, "The History Behind 'When the Looting Starts, the Shooting Starts,'" NPR, May 29, 2020, https://www.npr.org/2020/05/29/864818368/the-history-behind-when-the-looting-starts-the-shooting-starts.

173 See Nicholas Bogel-Burroughs, Shaila Dewan, and Kathleen Gray, "F.B.I. Says Michigan Anti-government Group Plotted to Kidnap Gov. Gretchen Whitmer," *New York Times*, October 8, 2020, https://www.nytimes.com/2020/10/08/us/gretchen-whitmer-michigan-militia.html.

174 See Facebook Oversight Board, Case Decision 2021-001-FB-FBR, 21.

175 Bizarrely, Facebook claimed that it "has never applied the newsworthiness allowance to content posted by the Trump Facebook page or Instagram account." The company told the Oversight Board that 20 pieces of content from Trump's accounts were initially marked as violating Facebook rules but were ultimately determined (by an unnamed person or group of persons) not to be violations. See Facebook Oversight Board, Case Decision 2021-001-FB-FBR, 12.

176 See YouTubeInsider (@YouTubeInsider), "1/ After review, and in light of concerns about the ongoing potential for violence, we removed new content uploaded to Donald J. Trump's channel for violating our policies. It now has its 1st strike & is temporarily prevented from uploading new content for a *minimum* of 7 days," Twitter, January 12, 2021, 8:04 p.m., https://twitter.com/YouTubeInsider/status/1349205688694812672.

177 See Richard Nieva, "YouTube Extends Trump's Suspension for a Second Time," *CNET*, January 26, 2021, https://www.cnet.com/news/youtube-extends-trumps-suspension-for-a-second-time.

178 See Elizabeth Culliford and Paresh Dave, "YouTube Will Lift Ban on Trump Channel When Risk of Violence Decreases: CEO," Reuters, March 4, 2021, https://www.reuters.com/article/us-youtube-trump-suspension/youtube-will-lift-ban-on-trump-channel-when-risk-of-violence-decreases-ceo-idUSKBN2AW2LI.

179 See Twitter Inc., "Permanent Suspension of @realDonaldTrump," Twitter Blog, January 8, 2021, https://blog.twitter.com/en_us/topics/company/2020/suspension.html.

180 Facebook Oversight Board, "Oversight Board Upholds Former President Trump's Suspension, Finds Facebook Failed to Impose Proper Penalty," May 2021, https://oversightboard.com/news/226612455899839-oversight-board-upholds-former-president-trump-s-suspension-finds-facebook-failed-to-impose-proper-penalty.

181 Facebook Oversight Board, "Oversight Board Upholds Former President Trump's Suspension."

182 See Facebook Oversight Board, Case Decision 2021-001-FB-FBR, 34–37.

183 See Facebook, "Facebook Responses to Oversight Board Recommendations," 9.

184 Facebook, "Facebook Responses to Oversight Board Recommendations," 9.

185 Megan Brown and Zeve Sanderson, "How Trump Impacts Harmful Twitter Speech: A Case Study in Three Tweets," Brookings Institution, October 22, 2020, https://www.brookings.edu/techstream/how-trump-impacts-harmful-twitter-speech-a-case-study-in-three-tweets.

186 See Rosalind S. Helderman, Spencer S. Hsu, and Rachel Weiner, "'Trump Said to Do So': Accounts of Rioters Who Say the President Spurred Them to Rush the Capitol Could Be Pivotal Testimony," *Washington Post*, January 16, 2021, https://www.washingtonpost.com/politics/trump-rioters-testimony/2021/01/16/01b3d5c6-575b-11eb-a931-5b162d0d033d_story.html.

187 See Helderman et al., "'Trump Said to Do So.'"

188 See Philip Bump, "Ted Cruz's Electoral Vote Speech Will Live in Infamy," *Washington Post*, January 6, 2021, https://www.washingtonpost.com/politics/2021/01/06/ted-cruzs-electoral-vote-speech-will-live-infamy; Danny Hakim and Elaina Plott, "Josh Hawley, Vilified for Exhorting Jan. 6 Protesters, Is Not Backing Down," *New York Times*, March 8, 2021, https://www.nytimes.com/2021/03/08/us/politics/josh-hawley-vilified-for-exhorting-jan-6-protesters-is-not-backing-down.html; and Lis Power, "In 2 Weeks After It Called the Election, Fox News Cast Doubt on the Results Nearly 800 Times," Media Matters for America, January 14, 2021, https://www.mediamatters.org/fox-news/2-weeks-after-it-called-election-fox-news-cast-doubt-results-nearly-800-times.

189 This is reflected in the creation of Facebook's Oversight Board, as well as Twitter's addition of an in-app appeal process. See, e.g., Brent Harris, "Oversight Board to Start Hearing Cases," Facebook Newsroom, October 22, 2020, www.about.fb.com/news/2020/10/oversight-board-to-start-hearing-cases (describing the launch of the

Facebook Oversight Board); and @dhicks and David Gasca, "A Healthier Twitter: Progress and More to Do," Twitter Blog, April 16, 2019, https://blog.twitter.com/en_us/topics/company/2019/health-update.html (describing Twitter's 60 percent faster response to appeals requests its new in-app appeal process).

190 For example, Facebook does not allow appeals for violations that it determines have "extreme safety concerns." The company has recently stated that it will provide appeals for content that was reported but not acted on, but it has provided little information about how this will work or in what circumstances it will be available. Facebook Transparency Center, "What Can Be Appealed: Understanding the Community Standards Enforcement Report," accessed July 6, 2021, https://transparency.fb.com/policies/improving/appealed-content-metric/.

191 The lack of information provided to users in Facebook's appeal process was highlighted in almost all the Oversight Board's initial decisions. Facebook Oversight Board, "Announcing the Oversight Board's First Case Decisions," January 2021, https://www.oversightboard.com/news/165523235084273-announcing-the-oversight-board-s-first-case-decisions. See also Faiza Patel and Laura Hecht-Felella, "Oversight Board's First Rulings Show Facebook's Rules Are a Mess," *Just Security*, February 19, 2021, https://www.justsecurity.org/74833/oversight-boards-first-rulings-show-facebooks-rules-are-a-mess.

192 See, e.g., Twitter Inc., "Insights from the 17th Twitter Transparency Report," Twitter Blog, January 11, 2021, https://blog.twitter.com/en_us/topics/company/2020/ttr-17.html ("The COVID-19 pandemic severely impacted business operations for all of us around the world. Given the changes in workflows, coupled with country specific COVID-19 restrictions, there was some significant and unpredictable disruption to our content moderation work and the way in which teams assess content and enforce our policies — a disruption that is reflected in some of the data presented today. We increased our use of machine learning and automation to take a wide range of actions on potentially abusive and misleading content, whilst continually focusing human review in areas where the likelihood of harm was the greatest."); Facebook Transparency Center, *Community Standards Enforcement Report December 2020*, accessed June 22, 2021, https://about.fb.com/wp-content/uploads/2021/02/CSER-Q4-2020-Data-Snapshot.pdf ("Note: Due to a temporary reduction in our review capacity as a result of COVID-19, we could not always offer our users the option to appeal. We still gave people the option to tell us they disagreed with our decision, which helped us review many of these instances and restore content when appropriate."); and Guy Rosen, "Community Standards Enforcement Report, Fourth Quarter 2020," Facebook Newsroom, February 11, 2021, https://about.fb.com/news/2021/02/community-standards-enforcement-report-q4-2020 ("We're slowly continuing to regain our content review workforce globally, though we anticipate our ability to review content will be impacted by COVID-19 until a vaccine is widely available. With limited capacity, we prioritize the most harmful content for our teams to review, such as suicide and self-injury content."). See also Facebook Oversight Board, Case Decision 2021-003-FB-UA, April 29, 2021, https://oversightboard.com/decision/FB-H6OZKDS3 (noting that "while the user appealed the decision to Facebook, they were informed that Facebook could not review the post again due to staff shortages caused by COVID-19. . . . While the Board appreciates these unique circumstances, it again stresses the importance of Facebook providing transparency and accessible processes for appealing their decisions. . . . To ensure users' access to remedy, Facebook should prioritize the return of this capacity as soon as possible.").

193 See Duarte et al., *Mixed Messages?*, 15.

194 Tomiwa Ilori, "Facebook's Content Moderation Errors Are Costing Africa Too Much," *Slate*, October 27, 2020, https://slate.com/technology/2020/10/facebook-instagram-endsars-protests-nigeria.html.

195 Ilori, "Facebook's Content Moderation Errors."

196 Alice Wu, "Strike You're Out! Or Maybe Not?" YouTube Official Blog, July 2, 2010, https://blog.youtube/news-and-events/strike-youre-out-or-maybe-not.

197 Twitter Help Center, "Our Range of Enforcement Options"; and Bickert, "Publishing Our Internal Enforcement Guidelines."

198 Originally, appeals were limited to posts removed for nudity or sexual activity, hate speech, or graphic violence. In these contexts, users were notified of the removal and given the option to request an additional review. Facebook would restore the post if it determined that its content reviewers had made a mistake. Later, in 2018, Facebook expanded appeals to removals under its bullying and harassment policy. By May 2019, Facebook was allowing appeals for takedowns of spam, terrorist content, and content depicting regulated goods like drugs and firearms. Users can also appeal disabled accounts. Facebook has recently stated that it will provide appeals for content that was reported but not acted on. However, it has provided little information about how this will work or in what circumstances it will be available. Facebook Transparency Center, *Community Standards Enforcement Report December 2020*.

199 "The Santa Clara Principles on Transparency and Accountability in Content Moderation," May 7, 2018, https://santaclaraprinciples.org. See also Sam Levin, "Civil Rights Groups Urge Facebook to Fix 'Racially Biased' Moderation System," *Guardian*, January 18, 2017, https://www.theguardian.com/technology/2017/jan/18/facebook-moderation-racial-bias-black-lives-matter; Electronic Frontier Foundation, "EFF, Human Rights Watch, and Over 70 Civil Society Groups Ask Mark Zuckerberg to Provide All Users with Mechanism to Appeal Content Censorship on Facebook," press release, November 13, 2018, https://www.eff.org/press/releases/eff-human-rights-watch-and-over-70-civil-society-groups-ask-mark-zuckerberg-provide; Angwin and Grassegger, "Facebook's Secret Censorship Rules"; and Ariana Tobin, Madeleine Varner, and Julia Angwin, "Facebook's Uneven Enforcement of Hate Speech Rules Allows Vile Posts to Stay Up," ProPublica, December 28, 2017, https://www.propublica.org/article/facebook-enforcement-hate-speech-rules-mistakes.

200 Nick Clegg, "Charting a Course for an Oversight Board for Content Decisions," Facebook, January 28, 2019, https://about.fb.com/news/2019/01/oversight-board.

201 Facebook Oversight Board, "Announcing the Oversight Board's First Cases and Appointment of Trustees," December 2020, https://www.oversightboard.com/news/719406882003532-announcing-the-oversight-board-s-first-cases-and-appointment-of-trustees; and Facebook Oversight Board, "Announcing the Oversight Board's First Case Decisions."

202 Patel and Hecht-Felella, "Oversight Board's First Rulings."

203 Faiza Patel and Laura Hecht-Felella, "Evaluating Facebook's New Oversight Board for Content Moderation," *Just Security*, November 19, 2019, https://www.justsecurity.org/67290/evaluating-facebooks-new-oversight-board-for-content-moderation.

204 Google, "Appeals," *Google Transparency Report*, accessed June 22, 2021, https://transparencyreport.google.com/youtube-policy/appeals?hl=en.

205 Jillian C. York and David Greene, "How to Put COVID-19 Content Moderation into Context," Brookings Institution, May 21, 2020, https://www.brookings.edu/techstream/how-to-put-covid-19-content-moderation-into-context.

206 Facebook Help Center, "I don't think Facebook should have taken down my post," accessed June 22, 2021, https://www.facebook.com/help/2090856331203011/?helpref=search&query=appeal&search_session_id=e388c7670b7e0af63910f-3de2257f7df&sr=16.

207 Twitter Help Center, "Appeal an Account Suspension or Locked Account," accessed June 22, 2021, https://help.twitter.com/forms/general?subtopic=suspended; and Twitter Safety (@TwitterSafety), "We move quickly to enforce our rules, but sometimes we don't have the full context and can make mistakes. To fix that, we added a way

for people to appeal our decision in the app and have been able to get back to people 60% faster than before," Twitter, April 2, 2019, 11:00 a.m., https://twitter.com/TwitterSafety/status/1113139073303089152.

208 In April 2021, Facebook announced that users would have the ability to appeal other people's content that has been left up to the Oversight Board. See Guy Rosen, "Users Can Now Appeal Content Left Up on Facebook or Instagram to the Oversight Board," Facebook Newsroom, April 13, 2021, https://about.fb.com/news/2021/04/users-can-now-appeal-content-left-up-on-facebook-or-instagram-to-the-oversight-board.

209 Patel and Hecht-Felella, "Oversight Board's First Rulings." See also Murphy et al., *Facebook's Civil Rights Audit*, 48.

210 See, e.g., Twitter, "Rules Enforcement"; and Google, "YouTube Community Guidelines Enforcement."

211 Spandana Singh and Leila Doty, "The Transparency Report Tracking Tool: How Internet Platforms Are Reporting on the Enforcement of Their Content Rules," Open Technology Institute, New America, April 8, 2021, https://www.newamerica.org/oti/reports/transparency-report-tracking-tool.

212 Google, "Featured Policies: Hate Speech," *Google Transparency Report*, accessed April 29, 2021, https://transparencyreport.google.com/youtube-policy/featured-policies/hate-speech?hl=en.

213 Google, "Featured Policies: Hate Speech" ("Across our policy areas, we continue to invest in the network of over 180 academics, government partners, and NGOs who bring valuable expertise to our enforcement systems, including through our Trusted Flagger program. In the context of hate speech, this includes global and local partners like Faith Matters, JFDA, Licra, and Observatorio Web.").

214 See, e.g., Google, "Featured Policies: Violent Extremism," *Google Transparency Report*, accessed April 29, 2021, https://transparencyreport.google.com/youtube-policy/featured-policies/violent-extremism?hl=en ("This protocol has been tested and proven effective, for example following the attack on a synagogue in Halle, Germany (October 2019) and following a shooting in Glendale, Arizona, US (May 2020).").

215 We make a number of recommendations in section V. A coalition of civil rights organizations has proposed a similar list of requested disclosures to better understand the impact of hateful activities online. Although some of the recommendations diverge somewhat, they all reflect a unified concern regarding platforms' inadequacy in reporting on the disparate effects of their content moderation practices. See Center for American Progress et al., *Recommended Internet Company Corporate Policies and Terms of Service to Reduce Hateful Activities*, October 2018, 6–7, https://assets.website-files.com/5bba6f4828dfc3686095bf6b/5bd0e36186e28d35874f0909_Recommended%20Internet%20Company%20Corporate%20Policies%20%20Terms%20of%20Service_final-10-24.pdf.

216 See, e.g., Twitter, "Rules Enforcement"; and Facebook Transparency Center, "Hate Speech," *Community Standards Enforcement Report*, February 2021, https://transparency.facebook.com/community-standards-enforcement#hate-speech.

217 Facebook Transparency Center, "Dangerous Organizations."

218 Google, "Appeals."

219 Section 230 provides two main protections. Section 230(c)(1) shields "interactive computer services" (broadly covering everything from a major social media company to a small web publisher) from liability for content uploaded by users. Section 230(c)(2) protects those same service providers from liability for voluntarily taking actions in good faith to restrict access to objectionable content. Despite these considerable protections, the law is not a complete liability shield. It does not immunize companies from liability for content that violates federal criminal laws, intellectual property laws, electronic communications privacy laws, and the most recent carve-outs for sex trafficking laws. In addition, a company can only rely on a defense under Section 230(c)(1) if it is not an "information content provider," which means that the platform must not be "responsible, in whole or in part, for the creation or development" of content. Communications Decency Act (CDA) of 1996, 47 U.S.C. § 230(c)–(f).

220 Republican and Democratic politicians alike have expressed frustration with Section 230 and signaled that the law should be revisited as a method of imposing accountability for large online platforms. The political parties are united in their belief that platforms are not appropriately moderating content, but their respective approaches as to the correct way to do so differ substantially. More than 20 proposals to update Section 230 are currently on the table, with more on the way. Some bills propose eliminating Section 230 entirely. See, e.g., Abandoning Online Censorship Act, H.R. 874, 117th Cong. (2021) (previously H.R. 8896, 116th Cong. (2020)). Others seek to limit Section 230 immunities to protect only removals of illegal content. See, e.g., Stop Suppressing Speech Act, S. 4828, 116th Cong. (2020). And still others seek to categorically exclude certain claims from Section 230's immunities. See, e.g., Safeguarding Against Fraud, Exploitation, Threats, Extremism and Consumer Harms (SAFE TECH) Act, S. 299, 117th Cong. (2021).

221 Addressing these recommendations via stand-alone regulation must consider the appropriate liability and legal remedies to ensure platform compliance.

222 For example, a full repeal of Section 230 could have serious implications for the ability of marginalized communities to communicate and organize online. Section 230 allows the internet to remain an important organizing tool for communities pushing back against the structural systems that stratify American society across race and gender without relying on traditional media's attention and coverage. Eliminating Section 230 could also further sacrifice social media's role in collecting and sharing documentation of government abuses that challenges official narratives, as it could prompt over-removal of content that risks liability for platforms. A second approach to reform, which looks to impose "politically neutral" moderation, would be a significant step backward for platforms that have made progress in addressing hate speech and harassment because these removals would no longer be immunized and may even be prohibited.

223 Joan Donovan, "Navigating the Tech Stack: When, Where and How Should We Moderate Content?," Centre for International Governance Innovation, October 28, 2019, https://www.cigionline.org/articles/navigating-tech-stack-when-where-and-how-should-we-moderate-content.

224 See Cindy Cohn, "Bad Facts Make Bad Law: How Platform Censorship Has Failed So Far and How to Ensure That the Response to Neo-Nazis Doesn't Make It Worse," *Georgetown Law Technology Review* 2, no. 2 (2018): 437–38 ("While each indirect intermediary provides a slightly different service, they are similar in that they are often unable to remove only a single post, and instead, they often remove an entire website, domain, or worse, a set of domains. They also generally have only limited technical interactions with their customers.").

225 See Peter Guest, Emily Fishbein, and Nu Nu Lusan, "TikTok Is Repeating Facebook's Mistakes in Myanmar," *Rest of World*, March 18, 2021, https://restofworld.org/2021/tiktok-is-repeating-facebooks-mistakes-in-myanmar. See also Jacob Schulz, "Substack's Curious Views on Content Moderation," *Lawfare*, January 4, 2021, https://www.lawfareblog.com/substacks-curious-views-content-moderation; and Diyora Shadijanova, "The Problem with Clubhouse," *Vice*, February 10, 2021, https://www.vice.com/en/article/z3vkde/clubhouse-app-misinformation-problem.

226 For example, Facebook's newsworthiness policy is most thoroughly explained in a legal analysis published by the Knight First Amendment Institute. See Klonick, "Facebook v. Sullivan."

227 YouTube already provides a limited version of this type of disclosure for particular rules. Google, "Featured Policies: Violent Extremism."

228 It should be noted that the extent to which platforms already

track users based on race remains unclear. See, e.g., Hao, "How Facebook Got Addicted to Spreading Misinformation" ("All Facebook users have some 200 'traits' attached to their profile. These include various dimensions submitted by users or estimated by machine-learning models, such as race, political and religious leanings, socioeconomic class, and level of education.").

229 See, e.g., YouTube Help, "Hate Speech Policy." YouTube's harassment policy, which prohibits "content that targets an individual with prolonged or malicious insults based on intrinsic attributes," also should facilitate greater insight into the removals it makes via its harassment policy based on targeting a person's intrinsic attributes. See YouTube Help, "Harassment and Cyberbullying Policies."

230 See, e.g., Issie Lapowsky, "Platforms vs. PhDs: How Tech Giants Court and Crush the People Who Study Them," *Protocol*, March 19, 2021, https://www.protocol.com/nyu-facebook-researchers-scraping. See also Leon Yin and Aaron Sankin, "Google Blocks Advertisers from Targeting Black Lives Matter YouTube Videos," *Markup*, April 9, 2021, https://themarkup.org/google-the-giant/2021/04/09/google-blocks-advertisers-from-targeting-black-lives-matter-youtube-videos.

231 The choice of body is important because it must act as a check against dishonest researchers but also ensure that the government does not exert undue influence over content moderation or place data in the hands of law enforcement and intelligence agencies.

232 For example, some projects may require personal data, whereas aggregate or anonymized data may be sufficient in other circumstances. The organization handling the research data sets must also maintain particularly rigorous data security protocols, given the potential for hacking or malicious interference.

233 The commission must also evaluate whether voluntary disclosure provisions in the Electronic Communications Privacy Act of 1986 are sufficient to facilitate the disclosures of user records for a given research proposal.

234 According to one researcher, several examples of this protection exist, ranging from "clean rooms" used for scientific research, to "data rooms" used during European Union antitrust and merger investigations, to rooms set up by platforms like Facebook to facilitate sharing user information with advertisers. See Mathias Vermeulen, "The Keys to the Kingdom: Overcoming GDPR-Concerns to Unlock Access to Platform Data for Independent Researchers," 16–17, draft submitted to Columbia University's Knight First Amendment Institute Data and Democracy Symposium, October 15–16, 2020, https://osf.io/vnswz.

235 This effort should consider protections against government agencies seeking to obtain access directly as well as through the co-option of academic research.

236 This measure may necessitate the nullification of nondisclosure agreements that prevent individuals from speaking out for fear of retaliation, and it may require a notice of immunity to employees and contractors for the disclosure of confidential information made in confidence and solely for the purpose of reporting or investigating a suspected violation of law.

237 Chloe Colliver and Jennie King, *The First 100 Days: Coronavirus and Crisis Management on Social Media Platforms*, Institute for Strategic Dialogue, June 2020, https://www.isdglobal.org/wp-content/uploads/2020/06/First-100-Days.pdf.

238 See, e.g., Sap et al., "Risk of Racial Bias," 1668–78.

239 Susan Benesch, "What Is Dangerous Speech?," Dangerous Speech Project, accessed June 23, 2021, https://dangerousspeech.org/about-dangerous-speech.

240 See Facebook Oversight Board, Case Decision 2021-001-FB-FBR, 35.

241 See Twitter Help Center, "About Public-Interest Exceptions."

242 See Oversight Board, Case Decision 2021-001-FB-FBR ("However, international human rights standards expect state actors to condemn violence (Rabat Plan of Action), and to provide accurate information to the public on matters of public interest, while also correcting misinformation (2020 Joint Statement of international freedom of expression monitors on COVID-19.").

243 Lewis, *Alternative Influence*.

244 See Accountable Tech, "Election Integrity Roadmap for Social Media Platforms," September 2020, https://accountabletech.org/wp-content/uploads/2020/09/Election-Integrity-Roadmap-for-Social-Media-Platforms.pdf.

245 Compare Facebook, "Facebook Responses to Oversight Board Recommendations," 6 ("We ensure that content reviewers are supported by teams with regional and linguistic expertise, including the context in which the speech is presented. And we will continue to provide adequate resources to support that work.") with Ilori, "Facebook's Content Moderation Errors."

246 See Facebook Oversight Board, Case Decision 2021-001-FB-FBR, 36.

247 For example, the Facebook Oversight Board cautions against incitement cushioned between "superficial encouragement to act peacefully or lawfully." See Facebook Oversight Board Case Decision, 2021-001-FB-FBR, 35.

248 See Facebook, "Facebook Responses to Oversight Board Recommendations," 2 ("During especially high-risk events, such as elections and large-scale protests, Facebook regularly establishes an Integrity Product Operations Center ('IPOC'), which is a working group composed of subject matter experts from our product, policy, and operations teams. This structure allows these experts to more quickly surface, triage, investigate, and mitigate risks on the platform.").

249 Global Internet Forum to Counter Terrorism, *GIFCT Transparency Report*.

250 See Vishal Manve, "Twitter Tells Kashmiri Journalists and Activists That They Will Be Censored at Indian Government's Request," Global Voices Advox, September 14, 2017, https://advox.globalvoices.org/2017/09/14/kashmiri-journalists-and-activists-face-twitter-censorship-at-indian-governments-request.

251 See Rajesh Roy and Newley Purnell, "India Threatens Twitter with Penalties If It Doesn't Block Accounts," *Wall Street Journal*, February 3, 2021, https://www.wsj.com/articles/india-threatens-twitter-with-penalties-if-it-doesnt-block-accounts-11612364787; and Newley Purnell, "India Accused of Censorship for Blocking Social Media Criticism amid Covid Surge," *Wall Street Journal*, April 26, 2021, https://www.wsj.com/articles/india-accused-of-censorship-for-blocking-social-media-criticism-amid-covid-surge-11619435006.

252 See Glenn Greenwald, "Facebook Says It Is Deleting Accounts at the Direction of the U.S. and Israeli Governments," *Intercept*, December 30, 2017, https://theintercept.com/2017/12/30/facebook-says-it-is-deleting-accounts-at-the-direction-of-the-u-s-and-israeli-governments.

253 See Yoni Kempinski, "Benny Gantz to Facebook and TikTok Executives: You Must Take Action," *Israel National News*, May 14, 2021, https://www.israelnationalnews.com/News/News.aspx/306224.

254 הנידמה תוטילקרפ (@praklitut.gov.il) (Israel State Attorney's Office Facebook profile), Facebook, May 19, 2020, 9:42 a.m., https://www.facebook.com/283445232509318/posts/916928559160979.

255 See Evelyn Douek, "The Free Speech Blind Spot: Foreign Election Interference on Social Media," in *Defending Democracies: Combating Foreign Election Interference in a Digital Age*, ed. Duncan B. Hollis and Jens David Ohlin (New York: Oxford University Press, 2021), 265–92.

ABOUT THE AUTHORS

▶ **Ángel Díaz** is counsel in the Liberty and National Security Program at the Brennan Center for Justice and an adjunct professor of clinical law at NYU School of Law. His work focuses on the intersection of technology with civil rights and civil liberties. He is active on issues related to online speech and content moderation, as well as matters related to police surveillance. Díaz authored or coauthored the following Brennan Center white papers and resources: *Law Enforcement Access to Smart Devices* (2020), *When Police Surveillance Meets the 'Internet of Things'* (2020*), Automatic License Plate Readers: Legal Status and Policy Recommendations for Law Enforcement Use* (2020), and *New York City Police Department Surveillance Technology* (2019). His work and commentary have been featured in outlets such as the Associated Press, NPR, the *Intercept*, *Slate*, the *New York Daily News*, *City & State*, the *Chicago Reporter*, *Just Security*, and Univision. Díaz received his BA and JD from the University of California, Berkeley.

▶ **Laura Hecht-Felella** is the George A. Katz Fellow with the Brennan Center's Liberty and National Security Program. She focuses on issues related to civil rights and technology, as well as content moderation and online speech. Prior to joining the Brennan Center, Hecht-Felella was an attorney at Brooklyn Legal Services, where she represented low-income New Yorkers in litigation seeking to prevent displacement and preserve affordable housing. Previously, she worked on reproductive justice issues at National Advocates for Pregnant Women. She is a graduate of the Macaulay Honors College of the City University of New York and NYU School of Law.

ABOUT THE BRENNAN CENTER'S LIBERTY AND NATIONAL SECURITY PROGRAM

The Brennan Center's Liberty and National Security Program works to advance effective national security policies that respect constitutional values and the rule of law, using research, innovative policy recommendations, litigation, and public advocacy. The program focuses on reining in excessive government secrecy, ensuring that counterterrorism authorities are narrowly targeted to the terrorist threat, and securing adequate oversight and accountability mechanisms.

ACKNOWLEDGMENTS

The Brennan Center gratefully acknowledges the Bauman Foundation, CS Fund/Warsh-Mott Legacy, and Open Society Foundations for their generous support of our work.

The authors would like to express their gratitude to the Brennan Center's Faiza Patel for her sharp analysis and drafting guidance; to Raya Koreh, Priyam Madhukar, and Kaylana Mueller-Hsia for their research and cite-checking contributions; to Stephanie Sykes, Mireya Navarro, Zachary Laub, Lisa Benenson, and Alden Wallace for their editing and communications assistance; and to Michael Waldman and John Kowal for their guidance and support. They also thank David Brody, Daphne Keller, and Spandana Singh for their invaluable review and subject-matter expertise. Finally, the authors thank the numerous journalists, civil society organizations, and advocates who have documented and shared the online experiences of marginalized communities.

BRENNAN CENTER FOR JUSTICE

Brennan Center for Justice at New York University School of Law
120 Broadway // 17th Floor // New York, NY 10271
www.brennancenter.org