TOXIC TWITTER:

THE STATE OF HATEFUL ACTIVITIES ON THE PLATFORM

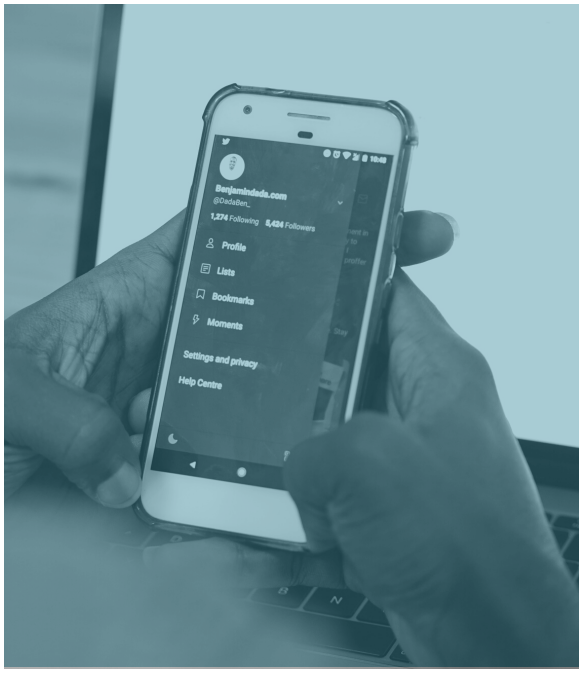
BY GAURAV LAROIA AND CARMEN SCURATO

NOVEMBER 2019

Free Press

> We underestimated the level of bad actors that we would see and the level of impact they would have.

EV WILLIAMS[1]
TWITTER CO-FOUNDER
May 2019



"We must ensure that all voices can be heard. We must continue to make improvements to our service so that everyone feels safe participating in the public conversation — whether they are speaking or simply listening. And we must ensure that people can trust in the credibility of the conversation and its participants."

—Twitter CEO Jack Dorsey, September 2018[2]

BACKGROUND & INTRODUCTION

Twitter struggles with pervasive hate speech and harassment.[3] While it lives in the shadow of larger social networks like Facebook,[4] it has an outsized effect on U.S. discourse as perhaps *the* platform where public intellectuals, journalists, activists and public officials gather to read, make and announce the news.[5]

Twitter's reach extends beyond its own platform. Tweets often make the rounds on platforms like Facebook and Instagram in the form of screenshots and many Facebook and Instagram accounts are dedicated to sharing screenshots from Twitter. Despite the site's prominence as a central place for public debate, it's failed to meaningfully address the prevalence of white supremacists and hateful activities on its network.

On Oct. 25, 2018, Change the Terms,[6] a coalition of racial-justice and civil-rights groups, launched a set of recommended corporate policies[7] to reduce hateful activities on internet platforms. Change the Terms' mission is to encourage these sites to take a stand against the kinds of online activities that makes these platforms dangerous and toxic places for people of color, women and other marginalized groups and to ensure that the companies' policies are enforced in an equitable, culturally relevant and transparent way. Coalition members have met with Twitter representatives multiple times over the past year and provided guidance and feedback on its policies.

> "People are being taken down who are protesting racism and people are staying up who are wildly racist and organizing racist rallies using social media and using Twitter, in particular.
>
> "Twitter needs to do a wholesale reform of its content-moderation policy. We can't have this happen piecemeal. It's offensive that they're not going head on after white supremacy, and we think they ought to."
>
> —Free Press Vice President and Change the Terms co-founder Jessica J. González, Gizmodo, July 2019[8]

Yet Twitter has failed to seriously engage with recommendations for meaningful change to decrease hateful activities on its site.

This is not for lack of ability; in fact, Twitter boasts of effective efforts to squash international terrorism and remove ISIS content.[9] Yet here, in its country of origin, it's failed to grapple with the prolific threats posed by domestic white-supremacist terrorist groups. The company's attempts to understand and abate white supremacy have done little to make the site safer for the targets of abuse. For example, CEO Jack Dorsey committed to a civil-rights audit of the platform during his congressional testimony in September 2018,[10] but the company has failed to initiate this independent review more than a year later.

This paper will examine the extent to which Twitter has revised its policies since the Change the Terms coalition's launch and makes recommendations for needed further improvements. To measure how Twitter has changed the terms for the better, we've evaluated the seven categories listed in the coalition's Model Corporate Policies to Prevent Hateful Activities:

1. Terms of service and acceptable-use policies
2. Enforcement
3. Right of appeal
4. Transparency
5. Evaluation and training & governance and authority
6. State actors, bots and troll campaigns

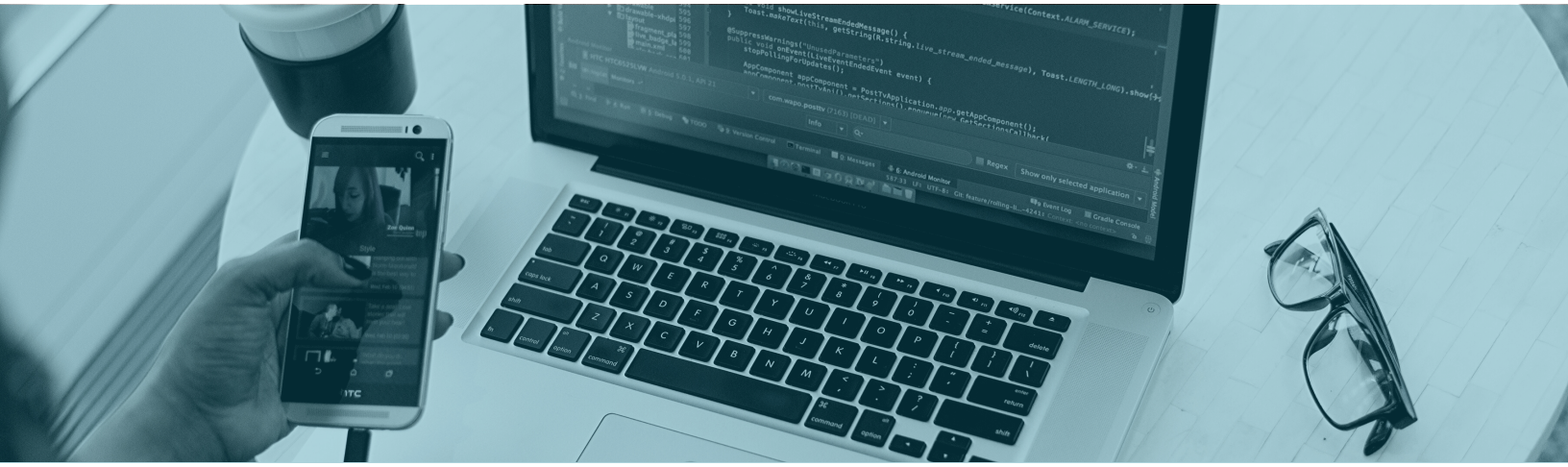


Photo credit: Flickr user WOCinTech Chat

TWITTER'S "HEALTH" FOCUS & DEFINITIONS

In early 2018, Twitter CEO Jack Dorsey publicly committed to "increase the collective health, openness, and civility of public conversation,"[11] pledging that "Twitter's health will be built and measured by how we help encourage more healthy debate, conversations, and critical thinking."[12] In July 2018, the company announced that it would initiate research projects focusing on conversational health.[13] The results of these studies have yet to be released.

The company's content-moderation efforts have focused on its "health" paradigm instead of on combating racism and harassment on the platform. Yet Twitter's view on health is both too vague and too narrow. For instance, it fails to consider how abusive hate speech and bigotry impact the health of users and dialogue. For instance, Twitter doesn't seem to recognize that hate speech and harassment can silence its targets.

By definition Twitter allows 'hateful activities' and hateful people and entities

Change the Terms defines hateful activities as "activities that incite or engage in violence, intimidation, harassment, threats, or defamation targeting an individual or group based on their actual or perceived race, color, religion, national origin, ethnicity, immigration status, gender, gender identity, sexual orientation, or disability."[14] Twitter has failed to adopt the letter — much less the spirit — of this recommendation.

Twitter CEO Jack Dorsey has failed to protect the targets of abuse on his platform.

Photo credit: Flickr user JD Lasica

Twitter's definitions[15]

"Abuse/harassment: You may not engage in the targeted harassment of someone, or incite other people to do so. This includes wishing or hoping that someone experiences physical harm."

"Hateful conduct: You may not promote violence against or directly attack or threaten other people on the basis of race, ethnicity, national origin, sexual orientation, gender, gender identity, religious affiliation, age, disability, or serious disease. We also do not allow accounts whose primary purpose is inciting harm towards others on the basis of these categories."

"Hateful imagery and display names: You may not use hateful images or symbols in your profile image or profile header. You also may not use your username, display name, or profile bio to engage in abusive behavior, such as targeted harassment or expressing hate towards a person, group, or protected category."

The scope of protections Twitter promises its users falls short of Change the Terms' recommendations and even the company's own supposed focus on "healthy conversations." The company's focus on "direct" attacks on members of protected classes means that vast amounts of hateful activities and unhealthy conversations are permitted on its site. Its interpretation of incitement is so narrow as to preclude almost any action against dehumanizing rhetoric that we know can lead to violence.

Twitter also falls short by only banning accounts whose "primary purpose" is to inflict harm. This also sets the company up for failure. It isn't clear how often a Twitter user must engage in hateful activities to be banned under this provision. Many users engage in conversations devoid of hateful activities on Twitter most of the time but will still engage in hateful activities enough to seriously affect the health of conversations on the platform.

It isn't enough for Twitter to only ban accounts that are designed to engage in hateful activities. The company must also take "casual" hate seriously.

CHANGE THE TERMS' MODEL CORPORATE POLICIES

Model Policy #1: Terms of service and acceptable-use policies

Change the Terms asserts that terms of service "should, at a minimum, make it clear that using the service to engage in hateful activities on the service or to facilitate hateful activities off the service shall be grounds for terminating the service for a user."[16] We recommend that platforms adopt model language stating that "users may not use these services to engage in hateful activities or use these services to facilitate hateful activities engaged in elsewhere, whether online or offline."[17]

TWITTER'S PROGRESS

As with other social-media companies, Twitter's written policies meet some of the Change the Terms guidelines.

In September 2018, the company announced its intention to ban dehumanizing language based on a person's membership in an identifiable group "as this speech can lead to offline harm."[18] Twitter didn't specify which groups it would apply this prohibition to, but reports suggested the company was considering applying it to all tweets regarding all groups. Internal meetings reportedly featured a discussion of President Trump's tweet condemning countries like Haiti as an example of the kind of rhetoric it would ban. However, Twitter quickly backtracked away from a broad interpretation of its own policy.[19]

> "The scaling back of Twitter's efforts to define dehumanizing speech illustrates the company's challenges as it sorts through what to allow on its platform. While the new guidelines help it draw starker lines around what it will and will not tolerate, it took Twitter nearly a year to put together the rules — and even then they are just a fraction of the policy that it originally said it intended to create."[20]
>
> — *The New York Times*, Sept. 7, 2019

In November 2018, Twitter updated its policies to include bans on the misgendering and deadnaming of transgender individuals.[21]

In March 2019, under its "terrorism and violent extremism policy," the company stated that it "examine[s] a group's activities both on and off Twitter to determine whether they engage in and/or promote violence against civilians to advance a political, religious and/or social cause."[22] Twitter asserted that it will "immediately and permanently suspend any account"[23] that violates this policy.

Then in July 2019, Twitter updated its policy on dehumanizing language, but applied it *only* to dehumanizing language against religious groups — a narrow experiment the company said it is trying as it evolves.[24] It announced that:

> "We create our rules to keep people safe on Twitter, and they continuously evolve to reflect the realities of the world we operate within. Our primary focus is on addressing the risks of offline harm, and research shows that dehumanizing language increases that risk. As a result, after months of conversations and feedback[25] from the public, external experts and our own teams, we're expanding our rules against hateful conduct[26] to include language that dehumanizes others on the basis of religion."[27]

More than 50 CIVIL-RIGHTS, HUMAN-RIGHTS, TECHNOLOGY-POLICY, AND CONSUMER-PROTECTION ORGANIZATIONS have signed on in support of Change the Terms' recommended policies for corporations to adopt and implement to reduce hateful activities on their platforms.

Learn more at changetheterms.org.

Twitter justified this narrow application of its policy by claiming it was clear across languages and contexts and would aid in creating consistent enforcement. Twitter also asserted that a broader application would stymie efforts to reach out to hate groups and political groups.[28] In a blog post, the company explained that including only religious groups in its policy was important:

> "Clearer language — Across languages, people believed the proposed change could be improved by providing more details, examples of violations, and explanations for when and how context is considered. We incorporated this feedback when refining this rule, and also made sure that we provided additional detail and clarity across all our rules.
>
> "Narrow down what's considered — Respondents said that 'identifiable groups' was too broad, and they should be allowed to engage with political groups, hate groups, and other non-marginalized groups with this type of language. Many people wanted to 'call out hate groups in any way, any time, without fear.' In other instances, people wanted to be able to refer to fans, friends and followers in endearing terms, such as 'kittens' and 'monsters.'
>
> "Consistent enforcement — Many people raised concerns about our ability to enforce our rules fairly and consistently, so we developed a longer, more in-depth training process with our teams to make sure they were better informed when reviewing reports. For this update it was especially important to spend time reviewing examples of what could potentially go against this rule, due to the shift we outlined earlier."

In the months since Twitter announced this restrictive policy we have yet to hear how well it worked or if it met the metrics that the company articulated above. As the 2020 U.S. election nears we anticipate the level of dehumanizing rhetoric to vastly increase. The company must reconsider this decision and be prepared to tackle this rhetoric in the difficult months ahead.

RECOMMENDATIONS

Unlike other social-media sites,[29] Twitter clearly explains its rules and policies in one easy-to-access page, with further explanations linked to each of its policies. At least on its face, Twitter's definition of hateful conduct and its policy to examine a group's offline activities are superficially similar to Change the Terms' recommendations. But by neglecting to include all protected categories — and by failing to take incitement and dehumanization seriously — the company isn't promoting the kinds of conversations it's promised its users and shareholders.

Twitter must also reevaluate its rules on violent extremism to properly contextualize white supremacy. The company should permanently suspend white supremacists for "promot[ing] violence against civilians to advance a political, religious and/or social cause."[30]

Its policy on dehumanizing language falls far short of Change the Terms' guidelines. Twitter must expand its hateful-conduct policy and examine activities off the platform in this context just as it does in the context of its anti-terrorism policy. It must also expand its ban against dehumanizing language to all protected classes of people. Its stated reasons for slow-walking the full realization of this policy are unacceptable. The company must take a stand against hate speech and dehumanization against all the protected classes identified in Change the Terms' "hateful activities" definition.

Twitter's failure to broaden its dehumanization policy beyond protection of religious groups shows a lack of courage and a refusal to directly grapple with the need to hold accountable government officials and other powerful individuals whose posts harm marginalized groups.

Well-designed and articulated policies and rules are meaningless if a company isn't willing to enforce those rules to combat hateful activities on its platform.

Photo credit: Flickr user WOCinTech Chat

Model Policy #2: Enforcement

Change the Terms recommends that internet companies have enforcement strategies that adequately reflect the scope of hateful activities on their platforms. There are several specific recommendations,[31] including that users be allowed to flag hateful activities and that companies create "trusted-flagger programs" to empower outside groups to help these platforms enforce their policies.

Change the Terms enforcement guidelines
The Internet Company will do the following:

- Provide a well-resourced enforcement mechanism that combines technological solutions with staff responsible for reviewing usage of services to ensure that hateful activities are not present.
- Allow for individuals and organizations — but not government actors — to flag hateful activities, as well as flag groups and individuals engaged in hateful activities.
- Create a trusted-flagger program for vetted, well-established civil and human rights organizations to expedite review of potential hateful activities.
- Inform flaggers of the results of the company's review of the flagging, including what actions, if any, were taken and why the actions were or were not taken.

In addition, Change the Terms recommends that content moderation involve a combination of technological solutions as well as human review, with regular audits of both the technology and human efforts. The coalition also recommends that platforms notify flaggers about "what actions the internet company has taken and why, including if the internet company has chosen to take no action. This clarity encourages flagging of hateful activities, increases company accountability, and allows users to know whether their understanding of what hateful activities are is shared by the internet companies and services that they use."[32]

Finally, government actors should not be allowed to use these tools to flag content that is legal.

TWITTER'S PROGRESS

Twitter has various types of enforcement actions,[33] such as tweet-level enforcement, direct message-level enforcement and account-level enforcement. In its Help Center page on "reporting abusive behavior," Twitter explains that it will provide follow-up emails and notifications to affected users regarding the report, as well as "recommendations for additional actions you can take to improve your Twitter experience."[34]

Twitter allows individual users to flag content but hasn't established a trusted-flagger program. Other social-media companies, notably Facebook, have experimented with this model. In the absence of such programs, groups based in Washington, D.C., and other world capitals have more access to corporate decision-makers than organizations in the global South. We've found that trusted-flagger programs can make access more equitable.

As recommended by Change the Terms, Twitter does use a combination of human reviewers and technological tools to moderate content, but does not explain how this process works. It notes only that the information provided in its "abusive behavior report" does not include content removed through technological tools.[35] We'd like to see much more transparency about how this process works.

RECOMMENDATIONS

Though Twitter clearly states that it examines activities off of its platform, the enforcement of this policy is substandard. Though the policy as written would capture white supremacists calling for violence and genocide, Twitter routinely allows prominent white supremacists to spread hate and vitriol through its platform relatively unchecked.

A recent example was Twitter's temporary suspension of former KKK Grand Wizard David Duke's account for violating its hateful-conduct policy.[36] However, it's unclear how long this suspension lasted and what additional line Duke must cross for Twitter to permanently suspend his account.

In May 2017, Katie Hopkins called for a "final solution" in a tweet following the terrorist attack at the Ariana Grande concert in Manchester, England. Users reported it and Hopkins deleted the tweet yet she's still on Twitter. She's also compared migrants to cockroaches and has called for gunships to sink boats carrying refugees.

Change the Terms recommends that Twitter consistently enforce its rules and reach out to experts and civil-rights groups to incorporate a racial-justice analysis in its enforcement practices.

Model Policy #3: Right of appeal

We want to ensure that those protesting or reporting on hate and bigotry should have the right to appeal any material impairment, suspension or termination of service, whether that impairment, suspension or termination represents a permanent ban or a temporary one. We want to ensure that anyone protesting or reporting on bigotry who is swept up in enforcement has a quick and easy means to appeal.

This right should allow a user to make an appeal to a neutral decision-maker — someone other than the person who made the initial determination. That decision-maker should have knowledge of the context and social, political and cultural history of the user's country or countries.[37] The user filing the appeal should be permitted to present information to advocate for their position.

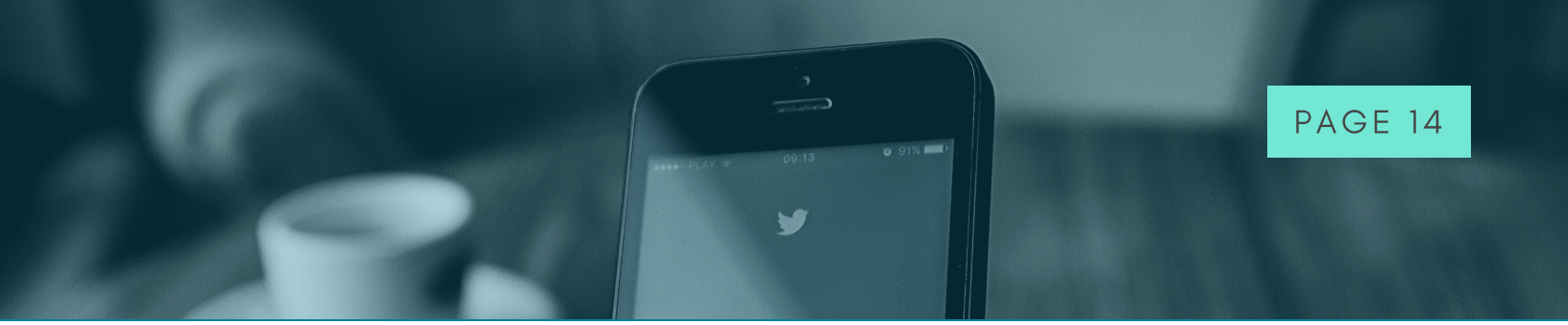
TWITTER'S PROGRESS

Twitter's appeals process mostly involves users writing to the site through its help center and using an online form to ask the company to revisit its content-moderation decisions. On April 2, 2019, Twitter rolled out a feature allowing users to appeal within the Twitter app instead of just an online form.[38] This change is welcome and long overdue.

RECOMMENDATIONS

The company has paid little attention to its appeals process. There's much we don't know, such as how many people work on appeals at Twitter or how the appeals process feeds back into the company's content-moderation policy decisions.

It's all too apparent that the company has failed to invest in its appeals and content-moderation processes. This is evident in the perennial stories about the company's under- and overenforcement of its policies and the lack of a fast appeals process to correct those mistakes. If the company wants to promote healthy conversations, it must have a clear and transparent process to fix any errors.



Model Policy #4: Transparency

Change the Terms offers several recommendations for increased transparency.

We suggest additional data points to evaluate what hateful activities are occurring on the platform and how Twitter is addressing those activities. The coalition asks that internet companies be transparent with the enforcement actions they're taking and provide explanations of what they're doing and who their policies affect. We ask that data be made available publicly in forms that are both human and machine readable. We ask that this information be made available to researchers and scholars so comparative studies can be undertaken and the companies' processes improved.

TWITTER'S PROGRESS

Twitter has been releasing biannual transparency reports since July 2012. It released its 15th major report in October 2019, detailing its enforcement numbers through June 2019 and comparing those to the data in its previous reporting period.[39]

Twitter has reported that:

- More than 50% of tweets Twitter takes action on for abuse are now proactively surfaced using technology, rather than relying on reports to Twitter;
- 105% increase in accounts actioned by Twitter (locked or suspended for violating the Twitter Rules);
- Continuing a year-on-year trend, a 30% decrease in accounts suspended for the promotion of terrorism; and
- 67% more global legal demands, originating from 49 different countries.[40]

In a separate report chronicling its enforcement actions from January through June 2019, the company stated that, "15,638,349 unique accounts were reported for possible violations of the Twitter Rules, amounting to a 42 percent increase compared to the prior reporting period." The company also noted that it "actioned 395,917 accounts under abuse policies, 584,429 accounts under hateful conduct policies, 43,536 under sensitive media policies, 30,107 under CSE [child sexual exploitation] policies, 124,339 under impersonation policies, 19,679 under private information policies, and 56,219 under violent threats policies."[41]

RECOMMENDATIONS

Twitter's transparency reports are high level and do not have the granularity needed to assess the breadth and depth of its content-moderation challenges. Twitter should provide regular updates about the number of hateful activities the company identifies. These updates should be broken down by protected characteristics, the types of targets, how and by whom the content was initially flagged, how many people have been denied services for their hateful activities, and the success rate of appeals.[42] For instance, though we know that Twitter "actioned" more than 500,000 accounts for hateful activities during this reporting period, we don't have any insight into the activities users were sanctioned for. We also don't have a granular-level breakdown of the kinds of enforcement activities the company took against those accounts.[43]

Twitter should publish this information in a format that protects users' personally identifiable information and should make this content available in formats that both people and machines can read, with clear date tags identifying the reporting period the numbers pertain to. Despite Twitter's assertion that "The data in these reports is as accurate and comprehensive as possible,"[44] the company hasn't yet released reports with the kinds of specific details Change the Terms calls for.

Model Policy #5: Evaluation and training & governance and authority

Evaluation and training

Change the Terms recommends that online platforms "establish a team of experts on hateful activities with the requisite authority to train and support programmers and assessors working to enforce anti-hateful activities programs of the terms of service, develop training materials and programs, as well as create a means of tracking the effectiveness of any actions taken to respond to hateful activities."

The coalition also urges each platform to assign a member of its executive team to serve as a senior manager focused on overseeing how the company addresses hateful activities. The senior manager would need to "approve all training materials, programs, and assessments."

Change the Terms recommends that platforms "routinely test any technology used to identify hateful activities to ensure that such technology is not biased against individuals or groups … make the training materials available to the public for review; locate assessment teams enforcing the hateful activities rules within affected communities to increase understanding of cultural, social, and political history and context."

Governance and authority

Change the Terms recommends that a company "integrate addressing hateful activities into [their] corporate structures in three ways":

1. Assign a committee comprised of members from a platform's corporate board to assess management efforts to stop hateful activities on their services.
2. Assign an executive-team member to serve as a senior manager to oversee addressing hateful activities. Name that person publicly and ensure they have adequate resources and authority.
3. Create a committee of outside advisers with expertise in identifying and tracking hateful activities who will produce an annual report on the effectiveness of the steps the company has taken.

TWITTER'S PROGRESS

Twitter provides little information about the investment it's making in evaluating and training its content-moderation teams or how those staff members fit into the company's corporate structure. Twitter has formed a few key partnerships and launched a few research projects — as when it partnered with UC Berkeley on machine learning — that seek to improve its content-moderation policies and "health" on its platform.[45]

Following the terrorist attack on two mosques in Christchurch, New Zealand, the company joined the May 2019 Christchurch call[46] committing itself to updating its terms of use; making it easy for users to report terrorist and violent extremist content; using the latest technology to aid this process; checking livestreams for violent content; and committing to releasing regular transparency reports.[47]

However, Twitter has failed to communicate to the public about any steps it's taken to get the right team in place to deal with mounting hate and racism that will only get worse during the 2020 U.S. election cycle.

RECOMMENDATIONS

While Twitter has stated its intention to combat violent extremism and promote "healthy conversations," we see little progress or disclosure in how the company intends to follow through on these promises.

The company has made investments in countering Islamic terrorist content but we haven't seen similar energy or investment in combating white-supremacist content. To our knowledge, the company hasn't initiated a civil-rights audit, nor are we aware of it promoting high-level corporate officers to see such an effort through. Twitter has assured the Change the Terms team that it has executives and other decision-makers in place who hold a deep understanding of race equity. But we have no data about the number of people at Twitter who belong to the groups most negatively impacted by rampant bigotry on its site and who have actual power to make policy change at the company.

Model Policy #6: State actors, bots and troll campaigns

Change the Terms recommends that platforms ban the use of bots or teams of individuals for coordinated campaigns that engage in hateful activities.

TWITTER'S PROGRESS

Social-media sites struggled to combat coordinated disinformation campaigns in the leadup to the 2016 presidential election. The Russian propaganda campaign on Twitter was more sophisticated than previously realized, with some Russian bots even earning money through promoting disinformation on the platform.[48]

On Oct. 30, 2019, Twitter CEO Jack Dorsey announced a "decision to stop all political advertising on Twitter globally." Dorsey cited his belief that political advertising presents vastly different equities than commercial advertising and that online political advertising "brings significant risks to politics, where it can be used to influence votes to affect the lives of millions."[49]

RECOMMENDATIONS

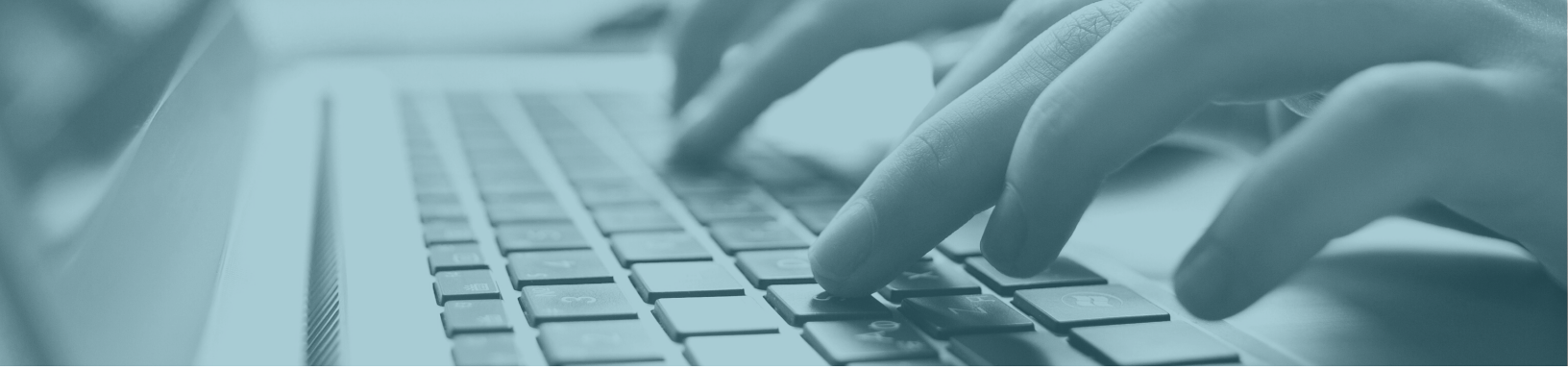
There's an urgent need to combat disinformation and racist propaganda campaigns on internet platforms in the leadup to the 2020 U.S. election. We recognize Twitter's decisive stand against political advertising on its platform. But we're concerned that replicating that decision across the web, where political and issue advertising can be instrumental in getting the message out about racial-justice issues, will actually stifle political participation and awareness.

CONCLUSION

Twitter hasn't seriously grappled with either ongoing harassment or white supremacy on its site. The company has made repeated presentations to its shareholders about encouraging "healthy" conversations and recently promised that it has a goal of "reducing the burden on victims of abuse" in conversations. Much of this is far too little and too late. The company's race-neutral approach to combating hateful activities on its site obscures the real harms inflicted on marginalized communities.

While Twitter has succeeded in getting press for high-profile one-off actions like banning right-wing conspiracy theorist Alex Jones or political advertisements, both the company and investors need to reckon with the fact that the platform has failed to address safety in a systematic way. It hasn't begun a promised civil-rights audit or taken any steps to integrate anti-discrimination and anti-harrassment programs at a senior level. Moreover, Twitter's lack of transparency regarding these efforts has stymied the ability of watchdog groups and civil-society organizations to meaningfully engage with the company to protect our communities.

If Twitter is to retain its position as the preeminent platform for important national discourse, it must make these investments and protect users from abuse.



ABOUT THE AUTHORS

GAURAV LAROIA

Gaurav works alongside the Free Press policy team on topics ranging from internet-freedom issues like Net Neutrality and media ownership to consumer privacy and government surveillance. Gaurav's human-rights and civil-liberties work has taken him from Capitol Hill to Uganda, India and Liberia. Before joining Free Press, he worked at the Government Accountability Project protecting the rights of national-security whistleblowers like Edward Snowden, and prior to that as a legislative counsel at the American Civil Liberties Union. He earned both his B.A. in international affairs and his J.D. from the George Washington University.

CARMEN SCURATO

Carmen works to protect the open internet, prevent media and telecom-industry concentration, promote affordable internet access and foster media diversity. She also coordinates responses to regulatory proposals that threaten to widen the digital divide and has authored pieces on the importance of Net Neutrality and the Lifeline program for communities of color. Before joining Free Press, Carmen led the policy team at the National Hispanic Media Coalition, where she was the architect of a series of prominent federal-records requests that compelled the FCC to release more than 50,000 previously undisclosed consumer complaints about Net Neutrality violations. Earlier in her career, she worked at the Department of Justice. Carmen, a native of Puerto Rico, earned her J.D. from Villanova University School of Law and her B.A. from New York University.

ENDNOTES

1. Alsup, Blake, "Twitter Co-Founder: Donald Trump Is a 'Master of the Platform,'" *Boston Herald*, May 23, 2019: https://www.bostonherald.com/2019/05/23/twitter-co-founder-donald-trump-is-a-master-of-the-platform/
2. Testimony of Twitter CEO Jack Dorsey in hearing on "Twitter: Transparency and Accountability," House Committee on Energy & Commerce, Sept. 5, 2019: https://energycommerce.house.gov/sites/democrats.energycommerce.house.gov/files/documents/Testimony%20-Dorsey-FC-Hrg-on-Twitter-Transparency-and-Accountabilit-2018-09-05.pdf
3. Scurato, Carmen, *Facebook vs. Hate*, Free Press, September 2019: https://www.freepress.net/sites/default/files/2019-09/facebooks_vs_hate_free_press_report.pdf"
4. "Social Media Fact Sheet," Pew Research Center, June 12, 2019: https://www.pewresearch.org/internet/fact-sheet/social-media/
5. Ingram, Matthew, "Do Journalists Pay Too Much Attention to Twitter?" *Columbia Journalism Review*, Oct. 10, 2018: https://www.cjr.org/the_media_today/journalists-on-twitter-study.php
6. Change the Terms is a coalition of more than 50 racial-justice and civil-rights groups. The core contributors are the Center for American Progress, Color Of Change, Free Press, the Lawyers' Committee for Civil Rights Under Law, the National Hispanic Media Coalition and the Southern Poverty Law Center: https://www.changetheterms.org/coalition (last visited on Nov. 12, 2019)
7. *See generally* "Change the Terms Recommended Internet Company Corporate Policies and Terms of Service to Reduce Hateful Activities," http://bit.ly/2lmfeUO ("Change the Terms Model Policies").
8. Cameron, Dell, "Civil Rights Groups Mostly Unimpressed by New Twitter Policy Against 'Dehumanizing' Language," Gizmodo, July 9, 2019: https://gizmodo.com/civil-rights-groups-mostly-unimpressed-by-new-twitter-p-1836227745
9. Broderick, Ryan and Hall, Ellie, "Tech Platforms Obliterated ISIS Online. They Could Use The Same Tools On White Nationalism," BuzzFeed News, March 20, 2019: https://www.buzzfeednews.com/article/ryanhatesthis/will-silicon-valley-treat-white-nationalism-as-terrorism"
10. "House Committee on Energy & Commerce Members Pallone and Rush Push Twitter to Conduct a Civil Rights Audit," Sept. 24, 2018: https://energycommerce.house.gov/newsroom/press-releases/pallone-rush-push-twitter-ceo-jack-dorsey-on-commitment-for-civil-rights
11. Dorsey, Jack (@jack), "We're committing Twitter to help increase the collective health, openness, and civility of public conversation, and to hold ourselves publicly accountable towards progress," posted March 1, 2018: https://twitter.com/jack/status/969234275420655616
12. Id.
13. Gadde, Vijaya and Gasca, David, "Measuring Healthy Conversation," Twitter blog, July 30, 2018: https://blog.twitter.com/en_us/topics/company/2018/measuring_healthy_conversation.html
14. Change the Terms Model Policies, p. 2: http://bit.ly/2lmfeUO
15. *See* Twitter, "The Twitter Rules": https://help.twitter.com/rules-and-policies/twitter-rules (last visited on Nov. 12, 2019)
16. Change the Terms Model Policies, p. 3
17. Change the Terms Model Policies, p. 4 (emphasis added)
18. Gadde, Vijaya and Harvey, Del, "Creating New Policies Together," Twitter blog, Sept. 25, 2018: https://blog.twitter.com/en_us/topics/company/2018/Creating-new-policies-together.html
19. Conger, Kate, "Twitter Backs off Broad Limits on 'Dehumanizing' Speech," *The New York Times*, Sept. 7, 2019: https://www.nytimes.com/2019/07/09/technology/twitter-ban-speech-dehumanizing.html
20. Id.
21. "Twitter Has Banned Misgendering or 'Deadnaming' Transgender People," The Verge, Nov. 27, 2018: https://www.theverge.com/2018/11/27/18113344/twitter-trans-user-hateful-content-misgendering-deadnaming-ban
22. *See* Twitter, "Terrorism and Violent Extremism Policy," March 2019: https://help.twitter.com/en/rules-and-policies/violent-groups (emphasis added)
23. Id.
24. Cameron, Dell, "Civil Rights Groups Mostly Unimpressed by New Twitter Policy Against 'Dehumanizing' Language," Gizmodo, July 9, 2019: https://gizmodo.com/civil-rights-groups-mostly-unimpressed-by-new-twitter-p-1836227745
25. Gadde, Vijaya and Harvey, Del, "Creating New Policies Together," Twitter blog, Sept. 25, 2018: https://blog.twitter.com/en_us/topics/company/2018/Creating-new-policies-together.html
26. *See* Twitter, "The Twitter Rules": https://help.twitter.com/en/rules-and-policies/twitter-rules
27. *See* Twitter, "Updating Our Rules Against Hateful Conduct," July 9, 2019: https://blog.twitter.com/en_us/topics/company/2019/hatefulconductupdate.html
28. Id.
29. Scurato, Carmen, *Facebook vs. Hate*, Free Press, September 2019: https://www.freepress.net/sites/default/files/2019-09/facebooks_vs_hate_free_press_report.pdf
30. *See* Twitter, "Terrorism and Violent Extremism Policy," March 2019: https://help.twitter.com/en/rules-and-policies/violent-groups
31. *See* Change the Terms, Enforcement, p. 4 (full list of enforcement recommendations)
32. Id.
33. *See* Twitter, "Our Range of Enforcement Options": https://help.twitter.com/en/rules-and-policies/enforcement-options (last visited on Oct. 10, 2019)
34. *See* Twitter Help Center, "Report Abusive Behavior": https://help.twitter.com/en/safety-and-security/report-abusive-behavior (last visited on Oct. 10, 2019)

ENDNOTES

35. Id. "This data does not take into account content that has been actioned using technological tools, the goal of which is to limit the reach and spread of potentially abusive content."
36. *See* Sleeping Giants, "People are reporting that David Duke has been given a suspension by @TwitterSafety. Which is good. Unless you consider that HE'S THE NATION'S FOREMOST WHITE SUPREMACIST AND HE'S STILL ON THIS FUCKING PLATFORM," posted on Oct. 2, 2019: https://twitter.com/slpng_giants/status/1179625531563659264?s=21
37. *See* Change the Terms, Right of Appeal, p. 6
38. Perez, Sarah, "Twitter Now Lets Users Appeal Violations Within Its App," TechCrunch, April 2, 2019: https://techcrunch.com/2019/04/02/twitter-now-lets-users-appeal-violations-within-its-app/
39. *See* Twitter, "Twitter Rules Enforcement": https://transparency.twitter.com/en/twitter-rules-enforcement.html (last visited on Nov. 12, 2019)
40. *See* Twitter, "15th Transparency Report: Increase in Proactive Enforcement on Accounts," Oct. 31, 2019: https://blog.twitter.com/en_us/topics/company/2019/twitter-transparency-report-2019.html
41. *See* Twitter, "Twitter Rules Enforcement": https://transparency.twitter.com/en/twitter-rules-enforcement.html (last visited on Nov. 12, 2019)
42. *See* Change the Terms, Transparency, pp. 6-7 (recommendation that platforms collect detailed information and make that data easily accessible to the general public)
43. *See* Twitter, "Twitter Rules Enforcement"" https://transparency.twitter.com/en/twitter-rules-enforcement.html (last visited on Nov. 12, 2019)
44. Id.
45. Erkan, Naz and Pandey, Sandeep, "Partnering with Researchers at UC Berkeley to Improve the Use of ML," Twitter blog, Jan. 29, 2019: https://blog.twitter.com/en_us/topics/company/2019/ucberkeley-twitter-ml.html
46. "Christchurch Call: To Eliminate Terrorist and Violent Extremist Content Online," New Zealand Ministry of Foreign Affairs and Trade: https://www.christchurchcall.com/
47. "Addressing the Abuse of Tech to Spread Terrorist and Extremist Content," Twitter blog, May 15, 2019: https://blog.twitter.com/en_us/topics/company/2019/addressing-the-abuse-of-tech-to-spread-terrorist-and-extremist-c.html
48. Starks, Tim and Cerulus, Laurens and Scott, Mark, "Russia's Manipulation of Twitter Was Far Vaster Than Believed," *Politico*, June 5, 2019: https://www.politico.com/story/2019/06/05/study-russia-cybersecurity-twitter-1353543
49. Dorsey, Jack (@jack), "We've made the decision to stop all political advertising on Twitter globally. We believe political message reach should be earned, not bought. Why? A few reasons … " posted on Oct. 30, 2019: https://twitter.com/jack/status/1189634360472829995