Stop Hate for Profit One Year Later: How Well Are Social Media Platforms Doing?

One year ago, the Stop Hate for Profit (SHFP) coalition called for Facebook to address the prevalence of hate, racism, and misinformation on their platforms. Led by prominent civil rights groups and nonprofit organizations including ADL, Color Of Change, Common Sense, Free Press, LULAC, Mozilla, NAACP, National Hispanic Media Coalition, and Sleeping Giants, Stop Hate for Profit garnered the support of thousands of businesses that paused spending on Facebook and Instagram advertisements in July 2020.

Over the past 12 months, it has become clear that the concerns raised by Stop Hate for Profit have only increased in urgency. We've seen a surge of extremist conspiracy theories, hate, and violence. The coronavirus pandemic was twisted to justify anti-Asian and antisemitic attacks, the country was roiled by unprecedented political polarization and a divisive election, and thousands attacked the U.S. Capitol in an attempt to violently overturn a valid presidential election—an event that was planned online in plain view. But we still lack adequate regulation, bipartisan political will, and, seemingly, the marketplace power to hold social media companies accountable for their role in amplifying hate and racism.

Did Stop Hate for Profit work? The campaign garnered enormous attention and support from many, including Facebook employees, platform advertisers, celebrities, civil society groups, and Congress. It succeeded in getting a number of incremental improvements from Facebook when nothing else had worked, but not the bold structural change that is needed. Some of the biggest wins came not from Facebook, but from other social media platforms that were not targeted by the campaign—but clearly hoped to avoid such targeting.

PROGRESS REPORT

Have social media companies risen to meet these pressing challenges?

The following is a summary of the progress Facebook and a number of other large social media companies have made over the past year in addressing hate and racism on their platforms. ADL and other organizations compared Stop Hate for Profit's demands to technology companies' policy changes. We then looked at the adequacy of their policy enforcement.

Red: Incremental changes
Orange: A few significant changes
Yellow: Many significant changes
Green: Significant structural changes

Facebook Overall Rating: Red

Facebook is the world's biggest social media platform with over two billion users, exerting an outsized influence over how people communicate and what information they receive. During the first quarter of 2021, Facebook earned more than $26 billion in revenue. Its dominance in the industry yields the company handsome dividends but inflicts harm upon society. The platform's algorithms recommend and spread inflammatory, divisive and inaccurate content to maximize user engagement and drive up advertising revenue. Multiple reports have found that Facebook continues to refuse to address the problem effectively. As a result, millions of people are targeted with, or impacted by, online hate and harassment, shaping how they view the world, and how they work, socialize, and communicate.

Facebook enacted some common-sense changes to its platform in response to the pressure of the Stop Hate for Profit campaign and others but these efforts have not yielded meaningful change to the way the platform operates.

Orange: CIVIL RIGHTS INFRASTRUCTURE

- SHFP demanded that Facebook create a civil rights department to report to its top executives. The coalition asked the company to establish a permanent civil rights infrastructure including a C-suite level executive with civil rights expertise to evaluate products and policies for discrimination, bias, and hate. This person would make sure that the design and decisions of this platform considered the impact on all communities and the potential for radicalization and hate.
    - The company hired civil rights attorney Roy Austin in January 2021, but appointed him as a vice president who would not report to the senior staff.
    - It also hired Cynthia Deitle, director of civil rights reform at the Matthew Shepard Foundation, as director and associate general counsel for civil rights.

Red: INDEPENDENT AUDITS

- SHFP demanded that Facebook submit to regular third-party audits on hate and misinformation.
    - While Facebook committed to auditing from the advertising industry in 2020, this has yet to be completed. Facebook also committed to an audit of its transparency metrics in August 2020 and announced in May 2021 that the accounting firm Ernst & Young would undertake this audit. The results of that audit remain to be seen.

Red: ELECTION MISINFORMATION

- On January 8, 2021, SHFP called for social media platforms to remove Donald Trump permanently. Immediately following the insurrection, Facebook blocked Trump's Facebook and Instagram accounts "indefinitely and for at least two weeks.". Facebook then extended the ban indefinitely and assigned its Oversight Board to adjudicate over Trump's fate. The board upheld that suspension, but declined to determine its duration, instead kicking that decision back to the company. In earlyJune, the company suspended Trump's account until early January 2023 -- two years after the initial suspension and a couple of months after the November 2022 midterm elections. The coalition believes this is insufficient given Trump's history of repeatedly using hate speech, inciting violence, and spreading disinformation.

- SHFP demanded that Facebook help ensure accuracy in political and voting matters by eliminating the newsworthiness exemption for politicians, which was based on the argument that the speech of political leaders was valuable to the public even if it was abusive; removing misinformation related to voting; and prohibiting calls to violence by politicians in any format.
    - Facebook announced it was ending the newsworthiness exemption in June 2021 and confirmed the SHFP coalition's belief that the platform had accorded special treatment to public figures who engaged in spreading hate and inciting violence on the platform. We have yet to see Facebook enforce the policy change. Additionally Facebook has not changed its stance on not fact-checking political leaders.

- More broadly, a report from the online advocacy group Avaaz showed how Facebook failed to curb misinformation around the 2020 U.S. election, allowing the "Stop the Steal" conspiracy theory to proliferate before taking any action.

Orange: GROUPS AND RECOMMENDATIONS

- SHFP demanded that Facebook address issues regarding Facebook Groups. Specifically, the coalition asked Facebook to find and remove public and private groups tied to white supremacy, anti-government extremism, antisemitism, violent conspiracies, vaccine misinformation, and climate denialism. SHFP also called for Facebook to stop recommending or amplifying groups or content from groups associated with hate, misinformation, or conspiracies. Additionally, the coalition demanded Facebook create an internal mechanism to automatically flag hateful content in private groups for human review.
  - Facebook made two major updates to its Groups product in September 2020 and March 2021:
    - In the September 2020 update, Facebook announced that the platform would limit "the spread of [groups tied to violence] by removing them from recommendations, restricting them from search, and soon reducing their content in News Feed." While it previously had taken some action on "groups tied to violence" such as QAnon and militias, this was a significant change.
    - In the March 2021 update, Facebook announced it would remove all civic and political Facebook groups from platform recommendations. Rather than finding a thoughtful way to address the issue of hate and misinformation in Facebook Groups, this broad-brush solution is likely to harm those who use the Facebook Groups product for legitimate political and civic activities.
  - After both of these announcements, NBC News found evidence of members of U.S. special operations forces using private Facebook groups to spread hate, misinformation and conspiracy theories such as QAnon. While the investigation did not speak to the prevalence of these groups, its findings are concerning given the changes Facebook stated it implemented.

Red: SUPPORT FOR TARGETS OF HARASSMENT

- SHFP demanded that Facebook enable individuals facing severe hate and harassment to connect with a live Facebook employee. In no other sector does a company not have a way for victims of its product to seek help.
  - In 2017, Facebook announced new tools to connect individuals at risk of suicide with live chat support, indicating that providing live support to individuals in specific content areas on Facebook's platforms is possible. In 2019, experts questioned the efficacy of these efforts, stating the Facebook lacked "transparency and ethics" around its suicide prevention work.
  - To date, we are unaware of any efforts by Facebook to address this concern meaningfully.

Red: CONTENT MODERATION AROUND HATE, ANTISEMITISM AND RACISM

- SHFP demanded that Facebook create expert teams to review submissions of identity-based hate and harassment.
  - Replying to SHFP's demands last summer, Facebook stated that it "automatically sends hate speech reports to reviewers with specific training in identity-based

hate policies, in 50 regions covering 30 languages." It also said that content moderators undergo "a comprehensive training program that includes at least 80 hours of live instructor-led training, as well as hands-on practice for all of our reviewers."

- A 2019 Reuters investigation found that Facebook supports 111 languages, meaning that its team of reviewers with training in identity-based hate covers less than a third of the languages spoken on the platform.
- Seventy-five percent of those who experienced online harassment reported that at least some of the abuse occurred on Facebook. Much of this harassment is based on an individual's identity. According to ADL's 2021 Online Hate and Harassment survey, a third of Americans who reported harassment say they experienced identity-based harassment. LGBTQ+ Americans suffer higher rates of overall harassment by group, at 64%. Our survey also showed that 59% of Black Americans who reported identity-based harassment reported they were harassed online because of their race.
- In October 2020, the company finally changed its hate speech policy to ban content that denies or distorts the Holocaust after a decade of advocacy from ADL and the Jewish community, only to enforce its policy inadequately. ADL's Holocaust denial report card, issued on January 2021 gave Facebook a "D" for not taking down any of the explicit Holocaust denial posts ADL reported to the company.

Stated efforts by Facebook to curb hate and racism have been insufficient. Facebook's attempts to address hate have been band-aids over larger, more necessary changes the company should undertake.

Efforts by Other Social Media Platforms

Stop Hate for Profit's initial focus on Facebook prompted other social media platforms to address the proliferation of hate, racism, and misinformation on their platforms. The campaign also spurred businesses to pause their advertising on all social media platforms. In January 2021, Stop Hate for Profit demanded that Twitter, Alphabet (the parent company of Google and its subsidiaries, including YouTube), and other platforms remove then-president Donald Trump permanently. The following is an overview of changes these platforms have made since the launch of Stop Hate for Profit.

Orange: Twitter

During and after SHFP's launch, Twitter made numerous changes to address hate, racism, and misinformation on its platform. Most notably, the company changed its policies in July 2020 to take down posts with links to external websites that contain content violating Twitter's rules. This change led to the deplatforming of the infamous white supremacist David Duke. In addition, the platform took a robust approach to election integrity efforts. Despite these actions, CEO Jack Dorsey admitted that Twitter "played a role" in the attack on the U.S. Capitol on January 6. Twitter allowed support for QAnon to run rampant on its platform. While the platform took action on Donald Trump's accounts during the election cycle by labeling some of his posts as misinformation, it was only *after* the insurrection that Twitter deplatformed the former president.

More recently, ADL also noted the high levels of antisemitism on the platform during the recent conflict in the Middle East.

Additionally, to date, the company has not made substantive changes to help targets of wide-scale harassment on the platform, such as improving its reporting tools to allow users to flag more than five posts at a time. Moreover, as of this writing, Twitter does not allow users reporting hate or harassment to state whether they were targeted for their identity, or to express the specific identity they were targeted for (including their actual or perceived race, religion, gender identity, and sexual orientation). Without knowing details about how hate manifests and impacts users, it is difficult to manage it effectively.

Orange: YouTube

During and after SHFP's launch, YouTube took action in June 2020 against six prominent white supremacists, including Richard Spencer, Stefan Molyneux, and David Duke. However, YouTube did not implement any large-scale policy or product changes ahead of the 2020 election. After the election, instead of robustly moderating misinformation content, the platform provided the unclear label of "Results may not be final" to harmful misinformation.

In February 2021, an ADL study by Belfer Fellow Brendan Nyhan found that despite the platform claiming it changed its recommendation system in 2019, YouTube still recommended extremist and alternative content to users, especially those with already high levels of "racial resentment."

In April 2021, YouTube introduced the Violative View Rate, a metric for how often violative content is viewed on the platform. An April 2021 report from The Markup found that YouTube allowed hateful terminology such as "White Lives Matter" and "White Power" to be used in ad targeting on the platform, after which Google took action to block those keywords.

YouTube implemented some notable recommendations in December 2020, such as placing a warning ahead of posting when users attempt to comment with potentially violative language and allowing creators to provide YouTube users with their identity characteristics to better track hate and harassment against marginalized communities. Nevertheless, larger structural issues remain.

Orange: TikTok

In August 2020, not long after SHFP launched, TikTok released its first transparency numbers around hate speech. It also banned QAnon and various white nationalist ideologies from the platform as of October 2020. Most recently, in May, TikTok released product features to help targets of hate and harassment; in particular, it will allow targets to flag up to 100 comments at the same time for review, the largest amount of any platform to date. This is an important step and something SHFP specifically asked for in its product recommendations.

Even so, TikTok has had numerous incidents of antisemitism in recent months targeting Jewish creators and a 97-year-old Holocaust survivor. Recently, it was reported that over- enforcement measures targeted Jewish influencers. ADL has reported on this trend both back in 2020 and in June. Thus, while TikTok has made progress this year, the spate of antisemitism shows it still has work to do.

Yellow: Reddit

Reddit released its first cross-platform hate policy in June 2020, followed by its ban of the subreddit r/The_Donald and the release of a report on specific communities targeted by hate in August 2020. In January 2021, Reddit acted swiftly to deplatform r/DonaldTrump for spreading misinformation about the election and the January 6 insurrection. While Reddit had a high-profile antisemitic incident on r/Wallstreetbets, its handling of the abuse through platform and community moderation was admirable in the sense that both platform and community moderation efforts acted swiftly to remove antisemitic comments. At the same time, the r/Wallstreetbets community itself pushed back on antisemitic comments by downvoting them and engaging in counterspeech denouncing them.

Conclusion

Despite the events of the past year presenting grave threats to our democratic institutions and to the safety of marginalized communities, and despite the clear evidence of how widely hate and incitement are spread online, no platform has enacted the large structural changes needed to address the consequential issues raised by Stop Hate for Profit. Thus far, the platforms doing the most (namely, Tik Tok and Reddit) are much smaller than the world's biggest three social media companies: Facebook, Twitter, and YouTube. Content moderation at the massive scales of these companies necessitates the use of automated tools, but as we've seen, artificial intelligence is nowhere near capable of effectively stamping out online hate speech and disinformation.

Transparency remains frustratingly elusive at the three largest platforms. Their moderation operations are unclear and we still lack data on those harmed by violative content. These companies' slowness, even reticence, to act boldly and at the appropriate scale guarantees that hateful content, conspiracy theories, and misinformation will keep growing relatively unabated to the detriment of all.

Additional Resources

For more information on reducing hate, racism, and misinformation online, see ADL's REPAIR Plan, a six-part framework for lawmakers and technology companies. To learn about the experience of Americans navigating the internet, read our annual Online Hate and Harassment survey.