glaad

SOCIAL MEDIA SAFETY INDEX

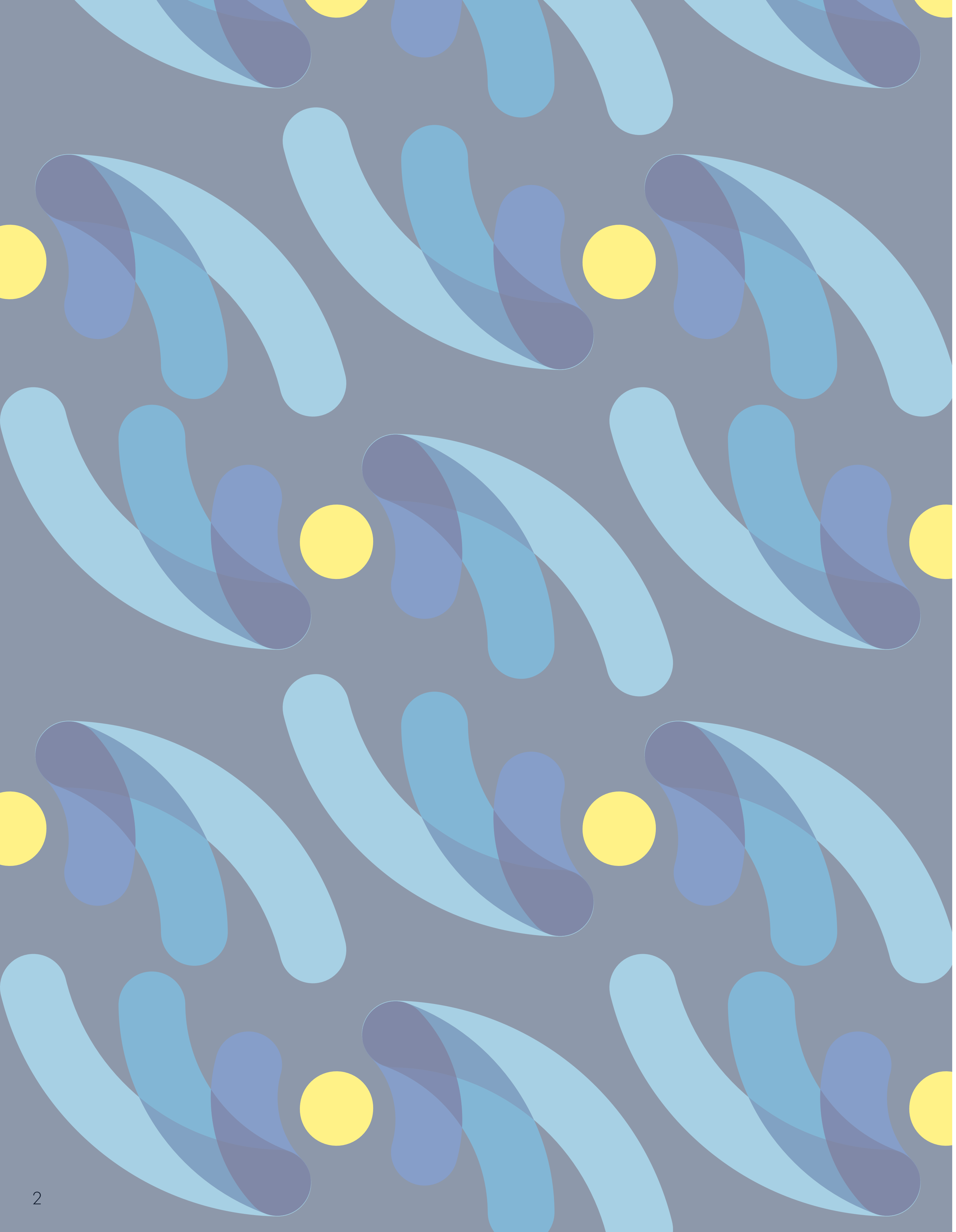


TABLE OF CONTENTS

LETTER FROM GLAAD PRESIDENT & CEO SARAH KATE ELLIS

At a time when talk of regulation around content and ads on social media is rapidly escalating and social media platforms consider the critical and urgent calls for transformation from other marginalized communities, the unique needs of LGBTQ people have largely been invisible or fall low on the priority list. *The GLAAD Media Institute's Social Media Safety Index* (SMSI) aims to change that by creating an annual form of industry accountability to the LGBTQ community. In addition to documenting current threats to LGBTQ safety, the *Social Media Safety Index* sets out a roadmap for change and marks the launch of a renewed commitment to ongoing advocacy across the industry.

For over 35 years, GLAAD has been the leader in creating safe and inclusive environments in Hollywood, journalism, and across our culture. Our founders were visionaries who understood that what people see and hear in the media affects the decisions made in schools, offices, living rooms, courtrooms and ballot boxes. Because of GLAAD's media work—and the work of so many content creators and media industry leaders—the world came to know LGBTQ people and to accept us. By ensuring LGBTQ people were included and represented in fair and accurate ways, GLAAD's work changed hearts and minds and LGBTQ acceptance grew.

We continue to innovate to keep step with a rapidly and ever-changing media landscape. GLAAD's advocacy and consulting have expanded into sports, political media, kids and family media, advertising, and video games. Three years ago, during a LGBTQ event held in Davos during the World Economic Forum's Annual Meeting, we launched the GLAAD Media Institute (GMI) with a grant from the Ariadne Getty Foundation to house all of this behind-the-scenes consulting work and to expand our thought leadership and research reports. For decades, our reports on television and film have leveraged research, created accountability, and generated ongoing advocacy which has increased LGBTQ representation in front of and behind the camera.

The SMSI is new research from the GLAAD Media Institute that follows in the successful tradition of our reports by manifesting tangible accountability and shining a light on the frequent disconnect between a platform's policies and the actual user experiences. We also hope the GMI can be a resource to help execute many of the recommendations in the Index and to hold ongoing constructive conversations with each platform throughout the year. GLAAD has already played a significant role in advocacy and consulting with social media. From successfully advocating Facebook to add gender options in 2014 for trans and gender non-conforming users (and to add 'in a civil union' in 2011 when such relationships were the only legal option for families like mine), to consulting with dating apps like Tinder to welcome trans users safely, to working behind the scenes to unblock safe and appropriate LGBTQ content on numerous platforms, GLAAD can be a vital resource for social media policy and engineering teams.

Over the past few years, the growth of violent speech and the spread of misinformation across social media has become one of the greatest barriers to full LGBTQ equality and acceptance. Taking leadership to assist social media companies in addressing this problem, we are also launching the GLAAD Listing of Anti-LGBTQ Online Hate Speech, a resource to assist platforms and social media users in mitigating hateful content and conduct. We have continually witnessed that LGBTQ-inclusive content policies and community standards do not align with the user experience. Further, those policies and standards vary tremendously across the platforms.

Our esteemed advisory committee members, along with the Gill Foundation and Craig Newmark Philanthropies, share these concerns and we are thankful for their critical leadership and guidance in the development of this inaugural Social Media Safety Index.

The simple overarching recommendation of this report is that decision makers and policy leads at social media platforms must act immediately to improve social media safety for LGBTQ people and for other historically marginalized groups. The safety of LGBTQ people on social media platforms is an urgent public safety issue. If we approach this issue using a lens of public health and public safety, it is clear that companies have an inherent responsibility to make their products not merely *safer* but actually truly *safe* — for LGBTQ users, and for everyone.

SARAH KATE ELLIS
President & CEO, GLAAD

EXECUTIVE SUMMARY

Recognizing the urgent need to push major social media platforms to make their products safer, this inaugural edition of the GLAAD Social Media Safety Index establishes an initial baseline exploration of the social media landscape for LGBTQ users. Our over-arching recommendation is that decision makers, product teams and policy leaders at social media platforms must urgently make their products safe — for LGBTQ people, and for other historically marginalized groups.

This report draws on extensive input from leaders at the intersection of tech and LGBTQ advocacy, as well as a broad literature review distilling other reports, articles, research and journalism; and a review of platform policies and analysis of how they match up (or don't match up) with actual LGBTQ user-experience.

Surveying the current landscape of leading social media platforms, the entire sector is effectively unsafe for LGBTQ users.

Of special concern, the prevalence and intensity of hate speech and harassment stands out as the most significant problem in urgent need of improvement. Problems include: inadequate content moderation, polarizing algorithms, and discriminatory AI which disproportionately impacts LGBTQ users and other marginalized communities who are uniquely vulnerable to hate and harassment and discrimination. This index identifies these and other problem areas and offers dozens of recommendations and urgings — both concrete and general.

LGBTQ hate speech and misinformation is a public health and safety issue. Some of the urgent recommendations across platforms include:

- Stop allowing algorithms to fuel extremism and hate. Similarly, confront the problem of bias in AI which disproportionately impacts LGBTQ people and other marginalized communities.
- Make it easier for users to report problematic content, be transparent in content moderation, and use more human moderators.
- Employ a dedicated LGBTQ policy lead.
- Respect data privacy, especially where LGBTQ people are vulnerable to serious harms and violence.
- Only select platforms currently take any kind of action on violent speech and misinformation, with tactics including monitoring trending topics for misinformation, restricting hashtags or shares, or having labels on misinformation, but when it comes to anti-LGBTQ misinformation, enforcement is arbitrary at best.

While there is a broad tangle of overlapping issues, requiring an array of approaches to mitigation and solutions, our research makes it clear that these companies can — and must — do better.

Upon release of the 2021 Social Media Safety Index report, GLAAD will offer briefings for each platform to review issues that LGBTQ users face, and to go over the recommendations described here. In future releases of the Social Media Safety Index, GLAAD looks to the platforms to provide updates on improvements, achievements, or progress on any and all LGBTQ safety measures or ways they are addressing the concerns of this report.

Such information will be documented in the next SMSI report and we are hopeful that each platform will implement meaningful changes to make their platforms safer for LGBTQ users. GLAAD also looks forward to sharing the general recommendations sections of the SMSI with other leading platforms, apps, and messaging programs. Through a series of presentations at conferences and events, GLAAD will continue to maintain an ongoing dialogue about LGBTQ platform safety amongst industry colleagues throughout 2021. GLAAD will also continue to spotlight new and existing safety issues facing LGBTQ users in real-time both directly to the platforms and to the press and public.

Instilling the ambition to improve their products, we plan to expand future annual editions of the SMSI to include a scorecard system, rating individual platforms on their performance.

Social media platforms, and tech companies in general, have come into existence so swiftly that corresponding public policy (and regulatory mechanisms) to understand the ramifications of their business-models has simply not kept up. There is no question that the impact of these platforms on our society is enormous.Industries will resist regulations that *increase* the cost of doing business or *decrease* their profits. There is nothing surprising or shocking about this. It is the nature of industry and for-profit business. Corporations are corporations and there is no point appealing to consciences that they do not have. Hence the inclusion of this final line in each section below:

In concluding our recommendations, we urge *every individual* in a position of leadership at these companies to find ways to take meaningful action now to make these platforms safe for their LGBTQ users.

GLAAD Social Media Safety Index Advisory Committee

Kara Swisher
Contributing writer and host of the 'Sway' podcast at The New York Times

Maria Ressa
Journalist & CEO, Rappler

Brandi Collins-Dexter
Senior Fellow, Color of Change & Visiting Fellow, Shorenstein Center

Liz Fong-Jones
Principal Developer Advocate for SRE & Observability, Honeycomb

Dr. Sarah T. Roberts
Co-Director, UCLA Center for Critical Internet Inquiry

Marlena Wisniak
Co-Director, Taraaz

Lucy Bernholz
Director, Digital Civil Society Lab at Stanford University

Leigh Honeywell
CEO and Co-Founder, Tall Poppy

Tom Rielly
Founder, TED Fellows program & Founder, PlanetOut.com

Jenni Olson
Co-Founder PlanetOut.com & Senior Project Consultant

Rich Ferraro
GLAAD Chief Communications Officer.

INTRODUCTION & METHODOLOGY

In preparing this report, GLAAD reviewed thought leadership, reports, journalism, and findings across the field of social media safety—as well as consulting with our GLAAD SMSI advisory committee and many other organizations and leaders in technology and social justice. As reflected in our *SMSI Articles & Reports Appendix*[1], there are constant ongoing developments regarding the real-world impact of social media platforms on individual user safety and on public health and safety as a whole.

This report begins with a selection of broad recommendations, relevant to all platforms, tackling such realms as: LGBTQ self-expression; privacy and outing; LGBTQ hiring, inclusion and leadership; civil discourse around LGBTQ issues; content moderation; mitigating anti-LGBTQ hate; disinformation and misinformation; transparency and accountability; algorithms and AI; and more.

Our general recommendations are followed by a series of platform-specific recommendations for Facebook, Twitter, YouTube, Instagram, and TikTok respectively. Even as other approaches are currently being put forth (such as a regulatory strategy, the current antitrust suits, etc. — see sidebars) the SMSI emphasizes what these social media companies can and must do themselves right now to address these problems[2].

Thankfully, social media platforms are making adjustments and improvements to their products every day. Since the aspiration of this report is for as many of these recommendations as possible to be implemented, it will be counted as an achievement if any of the items below are outdated as we go to press (as of March 31, 2021). See for example, Facebook's March 31, 2021 new features announcement which includes enabling users to limit who can comment on their posts and making it easier for users to adjust the algorithm of the news feed.

GLAAD'S Role

GLAAD has a long history of consulting directly with apps and social media platforms on some of their most significant LGBTQ policy and product updates. In addition to early involvement with YouTube's Trusted Flagger program, GLAAD is also a current member of Twitter's Trust & Safety Council and Facebook's Network of Support, an advisory coalition it helped create in 2010.

GLAAD's recommendations here are focused on these five leading social media companies: Facebook, Twitter, YouTube, (Facebook-owned) Instagram, and TikTok. We direct our recommendations to these platforms, and urge that other social media companies review these guidelines as well. The Index includes overall observations for each of these five platforms, along with guidelines for improvement, in the hope that our ongoing annual evaluations will assist these companies in making their platforms safe for LGBTQ customers and constituents—and for everyone.

[1] SMSI — Articles & Reports Appendix spreadsheet

[2] Even setting aside an ethical perspective, platforms might consider a business argument for making these improvements. As one astute November 2018 Business Insider article points out: "users feeling unsafe on Instagram would spell bad news for brands advertising on the platform—48% of respondents said the degree of safety they feel on a given platform is either very or extremely impactful on their decision to interact with ads and sponsored content."

Online Hate

In our exploration of the social media landscape for LGBTQ people, the prevalence and intensity of hate speech and harassment stands out as the most urgent problem.

According to the ADL, in a 2021 report on online hate and harassment:

> LGBTQ+ respondents in particular continued to report disproportionately higher rates of harassment than all other identity groups at **64%** [compared to 41% for the general population overall], no significant change from the 65% in the previous 2020 survey [...] As was the case with overall reports of harassment, LGBTQ+ respondents, at **52**%, experienced far higher rates of severe harassment than all other groups[3].

Describing these figures as "dismayingly high," the ADL report further specifies the online locations of these hate and harassment incidents:

> Facebook, the largest social media platform in the world, was implicated in the highest percentage of online harassment reports, with three-quarters (75%) of those who experienced online harassment reporting that at least some of that harassment occurred on Facebook. Smaller shares experienced harassment or hate on Twitter (24%), YouTube (21%), Instagram (24%) [...] [and TikTok (9%)].

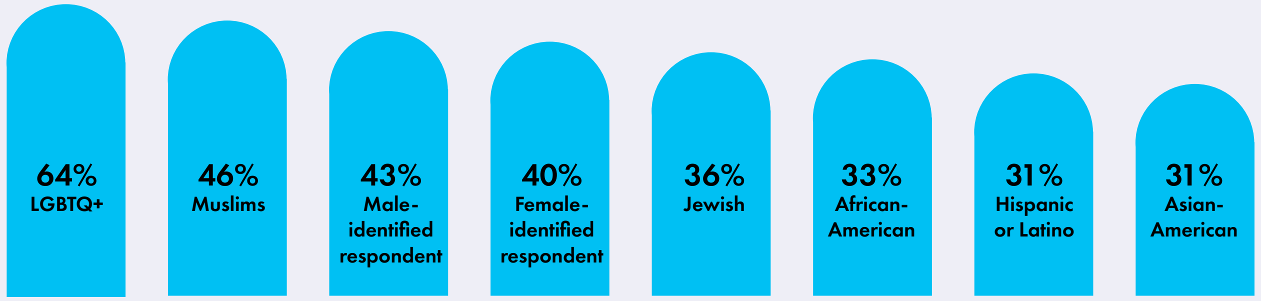
These figures reflect a disturbing reality for LGBTQ and other social media users. As the report forcefully concludes: "Encountering hate and harassment online has become a common experience for millions of Americans—and that experience does not appear to be getting safer. Technology companies are not resourcing to handle the magnitude of the problem, regardless of what their public-facing statements say to the contrary."

These alarming findings, and the pervasive anti-LGBTQ content and conduct documented in this report, led us to create the new GLAAD Listing of Anti-LGBTQ Online Hate Speech.

Online Hate and Harassment

Demographics of Harassment

(Total harassment experienced by group)



Source: ADL, Online Hate and Harassment Report: The American Experience 2021

[3] "Online Hate and Harassment Report: The American Experience 2021." ("A survey of 2,251 individuals was conducted on behalf of ADL by YouGov, a leading public opinion and data analytics firm, examining Americans' experiences with and views of online hate and harassment [...] Surveys were conducted from January 7, 2021, to January 15, 2021.")

ABOUT HATE

"We can disagree and still love each other unless your disagreement is rooted in my oppression and denial of my humanity and right to exist."

— Robert Jones, Jr. (@SonOfBaldwin)

All of our advisory committee members pointed to the overarching **category of hate speech as the single most important aspect of social media safety for LGBTQ people**. The bulk of this report is correspondingly devoted to the topic, and this has also prompted a new initiative: the GLAAD Listing of Anti-LGBTQ Online Hate Speech.

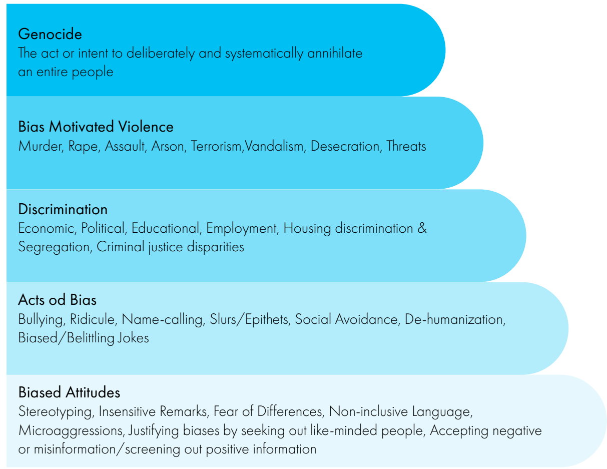
From an international human rights perspective, hate speech is defined in the 2019 *UN Strategy and Plan on Hate Speech* as communication that: "attacks or uses pejorative or discriminatory language with reference to a person or a group on the basis of who they are, in other words, based on their religion, ethnicity, nationality, race, color, descent, gender, or other *identity factor*."[4]

We use the term "Hate Speech" here to mean not just conventional words and language but also the interconnected array of conduct, behaviors, actions, and tactics that have come to be commonly weaponized against LGBTQ people in the social media landscape. This includes, but is not limited to, things like trolling, keyword squatting, impersonation (fake profiles, troll armies, coordinated inauthentic behavior, etc.), doxing, viral sloganeering, and memes.[5]

A January 2021 Pew Research survey reports the alarming statistic that 68% of LGB adults had encountered online hate and harassment (please see sidebar for more details).

The State of Online Harassment, a Pew Research report released in January 2021, shares that: "Lesbian, gay or bisexual [LGB] adults are particularly likely to face harassment online. Roughly seven-in-ten [68%] have encountered any harassment online and fully 51% have been targeted for more severe forms of online abuse. By comparison, about four-in-ten straight adults [41%] have endured any form of harassment online, and only 23% have undergone any of the more severe behaviors." The report adds that, "Fully 91% of Americans say people being harassed or bullied online is a problem, including 55% who describe this as a major problem[...]. As online harassment permeates social media, the public is highly critical of the way these companies are tackling the issue. Fully 79% say social media companies are doing an only fair or poor job at addressing online harassment or bullying on their platforms."

These expressions of hate, both online and off, are violent, dangerous, and harmful. There are very real harms and impacts on LGBTQ people, and on society as a whole, arising especially from the anti-LGBTQ hate speech, conduct, and content—including both disinformation and misinformation—that abound on social media platforms.[6]



Genocide
The act or intent to deliberately and systematically annihilate an entire people

Bias Motivated Violence
Murder, Rape, Assault, Arson, Terrorism,Vandalism, Desecration, Threats

Discrimination
Economic, Political, Educational, Employment, Housing discrimination & Segregation, Criminal justice disparities

Acts od Bias
Bullying, Ridicule, Name-calling, Slurs/Epithets, Social Avoidance, De-humanization, Biased/Belittling Jokes

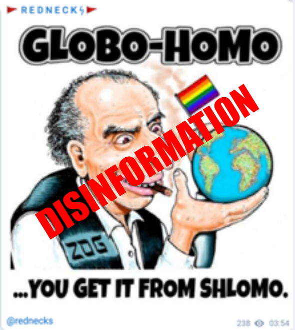
Biased Attitudes
Stereotyping, Insensitive Remarks, Fear of Differences, Non-inclusive Language, Microaggressions, Justifying biases by seeking out like-minded people, Accepting negative or misinformation/screening out positive information

© 2018 Anti-Defamation League

One of the most powerful representations of the mechanics of hate speech, and the real-world dangers and harms that can arise from these kinds of biased attitudes and behaviors, is the ADL's Pyramid of Hate, pictured above.

Words as weapons

To be clear, when we talk about online hate speech, we are not referring to the occasional epithet or even garden-variety homophobia—though that is also certainly part of it. Have you ever heard the term *Globohomo? CloverGender? LGBTP? Transvestigations?* These are just a few of the many creative examples of new, dangerous, weaponized anti-LGBTQ content that circulate freely and widely across social media platforms causing online and offline harms for LGBTQ people and contribute to an overall atmosphere of disrespect, dehumanization, hate and violence (see the new *GLAAD Listing of Anti-LGBTQ Online Hate Speech for definitions and contextual examples*).



This right-wing troll invention is a uniquely vicious combination of homophobia and anti-Semitic conspiracy theory (this specific example was posted on Instagram and promotes a vicious extremist Telegram account called Rednecks).

Discussing their 2020 report about the very serious real-world impacts of the massive volume of right-wing anti-trans content on Facebook, Media Matters observes that:

> Harmful narratives divert attention from important issues facing the community such as employment discrimination and high rates of violence. And when trans youth and their families use these platforms, they are fed a stream of disinformation that could result in parents denying their children critical care or rejecting their identities, which can harm trans kids' physical and mental well-being.

While Facebook, Twitter, YouTube, Instagram, TikTok and others must balance concerns around free expression, it cannot be stated strongly enough that social media platforms must take substantive, meaningful, and far more aggressive action to prioritize the safety of their LGBTQ users and to staunch the epidemic of hate and extremism. These efforts must also address the spread of disinformation (the intentional spreading of false or misleading information) as well as misinformation (misleading information believed to be true but not necessarily intending to cause harm).

Strategies to Mitigate Hate

There are many ways for platforms to curb anti-LGBTQ conduct and content, including adding context or links (in the same way that platforms add an official voter information link to any posts that include the word "vote" or "election"); removing content; demonetizing or suspending accounts (some platforms apply a "three strikes and you're out" policy); and banning/de-platforming (individuals or organizations will not be allowed to create new accounts or pages on a given platform).

There are also numerous strategies—like speed-bumps or circuit breakers that throttle viral content—that have been used effectively to *slow* the spread of misinformation, including anti-LGBTQ hateful content. These particular strategies have been developed mainly in relation to public health issues, especially around the Coronavirus and vaccine. It is worth noting that if policy makers were to reframe the discussion of online dis/misinformation and hate speech as a public safety/public health issue, then social media platforms might be held to a higher standard.[7]

The Issue Of Censorship

On the one hand, LGBTQ individuals are *vulnerable to hate speech* and other manifestations of online homophobia, biphobia and transphobia—acts which have very real offline impacts and harms. On the other hand, we are also *vulnerable to censorship* and disproportionate limitations of free expression related to our identities.

The most succinct explanation of this vulnerability is that because our sexuality is a defining aspect of LGBTQ identity there are greater opportunities for these characteristics to be flagged. It is also the case that actual homophobia, biphobia and transphobia can come into play on the part of AI and human content moderators and result in disproportionate suppression of LGBTQ expression.

In addition to being yet another thread in a social fabric of marginalization, bias, and oppression, these examples of bias (whether in human content moderators or in AI systems) create real harms and obstructions for LGBTQ people—including impacting our right to freely organize online, to access information, and to exercise our economic, social and cultural rights.

[4] The UN Strategy and Plan further adds that: "Rather than prohibiting hate speech as such, international law prohibits the incitement to discrimination, hostility and violence (referred to here as 'incitement')." These perspectives on hate speech build on Article 20, paragraph 2 of the International Covenant on Civil and Political Rights (ICCPR), which states that: "any advocacy of national, racial or religious hatred that constitutes incitement to discrimination, hostility or violence shall be prohibited by law."

[5] These are just a few examples. For many more, and for concise definitions of all these scary things, please see the impressive Media Manipulation Casebook list created by the Shorenstein Center on Media, Politics and Public Policy.

[6] Some deeper perspectives on hate speech include the Social Science Research Council's "Hate-Speech Intensity Scale" and the work of the Dangerous Speech Project.

[7] Amnesty International's 2018 Toxic Twitter" report on online violence against women offers a broad set of recommendations which could also be applied to anti-LGBTQ online hate.

RECOMMENDATIONS FOR ALL PLATFORMS

> *"These companies need to internalize the costs of effectively moderating their platforms and stop externalizing these costs onto the bodies and lives of vulnerable people and groups."*
>
> — Leigh Honeywell (Founder, Tall Poppy & GLAAD SMSI Advisory Committee member)

About Our Recommendations

In addition to these recommendations for *all* platforms, this report offers a variety of specific recommendations below for the five major platforms. Facebook, Twitter, YouTube, Instagram, and TikTok.

Subsequent annual reports will offer ratings of platform progress in the specific areas that we outline.

GLAAD Social Media Safety Index — Platform Responsibility Checklist

- Protection of LGBTQ users in community guidelines
- Mitigating algorithmic bias and bias in artificial intelligence (AI)
- Addressing privacy and outing (including data-privacy & micro-targeting)
- Promotion of civil discourse
- Overall transparency and accountability
- Content moderation (and multiple related and overlapping areas including hate speech and misinformation mitigation, enforcement, transparency and accountability, user-reporting systems, self-expression, etc.)
- LGBTQ hiring, inclusion and leadership
- Engagement of independent researchers and social scientists
- Engagement of affected users/communities, especially underrepresented groups
- Innovation
- Corporate responsibility

These very broad general categories are but a few of the top-level concerns that social media platforms must address in making their products safe for LGBTQ users. For a much more thorough taxonomy please see the 2020 Indicators list produced by Ranking Digital Rights.

Improve Protections of LGBTQ Users in Community Guidelines & in Hate Speech Definitions

Platforms should expand AI flagging to incorporate words and phrases that have been identified as anti-LGBTQ hate speech by leading NGOs, human rights groups and other specialists in the field (and attention should be given to terms and phrases in multiple languages/dialects other than English). See the new *GLAAD Listing of Anti-LGBTQ Online Hate Speech* for a list of terms and phrases to be added. GLAAD also urges all platforms to follow the lead of Twitter's Policy on Hateful Conduct, which includes a specific prohibition against misgendering and deadnaming: "We prohibit targeting individuals with repeated slurs, tropes or other content that intends to dehumanize, degrade or reinforce negative or harmful stereotypes about a protected category. *This includes targeted misgendering or deadnaming of transgender individuals.*" (Misgendering is referring to a transgender person with the wrong gender. Deadnaming is referring to a trans person by a former name, usually one assigned to them prior to transitioning, without their consent).

Elevate Legitimate Voices

Given the extraordinary amount of ingenuity and resources social media companies have deployed in creating sophisticated algorithms so successfully focused on maximizing revenue, it is reasonable to suggest that this same brilliance should be applied to product safety for LGBTQ people, and for everyone. Platforms should especially **implement tools and policies to efficiently elevate legitimate and diverse voices and to moderate extremists and reduce anti-LGBTQ misinformation, hate and threats of violence**. In fact, Facebook has already implemented such simple and successful public safety measures designed to combat political misinformation and hate speech. But, as the *New York Times* reports, Facebook executives rolled back such measures after the 2020 election, as they have done repeatedly in the past, "either because they hurt Facebook's usage numbers or because executives feared they would disproportionately harm right-wing publishers." In February 2021, Wikimedia announced a new platform policy that promises to make their products safer for LGBTQ users, and for everyone. The Universal Code of Conduct for Wikipedia: "creates binding standards to elevate conduct on the Wikimedia projects, and empower our communities to address harassment and negative behavior across the Wikimedia movement." GLAAD echoes the recommendations outlined in the 2020 **Stop Hate For Profit** initiative ("a diverse and growing coalition that wants social media companies to take common-sense steps to address the rampant racism, disinformation and hate on its platform.") Ditto the 2018 **Change The Terms** report (which focuses on "recommended corporate policies and terms of service to ensure that social media platforms, payment service providers, and other internet-based services are not places where hateful activities and extremism can grow.") GLAAD also supports the recommendations of the 2020 **Ranking Digital Rights Corporate Accountability Index**.

Be Accountable & Transparent

Platforms should achieve **accountability and transparency** across all levels. This includes undergoing regular independent audits, providing researchers open access to data, and working with relevant stakeholders in creating platform policies (GLAAD has consulted with platforms over the years with regard to LGBTQ-related policies and product updates and will continue to do so). There are many other reports and campaigns calling for these fundamental commonsense basics — including the *Ranking Digital Rights Corporate Accountability Index, Access Now's 26 Recommendations on Content Governance*" and the GMF's *Safeguarding Digital Democracy: Digital Innovation and Democracy Initiative Roadmap*. As the *2020 Mozilla Internet Health Report* summarizes: "With increased transparency about the algorithms, governance, and community dynamics of large platforms, a broader set of stakeholders can engage in more fruitful conversations about strategies for the future."

Public Health & Safety Must Guide Product Design & Policy Decisions

That's It. That's the Tweet.

Innovate!

The vision and creativity of platform engineers could surely yield brilliant and exciting new tools, systems, and designs to make their products safer and improve the experience for their LGBTQ users, advertisers, and *all* stakeholders—especially mitigating hateful content and conduct while also balancing concerns such as privacy and free expression (issues which are also of special relevance for LGBTQ people who are often disproportionally impacted by restrictions on our own use of language). Platforms (and policy makers) can look to projects like the MIT Media Lab's Cortico AI which explores the concept of measuring "conversational health" via four indicators: shared attention, shared reality, variety of opinion, and receptivity (Twitter worked with Cortico in March 2018) or the MIT Media Lab's new Center for Constructive Communication, which looks to "better understand current social and mass media ecosystems and design new tools and communication networks capable of bridging social, cultural, and political divides." Or Civic Signals, the ambitious and exciting new two-year research project from NewPublic.org, which explores how to build better digital public spaces ("A flourishing digital public space should be welcoming and safe for diverse publics, help us understand and make sense of the world, connect people near and far across divides and hierarchies, and enable us to act together.") There's also the early 2021 announcement of the Digital Trust & Safety Partnership, to develop an industry framework for handling harmful online content and conduct. Another promising initiative is the September 2020 collaboration between the major platforms and the World Federation of Advertisers (WFA) to adopt a common framework for defining harmful content and create "the first global brand safety and sustainability framework for the advertising industry." Let's also look at thought leadership like Nobel-winning economist Paul Romer's idea of a levy on targeted ad revenue, or Siva Vaidhyanathan's ambitious array of lenses in his January 2021 overview for *The New Republic*. Vaidhyanathan touches on a variety of perspectives ranging from regulatory and antitrust approaches; the community-reliant moderation model of Wikipedia (others point as well to Reddit); a reconsideration of public-service media; a tax on data collection; European models of oversight such as the EU Digital Services Act and Digital Markets Act (see sidebar). These are just a few of the many, many visions and opportunities for innovation and change.

Make it Easier for Users to Report Problematic Content

There are many opportunities for platforms to find ways to encourage users to report problematic content, and to make the process more transparent. Note that GLAAD's new *Listing of Anti-LGBTQ Online Hate Speech* includes a resource section on "How to Report LGBTQ Online Hate Speech," which features links to relevant reporting guidelines for the major social media platforms. The ADL Cyber-Safety Action Guide summarizes the hate speech policies of social media platforms and other top websites and includes links for reporting such speech. The World Health Organization also offers a helpful page, "How to report misinformation online." And this June 2020 article from *PC Magazine*, "How to Report Abuse on Social Media" features an illustrated guide to reporting things on Facebook, Twitter and Instagram.

Employ a Dedicated LGBTQ Policy Lead

There should be a dedicated LGBTQ policy lead at each platform, to drive ongoing platform research work shaping these policies and to liaise with GLAAD and other nonprofits and NGOs in the field. This point of contact can then present such platform-driven research to LGBTQ organizations and experts for input and perspective. In the meantime, GLAAD will be monitoring each platform and leveraging our network of contacts, including the SMSI advisory committee, to continue to evaluate platform performance annually.

Stop Allowing Algorithms to Fuel Extremism and Hate

Improving safety for LGBTQ users is a complex challenge and algorithms are a key component of the battle. Platforms must **fundamentally change the ways that algorithms work to prioritize content according to criteria other than the maximization of ad revenue**. Currently, social media algorithms tend to push people further into silos of experience, sending them ever deeper into echo chambers of racism, anti-Semitism, Islamophobia, sexism, homophobia/transphobia, xenophobia and hate—a phenomenon that is well-documented by both researchers and journalists.

A damning internal Facebook report from way back in 2016 included the astounding statistic that: "64% of all extremist group joins are due to our recommendation tools [...]. Our recommendation systems grow the problem." (In March 2021 Facebook announced their intent to penalize Facebook Groups that violate their community standards.) Further on the topic of algorithms, note that Facebook's own Civil Rights Audit from July 2020 explains that the algorithms used by Facebook:

> fuel extreme and polarizing content [...]. Facebook should do everything in its power to prevent its tools and algorithms from driving people toward self-reinforcing echo chambers of extremism, and that the company must recognize that failure to do so can have dangerous (and life-threatening) real-world consequences.

Facebook's January 2021 promise to ban political Facebook Groups was met with understandable skepticism. In a January 2021 article for *Politico*, Elena Schneider and Cristiano Lima cite additional concerns for possible unanticipated impacts on progressive and social justice organizations, including those of Evan Greer from digital rights group Fight for the Future, who says:

> The decision about what is or isn't political is a very political decision in and of itself [...]. Will they consider a local veterans group to be political? If so, will they not consider a local anti-war group to be political? Would they consider an LGBTQ support group to be political? Frankly, all of those things are political.

Use "Friction" to Slow the Spread of Hate

Along with redesigning algorithms, researchers have advocated for the benefits of **introducing "friction"** in the user experience as a way to slow the spread of mis/disinformation (as well as extremism and hate, including anti-LGBTQ content). Examples include the introduction of viral circuit-breakers, fact-check panels, labeling of posts, scan and suggest technology, limiting auto-play of videos, etc. An August 2020 Center for American Progress report, *Fighting Coronavirus Misinformation and Disinformation: Preventive Product Recommendations for Social Media Platforms*, offers an excellent appendix of such recommendations. The day after the 2020 election, the *New York Times* ran an article with the headline: "On Election Day, Facebook and Twitter Did Better by Making Their Products Worse." As reporter Kevin Roose explains:

> For months, nearly every step these companies have taken to safeguard the election has involved slowing down, shutting off or otherwise hampering core parts of their products—in effect, *defending democracy by making their apps worse*." [emphasis added]

Indeed, as is true of so many healthier choices in life, the solution sounds inconvenient—"worse." For its part, Facebook implemented a "virality circuit-breaker" and generally added more "friction" to slow down the spread of viral posts so fact-checkers could verify claims or add warning labels. The platform also shut off the recommendation algorithms for some private groups and restricted certain hashtags. Facebook-owned Instagram also restricted hashtags. Twitter made similar changes, including monitoring their trending topics, restricting hashtags, disabling sharing features on tweets labeled as misinformation, introducing a user alert suggesting that one might want to actually read the content of an article before sharing it, and defaulting the act of re-tweeting to make it a two-step process (aka a "timeout"). These solutions can be repurposed in numerous ways to combat anti-LGBTQ hate speech and misinformation.

Confront the Problem of Bias in Algorithms & Artificial Intelligence (AI)

The multitude of harms wrought by Artificial Intelligence (AI) and algorithmic biases continues to disproportionately impact historically marginalized individuals and communities—including LGBTQ people. GLAAD urges all platforms to **devote resources to remedying the very serious and well-documented problem of AI bias and algorithmic bias** in their products.[8] In February 2021, journalists at *The Markup* discovered an alarming example of this bias in which: "Companies trying to run ads on YouTube or elsewhere

[8] From the AI Now April 2019 report, *Discriminating Systems: Gender, Race, and Power in AI: "Both within the spaces where AI* is being created, and in the logic of how AI systems are designed, the costs of bias, harassment, and discrimination are borne by the same people: gender minorities, people of color, and other under-represented groups. Similarly, the benefits of such systems, from profit to efficiency, accrue primarily to those already in positions of power, who again tend to be white, educated, and male [we would also add to this: straight and cis — among other categories]."

on the web could direct Google not to show those ads to people of 'unknown gender'—meaning people who have not identified themselves to Google as 'male' or 'female.'" What this meant was that "Google's advertising system allowed employers or landlords to discriminate against nonbinary and some transgender people." Alerted to the issue, Google promised to fix the problem. Another example of algorithmic bias comes from TikTok, which implemented a policy in 2019 where, in an effort to supposedly reduce bullying, the platform decided to tag certain accounts (chiefly people with disabilities and LGBTQ people) as vulnerable to bullying —and then proceeded to reduce the viral circulation of their posts. This algorithmic policy may have reduced users' exposure to bullying, but it unjustly suppressed their accounts. The company stated that it ceased employing this strategy as of December 2019. In a headline-grabbing 2017 example of anti-LGBTQ AI bias, Google's Cloud Natural Language enterprise software, "ended up having a considerably negative reaction to words and phrases that are about homosexuality. For example, the AI rated the phrase 'I'm straight' a 0.1 and the phrase 'I'm homosexual,' a -0.4." At that time, Google's parent-company, Alphabet, enlisted the assistance of GLAAD to help train the AI to make it less homophobic. Safiya Umoja Noble's groundbreaking book, *Algorithms of Oppression*, is a must-read on this topic, especially illuminating the breadth of racism and sexism embedded in algorithms and AI.

Stop Demonetizing LGBTQ Content in Ad Services

Social media companies have a history of blocking and/or demonetizing legitimate LGBTQ content in the realm of ad services. According to the *Advocate*, a September 2019 survey found that 73 percent of articles served by online ad services from LGBTQ news sites were getting blacklisted for advertisers—meaning that LGBTQ media outlets were unfairly unable to earn ad revenue. Platforms must **implement ongoing transparent research efforts to identify and address these kinds of problems in ad services** (including providing transparent documentation of these processes). Also, see sidebar on LGBTQ user account demonetization.

Use More Human Moderators

Platforms have implemented a variety of AI strategies to reduce the posting and spread of anti-LGBTQ hate speech, extremist rhetoric, and dis/misinformation. Much more **content moderation needs to be done by actual human moderators to successfully address anti-LGBTQ content and all forms of hate speech**. Platforms should also provide transparency on how human moderators are trained to detect online abuse against LGBTQ users (if such training exists at all). Not only are AI solutions flawed and limited, bad actors have learned how to game these systems. While AI is a valuable tool, it is not the singular solution. There is a need for human moderation — as well as a corresponding need for ethical and responsible employment practices in relation to these workers (see also Mary L. Gray and Siddharth Suri's *Ghost Work*).

Be Transparent in Content Moderation

Hand in hand with the above, **accountability and transparency in content moderation** are two of the most important needs. This includes visibility into the reporting process and effective consequences for violations. When reporting anti-LGBTQ and other kinds of hate speech or content that violates the platform's community standards, the reporting user should experience as much transparency as possible from the platform — including messaging that the report was received and how it is being responded to (conversely, if the user is being punished, they should be told why and how, and be given as much detail as possible to understand, including being given a transparent and timely process for appeal). On the flip-side of this topic, one of the most disturbing types of anti-LGBTQ conduct on social media is the well-documented practice of trolls reporting legitimate LGBTQ users in an effort to have their accounts de-platformed—with no reason conveyed to the user. See, for example, this *Los Angeles Blade* story about the case of Rosalynne Montoya, a Latina trans woman whose TikTok account was taken down after being reported by trolls, though it had not actually violated any guidelines (and which has subsequently been restored). Montoya's Change.org petition to "Change TikTok's Community Guidelines Algorithm" had more than 17,000 signatures as of mid-March 2021. Platforms should provide greater transparency on how decisions are made and what recourse users have for swift appeal and account restoration. For a more detailed articulation of suggested best practices see the widely known Santa Clara Principles On Transparency and Accountability in Content Moderation. Also see Amnesty International's *Twitter Scorecard* recommendations.

Make the Consequences Count

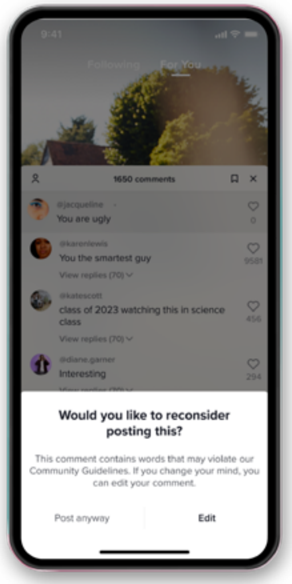
With regard to **enforcement**, consequences for violation must be effective. Repeat bad actors should be more effectively punished to genuinely and effectively halt the dissemination of hate and misinformation. Actual threats of violence must be swiftly identified and addressed and users who report content should have a closed-loop system where they're given the opportunity to provide feedback on the reporting process itself.

Make More Effective Use of Community Guidelines

Additionally, in the category of **enforcement**, platforms should take up the practice of **applying community guidelines across multiple categories of potential violation** when a piece of content is reported. For example, if a post promoting the discredited practice of "conversion therapy" is reported as hate speech, it should also be reviewed in the category of "medical misinformation," where it can be clearly identified and removed.

Apply Lessons Learned from Other Scripts & Algorithms

Some social media sites have trained their systems to recognize certain key phrasings or words as indicators that a user may be contemplating suicide. A script then proactively offers resources and messaging to provide help. This kind of system could also recognize radicalizing language or hate speech and direct people away from these antisocial behaviors while offering mitigating content or messaging. In fact, Instagram introduced a "Comment Warning" system in July 2019 ("Are you sure you want to post this? [...] We're asking people to rethink comments that seem similar to others that have been reported.") Twitter began testing a similar system in May 2020 ("Want to revise this? Language like this could lead to someone reporting your reply. But you can change it before sending.") Facebook's version appeared shortly thereafter ("Your comment may go against our community standards. It looks similar to others that we removed for bullying or harassment.") YouTube introduced a feature like this in December 2020 ("To encourage respectful conversations on YouTube, we're launching a new feature that will warn users when their comment may be offensive to others, giving them the option to reflect before posting.") And TikTok introduced a new feature in March 2021 which appears to be pursuing this strategy as well (see image below). As explained on the TikTok newsroom page: "A new comment prompt now asks people to reconsider posting a comment that may be inappropriate or unkind. It also reminds users about our Community Guidelines and allows them to edit their comments before sharing." Note that attention should be given to terms and phrases in multiple languages/dialects other than English. And of course we also urge platforms to ensure these kinds of features do not discriminate against or burden LGBTQ users and other marginalized communities who are disproportionately impacted by the widespread and well-documented phenomenon of AI bias.



Would you like to reconsider posting this?

This comment contains words that may violate our Community Guidelines. If you change your mind, you can edit your comment.

Illustration of TikTok's "comment prompt" feature introduced in March 2021. Image source: TikTok.

Respect Data Privacy

Data privacy, and the lack thereof, has many very real impacts on individual user safety. In the case of LGBTQ people **it is essential that users have transparent control over choices of how their user data is used by platforms**. The sexuality or gender identity of an individual user is one of many pieces of private information. Users should be able to decide (in an easy, transparent way) whether they want to share personal information with platforms or not. Users should never experience micro-targeted ads or be subject to data-driven user-history algorithms unless they proactively opt-in to them. The array of additional unique concerns confronting LGBTQ social media users around the world must be prioritized and addressed by platforms, especially in countries where LGBTQ people are vulnerable to serious harms and violence for their sexual orientation or gender identity.

Source: www.lovehasnolabels.com

Take Leadership in Civil Discourse

Facebook, Twitter, YouTube, Instagram, TikTok and other platforms are uniquely positioned to **serve as sources of information and education on civil discourse and LGBTQ allyship** for all of their users. GLAAD urges all social media platforms to take leadership in this regard—whether by creating such PSA campaigns themselves or in partnership with LGBTQ organizations or leaders, or providing *pro bono* promotion or exposure for non-profits like the Ad Council (and their acclaimed "Love Has No Labels" PSA campaign, with its celebratory message of diversity and inclusion which actively addresses bias, discrimination, and hate). A shout-out to TikTok for their 2020 "Be Informed" media literacy PSA series in partnership with the National Association for Media Literacy Education (NAMLE). Though platforms are not (yet) beholden to regulations requiring them to put forth such messages, the proactive assertion of these values could greatly improve the product experience for LGBTQ users, content creators, advertisers—and everyone.

Rely on Independent Research

All platforms should **use independent researchers and social scientists (especially LGBTQ researchers and social scientists)** to explore what is happening on their platforms (anti-LGBTQ content and conduct and hate speech, radicalization, misinformation) and to look for ways to off-ramp bad actors and expeditiously mitigate their harms—as well as educating users and contributing to constructive engagement around civil society. Facebook's announced research initiative on the 2020 Election is one potentially promising example of this.

Remain Diligent & Committed to LGBTQ User Safety

True dedication to the safety of LGBTQ users is an ongoing process. Companies must make **ongoing commitments to diligently and effectively seek to prevent harms and address all threats to LGBTQ safety on their platforms as they evolve over time**. Their responses to the multitude of these issues should be ongoing and adaptive, proactive and responsive. Just as we are now experiencing the impacts of swiftly-developing technologies and playing catch-up on the public health and safety consequences that social media has for our society, there will continue to be new and complex challenges and choices, especially as social media data interacts with physical data (addresses, CCTV data, and transport data) in ways that further blur the distinctions between online and offline. The need to continually adapt is critical.

Be Ethical. Be Responsible

As is true of Big Tobacco before it, **Big Tech must arrive at ethical and responsible business practices**. Social media platforms object that it is unreasonable to expect them to make changes to their algorithms in ways that reduce revenue; and that it is burdensome to moderate content, and to provide transparency, and to be truly accountable for the impact their products have on society. There are any number of corporate responsibility equivalents that can be cited to point out the speciousness of this resistance. The redesign of cars to include technologically advanced seat belts and to mitigate exhaust impacts and improve fuel efficiency presented enormous costs and hassles for automobile manufacturers; adding warning labels to cigarettes certainly had a huge negative impact on profits for tobacco companies; and halting the practice of simply dumping toxic waste into our rivers and public waterways cut into the business models of corporations and industries of all kinds. But because public health is at stake, society and policy makers have agreed that companies should bear at least some of these expenses as part of the cost of doing business. It would be valuable for platforms to review the *UN Guiding Principles on Business and Human Rights*, (which Facebook, in its new March 2021 *Corporate Human Rights Policy*, has said it will "strive to respect") and especially to consider the UN OHCHR (Office of the High Commissioner on Human Rights) B-Tech paper, "Addressing Business Model Related Human Rights Risks," on the responsibilities of tech companies to conduct human rights due diligence across all of their business activities and relationships:

> This implies that they should: *i)* pro-actively identify when their business model-driven practices, and related technology designs, create or exacerbate human rights risks; and *ii)* take action to address these situations - whether by mitigating risks within existing business models or by innovating entirely new ones.

Don't Implement Policies That are "Bad for the World."

Recent Facebook policy implementations like banning posts about Holocaust denial, removing QAnon groups, and halting political ads before the 2020 U.S. elections have all drawn attention to the fact that Facebook and other social media platforms are perfectly capable of making major changes to their products when they decide that they want to do so. The most striking example of this fundamental capacity for change is illustrated in the Nov 25 2020 *New York Times* story, "Facebook Struggles to Balance Civility and Growth," which describes how, in the days after the US Presidential election, the platform implemented an algorithm to demote posts it had determined were: "bad for the world," but that, because of the resulting reduction in site engagement, the decision was made to "less stringently demote such content."

Good for the World:
An Invitation to Partnership

In the following sections we offer our evaluation and recommendations for each specific platform. As part of this Index, GLAAD hopes to meet with policy departments, as well as product designers and engineers, to work with them on implementing recommendations, improving their products and company policies, and then reporting on their achievements in future annual releases of the GLAAD Social Media Safety Index.

GLOBAL PERSPECTIVES & OTHER FRAMEWORKS

Our primary intent with the GLAAD Social Media Safety Index is to present recommendations to companies urging them to voluntarily undertake measures to improve their platforms. Other approaches to this problem include the current US Department of Justice **antitrust** lawsuit against Google and Federal Trade Commission antitrust case against Facebook (for further reading see Shoshana Zuboff's *The Age of Surveillance Capitalism* and also Cory Doctorow's How To Destroy Surveillance Capitalism). In February 2021 Senator Amy Klobuchar also introduced her Competition and Antitrust Law Enforcement Reform Act of 2021. Of course there is also the long-recommended argument for a **regulatory** "Digital Platform Agency" as well as **privacy regulation** approaches like the California Consumer Privacy Act (CCPA).

Among the various perspectives for looking at social media safety and platform responsibility, approaches foregrounding an **international human rights framework** offer a valuable perspective. Unanimously endorsed by the United Nations Human Rights Council in 2011, the *United Nations Guiding Principles on Business and Human Rights* is the leading international framework establishing corporate responsibility to respect human rights. While most of the companies analyzed herein are American-led, their reach is global. In fact, the majority of their users are non-American. As such, social media platforms must abide by the laws of other governments, and their policies around content moderation must ensure the safety of all users online, regardless of where they reside in the world. Voluntary global standards should also guide the content moderation policies and practices of platforms.

In March 2021 Facebook released a new *Corporate Human Rights Policy* which: "sets out the human rights standards [they] will strive to respect as defined in international law." The policy is a step in the right direction. As with so many of the company's policies, implementation—not just aspiration—is key. In their analysis of the policy, leading human rights and technology NGO Access Now expressed a blend of encouragement mixed with significant skepticism: "We welcome Facebook's new human rights policy, a necessary step for every company seeking to respect human rights. But 17 years is too long to wait for this basic declaration, especially from a huge and powerful firm like Facebook," said Peter Micek, Access Now's General Counsel. "The company's many failures in safeguarding data, respecting free expression, and protecting vulnerable users show Facebook adrift, far downstream, and paddling against inertia. If Facebook CEO Mark Zuckerberg signed off on this policy, he must ensure its implementation, respecting calls from civil society while complying with rights-respecting regulation, to chart an entirely new direction at Facebook."

It is extremely important and useful to be reminded that other governments around the world have taken vastly different, and often much more rigorous, approaches to public social media safety — especially in prioritizing the safety of individual citizens over the business interests of corporations. The "EU Code of Conduct on Countering Illegal Hate Speech Online" is just one example of this kind of policy approach. Leading platforms (including Facebook, Twitter, YouTube, Instagram, TikTok and others) have participated in this EU initiative and submitted to external monitoring of their progress in effectively preventing and removing hate speech from their platforms.

In the most recent monitoring report, issued in June 2020 and covering the previous year of 2019 (a period which does not include TikTok since the platform only joined in September 2020), sexual orientation was the most commonly reported type of hate speech — with 33.1% of users flagging such content. (Note that the report qualifies this number with the observation that: "In this monitoring round, organisations working on LGBTI rights have been more active in flagging content, in relative terms.")

These annual reports offer evaluation in the following areas: "Notifications of illegal hate speech," "Time of assessment of notifications," "Removal rates," "Feedback to users and transparency," and "Grounds for reporting hatred." Related to this approach, December 15, 2020 marked the announcement of the **EU Digital Services Act (DSA) and Digital Markets Act** — "a legislative reform that holds the promise of systemic regulation of large online platforms."

As former FCC Commissioner Susan Ness observes in a December 2020 article for Slate:

> These European rules could become the standard for the global net—leaving the U.S. behind. We have seen this before. American policymakers sat on the sidelines while the EU enacted its General Data Protection Regulation, which has become the de facto global standard. If America wants to help shape the rules of the road governing online discourse, it must step up and engage now.

Other relevant initiatives on the European level and focusing on AI include the Council of Europe: Ad Hoc Committee on Artificial Intelligence (CAHAI) and the *EU Commission: White Paper on AI*.

In a February 2021 Atlantic Council post pointing out one of the defining differences between the United States vs. European approaches, Frances Burwell notes that "the real question is why [...] private companies have been the key decision-makers. **Rather than relying on CEOs [...], the U.S. government—especially Congress and the courts—should make clear what type of speech is acceptable online and what type of speech is not.**"

The UN's 2018 "Report of the Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression on content moderation" states that:

> Companies should recognize that the authoritative global standard for ensuring freedom of expression on their platforms is human rights law, not the varying laws of States or their own private interests, and they should re-evaluate their content standards accordingly. Human rights law gives companies the tools to articulate and develop policies and processes that respect democratic norms and counter authoritarian demands. This approach begins with rules rooted in rights, continues with rigorous human rights impact assessments for product and policy development, and moves through operations with ongoing assessment, reassessment and meaningful public and civil society consultation. The Guiding Principles on Business and Human Rights, along with industry-specific guidelines developed by civil society, intergovernmental bodies, the Global Network Initiative and others, provide baseline approaches that all Internet companies should adopt.

While this GLAAD Social Media Safety Index report is primarily focused on U.S. examples and situations, the Council on Foreign Relations offers a helpful brief overview with more international context in their backgrounder, "Hate Speech on Social Media: Global Comparisons"). Also see the 2015 UNESCO report, "*Countering Online Hate Speech*."

FACEBOOK

Note: GLAAD is an organizational member of Facebook's Network of Support, a coalition of LGBTQ organizations that advise on select policy and product updates.

> *"Tech companies show an incredible ability to adapt their algorithms to boost engagement and profits. They need to devote similar energy to creating algorithms that minimize hate and harassment—for their sake and for society's"*
>
> — Ina Fried, Axios

Facebook has implemented several responsive platform changes in recent months, but numerous aspects of the platform continue to threaten public safety in general, and LGBTQ safety in particular. In so many instances where we may pause to commend a responsive change, the company has subsequently backtracked on such measures, failed to "operationalize" promised changes, or simply expressed theoretical fixes without offering concrete plans for implementation.

The Facebook Oversight Board (FOB), a body of experts to provide independent review, has been one of the company's approaches to these problems. For a deeper dive into that process see the Board's first round of January 2021 recommendations and Facebook's response.

The 2020 Ranking Digital Rights Corporate Accountability Index offers an in-depth evaluation of Facebook's overall performance on numerous metrics, including relevant social media safety indicators. There is so much reporting and research on Facebook's corrosive impacts — one more report worth calling out is Amnesty International's Surveillance Giants: How The Business Model of Google and Facebook Threatens Human Rights.

Below are some of our specific recommendations for Facebook. We urge Facebook to also attend to the general recommendations and Platform Responsibility Checklist in the first part of this report, including items related to protection of LGBTQ users in community guidelines, algorithmic bias and bias in AI, privacy and outing, promoting civil discourse, and more.

RECOMMENDATIONS

Content Moderation

Facebook Community Standards regarding "Objectionable Content" state:

> We do not allow hate speech on Facebook because it creates an environment of intimidation and exclusion and in some cases may promote real-world violence. We define hate speech as a direct attack on people based on what we call protected characteristics—race, ethnicity, national origin, religious affiliation, sexual orientation, caste, sex, gender, gender identity, and serious disease or disability[...]. We define attack as violent or dehumanizing speech, harmful stereotypes, statements of inferiority, or calls for exclusion or segregation.

While GLAAD acknowledges the expansiveness of these guidelines, Facebook must achieve considerable improvement in the enforcement of these policies to make the platform safe for LGBTQ users, and for everyone.

Some areas to be addressed in the broad realm of **content moderation**, and some recommendations for improvement, include the following items:

Protection of LGBTQ Users in Community Guidelines

See numerous items in our "Recommendations for All Platforms" above.

Disinformation/Misinformation

RECOMMENDATION:
Label Content and Identify Trusted Sources.

Many instances of anti-LGBTQ content fall under the heading of dis/misinformation. The platform could make better use of dis/misinformation mitigation tools such as labeling certain kinds of content and/or pointing users to other trusted sources. Facebook does have a False News reporting option that enables users to report dis/misinformation.

RECOMMENDATION:
Third-Party Fact Checking.

In 2016, Facebook started its third-party fact-checking to rate and review the accuracy of content on the platform. Facebook states that "when misinformation is identified by our fact-checking partners, we reduce its distribution within News Feed and other surfaces" and that Facebook applies "strong warning labels and notifications on fact-checked content." Shockingly, one of the ten independent fact checkers in the U.S. is 'Check Your Fact,' a for-profit subsidiary wholly owned by The Daily Caller, one of the most virulently anti-LGBTQ 'news' sites. As noted in our case study, appealing to the fact checking organizations has resulted in ads and content with misleading and inaccurate information being removed, but findings from fact checking organizations should better inform the Community Standards that govern Facebook content, so action is taken *before* such content reaches unsuspecting users. Facebook chose to give a large amount of power and decision-making to its independent fact-checkers and would benefit by adding LGBTQ-focused news outlets or organizations which can be tapped for information about our lives.

Transparency and Accountability

RECOMMENDATION:
Improve the Process of User Reporting.

User reporting is a key to helping fight anti-LGBTQ content and conduct, but the tools for reporting content, comments, and accounts need to be more robust. While the Facebook Support InBox does a good job at offering access to correspondence about reports, it does not provide a link to the reported content if it is a post or comment (only if it is a page or group). While these 2020 Facebook Newsroom releases, "How We Review Content" and "Measuring Our Progress Combating Hate Speech" are a step in the right direction, Facebook should provide much greater transparency on how decisions are made and what recourse users have for appeals. For further details on best practice recommendations see the Santa Clara Principles On Transparency and Accountability in Content Moderation and their proposal of: "initial steps that companies engaged in content moderation should take to provide meaningful due process to impacted speakers and better ensure that the enforcement of their content guidelines is fair, unbiased, proportional, and respectful of users' rights."

RECOMMENDATION:
Bring Transparency to the System of Enforcement.
With regard to enforcement of community guidelines violations — among the many demands of the coalition of independent researchers, activists, and academics known as "The Real Facebook Oversight Board" GLAAD concurs with their call for: "A public codified system that makes transparent what Facebook's system of enforcement is. There needs to be a clearly defined strike-out system in line with Twitter's Civic Integrity update [of January 11 2021]."

Incorrect Blocking of LGBTQ Content

RECOMMENDATION:
Use Qualified Human Moderators.

Facebook should increase use of qualified human moderators to more accurately interpret legitimate use of LGBTQ terms and to distinguish legitimate accounts and posters from trolls and bad actors. Make corresponding improvements to AI systems. Also see below item on self-expression.

Algorithms

In addition to the notes on algorithms made in the general recommendations above, **Facebook must prioritize improved practices and systems to reduce anti-LGBTQ hate and extremist content — including adjusting both their current content moderation systems and their algorithms which appear to escalate the dissemination of such content**.

As is true of other platforms, when a Facebook user looks at the feed of a hateful profile or group they are given further such recommendations of similar profiles or groups to follow. These recommendation algorithms drive users further into what researchers call "information silos," effectively eliminating other perspectives (see sidebar "Straight Pride" example). Adjusting algorithms is just one way for platforms to enhance safety, a practice that has already been demonstrated to be effective by Facebook itself.

As journalists and researchers have repeatedly observed, there are many known solutions to the problems that algorithms create, solutions that the platforms themselves will often briefly implement and then retract. While Facebook's January 2021 promise to stop recommending political groups sounds great on the face of it, observers are understandably pessimistic about the company's good faith and actual implementation of this.

Here's one of many instances that illustrate why such suspicion is warranted: In the weeks leading up to the 2020 election, Facebook chose to temporarily implement a viral "circuit-breaker" content-review system as a way of halting the spread of a fake story related to Hunter Biden. In an October 18, 2020 Fortune.com article, tech beat reporter Jeff John Roberts points out that:

> The tool has enormous potential to limit a tsunami of false or misleading news on topics like politics and health. The circuit breaker tactic is a common sense way for the social network to fix its fake news problem, but it may also run counter to Facebook's business interest [...]. *The company, meanwhile, has yet to offer a convincing answer about how it plans to reconcile [the] tension between an ethical duty to limit the spread of misinformation, and the fact it makes money when such misinformation goes viral.*

Self-expression & LGBTQ inclusion

RECOMMENDATION:
Be Diligent in Protecting LGBTQ Self-Expression

With regard to LGBTQ self-expression, it is good to see that Facebook's robust policies reflect the understanding that individuals who belong to protected groups may use self-referring terminology which might otherwise be considered offensive ("In some cases, words or terms that might otherwise violate our standards are used self-referentially or in an empowering way.") GLAAD also reminds the platform of the need for continued diligence in the implementation and enforcement of these policies, lest legitimate LGBTQ content be over-policed or unfairly removed. GLAAD also strongly urges Facebook to devote resources to gathering and releasing data on the current state of LGBTQ self-expression on the platform. Additional research is needed to determine all the ways that LGBTQ users are currently being impacted in this area.

LGBTQ hiring, inclusion, & leadership

RECOMMENDATION:
Continue to Diversify the Workforce.

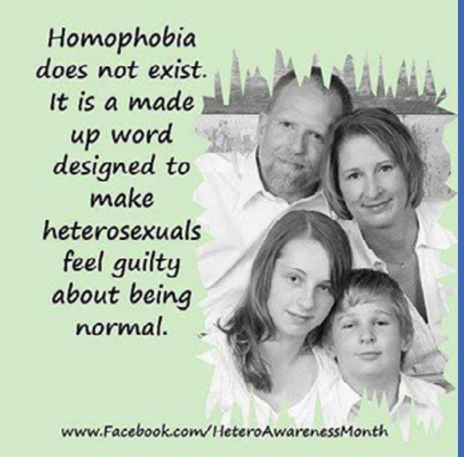
GLAAD strongly urges Facebook to continue to diversify its hiring of LGBTQ employees, especially in positions in the engineering and product teams, to shape the platform at its highest level. The 2020 Facebook Diversity Report indicates that "about 8% of US-based Facebook employees identify as LGBTQ+, based on a voluntary survey." It is encouraging to see that as of February 2021 Facebook is hiring a Director of Diversity and Inclusion. Diverse workforces are essential in serving the needs of a diverse array of users. It is also essential to hire LGBTQ content moderators and to train all content moderators to understand the needs of LGBTQ users.

This is only a partial list. Please see the Platform Responsibility Checklist and general recommendations section above.

THUMBS DOWN

Straight Pride

The homosexual agenda is the biggest threat to the right of free speech today.
— SPWW, "Straight Pride World Wide" Facebook page description


Homophobia does not exist. It is a made up word designed to make heterosexuals feel guilty about being normal.
www.Facebook.com/HeteroAwarenessMonth

Example of meme featured on SPWW, Straight Pride World Wide Facebook Group page.

There are a handful of types of expression we associate with social media platforms: postings of text, images, videos, links; comments on these items; advertising; individual profile pages (profile image, description, etc.); businesses or causes or organizational pages (ditto); and, of course, in the case of Facebook, groups dedicated to certain themes or interests (for instance, "LGBT Pride" on the one hand, or "Straight Pride," on the other — and yes, this is a wonderful illustration of how effortless it is to create inherently hateful content that does not technically violate the community guidelines).

If we consider all of these "expressions" in terms of community guidelines, the main question in each case—whether "speech" is evaluated by human or artificial intelligence—is whether an item is acceptable or offensive, malignant or benign. "Straight Pride" may sound innocuous enough, and our first instinct as Americans is the desire to protect even the free speech of people who hate us. Facebook being the dark landscape that it is, though, we can land on hundreds of such pages in a single click to discover that "SPWW, Straight Pride World Wide" displays a main banner image comparing homosexuals to Satan worshippers, rapists, and murderers (as of early October 2020)[9]. Facebook's meticulously designed recommendation algorithms will then direct us to "Related Pages" like: "Homosexuality is Wrong, Enough is Enough—Stop Homosexual Promotion" and "Child Protection League" (an extreme right-wing organization warning of the dangers of gender neutral bathrooms and sex education in schools).

While the tension between hate speech and free speech is one of the most persistent dilemmas of our time, the fact remains that social media companies actively exploit this tension with the sole purpose of reaping enormous profits every single day. As the disproportionate targets of that hate, LGBTQ people and other marginalized individuals are the ones paying the price.

[9] Description of "SPWW, Straight Pride World Wide" from their About section: The Homosexual Agenda is a self-centered set of beliefs and objectives designed to promote and even mandate approval of homosexuality and homosexual ideology, along with the strategies used to implement such. The goals and means of this movement include indoctrinating students in public school, restricting the free speech of opposition, obtaining special treatment for homosexuals, distorting Biblical teaching and science, and interfering with freedom of association. Advocates of the homosexual agenda seek special rights for homosexuals that other people don't have, such as immunity from criticism (see hate speech, hate crimes). Such special rights will necessarily come at the expense of the rights of broader society. The homosexual agenda is the biggest threat to the right of free speech today.

THUMBS DOWN

Ads Harmful to LGBTQ People Proliferate In Spite of Expert Reporting

In September 2019, public health advocate and HIV activist Peter Staley contacted GLAAD regarding ads on Facebook and Instagram, placed by a variety of personal injury law firms, with misinformation about the use of Pre-Exposure Prophylaxis (PrEP) as a preventative measure against HIV. After several conversations with Facebook public policy and advertising policy, GLAAD, PrEP4All, and more than 50 LGBTQ, HIV, and public health groups sent an open letter to Facebook which called for the ads to be removed, noting: “Leading public health officials, medical professionals, and dedicated PrEP navigators and outreach coordinators have shared that these advertisements on Facebook and Instagram are being directly cited by at-risk community members expressing heightened fears about taking PrEP. This issue goes beyond misinformation, as it puts real people’s lives in imminent danger.”

GLAAD reported several ads to six independent fact-checking agencies which review ads running on Facebook products for misinformation. Weeks later, one of those agencies, Science Feedback, confirmed that an ad featuring misinformation was ‘misleading’ and would be removed from Facebook. According to Facebook’s Advertising Library at the time, additional ads with similar language were also removed. Facebook confirmed to the *Washington Post*: “After a review, our independent fact-checking partners have determined some of the ads in question mislead people about the effects of Truvada [for PrEP]. As a result we have rejected these ads and they can no longer run on Facebook.”

In a January 2020 response letter, Facebook also noted that: “While this does not mean we’ve rejected all the ads that have been surfaced across different advertisers—as some make variations of the claim and require separate assessment—we will continue to surface ads that appear to make similar claims to factcheckers and will reject any that are debunked.”

Unfortunately, ads with nearly identical language and imagery to those previously removed continue to run on Facebook nearly a year and a half later— and after follow-up discussions. As long as Facebook has no clear mechanism to detect and bar harmful ads, the onus is on individuals and groups like GLAAD to search Facebook’s Advertising Library and report content to agencies like Science Feedback. In April 2021, GLAAD and PrEP4All sent a follow-up letter to Facebook asserting that these ads are still harming public health and that more action is urgently needed.

THUMBS UP

Affirming LGBTQ Original Content & Visibility

Facebook (and Facebook-owned Instagram) put significant resources into creating campaigns, product features, and original content that spotlight LGBTQ people and provide creative ways for LGBTQ self-expression. In 2020, Facebook added new Pride stickers and frames for Facebook and Messenger, the latest in consistent updates for Pride and other LGBTQ days of visibility. In 2019, Instagram rolled out additional Pride stickers, GIFs, and displayed popular hashtags that the community uses to connect in a rainbow gradient, a project that Instagram collaborated with GLAAD on. Instagram has collaborated with GLAAD, The Trevor Project and other organizations on content and campaigns including a 2021 Guide that highlighted trans comedians, which was featured in front of the 370M+ followers of the @Instagram handle on Transgender Day of Visibility. The LGBTQ@Facebook page counts 20M+ followers and regularly shares content and resources from LGBTQ organizations. The page, run by Facebook staff, is an important LGBTQ news source and spot for community. Facebook Watch Originals is also home to LGBTQ content including a powerful 2021 episode of “Peace of Mind with Taraji” about Black transgender women, a 2020 Coming Out special with Demi Lovato and Tan France, as well as a 2020 episode of “Red Table Talk: The Estefans” featuring Emily Estefan’s coming out story, which was nominated for a GLAAD Media Award. As platforms create more original content, we hope to see LGBTQ representation in that programming rise. Proactive LGBTQ campaigns and content, which all companies in this report produce, are important for LGBTQ users. What’s also critical is to showcase this content in places that are not LGBTQ exclusive (@Instagram, Facebook Watch) so non-LGBTQ audiences can also interact with the content.

Source: www.facebook.com/watch

Facebook Community Standards Enforcement Report — Q2 & Q3 2020

Here are the Q2 2020 numbers for hate speech on Facebook according to the company’s Community Standards Enforcement Report:

Facebook removed 22.5 million posts during the second quarter that violated its rules against hate speech, more than double the number during the first quarter and nearly quadruple from the same period in 2019.” And here are the Q3 2020 numbers: “22.1 million pieces of hate speech content, about 95% of which was proactively identified.

Our colleagues in the field (especially the Change the Terms and Stop Hate for Profit initiatives spearheaded by organizations including the ADL, Color of Change, Center for American Progress, Southern Poverty Law Center and others) have worked tirelessly in recent years to expose the extraordinary levels of hate and misinformation on Facebook and to maintain pressure on the company to remedy these problems. GLAAD echoes the positions outlined in the groundbreaking Stop Hate for Profit and Change the Terms reports as well as the more recent efforts of “The Real Facebook Oversight Board.”

With regard to Facebook, the ADL’s scathing November 2020 analysis of the platform’s most recent transparency reporting on hate speech takes Facebook to task on the relative *lack* of transparency in the numbers released in the report. The ADL’s analysis—“Facebook’s Transparency Reporting Continues to Obscure the Scope and Impact of Hate Speech”—is worth reading in full but this concluding point is especially powerful:

> Facebook needs to report on the prevalence of hate speech targeting specific communities, the experiences that distinct groups are having on its platform and the numbers for the different kinds of hate being spread. For example, how many antisemitic, anti-Black, anti-Muslim and anti-LGBTQ+ pieces of content required actioning? Without specifying these numbers and the types of content attacking each vulnerable group, it is difficult for civil rights groups to propose solutions to these problems. Facebook can follow the example set by Reddit by conducting a study on hate and abuse on its platform and making its findings public. The company should also conduct another independent audit, specifically focused on its lack of transparency.
>
> Greater transparency and active data collection around online hate speech should be accompanied by evidence-based policies and enforcement mechanisms. To show they are taking real steps to reduce hate speech, platforms must try to understand the scope of the problem by collecting the relevant data and using rigorous research methods. Failure to do so will result in vulnerable groups continuing to be at the mercy of toxic users on social media.[10]

In concluding our recommendations, we urge every individual in a position of leadership at Facebook to find ways to take meaningful action now to make the platform safe for its LGBTQ users.

[10] See also: “Understanding Hate on Reddit.”

TWITTER

Note: GLAAD is an organizational member of the Twitter Trust & Safety Council and has been advising the platform on these matters since 2016.

> *"Internet companies can no longer neglect how the hate speech of the few silences the voices and threatens the lives of the marginalized many."*
>
> **— Jessica González, Co-founder, Change the Terms & Co-CEO, Free Press**

Researchers and journalists evaluating Twitter's approach to hate speech, including anti-LGBTQ content and conduct, and dis/misinformation (especially in the intensely fraught landscape of late 2020 and into 2021) have noted the platform's many improvements in regard to user safety. These include the addition of warning labels to tweets determined to contain misinformation; the blocking of hashtags known to be used for hate and extremism (for example: #ProudBoys)[11]; the monitoring and slowing of trending topics to reduce the likelihood of viral spread of misinformation and hate; and other techniques and strategies including hiding or removing tweets, limiting tweet visibility, and de-platforming repeat violators of community guidelines.

Under the heading, "Twitter, the best of the worst," the *2020 Ranking Digital Rights Corporate Accountability Index* offers an in-depth evaluation of Twitter's overall performance on multiple metrics, including relevant social media safety indicators. Also see the *Amnesty International Toxic Twitter Scorecard* for valuable notes on transparency, reporting mechanisms, the abuse report review process, and privacy and security features. There are many, many changes the platform can implement to make their product safer for LGBTQ users. Below are some of our specific recommendations for Twitter. We urge Twitter to also attend to the general recommendations and Platform Responsibility Checklist in the first part of this report, including items related to: Protection of LGBTQ users in community guidelines; Algorithmic bias and bias in AI; Privacy and outing; Promoting civil discourse; and more.

RECOMMENDATIONS

Content Moderation

Twitter's Hateful Conduct Policy states: "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 community guidelines further prohibit posts, images and display names that promote: "violence against or directly attack or threaten other people on the basis of race, ethnicity, national origin, caste, *sexual orientation, gender, gender identity,* religious affiliation, age, disability, or serious disease." They also prohibit: "hateful images or symbols in profile image or profile header" and/or using "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.*"

These policies are expansive (especially their specific prohibition of misgendering and deadnaming — see sidebar). The company must also continue to enhance enforcement of these policies to make the platform safe for LGBTQ users. Some areas to be addressed in the broad realm of **content moderation**, and some recommendations for improvement, include the following items below.

Protection of LGBTQ Users in Community Guidelines

RECOMMENDATION:
Expand Current Prohibitions on "Language That Dehumanizes Others" to Include Anti-LGBTQ Language.

On top of its regular Hateful Conduct policy ("You may not promote violence against, threaten, or harass other people on the basis of [...]"), in July 2019 Twitter also added a prohibition against "language that dehumanizes others" based on religion or caste. In March 2020 this was further expanded to include language that dehumanizes others on the basis of age, disability or disease, and in December 2020 to include race, ethnicity or national origin. GLAAD strongly urges Twitter to add a prohibition against language that dehumanizes others on the basis of sexual orientation and gender identity— as soon as possible in 2021. (As of mid-March, 2021, this specific policy still reads: "We also prohibit the dehumanization of a group of people based on their religion, caste, age, disability, serious disease, national origin, race, or ethnicity.").

Disinformation/Misinformation

RECOMMENDATION:
Implement Substantive Solutions to Disinformation/Misinformation.

Many instances of anti-LGBTQ content and conduct fall under the heading of dis/misinformation. Twitter should make better use of dis/misinformation mitigation tools such as labeling certain kinds of content or pointing users to other trusted sources, and restricting engagement with tweets labeled as misinformation (including restricting them from being retweeted, replied to, or liked). Also, Twitter could easily add a "False Information" option to its menu of user reporting categories. (The closest options currently are the "suspicious or spam" and the "It's something else" categories.) Also, see the sidebar on the new Birdwatch pilot.

Transparency & Accountability

RECOMMENDATION:
Improve Transparency, Accountability, and the User-Reporting Process.

The platform could improve the system of user-reporting (of content, comments, and accounts). Twitter's Rules Enforcement report proclaims that it supports "the spirit" of the Santa Clara Principles on Transparency and Accountability in Content Moderation, and promises that the company is, "committed to sharing more detailed information about how we enforce the Twitter Rules in the future." We urge Twitter to provide greater levels of transparency and granular data on violated policies— now, and to continue to make improvements—including providing a more robust experience of transparency to users issuing reports. (Twitter currently offers only a basic notification for users issuing reports: "We received your report over the past hour[...]. If we take further action, we'll let you know.") While it is still insufficient, Twitter does provide some added info to users on their Help Center page:

> 5. We will include the text of the Tweets you reported in our follow-up emails and notifications to you. To opt-out of receiving this information, please uncheck the box next to Updates about this report can show these Tweets. 6. Once you've submitted your report, we'll provide recommendations for additional actions you can take to improve your Twitter experience.

Incorrect Blocking of LGBTQ Content

RECOMMENDATION:
Use Qualified Human Moderators.

Twitter should increase use of qualified human moderators to more accurately interpret legitimate use of LGBTQ terms and to distinguish legitimate accounts and posters from trolls and bad actors. The company should make corresponding improvements to AI systems. Also see below on self-expression.

[11] A brief creative anecdote of activists turning the tables on hate: In October 2020 (on the day after Donald Trump issued his presidential debate stage call for the white supremacist hate group, urging the Proud Boys to "stand back and stand by"), LGBTQ Twitter users instigated a takeover of the #ProudBoys hashtag and flooded the platform with proud gay imagery.

Algorithms

RECOMMENDATION:
Keep Refining Systems to Reduce Hate.

In addition to the notes on algorithms made in the general recommendations above, Twitter must prioritize improved practices and systems to reduce anti-LGBTQ hate and extremist content. It is notable that Twitter has implemented a variety of strategies to reduce the posting and spread of anti-LGBTQ hate speech, extremist rhetoric, and dis/misinformation—including Twitter's commitment to, "Focus [more] on how content is discovered + amplified, less on removal alone."

Self-expression & LGBTQ inclusion

RECOMMENDATION:
Report on the Current State of LGBTQ Self-Expression.

With regard to LGBTQ self-expression, it is good to see that Twitter's robust policies reflect the understanding that individuals who belong to protected groups may use self-referring terminology that might otherwise be considered offensive (From the Twitter Hateful Conduct policy: "Some Tweets may appear to be hateful when viewed in isolation, but may not be when viewed in the context of a larger conversation. For example, members of a protected category may refer to each other using terms that are typically considered as slurs. When used consensually, the intent behind these terms is not abusive, but a means to reclaim terms that were historically used to demean individuals.") GLAAD recommends continued diligence in the implementation and enforcement of these policies, lest legitimate LGBTQ content be over-policed or unfairly removed. GLAAD also strongly urges Twitter to devote resources to gathering and releasing data on the current state of LGBTQ self-expression on the platform. Additional research is needed to determine all the ways that LGBTQ users are currently being impacted in this area.

LGBTQ hiring, inclusion & leadership

RECOMMENDATION:
Stay Committed to Diverse Hiring.

GLAAD strongly urges Twitter to continue to diversify its hiring of LGBTQ employees, especially in positions in the engineering and product teams to shape the platform at its highest level. In March 2020 the company announced it would begin publicly reporting progress on workforce representation of LGBTQ employees at the end of the year — the December 2020 report however does not include this data. Diverse workforces are essential in serving the needs of a diverse array of users. It is also essential to hire LGBTQ content moderators and to train all content moderators to understand the needs of LGBTQ users.

This is only a partial list. Please see the Platform Responsibility Checklist and general recommendations section above.

THUMBS UP

Prohibition Against Misgendering/Deadnaming

Twitter offers a thoughtful rationale contextualizing their Hateful Conduct policies and includes a specific prohibition against misgendering and deadnaming: "We prohibit targeting individuals with repeated slurs, tropes or other content that intends to dehumanize, degrade or reinforce negative or harmful stereotypes about a protected category. *This includes targeted misgendering or deadnaming of transgender individuals.*" While GLAAD's attempt to flag select content that misgendered did not result in removal of tweets, GLAAD urges other platforms to adopt and enforce this policy as well. Twitter also does an excellent job at thoroughly characterizing the values behind these policies: "We are committed to combating abuse motivated by hatred, prejudice or intolerance, particularly abuse that seeks to silence the voices of those who have been historically marginalized. For this reason, we prohibit behavior that targets individuals with abuse based on protected category."

THUMBS UP

Birdwatch

Twitter's Birdwatch pilot—the community-reliant moderation initiative launched in January 2021—is an innovative effort that could play a critical role in slowing the spread of misinformation. From Twitter: "Birdwatch is a community-driven approach to address misinformation on Twitter. Participants can identify Tweets they believe are misleading, write notes that provide context to the Tweet, and rate the quality of other participants' notes." While Birdwatch is focused on misinformation, this kind of functionality could also help mitigate hate and harassment. Offering input for this report, numerous SMSI advisory committee members pointed to the community-reliant moderation approaches of both Wikipedia and Reddit as possible models that platforms should look to.[12] Twitter also incorporates transparency into the Birdwatch initiative: all data contributed will be publicly available, and the company claims that they "aim" to make the new algorithms publicly available as well. Birdwatch also appears to be following a variety of other best practices and recommendations including integrating social science and academic perspectives in the product development process as well as aspiring to reflect diverse perspectives to avoid bias. While there will be big questions to answer as the pilot rolls out, it is promising to see the platform investing in exploring innovative solutions.

Twitter Rules Enforcement Data — January-June 2020

The following numbers are drawn from the Twitter Rules Enforcement for January–June 2020. During this period, Twitter reports there were 1.9 million accounts actioned, 925.7 thousand accounts suspended, and 1.9 million items of content removed.

Twitter's Rules Enforcement report does not include specific data on anti-LGBTQ content removals (as is true of other platforms, Twitter should share disaggregated data). But in the broader categories of Abuse/Harassment and Hateful Conduct, the numbers are as follows:

Hateful conduct:
accounts actioned: 635,415
accounts suspended: 127,954
content removed: 955,212

Abuse/harassment
accounts actioned: 398,057
accounts suspended: 72,139
content removed: 609,253

While this Jan-June 2020 reporting period shows a 16% decrease in accounts actioned compared to the previous reporting period, in that previous July-December 2019 report, according to Twitter: "there was a 95% increase in the number of accounts actioned for violations."

In concluding our recommendations, we urge every individual in a position of leadership at Twitter to find ways to take meaningful action now to make the platform safe for its LGBTQ users.

[12] For more on the three main models of content moderation see this Data & Society report, Content or Context Moderation? Artisanal, Community-Reliant and Industrial Approaches.

YOUTUBE

> *"There is often a great disconnect between what actions YouTube says it is taking and what users and creators actually experience. This is in part because these actions mean little if the platform has no clear idea of how it defines hate speech, extremism, harassment or borderline content and what values it seeks to uphold in its actions. Indeed, YouTube has often backed itself into a corner by attempting to stay as "apolitical" as possible… The great irony is that by attempting to stay apolitical, YouTube consistently makes the political choice not to care about or protect vulnerable communities."*
>
> — Becca Lewis, *The Guardian*

There are countless individual YouTube channels with enormous followings that traffic in hateful rhetoric, including anti-LGBTQ sentiment. While the platform continues to effectively remove select instances of anti-LGBTQ hate speech, many others remain (see sidebar excerpt from the November 2020 *Media Matters* report on PragerU).

The 2020 Ranking Digital Rights Corporate Accountability Index offers an in-depth evaluation of YouTube parent company Google's overall performance on numerous metrics, including relevant social media safety indicators.

Below are some of our specific recommendations for YouTube. We urge YouTube to also attend to the general recommendations and Platform Responsibility Checklist in the first part of this report, including items related to: Protection of LGBTQ users in community guidelines; Algorithmic bias and bias in AI; Privacy and outing; Promoting civil discourse; and more.

RECOMMENDATIONS

Content Moderation

YouTube community guidelines prohibit: "Content promoting violence or hatred against individuals or groups based on any of the following 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, Veteran Status." YouTube also does not allow: "content that targets an individual with prolonged or malicious insults based on intrinsic attributes, including their protected group status [the same list as above]." In the platform's Harassment and Cyberbullying policy, YouTube prohibits: "Content that features prolonged name calling or malicious insults (such as racial slurs) based on their intrinsic attributes. These attributes include their protected group status[...]."

These guidelines are expansive, but the company must also achieve considerable improvement in the enforcement of these policies to make the platform safe for LGBTQ users. Areas to be addressed in the broad realm of **content moderation**, and recommendations for improvement, include the following items below.

Protection of LGBTQ Users in Community Guidelines

RECOMMENDATION:
Label Content and Point Users to Trusted Sources.

Many instances of anti-LGBTQ content and conduct fall under the heading of dis/misinformation. YouTube should increase the use of mitigation tools such as labeling certain kinds of content or pointing users to other trusted sources. YouTube does have a Spam or Misleading option that can be used to report dis/misinformation on the platform.

Transparency & Accountability

RECOMMENDATION:
Improve Transparency, Accountability, and the User-Reporting Process.

YouTube should improve their system of user-reporting (of content, comments, and accounts) to provide full transparency to the user issuing the report (YouTube does offer transparency to the user whose material is being reported). The platform should also provide transparency on how decisions are made and what recourse users have. See the Santa Clara Principles On Transparency and Accountability in Content Moderation for best practice recommendations. (See sidebar case study example of our attempt to report the viral transphobic "It's Ma'am" video; also see other related YouTube sidebars).

Incorrect Blocking of LGBTQ Content

RECOMMENDATION:
Use Qualified Human Moderators.

YouTube should increase the use of qualified human moderators to more accurately interpret legitimate use of LGBTQ terms and to distinguish legitimate accounts and posters from trolls and bad actors. The platform should make corresponding improvements to AI systems. GLAAD is cautiously optimistic about YouTube's announced 2021 survey to identify LGBTQ content creators in an effort to evaluate any problems in, "possible patterns of hate, harassment, and discrimination" and to look at, "how content from different communities is treated in our search and discovery and monetization systems." (Also, see below item on self-expression. And see sidebar: "Unfair Demonetization and Removal of LGBTQ Content on YouTube.")

RECOMMENDATION:
Improve Content Moderation.

According to its own transparency reports, the majority of takedowns on YouTube are facilitated via AI. The platform has a history of both over-policing legitimate LGBTQ content (see sidebar: "Unfair Demonetization and Removal of LGBTQ Content on YouTube") as well as a history of failing to remove actual anti-LGBTQ content, comments, and accounts (see sidebar "YouTube removed anti-trans PragerU videos for violating hate speech policies" and links below in next item). Both of these problems must be addressed.

Algorithms

RECOMMENDATION:
Refine Algorithms to Reduce Hate, Not Spread It.

In addition to the notes on algorithms made in the general recommendations above, YouTube must prioritize improved practices and systems to reduce anti-LGBTQ hate and extremist content, including adjusting both their current content moderation systems—and their algorithms, which appear to escalate the dissemination of such content. A 2019 UK study investigating extremist content ("Radical Filter Bubbles: Social Media Personalisation Algorithms and Extremist Content") found that YouTube's algorithms introduce increasingly extreme content to users who have previously engaged with less extreme content. YouTube could do much more to fight the spread of anti-LGBTQ hate-driven content and conduct on the platform. Adjusting recommendation algorithms is just one example. As a December 2020 *USA Today* feature describes: "For years, YouTube executives ignored staff's warnings that its recommendation feature, which aimed to boost time people spend online and generate more advertising revenue, ignited the spread of extremist content, according to published reports."

Self-Expression & LGBTQ Inclusion

RECOMMENDATION:
Stop Blocking Words Like "Gay" or "Transgender."

Closely related to content moderation bias and over-policing of legitimate LGBTQ content, the platform has a history of blocking some simple uses of the words gay, lesbian, and transgender. (See Unfair Demonetization sidebar). GLAAD strongly urges YouTube to devote product team resources to fixing these problems and urges the platform to release a report on the current state of the situation. Additional research is needed to determine all the ways that LGBTQ users are currently being impacted in this area.

RECOMMENDATION:
Note Context.

YouTube's hate speech policy offers an explanation of the importance of context. Unfortunately this policy as outlined (and as implemented) falls short with regard to LGBTQ self-expression. GLAAD recommends that the platform follow the lead of Twitter, Facebook, and Instagram—all of which have more robust policies expressing the understanding that individuals who belong to protected groups may use self-referring terminology which might otherwise be considered offensive.

LGBTQ hiring, inclusion & leadership

RECOMMENDATION:
Continue to Diversify Hiring.

GLAAD strongly urges YouTube to diversify its hiring of LGBTQ employees especially in positions in the engineering and product teams to shape the platform at its highest level. The most recent diversity report from YouTube's parent company, Google indicates that 7.1% of Google workers self-identify as LGBQ+ and/or Trans+. Diverse workforces are essential in serving the needs of a diverse array of users. It is also essential to hire LGBTQ content moderators and to train all content moderators to understand the needs of LGBTQ users.

RECOMMENDATION:
Consult Expert LGBTQ Advisors.

GLAAD urges YouTube to establish a council of external experts to advise on policies and product updates, including LGBTQ content moderation. This body could be modeled on Facebook's Network of Support (a group of LGBTQ organizations that Facebook contacts around potential product and policy updates) and Twitter's Trust & Safety Council (an intersectional collection of organizations representing diverse communities from around the globe that provide feedback on policy and product updates).

This is only a partial list. Please see the Platform Responsibility Checklist and general recommendations section above.

THUMBS UP

LGBTQ Education & Allyship

YouTube has a strong history of original content which fosters education and acceptance, including showcasing the ACLU LGBT Rights Project on the YouTube Social Impact page. The ACLU's short video: "A Boy Named Gavin" (about teen trans-rights activist Gavin Grimm) is presented with the caption: "The ACLU uses animation to help their audience emotionally connect with an individual story while generating empathy for LGBT rights."

YouTube Originals has produced and distributed award-winning LGBTQ-focused content including the documentaries 'This Is Everything: Gigi Gorgeous,' which was nominated for a GLAAD Media Award in 2018, and 'State of Pride,' which received the GLAAD Media Award for Outstanding Documentary in 2020. Google and YouTube were official streaming partners for 2020 Global Pride, which featured content from more than 500 Pride and community organizations from 91 countries. YouTube Originals also recently announced a multi-hour livestream event for Pride 2021 to benefit The Trevor Project.

Prominently featuring and sharing, as well as proactively creating, fair and accurate LGBTQ content should continue to be a priority, not only for YouTube, but for all companies in this report. It is also important to ensure such content reaches non-LGBTQ audiences by promoting original content and campaigns via main pages and other spots where the general public visits, as well as suggesting such content to viewers of non-LGBTQ content.

THUMBS UP

AI-driven "Comment Reminder"

YouTube recently implemented an AI-driven "comment reminder" urging commenters whose posts are perceived to be in possible violation of community guidelines to reconsider whether they're sure they want to post.

Of course we also urge YouTube to ensure this AI feature does not discriminate against or burden LGBTQ users and other marginalized communities who are disproportionately impacted by the widespread and well-documented phenomenon of AI bias (see our recommendations below).

YouTube Content Moderation Data — Q3 2020

The following numbers are drawn from YouTube's transparency reporting for Q3 2020 (July-September 2020), which offers reporting related to **violations and removals of comments, videos, and channels**. The report states that 99.6% of community guidelines violating comments removed from the platform were detected by their automated flagging system (.4% were reported by human flaggers). YouTube does not report specifically on anti-LGBTQ videos or comments but does distinguish several relevant subcategories of removals within which anti-LGBTQ conduct and content would appear (the most relevant are: "Hateful or abusive", "Harassment and cyberbullying," "Promotion of violence," "Harmful or dangerous"; anti-LGBTQ conduct or content could also appear under other categories). Although the combined percentage of removals (of videos and of comments) across these categories is relatively small amidst the far larger percentage of takedowns in other categories, these numbers are still quite substantial (the number of total comments removed in this three month period was: 1,140,278,887—more than one billion!). As is true of other platforms, YouTube should share disaggregated data to afford researchers clear visibility into anti-LGBTQ activity on the platform.

Comment removals:
Hateful or abusive (1.1%), Harassment and cyberbullying (17.3%) — total: 18.4%. Other removals are as follows: Nudity or sex (0.2%); Spam/misleading (51.4%), Child safety (26.9%).

Video removals:
Hateful or abusive (4.1%), Harassment and cyberbullying (0.6%), Promotion of violence and violent extremism (2.5%), Harmful or dangerous (2.5%) — total: 6.7%. Other removals are as follows: Violence or graphic imagery (14.2 %); Nudity or sex (0.2%); Spam/misleading (25.5%), Child safety (31.7%); Other (1.9%).

Channel removals:
Hateful or abusive (3.0%), Harassment and cyberbullying (0.8%), Promotion if violence and violent extremism (0.5%) — total: 4.3%. Other removals are as follows: Nudity or sex (6.5%); Spam/misleading (85.4%), Child safety (2.0%); Impersonation (1.3%); Multiple violations (0.3%).

From
Media Matters (November 23, 2020):

YouTube removed anti-trans PragerU videos for violating hate speech policies

YouTube said the videos violated rules that forbid claims that trans people "are physically or mentally inferior, deficient, or diseased." Other videos that make similar claims remain on the platform.

On November 20, YouTube removed two anti-trans videos from right-wing propaganda network PragerU for violating its hate speech policy, which forbids claims "that individuals or groups are physically or mentally inferior, deficient, or diseased" based on sex or gender, among other categories. Several similar videos from repeat bad actors remain on the platform.

YouTube's hate speech policy states that the platform will "remove content promoting violence or hatred against individuals or groups based on" attributes including sexual orientation, gender identity and expression, and sex or gender.

Media Matters has identified several other videos that remain on the platform that claim being trans is a mental illness, including another one from PragerU as well as videos from the Heritage Foundation and Joe Rogan's podcast The Joe Rogan Experience.

These videos spread harmful misinformation about trans people and have earned millions of combined views.

THUMBS DOWN

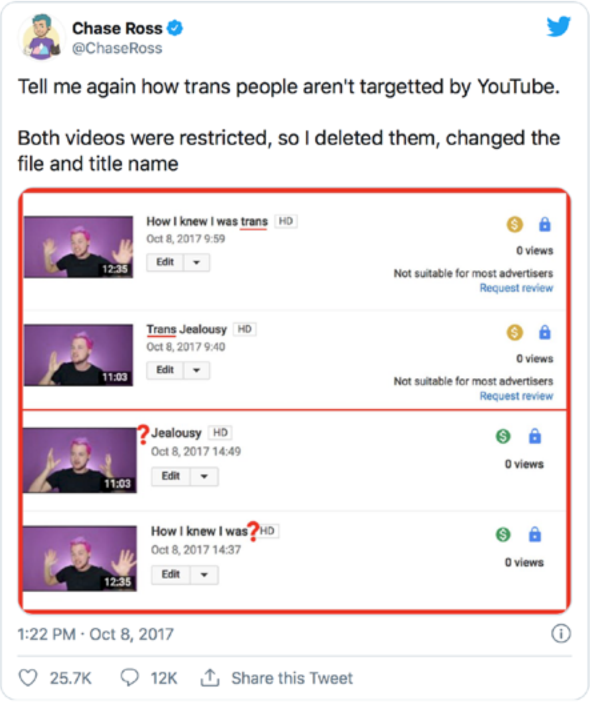
Unfair Demonetization & Removal of LGBTQ Content on YouTube

On YouTube, as with other social media platforms, legitimate LGBTQ content is frequently removed, filtered, or demonetized for alleged violation of community guidelines. But it can be nearly impossible to determine what guidelines are being invoked to determine violation because of the relative lack of transparency on the part of the platforms. Analysts are hamstrung by the refusal of social media companies to share granular data (current YouTube "transparency" reports only convey percentages in broad categories). One of YouTube's unique strategies for mitigating hate speech and content that may violate community guidelines is the ability to demonetize content creators' accounts (so that at least they are not earning profits from their hate).

In an October 2019 investigative feature, ("A Group of YouTubers is Trying to Prove the Site Systematically Demonetizes Queer Content") Vox highlights examples from an ambitious research project in which LGBTQ YouTubers analyzed how uploads featuring even such innocuous words as "gay" or "lesbian" or "LGBTQ" were systematically demonetized by YouTube's AI.[13] YouTube has had a history of similar problems over the years.[14] According to *Vox*, "the researchers found that 33 percent of the videos they tested with queer content in the titles were automatically demonetized." The LGBTQ YouTubers subsequently sued the platform claiming that YouTube was unfairly discriminating against LGBTQ content and creators. (The 84-page class-action complaint can be seen in its entirety at Court House News.)

Speaking to *Buzzfeed* about the lawsuit in August 2019, a YouTube spokesperson asserted that, "Our policies have no notion of sexual orientation or gender identity and our systems do not restrict or demonetize videos based on these factors or the inclusion of terms like 'gay' or 'transgender.'"

It was just a year prior —in June 2018—that YouTube had officially apologized to LGBTQ creators for this exact same issue. Although YouTube's apology expressed enthusiastic support of the LGBTQ community (an enthusiasm that seems genuine), the statement they made was vague and did not offer concrete commitments: "And when we hear concerns about how we're implementing our monetization policy, we take them seriously and make improvements if needed[...]. We're sorry and we want to do better [...] we are committed to working with you to get this right," it said in part.



Chase Ross
@ChaseRoss

Tell me again how trans people aren't targetted by YouTube.

Both videos were restricted, so I deleted them, changed the file and title name

How I knew I was trans HD
Oct 8, 2017 9:59
Edit
0 views
Not suitable for most advertisers
Request review

Trans Jealousy HD
Oct 8, 2017 9:40
Edit
0 views
Not suitable for most advertisers
Request review

?Jealousy HD
Oct 8, 2017 14:49
Edit
0 views

How I knew I was? HD
Oct 8, 2017 14:37
Edit
0 views

1:22 PM · Oct 8, 2017

25.7K 12K Share this Tweet

From them (Jan 8, 2021): "Chase Ross, one of the plaintiffs attached to the suit, highlighted the alleged double standard to which LGBTQ+ creators are subjected in a 2017 tweet. Ross [...] noted that when the word 'trans' was included in the titles of his videos, they were deemed 'not suitable for most advertisers.' That meant he was unable to profit from those videos, impacting the money he was able to make from his channel."

A January 2021 follow-up piece by reporter Nico Lang on *them* delves deeper into the details of the lawsuit (spoiler alert: the LGBTQ YouTubers lost) and points out that, "The mea culpa has not prevented LGBTQ+ videos from continuing to be demonetized[...]."

Lang continues: "While Magistrate Judge Virginia K. DeMarchi did not rule on the validity of what LGBTQ+ creators had experienced [...] she dismissed claims that YouTube had violated their free speech rights." The article concludes that, "YouTube and its parent company, Google, have routinely denied allegations that it purposefully discriminates against LGBTQ+ creators [...]. Although the ruling was a setback for LGBTQ+ creators, the case is not over. DeMarchi's decision allowed the plaintiffs to amend their claims that marking neutral LGBTQ+ content as 'restricted' amounts to false advertising."

GLAAD urges YouTube to release a report on the current state of the situation and to substantively assure LGBTQ users that these issues of discrimination have been remedied.

THUMBS DOWN

"It's Ma'am" — YouTube Hateful Content Reporting Case Study

A deeply disturbing example of transphobia, the "It's Ma'am" viral video, began circulating on YouTube at the end of 2018. As seen in the video, shot by a customer in line behind her, when a trans woman is misgendered by a retail clerk at a Gamestop store she angrily corrects him. In the many repostings of the video on YouTube, she is mocked for her request to be correctly gendered, "It's Ma'am." The video itself is less than two-minutes in length. In virtually every appearance of the video online (and in the subsequent memes and quotations — which became so popular they spawned a trend of transphobic "It's Ma'am" t-shirts now sold on Amazon and other e-commerce platforms) it is clear that the posting serves as a vehicle for transphobia and the mocking of trans women. GLAAD chose one example of the video on YouTube as a test case to experience the YouTube reporting system.

With 1,462,333 views as of December 22, 2020 when we first reported it, the YouTube video: "Trans gender ma'am goes off on a GameStop" (posted by Jegan Gaming on Dec 28, 2018) features the dehumanizing and malicious description, "Tranny doesn't like being called sir...lol." The overall YouTube community response to the video echoes the transphobia of the description — with more than 18,000 "thumbs up" reactions (and, thankfully, 676 thumbs downs). A skim of the 6,202 comments reveals that the most popular reply echoes the LOL of the description, "LMFAO." This response clearly underscores the intent of the video — to mock and maliciously insult and dehumanize trans women.

[13] The project was started by the YouTuber Sybreed and later automated by Sealow of Ocelot AI.

[14] In March 2017, LGBTQ YouTubers discovered that their videos were being filtered into Restricted Mode by the platform thereby hiding "hundreds of thousands [of videos] featuring LGBTQ+ content." In their April 21, 2017 response assuring that they had remedied the problem, YouTube offered a heartening reassurance that: "Restricted Mode should not filter out content belonging to individuals or groups based on certain attributes like gender, gender identity, political viewpoints, race, religion or sexual orientation."

GLAAD SMSI senior project consultant Jenni Olson attempted repeatedly to report the video but YouTube content moderators did not to remove it, despite its clear violation of YouTube community guidelines. Those guidelines expressly prohibit: "content promoting violence or hatred against individuals or groups based on any of the following attributes [...] Gender Identity and Expression." YouTube also does not allow: "content that targets an individual with prolonged or malicious insults based on intrinsic attributes, including their protected group status [...] including Gender Identity and Expression." The site's harassment and cyberbullying policy adds: "We take a harder line on content that maliciously insults someone based on their protected group status.")

First report — Dec 22 2020

1,462,333 views; Thumbs up: 18K; Thumbs down: 676; 6,202 Comments

Category of report: Hateful or abusive content > Promotes hatred or violence.

Text of report: This video is clearly being presented in a way that is hateful and dehumanizing to transgender individuals - it is in violation of your community guidelines. Note that the channel owner description for the video utilizes a well-known offensive epithet for transgender people and is mocking in its tone: "Tranny doesn't like being called sir...lol." Of the more than 6,000 comments most are extremely transphobic and hateful. This has been up for 2 years and has more than 1.4 million views.

> **YouTube auto-reply:**
>
> Thanks for reporting. If we find this content to be in violation of our Community Guidelines, we will remove it.

Second report — Dec 28 2020

Category of report: Hateful or abusive content > Abusive title or description

Text of report: This video description ("Tranny doesn't like being called sir...lol.") is hate speech and violates YouTube's community guidelines. According to GLAAD the word "tranny" is derogatory and dehumanizing. See the bottom of this page for more info: https://www.glaad.org/reference/transgender. "Defamatory: 'tranny' — These words dehumanize transgender people and should not be used." Please remove this video.

Third report — Jan 5 2021

Category of report: Hateful or abusive content > Abusive title or description

Text of report: "Tranny" is a derogatory epithet and is considered violent hate speech by GLAAD. The description of this video ("Tranny doesn't like being called sir...lol.") is in violation of YouTube Community Guidelines which clearly state that: "Hate speech is not allowed on YouTube. We remove content promoting violence or hatred against individuals or groups based on any of the following attributes: Gender Identity and Expression." Please take down this video.

Fourth report — Jan 7 2021 (at this point: 1,479,697 views)

Category of report: Hateful or abusive content > Bullying

Text of report: This description: ("Tranny doesn't like being called sir...lol.") is a malicious insult. YouTube also does not allow: "content that targets an individual with prolonged or malicious insults based on intrinsic attributes, including their protected group status (including Gender Identity and Expression)." The Harassment and cyberbullying policy says: "We take a harder line on content that maliciously insults someone based on their protected group status." Please remove this video.

As of **late-March 2021** the video is approaching 1.6 million views and has still not been removed.

GLAAD urges YouTube to look to GLAAD's new *Listing of Anti-LGBTQ Online Hate Speech* as a resource and to add the word "tranny" to its list of "malicious insults based on[...] protected group status," while remaining aware that, like many such words, it can also be used by the community as a self-identifying term and must always be evaluated in context. GLAAD further urges YouTube to hire more content moderators versed in LGBTQ issues, and to fully train all content moderators in sensitivity to anti-LGBTQ bias to enable them to recognize this kind of hateful content for what it is and to deal with it accordingly.

A Different Approach: Public Advertiser Brand-Shaming

On January 7, 2021 we tried a different approach on one other iteration of the same video, "Transgender at Gamestop goes BERSERK", which was posted on Dec 29, 2018 and as of January 7, 2021 had 448,883 views, 31,000 thumbs ups, 1,400 thumbs downs, and 8,653 comments. If there is any doubt about the viciousness of the hate that this video elicits, a few examples of user comments include ***LANGUAGE WARNING***: "We need to go back to the days of beating up perverts and weirdos. There were way less offenses committed back in the day." And: "If you put pink panties on that dick and balls, its still a dick and balls." And: "Swift kick to the nuts will remind "ma'am" what it really is." The video is monetized with pre-roll ads (meaning that, in addition to YouTube bringing in ad revenue, the YouTube user is also making money from the video and various advertisers are having their brands associated with the content). Posting via Twitter, GLAAD SMSI senior project consultant Jenni Olson publicly alerted the streaming movie platform MUBI that their ads were being run on the video:

> FYI @mubi - @YouTube is running your ads on transphobic videos that violate their own community guidelines against "malicious insults based on intrinsic attributes, including protected group status (including Gender Identity and Expression)."

There was no reply from YouTube, but MUBI replied within 20 minutes to say: "Thank you for bringing this to our attention. We would never intentionally support or align with this content. We are immediately taking steps to ensure this will not happen again."

As of March 19, 2021 (when we reported it again) the video was still up (with more than 450,000 views) and continued to be preceded by ads from such major national brands as MasterClass and Monday.com. When we checked on March 26, 2021 the video had been set to Private and is now unavailable. The archived version we created can be seen here.

Anti-LGBTQ Hate is Not a Joke

The "It's Ma'am" viral transphobia phenomenon is an emblematic example of the kinds of vicious hateful content posted on social media platforms that can be easily recognized by those sensitized to anti-LGBTQ hate as violating community guidelines, while clearly untrained platform moderators see such material as merely a harmless joke. While these determinations are not always easy to make, the conflict of interest here for YouTube is obvious: the platform leans towards *allowing* extreme and hateful content to remain and generate views and profits. While the main argument YouTube and other platforms invoke is that they believe in "free speech," the underlying reality is that platforms derive enormous profits from this *laissez faire* approach as LGBTQ people, and society as a whole, suffer the dangerous consequences of bias, hate, and violence.

What's more, because of the ways that YouTube's recommendation systems work, it is virtually guaranteed that even more extreme anti-LGBTQ content is offered up as additional viewing options to YouTube users who happen upon even one of these hateful videos.

In concluding our recommendations, we urge every individual in a position of leadership at YouTube to find ways to take meaningful action now to make the platform safe for its LGBTQ users.

INSTAGRAM

Note: Instagram is owned by Facebook. GLAAD is an organizational member of Facebook's Network of Support, an advisory coalition of LGBTQ organizations that advise on policy and product updates.

> ***"It is the duty of platform companies to curate content on contentious topics so that their systems do not amplify hate or make it profitable. Tech companies that refuse to adapt for the culture will become obsolete."***
>
> **— Joan Donovan, PhD, Research Director, Shorenstein Center**

As is true of the other large social networks, Instagram is a divided landscape. As documented by researchers and journalists, there are powerful platform-driven algorithms that control what we see—and which create silos of experience that eliminate other perspectives. It is of course incredibly ironic that this actual truth sounds so much like a conspiracy theory that one is inclined to mistrust its veracity. Like its parent company Facebook, Instagram is in urgent need of product improvements in many areas (see our GLAAD SMSI Articles & Reports Appendix for more context).[15]

The 2020 Ranking Digital Rights Corporate Accountability Index offers an in-depth evaluation of Instagram parent company Facebook's overall performance on numerous metrics, including relevant social media safety indicators. There are many, many changes platforms can implement to make their products safer for LGBTQ users.

Below are some of our specific recommendations for Instagram. We urge Instagram to also attend to the general recommendations and Platform Responsibility Checklist in the first part of this report, including items related to: Protection of LGBTQ users in community guidelines; Algorithmic bias and bias in AI; Privacy and outing; Promoting civil discourse; and more.

RECOMMENDATIONS

Content Moderation

Instagram's hate speech policies are identical to those of its parent company, Facebook, but there are additional overlays of policy related to the unique culture of Instagram. The platform's specific community guidelines link to Facebook's hate speech policy, but offer the following additional framing:

Respect other members of the Instagram community.

> We want to foster a positive, diverse community. We remove content that contains credible threats or hate speech, content that targets private individuals to degrade or shame them, personal information meant to blackmail or harass someone, and repeated unwanted messages...
>
> It's never OK to encourage violence or attack anyone based on their race, ethnicity, national origin, sex, gender, gender identity, sexual orientation, religious affiliation, disabilities, or diseases. When hate speech is being shared to challenge it or to raise awareness, we may allow it. In those instances, we ask that you express your intent clearly.

These guidelines are expansive, but the company must also achieve considerable improvement in the enforcement of these policies to make the platform safe for LGBTQ users. Some areas to be addressed in the broad realm of **content moderation**, and some recommendations for improvement, include the following items below.

Protection of LGBTQ Users in Community Guidelines

See numerous items in our "Recommendations for All Platforms" above.

Disinformation/Misinformation

RECOMMENDATION:
Implement Substantive Solutions to Disinformation/Misinformation

Many instances of anti-LGBTQ content and conduct fall under the heading of dis/misinformation.

Instagram should make greater use of mitigation tools such as labeling certain kinds of content or pointing users to other trusted sources. Instagram does have a False Information option for users to report dis/misinformation on the platform.

Transparency and Accountability

RECOMMENDATION:
Improve the Process of Reporting, and of Appealing Reports

Instagram should improve the system of user-reporting (of content, comments, and accounts) to provide greater transparency. Instagram messages users in the activity feed with "support request" updates on reported items (either the content is removed or it is not). The process for appealing reports must also be significantly improved: see this December 2020 *Los Angeles Blade* story about a legitimate LGBTQ account (gay couple Matthew Olshefski & Paul Castle) being reported by right-wing trolls and disabled by Instagram (thankfully, it was subsequently restored). Also, see the sidebar case study example of our attempt to report hate speech on Instagram.

Incorrect Blocking of LGBTQ Content

RECOMMENDATION:
Use Human Moderators

Instagram should increase use of qualified human moderators to more accurately interpret legitimate use of LGBTQ terms and to distinguish legitimate accounts and posters from trolls and bad actors. The platform should make corresponding improvements to AI systems and evaluation of user reported content. Also see below item on self-expression.

Algorithms

RECOMMENDATION:
Refine Algorithms to Reduce Hate, Not Spread It

In addition to the notes on algorithms made in the general recommendations above, Instagram must prioritize improved practices and systems to reduce anti-LGBTQ hate and extremist content — including adjusting both their current content moderation systems and their algorithms which appear to escalate the dissemination of such content.

Self-expression & LGBTQ inclusion

RECOMMENDATION:
Update Community Guidelines to Reflect Context

With regard to LGBTQ self-expression, Instagram incorporates some degree of nuance about hate speech in the site's community guidelines page: ("When hate speech is being shared to challenge it or to raise awareness, we may allow it."); GLAAD encourages Instagram to also add to this page on their site the following more robust language from the Facebook community guidelines policy, which reflects the understanding that individuals who belong to protected groups may use self-referring terminology which might otherwise be considered offensive ("In some cases, words or terms that might otherwise violate our standards are used self-referentially or in an empowering way.") Instagram's "Transparency report on Hate Speech" incorporates the following additional language to accommodate further nuances: "We do not allow hate speech on Instagram. We define hate speech as violent or dehumanizing speech, statements of inferiority, calls for exclusion or segregation based on protected characteristics, or slurs. These characteristics include race, ethnicity, national origin, religious affiliation, sexual orientation, caste, sex, gender, gender identity, and serious disability or disease. *When the intent is clear, we may allow people to share someone else's hate speech content to raise awareness or discuss whether the speech is appropriate to use, to use slurs self-referentially in an effort to reclaim the term, or for other similar reasons.*" GLAAD also reminds the platform of the need for continued diligence in the implementation and enforcement of these policies, lest legitimate LGBTQ content be over-policed or unfairly removed. GLAAD also strongly urges Instagram to devote resources to gathering and releasing data on the current state of LGBTQ self-expression on the platform. Additional research is needed to determine all the ways that LGBTQ users are currently being impacted in this area.

[15] SMSI — Articles & Reports Appendix spreadsheet

LGBTQ hiring, inclusion, & leadership

RECOMMENDATION:
Continue Commitment to Diverse Hiring

GLAAD strongly urges Instagram to continue to diversify its hiring of LGBTQ employees, especially in positions in the engineering and product teams to shape the platform at its highest level. Instagram should consider following the lead of Google/YouTube, which solicits voluntary demographic data from LGBTQ employees on how they self-identify. In September 2020 Instagram announced the creation of a new Equity team and created a new Director of Diversity and Inclusion position, which was filled in late 2020. Diverse workforces are essential in serving the needs of a diverse array of users. It is also essential to hire LGBTQ content moderators and to train all content moderators to understand the needs of LGBTQ users.

This is only a partial list. Please see the Platform Responsibility Checklist and general recommendations section above.

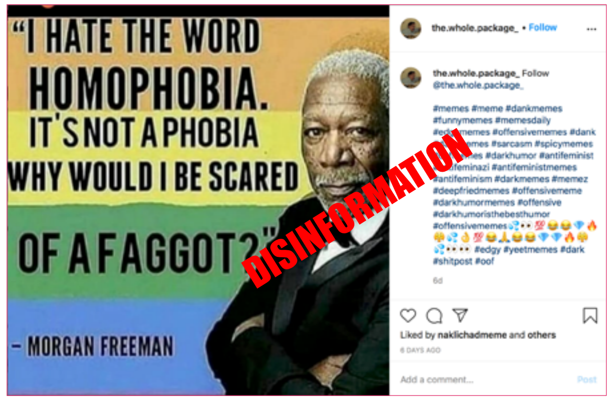
Instagram Hate Speech Reporting Case Study

THUMBS UP

1) Reported post (Jan 11, 2021):

At the time we reported the post, from account The Whole Package, it had been up for 6 days.

Meme: "I hate the word homophobia. It's not a phobia. Why would I be scared of a faggot?" — Morgan Freeman



the.whole.package_ • Follow

the.whole.package_ Follow @the.whole.package_

#memes #meme #dankmemes #funnymemes #memesdaily #edgymemes #offensivememes #dank [illegible] #sarcasm #spicymemes [illegible] #darkhumor #antifeminist #antifeminazi #antifeministmemes #antifeminism #darkmemes #memez #deepfriedmemes #offensivememe #darkhumormemes #offensive #darkhumoristhebesthumor #offensivememes💦👀💯😂😂💎🔥😤💦👌💯😂🙏😂😂💎💎🔥😤💦👀👀 #edgy #yeetmemes #dark #shitpost #oof

6d

Liked by naklichadmeme and others

6 DAYS AGO

Add a comment... Post

Note: This meme is actually a trolling riff on a previously existing fake Morgan Freeman quote: "I hate the word homophobia. It's not a phobia. You are not scared; you are an asshole." Which he also didn't actually say.

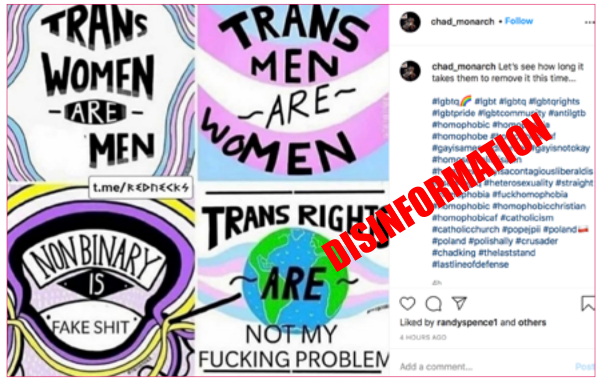
Subsequent status update received from Instagram:

"This post is no longer available." [It seems likely that the post was reported by someone else and removed in the interim of our report being reviewed].

THUMBS DOWN

2) Reported post (Jan 11, 2021):

At the time we reported this post, from the account "Chad Monarch," it had been up for 4 hours.



chad_monarch • Follow

chad_monarch Let's see how long it takes them to remove it this time...

#lgbtq🏳️‍🌈 #lgbt #lgbtq #lgbtqrights #lgbtpride #lgbtcommunity #antilgtb #homophobic #homophobia #homophobe #homophobicaf #gayisamentaldisorder #gayisnotokay [illegible] #homosexualityisacontagiousliberaldis [illegible] #heterosexuality #straight #homophobia #fuckhomophobia #homophobic #homophobicchristian #homophobicaf #catholicism #catholicchurch #popejpii #poland🇵🇱 #poland #polishally #crusader #chadking #thelaststand #lastlineofdefense

4h

Liked by randyspence1 and others

4 HOURS AGO

Add a comment... Post

Meme: "Trans Women are Men; Trans Men are Women; Non-Binary is Fake Shit; Trans Rights are Not My Fucking Problem."

Note: The meme includes a URL which is the address of the "Rednecks" Telegram profile page which features ***LANGUAGE WARNING*** the following profile description: "Black humor meme and others nationalist things. If you are a n***er, k*ke, LGBTHIV+ supporter, feminist, politically correct, globalist and communist... this page is not for you." (The asterisks are not in the original).

That Telegram page offers a cesspool of extremist hate memes combining vicious homophobia and transphobia with vicious racism, anti-Semitism, and conspiracy theories. This is an emblematic example of a dehumanizing "humorous" meme devolving to extremist hate. As the ADL's "Pyramid of Hate" illustrates, this dehumanization ultimately may lead to violence and abuse offline.

It is vitally important to see how this kind of online hate is not just offensive but dangerous.

Note: The Instagram post also includes the defiant statement: "Let's see how long it takes them to remove it this time [...]" which itself testifies to the inadequacies of the platform's current systems for mitigating hate. The statement is accompanied by a barrage of hateful hashtags (such as: #homophobicaf #gayisamentaldisorder #gayisnotokay #homosexualityisacontagiousliberaldisease) which offer access to yet more hate; as well as an array of standard LGBTQ hash tags (like #lgbtq #lgbtqrights #lgbtpride) that maliciously troll legitimate LGBTQ hashtags to fill them with hateful content.

Subsequent status update received from Instagram on Jan 13 2021: "Report Reviewed. You anonymously reported chad_monarch's photo for hate speech or symbols. **We didn't remove chad_monarch's photo. We found that this photo likely doesn't go against our Community Guidelines.** If you think we made a mistake, please report it again. Because Instagram is a global community, we understand that people may express themselves differently. We'll use your feedback to make this experience better for everyone. If you don't want to see chad_monarch on Instagram, you can unfollow, mute or block them to hide their posts and comments from your feed."

Note: Subsequent to this, chad_monarch's account settings were switched to private so the account is no longer accessible. The archived version we created is here.

Instagram Community Standards Enforcement Report, February 2021

According to their reports for 2020, Instagram took action on 3.2 million pieces of hate speech in Q2 2020 (quadruple the quantity of hate speech posts — 800,000 — in Q1). That quantity of actioned content items then doubled again from Q2 to Q3 (hitting 6.5 million) and rose to 6.6 million for Q4. The Q4 report also cites that there were 5 million pieces of bullying and harassment content actioned (this number rose steadily from 1.5 million in Q1 to 2.3 in Q2 to 2.6 million in Q3). The platform attributes the increases in actioned content to improvements in their proactive detection technology. GLAAD calls on Instagram to offer greater transparency as to the exact nature of the hate speech/bullying and harassment represented in these reports (anti-LGBTQ, anti-Semitism, racism, etc.). Beginning with their Q3 2020 reports, Facebook and Instagram released other data about content appeals and restoration as well as prevalence of hate speech on Facebook (though not on Instagram). While this greater transparency is in line with the Santa Clara Principles On Transparency and Accountability in Content Moderation, there is still a very long way to go towards fully meeting those best practices. As is true of other platforms, Instagram should share disaggregated data to enable researchers clear visibility into anti-LGBTQ activity on the platform. This November 20, 2020 ADL blog post ("Facebook's Transparency Reporting Continues to Obscure the Scope and Impact of Hate Speech") offers a very useful critique and analysis. GLAAD urges Facebook and Instagram to respond—swiftly and completely—to the ADL's requests for action, the conclusion of which we reprint here:

> Finally, as ADL has long demanded, Facebook needs to report on the prevalence of hate speech targeting specific communities, the experiences that distinct groups are having on its platform and the numbers for the different kinds of hate being spread. For example, how many antisemitic, anti-Black, anti-Muslim and anti-LGBTQ+ pieces of content required actioning? Without specifying these numbers and the types of content attacking each vulnerable group, it is difficult for civil rights groups to propose solutions to these problems. Facebook can follow the example set by Reddit by conducting a study on hate and abuse on its platform and making its findings public. The company should also conduct another independent audit, specifically focused on its lack of transparency.
>
> Greater transparency and active data collection around online hate speech should be accompanied by evidence-based policies and enforcement mechanisms. To show they are taking real steps to reduce hate speech, platforms must try to understand the scope of the problem by collecting the relevant data and using rigorous research methods. Failure to do so will result in vulnerable groups continuing to be at the mercy of toxic users on social media.

In concluding our recommendations, we urge every individual in a position of leadership at Instagram to find ways to take meaningful action now to make the platform safe for its LGBTQ users.

TIKTOK

"Lies laced with anger and hate spread faster and further than facts."

— Maria Ressa, Journalist & CEO, Rappler (GLAAD SMSI advisory committee member)

Like other platforms TikTok has stepped forward at various junctures to make improvements to product safety for LGBTQ users and other vulnerable groups, including an announced crackdown on hate speech in October 2020.

Comparing the platform to Facebook and YouTube, the social media watchdog organization Sleeping Giants observed in an October 2020 Guardian article about these policy changes, "The real test, as always, will be enforcement." While the platform has made many meaningful product and policy updates to support LGBTQ users and our issues, TikTok has also made news for wrongfully suppressing LGBTQ users and content (see below).

There are many, many changes the platform can implement to make their product safer for LGBTQ users. Below are some of our specific recommendations for TikTok. We urge TikTok to also attend to the general recommendations and Platform Responsibility Checklist in the first part of this report, including items related to: Protection of LGBTQ users in community guidelines; Algorithmic bias and bias in AI; Privacy and outing; Promoting civil discourse; and more.

RECOMMENDATIONS

Content Moderation

TikTok's Community Guidelines concerning "Hateful behavior" specifically prohibit attacks on the basis of "protected attributes," "Slurs," and "Hateful ideology." The guidelines state that: "TikTok is a diverse and inclusive community that has no tolerance for discrimination. We do not permit content that contains hate speech or involves hateful behavior and we remove it from our platform. We suspend or ban accounts that engage in hate speech violations or which are associated with hate speech off the TikTok platform[...]. We define hate speech or behavior as content that attacks, threatens, incites violence against, or otherwise dehumanizes an individual or a group on the basis of the following protected attributes: Race, Ethnicity, National origin, Religion, Caste, Sexual orientation, Sex, Gender, Gender identity, Serious disease, Disability, Immigration status."

Having come under pressure similar to other platforms around hate speech, disinformation and other issues, TikTok joined the voluntary EU Code of Conduct on Countering Illegal Hate Speech Online in September 2020, and implemented a European Safety Advisory Council, partly tasked with addressing bias and hate, in March 2021 (a US Content Advisory Council was created in March 2020). It seems that much of this recent planning is in direct response to the EU's Digital Services Act (see sidebar). As reporter Natasha Lomas phrases it, in a March 2021 *TechCrunch* article: "Ahead of that oversight regime coming in, platforms have increased incentive to up their outreach to civil society in Europe so they're in a better position to skate to where the puck is headed."

As TikTok continues to make improvements to Community Guidelines, the company must also continue to improve the enforcement of these policies to make the platform safe for LGBTQ users. Some areas to be addressed in the broad realm of **content moderation**, and some recommendations for improvement, include the following items below.

Protection of LGBTQ Users in Community Guidelines

See numerous items in our "Recommendations for All Platforms" above.

Disinformation/Misinformation

RECOMMENDATION:
Continue to Explore and Implement Tools for Mitigation of Disinformation and Misinformation

Many instances of anti-LGBTQ content and conduct fall under the heading of dis/misinformation. In February 2021, TikTok added various friction policies and functionality to slow the spread of misinformation including not promoting videos in the main "For You" feed if they have been flagged by TikTok's fact-checking partners as unverified content. If a user attempts to share an unverified video they are shown a prompt: "Are you sure you want to share this video? This video was flagged for unverified content." According to TikTok, "viewers decreased the rate at which they shared videos by 24%, while likes on such unsubstantiated content also decreased by 7%." In March 2021, TikTok implemented a similar functionality, urging users to pause before commenting if the post uses words that the platform's AI recognizes as possibly unkind or in violation of community guidelines. TikTok does also have a robust Misleading Information option for users to report various categories of dis/misinformation. The platform implemented a 2020 "Be Informed" media literacy PSA series about, among other things, how to identify (and refrain from sharing) misinformation. TikTok should continue to explore and implement even greater utilization of dis/misinformation mitigation tools.

Transparency & Accountability

RECOMMENDATION:
Improve Transparency, Accountability, and the User-Reporting Process

Also key to helping fight anti-LGBTQ content and conduct, the platform could improve the system of user-reporting (of content, comments, and accounts). The TikTok InBox offers centralized communications with regard to user-reported posts. TikTok should also provide greater transparency on how decisions are made and what recourse users have when their posts have been flagged. For further details on best practice recommendations see the Santa Clara Principles On Transparency and Accountability in Content Moderation and their proposal of "initial steps that companies engaged in content moderation should take to provide meaningful due process to impacted speakers and better ensure that the enforcement of their content guidelines is fair, unbiased, proportional, and respectful of users' rights." In March 2020, TikTok announced the forthcoming launch of a Transparency Center (which continues to be delayed due to Covid) and released their first Transparency Report. GLAAD looks forward to continuing efforts at greater transparency from TikTok.

TikTok currently offers only a basic notification for users issuing reports: "Thank you for helping to keep our community safe. You will receive a notification when the review is complete." Subsequent to reporting a video that reported video is hidden from the user, and the user is given a message that "We'll show you fewer videos like this." For the user whose video is being reported a notice is sent alerting them to the flag and offering an appeals process. We urge TikTok to continue to provide greater levels of transparency and granular data on violated policies and to continue to make improvements — including providing a more robust experience of transparency to users issuing reports.

Incorrect Blocking of LGBTQ Content

RECOMMENDATION:
Use Human Moderators

TikTok should increase use of qualified human moderators to more accurately interpret legitimate use of LGBTQ terms and to distinguish legitimate accounts and posters from trolls and bad actors. The company should also make corresponding improvements to AI systems. Also see below item on self-expression.

RECOMMENDATION:
Don't De-Platform Legitimate Users

One of the most disturbing types of anti-LGBTQ conduct on social media is the well-documented practice of trolls reporting legitimate LGBTQ users in an effort to have their accounts de-platformed.

The reason for account or content removal is often not conveyed to the user. See for example the *Los Angeles Blade* story about the case of Rosalynne Montoya, a Latina trans woman whose TikTok account was taken down after being reported by trolls, though it had not actually violated any guidelines (and which has subsequently been restored). Montoya's Change.org petition to "Change TikTok's Community Guidelines Algorithm" had more than 17,000 signatures as of mid-March, 2021. TikTok should provide greater transparency on how decisions are made and what recourse users have. See the Santa Clara Principles On Transparency and Accountability in Content Moderation for further detail on best practice recommendations.

Algorithms

RECOMMENDATION:
Stay Vigilant to Protect User Safety

In addition to the notes on algorithms made in the general recommendations above, TikTok must prioritize improved practices and systems to reduce anti-LGBTQ hate and extremist content. It is notable that TikTok quickly implements measures to reduce the posting and spread of anti-LGBTQ hate speech and dis/misinformation. One recent high-profile example of this is TikTok's responsiveness in addressing the emergence of the transphobic trolling #SuperStraight hashtag in March 2021 (including deplatforming the initiator of the trend and shadow banning the hashtag).

Self-expression & LGBTQ inclusion

RECOMMENDATION:
Take Context Into Account

With regard to LGBTQ self-expression, it is good to see that TikTok's policies reflect the understanding that individuals who belong to protected groups may use self-referring terminology which might otherwise be considered offensive. From the TikTok Community Guidelines: "Slurs are defined as derogatory terms that are intended to disparage an ethnicity, race, or any other protected attributes listed above. To minimize the spread of egregiously offensive terms, we remove all slurs from our platform, *unless the terms are reappropriated, used self-referentially (e.g., in a song), or do not disparage."* GLAAD also reminds the platform of the need for continued diligence in the implementation and enforcement of these policies, lest legitimate LGBTQ content be over-policed or unfairly removed. GLAAD also strongly urges TikTok to devote resources to gathering and releasing data on the current state of LGBTQ self-expression on the platform. Additional research is needed to determine all the ways that LGBTQ users are currently being impacted in this area.

LGBTQ hiring, inclusion & leadership

RECOMMENDATION:
Continue to Diversify Hiring

GLAAD strongly urges TikTok to continue to diversify its hiring of LGBTQ employees, especially in positions in the engineering and product teams to shape the platform at its highest level. TikTok should consider following the lead of Google/YouTube which solicits voluntary demographic data from LGBTQ employees on how they self-identify. Diverse workforces are essential in serving the needs of a diverse array of users. TikTok's Education and Philanthropy team has launched several impactful and proactive LGBTQ public education and awareness campaigns with GLAAD and other LGBTQ organizations, and the team amplifies voices of LGBTQ creators. In addition to departments related to content creation and social impact, it is also essential to hire LGBTQ content moderators and to train all content moderators to understand the needs of LGBTQ users.

This is only a partial list. Please see the Platform Responsibility Checklist and general recommendations section above.

THUMBS UP

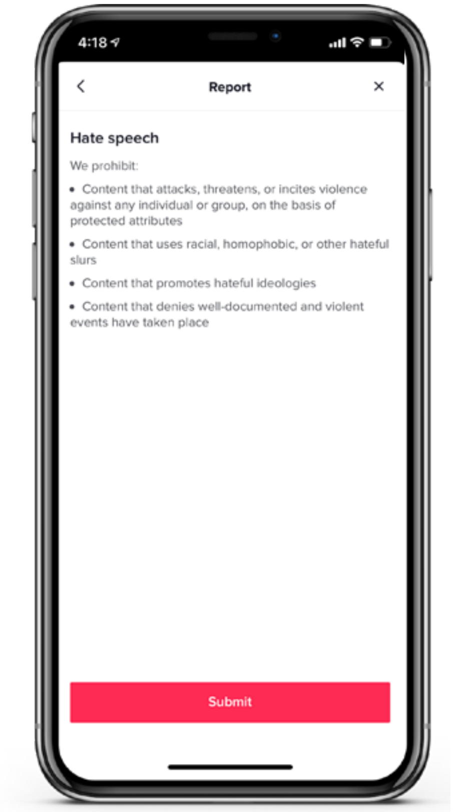
Categorization of Conversion Therapy as Hurtful & Hateful

While broad policies against hate have been adopted by all of the major social media platforms it is also necessary for companies to come forward with more specific positions on some of the unique ways in which LGBTQ people are targeted. It is good to see that TikTok, in an October 2020 policy statement, "Countering Hate on TikTok," has specifically denounced content promoting conversion therapy: "We're also removing content that is hurtful to the LGBTQ+ community by removing hateful ideas, including content that promotes conversion therapy and the idea that no one is born LGBTQ+."

THUMBS DOWN

Opportunity for Improvement: Incomplete Hate Speech Description in User Reporting Process

The alert that a user receives when reporting hate speech only offers an incomplete bullet point summary of what things are prohibited by the TikTok Community Guidelines. This current iteration does not instill confidence that content moderators are utilizing the full Community Guidelines to evaluate reports. TikTok should at the very least include mention here of the full Community Guidelines (and a link to them if possible). An additional suggestion would be to expand the language of item 2 to include transphobic slurs as well as adding a more complete list (anti-Semitic, Islamophobic, sexist, xenophobic, etc.). This is also a good place to remind TikTok that the Santa Clara Principles On Transparency and Accountability in Content Moderation urge companies to offer full transparency on the content moderation process.



Hate speech

We prohibit:

- Content that attacks, threatens, or incites violence against any individual or group, on the basis of protected attributes
- Content that uses racial, homophobic, or other hateful slurs
- Content that promotes hateful ideologies
- Content that denies well-documented and violent events have taken place

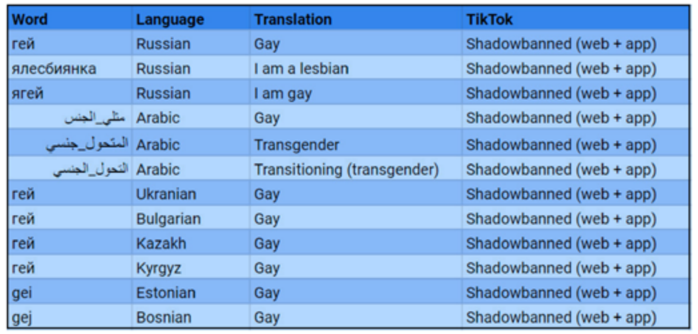
TikTok alert from user-reporting of hate speech process.

THUMBS DOWN

Shadow Banning LGBTQ hashtags

As noted in the general recommendations with regard to its 2019 suppression of LGBTQ accounts, TikTok has had a history of problematic policies and practices with regard to LGBTQ users. As recently as September 2020, reports indicate that TikTok has been censoring LGBTQ material on the platform by "shadow banning" certain legitimate LGBTQ hashtags.[16] In a September 8, 2020 Quartz article on a report from the Australian Strategic Policy Initiative (ASPI), reporter Jane Li offers this alarming summary: "Try searching for hashtags related to LGBT issues in countries like Russia, Bosnia, and Jordan on TikTok and you might find no results, even if you were able to see it on a friend's post. That's because the app is now shadow banning such hashtags, including the word "gay" in languages including Russian (гей), Arabic (الجنس_مثلي), and Bosnian (gej)[...] Hashtags like "#transgender" in Arabic (#جنسي المتحول) and #I am a gay/lesbian" in Russian (#ягей/#ялесбиянка) are also suppressed."

Figure 4: List of shadow-banned LGBTQ+ hashtags on TikTok



Word	Language	Translation	TikTok
гей	Russian	Gay	Shadowbanned (web + app)
ялесбиянка	Russian	I am a lesbian	Shadowbanned (web + app)
ягей	Russian	I am gay	Shadowbanned (web + app)
مثلي_الجنس	Arabic	Gay	Shadowbanned (web + app)
المتحول_جنسي	Arabic	Transgender	Shadowbanned (web + app)
التحول_الجنسي	Arabic	Transitioning (transgender)	Shadowbanned (web + app)
гей	Ukranian	Gay	Shadowbanned (web + app)
гей	Bulgarian	Gay	Shadowbanned (web + app)
гей	Kazakh	Gay	Shadowbanned (web + app)
гей	Kyrgyz	Gay	Shadowbanned (web + app)
gei	Estonian	Gay	Shadowbanned (web + app)
gej	Bosnian	Gay	Shadowbanned (web + app)

Source: ASPI ICPC.

The September 2020 ASPI report ("TikTok and WeChat: Curating and Controlling Global Information Flows") offers a lengthy analysis of the problem and includes the following response from a TikTok spokesperson:

> As part of our localised approach to moderation, some terms that the ASPI provided were partially restricted due to relevant local laws. Other terms were restricted because they were primarily used when looking for pornographic content [...]. We also identified, and fixed, an issue where some compound phrases in Arabic were being incorrectly moderated because part of the phrase may relate to pornography. Separately, a couple of English phrases were incorrectly moderated, and we have resolved the error. We are currently conducting a review of those terms that were moderated in error and will look for ways to improve our processes to avoid similar issues in the future. In addition, we want to be crystal clear that TikTok strongly supports our LGBTQ creators around the world and is proud that LGBTQ content is among the most popular category [sic] on the platform with billions of views.

The ASPI report further clarifies that: "Our research shows, for example, that hashtags related to LGBTQ+ issues are suppressed on the platform in at least 8 languages. This blunt approach to censorship affects not only citizens of a particular country, but all users speaking those languages, no matter where in the world they live." (See ASPI chart).

GLAAD looks forward to further information and action from TikTok towards resolving these anti-LGBTQ shadowbans.

[16] Note: Shadow banning is when a platform restricts visibility or suppresses content from being seen by other users without alerting the user that their post is being banned.

TikTok Community Guidelines Enforcement Data — July-December 2020

The following numbers are drawn from the TikTok Community Guidelines report for July–December 2020 During this period, TikTok reports that: "**89,132,938** videos were removed globally for violating our Community Guidelines or Terms of Service, which is less than 1% of all videos uploaded on TikTok." TikTok's Community Guidelines report does not include specific data on anti-LGBTQ content removals (as is true of other platforms, TikTok should share disaggregated data).

TikTok offers the following comments on the two most relevant headings in which anti-LGBTQ hate would be most likely to be categorized:

Harassment and bullying: 6.6%

"We believe in an inclusive community and individualized expression without fear of abuse and do not tolerate members of our community being shamed, bullied, or harassed. Of the videos we removed, **6.6%** violated this policy, which is up from **2.5%** in the first half of 2020. This increase reflects adjustments to policies around sexual harassment, threats of hacking, and targets of bullying statements, which are now more comprehensive. Additionally we saw modest improvements in our abilities to detect harassment or bullying proactively which still *remains a challenge with linguistic and cultural nuances."*

Hateful behavior: 2%

"TikTok is a diverse and inclusive community that has no tolerance for hateful behavior. Last year we changed this policy from 'hate speech' to its current name 'hateful behavior' to take a more comprehensive approach to combating hateful ideologies and off-platform activities. As a result, **2%** of the videos we removed violated this policy, up from **.8%** in the first half of 2020. We have systems to detect hateful symbols, like flags and icons, but *hate speech remains a challenge to proactively detect and we continue to make investments to improve."*

As noted above, TikTok should prioritize a variety of mitigation strategies around anti-LGBTQ hate speech including adding additional terms and phrases to AI scripts (in all languages, not just in English).

In concluding our recommendations, we urge every individual in a position of leadership at TikTok to find ways to take meaningful action now to make the platform safe for its LGBTQ users.

IN CONCLUSION

GLAAD President and CEO Sarah Kate Ellis wrote in her Summer 2020 Black Lives Matter Pride statement, "There can be no Pride if it is not intersectional." In the creation of this report, the path of researching anti-LGBTQ hate online was strewn with extraordinary volumes of vicious racism, extreme anti-Semitism and anti-Muslim hate, shocking misogyny, dehumanization of people with disabilities, and all varieties of xenophobic ignorance and intolerance — more often than not, all of it blended together in the same posts, comments, and accounts. These intersectional flaws in human character illuminate all the more clearly the need for our intersectional social justice movements. **As it cries out for equity, inclusion, and justice for LGBTQ people, may this report be one more voice in the chorus demanding we achieve justice for *all*.**

Addressing the many failures and problematic aspects of social media platforms will require a complex array of approaches. This report offers suggestions and recommendations for Facebook, Twitter, YouTube, Instagram, and TikTok — and to all platforms. We call upon the leadership of these companies to take immediate action, to implement these urgently needed changes in their products and policies and to prioritize researching new and different ideas and solutions.

We appeal to their sense of responsibility both to their customers and to our society as a whole.

Even more emphatically, we urge our policy makers in Washington to prioritize the admittedly long and complex process of finding new approaches — including creating regulatory oversight that will require these companies to be accountable.

Social media companies have had years, decades even, to demonstrate responsible curation and moderation of content. But they have not risen to the challenge, choosing to prioritize profit over public safety. This is a reality of corporate America, and not a surprising one. The EPA, FDA, SEC, OSHA came into existence for these very reasons. Creating guidelines and oversight to ensure the public health and safety of the American people is not a radical idea: it is a reasonable, commonsense solution. Knowing also that it is the nature of any industry to want to avoid external regulation (or to solicit regulation in forms they can manipulate), we would do well to remember that the smokescreen of rhetoric about neutrality and freedom of speech exuded from Silicon Valley represents self-serving false arguments designed to maintain the status quo. In an illuminating 2017 article about this ongoing platform resistance, "Why Media Companies Insist They're Not Media Companies, Why They're Wrong, and Why it Matters," researchers Philip Napoli and Robyn Caplan point towards a simple idea of new, "norms and governance structures that better serve the public interest." The time for serving the public interest is now.

Acknowledgements

GLAAD is grateful to the many organizations and individuals doing this important work. We especially want to acknowledge our advisory committee: Kara Swisher, contributing writer and host of the 'Sway' podcast at *The New York Times*; co-host of Pivot podcast at New York Media; Maria Ressa, Journalist & CEO, Rappler; Brandi Collins-Dexter, Senior Fellow, Color of Change & Visiting Fellow, Shorenstein Center; Liz Fong-Jones, Principal Developer Advocate for SRE & Observability, Honeycomb; Dr. Sarah T. Roberts, Co-Director, UCLA Center for Critical Internet Inquiry; Marlena Wisniak, Co-Director, Taraaz; Lucy Bernholz, Director, Digital Civil Society Lab at Stanford University; Leigh Honeywell, CEO and Co-Founder, Tall Poppy; Tom Rielly, founder, TED Fellows program & founder, PlanetOut.com; Jenni Olson, co-founder PlanetOut.com & Senior Project Consultant; Rich Ferraro, GLAAD Chief Communications Officer.

In addition to the advisory committee, we are grateful to: the amazing team at the Shorenstein Center (in particular Brian Friedberg and Joan Donovan), the ADL (Mark Pitcavage, Mike Salamon and colleagues), Media Matters (Brennan Suen and Rachel Tardiff), and Color of Change. Thank you also to David Hornik, Christian Williams, Mary Gray; and to copy-editor Lisa Webster and Rodolfo Mustafé, Dustin Hood, and Abdool Corlette for designing the report. Thanks also to GLAAD staff including Anthony Shallenberger, Georgia Davis, Louise Prollamante and Bill McDermott. The work of countless journalists, researchers, activists, and others continues to drive change forward. We are indebted to them all.

This report would not have been possible without the support of: Craig Newmark Philanthropies and the Gill Foundation. We are also grateful to Kara Swisher and to the Stanford Center on Philanthropy and Civil Society for early support of this project.

Note from GLAAD: On the Firewall Between Financial Sponsorship & Our Advocacy

Several of the companies that own products and platforms listed in this report, including Facebook, TikTok, and Google, are current financial sponsors of GLAAD, a 501(c)3 non-profit. A firewall exists between GLAAD's advocacy work and GLAAD's sponsorships and fundraising. As part of our media advocacy and work as a media watchdog, GLAAD has and will continue to publicly call attention to issues that are barriers to LGBTQ safety, as well as barriers to fair and accurate LGBTQ content and coverage—including issues originating from companies that are current financial sponsors.

glaad

GLAAD.org

@GLAAD