THE FUTURE OF FREE SPEECH

REBUILDING THE BULWARK OF LIBERTY

Scope Creep: An Assessment of 8 Social Media Platforms' Hate Speech Policies

Jacob Mchangama, Abby Fanlo and Natalie Alkiviadou

"Scope Creep: An Assessment of 8 Social Media Platforms' Hate Speech Policies"

WHO WE ARE

Justitia
Founded in August 2014, Justitia is Denmark's first judicial think tank. Justitia aims to promote the rule of law and fundamental human rights and freedom rights both within Denmark and abroad by educating and influencing policy experts, decision-makers, and the public. In so doing, Justitia offers legal insight and analysis on a range of contemporary issues.

The Future of Free Speech Project
The Future of Free Speech is a collaboration between the global judicial think tank Justitia, Vanderbilt University, and researchers from Aarhus University's Department of Political Science.

The project is sponsored by:

Table of Contents

Executive Summary

At the turn of the 21st century, academics, civil society organizations, and governments hailed the promise of the Internet to eliminate any centralized control over speech. A few short decades later, however, this tech utopianism has disappeared. Dominant social media platforms have become de facto gatekeepers of global information and communication flows, giving private entities the ability to moderate the speech of billions of people. As social media companies gained this power, panic about the virality and volume of objectionable content online grew. The internet was painted as a sort of "Wild West" of toxic speech, resulting in monetary, reputational, and regulatory pressure on platforms to remove broad categories of content - including hate speech. Governments around the world also began exploring ways to intervene in platforms' moderation practices, including by requiring platforms to remove certain types of specified content.

Despite these developments, however, to date, there has been no cross-platform, cross-temporal, systematic analysis of the way that platforms treat hate speech. What categories of content do platforms' hate speech policies cover? How have platforms' hate speech policies changed since their initial inception? Do platform hate speech policies align with international human rights law, given many of the major social media companies have publicly committed to respect human rights, under the United Nations (UN) Guiding Principles on Business and Human Rights? This report seeks to answer those questions. To do so, we collect original data on the hate speech policies of eight major platforms since their founding: Facebook, Instagram, Reddit, Snapchat, TikTok, Tumblr, Twitter, and YouTube. We then analyze how the policies have changed in scope over time, both within each individual platform and across all eight, and the extent to which they accord with Articles 19 and 20(2) of the International Covenant on Civil and Political Rights (ICCPR).

The results demonstrate a substantial increase in the scope of the platforms' hate speech policies over time, both in the content and the protected characteristics covered. Platforms have gone from prohibiting the promotion of hatred or racist speech, in the mid-aughts and early 2010s, to introducing prohibitions on a long list of potential forms of hate speech, including harmful stereotypes, conspiracy theories, and curses targeting protected groups, over the past several years. In addition, the average number of protected characteristics listed in platform policies has more than doubled since 2010, with platforms protecting identities as wide-ranging as caste, pregnancy, veteran status, and victims of a major event. These developments do not align with the international human rights standards that most of the analysed platforms, with the exception of Tumblr and Reddit, have committed to respect under the UN Guiding Principles for Business and Human Rights. In particular, the scope creep of platforms' hate speech policies goes far beyond the mandatory prohibition on hatred in Article 20(2) of the ICCPR. Moreover, the often vague nature of many of the platforms' policies falls afoul of the requirement that restrictions on

freedom of expression and access to information comply with the strict requirement of "legality" in ICCPR Article 19(3).

While current restrictions on researcher access to platform data make it impossible to causally identify the impact of this scope creep, we document many cases in which hate speech policies, when erroneously and inaccurately enforced, actually led to the inadvertent repression of minority speech. Most platforms argue hate speech can silence minority voices, but the non-exhaustive list of examples included in this report raise questions about the extent to which hate speech policies achieve their objectives. Moreover, the findings of this report challenge the prevailing narrative that platforms have been indifferent to hate speech and that social media constitutes an "unregulated Wild West" where hatred is allowed to spread freely. In fact, viewed from a human rights perspective, there are strong reasons to believe that platforms tend to err on the side of restrictions, rather than expression, when formulating hate speech policies.

Addressing hate speech is not an easy task for platforms, given they have diverse, global user bases with varying norms surrounding hate speech, face varied domestic laws that address the topic, and must rely on artificial intelligence to moderate the unprecedented amount of speech they host daily. However, the status quo approach to hate speech at all eight of the analyzed platforms goes far beyond globally accepted norms surrounding legitimate restrictions on freedom of expression, despite most of the platforms publicly committing to uphold these standards. We therefore present two potential complementary alternatives to this status quo – tying hate speech policies to international human rights law (IHRL) and/ or decentralizing content moderation - and discuss their benefits and drawbacks. Ultimately, however, we believe both paths forward are preferable to the status quo and recommend platforms adopt one or both. Under the former option, platforms would prohibit hate speech consistent with Articles 19 and 20 (2) of the International Covenant on Civil and Political Rights. Decentralizing content moderation would involve allowing third parties to develop their own filters for content, which users could choose between based on their own values and tolerance levels. A combination of the two might look like platforms allowing third parties to develop their own content moderation and curation systems but requiring that all of these still abide by IHRL standards when it comes to hate speech.

Introduction

In 1999, Harvard professor Laurence Lessig declared that the internet would make it almost impossible to regulate speech. "Relative anonymity, decentralized distribution, multiple points of access, no necessary tie to geography, no simple system to identify content, tools of encryption—all these features and consequences of the Internet protocol make it difficult to control speech in cyberspace," Lessig explained, emphasizing that "the architecture of cyberspace is the real protector of speech."[1] Lessig was not alone in his optimism; democracies, media institutions, and civil society organizations hailed the promise of technology to spread freedom of expression to new corners of the world and combat traditional authoritarian censorship methods. However, this tech utopianism, which characterized the early days of the World Wide Web, has largely disappeared. Today, centralized social media platforms dominate the internet, deciding what speech is and is not permissible, and governments are introducing legislation aiming to directly regulate those platform decisions. Six tech giants and their subsidiaries: Google (including YouTube), Netflix, Facebook (including Instagram and WhatsApp), Microsoft (including Skype and LinkedIn), Apple, and Amazon account for close to half of all internet traffic.[2] These dominant platforms have become *de facto* gatekeepers of global information streams, guided by their own Terms of Service and Community Guidelines, as well as opaque algorithmic content moderation and distribution systems. As former UN Special Rapporteur on Freedom of Expression and Opinion David Kaye noted, "a centralizing internet dominated by corporative imperatives …. is friendlier to censorship …. than the horizontal web of blogs and websites."[3]

As platforms have become more important in everyday life, public and elite concern about the virality and volume of toxic content online – and the ability of social media platforms to profit from it - has grown. During a speech to the Internet Governance Forum in 2018, French President Emmanuel Macron said that democracies could "not tolerate much longer the torrents of hate coming over the Internet from authors protected by anonymity."[4] In July 2020, over 1,200 businesses and civil society organizations took part in the Stop Hate for Profit ad pause, which aimed to send a clear message to Facebook to "stop valuing profits over hate, bigotry, racism, antisemitism, and disinformation."[5] In January 2022, EU Commissioner Thierry Breton characterized the online ecosystem as a 'Wild West' of "uncontrolled hate speech, incitement to violence,

[1] Lawrence Lessig, *'Code: And Other Laws of Cyberspace'* 1st edn. Basic Books 1999, 166.
[2] Cam Cullen, "Over 43% of the internet is consumed by Netflix, Google, Amazon, Facebook, Microsoft, and Apple: Global Internet Phenomena Spotlight." Sandvine, August 30, 2019, https://www.sandvine.com/blog/netflix-vs.-google-vs.-amazon-vs.-facebook-vs.-microsoft-vs.-apple-traffic-share-of-internet-brands-global-internet-phenomena-spotlight .
[3] David Kaye, 2018: Introduction
[4] Emmanuel Macron, "Internet Governance Forum 2018 Speech," Internet Governance Forum, https://www.intgovforum.org/en/content/igf-2018-speech-by-french-president-emmanuel-macron.
[5] Anti-Defamation League, "Stop Hate for Profit, " https://www.adl.org/stop-hate-profit-0.

disinformation, and destabilization strategies" from which social media companies "have - it must be admitted – largely profited."[6]

Both "platformization" and public sentiment have put pressure on governments to address harmful speech. Even liberal democracies are now mandating that platforms moderate certain categories of content. In that sense, a mix of centralized private power and public power is rendering social media platforms to be arbiters of truth, fact, and law. These national and regional legislative measures, such as the German Network Enforcement Act (NetzDG) and the Digital Services Act (DSA) of the European Union (EU), significantly enhance platform responsibilities for user-generated content. To meet these legal obligations and avoid hefty fines, social media platforms are arguably adopting a "better safe than sorry approach," increasingly relying on artificial intelligence to proactively remove contentious areas of speech and erring on the side of caution when addressing sensitive categories of content, such as hate speech. Across the Atlantic, American policymakers on the left and the right are trying to amend the platform immunity enshrined in Section 230 of the United States' Communications Decency Act, to address competing concerns that platforms are amplifying harmful content[7] or censoring conservative speech.[8] Lawmakers in Texas and Florida have also passed legislation that would force platforms to host certain types of speech, which industry groups,[9] free speech activists, and several academics[10] argue violate the First Amendment. These developments, especially the DSA, are likely to have global impacts. Courtesy of the 'Brussels Effect,' legislators around the world often base their norms and policies on EU rules.[11]

[6] Thierry Breton, "Speech by Commissioner Breton on the Digital Services Act," transcript of speech delivered at the European Commission, Brussels, January 19, 2022, https://ec.europa.eu/commission/presscorner/detail/en/speech_22_431 .

[7] See, for example, the Protecting Americans from Dangerous Algorithms Act at "Reps. Eshoo and Malinowski Introduce Bill to Hold Platforms Liable for Algorithmic Promotion of Extremism," October 20, 2020, https://eshoo.house.gov/media/press-releases/reps-eshoo-and-malinowski-introduce-bill-hold-tech-platforms-liable-algorithmic.

[8] See discussion draft introduced in 2021 by Republican House members at "McMorris Rodgers Leads Aggressive Effort to Hold Big Tech Accountable, Announces Next Steps to Reform Section 230," July 28, 2021, https://mcmorris.house.gov/posts/mcmorris-rodgers-leads-aggressive-effort-to-hold-big-tech-accountable-announces-next-steps-to-reform-section-230.

[9] Adam Liptak, "Supreme Court Puts Off Considering State Laws Curbing Internet Platforms," *New York Times*, January 23, 2023, https://www.nytimes.com/2023/01/23/us/scotus-internet-florida-texas-speech.html .

[10] See Knight First Amendment Institute, "Amicus Brief: NetChoice v. Paxton," https://knightcolumbia.org/cases/netchoice-llc-v-paxton ; Jeff Kosseff, "State Legislatures Threaten Right to Anonymous Speech," *Lawfare*, March 8, 2023, https://www.lawfareblog.com/state-legislatures-threaten-right-anonymous-speech ; Jeff Kosseff, "9 People Hold the Internet's Fate in Their Hands," *Wired*, February 24, 2023, https://www.wired.com/story/scotus-section-230/ ; Eric Goldman, Amicus Brief in NetChoice v. Florida Attorney General, *Santa Clara Univ. Legal Studies Research Paper No. 4289070*, November 23, 2022, https://ssrn.com/abstract=4289070 or http://dx.doi.org/10.2139/ssrn.4289070 ; Mike Masnick, "5th Circuit Rewrites a Century of 1st Amendment Law to Argue Internet Companies Have No Right to Moderate," *TechDirt*, September 16, 2022, https://www.techdirt.com/2022/09/16/5th-circuit-rewrites-a-century-of-1st-amendment-law-to-argue-internet-companies-have-no-right-to-moderate/ .

[11] Anu Bradford, *The Brussels Effect: How the European Union Rules the World* (New York, 2020; online edn, Oxford Academic, 19 Dec. 2019), https://doi.org/10.1093/oso/9780190088583.001.0001, accessed 30 Apr. 2023.

These reputational, financial, and regulatory pressures may have led social media platforms to police a broader scope of potentially objectionable content over time, beyond what is prohibited in Article 20(2) of the International Covenant on Civil and Political Rights (ICCPR), despite committing to align with international human rights standards. To date, however, there has been no cross-platform, cross-temporal assessment of the scope of platform hate-speech policies.[12] In this report, we explore how the content covered by and the protected characteristics stipulated in platforms' hate speech policies have evolved over time. While we hypothesize the scope of the policies has grown over time, due to the pressures outlined above, we only assess the first part of the hypothesis. In other words, the report is a descriptive endeavor, which maps the evolution of platform policies without assessing the causes of any changes. However, by documenting changes in platforms' hate speech policies, we hope to lay the groundwork for future research on the causes.

To assess whether platforms' hate speech policies have expanded in scope over time, we collected information from the Terms of Service and Community Guidelines/Standards/Rules of eight platforms: Facebook, Instagram, Reddit, Snapchat, TikTok, Tumblr, Twitter, and YouTube. We documented the evolution of these policies to assess if the scope of content they cover has changed from inception to the current day, and if so, the extent of that change. We also assess whether these policies align with the framework for assessing the limits between freedom of expression on the one hand and the mandatory and permitted prohibitions on hate speech in International Human Rights Law (IHRL), specifically Articles 19(3) and 20(2) of the International Covenant on Civil and Political Rights (ICCPR). We assess platform policies against these standards because most of the platforms have explicitly committed themselves to these standards in their human rights policies, including by acknowledging the UN Guiding Principles on Business and Human Rights. Moreover, the UN Special Rapporteur on freedom of opinion and expression has called for platforms to use IHRL to guide their policies and developed interpretive frameworks for doing so.[13] Nevertheless, we recognize that making IHRL the basis for content moderation would not be a panacea for the many issues with the practice, and we discuss the challenges with this approach in the conclusion.

Our analysis reveals significant scope creep in platforms' hate speech policies over time, both in the content and the protected characteristics covered. Platforms have gone from prohibiting the promotion of hatred or racist speech to prohibiting a long list of potential forms of hate speech,

[12] While there have been analyses of hate speech policies at one specific point in time, they did not track the evolution of hate speech policies over time nor document the full extent of content or protected characteristics covered by the policies. See Adriana Stephan, "Comparing Platform Hate Speech Policies: Reddit's Inevitable Evolution," Freeman Spogli Institute, Stanford Internet Observatory, July 8, 2020, https://fsi.stanford.edu/news/reddit-hate-speech.

[13] Special Rapporteur on freedom of opinion and expression, "Report on content regulation," A/HRC/38/35, April 6, 2018, https://www.ohchr.org/en/calls-for-input/report-content-regulation.

including harmful stereotypes, conspiracy theories, and cursing targeting protected groups. The average number of protected characteristics listed in platform policies has more than doubled since 2010, with platforms protecting identities as wide-ranging as veteran status, pregnancy, and victims of a major event. These developments do not align with international human rights standards, notably Article 19 of the International Covenant on Civil and Political Rights.[14] Moreover, while current restrictions on researcher access to platform data make it impossible to causally identify the impact of this scope creep, we provide anecdotal evidence that platforms often silence minority speech by enforcing the hate speech policies that are designed to protect it. This reality suggests that these policies often result in outcomes that contradict their stated objectives. We conclude by discussing the benefits and drawbacks of two potential solutions to this problem: tying content policies more directly to international human rights law and/or moving to a decentralized model of content moderation and curation.

Motivation & Hypotheses

As of April 2023, 5.18 billion people use the internet,[15] and 4.8 billion of those people use social media.[16] Social media platforms are communities that allow users to share ideas and opinions on themes ranging from politics to fashion, revolutionizing social interaction in a manner unimaginable just only a decade ago. Nevertheless, the egalitarian model of social media, and of the internet more generally, also provides a channel for the expression of hatred and extremism. This reality has contributed to significant concern and even panic among both governments, academics, civil society, and traditional media.

Social media companies are also often accused of profiting off hateful content. For example, in June 2020, the Stop Hate for Profit campaign, which included civil society organizations like the Anti-Defamation League, Color of Change, Common Sense, Free Press, and the NAACP, asked businesses to take a stand against "hate and disinformation being spread by Facebook" by temporarily pausing advertising on Facebook and Instagram.[17] Over 1,000 businesses joined the campaign. In response to an allegation by former Meta-employee Frances Haugen that Facebook only removed 3 to 5% of hate speech, the President and CEO of the NAACP, Derrick Johnson, told Bloomberg News that white supremacism is rampant and expressed dismay at the way platforms

[14] See Special Rapporteur on freedom of opinion and expression, "Report on Content Regulation", and Special Rapporteur on freedom of opinion and expression, "A/74/486: Report on online hate speech," A/74/486, https://www.ohchr.org/en/documents/thematic-reports/a74486-report-online-hate-speech.

[15] "Digital Around the World," *Datareportal*, https://datareportal.com/global-digital-overview (accessed April 30, 2023.)

[16] "Global Social Media Statistics," *Datareportal*, https://datareportal.com/social-media-users (accessed April 30, 2023).

[17] "More than 1,000 companies pause advertising on Facebook as part of civil society campaign to stop spread of hate & discrimination on the platform," *Business & Human Rights Resource Centre*, July 3, 2020, https://www.business-humanrights.org/en/latest-news/more-than-1000-companies-pause-advertising-on-facebook-as-part-of-civil-society-campaign-to-stop-spread-of-hate-discrimination-on-the-platform/

were "profiting on hate and disinformation."[18] In February 2021, Jonathan Greenblatt, the CEO of the Anti-Defamation League, said it was "far too easy for individuals interested in extremist content" to find it on YouTube, and he called for the company "to be held accountable for instances when their systems, built to engage users, actually amplify dangerous content that leads to violence."[19] Thierry Henry, a former football (soccer) star, led a boycott of social media platforms in 2021 to protest platforms profiting from hate. He told The Guardian that "people shouting abuse in the street (will) be arrested," but online "it seems you can do whatever you want."[20]

In response to concerns like this, many social media companies have developed usage terms and content policies that prohibit certain forms of hateful or discriminatory content, and governments around the world have introduced legislation seeking to regulate those platform policies. In 2016, the European Commission agreed with Facebook, Microsoft, Twitter, and YouTube on a Code of Conduct on Illegal Hate Speech,[21] which requires platforms to "voluntarily" remove hate speech within 24 hours. Since then, eight more companies – including Instagram, Snapchat, and TikTok – have signed on.[22] It has been signed by all of the platforms analyzed in this report except for Tumblr and Reddit. In 2017, Germany adopted the Network Enforcement Act (NetzDG), which imposed hefty fines on social media platforms that did not remove content that violated provisions of the German Criminal Code, including insult, incitement, and religious defamation. Social media platforms have 24 hours to remove "manifestly unlawful content" and up to seven days for merely "unlawful content".[23] This template has been replicated in more than twenty countries around the world, including many authoritarian states.[24] In May 2020, France passed the Avia Law, which obligated social media companies to remove "manifestly illicit" hate speech within 24 hours or face fines of up to 1.25 million

[18] Cameron Jenkins, "NAACP calls for meeting with Zuckerberg after hate speech revelations on Facebook," *The Hill*, October 6, 2021, https://thehill.com/policy/technology/575493-naacp-calls-for-meeting-with-zuckerberg-after-hate-speech-revelations-on/ .

[19] "Despite remediation efforts, ADL fins YouTube Still Amplifies Extremist Content," *Anti-Defamation League*, February 11, 2021, https://www.adl.org/resources/press-release/despite-remediation-efforts-adl-finds-youtube-still-amplifies-extremist .

[20] Dan Milmo, "Social media companies 'make money from hate', says Thierry Henry," *The Guardian*, November 2, 2021, https://www.theguardian.com/football/2021/nov/02/social-media-companies-make-money-from-hate-says-thierry-henry.

[21] European Commission, "The EU Code of conduct on countering illegal hate speech online," https://ec.europa.eu/info/policies/justice-and-fundamental-rights/combatting-discrimination/racism-and-xenophobia/eu-code-conduct-countering-illegal-hate-speech-online_en.

[22] European Commission, "The EU Code of conduct on countering illegal hate speech online."

[23] Network Enforcement Act (Netzdurchsetzunggesetz, NetzDG), *German Law Archive*, https://germanlawarchive.iuscomp.org/?p=1245 .

[24] See Jacob Mchangama and Joelle Fiss, "The Digital Berlin Wall: How Germany (Accidentally) Created a Prototype for Global Online Censorship," *Justitia*, November 2019, https://justitia-int.org/en/the-digital-berlin-wall-how-germany-created-a-prototype-for-global-online-censorship/; and Jacob Mchangama and Natalia Alkiviadou, "The Digital Berlin Wall: How Germany (Accidentally) Created a Prototype for Global Online Censorship – Act Two," *Justitia*, September 2020, https://justitia-int.org/en/the-digital-berlin-wall-act-2-how-the-german-prototype-for-online-censorship-went-global-2020-edition/.

Euros.[25] In June 2020, however, France's Constitutional Council ruled that this law limited freedom of expression in an unnecessary, inappropriate, and disproportional manner.[26]

In July 2022, the European Parliament officially adopted the DSA, the EU's landmark legislation for the digital sphere,[27] and the European Council approved the text a few months later, in October 2022.[28] While the DSA will not apply across the EU until January 2024, certain obligations went into force in 2023.[29] The DSA establishes a "notice and action" process that requires hosting services to act "without undue delay, taking into account the type of illegal content that is being notified and the urgency of taking action."[30] It notes that illegal content should cover "hate speech" and "unlawful discriminatory content," but these terms are to be defined in national and EU law. This vagueness could push platforms to be overly cautious.[31] Although the DSA does not impose general monitoring obligations on hosting providers, it achieves enhanced liability through other means. The due diligence rules for very large online platforms (VLOPs), including annual risk assessments under the close eye of the Commission and the possibility of fines for non-compliance, arguably still dilute the liability exemption directly.[32] As Joan Barata, a Senior Fellow at the Future of Free Speech Project and the Stanford Center for Platform Regulation, argues, the mere notification of alleged illegality should not create knowledge or awareness to kick start the notice and action process "unless the notified content reaches a certain threshold of obviousness of illegality."[33]

[25] "Assemblée Nationale, Session Ordinaire de 2019-2020, 22 Janvier 2020," http://www.assemblee-nationale.fr/dyn/15/textes/l15t0388_texte-adopte-seance.

[26] "French law on illegal contnet online ruled unconstitutional: Lessons for the EU to learn," *Patrick Breyer Press Release,* June 18, 2020, https://www.patrick-breyer.de/?p=593729&lang=en.

[27] Pim ten Thije, "The Digital Services Act: Adoption, Entry into Force and Application Dates," *DSA Observatory*, https://dsa-observatory.eu/2022/09/12/digital-services-act-adoption-entry-into-force-application-dates-dsa/.

[28] European Council, "DSA: Council gives final approval to the protection of users' rights online," *Council of the EU Press Release*, October 4, 2022, https://www.consilium.europa.eu/en/press/press-releases/2022/10/04/dsa-council-gives-final-approval-to-the-protection-of-users-rights-online/ .

[29] "The Digital Services Act package," *European Commission*, https://digital-strategy.ec.europa.eu/en/policies/digital-services-act-package.

[30] European Parliament, "Digital Services Act: Regulation platforms for a safer online space for users," *European Parliament Press Releases*, January 20, 2022, https://www.europarl.europa.eu/news/en/press-room/20220114IPR21017/digital-services-act-regulating-platforms-for-a-safer-online-space-for-users .

[31] Joan Barata, "The Digital Services Act and Its Impact on the Right to Freedom of Expression: Special Focus on Risk Mitigation Obligations," *Plataforma por la Libertad de Información*, https://libertadinformacion.cc/wp-content/uploads/2021/06/DSA-AND-ITS-IMPACT-ON-FREEDOM-OF-EXPRESSION-JOAN-BARATA-PDLI.pdf .

[32] The European Commission designated six of the eight platforms analyzed in this report – Facebook, Instagram, Snapchat, TikTok, Twitter, and YouTube – as VLOPs. See "Digital Services Act: Commission designates first set of Very Large Online Platforms and Search Engines," *European Commission Press release*, April 25, 2023, https://ec.europa.eu/commission/presscorner/detail/en/IP_23_2413.

[33] Barata, "The Digital Services Act and Its Impact on the Right to Freedom of Expression: Special Focus on Risk Mitigation Obligations."

Officials in Brazil are looking to Europe as they seek to construct a new framework for platform regulation.[34] Since taking office in early 2023, the administration of Brazilian President Luiz Inácio Lula da Silva has "been actively pursuing measures to enhance the responsibility of intermediaries in moderating harmful online content on multiple fronts."[35] In March, the Secretary for Digital Policies, Joao Brant, said that the government is seeking to make social media companies responsible for preventing the spread of misinformation and hate speech, by monitoring their efforts to protect the digital environment overall.[36] On April 12, 2023, the Ministry of Justice and Public Safety issued an ordinance combating illegal, harmful, or damaging content on social media platforms.[37] The government is also proposing changes to the draft "Fake News Bill," including amendments that would require platforms to "take preventive action against 'potentially illegal content' generated by third parties."[38]

The landscape in the United States is somewhat different. In the last few years, new laws in Texas and Florida attempted to force platforms to host certain types of speech that they might otherwise have removed. Texas HB 20 prohibits platforms from censoring "a user, a user's expression, or a user's ability to receive the expression of another person based on the viewpoint of the user or another person,"[39] while Florida SB 7072 introduced fines on social media companies if they removed political candidates from their platforms for violating usage policies.[40] Two industry groups, NetChoice and the Computer & Communications Industry Association, challenged the constitutionality of the laws, arguing that the First Amendment "prohibits the government from telling private companies whether and how to disseminate speech."[41] Competing rulings from the relevant Circuit Courts have set up a potential Supreme Court showdown on these social media laws.[42]

The First Amendment not only prohibits the U.S. government from regulating platforms' speech via regulation of their content policies, but it also prohibits the government from requiring

[34] Beatriz Kira, "In Brazil, Platform Regulation Takes Center Stage," *Tech Policy Press,* April 24, 2023, https://techpolicy.press/in-brazil-platform-regulation-takes-center-stage/
[35] Kira, "In Brazil, Platform Regulation Takes Center Stage."
[36] Victor Pinheiro, "Brazil looks to regulate monetized content on the internet," *Reuters*, March 17, 2023, https://www.reuters.com/world/americas/brazil-looks-regulate-monetized-content-internet-official-2023-03-17/
[37] "Diario Oficial Da União," *Ministério da Justiça e Segurança Pública/Gabinete do Ministro*, Abril 12, 2023, https://www.in.gov.br/en/web/dou/-/portaria-mjsp-n-351-de-12-de-abril-de-2023-476702096.
[38] Kira, "In Brazil, Platform Regulation Takes Center Stage."
[39] John Villasenor, "Texas's new social media law is likely to face an uphill battle in federal court," *Brookings*, November 9, 2021, https://www.brookings.edu/blog/techtank/2021/11/09/texass-new-social-media-law-is-likely-to-face-an-uphill-battle-in-federal-court/
[40] Liptak, "Supreme Court Puts Off Considering State Laws Curbing Internet Platforms."
[41] Liptak, "Supreme Court Puts Off Considering State Laws Curbing Internet Platforms."
[42] Andrew Chung, "U.S. Supreme Court seeks Biden administration view on Florida, Texas social media laws," *Reuters*, January 24, 2023, https://www.reuters.com/legal/us-supreme-court-seeks-biden-administration-view-florida-texas-social-media-laws-2023-01-23/

platforms to ban protected speech. It does not, however, prohibit platforms from independently deciding to block that protected speech. Section 230 of the Communications Decency Act also assists platforms if they choose to go that route, by providing immunity for efforts to block objectionable content. In other words, platforms cannot be held liable for some content that they host simply because they have chosen to remove other types of content.[43] The Supreme Court recently heard oral arguments in two cases related to Section 230, *Gonzalez v. Google and Twitter v. Taamneh.* In both cases, the plaintiffs argued that the platforms recommended ISIS content to users and did not adequately enforce anti-terrorism content policies, thereby aiding ISIS in their terrorist attacks; the platforms are defending themselves on Section 230 grounds.[44] The Court's decision to grant cert in these cases raised concern among scholars of internet law and technology policy experts that the Justices would unravel the legal foundation of the modern internet. However, in *Taamneh,* the Court ruled that the plaintiffs (petitioners) failed to demonstrate that Twitter aided the terrorist attack in question, and, in light of that decision, the Court remanded *Gonzalez* to the Ninth Circuit.[45] Even though the Court did not upend Section 230 with these rulings, pressure within the U.S. political system to reform the statute is likely to continue. In recent years, Americans on both sides of the aisle have called for Section 230 reform.[46] Conservatives believe that the statute allows social media companies to censor speech based on viewpoint, while liberals are frustrated that Section 230 allows platforms to profit from harmful and objectionable speech.[47]

Despite concerns about the proliferation of online hate speech – and associated legal and legislative battles, several studies suggest hate speech comprises a relatively small proportion of social media content. A recent study, which assessed whether Trump's 2016 campaign and its aftermath contributed to a rise in hate speech, found that only between 0.001% and 0.003% of

[43] Jeff Kosseff, "A User's Guide to Section 230, and a Legislator's Guide to Amending It (or Not)," *Berkeley Technology Law Journal* 37, no. 2 (2022).

[44] Quinta Jurecic, Alan Z. Rozenshtein, and Benjamin Wittes, "Have the Justices Gotten Cold Feet About 'Breaking the Internet?'," *Lawfare*, February 24, 2023, https://www.lawfareblog.com/have-justices-gotten-cold-feet-about-breaking-internet.

[45] Hyemin Han, "Supreme Court Rules in Favor of Twitter in Taamneh, Remands Gonzelez," *Lawfare,* May 18, 2023, https://www.lawfareblog.com/supreme-court-rules-favor-twitter-taamneh-remands-gonzalez#:~:text=Supreme%20Court%20Rules%20in%20Favor%20of%20Twitter%20in%20Taamneh%2C%20Remands%20Gonzalez,-By%20Hyemin%20Han&text=On%20May%2018%2C%20the%20Supreme,Twitter%20and%20sent%20Gonzalez%20v.

[46] See "Reps. Eshoo and Malinowski Introduce Bill to Hold Platforms Liable for Algorithmic Promotion of Extremism," October 20, 2020, https://eshoo.house.gov/media/press-releases/reps-eshoo-and-malinowski-introduce-bill-hold-tech-platforms-liable-algorithmic, and "McMorris Rodgers Leads Aggressive Effort to Hold Big Tech Accountable, Announces Next Steps to Reform Section 230," July 28, 2021, https://mcmorris.house.gov/posts/mcmorris-rodgers-leads-aggressive-effort-to-hold-big-tech-accountable-announces-next-steps-to-reform-section-230.

[47] Danielle Keats Citron and Mary Anne Franks, "The internet as a speech machine and other myths confounding section 230 reform," *U. Chi. Legal F.* (2020): 45.

1.2 billion analyzed tweets from this period contained such content."[48] A joint study from the University of Oxford and Addis Ababa University demonstrated a similarly low prevalence of hate speech content on Facebook in Ethiopia, finding that only 0.4% of 13,000 statements in the sample incited others to discriminate or act against individuals based on ethnicity, religion, or gender.[49] A 2020 Justitia study found that, for each criminally sanctionable comment removed from the Facebook pages of Danish news media, 36 non-hateful and non-offensive comments on issues such as politics were removed.[50]

This level of online hate speech can still pose pain, and in some cases, real life harm to the vulnerable communities it targets, and there are documented instances of online hate speech fueling offline violence – such as the Rohingya genocide in Myanmar.[51] Thus, one could argue that online hate speech should be regulated even more stringently than it currently is. However, hate speech moderation is not without collateral damage for free expression and the expression of minority voices. A 2022 Justitia study analyzed 2,400 Facebook comments labeled as "hateful attacks," using an algorithm that was trained to detect attacks and hate speech in Danish and Norwegian, and which operated by collecting and timestamping the comments to determine the number of comments deleted from all pages in the study and assess how many of these violated Danish Law.[52] The comments were a representative sample of over 900,000 hateful attacks found by analyzing 63 million comments on Facebook pages belonging to Danish politicians and media outlets. Justitia found that only 11 comments, or 0.066% of the total, could be considered illegal under Danish prohibitions on incitement and hate speech. These findings suggest that expansive definitions of hate speech may lead to the mass removal of legal content.

There is also evidence that hate speech restrictions, including prohibitions on content that expresses support for hate groups or hateful ideologies, have silenced marginalized voices. Twitter previously suspended the accounts of many Egyptian dissidents, due to an algorithm that flagged content involving Arabic swear words as hateful.[53] For example, one Twitter user was blocked after

[48] Alexandra Siegel, Evgenii Nikitin, Pablo Barberá, Joanna Sterling, Bethany Pullen, Richard Bonneau, Jonathan Nagler, and Joshua, "Trumping hate on Twitter? Online hate speech in the 2016 U.S. Election campaign and its aftermath," *Working Paper,* that has since been published in the Quarterly Journal of Political Science,) accessed via author's website, https://alexandra-siegel.com/wp-content/uploads/2019/08/qjps_election_hatespeech_RR.pdf.

[49] Iginio Gagliardone, Matti Pohjonen, Zenebe Beyene, Abdissa Zerai, Gerawork Aynekulu, Mesfin Bekalu, Jonathan Bright et al, "Mechachal: Online debates and elections in Ethiopia-from hate speech to engagement in social media," *Available at SSRN 2831369* (2016).

[50] Jacob Mchangama, "New report: Digital Freedom of Speech and Social Media," *Justitia*, May 29, 2020, http://justitia-int.org/en/new-report-digital-freedom-of-speech-and-social-media/

[51] "Myanmar: The social atrocity: Meta and remedy for the Rohingya," *Amnesty International,* September 29, 2022, https://www.amnesty.org/en/documents/ASA16/5933/2022/en/ .

[52] Jacob Mchangama, "The Wild West," *Justitia*, January 20, 2022, https://justitia-int.org/en/the-wild-west/ .

[53] Wael Eskander, "How Twitter is gagging Arabic users and acting as morality police, *Open Democracy*, October 23, 2019, https://www.opendemocracy.net/en/north-africa-west-asia/how-twitter-gagging-arabic-users-and-acting-morality-police/ .

commenting, "a few ass kissers appeared during the time of the Muslim Brotherhood, became state loyalists with the appearance of the ass kisser." Another was suspended for tweeting: "Tawadros (the Coptic pope) was an ass kisser," while still another lost his account because he cursed an Egyptian football club. In a 2020 open letter to Facebook, Twitter, and YouTube, an array of journalists, activists and human rights organizations urged the platforms to "stop silencing critical voices from the Middle East and North Africa."[54] In May 2021, Meta admitted that mistakes in their hate speech detection algorithms led to the inadvertent removal of millions of pro-Palestinian posts.[55]

In 2022, Facebook removed a post from a user in Latvia that cited alleged atrocities committed by Russian soldiers in Ukraine, along with text from a poem by Konstantin Simonov that included the lines: "kill the fascist.. Kill him! Kill him! Kill!" The post tried to draw a connection between the Nazi army and the Russian army in Ukraine. Meta's Oversight Board found that removing the post did not align with Meta's Community Standards nor its human rights obligations, noting that the post did not make a general accusation that "Russian soldiers are Nazis," but rather suggested they acted like Nazis at a specific time and place.[56] It is worth noting Meta´s original decision to remove the post would likely have been upheld if the Oversight Board had applied regional European, rather than international UN, human rights standards. In 2018, the European Court of Human Rights (ECtHR) found that a journalist who had called Russian security forces operating in Chechnya '"maniacs", "murderers," and otherwise criminally minded persons,' had overstepped the limits of freedom of expression.[57]

There are also sometimes inequalities in enforcement. As Eric Heinze, a Professor of Law at the University of London's Queen Mary School of Law, argues, historically viewpoint restriction has "overwhelmingly been one of repression of minority and dissenting voices."[58] For example, Facebook did not remove a post from a U.S. Congressman that called for the slaughter of "radicalised Muslims." "Kill them all," it read, "For the sake of all that is good and righteous. Kill them all."[59] However, Facebook did remove a post from a Black Lives Matter activist that read: "all white people are racist. Start from this reference point or you've already failed." According to ProPublica, Facebook allowed the first post because it referenced radicalized Muslims only and

[54] "Open Letter to Facebook, Twitter, and YouTube: Stop silencing critical voices from the Middle East and North Africa," *Access Now*, December 17, 2020, https://www.accessnow.org/facebook-twitter-youtube-stop-silencing-critical-voices-mena/
[55] Elizabeth Dwoskin and Gerrit De Vynck, "Facebook's AI treats Palestinian activists like it treats American Black activists. It blocks them.," *Washington Post*, May 28, 2021, https://www.washingtonpost.com/technology/2021/05/28/facebook-palestinian-censorship/
[56] "Russian poem," *Oversight Board,* 2022-008-FB-UA, https://www.oversightboard.com/decision/FB-MBGOTVN8/ .
[57] Stomakhin v Russia, App. No 52273/07 (ECHR 9 May 2018) Para.9
[58] Eric Heinze, "Hate speech and the normative foundations of regulation," *International Journal of Law in Context* 9, no. 4 (2013): 590-617.
[59] https://archive.is/95FO1

not Muslims in general – but removed the second because it referred to all white people.[60] The widespread use of automated content moderation systems also contributes to unequal enforcement, since natural language processing algorithms often amplify biases in training data, are more accurate for some languages than others,[61] and cannot recognize contextual or local nuances in speech[62]. For example, a recent study from Dias et. al found that automated content moderation technology developed by Jigsaw considered a significant number of drag queen Twitter accounts to have higher levels of toxicity than White nationalists and could not distinguish when LGBTQ people were using words that might conventionally be offensive to reclaim their power or in a self-referential way.[63] Accordingly, it is not clear that enforcing ever-expanding hate speech policies is the best way to achieve their stated objective of protecting minorities and vulnerable groups, while also safeguarding free expression.

Finally, different people, with different values, in different countries and cultures, have different ideas about what type of content should be protected and different degrees of tolerance towards potentially objectionable content.[64] In 2021, a Justitia survey fielded by YouGov in 33 countries, from every major continent, found that around 90% of people support free speech in principle.[65] However, support drops when free speech is put in tension with other values, such as protecting minority groups. expressing support for homosexuality or safeguarding the national economy. Overall, individuals in the U.S., Scandinavia, East Asia, Hungary, and Venezuela tend to be more supportive of the right to offend minority groups than individuals in Turkey, Pakistan, Indonesia, Tunisia, and Kenya.[66] Men are also more supportive of this right than women.[67] Moreover, in the U.S., Australia, and Europe, individuals' placement on the political spectrum impacts their support for the right to make statements that are offensive to minority groups, while this relationship is not as strong in developing countries.[68] These results suggest social media users are not a monolith in terms of their views on what constitutes unacceptable speech. Thus, platforms are

[60] Julia Angwin, ProPublica, and Hannes Grasseger, "Facebook's Secret Censorship Rules Protect White Men From Hate Speech But Not Black Children," *ProPublica*, June 28, 2017, https://www.propublica.org/article/facebook-hate-speech-censorship-internal-documents-algorithms .

[61] Natasha Duarte, Emma Llanso, Anna Loup, "Mixed Messages?: The Limits of Automated Social Media Content Analysis," *Center for Democracy and Technology*, November 2017, https://cdt.org/wp-content/uploads/2017/11/Mixed-Messages-Paper.pdf .

[58] Thiago Dias Oliva, Antonialli Dennys Marcelo, and Alessandra Gomes, "Fighting hate speech, silencing drag queens? artificial intelligence in content moderation and risks to lgbtq voices online," *Sexuality & culture* 25, no. 2 (2021): 700-732.

[63] Dias Oliva, et al, "Fighting hate speech, silencing drag queens? artificial intelligence in content moderation and risks to lgbtq voices online."

[64] Jacob Mchangama, "Report: Who Cares About Free Speech? Findings From a Global Survey of Free Speech," *Justitia*, June 7, 2021, https://justitia-int.org/en/report-who-cares-about-free-speech-findings-from-a-global-survey-of-free-speech/.

[65] Mchangama, "Report: Who Cares About Free Speech? Findings From a Global Survey of Free Speech."

[66] Svend-Erik Skaaning and Suthan Krishnarajan, "Who Cares about Free Speech? Findings from a Global Survey of Support for Free Speech," *Justitia*, May 2021, https://futurefreespeech.com/wp-content/uploads/2021/06/Report_Who-cares-about-free-speech_21052021.pdf, pg 9.

[67] Skaaning and Krishnarajan, "Who Cares about Free Speech? Findings from a Global Survey of Support for Free Speech," pg 21.

[68] Skaaning and Krishnarajan, "Who Cares about Free Speech? Findings from a Global Survey of Support for Free Speech," pg 22-23.

likely to face competing pressures from various users, and the organizations that represent them, to develop policies that reflect the views of different groups, all of which cannot simultaneously be consistently formulated much less enforced.

Many years ago, Marc Zuckerberg wrote that Facebook was enabling people to "make their voices heard on a different scale from what has historically been possible."[69] As platforms grew, however, they began implementing certain restrictions on speech to address floods of complaints about different kinds of content as they arose. In the early days, these rules were developed by American lawyers in American companies, via a "common law" approach that was not intended to create a harmonized code for regulating speech at scale.[70] As Kate Klonick has written, early iterations of Facebook's community standards were not "reflective of the norms of global society, or even reflective of the norms of Facebook users. Rather, the early rules reflected the norms of the drafters: Americans 'trained and acculturated in American free speech norms and First Amendment law.'"[71] International human rights law was probably not at the forefront of these lawyers' minds as they attempted to fashion practical and timely responses to the newest type of abuse on the platform or to quickly respond to failures of the existing moderation system. Nevertheless, once it became clear that platforms were developing rules to regulate speech across the world, and were increasingly in dialogue with institutions whose frame of reference is grounded in IHRL, many of them committed to aligning their policies with IHRL. In fact, six of the eight platforms[72] examined in this report, every platform except Reddit and Tumblr,[73] have committed themselves to following the standards set by IHRL, including the UN`s Guiding Principles on Business and Human Rights (Guiding Principles).[74]

At the same time, however, platforms have faced an increasing drumbeat of regulatory and civil society pressure to police objectionable content, incentivizing them to be increasingly risk-averse in their content moderation practices. We hypothesize that this caution has led platforms to abandon their stated, human-rights based approach to policy development – and instead to

[69] Josh Constine, "Facebook's S-1 Letter from Zuckerberg Urges Understanding Before Investment," *TechCrunch*, February 1, 2012, https://techcrunch.com/2012/02/01/facebook-ipo-letter/.

[70] The authors thank Alex Feerst for pointing out the common law approach to policy development that dominated the early years of content moderation and trust & safety work. This history is also explored in Alex Feerst, "A Natural History of Trust & Safety," *Medium,* May 28, 2023, https://feerst.medium.com/a-natural-history-of-trust-safety-c73066d04b86.

[71] Kate Klonick, "The Facebook Oversight Board: Creating an independent institution to adjudicate online free expression," *Yale Law Journal* 129, no. 2418 (2020): pg. 1448.

[72] See Meta: https://about.fb.com/wp-content/uploads/2021/03/Facebooks-Corporate-Human-Rights-Policy.pdf; Twitter: https://help.twitter.com/en/rules-and-policies/defending-and-respecting-our-users-voice; Snapchat: https://s25.q4cdn.com/442043304/files/doc_downloads/2021/05/Code-of-Conduct.pdf; TikTok: https://www.tiktok.com/transparency/en-us/upholding-human-rights/; Google (YouTube): https://about.google/human-rights/.

[73] We could not find any evidence that Reddit or Tumblr have publicly committed to the UN Guiding Principles on Business and Human Rights or to upholding the ICCPR.

[74] "Guiding Principles on Business and Human Rights," *United Nations Human Rights, Office of the High Commissioner,* https://www.ohchr.org/sites/default/files/documents/publications/guidingprinciplesbusinesshr_en.pdf.

develop policies that address the latest instance of harmful or offensive content that is gaining elite attention. The result of this ad-hoc approach has likely been gradual scope creep in platforms' hate speech policies - and the associated negative consequences for free speech documented by the studies above. To assess these hypotheses, we map the evolution of platforms' policies in this area, analyze changes in their scope over time, and compare their scope to prohibitions on hate speech codified in international law. The remainder of this section describes how we do so.

Research Design

Scope

We wanted to focus on social media platforms with a global and relatively large user base, publicly available content policies, and publicly available user-generated content (i.e. non-encrypted platforms). Given the temporal nature of the analysis, we also wanted to get a mix of social media platforms, in terms of size of the user base and popularity over time. Based on this scope, we identified eight platforms for analysis: Facebook, Instagram, Reddit, Snapchat, Twitter, Tumblr, TikTok, and YouTube. While we identified these eight platforms for analysis prior to this decision, the European Commission designated six of these platforms (all except for Reddit and Tumblr) as very large online platforms (VLOPs) in April 2023. Thus, most of the platforms that we analyze in this report will have to comply with the full set of new obligations under the DSA.[75]

The first section of the report describes the evolution of each platform's hate speech policies, from inception until March 2023, analyzes changes in their scope over time, and describes potential correlations between changes in policy scope and changes in enforcement volume. The second section provides a cross-platform analysis of scope changes in content policies that cover hate speech. To conduct these analyses, we had to (a) collect information on each platform's policies and (b) determine whether a particular platform policy addressed hate speech.

Methodology

We collected information about each platform's approach to hate speech via two sources: Terms of Service or Use, which are an agreement between a user and a service provider, and Community Guidelines/Standards/Rules, which are documents that go beyond the basic user agreement and incorporate information about the kind of content that is prohibited on the platform. It is important to note that the above documents do not necessarily provide full transparency regarding social media platforms' content moderation practices, since social networks also have internal guidelines that outline how to address different kinds of speech. For example, Facebook's internal implementation standards comprise an: "ever-changing wiki, roughly twelve thousand words long, with twenty-four headings—Hate Speech, Bullying, Harassment, and so on—each of

[75] See "Digital Services Act: Commission designates first set of Very Large Online Platforms and Search Engines," *European Commission Press release.*

which contains dozens of subcategories, technical definitions, and links to supplementary materials. These are located on an internal software system that only content moderators and select employees can access."[76] Nevertheless, the Terms and Community Guidelines are the most comprehensive, publicly available information about platforms' content rules.

We largely located the platforms' Terms of Service and Community Guidelines through the WayBack Machine,[77] a tool that takes a screenshot of a particular website on a regular basis (even up to several times a day). WayBack Machine has been described as a "viable research tool,"[78] and scholars have argued that "researchers can now have greater confidence in the data generated by the tool."[79] However, while the tool will capture any change on a platforms' policy website, the first time the tool captures a change may not necessarily reflect the official release date of a new policy. This problem is ameliorated if the company website indicates the effective date for each policy. Where an effective date is not listed on the company website for each relevant capture, the date used in the analysis may not be precise and may have a variation of up to 12 months.[80] In some, though not many, instances, platforms provided a change log that offered older versions of the material needed for this report. We used this information when it was available.

Identifying whether a particular policy addressed hate speech was more complicated. While many companies label certain rules as "hate speech policies," relevant provisions may also be listed under other titles. Therefore, we had to develop a coding rule that could be applied across platform policies to determine if they pertained to hate speech.

To do so, we began by examining definitions of hate speech in international human rights law. While most countries in the world have laws that prohibit hate speech,[81] these bans vary widely in scope and definition, and measuring hate speech policies of global platforms against the approach of one specific country would be inimical to the nature of platforms with users in countries all over the world. While IHRL standards are not formally legally binding on private corporations, in 2018,[82] the former UN Special Rapporteur on the Freedom of Opinion and Expression proposed an IHRL framework for content moderation that "puts human rights at the very center," since national laws are inappropriate given the geographical and cultural diversity of

[76] Andrew Marantz, "Why Facebook Can't Fix Itself," *The New Yorker,* October 12, 2020, https://www.newyorker.com/magazine/2020/10/19/why-facebook-cant-fix-itself .
[77] https://archive.org/web/
[78] Jamie Murphy, Noor Hazarina Hashim, and Peter O'Connor, "Take me back: validating the wayback machine," *Journal of Computer-Mediated Communication* 13, no. 1 (2007): 71.
[79] Ibid.
[80] Ibid, 64.
[81] "Global Handbook of Hate Speech Laws," *Future of Free Speech Project*, https://futurefreespeech.com/global-handbook-on-hate-speech-laws/ .
[82] Report of the Special Rapporteur on the Promotion and Protection of the Right to Freedom of Opinion and Expression (2018) A/HRC/38/35, para. 42.

digital users. The Special Rapporteur stressed that relying on IHRL to determine acceptable and unacceptable speech "enables forceful normative responses against undue State restrictions – provided companies play by similar rules." Moreover, as stated above, six of the platforms[83] examined in this report have committed themselves to respecting IHRL standards, including the UN`s Guiding Principles.[84] The Meta Oversight Board, which reviews content moderation decisions made by Meta (including those pertaining to hate speech) to assess whether the company acted in line with its policies, even conducts case analyses based partly on IHRL principles.[85]

The Guiding Principles set out "human rights [a]s a global standard of expected conduct for all business enterprises wherever they operate." They refer chiefly to the International Bill of Human Rights, which consists of a number of core human rights instruments including, the International Covenant on Civil and Political Rights (ICCPR),[86] which is the most directly relevant human rights convention when it comes to delineating the relationship between freedom of expression and hate speech and has been ratified by 173 countries of the United Nations. The Convention on the Elimination of all Racial Discrimination (ICERD) (ratified by 177 countries)[87] is also relevant for the definition of racist hate speech.

Article 19 (2) of the ICCPR ensures that "Everyone shall have the right to freedom of expression; this right shall include freedom to seek, receive and impart information and ideas of all kinds, regardless of frontiers, [..]through any [..] media of his choice." However, the rights to freedom of expression and access to information may be subject to restrictions that are "provided by law and are necessary" for "respect of the rights or reputations of others," "protection of national security," "public order," or "public health or morals." While Article 19 (3) of the ICCPR permits certain restrictions on freedom of expression, Article 20 (2) mandates that "any advocacy of national, racial, or religious hatred that constitutes incitement to discrimination, hostility, or violence shall be prohibited by law." An important contribution to the interpretation of the relationship between protected speech and impermissible hate speech under the ICCPR is the Rabat Plan of Action (RPA). The RPA was drafted in 2012 following a series of global expert workshops on the prohibition of incitement to national, racial, or religious hatred organized by the Office of the High

[83] See Meta: https://about.fb.com/wp-content/uploads/2021/03/Facebooks-Corporate-Human-Rights-Policy.pdf; Twitter: https://help.twitter.com/en/rules-and-policies/defending-and-respecting-our-users-voice; Snapchat: https://s25.q4cdn.com/442043304/files/doc_downloads/2021/05/Code-of-Conduct.pdf; TikTok: https://www.tiktok.com/transparency/en-us/upholding-human-rights/; Google (YouTube): https://about.google/human-rights/.

[84] "Guiding Principles on Business and Human Rights," *United Nations Human Rights, Office of the High Commissioner,* https://www.ohchr.org/sites/default/files/documents/publications/guidingprinciplesbusinesshr_en.pdf.

[85] "Case decisions and policy advisory opinions," *Meta Oversight Board*, https://www.oversightboard.com/decision/.

[86] UN General Assembly, *International Covenant on Civil and Political Rights*, 16 December 1966, United Nations, Treaty Series, vol. 999, p. 171, https://www.ohchr.org/en/instruments-mechanisms/instruments/international-covenant-civil-and-political-rights .

87 UN General Assembly, *International Convention on the Elimination of All Forms of Racial Discrimination*, December 21, 1965, United Nations, Treaty Series, vol. 660, p. 195, https://www.ohchr.org/en/instruments-mechanisms/instruments/international-convention-elimination-all-forms-racial ,

Commissioner for Human Rights. Importantly, the RPA establishes a six-part threshold test for context, speaker, intent, content and form, extent of dissemination, and likelihood of imminent harm that should be satisfied before speech falls outside the protection of Article 19. Relevant to this and as noted by Joan Barata, the Meta Oversight Board "usually takes [Article 19 and Article 20] as the main direct legal reference regarding the freedom of expression" while also using "widely recognized standards established within the context of the [RPA]."[88]

Article 4(a) of the CERD obliges ratifying states to punish by law "all dissemination of ideas based on racial superiority or hatred, incitement to racial discrimination." Article 4(b) provides that states must "declare illegal and prohibit organizations, and also organized and all other propaganda activities, which promote and incite racial discrimination." However, these restrictions must be adopted with "due regard" to other human rights, including freedom of expression.

The UN's Human Rights Committee (HRC), which monitors the implementation of the ICCPR, highlights that Article 20(2) secures the right of persons to be free from hatred and discrimination but underlined that it is "crafted narrowly" to ensure freedom of expression. It recalled that free speech may incorporate 'deeply offensive' speech and speech that is disrespectful for a religion.[89] Significantly, in 2011 the HRC highlighted that laws penalizing the expression of opinions about historical facts (such as the Holocaust) are incompatible with Article 19 of the ICCPR. [90]

In their policies, some platforms not only refer to international UN human rights instruments but also to regional human rights instruments, such as the European and Inter-American systems, and national bills of rights. For instance, Twitter´s human rights policy mentions both the US Bill of Rights and The European Convention on Human Rights (ECHR),[91] whereas Meta references the Charter of Fundamental Rights of the EU and the American Convention on Human Rights. The UN approach is set apart from the European human rights system, as the European Human Rights system is a mix of both EU and Council of Europe instruments, namely the Charter of Fundamental Rights of the European Union and the ECHR. Whilst the EU has not (yet) acceded to the ECHR, the latter impacts the EU framework with its Charter reaffirming the rights emanating from the ECHR as interpreted by the ECtHR. While hate speech cases have been dealt with by the Court of Justice of the European Union under the non-discrimination framework, the general approach of the EU

[88] Joan Barata, 'The Decisions of the Oversight Board from the Perspective of International Human Rights Law' (2023) *Global Freedom of Expression Columbia University,* p.10

89 Mohamed Rabbae, A.B.S and N.A v The Netherlands, Communication no. 2124/2011 (14 July 2016) CCPR/C/117/D/2124/2011, para. 10(4).

[90] General Comment No. 34: Article 19: Freedoms of Opinion and Expression <https://www2.ohchr.org/english/bodies/hrc/docs/gc34.pdf> Para.49

[91] "Defending and respecting the rights of people using our service," *Twitter Help Center*, https://help.twitter.com/en/rules-and-policies/defending-and-respecting-our-users-voice .

towards the hate speech in, for example, the DSA and the EU Code of Conduct on Illegal Hate Speech Online, mirrors the ECtHR's restrictive approach.

Specifically, the ECtHR provides significantly less protection to hate speech. It has taken a broad view of what hate speech entails, finding that permitted restrictions on freedom of expression include "not only calls for violence or other criminal acts" but also "insults, ridicule, and slander," constituting "prejudicial speech, moreover, the ECtHR routinely denies any protection to Holocaust denial."[92] This approach sets the ECtHR´s markedly apart from the much more robust tripartite test under the UN approach. Strikingly, in a Grand Chamber decision of May 2023, the ECtHR found that politicians have a responsibility to monitor hateful comments that may arise under their posts, thereby extending positive obligations to such individuals in relation to regulation this party posts.[93]

The DSA imposes obligations on the broadly defined area of "illegal content" which is to cover "hate speech" but also "unlawful discriminatory content" with no accompanying definitions. The associated "vagueness and broadness may trigger over-removals of content."[94]

The Code of Conduct on Illegal Hate Speech Online provides that removal notifications should be reviewed against their own rules and community guidelines and where necessary national laws transposing the EU Framework Decision on combating certain forms and expressions of racism and xenophobia.[95] The reference to the companies' own terms which differ, may lead to a non-uniform interpretation and application of the Code's application.

Thus, there are many different definitions and interpretations of hate speech in national, regional, and international human rights standards, with ICCPR Article 19 providing the strongest protection to free speech vis-à-vis hate speech and the European system providing the lowest level of protection to "hate speech", a concept which it has not authoritatively defined. To identify relevant platform policies in this analysis, we develop a coding rule that aggregates these definitions. By basing a coding rule on the full scope of these definitions, we are likely to capture

[92] Féret v. Belgium, no. 15615/07, para. 72 (ECHR 16 July 2009) as followed in Vejedland v Sweden, App. No 1813/07, para.54 (ECHR 9 May 2012).
[93] Sanchez v France, App. No 45581/15 (ECHR 29 June 2022) + 6][+/ ``
[94] Joan Barata, 'The Digital Services Act and its Impact on the Right to Freedom of Expression: Special Focus on Risk Mitigation Obligations' Platforma por la Libertad de Infoacion, 15
[95] Council Framework Decision 2008/913/JHA of 28 November 2008 on combating certain forms and expressions of racism and xenophobia by means of criminal law, *OJ L 328, 6.12.2008, p. 55–58 (BG, ES, CS, DA, DE, ET, EL, EN, FR, IT, LV, LT, HU, MT, NL, PL, PT, RO, SK, SL, FI, SV), Special edition in Croatian: Chapter 19 Volume 016 P. 141 - 144*, https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=celex%3A32008F0913

a broad range of platform provisions relevant to hate speech. *Thus, for the purposes of this analysis, we code a platform policy as a "hate speech" provision if any part of it*[96]*:*

- Uses the words "hate" or "hatred," in conjunction with speech or content.
- Mentions the following types of content AND specifies that such content is prohibited if it targets individuals or groups on the basis of particular characteristics related to identity.
  - Incitement to or threats of violence
  - Incitement or promotion of hatred
  - Attacks
  - Discriminatory language or calls for discrimination.
  - Pejorative language, such as slurs.

While we use the above coding rule to identify platform policies that are relevant to hate speech, we do not mean to endorse this coding rule as the appropriate definition of hate speech. Rather, this coding rule's breadth ensures that we will capture cases where platforms named and prohibited the concept of hate speech directly, as well as instances where companies described the concept of hate speech – and prohibited it - but did not name it. Importantly, it also ensures that we do not capture cases where a platform simply prohibited objectionable or harmful content regardless of whether it targeted individuals based on protected characteristics. This approach ensures that only policies that explicitly deal with hate speech, and not just with offensive content, are analyzed. It is important to note that this coding rule is used to identify relevant provisions, but the entirety of the content covered by that provision is recorded and analyzed, even if it goes beyond the content listed in the coding rule. In addition to recording the content covered by each platform's hate speech policies, we collect quantitative information on the number of characteristics that are protected from hate speech in the relevant provisions (also known as 'protected characteristics'). All the data we collected for this analysis are available online.

After mapping the evolution of each platform's hate speech policies, we analyze changes in the policies' scope since inception. We also compare these policies to Articles 19 and 20 of the ICCPR, given that six of the eight platforms analyzed have publicly committed to upholding international human rights standards and that the Special Rapporteur on freedom of expression and opinion has recommended platforms adopt IHRL as a framework for hate speech policies. Evelyn Aswad and David Kaye (the latter who served as UN Special Rapporteur on Freedom of Expression and Opinion from 2014-2020) have argued that "U.N. treaty bodies, along with expert opinions offered within the U.N. system, have developed an increasingly consistent set of rules governing the

[96] In some cases, we had to use our discretion to determine if an entire policy was relevant to hate speech - or if only some part of it was relevant. For example, if a platform policy is broken into multiple paragraphs, but only the first paragraph mentions hate speech, we would include the entire policy if it was clearly further description of the parts that mentioned hate speech. But if the other paragraphs dealt with entirely different issues, we would only include the paragraph relevant to hate speech.

appropriate boundaries for hate speech laws."[97] The key component of these interpretive rules is the so-called "three-part test" of "legality, necessity and legitimacy," which applies both when formulating and enforcing restrictions on freedom of expression. Legality means that the applicable restrictions on free expression shall be properly enacted and must not be overly vague or broad. Legitimacy indicates that restrictions must only pursue the aims enumerated in Article 19 (3). Necessity entails the restrictions to be the least intrusive means to achieve the legitimate objective and, that such restrictions be proportionate to the interest to be protected. Moreover, when it comes to the mandatory prohibition of hate speech in Article 20(2) the requirements of intent, incitement, and particular harms must be fulfilled. General Comment 34 notes that "Articles 19 and 20 are compatible with and complement each other. The acts that are addressed in Article 20 are all subject to restriction pursuant to Article 19, paragraph 3."[98] As noted by the Special Rapporteur on the Freedom of Opinion and Expression, mere "advocacy of hatred on the basis of national, racial or religious grounds is not an offence in itself."[99]

However, we also recognize that different regions and communities can have different values and norms around tolerance. That being said, we do not use domestic laws as a point of comparison, as it would be difficult to analyze the alignment of platform policies with every existing domestic hate speech law. Moreover, if governments request content removal on the basis of existing local law, platforms typically must comply or risk serious consequences that in some instances have resulted in being blocked or threatened with being banned from entire countries, such as in Turkey, India, and Nigeria. However, from the outset, their global Terms of Use and Community Guidelines represent the rules for all users, regardless of location, and thus should not necessarily be guided by individual local legislation. We do compare the overall scope of platform policies to European regional human rights standards in Part 2, however, though it is worth nothing that Aswad and Kaye argue that regional standards should not be able to supersede international counterparts on global, rather than regional, platforms.[100]

This analysis is aided by tables that map the content covered, as well as the characteristics protected, by a platform's hate speech policies in each year since inception. The rows in these tables represent categories of content prohibited by or characteristics protected by at least one of the eight platforms' hate speech policies. Thus, the tables provide a rough representation of the broadest potential scope that a hate speech policy could cover, thereby facilitating an

[97] Evelyn Aswad & David Kaye, 'Convergence & Conflict: Reflections on Global and Regional Human Rights Standards on Hate Speech.' (2022) 20 *Northwestern Journal of Human Rights* 3, pg. 168.
[98] General Comment No. 34: Article 19: Freedoms of Opinion and Expression <https://www2.ohchr.org/english/bodies/hrc/docs/gc34.pdf> Para.53
[99] Evelyn Aswad & David Kaye, 'Convergence & Conflict: Reflections on Global and Regional Human Rights Standards on Hate Speech.' (2022) 20 *Northwestern Journal of Human Rights* 3, pg. 212.
[100] Evelyn Aswad & David Kaye, 'Convergence & Conflict: Reflections on Global and Regional Human Rights Standards on Hate Speech.' (2022) 20 *Northwestern Journal of Human Rights* 3, pg. 168.

assessment of each platform's policies relative to their potential broadness. Separate tables for each platform are included in our analysis.

Finally, we assess whether there are any noticeable correlations between changes in policy scope and changes in enforcement volume, using public information about platform enforcement actions where available. It is important to note that these assessments are largely a descriptive enterprise; they are a discussion of whether noticeable changes in enforcement volume seem to occur after changes in policy scope. There are several factors that can impact changes in enforcement volume, however, including the overall prevalence of hate speech on a platform and changes in platforms' hate speech detection algorithms or human review capacity. Moreover, platforms may announce a policy change but take several months to ramp up enforcement, since it can take a long time to train classifiers.[101]

Because platforms do not provide external researchers with access to data on content actioned vs. content that is not actioned or changes in enforcement detection capabilities, it would be difficult to precisely identify the cause of any changes in enforcement volume. More broadly, it is impossible to know how companies are actually enforcing their policies. As Alex Feerst, the CEO of Murmuration Labs and a long-time Trust & Safety advisor, put it, no one really knows how platforms have enforced their publicly stated rules over time. "We have a set of random maybe but probably not representative anecdotes, some events that were big enough to catch attention on social media or old media or from government or civil society groups," Feerst explains, "But the gap in understanding between rules as written and rules as enforced is an inevitable problem with the fact that data sets about moderation as it has happened in reality are almost entirely kept within each company. In other words, [companies] make public the policy 'menu' and sometimes an abbreviated 'recipe,' or two, but not how each kitchen executes their recipes or trains line chefs over time or in relation to each other. Moreover, the various parties who see their role as calling attention to particular flawed or inconsistent moderation practices, including incumbent media organizations, independent researchers, civil society watchdogs, academic researchers, and government actors, each have their own set of biases and mixed incentives to emphasize anecdotal evidence convenient to their particular respective narratives and particular agendas."[102]

It follows that causally identifying the effect of changes in policy scope on enforcement volume would also be challenging without additional access to platform data. Regulation in the U.S. and E.U. seeks to introduce data-sharing requirements for platforms, and many observers have called on platforms to create libraries of removed content for research purposes. Thus, causal assessments of the impact of changes in policy scope may be possible down the line. We hope

[101] Paris Martineau, "YouTube Removes More Videos but Still Misses a lot of Hate," *Wired*, September 4, 2019, https://www.wired.com/story/youtube-removes-videos-misses-hate/ .
[102] Comments by Alex Feerst to Jacob Mchangama, May and June 2023.

that future researchers will use the data we collected to facilitate our analysis to aid in such analyses when they are possible.

Part I: Mapping the Evolution of Platform Hate Speech Policies

This section describes the evolution of the eight platforms' hate speech rules, via an analysis of each platform's Terms of Service and Community Standards, Guidelines, or Policies, and describes changes in enforcement volume that may be correlated with changes in policy scope.

1. Facebook

- **Release/Launch Date:** February 4, 2004
- **Number of Users/Visitors:** 2.910 billion monthly active users[103]
- **Short Overview of Content Moderation Process:** Content moderators review posts that have been flagged by AI or reported by users. The majority of this work is outsourced to third-party vendors.[104]
- **Signatory to the EU'S Code of Conduct on Illegal Hate Speech Online?** Yes

[103] "Most popular social networks worldwide as of January 2023, ranked by number of monthly active users," *Statista*, https://www.statista.com/statistics/272014/global-social-networks-ranked-by-number-of-users/ (accessed on April 30, 2023).
[104] John Koetsier, "Report: Facebook Makes 300,000 Content Moderation Mistakes Every Day," Forbes, June 9, 2020, https://www.forbes.com/sites/johnkoetsier/2020/06/09/300000-facebook-content-moderation-mistakes-daily-report-says/?sh=7edb3e6454d0#:~:text=Facebook%20employs%20about%2015%2C000%20content,meets%20or%20violates%20community%20standards.

Key Developments

Terms of Use

Facebook was originally a static page that could be accessed only by persons with harvard.edu emails,[105] but today, Facebook (under its parent company Meta) is a global social network giant with more than 2.7 billion monthly active users.[106] In the first few years of its existence, Facebook "lacked a robust team for removing problematic content" and, at the same time, "had no real content-moderation policy to speak of,"[107] though it did have Terms of Use. Facebook's first Terms of Use, which date to 2004, did not include a hate speech provision. While Facebook reserved the right to review and delete any content which "might be offensive, illegal, or that might violate the rights, harm, or threaten the safety of Members," the provision did not stipulate that offensive or harmful content is prohibited *if it targets people on the basis of specific identity-based characteristics*. In 2005, however, Facebook added a hate speech provision, prohibiting users from posting content deemed "hateful, or racially, ethnically or otherwise objectionable."[108]

In 2009, however, Facebook removed the above reference and overhauled the Terms of Use. In the new terms, under the "Safety" section, Facebook prohibited posting "content that is hateful" (see Figure 1). The reference to content that was objectionable on racial and ethnic terms disappeared, while the prohibition on 'hateful' content remained. In other words, the provision became more generic. By 2013, the wording of this provision had evolved to prohibit "hate speech." By September 2018, Facebook had removed all the above references, and the Facebook Terms no longer included a hate speech provision. Today, prohibitions on this type of content are covered by Facebook's Community Standards.

[105] David Kirkpatrick, *The Facebook Effect: The Inside Story of the Company that is Connecting the World*, Simon and Schuster, 2011, 82-83.
[106] "Most popular social networks worldwide as of January 2023, ranked by number of monthly active users."
[107] Kate Klonick, "The Facebook Oversight Board: Creating an independent institution to adjudicate online free expression," *Yale Law Journal* 129, no. 2418 (2020), 2436.
[108] https://web.archive.org/web/20050826155708/http://www.thefacebook.com/terms.php

Figure 1

3. Safety

We do our best to keep Facebook safe, but we cannot guarantee it. We need your help in order to do that, which includes the following commitments:

1. You will not send or otherwise post unauthorized commercial communications to users (such as spam).

5. You will not bully, intimidate, or harass any user.

6. You will not post content that is hateful, threatening, pornographic, or that contains nudity or graphic or gratuitous violence.

8. You will not use Facebook to do anything unlawful, misleading, malicious, or discriminatory.

Community Standards

Despite some relevant provisions in early versions of Facebook's Terms of Use, the key policy developments relevant to this report exist in Facebook's Community Standards. The first traceable Community Standards are from 2011, and they began with an acknowledgement of the challenging line between protecting free expression and protecting the rights of others. This initial version of the Community Standards included a prohibition on hate speech, which implied the concept was defined by "singling out" people on the basis of nine identity-related characteristics (see Figure 2). This threshold for content to be considered hate speech is significantly lower than the ICCPR prohibition on advocacy to national, racial, or religious hatred that constitutes incitement to discrimination, hostility, or violence.

Figure 2

Hate Speech	Facebook does not tolerate hate speech. Please grant each other mutual respect when you communicate here. While we encourage the discussion of ideas, institutions, events, and practices, it is a serious violation of our terms to single out individuals based on race, ethnicity, national origin, religion, sex, gender, sexual orientation, disability, or disease.

Facebook's hate speech prohibition narrowed slightly a year later, when the company updated the rule to prohibit "attacks" based on protected characteristics (see Figure 3), arguably a higher threshold than a prohibition on "singling out" an individual based on their identity. While Facebook made a minor revision in 2013, recognizing the existence of humorous speech, the next major update to the hate speech provision occurred in 2015. While the company added a sentence

banning organizations dedicated to promoting hatred, the changes mostly involved a discussion of the company's approach to educational content and satire. Facebook acknowledged that people might share content "containing someone else's hate speech" to raise awareness or educate others about that harmful speech, in which case the company expected the user to clearly indicate the purpose of sharing that content. Facebook also asked users to associate their name and profile with any satire related to hate speech, since people tend to be more responsible when they can be held accountable for potentially insensitive content.

Figure 3

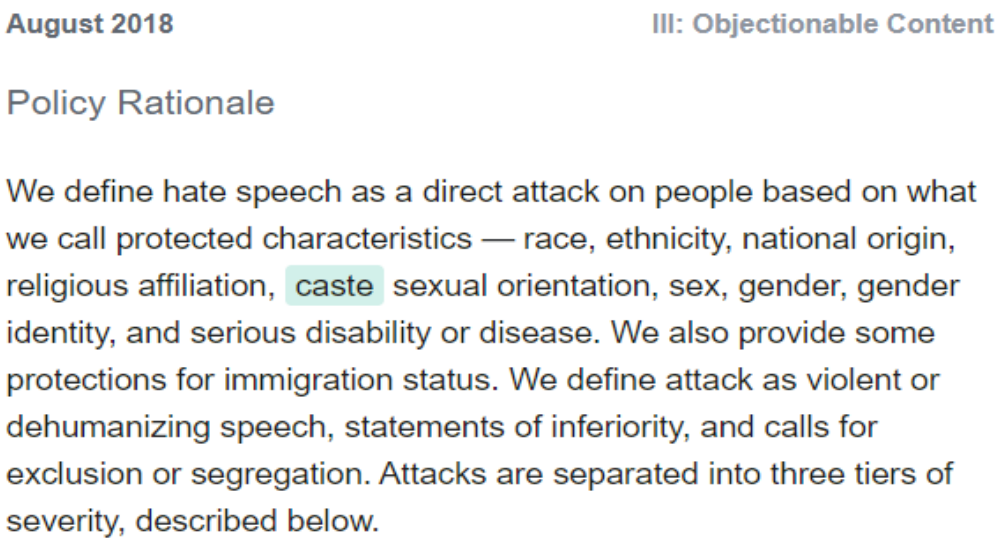
The next notable change in Facebook's hate speech provisions came in August 2018, when the company first defined "attack" in the context of its hate speech prohibition (see Figure 6). "We define attack as violent or dehumanizing speech, statements of inferiority, and calls for exclusion or segregation," the updated provision read. The company also explained that it separated attacks based on protected characteristics into three tiers of severity (see Figure 4).[109] All tiers included attacks targeting persons or groups with one or more protected characteristics, but they differed in the way attack was defined. In the first tier, attacks are defined as any violent speech, dehumanizing speech, or efforts to mock hate crimes or their victims. Tier 2 included attacks defined as statements of inferiority, expressions of contempt, or expressions of disgust (including cursing). Tier 3 included attacks defined as calls to exclude or segregate, except for in the context of criticizing immigration policies, and content that describes or negatively targets people with slurs. Over the next four years, the specific outline of each tier underwent several changes, though Facebook's conceptualization of hate speech as an attack remained. Facebook's updates to the specifics of each tier are available in the Change Log for the Hate Speech policy, available on Meta's Transparency Center.

While Facebook initially stated that the tiers corresponded to levels of severity, that sentence has now been removed from the policy rationale. Moreover, Facebook never explained whether it applied any differential enforcement mechanisms to hate speech based on the relevant severity

[109] Heather Kelly, "Facebook reveals its internal rules for removing controversial posts," *CNN Money*, April 24, 2018, https://money.cnn.com/2018/04/24/technology/facebook-community-standards/ .

tier. In fact, the company explicitly states that users should not post content in any of the tiers. Thus, while the addition of the tiers in the hate speech policy provides more specifics about the precise types of content that are covered by the policy, it does not provide insight into why Facebook categorizes hate speech into these tiers.

Figure 4



In August 2020, Facebook expanded its definition of hate speech to include "harmful stereotypes," in addition to violent and dehumanizing speech, statements of inferiority, and calls for exclusion or segregation.[110] The next month, in September 2020, Facebook listed "expressions of contempt, disgust or dismissal," as well as "cursing," in the explicit definition at the beginning of the policy and listed them as Tier 2 attacks, while these types of content had previously only existed under Tier 2 attacks.[111] Later that year, in October 2020, Facebook also added a prohibition on "any content that denies or distorts the Holocaust" to the hate speech policy.[112] Interestingly, this move contradicted CEO Mark Zuckerberg's previous position that such content should not be banned. In an earlier public Facebook post, Zuckerberg had written:

[110] "Hate speech," Meta Transparency Center, https://transparency.fb.com/policies/community-standards/hate-speech/ (accessed February 1, 2023).
[111] "Hate speech," Meta Transparency Center, https://transparency.fb.com/policies/community-standards/hate-speech/ (accessed February 1, 2023).
[112] "Hate speech," Meta Transparency Center, https://transparency.fb.com/policies/community-standards/hate-speech/ (accessed February 1, 2023).

"I've struggled with the tension between standing for free expression and the harm caused by minimizing or denying the horror of the Holocaust. My own thinking has evolved as I've seen data showing an increase in anti-Semitic violence, as have our wider policies on hate speech. Drawing the right lines between what is and isn't acceptable speech isn't straightforward, but with the current state of the world, I believe this is the right balance".[113]

Facebook did, however, take steps to limit the scope of its hate speech definition a year later, in June 2021. Facebook explained that, after much stakeholder consultation, it had decided to "define hate speech as a direct attack against people – rather than concepts or institutions."[114] The update also explained that the company would require additional information or content to remove "content attacking concepts, institutions, ideas, practices, or beliefs associated with protected characteristics, which are likely to contribute to imminent physical harm, intimidation or discrimination against the people associated with that protected characteristic."[115] Previously, the policy rationale stated: we "define hate speech as a direct attack against people," so this annotation introduced a previously unspecified limit to the company's definition.

Moreover, in November 2021, Facebook introduced a satirical exemption to the prohibition against hate speech on Facebook.[116] This exemption provides for Facebook to allow content that may otherwise violate the Community Standards when the company determines that the content is satirical. Content will only be allowed if the violating elements of the content are being satirized or attributed to something or someone else in order to mock or criticize them. The change was in response to a decision by the Oversight Board overturning Facebook's decision to remove a meme criticizing the Turkish government in relation to the Armenian Genocide.[117]

In July 2022, the company updated the hate speech policy rationale to clarify elements of enforcement surrounding slurs.[118] While the company does not tolerate slurs used to attack people on the basis of protected characteristics, it recognized that "people sometimes share content that includes slurs or someone else's hate speech to condemn it or raise awareness" or in a "self-referential" or "empowering" way. In those cases, Facebook required users to make their intentions clear. This change essentially updated the previous acknowledgment that people might share "someone else's hate speech" for educational purposes to include sharing "slurs" in an

[113] Facebook Post from Mark Zuckerberg, October 12, 2020, https://www.facebook.com/zuck/posts/10112455086578451
[114] "Hate speech," Meta Transparency Center, https://transparency.fb.com/policies/community-standards/hate-speech/ (accessed February 1, 2023).
[115] https://transparency.fb.com/policies/community-standards/hate-speech/
[116] "Hate speech," Meta Transparency Center, https://transparency.fb.com/policies/community-standards/hate-speech/ (accessed February 1, 2023).
[117] "Case on a comment related to the Armenian people and the Armenian Genocide," Meta Transparency Center, Jule 13, 2022, https://transparency.fb.com/en-gb/oversight/oversight-board-cases/comment-related-to-armenian-people-and-the-armenian-genocide/.
[118] "Hate speech," Meta Transparency Center, https://transparency.fb.com/policies/community-standards/hate-speech/ (accessed February 1, 2023).

educational or self-referential way. Thus, this change reflected an additional exception to the enforcement of the hate speech policy. Since then, Meta has made small tweaks to the wording of the policy, but there have been no major changes as of April 1, 2023.

Facebook's Dangerous Individuals and Organizations policy has also included provisions relevant to hate speech. In 2017, the company began banning content that expresses support for organized hate groups, including support or praise for the leaders of these organizations.[119] By 2019, the policy offered a definition of organized hate, stipulating that a hate organization was "any association of three or more people that is organized under a name, sign, or symbol and that has an ideology, statements, or physical actions that attack individuals based on characteristics, including race, religious affiliation, nationality, ethnicity, gender, sex, sexual orientation, serious disease, or disability."[120] In addition to banning content that expressed support for the group or its leadership, Facebook introduced a ban on symbols that represent hate groups. In 2020, the company introduced a prohibition on content that supports hateful ideologies, defined as "beliefs that are inherently tied to violence and attempts to organize people around calls for violence or exclusion of others based on their protected characteristics," including Nazism, White Supremacy, White Nationalism, and White Separatism.[121]

The list of protected characteristics covered by Facebook's hate speech policy has also changed several times over the years. The hate speech provision in the 2005 Terms of Use mentioned race and ethnicity, but this reference disappeared in later versions of the Terms. Moreover, the hate speech provision in Facebook's initial Community Standards referenced seven additional protected characteristics: national origin, religion, sex, gender, sexual orientation, disability, and disease - suggesting the scope of hate speech prohibited by Facebook had increased by 2011. In 2015, Facebook added gender identity as a protected characteristic, and by 2018, the company had also added caste and immigration status to the list.

The protected characteristics list further expanded in March 2020, when "age" was added if it was "paired with another protected characteristic."[122] In September 2020, protection was extended to "occupation" when "occupation" is referenced alongside another protected characteristic.[123] It is unclear why these characteristics are not protected on their own; moreover, if they need to be referenced alongside another protected characteristics to be protected, it is not clear why they

[119] https://web.archive.org/web/20171120221946/https://www.facebook.com/communitystandards/
[120] "Dangerous Organizations and Individuals" Meta Transparency Center, https://transparency.fb.com/policies/community-standards/dangerous-individuals-organizations/ (accessed February 1, 2023).
[121] "Dangerous Organizations and Individuals" Meta Transparency Center, https://transparency.fb.com/policies/community-standards/dangerous-individuals-organizations/ (accessed February 1, 2023).
[122] "Hate speech," Meta Transparency Center, https://transparency.fb.com/policies/community-standards/hate-speech/ (accessed February 1, 2023).
[123] "Hate speech," Meta Transparency Center, https://transparency.fb.com/policies/community-standards/hate-speech/ (accessed February 1, 2023).

are even listed as part of the policy. In September 2020, Facebook also provided more specifics about protections for 'immigration status,' replacing the term with "refugees, migrants, immigrants, and asylum seekers."[124]

Interestingly, in December 2020, several news outlets reported that Facebook was no longer assessing all protected characteristics equally. According to the Washington Post, the effort was aimed at overhauling the company's hate speech detection algorithms, which had regularly removed slurs against White people while flagging and removing innocuous posts from people of color.[125] Thus, Facebook began prioritizing the removal of anti-Black hate speech over hate speech directed at white people, men and Americans, to address the disproportionate effects that hate speech has on minority groups. The changes were also directed at tackling hate speech against Muslims, Jews, and members of the LGBTQ+ community.[126] A company spokesperson told the Washington Post: "We know that hate speech targeted towards underrepresented communities can be the most harmful, which is why we have focused our technology on finding the hate speech that users and experts tell us is the most serious."

However, a group that is underrepresented in one state may not be underrepresented in another. For example, Muslims are not underrepresented in the 40+ countries where Muslims make up over 50% of the population, and Jews are not underrepresented in Israel. In the United States, 75.8% of the population is white, so people of color (defined as someone who is not white)[127] are a minority.[128] However, in many countries around the world, people of color (in itself a nebulous term) constitute an overall majority, though certain non-white racial or ethnic groups may still be a minority. Thus, what distinguishes a vulnerable population in one state or region is different from what constitutes a vulnerable population in another, but the enforcement change that Meta allegedly introduced does not appear to reflect that.

Analysis of Policy Scope

Over time, Facebook has offered far more specificity in the content covered by its hate speech prohibitions. Though heightened specificity was associated with narrowed scope in a few cases, adding details typically corresponded to broader policy coverage. Early versions of Facebook's hate speech provisions banned "hateful" and "racially or ethnically objectionable" content, without

[124] "Hate speech," Meta Transparency Center, https://transparency.fb.com/policies/community-standards/hate-speech/ (accessed February 1, 2023).
[104] Elizabeth Dwoskin, Nitasha Tiku, and Heather Kelly, "Facebook to start policing anti-Black hate speech more aggressively than anti-White comments, documents show," *Washington Post*, December 3, 2020, https://www.washingtonpost.com/technology/2020/12/03/facebook-hate-speech/ .
[126] Nick Statt, "Facebook is stepping up moderating against anti-Black hate speech," *The Verge*, December 3, 2020, https://www.theverge.com/2020/12/3/22150964/facebook-moderation-anti-black-hate-speech-policy-change
[127] Merriam-Webster Dictionary, "Person of Color," https://www.merriam-webster.com/dictionary/person%20of%20color
[128] "Census Facts," United States Census Bureau, https://www.census.gov/quickfacts/fact/table/US/PST045221 (accessed April 15, 2023).

offering any details about what those terms meant in practice. This lack of clarity provided little information about the scope of content prohibited under the hate speech policy. In 2011, however, the company implied that hate speech involved "singling out" individuals based on protected characteristics, a very broad conceptualization of the term. A year later, in 2012, the definition narrowed to "attacks" based on protected characteristics. By 2023, however, the policy had expanded to cover incitement to violence, attacks, praise or support for organized hate groups, dehumanizing speech, statements of inferiority, expressions of contempt and disgust, mocking historical atrocities, calls for exclusion and segregation, slurs, harmful stereotypes, and cursing. Though the company recently clarified that its prohibitions generally only apply to attacks on people, rather than on concepts, Facebook's hate speech policy is considerably broader than it was in 2012. Table 1 illustrates these changes.

Table 1

Content Explicitly Covered by Facebook's Hate Speech Policies		2005	2006	2007	2008	2009	2010	2011	2012	2013	2014	2015	2016	2017	2018	2019	2020	2021	2022	2023
Hate(ful) speech/ content		X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X
Promotion of Hatred												X	X	X	X					
Support for Organized Hate (Including Symbols)														X*	X*	X*	X*	X*	X*	X*
On the basis of protected characteristics	Incitement to or Threats of Violence														X	X	X	X	X	X
	Attacks								X	X	X	X	X	X	X	X	X	X	X	X
	Statements of inferiority or content that demeans														X	X	X	X	X	X
	Dehumanization														X	X	X	X	X	X
	Expressions of contempt or disgust														X	X	X	X	X	X
	Calls for exclusion or segregation														X	X	X	X	X	X
	Discrimination																			
	Denying or mocking historical atrocities, or valorizing the perpetrators														X	X	X	X	X	X
	Slurs														X	X	X	X	X	X
	Harmful Stereotypes																X	X	X	X
	Conspiracy Theories																			
	Cursing														X	X	X	X	X	X

* Support for organized hate is banned by Facebook's Dangerous Individuals and Organizations policy.

The content currently covered by Facebook's hate speech policy covers the full range of content described by Article 20(2). In addition to violent speech, attacks, and calls for exclusion (a form of discriminatory language), which align with Article 20 (2), Facebook prohibits other forms of content, such as slurs, denying historical events, and cursing at members of protected groups that is neither covered by the mandatory prohibition of hate speech in Article 20(2), nor aligned with the permissible restrictions on free speech under Article 19 and the strict requirements of legality, legitimacy, and necessity.

Table 2 demonstrates that the scope of Facebook's protected characteristics has also expanded over time. Since 2005, Facebook has added protections for national origin, religion, sex, gender, sexual orientation, disability, disease, gender identity, immigration status, caste, age, and occupation to the platform's initial protections for race and ethnicity. Moreover, since the creation of the Community Standards in 2011, the scope of protected characteristics covered by the hate speech policy has been broader than those listed in Article 20(2). Facebook's 2011 hate speech policy included several characteristics not mentioned in those definitions of hate speech, namely sex, sexual orientation, disability, and disease. Since then, Facebook has also added caste, as well as age and occupation when paired with another characteristic, to the list, further expanding the scope of hate speech prohibited by Facebook beyond Article 20(2).

Table 2

Characteristics Protected in Facebook's Hate Speech Policies																			
	2005	2006	2007	2008	2009	2010	2011	2012	2013	2014	2015	2016	2017	2018	2019	2020	2021	2022	2023
Total	**2**	**2**	**2**	**2**	**0**	**0**	**9**	**9**	**9**	**9**	**10**	**10**	**10**	**12**	**12**	**14**	**13**	**13**	**13**
Race	X	X	X	X			X	X	X	X	X	X	X	X	X	X	X	X	X
Ethnicity	X	X	X	X			X	X	X	X	X	X	X	X	X	X	X	X	X
National Origin							X	X	X	X	X	X	X	X	X	X	X	X	X
Religion							X	X	X	X	X	X	X	X	X	X	X	X	X
Gender							X	X	X	X	X	X	X	X	X	X			
Color																			
Immigration Status														X	X	X	X	X	X
Sex							X	X	X	X	X	X	X	X	X	X	X	X	X
Gender Identity											X	X	X	X	X	X	X	X	X
Sexual Orientation							X	X	X	X	X	X	X	X	X	X	X	X	X
Age																X	X	X	X
Disability							X	X	X	X	X	X	X	X	X	X	X	X	X
Disease/ Medical Condition							X	X	X	X	X	X	X	X	X	X	X	X	X
Veteran Status																			
Occupation																X	X	X	X
Weight																			
Pregnancy																			
Caste														X	X	X	X	X	X
Victims of a Major Event																			
Socio-economic Status																			
Culture																			
Tribe																			

Notes: An X indicates the company's hate speech policies covered that protected characteristic for at least one month during the given year.

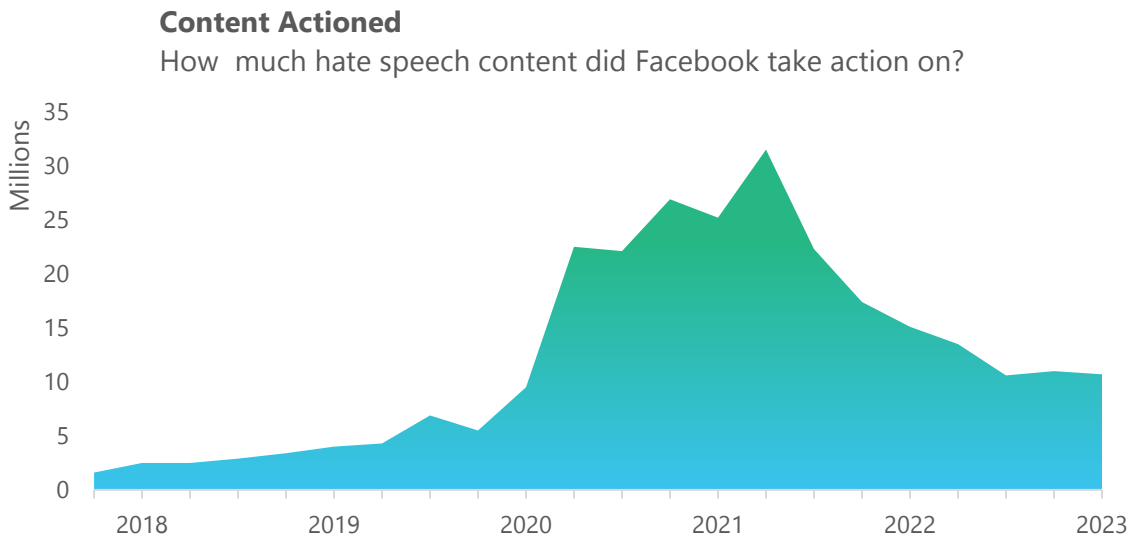
Changes in Enforcement Volume

Facebook regularly publishes a Community Standards Enforcement Report, which shares metrics related to the prevalence of violating content, the amount of content actioned for violating policies, and the volume of enforcement actions that are appealed and/or overturned. As Figure 5 demonstrates, the amount of content that Facebook removed due to hate speech violations went from below 5 million in late 2017, to above 30 million in early 2021, to a little over 10 million

in Q4 2022. It is not clear that changes in policy scope drove these changes. As detailed in the previous section, Facebook's hate speech policies significantly expanded in scope in August 2018 and then again in August 2020. Figure 5 does not show a large increase in the amount of content actioned under the hate speech policy around August 2018. While Facebook removed far more content for hate speech violations in the second quarter of 2020, compared to previously, the August 2020 addition of harmful stereotypes occurred in Q3 2020. For its part, Facebook attributed the 2020 increase in hate speech removals to improvements in hate speech classifiers.[129] From Q3 2021 to Q3 2022, there were consistent reductions in the amount of hate speech actioned on Facebook, but Facebook also estimated that the prevalence of hate speech on Facebook fell during this time.[130]

All of this information suggests that a variety of factors can impact the amount of content Facebook removes under its hate speech policies. Thus, for external researchers to assess how the 2018 and 2020 increases in policy scope impacted enforcement volume, Facebook would need to give researchers access to data on actioned and non-actioned content, as well as information about changes to hate speech classifiers and human review capacity.

Figure 5[131]



[129] See Guy Rosen, "Community Standards Enforcement Report, May 2020 Edition," *Meta Newsroom*, May 12, 2020, https://about.fb.com/news/2020/05/community-standards-enforcement-report-may-2020/, and Guy Rosen, "Community Standards Enforcement Report, November 2020," *Meta Newsroom*, November 19, 2020, https://about.fb.com/news/2020/11/community-standards-enforcement-report-nov-2020/ .

[130] "Hate speech," Meta Transparency Center, https://transparency.fb.com/policies/community-standards/hate-speech/ (accessed February 1, 2023).

[131] "Community Standards Enforcement Report," *Meta Transparency Center*, https://transparency.fb.com/data/community-standards-enforcement/ .

2. INSTAGRAM

- **Launch date:** October 6, 2010
- **Number of Users/Visitors:** 2 billion monthly active users [132]
- **Short Overview of Content Moderation Process:** Content moderators review posts that have been flagged by AI, reported by users or non-users. Non-users can file a report available on Instagram's Help Centre. Most of this work is outsourced to third-party vendors.
- **Signatory to the EU'S Code of Conduct on Illegal Hate Speech Online?** Yes

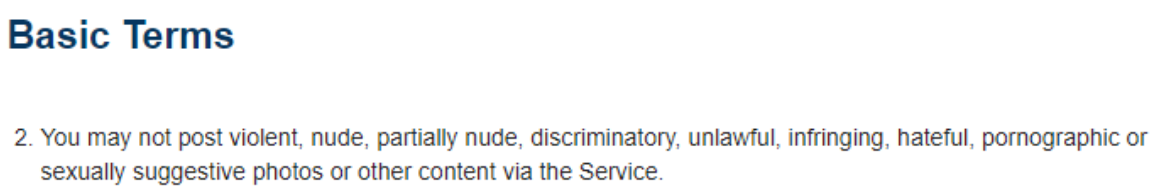
[132] "Most popular social networks worldwide as of January 2023, ranked by number of monthly active users," *Statista*, https://www.statista.com/statistics/272014/global-social-networks-ranked-by-number-of-users/ (accessed on April 30, 2023).

Key Developments

Terms of Use

While the first traceable Terms of Use for Instagram are from 2012, the document did not include a relevant provision on hate speech until 2013 (see Figure 6). The Terms did not define hateful content, however, nor reference any characteristics that Instagram protected from hateful content. By 2018, the provision was deleted from the Terms of Use.

Figure 6[133]



Community Guidelines

The first traceable Instagram Community Guidelines date to 2012, but the Guidelines did not include a hate speech provision until 2015. Under the subheading "respect other members of the Instagram community," the 2015 guidelines noted:

"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. We do generally allow stronger conversation around people who are featured in the news or have a large public audience due to their profession or chosen activities.

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."[134]

The above provision is an amalgamation of various types of objectionable content, ranging from harassment to hate speech, but the provision lacks a clear definition of any of these terms. The second paragraph, however, implies that Instagram considers hate speech to involve encouraging violence against or attacking individuals on the basis of protected characteristics. It is unclear, however, if the company intends to treat credible threats, hate speech, degrading content, blackmail, and harassment as separate types of content – or if the company considers all of these

[133] https://web.archive.org/web/20130123212202/http://instagram.com/about/legal/terms/updated/
[134] https://web.archive.org/web/20150825000805/https://www.facebook.com/help/instagram/477434105621119/

forms of speech to be hate speech, given the second paragraph focuses on hate speech specifically. In July 2020, however, Instagram added hyperlinks to this provision, which implied that the company considers these types of objectionable content separately.[135] The company added a link to the phrase "hate speech" that directed people to the hate speech policy in the Facebook Community Standards. The phrase "credible threats" linked to the Violence & Incitement policy in the Facebook Community Standards, while the phrase "degrade or shame them" linked to the Bullying & Harassment policy in the Facebook Community Standards. Thus, it appears that this provision of the Instagram Community Guidelines corresponds to several different policies within the Facebook Community Standards.

By adding these hyperlinks to the Instagram Community Guidelines, the company implied that Facebook's Community Standards apply to content on Instagram. The scope of the hate speech provision in Instagram's Community Guidelines differs from the hate speech policy in the Facebook Community Standards, however. Instagram's policy references ten protected characteristics, compared to the 16 that Facebook's policy covers. Thus, it is not clear whether Instagram prohibits hate speech against the ten protected characteristics listed in the Instagram Community Guidelines or the 16 listed in the Facebook Community Standards.

The Community Standards Enforcement Report page of the Transparency Center, however, states "Facebook and Instagram share content policies. What is violating on Facebook is violating on Instagram. Throughout this report, we link to our Community Standards, which include the most comprehensive description of these policies."[136] This statement suggests there is no need for two sets of policies and the Community Standards reflects the policies enforced on Instagram. Why then does Instagram continue to list the Community Guidelines on its website? Why are the Community Guidelines listed under "Other Policies" on the Meta Transparency Center? When did the Community Standards become the default rules for both platforms? There have been no updates to the relevant provision in Instagram's Community Guidelines since 2020, so perhaps sometime after that date? It is not clear.

This confusion is problematic in terms of the legality requirement in Article 19 (3) of the ICCPR. Users have no way of knowing what exactly the rules are for Instagram. Thus, even if the Community Standards are, in fact, the rules for both platforms, we still feel it is important to analyze Instagram's Community Guidelines in this report, since the Community Standards Enforcement Report is the only place where Meta clearly states that the Community Standards apply to both.

[135] https://web.archive.org/web/20200730024324/https://help.instagram.com/477434105621119
[136] https://transparency.fb.com/data/community-standards-enforcement/?source=https%3A%2F%2Ftransparency.faceb

Analysis of Policy Scope

As described in the previous section, the scope of content covered by Instagram's hate speech policy is somewhat unclear. The most basic interpretation of the Instagram policy guidance, however, suggests the company has defined hate speech as incitement to violence or attacks based on protected characteristics since 2015, as reflected in Table 3. This scope of covered content aligns with Article 20(2) of the ICCPR. Table 4 illustrates that the scope of characteristics covered by Instagram's policy also has not changed since 2015, though it is much broader than the list of characteristics covered by Article 20 (2).

Table 3

Content Explicitly Covered by Instagram's Hate Speech Policies		2013	2014	2015	2016	2017	2018	2019	2020	2021	2022	2023
Hate(ful) speech/ content		X	X	X	X	X	X	X	X	X	X	X
Promotion of Hatred												
Support for Organized Hate (Including Symbols)				X	X	X	X	X	X	X	X	X
On the basis of protected characteristics	Incitement to or Threats of Violence			X	X	X	X	X	X	X	X	X
	Attacks			X	X	X	X	X	X	X	X	X
	Statements of inferiority or content that demeans											
	Dehumanization											
	Expressions of contempt or disgust											
	Calls for exclusion or segregation											
	Discrimination											
	Denying or mocking historical atrocities, or valorizing the perpetrators											
	Slurs											
	Harmful Stereotypes											
	Conspiracy Theories											
	Cursing											

Table 4

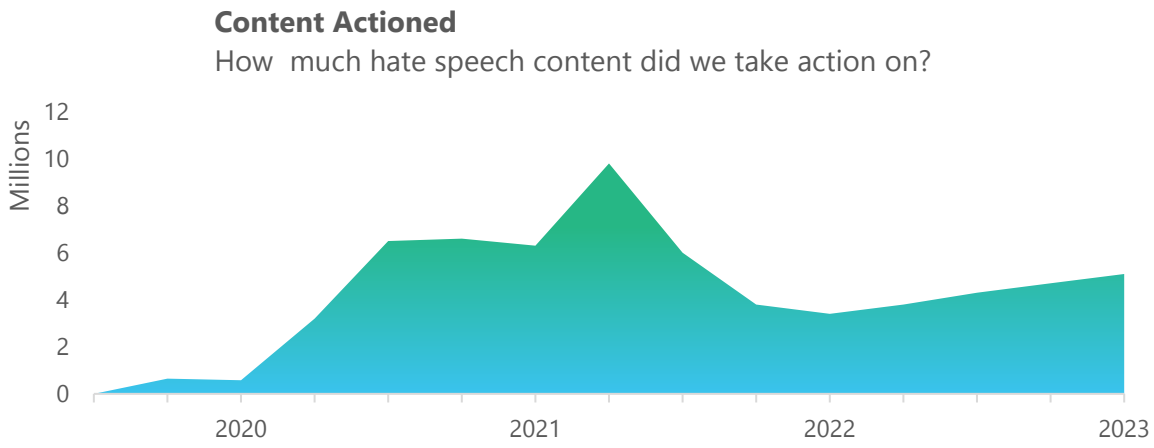
Characteristics Protected in Instagram's Hate Speech Policies											
	2013	2014	2015	2016	2017	2018	2019	2020	2021	2022	2023
Total	**0**	**0**	**10**	**10**	**10**	**10**	**10**	**10**	**10**	**10**	**10**
Race			X	X	X	X	X	X	X	X	X
Ethnicity			X	X	X	X	X	X	X	X	X
National Origin			X	X	X	X	X	X	X	X	X
Religion			X	X	X	X	X	X	X	X	X
Gender			X	X	X	X	X	X	X	X	X
Color											
Immigration Status											
Sex			X	X	X	X	X	X	X	X	X
Gender Identity			X	X	X	X	X	X	X	X	X
Sexual Orientation			X	X	X	X	X	X	X	X	X
Age											
Disability			X	X	X	X	X	X	X	X	X
Disease/ Medical Condition			X	X	X	X	X	X	X	X	X
Veteran Status											
Occupation											
Weight											
Pregnancy											
Caste											
Victims of a Major Event											
Socio-Economic Status											
Culture											
Tribe											

Notes: An X indicates the company's hate speech policies covered that protected characteristics for at least one month during the given year.

Changes in Enforcement Volume

As it does for Facebook, Meta provides a Community Standards Enforcement Report for Instagram. However, these reports do not exist for years prior to 2015, the year Instagram made its only major changes to the scope of its hate speech provisions.[137] Thus, it would be difficult to use this data to assess how changes in the scope of Instagram's Community Guidelines impacted enforcement volumes. Nevertheless, Figure 7, which reports the amount of Instagram content actioned due to violations of the hate speech prohibition, shows substantial changes in this metric over time. Because this report is part of the Community Standards Enforcement Report, it raises questions about whether Instagram's Community Guidelines, or Facebook's Community Standards, represent the final word on what content is and is not allowed on Instagram. Thus, changes in the Community Standards, as documented in the previous section, could possibly drive the changes in enforcement volume depicted in Figure 7. However, Figure 7 shows a large increase in the amount of content actioned in both Q2 and Q3 2020, and there was only a noticeable change in the scope of Facebook's hate speech policy in August 2020. In fact, Meta attributed these 2020 increases in Instagram content actioned for hate speech violations to improvements in proactive detection technology for the English, Spanish, and Arabic languages, and noted that they expected continued fluctuations in these numbers as the company adjusted to COVID-19 related workforce disruptions.[138]

Figure 7[139]



[137] "Hate Speech, Community Standards Enforcement Report," *Meta Transparency Center,* https://transparency.fb.com/data/community-standards-enforcement/hate-speech/instagram/ .

[138] See Guy Rosen, "Community Standards Enforcement Report, August 2020," *Meta Newsroom,* August 11, 2020, https://about.fb.com/news/2020/08/community-standards-enforcement-report-aug-2020/, and Guy Rosen, "Community Standards Enforcement Report, November 2020," *Meta Newsroom*, November 19, 2020, https://about.fb.com/news/2020/11/community-standards-enforcement-report-nov-2020/.

[139] "Community Standards Enforcement Report," Meta Transparency Center, https://transparency.fb.com/data/community-standards-enforcement/ .

3. REDDIT

- **Release/Launch Date:** June 2005
- **Number of Active Users:** 430 million[140]
- **Short Overview of Content Moderation Process:** Reddit has a centralised team of moderators who make up approximately 10% of Reddit's workforce. The majority of the platform's content moderation is decentralised to users who volunteer to moderate content on a particular subreddit.
- **Signatory to the EU'S Code of Conduct on Illegal Hate Speech Online?** No

[140] David Curry, "Reddit Revenue and Usage Statistics (2023)," Business of Apps, January 9, 2023, https://www.businessofapps.com/data/reddit-statistics/.

Key Developments

Reddit's initial User Agreement did not include a hate speech provision. In 2015, Reddit introduced a prohibition on content that "encourages or incites violence" or "threatens, harasses or bullies or encourages others to do so," which CEO Steve Huffman later argued amounted to an "implicit" prohibition on hate speech.[141] Nevertheless, the policy noted that Reddit "generally provides a lot of leeway in what content is acceptable," even if the nature of this content may be "offensive."

This policy was the status quo until June 29, 2020, when the company added an explicit prohibition on hate speech (see Figure 8). This revision was likely spurred by widespread international protests against the police killing of George Floyd, as well as a June 2020 open letter addressed to Huffman and Reddit's Board of Directors, signed by over 600 of the platform's groups (representing thousands of moderators and millions of Reddit subscribers.)[142] The open letter implored Huffman to "stand[] up to racism and hate . . . with real action" by "[e]nact[ing] a sitewide policy against racism, slurs, and hate speech [sic] targeted at protected groups." After the letter, Reddit revised its content policies and banned communities and users that promote hate based on identity or vulnerability.[143]

Figure 8[144]

> *Rule 1: Remember the human. Reddit is a place for creating community and belonging, not for attacking marginalized or vulnerable groups of people. Everyone has a right to use Reddit free of harassment, bullying, and threats of violence. Communities and people that incite violence or that promote hate based on identity or vulnerability will be banned.*

The new policy banned the promotion of hate, which was not defined, based on identity or vulnerability, which was defined as: "Groups based on their actual and perceived race, color, religion, national origin, ethnicity, immigration status, gender, gender identity, sexual orientation, pregnancy, or disability. These include victims of a major violent event and their families." It is worth noting that this 2020 policy initially excluded groups of people who are in the majority from protection. The rule on hate "does not protect all groups or all forms of identity," the policy read. "For example, the rule does not protect groups of people who are in the majority or who promote such attacks of hate."[145] The public, including Reddit users, however, noted that certain groups targeted by hate speech may be part of the majority, such as women. Reddit quickly tackled this

[141] Kevin Roose, "Reddit's C.E.O. on Why He Banned 'The_Donald' Subreddit," New York Times, June 30, 2020, https://www.nytimes.com/2020/06/30/us/politics/reddit-bans-steve-huffman.html.

[142] https://www.reddit.com/r/AgainstHateSubreddits/comments/gyyqem/open_letter_to_steve_huffman_and_the_board_of/

[143] https://www.reddit.com/r/announcements/comments/hi3oht/update_to_our_content_policy/

[144] https://web.archive.org/web/20200630002550/https://www.reddithelp.com/en/categories/rules-reporting/account-and-community-restrictions/promoting-hate-based-identity-or

[145] https://web.archive.org/web/20200630002550/https://www.reddithelp.com/en/categories/rules-reporting/account-and-community-restrictions/promoting-hate-based-identity-or

problem, and by July 1, 2020, the company had removed references to the "majority".[146] The revised policy explained that the rule on hate "does not protect those (groups) who promoted attacks of hate or who try to hide their hate in bad faith claims of discrimination."[147] Reddit's content policy has not changed since.

Analysis of Policy Scope

Reddit has historically been vocal in its support for uninhibited free speech and did not have an explicit hate speech prohibition for many years. In an "Ask Me Anything," CEO and co-Founder Steve Huffman famously replied to a user's question about slurs by saying:

"The way in which we think about speech is to separate behavior from beliefs… racism itself isn't against the rules . . . [but] I believe the best defense . . . instead of trying to control what people can and cannot say through rules . . . is to repudiate those views in a free conversation . . . We cannot control people's beliefs, but we can police their behaviors. And as it happens, communities dedicated to racist beliefs end up banned for violating rules we do have around harassment, bullying and violence".[148]

However, as mentioned above Reddit added a prohibition on the promotion of hatred in 2020, as also shown in Table 5. Though the scope of content covered aligns with Article 20 (2), except for the prohibition on statements of inferiority, Table 6 reveals that the spectrum of protected characteristics is much broader than the list referenced in the ICCPR, as it includes gender, color, immigration status, gender identity, disease, sexual orientation, pregnancy, and victims of a major event.

[146] Adriana Stephan, "Comparing Platform Hate Speech Policies: Reddit's Inevitable Evolution," Freeman Spogli Institute, Stanford Internet Observatory, https://fsi.stanford.edu/news/reddit-hate-speech .

[147] https://web.archive.org/web/20200702005456/https://www.reddithelp.com/en/categories/rules-reporting/account-and-community-restrictions/promoting-hate-based-identity-or

[148] "Reddit's 2017 transparency report and suspect account findings," *r/announcements*, https://www.reddit.com/r/announcements/comments/8bb85p/reddits_2017_transparency_report_and_suspect/dx5go62/.

Table 5

Content Explicitly Covered by Reddit's Hate Speech Policies		2020	2021	2022	2023
Hate(ful) speech/ content					
Promotion of Hatred		X	X	X	X
Support for Organized Hate (Including Symbols)					
On the basis of protected characteristics	Incitement to or Threats of Violence	X	X	X	X
	Attacks	X	X	X	X
	Statements of inferiority or content that demeans	X	X	X	X
	Dehumanization				
	Expressions of contempt or disgust				
	Calls for exclusion or segregation	X	X	X	X
	Discrimination				
	Denying or mocking historical atrocities, or valorizing the perpetrators				
	Slurs				
	Harmful Stereotypes				
	Conspiracy Theories				
	Cursing				

Table 6

Characteristics Protected in Reddit's Hate Speech Policies				
	2020	2021	2022	2023
Total	**12**	**12**	**12**	**12**
Race	X	X	X	X
Ethnicity	X	X	X	X
National Origin	X	X	X	X
Religion	X	X	X	X
Gender	X	X	X	X
Color	X	X	X	X
Immigration Status	X	X	X	X
Sex				
Gender Identity	X	X	X	X
Sexual Orientation	X	X	X	X
Age				
Disability	X	X	X	X
Disease/ Medical Condition				
Veteran Status				
Occupation				
Weight				
Pregnancy	X	X	X	X
Caste				
Victims of a Major Event	X	X	X	X
Socio-economic Status				
Culture				
Tribe				

Notes: An X indicates the company's hate speech policies covered that protected characteristics for at least one month during the given year.

Changes in Enforcement Volume

Reddit's 2020 content policy overhaul resulted in a huge purge of both content and groups, including its biggest Trump supporter community - "The Donald." This group was home to more

than 790,000 Reddit users. At the time, Reddit also banned approximately 2,000 other communities, including the leftist "Chapo Trap House" with 160,000 users.[149] Further, Reddit banned the "Gender Critical", prominent among radical feminists whose views on sex and gender were deemed hateful to members of the trans-community.[150] In fact, Reddit's 2020 transparency report noted the company removed 7,048 posts and 6,915 entire subreddits for hateful content.[151]

Interestingly, though the policy scope did not change again, the number of pieces of content removed for hate violations increased over the next two years while the number of subreddits banned for hate fell. According to Reddit's own transparency reports, the company removed 39,056 posts and comments involving hateful content in 2021[152] and 79,316 in 2022.[153] These increases mirrored a broader trend in increased content removals across policy categories, which Reddit attributed to evolving policies and enhanced enforcement abilities.[154] The number of entire subreddits banned for hateful content fell 93% in 2021, [155] to 467, and though it jumped back up slightly to 749 in 2022, [156] the number remained far lower than the 2020 level. It's possible that the 2020 purge got rid of most hateful subreddits and new ones did not emerge to replace them, while violations of the hateful content policy continued to appear across the platform and Reddit got better at detecting and removing them. Nevertheless, the data in Reddit's 2020, 2021, and 2022 transparency reports underscores the patterns revealed by the Facebook and Instagram enforcement data: changes in the amount of content actioned by a platform are not always closely correlated with announced changes in policy scope. Again, to truly understand the effect of scope creep in hate speech policies, better access to platform data will be necessary.

[149] Mike Isaac, "Reddit, Acting Against Hate Speech, Bans 'The_Donald' Subreddit," *New York Times*, June 29, 2020, https://www.nytimes.com/2020/06/29/technology/reddit-hate-speech.html .
[150] David Artavia, "Reddit Cracks Down on Hate Speech by Deleting TERF, Pro-Trump Forums," *Advocate*, June 29, 2020, https://www.advocate.com/news/2020/6/29/reddit-cracks-down-hate-speech-deleting-terf-pro-trump-forums .
[151] "Transparency Report 2020," *Reddit Inc.*, https://www.redditinc.com/policies/transparency-report-2020 .
[152] "Transparency Report 2021," *Reddit Inc.*, https://www.redditinc.com/policies/transparency-report-2021 .
[153] "Transparency Report 2022," *Reddit Inc.*, https://www.redditinc.com/policies/2022-transparency-report .
[154] "Transparency Report 2022," *Reddit Inc.*
[155] "Transparency Report 2021," *Reddit Inc.*
[156] "Transparency Report 2022," *Reddit Inc.*

4. SNAPCHAT

- **Release/Launch Date:** July 8, 2011
- **Number of Active Users:** 557 million [157]
- **Short Overview of Content Moderation Process:** Snapchat's content disappears within a 24-hour period. No information on whether this platform outsources content moderation is available. Content moderators review posts that have been flagged by AI, reported by users or non-users. Non-users can file a report available on Snapchat's Help Centre.
- **Signatory to the EU'S Code of Conduct on Illegal Hate Speech Online:** Yes

[157] "Most popular social networks worldwide as of January 2023, ranked by number of monthly active users," *Statista*, https://www.statista.com/statistics/272014/global-social-networks-ranked-by-number-of-users/ (accessed on April 30, 2023).

Key Developments

Terms of Use

Snapchat's first Terms of Use date to 2011, and they included a provision that addressed racist content. Under the heading "what you cannot do with the application," Snapchat prohibited "transmitting any material that could be considered racist, threatening or unlawful in any way."[158] According to the coding rules for this report, this provision addresses hate speech, because it prohibits material that discriminates on the basis of race – i.e. "racist" material. This provision was relatively narrow, since race was the only identity protected from discriminatory content. In 2013, Snapchat added another relevant provision to the Terms of Use, which directed users not to "send snaps that (their friends) don't want to receive (threats, harassment, racism etc.)."[159]

In 2014, Snapchat eliminated these provisions from its Terms of Use. The company removed the suggestion to avoid sending friends "threats, harassment, and racism" and deleted the prohibition on transmitting any material that could be considered racist. Instead, the company added a section titled "Prohibited Activities," which prohibited harassment and intimidation but made no reference to harassment or intimidation of individuals based on identity characteristics. The next year, however, Snapchat added a provision to the Terms of Use that explicitly prohibited content containing "hate speech," though it offered no definition of the concept.[160] In September 2021, Snapchat removed the hate speech provision from the Terms. Since then, Snapchat's Terms of Use have not addressed hate speech.

Community Guidelines

Snapchat also has Community Guidelines, which date back to 2014. They provide a general overview of prohibited content on the platform. The initial versions of these Guidelines prohibited harassment, bullying, and threats, but there was no mention of prohibiting such content if it targeted individuals based on their identity. Thus, these provisions did not address hate speech specifically. By 2017, however, the company had an explicit prohibition on hate speech in the Guidelines, which read: "Hate Speech: Don't post content that demeans, defames, or promotes discrimination on the basis of race, ethnicity, national origin, religion, sexual orientation, gender, disability, or veteran status."[161]

Snapchat has expanded the scope of its hate speech provisions in recent years, however. In 2021, Snapchat began addressing hate speech under a Guideline titled "Terrorism, Hate Groups, and Hate Speech." The rule stated: "Hate speech or content that demeans, defames, or promotes discrimination or violence on the basis of race, color, caste, ethnicity, national origin, religion,

[158] https://web.archive.org/web/20120711233922/http://www.snapchat.com/terms
[159] https://web.archive.org/web/20130417050031/http://www.snapchat.com/terms
[160] https://web.archive.org/web/20151122121643/https://www.snapchat.com/terms
[161] https://web.archive.org/web/20170127184718/https://support.snapchat.com/en-US/a/guidelines

sexual orientation, gender identity, disability, or veteran status, immigration status, socio-economic status, age, weight or pregnancy status is prohibited." While this revision did not update the 2017 definition of hate speech, it did cover seven additional protected characteristics. The rule also prohibited "hate groups," though the company did not provide a definition of such organizations.

In early 2023, Snapchat published an in-depth explanation of its hate speech, terrorism, and violent extremism policy, as part of a series of Community Guidelines explainers.[162] In this brief, the company expanded upon its definition of hate speech. Snapchat explained that, in addition to content that demeans or promotes discrimination against individuals on the basis of protected characteristics, "hate speech also extends to the valorization of perpetrators—or the denigration of victims—of human atrocities (such as genocide, apartheid, or slavery)" and "the use of hate symbols, which means any imagery that is intended to represent hatred or discrimination toward others (including those featured in the hate symbols database maintained by the Anti-Defamation League)."

Analysis of Policy Scope

Table 7 reveals that the scope of Snapchat's hate speech policy has increased substantially over the past ten years. Until 2023, Snapchat appeared to define hate speech as content that discriminates, defames, or demeans based on protected characteristics, but the recent explainer suggests the company also considers content that praises the perpetrators of genocide, apartheid, or slavery to be hate speech, which is broader than the previous conceptualization. This definition is much broader than the definition implied by Article 20.

Table 8 illustrates that the scope of characteristics protected under Snapchat's hate speech provisions has also expanded. In 2011, the platform simply banned racist content, but by 2021, the platform banned content that discriminated against or demeaned someone based on any one of fifteen different characteristics. Compared to the ICCPR, Snapchat's hate speech prohibition covers a broad range of characteristics. Specifically, Snapchat's protected characteristics include color, immigration status, gender identity, sexual orientation, age, disability, veteran status, socio-economic status, weight, and pregnancy, which are not covered in Article 20 (2). Arguably characteristics such as "socio-economic status", "veteran status", "pregnancy", and "weight" would not satisfy the necessity test required to pass muster under Article 19 (3), just as "demeaning" speech and the "denying" of historical crimes and atrocities constitute protected speech under ICCPR.

[162] "Hateful Content, Terrorism, and Violent Extremism: Community Guidelines Explainer Series," *Snap Privacy and Safety Hub, Transparency*, January 2023, https://values.snap.com/privacy/transparency/community-guidelines/hateful-content .

Table 7

Content Explicitly Covered by Snapchat's Hate Speech Policies		2011	2012	2013	2014	2015	2016	2017	2018	2019	2020	2021	2022	2023
Hate(ful) speech/ content						X	X	X	X	X	X	X	X	X
Promotion of Hatred														
Support for Organized Hate (Including Symbols)														X
On the basis of protected characteristics	Incitement to or Threats of Violence								X	X	X	X	X	X
	Attacks													
	Statements of inferiority or content that demeans							X	X	X	X	X	X	X
	Dehumanization													
	Expressions of contempt or disgust													
	Calls for exclusion or segregation													
	Discrimination	X	X	X	X	X		X	X	X	X	X	X	X
	Denying or mocking historical atrocities, or valorizing the perpetrators													X
	Slurs													
	Harmful Stereotypes													
	Conspiracy Theories													
	Cursing													

Table 8

Characteristics Protected in Snapchat's Hate Speech Policies													
	2011	2012	2013	2014	2015	2016	2017	2018	2019	2020	2021	2022	2023
Total	**1**	**1**	**1**	**1**	**0**	**0**	**8**	**9**	**8**	**8**	**15**	**15**	**15**
Race	X	X	X	X			X	X	X	X	X	X	X
Ethnicity							X	X	X	X	X	X	X
National Origin							X	X	X	X	X	X	X
Religion							X	X	X	X	X	X	X
Gender							X	X					
Color											X	X	X
Immigration Status											X	X	X
Sex													
Gender Identity								X	X	X	X	X	X
Sexual Orientation							X	X	X	X	X	X	X
Age											X	X	X
Disability							X	X	X	X	X	X	X
Disease/ Medical Condition													
Veteran Status							X	X	X	X	X	X	X
Occupation													
Weight											X	X	X
Pregnancy											X	X	X
Caste											X	X	X
Victims of a Major Event													
Socio-Economic Status											X	X	X
Culture													
Tribe													

Notes: An X indicates the company's hate speech policies covered that protected characteristic for at least one month during the given year.

Changes in Enforcement Volume

In Q2 2019, Snapchat began reporting information on the volume of content actioned for violations of the Community Guidelines. While the scope of the content covered by Snapchat's policy did not change dramatically after this point, the number of protected characteristics rose from 8 to 15 in January 2021. According to Snapchat's Transparency Report, the company took enforcement action on 77,587 Snaps, or 1.4% of total content actioned, between July and December 2020.[163] Between January and July 2021, Snapchat took enforcement action on 121,639 Snaps, or 1.9% of total content actioned during the period.[164] It is possible that the expansion in the scope of Snapchat's policies is related to the increase in content actioned in Q1 2021. However, the content actioned number fell to 93,341, or 1.5% of all content actioned, in Q2 2021,[165] suggesting the bump in Q1 did not persist. This observation suggests changes in the policy scope alone cannot account for the increase in enforcement action between Q2 2020 and Q1 2021. As with the data from the other platforms, this data underscores the need to give external researchers access to platform data, so they can rigorously analyze the impact of changes in policy scope on content removals and other enforcement actions.

[163] "Transparency Report: July 1, 2020 - December 31, 2020," *Snap Inc.*, July 1, 2021, https://www.snap.com/en-US/privacy/transparency/2020-12-31 .
[164] "Transparency Report: January 1, 2021 - June 30, 2021," Snap Inc., November 22, 2021, https://www.snap.com/en-US/privacy/transparency/2021-06-30 .
[165] "Transparency Report: July 1, 2021 - December 31, 2021," Snap Inc., April 1, 2022, https://www.snap.com/en-US/privacy/transparency/2021-12-31 .

5. TIKTOK

- **Launch date:** September 2016 (previously Musical.ly – April 2014)
- **Number of Active Users:** 1.051 billion[166]
- **Short Overview of Content Moderation Process:** TikTok redirects users who search for offensive content to Community Guidelines. It also refrains from showing results and removes related content. Content moderators review posts that have been flagged by AI, reported by users.
- **Signatory to the EU'S Code of Conduct on Illegal Hate Speech Online?** Yes

[166] "Most popular social networks worldwide as of January 2023, ranked by number of monthly active users," *Statista*, https://www.statista.com/statistics/272014/global-social-networks-ranked-by-number-of-users/ (accessed on April 30, 2023).

Key Developments

Terms of Use

The first Terms of Use for Musical.ly (the predecessor of TikTok) date back to 2014. The Terms warned users that they might encounter harmful or inaccurate content and limited Musical.ly's liability in such cases. They also prohibited a direct and specific threat of violence to others, as well as harassment and abuse, but they made no reference to prohibiting this objectionable content if it targeted specific identities. Thus, the initial Terms did not include an explicit hate speech provision. In mid-2015, however, the company added such a prohibition to the Terms. Under the heading "Restrictions on Content," Musical.ly informed users that they must "agree not to post any Content to the Platform that… is racially, ethnically or sexually discriminatory in any way, or that otherwise violates any right of others."[167] In December 2015, the company changed the Terms again, adding a section on "Objectionable Content." This part of the Terms prohibited content "that is or could be interpreted to be (i) abusive, bullying, defamatory, harassing, harmful, hateful, inaccurate, infringing, libelous, objectionable, obscene, offensive, pornographic, shocking, threatening, unlawful, violent, or vulgar" or "(ii) promoting bigotry, discrimination, hatred, racism, or inciting violence."[168]

This development is somewhat confusing. The first phrase prohibits hateful content, alongside several other types of objectionable content, but the second one prohibits content that promotes bigotry, discrimination, hatred, and racism – which might be forms of hateful content. This update also reduced the number of protected characteristics from three (race, ethnicity, and sex) to one (race). However, in early 2016, the company also added a somewhat vague reference to religion to this policy. The "Objectionable Content" section now included a prohibition on "SR Samples (… and the musical works therein), making a political message for or against any person, party, political belief or issue, of a religious nature."[169] It is not clear whether Musical.ly intended this phrase to prohibit religious messages entirely or to prohibit objectionable content of a religious nature.

In August 2018, ByteDance, a Chinese company that had purchased Musical.ly, merged the app with another product – TikTok. The combined app took the latter title, and it quickly gained popularity. TikTok's 2018 Terms of Service included a few different provisions relevant to hate speech.[170] These provisions remain in place today, and they instruct users not to "promote discrimination based on race, sex, religion, nationality, disability, sexual orientation or age" nor "use the Services to upload, transmit, distribute, store or otherwise make available in any way… material which is defamatory of any person, obscene, offensive, pornographic, hateful or

[167] https://web.archive.org/web/20150705030445/http://www.musical.ly/term.html#
[168] https://web.archive.org/web/20160114181828/http://musical.ly/term.html
[169] https://web.archive.org/web/20160402170821/http://musical.ly:80/term.html
[170] We did not find any Terms of Service for TikTok prior to the 2018 merger with Musical.ly on Wayback Machine.

inflammatory; [or] racist or discriminatory, including discrimination on the basis of someone's race, religion, age, gender, disability or sexuality."[171] The difference between the protected characteristics mentioned in the two provisions is puzzling.

It's also worth noting that the Terms of Service reserved TikTok the right to "remove or disable access to content at our discretion for any reason or no reason. Some of the reasons we may remove or disable access to content may include finding the content objectionable, in violation of these Terms or our Community Policy, or otherwise harmful to the Services or our users."[172] This provision suggests TikTok can remove content arbitrarily if they deem it necessary.

Community Guidelines

In 2016, Musical.ly created Community Guidelines, but they did not include any hate speech provisions. The first traceable TikTok Community Guidelines, which date to January 2020, prohibited content that "incites hatred against a group of people based on their race, ethnicity, religion, nationality, culture, disability, sexual orientation, gender, gender identity, age, or any other discrimination."[173] This policy added four protected characteristics to the list mentioned in the Terms of Service.

Later the same month, however, TikTok updated the Guidelines with a much more in-depth policy. Under the heading "Hate Speech," TikTok listed three categories of prohibited content: attacks on protected groups, slurs, and hateful ideologies. In the first section, the company defined hate speech as "content that does or intends to attack, threaten, incite violence against, or dehumanize an individual or group of people on the basis of protected attributes" (see Figure 9). The company also offered examples of the content covered by the policy, such as claims that persons with protected attributes are physically or morally inferior, criminals, or non-human entities (like animals).[174] In the second section, TikTok explained its prohibition on slurs, or "derogatory terms that are intended to disparage" people according to protected attributes, though the company noted that exceptions might be made for slurs used in a self-referential manner.[175] As the company later explained: "If a member of a disenfranchised group, such as the LGBTQ+, Latinx, Asian American and Pacific Islander, Black, and Indigenous communities, uses a slur as a term of empowerment, we want our moderators to understand the context behind it and not mistakenly take the content down. On the other hand, if a slur is being used hatefully, it doesn't belong on TikTok. Educating our content moderation teams on these important distinctions is ongoing work,

[171] https://web.archive.org/web/20180831013042/http://www.tiktok.com/i18n/terms/
[172] https://web.archive.org/web/20180831013042/http://www.tiktok.com/i18n/terms/
[173] https://web.archive.org/web/20200116003342/https://www.tiktok.com/community-guidelines?lang=en
[174] https://web.archive.org/web/20200122164447/https://www.tiktok.com/community-guidelines?lang=en
[175] https://web.archive.org/web/20200122164447/https://www.tiktok.com/community-guidelines?lang=en

and we strive to get this right for our users."[176] In the final section, TikTok outlined its prohibition on "content that promotes hateful ideologies," including content that "denies well-documented and violent events have taken place." [177] At this time, TikTok also removed culture from the list of protected characteristics and added disease, caste, and immigration status.

Figure 9[178]

Hate speech

We do not tolerate content that attacks or incites violence against an individual or a group of individuals on the basis of protected attributes. We do not allow content that includes hate speech, and we remove it from our platform. We also suspend or ban accounts that have multiple hate speech violations.

Attacks on protected groups

We define hate speech as content that does or intends to attack, threaten, incite violence against, or dehumanize an individual or a group of individuals on the basis of protected attributes. We also do not allow content that verbally or physically threatens violence or depicts harm to an individual or a group based on any of the following protected attributes:

- Race
- Ethnicity
- National origin
- Religion
- Caste
- Sexual orientation
- Sex
- Gender
- Gender identity
- Serious disease or disability

In December 2020, TikTok renamed the policy "hateful behavior" and changed the title of the first category to "attacks on the basis of protected attributes."[179] The company also added a sentence at the beginning of the policy guidance indicating they would even ban accounts engaged in or associated with hate speech off the platform. TikTok also offered the following definition of hateful

[176] Andrew Hutchinson, "TikTok Provides An Update on its Approach to Hate Speech and Offensive Content," *Social Media Today*, August 20, 2020, https://www.socialmediatoday.com/news/tiktok-provides-an-update-on-its-approach-to-hate-speech-and-offensive-cont/583905/
[177] https://web.archive.org/web/20200122164447/https://www.tiktok.com/community-guidelines?lang=en
[178] https://web.archive.org/web/20200122164447/https://www.tiktok.com/community-guidelines?lang=en
[179] https://web.archive.org/web/20201231234747/https://www.tiktok.com/community-guidelines?lang=en

ideologies: "those that demonstrate clear hostility toward people because of their protected attributes."[180] In February 2022, TikTok made further revisions to the policy, combining the "attacks on the basis of protected attributes" and "slurs" categories into one category titled "attacks and slurs on the basis of protected attributes."[181] The company also added references to specific prohibited hateful ideologies, such as white supremacist, misogynistic, anti-LGBTQ, and antisemitic beliefs. The blog post that accompanied this 2022 update suggested the hateful ideology category covers content like deadnaming, misgendering, and the promotion of conversion therapy programs.[182] In March 2023, TikTok announced a variety of changes to their Community Guidelines, including adding tribe as a protected characteristic under the hate speech and hateful behavior policy.[183] The overhaul also involved reorganizing the hate speech policy, though this reorganization resulted in no substantial changes to the types of content prohibited by the policy.

Analysis of Policy Scope

Table 9 and Table 10 demonstrate that TikTok's approach to hate speech has evolved considerably since Musical.ly's first Terms of Use, which prohibited discrimination based on race, ethnicity, and sex but did not go any further. Today, TikTok's definition of hate speech goes beyond discriminatory language and includes attacks, threats, dehumanization, and incitement against an individual or group based on any one of 12 different characteristics. In addition to being broader than Musical.ly's initial provision, TikTok's current hate speech policy explicitly covers several types of content, including conspiracy theories and the denial of violent events, and several more protected attributes, including gender, immigration status, gender identity, age, caste, sexual orientation, disease, and disability. The current wording includes significantly more protected categories than the mandatory prohibited categories in Article 20(2) of the ICCPR, and – taken as a whole - seems difficult to reconcile with ICCPR Article 19´s ban against overly vague and broad restrictions on free expression, as well as with the requirements of necessity and legitimacy.

[180] https://web.archive.org/web/20201231234747/https://www.tiktok.com/community-guidelines?lang=en
[181] https://web.archive.org/web/20220307104054/https://www.tiktok.com/community-guidelines-2022?lang=en#38
[182] Carmac Keenan, "Strengthening our policies to promote safety, security, and well-being on TikTok," *TikTok*, February 8, 2022, https://newsroom.tiktok.com/en-us/strengthening-our-policies-to-promote-safety-security-and-wellbeing-on-tiktok.
183 Julie de Bailliencourt, "Helping creators understand our rules with refreshed Community Guidelines," *TikTok*, March 21, 2023, https://newsroom.tiktok.com/en-us/community-guidelines-update .

Table 9

Content Explicitly Covered by Musical.ly's & TikTok's Hate Speech Policies		2015	2016	2017	2018	2019	2020	2021	2022	2023
Hate(ful) speech/ content		X	X	X	X	X	X	X	X	X
Promotion of Hatred		X	X	X	X	X	X	X	X	X
Support for Organized Hate (Including Symbols)							X	X	X	X
On the basis of protected characteristics	Incitement to or Threats of Violence				X	X	X	X	X	X
	Attacks						X	X	X	X
	Statements of inferiority or content that demeans						X	X	X	X
	Dehumanization						X	X	X	X
	Expressions of contempt or disgust									
	Calls for exclusion or segregation						X	X	X	X
	Discrimination	X	X	X	X	X	X	X	X	X
	Denying or mocking historical atrocities, or valorizing the perpetrators						X	X	X	X
	Slurs						X	X	X	X
	Harmful Stereotypes									
	Conspiracy Theories						X	X	X	X
	Cursing									

Table 10

Characteristics Protected by Musical.ly's and TikTok's Hate Speech Policies									
	2015	2016	2017	2018	2019	2020	2021	2022	2023
Total	**3**	**2**	**2**	**6**	**6**	**13**	**12**	**12**	**14**
Race	X	X	X	X	X	X	X	X	X
Ethnicity	X					X	X	X	X
National Origin				X	X	X	X	X	X
Religion		X	X	X	X	X	X	X	X
Gender						X	X	X	X
Color									
Immigration Status						X	X	X	X
Sex	X			X	X				X
Gender Identity						X	X	X	X
Sexual Orientation				X	X	X	X	X	X
Age				X	X	X	X	X	X
Disability						X	X	X	X
Disease/ Medical Condition						X	X	X	X
Veteran Status									
Occupation									
Weight									
Pregnancy									
Caste						X	X	X	X
Victims of a Major Event									
Socio-Economic Status									
Culture						X			
Tribe									X

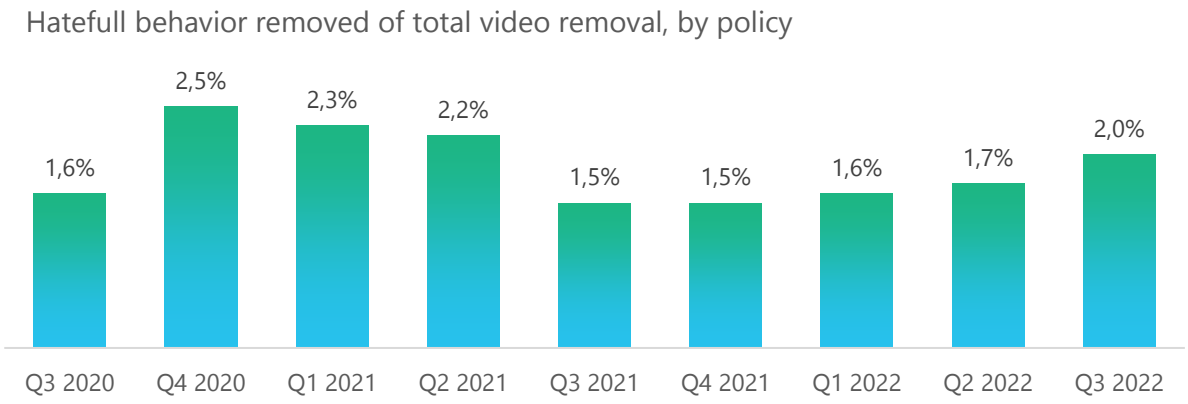
Notes: An X indicates the company's hate speech policies covered that protected characteristic for at least one month during the given year.

Changes in Enforcement Volume

As the previous section illustrates, TikTok expanded the scope of its hate speech prohibitions in January 2020, by adding both new categories of covered content and new protected

characteristics. In a blog post written in August 2020, TikTok reported that they had removed more than 380,000 videos for violating the hate speech policy, banned more than 1,300 accounts, and removed over 64,000 hate comments since the beginning of the year.[184] Unfortunately, we have no way of knowing whether this represented a large increase in enforcement, compared to before the January 2020 policy change. TikTok claims that their content moderation infrastructure did not enable them to provide information about video removal by policy type prior to December 2019,[185] so information on content enforcement by policy category is not available for periods prior to 2020. Thus, we cannot use TikTok's transparency reports to identify any potential correlations between the January 2020 change in policy scope and changes in enforcement volume. Though TikTok has edited the policy language since then, the scope of the provision has not changed substantially. That being said, Figure 10 does suggest the percentage of video removals due to hate speech has remained relatively steady since 2020.

Figure 10[186]



[184] Erik Han, "Countering hate on TikTok," *TikTok*, August 20, 2020, https://newsroom.tiktok.com/en-us/countering-hate-on-tiktok
[185] "Community Guidelines Enforcement Report: July 1, 2019 - December 31, 2019," *TikTok*, July 9, 2020, https://www.tiktok.com/transparency/en/community-guidelines-enforcement-2019-2/ .
[186] "Community Guidelines Enforcement Report: July 1, 2022 - September 30, 2022," *TikTok*, December 19, 2022, https://www.tiktok.com/transparency/en/community-guidelines-enforcement-2022-3/ .

6. TUMBLR

- **Release/Launch date:** 2007
- **Number of Active Visitors:** 292 million[187]
- **Short Overview of Moderation Process:** Content moderators remove blogs and re-blogs that have been reported by users/visitors or flagged by Artificial Intelligence
- **Signatory to the EU'S Code of Conduct on Illegal Hate Speech Online:** No

[187] "Worldwide visits to Tumblr.com from December 2021 to May 2022," *Statista*, https://www.statista.com/statistics/261925/unique-visitors-to-tumblrcom/ (Accessed on 21 December 2022).

Key Developments

Terms of Service

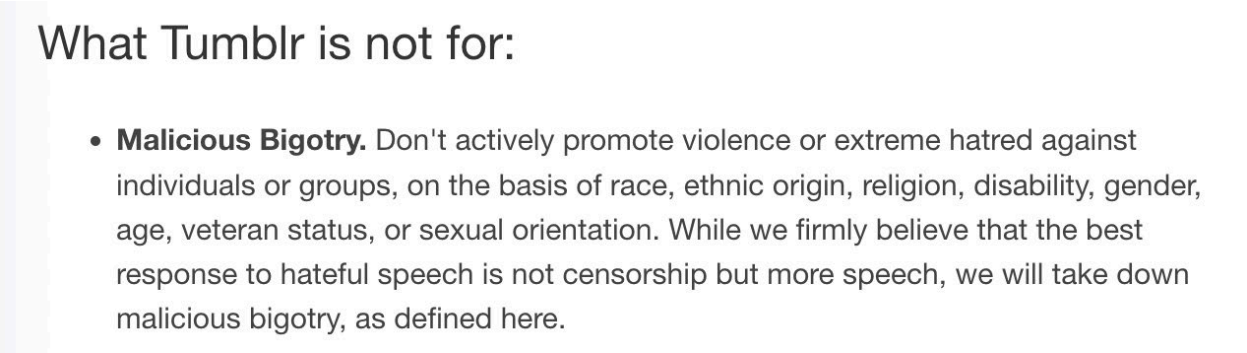
The first traceable Tumblr Terms of Service are from March 2007, but they did not include a provision on hate speech until October 2007. At that point, the Terms required subscribers to agree not to post content that is "hateful."[188] In 2010, Tumblr added an additional note on hate speech: "hate content don't belong on the web, and certainly don't belong on Tumblr."[189] By March 2012, Tumblr had removed both references to hateful content from the Terms and added policies that fall within the scope of this report the company's new Community Guidelines. In 2022, however, Tumblr added a reference to hate speech back into the Terms of Service, instructing users to "not give tips in exchange for – or to promote or encourage – content that is against our Community Guidelines or that is otherwise illegal, abusive towards others, hateful, or that could result in self-harm."[190] The hate speech provisions in Tumblr's Terms of Service have never referenced any specific protected characteristics, however.

Community Guidelines

In March 2012, Tumblr created a Content Policy that prohibited "hate speech and other objectionable content that is unlawful, defamatory, and fraudulent."[191] Within the same month, however, Tumblr launched the Community Guidelines, which began by stating that Tumblr as a "global platform for creativity and self-expression" that was "deeply committed to supporting and protecting freedom of speech."[192] However, in the same paragraph, the company noted that it drew "lines around a few narrowly-defined but deeply important categories of content and behavior that jeopardize our users, threaten our infrastructure, or damage our community." One such category of prohibited content was "malicious bigotry," or actively promoting "violence or extreme hatred against individuals or groups" based on eight different protected characteristics (see Figure 11). This provision implicitly equated hate speech and "malicious bigotry," defining them as the promotion of violence or hatred against specific identity groups. However, Tumblr did not provide any specifics about what actively promoting hatred would look like.

[188] https://web.archive.org/web/20071229031909/http://www.tumblr.com:80/terms_of_service
[189] https://web.archive.org/web/20100522002538/http://www.tumblr.com/terms_of_service
[190] https://web.archive.org/web/20220301053331/https://www.tumblr.com/policy/en/terms-of-service
[191] https://web.archive.org/web/20120319173638/http://www.tumblr.com/policy/en/community
[192] https://web.archive.org/web/20120731142112/http://www.tumblr.com/policy/en/community

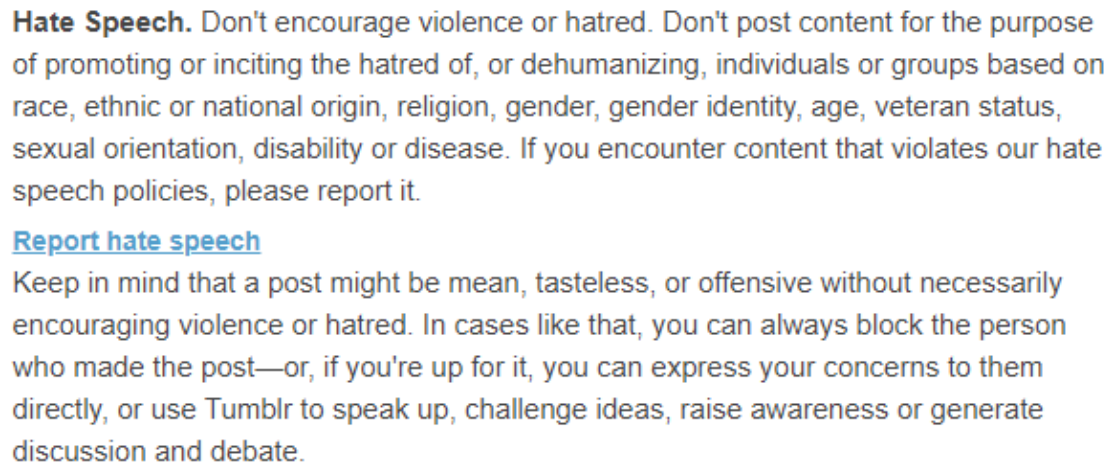
Figure 11[193]



By January 2014, Tumblr had changed the title of this provision from "Malicious Bigotry" to "Malicious Speech," and lowered the threshold from the "active promot[ion] of violence or extreme hatred" to the "encourage[ment] of violence or hatred."[194] Tumblr still provided no additional details about what encouraging hatred might look like. Interestingly, Tumblr also removed the reference to the company's firm belief that the best response to hateful speech is more speech and, instead encouraged users to "dismantle negative speech through argument rather than censorship." In 2016, Tumblr added more details about what it meant to encourage violence or hatred, telling users not to "make violent threats or statements that incite violence, including threatening or promoting terrorism," especially if such content threatened people based on nine protected characteristics.[195] Tumblr revised its policy in 2018 (see Figure 12), however, when the company retitled the provision "hate speech" and instructed users not to "encourage violence or hatred," nor "post content for the purpose of promoting or inciting the hatred of, or dehumanizing, individuals or groups based on race, ethnic or national origin, religion, gender, gender identity, age, veteran status, sexual orientation, disability, or disease." The 2018 policy also explicitly noted that content "might be offensive without necessarily encouraging violence or hatred." The policy has not changed since.

[193] https://web.archive.org/web/20120731142112/http://www.tumblr.com/policy/en/community
[194] https://web.archive.org/web/20140702023313/http://www.tumblr.com/policy/en/community
[195] https://web.archive.org/web/20160702110539/https://www.tumblr.com/policy/en/community

Figure 12[196]



- **Hate Speech.** Don't encourage violence or hatred. Don't post content for the purpose of promoting or inciting the hatred of, or dehumanizing, individuals or groups based on race, ethnic or national origin, religion, gender, gender identity, age, veteran status, sexual orientation, disability or disease. If you encounter content that violates our hate speech policies, please report it.

  Report hate speech

  Keep in mind that a post might be mean, tasteless, or offensive without necessarily encouraging violence or hatred. In cases like that, you can always block the person who made the post—or, if you're up for it, you can express your concerns to them directly, or use Tumblr to speak up, challenge ideas, raise awareness or generate discussion and debate.

Analysis of Policy Scope

Over time, Tumblr has added more details to its hate speech policies, but they have also increased in scope. The relevant provisions in the Terms of Service prohibited hateful content but never defined it, suggesting the policy could cover a wide range of content deemed to involve hatred. This lack of clarity is problematic in light of the legality requirement in Article 19 of the ICCPR. While the first version of the Community Guidelines prohibited the active promotion of hatred, by 2014, Tumblr had shifted to prohibiting the encouragement of hate, possibly representing a broader range of covered content. In 2018, Tumblr added a prohibition on content that dehumanizes individuals on the basis of protected characteristics, representing an additional expansion in the covered content. This prohibition is much broader than the mandatory prohibition set out Article 20 (2), as well as the permitted restrictions under Article 19, under the three-part test.

Additions to the list of protected characteristics have also represented expansions to the scope of Tumblr's hate speech policy. The hate speech provisions in the Terms of Service did not mention any protected characteristics, but the initial relevant policy in the Community Guidelines listed eight. Tumblr added "gender identity" in 2016 and "national origin" and "disease" in 2018. Tumblr's list is broader than the characteristics covered by Article 20(2), due to the inclusion of gender, gender identity, age, veteran status, sexual orientation, disability, and disease. Arguably specific restrictions of speech based on categories such as veteran status and age are not in conformity with Article 19.

[196] https://web.archive.org/web/20181001144639/https://www.tumblr.com/policy/en/community

Table 11

	Content Explicitly Covered by Tumblr 's Hate Speech Policies	2007	2008	2009	2010	2011	2012	2013	2014	2015	2016	2017	2018	2019	2020	2021	2022	2023
	Hate(ful) speech/ content	X	X	X	X	X	X	X					X	X	X	X	X	X
	Promotion of Hatred						X	X	X	X	X	X	X	X	X	X	X	X
	Support for Organized Hate (Including Symbols)																	
On the basis of protected characteristics	Incitement to or Threats of Violence						X	X	X	X	X	X	X					
	Attacks																	
	Statements of inferiority or content that demeans																	
	Dehumanization												X	X	X	X	X	X
	Expressions of contempt or disgust																	
	Calls for exclusion or segregation																	
	Discrimination																	
	Denying or mocking historical atrocities, or valorizing the perpetrators																	
	Slurs																	
	Harmful Stereotypes																	
	Conspiracy Theories																	
	Cursing																	

Table 12

Characteristics Protected in Tumblr's Hate Speech Policies																	
	2007	2008	2009	2010	2011	2012	2013	2014	2015	2016	2017	2018	2019	2020	2021	2022	2023
Total	**0**	**0**	**0**	**0**	**0**	**8**	**8**	**8**	**8**	**9**	**9**	**11**	**11**	**11**	**11**	**11**	**11**
Race						X	X	X	X	X	X	X	X	X	X	X	X
Ethnicity						X	X	X	X	X	X	X	X	X	X	X	X
National Origin												X	X	X	X	X	X
Religion						X	X	X	X	X	X	X	X	X	X	X	X
Gender						X	X	X	X	X	X	X	X	X	X	X	X
Color																	
Immigration Status																	
Sex																	
Gender Identity										X	X	X	X	X	X	X	X
Sexual Orientation						X	X	X	X	X	X	X	X	X	X	X	X
Age						X	X	X	X	X	X	X	X	X	X	X	X
Disability						X	X	X	X	X	X	X	X	X	X	X	X
Disease/ Medical Condition												X	X	X	X	X	X
Veteran Status						X	X	X	X	X	X	X	X	X	X	X	X
Occupation																	
Weight																	
Pregnancy																	
Caste																	
Victims of a Major Event																	
Socio-economic Status																	
Culture																	
Tribe																	

Notes: An X indicates the company's hate speech policies covered that protected characteristic for at least one month during the given year.

Changes in Enforcement Volume

Tumblr does not provide public information about content removals due to its Terms of Service or Community Guidelines.

7. TWITTER

- **Release/Launch Date:** 21 March 2006
- **Number of Users/Visitors:** 436 million[197]
- **Short Overview of Moderation Process:** Content moderators review posts that have been flagged by AI and reported by users. The majority of this work is outsourced to third-party vendors.
- **Signatory to the EU'S Code of Conduct on Illegal Hate Speech Online:** Yes

[197] "Most popular social networks worldwide as of January 2023, ranked by number of monthly active users," *Statista*, https://www.statista.com/statistics/272014/global-social-networks-ranked-by-number-of-users/ (accessed on April 30, 2023).

Key Developments

Terms of Service

Twitter's Terms of Service have never included a provision on hate speech. While they address harmful content, they have never referenced hate speech – nor referenced harmful content targeted at specific identity-based characteristics. Thus, the Terms of Service do not include provisions relevant to the scope of this report. However, as well as Terms of Service, Twitter also has "Twitter Rules."

Rules

Twitter first published "Rules" in 2009. While the company prohibited users from publishing or posting "direct, specific threats of violence against others," under the heading of "Content Boundaries and Use of Twitter," the Rules did not include a hate speech provision at this time. In fact, Twitter did not have an explicit prohibition on hate speech until 2017, when the company added a prohibition on hateful conduct and hateful imagery/display names to the Rules. The two relevant provisions in the 2017 version of the rules were as follows:

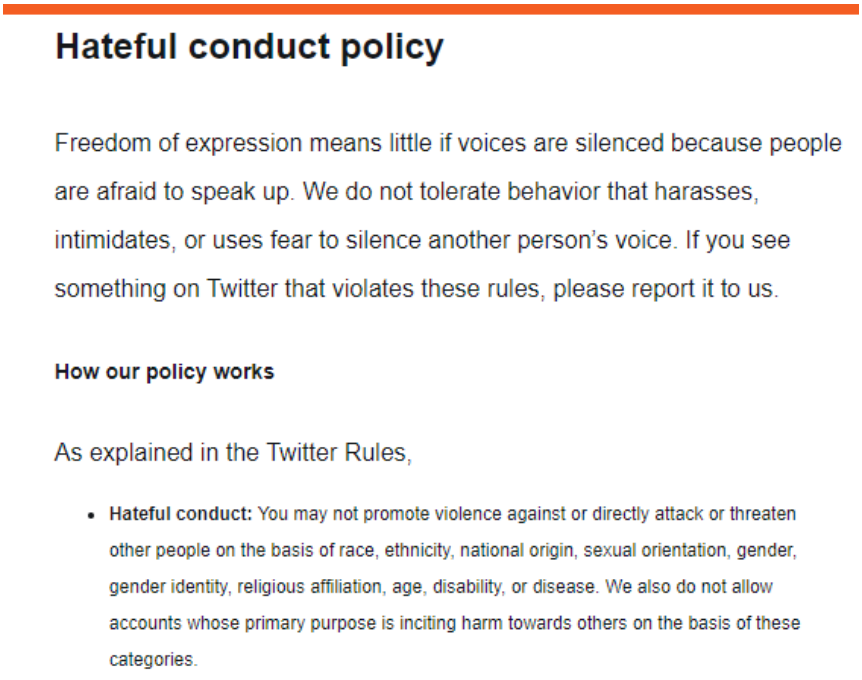
- "Hateful conduct: You may not promote violence against, threaten, or harass other people on the basis of race, ethnicity, national origin, sexual orientation, gender, gender identity, religious affiliation, age, disability, or serious disease.
- Hateful imagery and display names: You may not use hateful images or symbols in your profile image or profile header. You also may not use your username, display name, or profile bio to engage in abusive behavior, such as targeted harassment or expressing hate towards a person, group, or protected category."[198]

Since 2017, Twitter has also had an accompanying, in-depth explanation of the hateful conduct provision, which the company refers to as its "Hateful Conduct Policy." This accompanying document is a description of the policy's scope and application. In 2017, this document began by noting that "freedom of expression means little if voices are silenced because people are afraid to speak up" (see Figure 13). It also provided examples of content that the policy covered, including violent threats; wishes for the physical harm, death, or disease of individuals or groups; references to mass murder, violent events, or specific means of violence in which/with which such groups have been the primary targets or victims; behavior that incites fear about a protected group; and repeated and/or or non-consensual slurs, epithets, racist and sexist tropes, or other content that degrades someone. [199] The end of the hateful conduct policy included a section on enforcement,

[198] https://web.archive.org/web/20171218210508/https://help.twitter.com/en/rules-and-policies/twitter-rules
[199] https://web.archive.org/web/20171218171753/https://help.twitter.com/en/rules-and-policies/hateful-conduct-policy

which stated that Twitter would enforce the policy "when someone reports behavior that is abusive and targets an entire protected group and/or individuals who may be members."[200]

Figure 13[201]



Hateful conduct policy

Freedom of expression means little if voices are silenced because people are afraid to speak up. We do not tolerate behavior that harasses, intimidates, or uses fear to silence another person's voice. If you see something on Twitter that violates these rules, please report it to us.

How our policy works

As explained in the Twitter Rules,

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

In late 2018, Twitter added a specific section on hateful imagery to the list of covered content in the Hateful Conduct policy.[202] In addition to the hateful conduct policy, Twitter has an abusive profile information policy, which explains why, when, and how Twitter prohibits people from using hateful imagery or speech in their profile picture or display name. Interestingly, by 2019, the "Twitter Rules" no longer included a separate mention of hateful imagery/display names – suggesting the company felt the hateful conduct provision was sufficient to address hateful imagery in the "Twitter Rules."

Twitter also made other additions to the hateful conduct policy in late 2018, including adding a rationale section and in-depth explanations for each type of covered content. The rationale section expanded upon the brief paragraph included in the 2017 hateful conduct policy, which mentioned the meaningless nature of free expression if certain communities are silenced. In the late 2018 version, Twitter noted that "research has shown that some groups of people are disproportionately targeted with abuse online," including "women, people of color, lesbian, gay,

[200] https://web.archive.org/web/20171218171753/https://help.twitter.com/en/rules-and-policies/hateful-conduct-policy
[201] https://web.archive.org/web/20171218171753/https://help.twitter.com/en/rules-and-policies/hateful-conduct-policy
[202] https://web.archive.org/web/20181028023901/https://help.twitter.com/en/rules-and-policies/hateful-conduct-policy

bisexual, transgender, queer, intersex, asexual individuals, marginalized and historically underrepresented communities."[203] It went on to explain that abuse may be more common, more severe, and more impactful for individuals who identify with these underrepresented groups. Lastly, Twitter stated that it prohibited the abuse of individuals based on protected category because it was "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." [204] This section therefore implies that Twitter prohibits hate speech to protect the free speech of marginalized groups, though with specific assumptions about which groups are marginalized that reflect a US/Western centric lens.

The in-depth explanations of each type of covered content shed light on the scope of content that the policy prohibits. For example, under the heading "violent threats," Twitter stated: "we prohibit content that makes violent threats against an identifiable target."[205] Under the heading "wishing, hoping, or calling for serious harm on a person or group of people," Twitter explained that it prohibited content such as "hoping that someone dies as a result of a serious disease" or "saying that a group of individuals deserve serious physical injury."[206] Under the heading "Inciting fear about a protected category," Twitter noted: "we prohibit targeting individuals with content intended to incite fear or spread fearful stereotypes about a protected category, including asserting that members of a protected category are more likely to take part in dangerous or illegal activities."[207] In the same section, Twitter stated: "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" and "targeted misgendering or deadnaming of transgender individuals."[208] The policy is lengthy, so we do not include all the details here, but the excerpts demonstrate the breadth of speech covered by the policy.

The prohibition on "content that intends to dehumanize... a protected category" is worth briefly discussing in more depth. While our research suggests that Twitter's hateful conduct policy prohibited this type of speech as early as October 2018, Twitter published a blog post in July 2019 that appeared to announce such a prohibition for the first time.[209] The blog post noted, however, that the prohibition would only apply to one protected category – religion - for the time being, while the company assessed whether expanding the prohibition to other protected categories was necessary and proportionate to the potential severity of harm. In March 2020, Twitter updated the post to reflect expansion of the prohibition to content that dehumanizes on the basis of age,

[203] https://web.archive.org/web/20181028023901/https://help.twitter.com/en/rules-and-policies/hateful-conduct-policy
[204] https://web.archive.org/web/20181028023901/https://help.twitter.com/en/rules-and-policies/hateful-conduct-policy
[205] https://web.archive.org/web/20181028023901/https://help.twitter.com/en/rules-and-policies/hateful-conduct-policy
[206] https://web.archive.org/web/20181028023901/https://help.twitter.com/en/rules-and-policies/hateful-conduct-policy
[207] https://web.archive.org/web/20181028023901/https://help.twitter.com/en/rules-and-policies/hateful-conduct-policy
[208] https://web.archive.org/web/20181028023901/https://help.twitter.com/en/rules-and-policies/hateful-conduct-policy
[209] https://web.archive.org/web/20190710034657/https://blog.twitter.com/en_us/topics/company/2019/hatefulconductupdate.html

disability, or disease – in addition to religion.[210] For example, Tweets like "All [Age Group] are leeches and don't deserve any support from us" or "People with [Disability] are subhuman and shouldn't be seen in public" would be removed. In December 2020, Twitter expanded the prohibition to include content that dehumanizes on the basis of race, ethnicity, or national origin, providing examples like "There are too many [national origin/race/ethnicity] maggots in our country and they need to leave!"[211] In December 2021, the company announced that the ban on dehumanizing language now extended to all protected categories.[212]

Twitter made several additional changes to the hateful conduct policy after 2018. In some cases, these updates took the form of policy expansion. For example, in 2020, Twitter added caste to the list of protected characteristics.[213] In other cases, the updates involved clarifying Twitter's approach to enforcement. In October 2021, Twitter added a list of potential responses to violations of the hateful conduct policy, including downranking Tweets, making Tweets ineligible for amplification or recommendations, requiring Tweet removal, and suspending accounts.[214]

The most obvious recent change to Twitter's hateful conduct policy came in February 2023, after Elon Musk took over the company. In this update, the policy language was significantly pared down.[215] Nevertheless, despite Elon Musk's stated free speech policy, the breadth of content covered by the policy did not change dramatically. Previously, the policy prohibited promoting violence against, threatening, and wishing, hoping, or calling for serious harm against people based on protected characteristics. While these prohibitions disappeared in February 2023, Twitter added a note explaining that incitement to violence was covered by Twitter's Violent Speech policy. The policy still included prohibitions on hateful references to violent events where a protected category was the primary victim, incitement of fear, harassment, or economic discrimination, slurs, dehumanization, hateful imagery, and hateful profiles.[216]

210 https://web.archive.org/web/20200305193131/https://blog.twitter.com/en_us/topics/company/2019/hatefulconductupdate.html
211 https://web.archive.org/web/20201202183713/https://blog.twitter.com/en_us/topics/company/2019/hatefulconductupdate.html
212 https://web.archive.org/web/20211215194611/https://blog.twitter.com/en_us/topics/company/2019/hatefulconductupdate
213 https://web.archive.org/web/20210122154659/https://help.twitter.com/en/rules-and-policies/hateful-conduct-policy
214 https://web.archive.org/web/20211030195631/https://help.twitter.com/en/rules-and-policies/hateful-conduct-policy
215 https://web.archive.org/web/20230301042918/https://help.twitter.com/en/rules-and-policies/hateful-conduct-policy
216 On April 18, 2023, GLAAD reported that Twitter had removed its prohibition on targeted misgendering or deadnaming of transgender individuals from the Hateful Conduct Policy without public announcement, seemingly on April 8, 2023. This change is outside the temporal scope of our analysis (which ends on April 1, 2023), however, so we do not cover it in the main body of the report. See "GLAAD Responds to Twitter's Roll-Back of Long-Standing LGBTQ Hate Speech Policy," April 18, 2023, GLAAD Press Release, https://www.glaad.org/releases/glaad-responds-twitters-roll-back-long-standing-lgbtq-hate-speech-policy#:~:text=transgender-,GLAAD%20RESPONDS%20TO%20TWITTER'S%20ROLL%2DBACK%20OF%20LONG,STANDING%20LGBTQ%20HATE%20SPEECH%20POLICY&text=GLAAD%3A%20%E2%80%9CTwitter's%20decision%20to%20covertly,for%20users%20and%20advertisers%20alike.%E2%80%9D.

Analysis of Policy Scope

Table 13 illustrates that Twitter has prohibited a broad range of content under its hateful conduct policy since 2017, when the company first introduced an explicit prohibition on hate speech. Until early 2023, the company defined hateful conduct as promoting violence against, directly attacking, or threatening people based on certain identity-based characteristics, and provided detailed information about the types of content that definition covered: violent threats, expressing desire that others suffer serious harm, referring to violent events where protected groups were the primary victims, inciting fear about a protected category, and hateful imagery or profiles. This definition is broader than the prohibition on advocacy of hatred required by Article 20(2) and permitted under Article 19(3), and a more general prohibition against "slurs" and "harmful stereotypes" based on protected characteristics would likely fall a foul of the strict requirement of necessity, absent cases where such speech fulfills the requirements of "intent" and "imminence" of serious harm such as "discrimination", "hostility" or "violence". Though Twitter limited its definition of hate speech to be direct attacks on the basis of protected characteristics in early 2023, the scope of covered content remained relatively broad. The new definition eliminated incitement to violence and expressing wishes for harm, but these forms of content remain prohibited by Twitter's Violent Speech policy.

As Table 14 demonstrates, Twitter's policy also covers a wider range of protected characteristics than the prohibition on hatred in Article 20(2). In contrast to many other platforms, however, Twitter has not significantly expanded its list of protected characteristics over time. Twitter's first hate speech policy listed ten protected characteristics, and the only expansion in this list occurred in 2020, when the company added protection for caste. Nor has Twitter followed the trend towards prohibiting the denial of historical atrocities.

Table 13

Content Explicitly Covered by Twitter's Hate Speech Policies		2017	2018	2019	2020	2021	2022	2023
Hate(ful) speech/ content		X*	X*	X*	X*	X*	X*	X*
Promotion of Hatred								
Support for Organized Hate (Including Symbols)		X	X	X	X	X	X	X
On the basis of protected characteristics	Incitement to or Threats of Violence	X	X	X	X	X	X	X†
	Attacks		X	X	X	X	X	X
	Statements of inferiority or content that demeans	X	X	X	X	X	X	X
	Dehumanization		X	X	X	X	X	X
	Expressions of contempt or disgust							
	Calls for exclusion or segregation							
	Discrimination					X	X	X
	Denying or mocking historical atrocities, or valorizing the perpetrators							
	Slurs	X	X	X	X	X	X	X
	Harmful Stereotypes		X	X	X	X	X	X
	Conspiracy Theories							
	Cursing							

* The expression of hatred in profile bios is banned by Twitter's abusive profile information policy.

† Twitter removed the prohibition on incitement to violence against protected groups from the hateful conduct policy in February 2023, but they added a note explaining that such speech is covered by Twitter's violent speech policy.

Table 14

Characteristics Protected in Twitter's Hate Speech Policies							
	2017	2018	2019	2020	2021	2022	2023
Total	**10**	**10**	**10**	**11**	**11**	**11**	**11**
Race	X	X	X	X	X	X	X
Ethnicity	X	X	X	X	X	X	X
National Origin	X	X	X	X	X	X	X
Religion	X	X	X	X	X	X	X
Gender	X	X	X	X	X	X	X
Color							
Immigration Status							
Sex							
Gender Identity	X	X	X	X	X	X	X
Sexual Orientation	X	X	X	X	X	X	X
Age	X	X	X	X	X	X	X
Disability	X	X	X	X	X	X	X
Disease/ Medical Condition	X	X	X	X	X	X	X
Veteran Status							
Occupation							
Weight							
Pregnancy							
Caste				X	X	X	X
Victims of a Major Event							
Socio-Economic Status							
Culture							
Tribe							

Notes: An X indicates the company's hate speech policies covered that protected characteristic for at least one month during the given year.

Changes in Enforcement Volume

Twitter does not make information about Rules enforcement available for periods prior to the second half of 2018, which is unfortunate given the company introduced a hate speech policy for

the first time in late 2017 and made the most significant changes since in late 2018. Nevertheless, Twitter's enforcement reports purport to share information about the impact of changes in the scope of Twitter's hate speech policy. However, the reports' conclusions do not necessarily align with our research about the scope of Twitter's policy at different points in time.

For example, in July 2021, Twitter published a Transparency Report that showed a 77% increase in the number of accounts actioned for violations of the hateful conduct policy, from 635,415 to 1,126,990, for the period between July 1 to December 31, 2020.[217] To explain this increase, Twitter stated: "In September 2020, we began enforcing our hateful conduct policy against content that incites fear and/or fearful stereotypes about protected categories… in December 2020, we further expanded our hateful conduct policy to include content that dehumanizes on the basis of race, ethnicity, or national origin." This explanation is puzzling, since the data we collected from the WayBack machine suggest Twitter prohibited "content that intends to... reinforce negative or harmful stereotypes about a protected category," including content intended to "incite fear or spread fearful stereotypes," as early as October 2018.[218]

It is possible that Twitter did not start enforcing the fearful stereotypes prohibition until September 2020, two years after it appeared in the hateful conduct policy. Alternatively, the expansion of the dehumanization prohibition to race, ethnicity, and national origin in December 2020 - beyond religion, age, disease, and disability (as described above) - alone could account for the increase, though that would be surprising given it occurred in the final month of the reporting period. Lastly, it is possible that Twitter's explanation for the massive increase in hate speech content actioned is simply inaccurate. Regardless of the true explanation, this example illustrates the importance of researchers gaining access to platform data, so they can thoroughly assess and audit platforms' claims about policy enforcement. It is also problematic in terms of the human rights requirement of legality, since it is unclear when (and thus how) users would have been subject to the enforcement of the prohibition on fearful stereotypes of protected categories.

For the next reporting period, January through June 2021, Twitter reported a 2% decrease in the number of accounts actioned for violations of the hateful conduct policy. This decrease occurred even though Twitter expanded the scope of the policy during this time to include "content that incites others to discriminate by denying support to the economic enterprise of an individual or group because of their perceived membership in a protected category."[219] Because Twitter did not reduce the scope of its hate speech policy during this time, other factors, outside the policy's scope, likely contributed to the decline in content actioned. The lack of an increase in content

[217] "An Update to the Twitter Transparency Center," Twitter, July 14, 2021, https://blog.twitter.com/en_us/topics/company/2021/an-update-to-the-twitter-transparency-center .
[218] https://web.archive.org/web/20181028023901/https://help.twitter.com/en/rules-and-policies/hateful-conduct-policy
[219] "Rules Enforcement: Jan - Jun 2021," *Twitter, Transparency*, https://transparency.twitter.com/en/reports/rules-enforcement.html#2021-jan-jun .

actioned in this period could also indicate that there is not much economic discrimination posted on Twitter. On the other hand, if there was a significant drop in other forms of content that fall under this policy, Twitter could have actioned on quite a lot of economic discrimination – even though the amount of content actioned overall did not increase in this reporting period. All this speculation underscores how impossible it is to know how the policy is being enforced, as well as the impact of changes in policy scope on enforcement, without access to the company's data on content actioned under each provision in the hateful conduct policy.

As of April 1, 2023, the most recent transparency report available from Twitter covers the period from July through December 2021. Compared to the previous report, the report shows a 19% decrease in the number of accounts actioned for violations of the hateful conduct policy.[220] It also notes that the company expanded the prohibition on dehumanizing speech in December 2021 to include all protected categories, though the number of accounts suspended under this dehumanization prohibition from July 2021 to December 2021 amounted to 104,565, a 22% decrease since the last report.[221] This decrease suggests that factors other than changes in policy scope impacted the amount of content actioned.

[220] "Rules Enforcement: Jul - Dec 2021," *Twitter, Transparency*, https://transparency.twitter.com/en/reports/rules-enforcement.html#2021-jul-dec .
[221] "Rules Enforcement: Jul - Dec 2021," *Twitter, Transparency,*

8. YOUTUBE

- **Release/Launch Date:** 14 February 2005
- **Number of Active Users:** 2.562 billion [222]
- **Short Overview of Moderation Process:** Content moderators review posts that have been flagged by AI and reported by users. The majority of this work is outsourced to third-party vendors.
- **Signatory to the EU'S Code of Conduct on Illegal Hate Speech Online:** Yes (Google)

[222] "Most popular social networks worldwide as of January 2023, ranked by number of monthly active users," *Statista*, https://www.statista.com/statistics/272014/global-social-networks-ranked-by-number-of-users/ (accessed on April 30, 2023).

Key Developments

Terms of Use

YouTube's first traceable Terms of Use date back to 2005. They required that users not post material that is "hateful or racially, ethnically or otherwise objectionable."[223] Since 2007, however, YouTube's Terms of Use have not included a provision on hate speech.

Community Guidelines

Since 2006, however, YouTube has addressed hate speech in its Community Guidelines. Although the content covered by the policy has changed over time, the company has always stated its commitment to free expression before outlining its prohibitions on hate speech. The first version of the Community Guidelines underscored YouTube's commitment to defending "everyone's right to express unpopular points of view" but prohibited hate speech, defined as speech containing slurs or malicious stereotypes intended to attack or demean individuals on the basis of certain characteristics (see Figure 14). This initial conceptualization of hate speech was relatively broad, since it includes slurs and stereotypes rather than incitement to violence or threats on the basis of protected characteristics. In mid-2008, however, YouTube removed the reference to slurs and stereotypes, defining hate speech as speech that attacks or demeans a group based on certain characteristics.[224] Later that year, YouTube added an additional section to the Community Guidelines titled "Community Guideline Tips." There, the company defined hate speech as "content that promotes hatred against members of a protected group," such as "racist or sexist content."[225] It did not make any relevant updates to the Community Guidelines again until 2014. Thus, from late 2008 through 2014, YouTube's hate speech policies prohibited content that promoted hate, attacked, demeaned, or discriminated on the basis of protected characteristics.

Figure 14[226]

- We encourage free speech and defend everyone's right to express unpopular points of view. But we don't permit **hate speech** which contains **slurs** or the **malicious use of stereotypes** intended to attack or demean a particular gender, sexual orientation, race, religion, or nationality.

In late 2014, YouTube substantially revised the hate speech provision in its Community Guidelines. Instead of defining hate speech as content that attacks or demeans protected groups, the company defined it as content that "incites hatred against members of a protected group."[227] It also explicitly prohibited "content that promotes or condones violence against individuals or groups" based on protected characteristics.[228] While this conceptualization of hate speech remained the status quo for YouTube's Community Guidelines for the next six or so years,

[223] https://web.archive.org/web/20050428210756/http://www.youtube.com/terms.php
[224] https://web.archive.org/web/20080611231521/http://www.youtube.com/t/community_guidelines
[225] https://web.archive.org/web/20081112004550/http://www.youtube.com/t/community_guidelines
[226] https://web.archive.org/web/20061024061946/http://www.youtube.com/t/community_guidelines
[227] https://web.archive.org/web/20141105093019/https://www.youtube.com/t/community_guidelines
[228] https://web.archive.org/web/20141105093019/https://www.youtube.com/t/community_guidelines

YouTube also began addressing hate speech more thoroughly in an additional, accompanying policy. In March 2014, YouTube launched this "Hate Speech Policy," which appeared to be separate yet complementary to the Community Guidelines. The first version of this policy defined hate speech as "content that promotes violence or hatred against individuals or groups based on certain attributes," aligning closely with the relevant provision in the Community Guidelines.[229]

Throughout 2019, YouTube revised the definition of hate speech implied by its policy.[230] While the company maintained a prohibition on the promotion of violence or hatred against protected groups, it also added a list of examples of covered content to the policy language. This list included content that dehumanizes, states the inferiority of, calls for subjugation, and attacks on the basis of protected characteristics, as well as slurs, stereotypes, and conspiracy theories. It also prohibited the denial of well-documented events, such as claims that all the supposed victims of a crime were actors. This prohibition did not specify that the victims had to be members of a protected group, though one might infer that requirement from the structure of the policy language. Additionally, the policy prohibited "content containing hateful supremacist propaganda" or "music videos promoting hateful supremacism in the lyrics, metadata, or imagery."[231]

In early 2019, YouTube also added a section clarifying one policy exception. Under the heading "Educational Content," the company explained that hate speech might be allowed "if the primary purpose is educational, documentary, scientific, or artistic in nature."[232] By mid-2019, the section also explained that users had to clearly state the educational context in the video itself, noting that mentioning the educational nature of the hate speech in the title or description of the video was insufficient.[233] In 2020, YouTube deleted the freestanding hate speech provision from its Community Guidelines and just added a link to the broader hate speech policy. In 2021, YouTube added a line clarifying that the educational content exception also applied to external links provided in videos.[234]

In 2019, YouTube also added a provision relevant to hate speech to its policy on harassment and cyberbullying. "We do not allow content that targets individuals with prolonged or malicious insults based on intrinsic attributes, including their protected group status or physical traits."[235] The phrase "protected group status" included a hyperlink to the hate speech policy, implying that

[229] https://web.archive.org/web/20140329023647/https://support.google.com/youtube/answer/2801939?hl=en&ref_topic=2803176
[230] https://web.archive.org/web/20191114002846/https://support.google.com/youtube/answer/2801939?hl=en&ref_topic=9282436
[231] https://web.archive.org/web/20191114002846/https://support.google.com/youtube/answer/2801939?hl=en&ref_topic=9282436
[232] https://web.archive.org/web/20190407161800/https://support.google.com/youtube/answer/2801939?hl=en&ref_topic=2803176
[233] https://web.archive.org/web/20190605213123mp_/https://support.google.com/youtube/answer/2801939?hl=en
[234] https://web.archive.org/web/20211208081210/https://support.google.com/youtube/answer/2801939?hl=en&ref_topic=9282436#zippy=%2Ceducational-content%2Cother-types-of-content-that-violates-this-policy%2Cmore-examples
[235] https://web.archive.org/web/20191213125923/https://support.google.com/youtube/answer/2802268

the two policies protected the same characteristics. In addition to slurs and other forms of content already specified in the hate speech policy, the harassment and cyberbullying provision banned repeatedly showing pictures of someone and expressing disgust about their attributes, publishing nonpublic personal identifying information, and stalking or blackmail.

The list of protected characteristics covered by YouTube's hate speech provisions also changed over time. In 2005[236] and 2006[237], YouTube's Terms of Use prohibited "racially" or "ethnically objectionable" content, suggesting a ban on language that discriminated on the basis of race or ethnicity. YouTube's 2006 Community Guidelines prohibited slurs and malicious stereotypes that attacked or demeaned on the basis of gender, sexual orientation, race, religion, or nationality.[238] By 2008, the Guidelines no longer protected nationality, but they still prohibited hate speech on the basis of race, religion, sexual orientation, and gender, as well as new characteristics like age, veteran status, disability, and gender identity.[239] In 2014, YouTube added nationality back to the list.[240] The last major change in the list of protected characteristics came in 2019, when the company added protections for immigration status, sex, caste, and victims of a major event.[241] Thus, as of 2023, YouTube's hate speech provisions prohibit such content on the basis of fourteen different characteristics.

Analysis of Policy Scope

In contrast to most other platforms, which have shown a uniform expansion in the scope of their hate speech definitions over time, YouTube's conceptualization of hate speech initially shrunk in scope and then later expanded (see Table 15.) From 2006 to mid-2008, YouTube prohibited slurs and stereotypes that attacked or demeaned protected groups, an arguably broader definition of hate speech than the promotion/incitement of hatred or violence, the definition YouTube adopted a few years later. In 2019, however, YouTube returned to prohibiting slurs and stereotypes, while also adding several entirely new types of covered content, including dehumanization, statements of inferiority, hateful conspiracy theories, and calls for exclusion and discrimination. YouTube's definition of hate speech is much broader than the mandatory prohibition on hatred in the ICCPR's Article 20(2) and permitted restrictions under Article 19(3). Several categories such as "expression of contempt or disgust", "slurs," and "harmful stereotypes likely fall foul of the strict requirement of necessity, absent cases where such speech fulfills the requirements of "intent" and "imminence" of serious harm such as "discrimination", "hostility" or "violence." This is undoubtedly true for "denying or mocking historical atrocities" and "conspiracy theories."

[236] https://web.archive.org/web/20050428210756/http://www.youtube.com/terms.php
[237] https://web.archive.org/web/20060410020756/http://youtube.com/t/terms
[238] https://web.archive.org/web/20061024061946/http://www.youtube.com/t/community_guidelines
[239] https://web.archive.org/web/20080611231521/http://www.youtube.com/t/community_guidelines
[240] https://web.archive.org/web/20141105093019/https://www.youtube.com/t/community_guidelines
[241] https://web.archive.org/web/20190605213123mp_/https://support.google.com/youtube/answer/2801939?hl=en

Table 15

	Content Explicitly Covered by YouTube's Hate Speech Policies	2005	2006	2007	2008	2009	2010	2011	2012	2013	2014	2015	2016	2017	2018	2019	2020	2021	2022	2023
	Hate(ful) speech/ content	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X
	Promotion of Hatred				X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X
	Support for Organized Hate (Including Symbols)															X	X	X	X	X
On the basis of protected characteristics	Incitement to or Threats of Violence										X	X	X	X	X	X	X	X	X	X
	Attacks		X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X
	Statements of inferiority or content that demeans		X	X	X	X	X	X	X	X	X					X	X	X	X	X
	Dehumanization															X	X	X	X	X
	Expressions of contempt or disgust															X*	X*	X*	X*	X*
	Calls for exclusion or segregation															X	X	X	X	X
	Discrimination	X	X	X	X	X	X	X	X	X	X	X				X	X	X	X	X
	Denying or mocking historical atrocities, or valorizing the perpetrators															X	X	X	X	X
	Slurs		X	X	X											X	X	X	X	X
	Harmful Stereotypes		X	X	X											X	X	X	X	X
	Conspiracy Theories															X	X	X	X	X
	Cursing																			

* The expression of disgust regarding intrinsic attributes, including protected characteristics, is banned by YouTube's Harassment & Cyberbullying policy.

The scope of the protected characteristics covered by YouTube's hate speech policies has also grown significantly over time and certainly goes beyond the characteristics covered by the ICCPR definitions (see Table 16.) Between 2005 and 2009, YouTube's list of protected characteristics grew from two (race and ethnicity) to nine (race, ethnicity, religion, gender, gender identity, sexual orientation, age, disability, and veteran status). Thus, as early as 2009, YouTube's list of protected categories went far beyond the characteristics covered by Article 20(2). YouTube also added five more characteristics to the list by the end of 2019, including nationality, immigration status, sex, caste, and being a victim of a major event. Accordingly, YouTube´s hate speech policies raise serious concerns when measured against international human rights standards on freedom of expression and hate speech.

Table 16

Characteristics Protected in YouTube's Hate Speech Policies																			
	2005	2006	2007	2008	2009	2010	2011	2012	2013	2014	2015	2016	2017	2018	2019	2020	2021	2022	2023
Total	**2**	**6**	**5**	**9**	**9**	**9**	**9**	**9**	**9**	**10**	**10**	**10**	**10**	**10**	**14**	**14**	**14**	**14**	**14**
Race	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X
Ethnicity	X	X		X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X
National Origin		X	X							X	X	X	X	X	X	X	X	X	X
Religion		X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X
Gender		X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X
Color																			
Immigration Status															X	X	X	X	X
Sex															X	X	X	X	X
Gender Identity				X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X
Sexual Orientation		X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X
Age				X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X
Disability				X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X
Disease/ Medical Condition																			
Veteran Status				X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X
Occupation																			
Weight																			
Pregnancy																			
Caste															X	X	X	X	X
Victims of a Major Event															X	X	X	X	X
Socio-economic Status																			
Culture																			
Tribe																			

Notes: An X indicates the company's hate speech policies covered that protected characteristic for at least one month during the given year.

Changes in Enforcement Volume

YouTube provides information about channels and videos removed due to Community Guidelines violations, by removal reason, dating back to Q4 2018. According to our research, YouTube expanded the scope of content explicitly covered by its hate speech policy throughout 2019. According to YouTube's transparency reports, the company removed 18,950 videos for violating the hate speech policy in Q4 2018,[242] 0.2% of the 8,765,783 total videos removed during the period.[243] A year later, in Q4 2019, YouTube removed 88,589 videos because they violated the hate speech policy,[244] 1.5% of the 5,887,021 total videos removed.[245] This data suggests that the expansion in YouTube's hate speech policy over the course of 2019 is correlated with an increase in the number of accounts actioned, though we cannot make a causal claim. In Q4 2018, YouTube removed 26,867,027 comments because they were hateful or abusive, 1.4% of all comments removed.

That being said, in Q2 2020, the percentage of videos removed due to hate speech fell to 0.7%, or 80,033[246], of 11,401,696 total videos removed[247], but there was no noticeable reduction in the scope of YouTube's hate speech policy during this period. The decline may have had something to do with human review capacity during the early days of the pandemic, but there is no real way to know without more information from YouTube. Moreover, the percentage of videos removed due to hate speech was back to 1.1%[248] of 7,872,684 total videos removed[249] in Q3 2020. This data illustrates that content actioned under a particular policy can change for reasons other than policy

[242] "Featured Policies: Hate Speech, Oct 2018 - Dec 2018," *Google Transparency Report*, https://transparencyreport.google.com/youtube-policy/featured-policies/hate-speech?hl=en&policy_removals=period:2018Q4&lu=policy_removals .
[243] "YouTube Community Guidelines Enforcement: Oct 2018 - Dec 2018," *Google Transparency Report*, https://transparencyreport.google.com/youtube-policy/removals?hl=en&total_removed_videos=period:2019Q4;exclude_automated:all&lu=total_removed_videos .
[244] "Featured Policies: Hate Speech, Oct 2019 - Dec 2019," *Google Transparency Report*, https://transparencyreport.google.com/youtube-policy/featured-policies/hate-speech?hl=en&policy_removals=period:2019Q4&lu=policy_removals.
[245] "YouTube Community Guidelines enforcement: Oct 2019 - Dec 2019," *Google Transparency Report*, https://transparencyreport.google.com/youtube-policy/removals?hl=en&total_removed_videos=period:2019Q4;exclude_automated:all&lu=total_removed_videos .
[246] "Featured Policies: Hate Speech, Apr 2020 - Jun 2020," *Google Transparency Report*, https://transparencyreport.google.com/youtube-policy/featured-policies/hate-speech?hl=en&policy_removals=period:2020Q2&lu=policy_removals.
[247] "YouTube Community Guidelines enforcement: Apr 2020 - June 2020," *Google Transparency Report*, https://transparencyreport.google.com/youtube-policy/removals?hl=en&total_removed_videos=period:2020Q2;exclude_automated:all&lu=total_removed_videos&total_channels_removed=period:2020Q2
[248] "Featured Policies: Hate Speech, Jul 2020 - Sep 2020," *Google Transparency Report*, https://transparencyreport.google.com/youtube-policy/featured-policies/hate-speech?hl=en&policy_removals=period:2020Q3&lu=policy_removals .
[249] "YouTube Community Guidelines enforcement: Jul 2020 - Sep 2020," *Google Transparency Report*, https://transparencyreport.google.com/youtube-policy/removals?hl=en&total_removed_videos=period:2020Q3;exclude_automated:all&lu=total_removed_videos .

scope increases or decreases, underscoring the difficult of making any kind of causal claims about the impact of changes in policy scope without access to more robust platform data.

YouTube also reports data on the number of comments removed, but this data is not available prior to Q3 2019, which prevents a comparison of comments removed due to hate speech before and after the major changes in 2019.

Part 2: Cross-Platform Trends in the Scope of Hate Speech Policies

The previous section demonstrated scope creep, or a gradual increase in the types of content covered and protected characteristics mentioned, in the hate speech policies of most platforms, between the platform's founding and the current day. Every platform analyzed, excluding Reddit, has expanded the scope of its hate speech policy since first creating one. This section aggregates all of the information described in the previous section to identify overall trends in the scope of the eight platforms' hate speech policies. According to our research, more than half of the analyzed platforms went from having no explicit prohibition on hate speech (Reddit, Snapchat[250], Twitter, and TikTok- then Musical.ly) in 2014 to prohibiting a broad range of content targeting protected groups today, including speech that does not directly attack or call for violence. The results also reveal that the average number of protected characteristics has almost doubled in the past decade.

Table 17

Percent of Analyzed Platforms with Hate Speech Policies That Cover Listed Content		2014	2018	2023
Hate(ful) speech/ content		38%	88%	88%
Promotion of Hatred		25%	50%	50%
Support for Organized Hate (Including Symbols)		0%	38%	75%
On the basis of protected characteristics	Incitement to or Threats of Violence	25%	88%	88%
	Attacks	25%	50%	75%
	Statements of inferiority or content that demeans	13%	38%	75%
	Dehumanization	0%	38%	63%
	Expressions of contempt or disgust	0%	13%	25%
	Calls for exclusion or segregation	0%	13%	50%
	Discrimination	25%	25%	50%
	Denying or mocking historical atrocities, or valorizing the perpetrators	0%	13%	50%
	Slurs	0%	25%	50%
	Harmful Stereotypes	0%	13%	38%
	Conspiracy Theories	0%	0%	25%
	Cursing	0%	13%	13%

[250] Snapchat prohibited racism in 2014 - but not hate speech explicitly.

Table 17 lists the percentage of platforms that prohibited different types of content under their hate speech policies in three separate years: 2014 - the first year all eight platforms existed, 2018- the year after NetzDG was passed, and this year, 2023. In 2014, less than half of the analyzed platforms prohibited hate speech, but just four years later, that metric had climbed to 88%. In fact, since 2014, the percentage of platforms prohibiting all the categories of content listed in our tables has increased. Today, 88% of platforms prohibit incitement to violence and threats of violence based on protected characteristics, a form of hate speech explicitly prohibited by the ICCPR, up from 25% in 2014. More than half of all platforms now also prohibit statements of inferiority and dehumanization based on protected characteristics, up from less than 15% in 2014. Table 17 also reveals smaller increases in the percentage of platforms prohibiting the denial of historical atrocities, harmful stereotypes, conspiracy theories, and cursing.

However, the mandatory prohibition on advocacy of hatred in Article 20 of the ICCPR does not mention statements of inferiority, dehumanizing language, stereotypes, conspiracy theories, or cursing. Moreover, the permitted restrictions on freedom of expression under Article 19 and the three-part test raises serious questions about whether the general scope creep identified is compatible with the requirements of legality, legitimacy and necessity, which all the analyzed platforms [except Tumblr and Reddit] have committed to respecting through their human rights policies and adoption of the General Principles. The expanding scope of platform policies illustrated by Table 17 is more in line with the case law from the ECtHR described in the introduction, though scholars like Evelyn Aswad and David Kaye have argued regional standards should not supersede international ones.[251]

Table 17 also reveals that scope creep occurred both before and after 2018, suggesting platforms expanded their hate speech policies at different times. While the introduction of legislation like NetzDG may have contributed to the general trend, these legislative developments did not suddenly lead to a uniform decision across platforms to expand their hate speech prohibitions. Likely there are a variety of internal and external factors that influence when and how platforms change their policies. As mentioned earlier in the report, this analysis is a descriptive endeavor; it does not demonstrate what caused any changes in platform policies – but rather simply documents the changes that have occurred over time. Future research should investigate these factors more deeply, and platforms should provide researchers the necessary access to disaggregated data needed to more confidently establish (any) causal relationship between changes in platform policies and changes in enforcement. Access to such data could also facilitate

[251] Evelyn Aswad & David Kaye, 'Convergence & Conflict: Reflections on Global and Regional Human Rights Standards on Hate Speech.' (2022) 20 *Northwestern Journal of Human Rights* 3, pg. 168.

future research on the extent to which the enforcement of these policies conforms with the human rights standards that most of these platforms claim to respect.

Figure 15

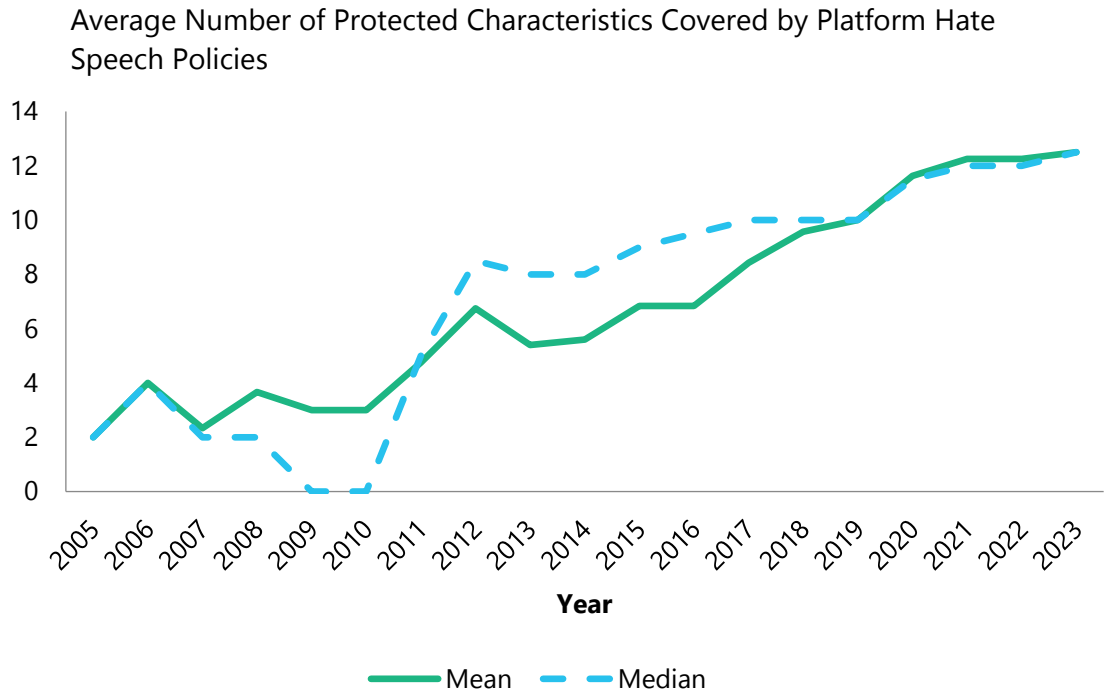


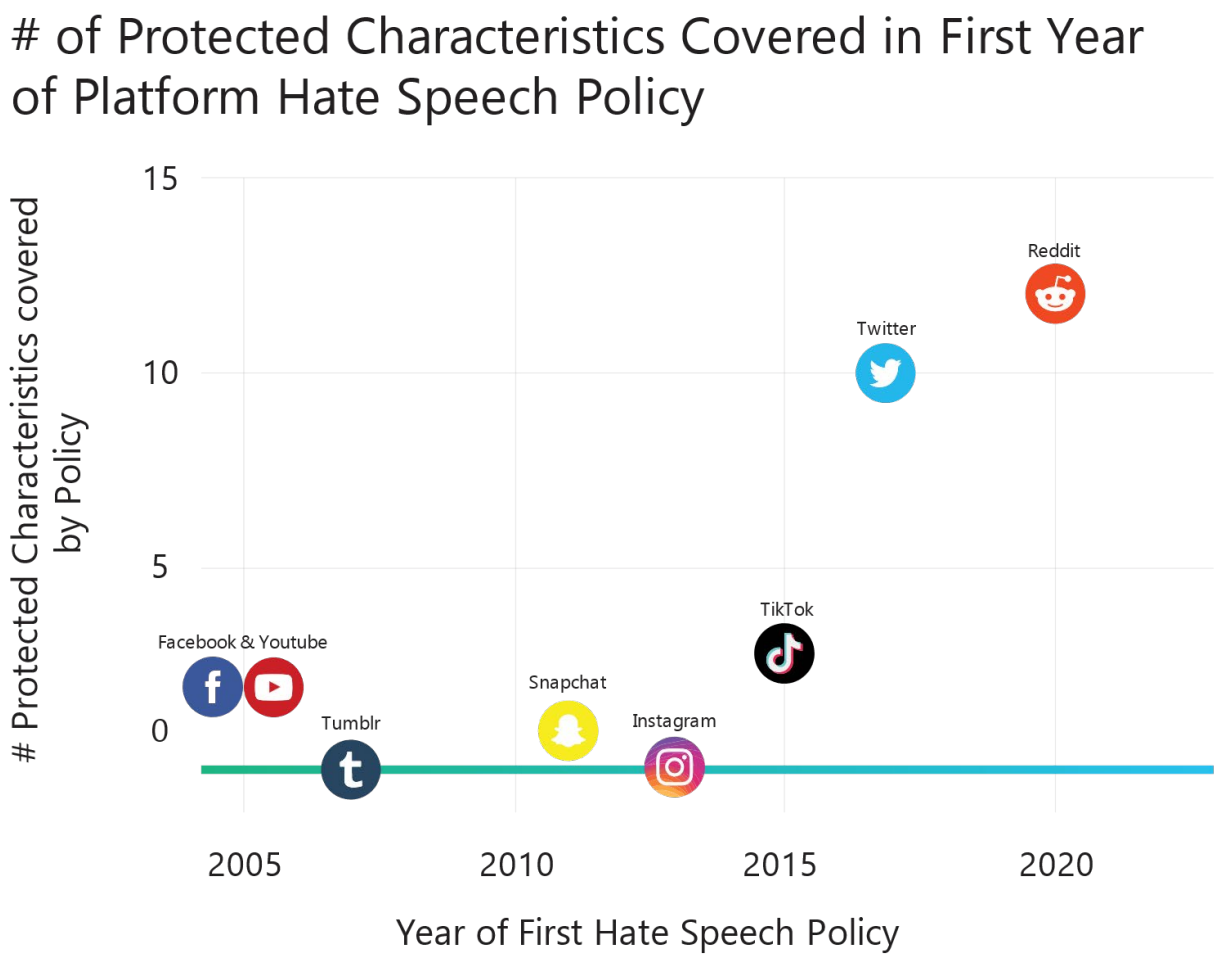
Figure 15 illustrates that the scope of protected characteristics in platform hate speech policies has also gradually increased over time. Prior to 2011, the mean number of protected characteristics in a hate speech policy was less than five, but by 2017, the mean had grown to eight. By 2020, the mean was twelve. This expansion in protected characteristics may be reflective of what Eugene Volokh calls "censorship envy", the sense that "if my neighbor gets to ban speech he reviles, why shouldn't I get to do the same?", and where different groups pressure platforms to afford them protection based on the inclusion on other groups, which makes it difficult for platforms to deny such without appearing biased or discriminatory.[252] Again, future research should investigate whether these pressures are directly responsible for the scope creep this report identifies.

Figure 16 reveals that the baseline expectation for platforms' hate speech policies, in terms of protected characteristics, may have shifted over time. Platforms that introduced a hate speech provision prior to 2015 included far fewer characteristics in their first policy than the two platforms

[252] Greg Lukianoff and Ryne Weiss, "The NYPost & Twitter Crash Into 'The Streisand Effect,' 'Censorship Envy,' and 'the Slippery Slope Tendency," *The Fire*, October 15, 2020, https://www.thefire.org/the-nypost-twitter-crash-into-the-streisand-effect-censorship-envy-and-the-slippery-slope-tendency/.

that introduced a prohibition after 2015, Twitter and Reddit. This finding could indicate that Twitter and Reddit simply looked to existing platform policies for a blueprint when they decided to ban hate speech, and by that point, most platforms were protecting a wide range of characteristics. Theoretically, it could also indicate that Twitter and Reddit had a more maximalist, underlying view of hate speech than the other platforms. This argument is less compelling, however, when one considers that both platforms did not ban hate speech for more than a decade after starting operations, and publicly advocated free speech positions more aligned with First Amendment principles until 2017 and 2020 respectively.

Figure 16



Part 3: Erroneous and Inconsistent Enforcement

This section provides anecdotal evidence of erroneous and inconsistent enforcement of four of the eight platforms' hate speech policies. We chose to focus on Facebook, Instagram, TikTok, and YouTube because they are the four largest platforms in our sample.[253] While the expansion in protected characteristics documented in this report likely contributed to a safer online environment for minorities in some instances, there are also negative consequences for minority expression associated with hate speech policy enforcement. Though each of these companies rationalized their policies by noting that hate speech can silence minority expression, the anecdotes in this section illustrate that those hate speech policies sometimes had the exact effect they were designed to address. Often, erroneous hate speech removals occur when vulnerable populations try to raise awareness about the abuse they have suffered, despite platforms' promises of exceptions for educational content.

Platforms regularly blame mistakes or inconsistencies in hate speech policy enforcement on poorly calibrated hate speech classifiers, which are often worse for languages that are less represented on the platform or in training data. Thus, minority populations are likely the populations most impacted by inaccurate hate speech detection algorithms, further underscoring that platform hate speech policies may not effectively protect the vulnerable and minority populations they claim to. As Casey Fiesler, a scholar of technology ethics at the University of Colorado, Boulder, puts it, both the under-enforcement and the over-enforcement of platform hate speech policies "harm the same people: those who are disproportionately targeted for abuse end up being algorithmically censored for speaking out about it."[254] While we recognize that platforms must rely on automated classifiers to enforce their policies, given the scale of the speech being moderated, it is important to highlight the negative consequences of enforcing very broad hate speech policies at this scale. By documenting this tradeoff, we hope to encourage platforms and external observers to more seriously weigh the benefits of enforcing the status quo hate speech policies at scale against the unintended consequences of doing so.

Facebook

The rationale for Facebook's hate speech policy clearly states that the company prohibits hate speech both to prevent offline violence *and* to create a safe space for minority expression. Facebook clearly states that hate speech is not allowed because it "creates an environment of intimidation and exclusion," noting that "people use their voice and connect more freely when they don't feel attacked on the basis of who they are."[255] However, Facebook has sometimes

[253] "Most popular social networks worldwide as of January 2023, ranked by number of monthly active users," *Statista*, https://www.statista.com/statistics/272014/global-social-networks-ranked-by-number-of-users/ (accessed on April 30, 2023).
[254] Abby Ohlheiser, "Welcome to TikTok's endless cycle of censorship and mistakes," *MIT Technology Review*, July 13, 2021, https://www.technologyreview.com/2021/07/13/1028401/tiktok-censorship-mistakes-glitches-apologies-endless-cycle/ .
[255] "Hate speech," Meta Transparency Center, https://transparency.fb.com/policies/community-standards/hate-speech/ (accessed February 1, 2023).

enforced its hate speech policies in a way that has resulted in the repression of minority voices. After Nigerian-American writer and author of *So You Want to Talk about Race*, Ijeoma Oluo, posted screenshots of racist messages and death threats she'd received on Facebook, her account was suspended and Facebook left the death threats unaddressed.[256] Similarly, Facebook disabled the account of Stacey Patton, a journalism professor at Morgan State University, after she posted a critique of the Black-on-Black crime trope, asking why "it's not a crime when White freelance vigilantes and agents of 'the state' are serial killers of unarmed Black people, but when Black people kill each other then we are 'animals' or 'criminals."[257]

Facebook has also enforced its hate speech policy inconsistently in the past. In 2017, ProPublica reported that Facebook paid little attention to the intersections and subgroups of protected classes, resulting in bizarre policies and training materials that left Black children unprotected but assigned protected class status to white men.[258] In March 2020, Facebook removed several accounts, pages, and groups associated with the NorthWest Front, a group promoting the establishment of a white nation-State in the U.S Pacific Northwest.[259] A 2020 Tech Transparency Project noted, however, that half of the 221 white supremacist organizations they identified had a Facebook presence.[260] Moreover, groups such as "Stalin Society," which seeks to "defend Stalin and his work on the basis of fact,"[261] continue to be allowed on Facebook, despite the well-documented crimes and millions of associated deaths that Stalin was responsible for. A few years ago, Facebook removed an English-language meme depicting a bruised barbie wearing a headscarf, with the caption "Sharia Barbie: come with hijab, bruises & Quran," for violations of the hate speech policy.[262] In 2018, however, Facebook failed to detect a number of accounts that spewed propaganda targeted at Myanmar's mostly Muslim, Rohingya community, which human rights groups have blamed for fueling violence toward and displacement of this group in Myanmar.[263]

Facebook's own data suggests that the company may have erroneously removed humorous speech under the hate speech policy for years. In November 2022, Meta released its quarterly

[256] Ijeoma Oluo, "Facebook's Complicity in the Silencing of Black Women," Medium, August 2, 2017, https://medium.com/@IjeomaOluo/facebooks-complicity-in-the-silencing-of-black-women-e60c34434181 .
[257] Angwin, ProPublica, and Grasseger, "Facebook's Secret Censorship Rules Protect White Men From Hate Speech But Not Black Children."
[258] *Ibid.*
[259] Kurt Wagner, "Facebook Removes Network of White Supremacist Accounts," *Bloomberg*, March 25, 2020, https://www.bloomberg.com/news/articles/2020-03-25/facebook-removes-network-of-white-supremacist-accounts?sref=VDXBDESF
[260] "White Supremacist Groups Are Thriving on Facebook," *Tech Transparency Project*, May 21, 2020, https://www.techtransparencyproject.org/articles/white-supremacist-groups-are-thriving-on-facebook.
[261] "Stalin Society," Facebook Page, https://www.facebook.com/stalinsociety/ (accessed May 1, 2023).
[262] Caitlin Ring Carlson and Haley Rousselle, "Report and repeat: Investigating Facebook's Hate Speech Removal Process," First Monday, https://firstmonday.org/ojs/index.php/fm/article/view/10288/8327 .
[263] Paul Mozur, "A Genocide Incited on Facebook, With Posts From Myanmar's Military," *New York Times*, October 15, 2018, https://www.nytimes.com/2018/10/15/technology/myanmar-facebook-genocide.html; https://time.com/6217730/myanmar-meta-rohingya-facebook/ .

Community Standards Enforcement report, which showed a 21% quarter-over-quarter decrease in hate speech content actioned on Facebook. The accompanying blog post attributed this decline, from 13.5 million to 10.6 million pieces of content actioned, to improvements "in the accuracy of our AI technology."[264] These improvements allowed the company to identify posts that previously "could have been removed by mistake without appropriate cultural context," such as "humorous terms of endearment used between friends" or "words that may be considered offensive or inappropriate in one context but not another."[265] This explanation suggests that Facebook may have misidentified banter as hate speech for years, potentially causing millions of pieces of content being removed erroneously.

Instagram

The provision in Instagram's Community Guidelines that addresses hate speech begins with the following sentence: "we want to foster a positive, diverse community." Thus, Instagram presumably prohibits hate speech and other forms of objectionable content in an effort to serve this goal. However, as the anecdotes in this section demonstrate, Instagram has previously enforced its hate speech policy in ways that do not align with this objective.

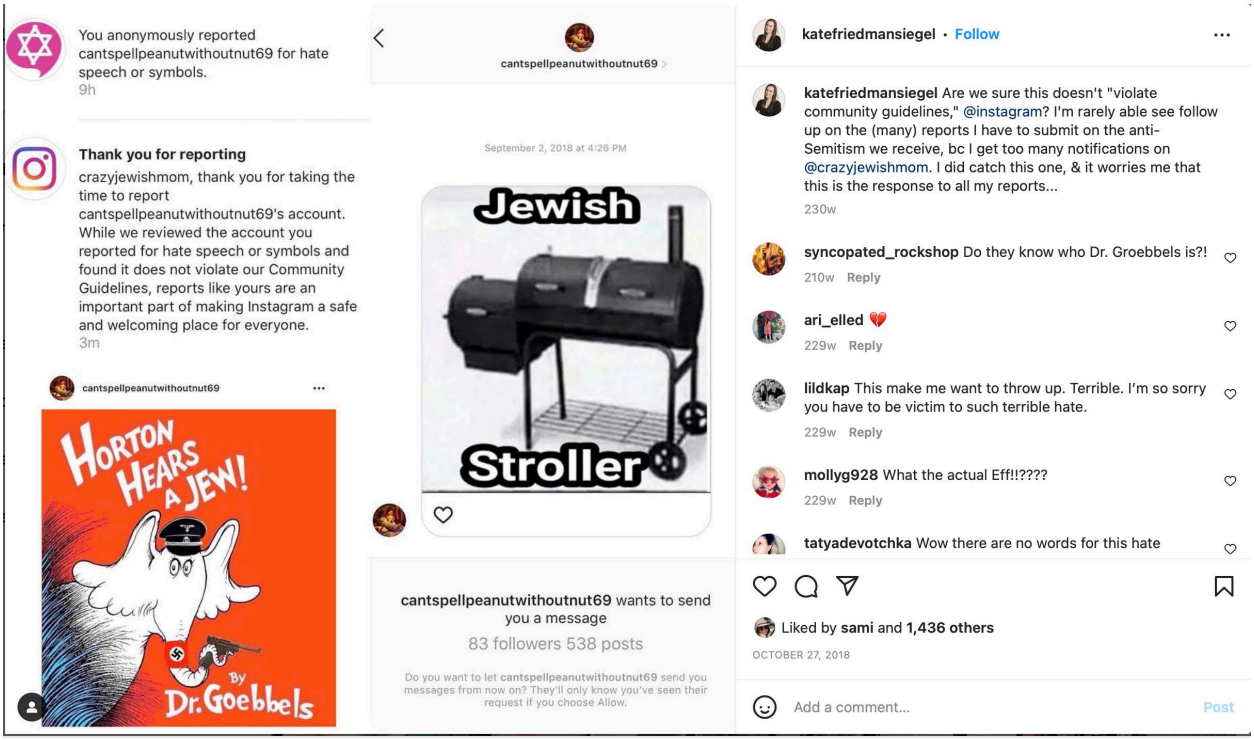
In 2018, for example, Instagram removed posts from *@CrazyJewishMom* that sought to raise awareness about Instagram's struggles to combat online antisemitism. The account, which is run by Kate Friedman-Siegel, mainly posts funny memes and conversations with Friedman-Siegel's Jewish mother. In the wake of the Pittsburgh synagogue shooting, however, Friedman-Siegel shared the post shown in Figure 17 to raise awareness about Instagram's ineffective policing of antisemitic content. The post included an image of a barbecue grill, labeled "Jewish Stroller," and a book cover depicting an elephant wearing a Swastika and holding a gun, designed to look like a Dr. Seuss novel written by Dr. Joseph Goebbels, Hitler's chief propagandist. Friedman-Siegel had previously received the first image in a DM from an account that had posted the second image. She reported the content to Instagram as hate speech, but the company ruled that it did not violate the Community Guidelines.

The caption that accompanied Freidman-Siegel's post clarified the intent behind it. The caption read: "Are we sure this doesn't "violate community guidelines," @instagram? I'm rarely able [to] see follow up on the (many) reports I have to submit on the anti-Semitism we receive, bc I get too many notifications on @crazyjewishmom. I did catch this one, & it worries me that this is the response to all my reports..." However, Instagram removed this post for violating the Community Guidelines and threatened to disable her page, even though Instagram's own Community

[264] Guy Rosen, "Integrity and Transparency Reports, Third Quarter 2022," *Meta Newsroom*, November 22, 2022, https://about.fb.com/news/2022/11/integrity-and-transparency-reports-q3-2022/ .
[265] *Ibid.*

Guidelines noted, "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.". After a surge in comments from her followers about the incident, Instagram reversed its decision.[266]

Figure 17[267]



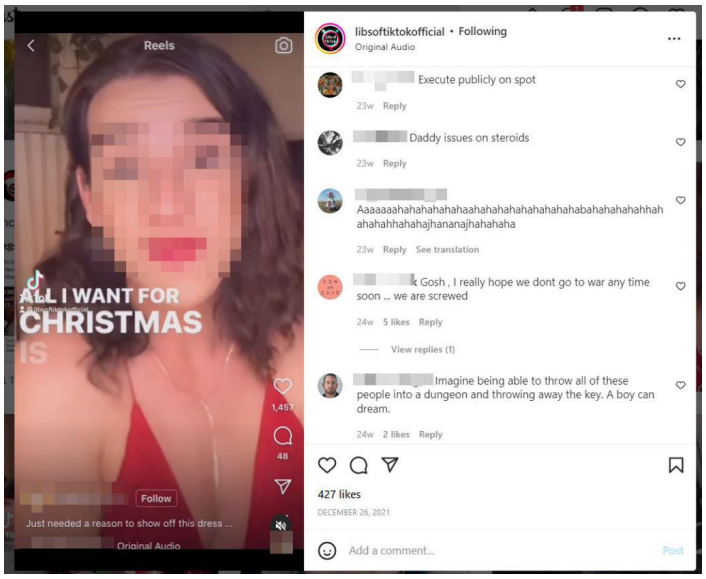
This anecdote is not the only instance of erroneous hate speech removals by Instagram. In September 2022, Business for Social Responsibility released the results of a human rights due diligence assessment of Meta's impacts in Israel and Palestine during the May 2021 conflict. BSR noted that Meta took many appropriate actions during the crisis, "seeking an approach to content removal and visibility based on necessary and proportionate restrictions consistent with the ICCPR's Article 19." [268] At the same time, however, BSR found that Meta over-enforced its policies on Arabic content, erroneously removing Palestinian voices. Moreover, BSR's analysis found that Meta over-enforced its policies on Arabic content more than it did so on Hebrew content, on a per-user basis, and that proactive detection rates for potentially violating Arabic content were

[266] Kate Friedman-Siegel, "I posted on Instagram about my anti-Semitic trolls and their persistent abuse. Instagram deleted my post: OPINION," *ABC News*, October 31, 2018, https://abcnews.go.com/Technology/posted-instagram-anti-semitic-trolls-persistent-abuse-instagram/story?id=58875150.
[267] https://www.instagram.com/p/Bpcq0ktlkcb/
[268] "Human Rights Due Diligence of Meta's Impacts In Israel and Palestine in May 2021," *Business for Social Responsibility,* September 2022, https://www.bsr.org/reports/BSR_Meta_Human_Rights_Israel_Palestine_English.pdf

much higher than for Hebrew content. According to BSR, this discrepancy may be attributed to the fact that Meta had an Arabic classifier for hostile speech but not a Hebrew classifier.

In addition to erroneous removals, there are also several documented instances of Meta failing to remove content that targets protected groups. BSR identified cases where Meta failed to remove violating content during the May 2021 crisis, including incitement to violence against Israelis.[269] BSR also noted that many Jewish organizations tracked antisemitic content that Meta failed to remove from the platform during this period, attributing this under-enforcement to "insufficient cultural competency on the part of content moderators" and "insufficient linguistic capacity in the range of languages (including small European languages) in which antisemitic content has appeared."[270] In July 2022, Media Matters for America accused Instagram of allowing a series of Instagram accounts – and their followers - to spew hate at LGBTQ people.[271] For example, the report shared a post from an account with the handle @garbagehuman4.0 that encouraged anti-LGBTQ+ hate in its comment section. [272] It also included screenshots of comments on a post by Libs of TikTok about transgender people, which said "execute publicly on spot" and "imagine being able to throw all of these people into a dungeon and throwing away the key," shown in Figure 18.[273]

Figure 18[274]



[269] *Ibid.*
[270] *Ibid,* 6.
[271] Camden Carter, "Instagram is allowing accounts to spew hate at LGBTQ people, while also claiming to support the community," *Media Matters for America*, July 6, 2022, https://www.mediamatters.org/facebook/instagram-allowing-accounts-spew-hate-lgbtq-people-while-also-claiming-support-community .
[272] *Ibid.*
[273] *Ibid.*
[274] *Ibid.*

TikTok

While concerns about the Chinese Communist Party's ability to censor TikTok content are dominating headlines today, concerns that TikTok represses minority voices have existed for years. In 2021, for example, Black influencer Ziggi Tyler typed in "Black lives matter" and "supporting Black success" into his bio on TikTok's Creator Marketplace and found that the company flagged the phrases as inappropriate.[275] However, the app allowed him to add "I am antisemitic" and "I am a neo nazi" to his bio. TikTok spokespeople told the MIT Technology Review that the error resulted from "an automatic filter set to block words associated with hate speech," which was "erroneously set to flag phrases without respect to order."[276] This incident fueled debate about racial bias in both TikTok's content moderation and recommendation algorithms.[277]

YouTube

There are also several anecdotes that depict erroneous or inconsistent hate speech enforcement by YouTube. In June 2019, YouTube rolled out several updates to its hate speech policy, including prohibitions on statements of inferiority, harmful conspiracy theories, calls for the subjugation of protected groups, and denials that well-documented events took place.[278] In the wake of the new policy, YouTube banned a few white nationalist accounts – but also removed a few channels set up by a Romanian teacher to provide educational content about Nazi propaganda (see Figure 19.)[279] The teacher, Scott Allsop, explained: "I'm a history teacher, not someone who promotes hatred. I share archive footage and study materials to help students learn about the past."[280] After Buzzfeed published a story about the erroneous removal, YouTube restored the channel.

[275] Ohlheiser, "Welcome to TikTok's endless cycle of censorship and mistakes."
[276] *Ibid.*
[277] Shirin Ghaffary, "How TikTok's hate speech detection tool set off a debate about racial bias on the app," Vox, July 7, 2021, https://www.vox.com/recode/2021/7/7/22566017/tiktok-black-creators-ziggi-tyler-debate-about-black-lives-matter-racial-bias-social-media.
[278] https://web.archive.org/web/20190605213123mp_/https://support.google.com/youtube/answer/2801939?hl=en
[279] Ryan Broderick, "This History Teacher Had His Educational YouTube Channel Banned For Hosting 'Hate Speech,'" BuzzFeed News, June 5, 2019, https://www.buzzfeednews.com/article/ryanhatesthis/history-teacher-scott-allsop-youtube-channel-banned-nazi .
[280] *Ibid.*

Figure 19

A trans creator, Chase Ross, found that YouTube systematically demonetized any videos he uploaded that were labeled with the word transgender.[281] In 2019, LGBTQ YouTubers claimed to have discovered over 900 words that would trigger the algorithm to demonetize minority content, among them "blacks" and "lesbians", "LGBTQ", and "gay".[282]

In 2023, YouTube suspended Michael Knowles, a conservative political commentator for The Daily Wire, after he received his second strike.[283] He received this strike for misgendering Dylan Mulvaney, a transgender individual. According to Knowles, he has been referring to transgender individuals with the pronoun that corresponds to the sex they were assigned at birth for years, but YouTube only started punishing creators for this behavior recently.[284] In fact, YouTube's hate speech policy does not explicitly prohibit misgendering, though the hate speech and harassment policies prohibit attacking or targeting an individual with prolonged or malicious insults based on protected group status, respectively. It is possible YouTube is enforcing that provision when they are punishing misgendering content, though it is not clear. YouTube has also publicly stated that

[281] Chris Stokel-Walker, "Why Has Transgender Become a Trigger Word for YouTube?," *The Daily Beast,* June 2, 2018, https://www.thedailybeast.com/why-has-transgender-become-a-trigger-word-for-youtube.
[282] Beurling & Ocelot AI, "Demonetization List Project," https://docs.google.com/document/d/18B-X77K72PUCNIV3tGonzeNKNkegFLWuLxQ_evhF3AY/edit .
[283] "The Michael Knowles Show – Ep. 1265," June 12, 2023, https://www.dailywire.com/episode/ep-1265-you-tube-suspended-me5.
[284] "The Michael Knowles Show – Ep. 1265," June 12, 2023, https://www.dailywire.com/episode/ep-1265-you-tube-suspended-me5.

misgendering and deadnaming could violate its monetization policies on hateful conduct but, as far as we can tell, has not addressed whether misgendering violates YouTube's content policies on hate speech and harassment directly.[285]

[285] Khadijah Khogeer, "YouTube demonetizes Candace Owens' anti-trans videos saying misgendering may fall under hateful conduct policy," NBC News, June 8, 2023, https://www.nbcnews.com/tech/internet/candace-owens-youtube-gender-pronouns-video-trans-announcement-rcna88175.

Conclusion & Recommendations

This report demonstrates that the scope of platforms' hate speech policies has grown over time, due to expansions in both the type of content and the number of protected characteristics covered by the policies. Ten years ago, less than 13% of the eight platforms we analyzed prohibited statements of inferiority, dehumanizing language, denial of atrocities, or slurs on the basis of protected characteristics, content that is outside the scope of the mandatory prohibitions included in ICCPR Article 20(2) and - in many cases - arguably also outside the scope of the permitted restrictions under Article 19(3). Today, 50% or more of the platforms do. The average number of protected characteristics included in platform hate speech policies has more than doubled since 2010. The lack of international consensus regarding a hate speech definition has likely made expansive platform policies possible, endowing private companies with the power to determine what definition of hate they want to adopt. Theoretically, the fact that most of the analyzed platforms have explicitly committed themselves to using international human rights law as a benchmark should lead them to align their policies with the ICCPR. However, this has arguably been undermined by the divergence between regional and international human rights standards on freedom of expression and hate speech, which permits platforms to "pick and choose" when designing and implementing their Terms of Service and Community Guidelines.

Much of the scope creep documented in this report likely reflects sincere efforts by social media companies to make their products safe for vulnerable communities. Hate speech is anathema to the functioning of society and a danger to values like solidary, equality, and respect. Hate speech can cause harm on a micro (inter-personal), meso (group), and macro (societal) level and is 'deeply rooted in the ideologies of racism, sexism, religious intolerance, xenophobia, and homophobia.'[286] Scholars have also argued that hate speech "initiates, perpetuates and aggravates socially accepted misrepresentation about outgroups."[287] The targets of hate speech may suffer "sadness, pain, distress,"[288] as well as, "humiliation, isolation and dignitary affront."[289] Publicly circulating hate speech may also give its speakers "the legitimacy of a global audience," further entrenching them in hateful beliefs."[290]

Nevertheless, as various examples of erroneous and inconsistent content removals show, platforms' hate speech policies often silence members of the very groups they are designed to

[286] Uladzislau Belavusau, *Freedom of speech: importing European and US constitutional models in transitional democracies,* Routledge, 2013: 41.
[287] Alexander Tsesis, *Destructive messages: How hate speech paves the way for harmful social movements,* Vol. 27. NYU Press, 2002: 138.
[288] Friedrich Kubler, "How Much Freedom for Racist Speech: Transnational Aspects of a Conflict of Human Rights," *Hofstra L. Rev.* 27 (1998): 335.
[289] The link between dignity and hate speech has been made by authors such as Richard L. Abel and Ryo Fujimoto in "Speaking Respect/Respecting Speech," *The Sociology of Law* 1998, no. 50 (1998): 214-234.
[290] LaShel Shaw, "Hate speech in cyberspace: Bitterness without boundaries," *Notre Dame JL Ethics & Pub. Pol'y* 25 (2011): 282.

protect. In fact, in the span of history, restricting viewpoints has overwhelmingly involved the "repression of minority and dissenting voices."[291] Prohibiting certain viewpoints may also have a boomerang effect, rendering "persecuted and criminalized ideas attractive" and thus "mak[ing] perpetrators into victims or martyrs."[292] Finally, consistently and fairly enforcing broad hate speech bans may be impossible, given global social media platforms host content from a plethora of cultures, societies, and religions with "few, if any, shared understandings as to what amounts to intolerable speech".[293]

Thus, the Future of Free Speech believes that neither free expression nor inclusivity can play second fiddle to the other. Instead, they must be treated as mutually reinforcing ideals. Though many platforms' content policy rationales restate this point, the platforms themselves have often enforced their rules in ways that threaten freedom of expression – especially for minorities. Moreover, new efforts at content regulation across the world may worsen these outcomes. Because legislation like the DSA empowers states to pressure private companies into quickly removing content at the risk of fines, it will likely spur further expansions in platform hate speech policies. By becoming even more risk-averse in content moderation, platforms will be able to proactively remove any content that any state might deem hateful at any point – reducing the risk of a scenario where content cannot be removed quickly enough and fines are imposed. As scholars have argued, hate speech bans result in different types of speech becoming "free to the extent compatible with the state's view".[294] Just as a social media platform overrun by threats and poisonous slurs will be un-free for individuals targeted by that content, however, a platform too quick to delete content will seriously limit public debate. Trends towards government legislation regarding content removals thus represent a serious threat to freedom of expression.

There is no simple or straightforward solution to this problem. Competing domestic legal structures surrounding hate speech, varied cultural norms surrounding expression and tolerance, and differences in vulnerable populations from region to region make it challenging to identify the right path forward. Nonetheless, we offer a few alternatives below and discuss their benefits and drawbacks.

First, platforms and governments could tie hate speech moderation to international human rights law, by ensuring their hate speech policies are legitimate, necessary, proportional, and within the boundaries of one of the grounds for restricting speech outlined in Article 19(3) of the ICCPR. A

[291] Heinze, "Hate speech and the normative foundations of regulation," 595.
[292] Veit Bader, "Free speech or non-discrimination as trump? Reflections on contextualised reasonable balancing and its limits," *Journal of ethnic and migration studies* 40, no. 2 (2014): 326.
[293] Uta Kohl, "Islamophobia,'gross offensiveness' and the internet," *Information & Communications Technology Law* 27, no. 1 (2018): 113.
[294] James Weinstein, "An Overview of American Free Speech Doctrine and its Application to Extreme Speech," in Ivan Hare and James Weinstein (eds) *Extreme Speech and Democracy*, Oxford University Press 2009: 82-83.

2018 report on content regulation from David Kaye, the U.N. Special Rapporteur on Freedom of Opinion and Expression, advised social media companies to adopt international human rights law – rather than varying domestic laws or their own private interests - as a framework for content moderation.[295] The report recommended that companies ensure content policies aligned with the standards of legality, necessity, and legitimacy that "bind State regulation of expression."[296] The next year, Kaye's office released a report explicitly focused on hate speech, which directed social media companies to "adopt content policies that tie their hate speech rules directly to international human rights law" and to "define the category of content that they consider to be hate speech with reasoned explanations for users and the public and approaches that are consistent across jurisdictions."[297] The report also instructed States to avoid prohibitions on hate speech, whether offline or online, "except in the gravest situations, such as advocacy of national, racial or religious hatred that constitutes incitement to discrimination, hostility or violence."[298]

These reports underscore that aligning hate speech policies with IHRL not only requires platforms to limit the scope of their rules - but also to clearly define the content these rules cover. Since the publication of these reports, Meta, the parent company of Facebook and Instagram, also established the Oversight Board, which reviews user appeals regarding the company's content moderation decisions. The Board's Charter and Bylaws require it to consider international human rights law standards in this work, and previous decisions have referenced the International Bill of Human Rights and the three-part test of legality, necessity, and legitimacy enshrined in Article 19 of the ICCPR.[299] As documented in this report, many platforms have also added significantly more detail to their hate speech policies, including long lists of examples of the types of content that might be removed according to the rule. Thus, there have been efforts by platforms to align their content moderation practices with IHRL.

Beyond protecting individuals' human rights, adopting this approach to content moderation carries several additional benefits. According to the Special Rapporteur, it would help platform policies more accurately reflect the values of their diverse and global user bases.[300] It would also reduce centralized, private power over expression. Rather than platforms deciding how to define hate speech for the purposes of policing it, platforms would place this power in the hands of existing bodies of international law. Using Articles 19 and 20(2) of the ICCPR as a basis for platform

[295] Special Rapporteur on freedom of opinion and expression, "Report on content regulation (A/HRC/38/35)."
[296] *Ibid*, [45].
[297] Special Rapporteur on freedom of opinion and expression, "A/74/486: Report on online hate speech."
[298] *Ibid*.
[299] "The Oversight Board: Operationalizing the UN Guiding Principles on Business and Human Rights, Submission to the Office of the High Commissioner for Human Rights, United Nations, on the practical application of the UNGPs to the activities of technology companies," *The Oversight Board*, February 2022, https://www.ohchr.org/sites/default/files/2022-03/Oversight-Board.pdf
[300] Special Rapporteur on freedom of opinion and expression, "Report on content regulation (A/HRC/38/35)," [41].

hate speech policies would also help cultivate a more transparent, legitimate, and speech-protective approach to handling online hate speech.

There are also challenges with this moderation approach, however. In the US, the prohibitions on hate speech outlined in the ICCPR cover speech that is protected by the First Amendment. Any requirement on platforms to adopt policies that align with Article 20 (2) would not align with the legal system in the United States. Second, this approach would not eliminate instances of erroneous and unequal hate speech policy enforcement. While the reduction in scope implied by this approach would likely streamline the process of hate speech moderation, automated moderation systems would not necessarily become more attuned to local or contextual nuance or unbiased.

Decentralized content moderation is an alternative or complementary path forward. There are a variety of existing proposals in this vein, but they all generally involve mandating platforms to allow third-party content moderation systems. The organization *Article 19* argues regulators should mandate platforms to unbundle "their hosting and content-curation functions" and to allow third parties to provide content curation to users.[301] Mike Masnick, the founder and editor of *Techdirt*, advocates for a social media landscape dominated by protocols rather than platforms, the same way that email services are interoperable with one another. "A user can use a non-Gmail email address within the Gmail interface," Masnick explains, or "use a Gmail account with an entirely different client, such as Microsoft Outlook or Apple Mail," and, "on top of that, it's possible to create new interfaces on top of Gmail itself, such as with a Chrome extension."[302] The Working Group on Platform Scale, convened by Stanford University's Program on Democracy and the Internet, advocated for a "middleware solution," where a "combination of regulation and new technology" would enable platforms to outsource content curation to third-party organizations.[303] This competitive content-curation layer would then enable users to "tailor their feeds to their own explicit preferences."[304] Platforms could also explore devolving content removal decisions to community leaders, the way Reddit does for sub-reddits and Facebook has done in certain cases.

Decentralized content moderation offers several benefits. First, different individuals have different values, leading to different views about what constitutes speech and different levels of tolerance. Different societies have varied histories and different vulnerable populations, suggesting that

[301] "Taming Big Tech: A pro-competitive solution to protect free expression," *Article 19*, 2021, https://www.article19.org/wp-content/uploads/2023/02/Taming-big-tech-UPDATE-Jan2023-P05.pdf.
[302] Mike Masnick, "Protocols Not Platforms: A Technological Approach to Free Speech," *Knight First Amendment Institute*, August 21, 2019, https://knightcolumbia.org/content/protocols-not-platforms-a-technological-approach-to-free-speech .
[303] Francis Fukuyama, Barak Richman, Ashish Goel, Marietje Schaake, Roberta R. Katz, Douglas Melamed, "Report of the Working Group on Platform Scale," *Stanford Program on Democracy and the Internet and Stanford Center on Philanthropy and Civil Society*, November 2020, https://cyber.fsi.stanford.edu/publication/report-working-group-platform-scale
[304] *Ibid.*

speech that is dangerous in one area may not be in another. Decentralization would allow users and communities to choose their own content moderation experience, based on their individual or communal values and histories. Second, decentralization would allow any one individual to avoid speech they find problematic or offensive – without censoring anyone completely, avoiding the backfire problem. Third, platforms would no longer have the power to decide who is allowed to speak and what they are allowed to speak about. Thus, decentralization would eliminate the paternalism and power concentration of current moderation frameworks.

Of course, there are also issues with this approach. There may be barriers to switching to new, decentralized services. This reality was highlighted by the network and technical challenges that many Twitter users faced when trying to switch to Mastodon, a decentralized social network, after Elon Musk's takeover of Twitter. However, regulations requiring all platforms to unbundle hosting and curation services, and to allow third party filtering systems, would address this problem. Decentralized moderation could also put targeted individuals or communities at risk, even if they are able to avoid seeing the offensive content. Dave Willner, the Head of Trust & Safety at OpenAI, has said: "The worse kinds of content are the things that others see about use (whether as individuals or members of a community) which change how they view us or act towards us. That's why the most violent hate speech is so dangerous – it's not just mean, it inspires action. Each of us individually filtering what we see does absolutely nothing to deal with those kinds of harms."[305] This problem could be partially solved by requiring middleware systems to adhere to the standards of international human rights law, which prohibits national, racial, and religious hate speech that constitutes incitement to violence. However, as explained above, such a mandatory requirement would risk running afoul of the First Amendment.

The Future of Free Speech acknowledges that neither decentralization nor the IHRL approach will resolve all the challenges associated with content moderation – and that both options carry their own challenges. Nevertheless, these alternatives provide a path forward that is more aligned with global norms of free expression and more transparent than the status quo. It is our hope that this report can contribute to the necessary debate, innovation, and policy development needed for such development.

[305] Tweet from Dave Willner (@dswillner), April 17, 2022, https://twitter.com/dswillner/status/1515854723923918849?s=20 .