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

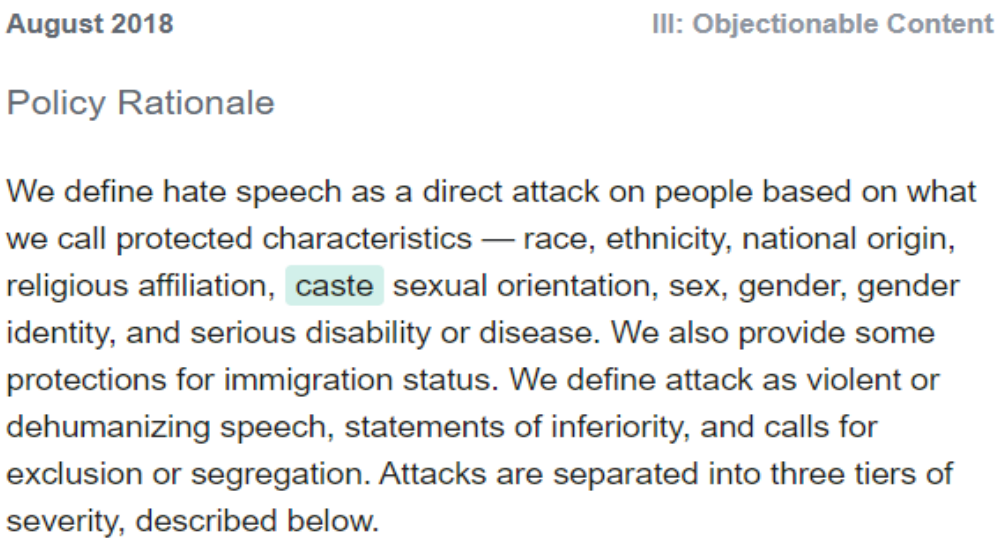
The next notable change in Facebook's hate speech provisions came in August 2018, when the company first defined "attack" in the context of its hate speech prohibition (see Figure 6). "We define attack as violent or dehumanizing speech, statements of inferiority, and calls for exclusion or segregation," 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.

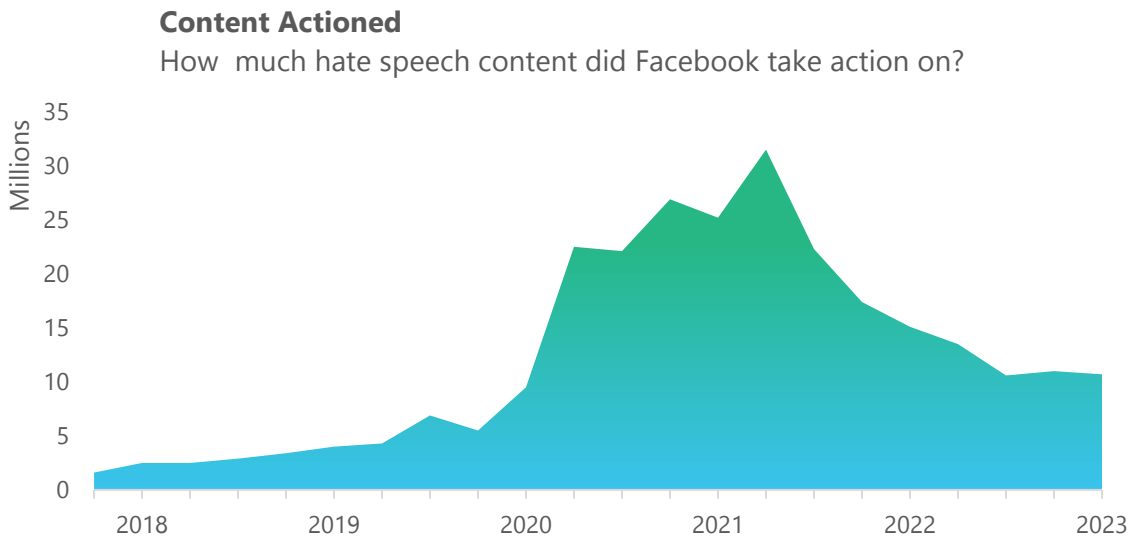
Changes in Enforcement Volume

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/ .