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:

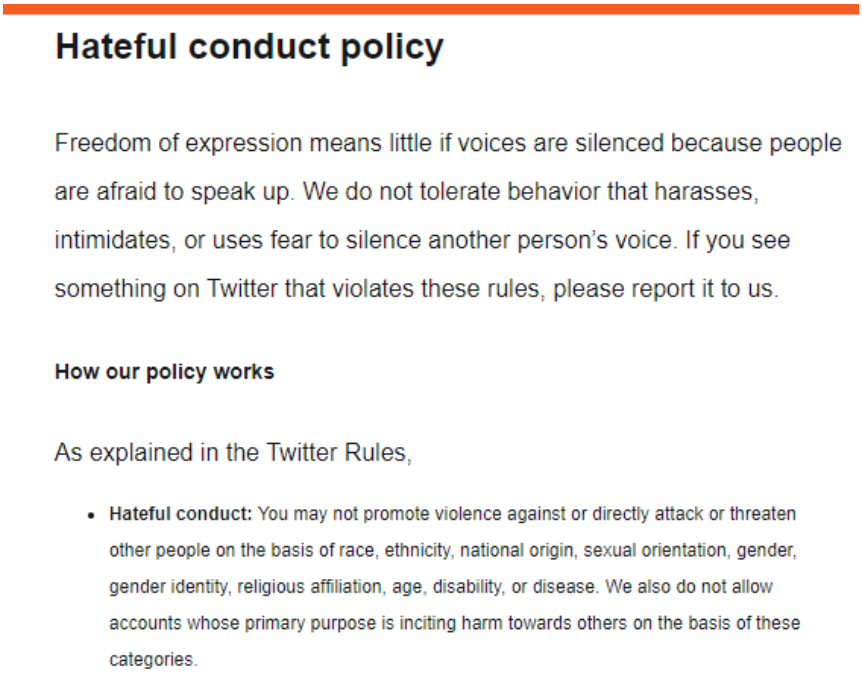
- "Hateful conduct: You may not promote violence against, threaten, or harass other people on the basis of race, ethnicity, national origin, sexual orientation, gender, gender identity, religious affiliation, age, disability, or serious disease.
- Hateful imagery and display names: You may not use hateful 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,*