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

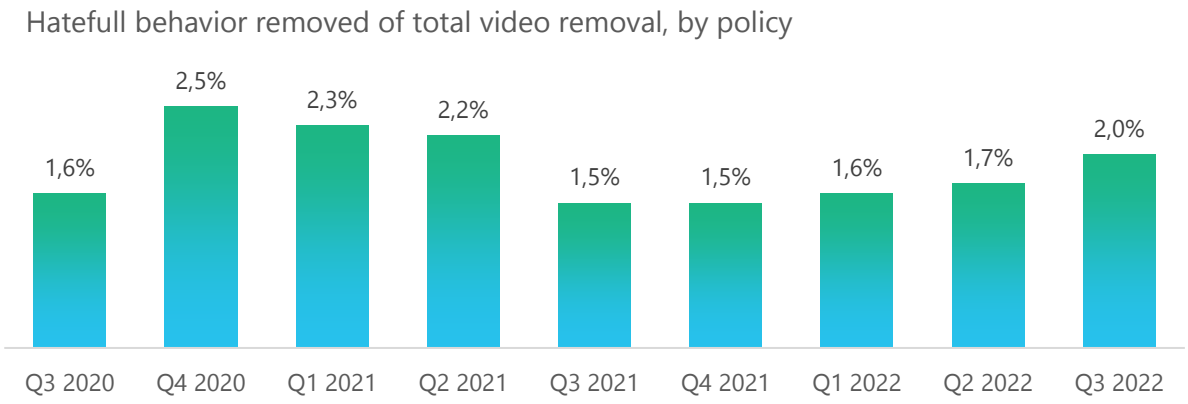
Notes: An X indicates the company's hate speech policies covered that protected characteristic for at least one month during the given year.

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/ .