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.