▶ YouTube

How YouTube protects our community from hate and harassment

At YouTube, we want to give everyone a voice. Hate speech and harassment discourages people from sharing their stories and experiences with the world, which is why we have developed policies that don't allow it on YouTube. As with our approach to other harmful content, we work quickly to remove content that violates our hate and harassment policies.

Who do YouTube's policies protect?

We distinguish hate speech and harassment by the intended target of malicious content.

Hate speech policies protect groups
Our hate speech policy restricts any content that incites violence or provokes hatred against groups defined by certain protected attributes.

Harassment policies protects individuals
Our harassment policy restricts harmful threats or insults targeting identifiable individuals based on intrinsic attributes, including their protected attributes or physical traits.

Protected attributes
We've based our definition of these groups on existing laws, as well as in consultation with experts in areas like civil rights and hate speech.

- ✓ Race
- ✓ Ethnicity
- ✓ Nationality
- ✓ Religion
- ✓ Disability
- ✓ Sex and gender
- ✓ Sexual orientation
- ✓ Immigration status
- ✓ Age
- ✓ Caste
- ✓ Veteran status
- ✓ Status as a victim of a major violent event (or their families)

What do YouTube's policies protect against?

Below are examples of what is not permitted on YouTube; reference our full <u>hate speech</u> and <u>harassment</u> policies online.

Hate speech
- ✕ Conspiracy theories that have been used to justify real-world violence
- ✕ Dehumanization
- ✕ Inciting violence
- ✕ Inferiorization
- ✕ Supremacism
- ✕ Threats
- ✕ Violent event denial

Harassment
- ✕ Doxxing
- ✕ Simulations of violence
- ✕ Stalking
- ✕ Threats
- ✕ Unwanted sexualization
- ✕ Violent event victim harassment
- ✕ Wishes for death
- ✕ Dehumanization
- ✕ Egregious insults
- ✕ Focused insults
- ✕ Targeting vulnerable individuals

How do we keep up with evolving trends in hate speech and harassment?

As with all language, expressions of hate speech and harassment are constantly changing. We take steps to ensure that our policies evolve to stay ahead of new trends, coded language, and targets of discrimination.

- **Intelligence Desk** is an internal team monitoring news sources, social media, and user reports
- **We collaborate with external experts** in areas like violent extremism, supremacism, civil rights, cyberbullying, and free speech
- **We consult with a wide array of creators** on an ongoing basis about their experiences to inform policy development and updates

How does YouTube enforce its policies?

Our community and machine learning systems help us detect hate speech and harassment

With more than 500 hours of video uploaded to YouTube every minute, we empower viewers and creators to take action, as well as invest in machine learning systems to help detect problematic content at scale.

Trusted Flagger Program for government partners and NGOs that have high accuracy rates in reporting hate speech and harassment

Creators can block certain words or individuals in the comments section with comment moderation tools

Viewers can flag videos and comments

Machine learning enables us to detect violative content at scale. Our systems are trained to identify harmful content across the platform and around the clock, meaning we're able to review hundreds of thousands of videos in a fraction of the time it would take a person to do the same

Content reviewers evaluate flagged content

Flagged content is not automatically removed. Our team of expert reviewers across the world decide what to remove based on our hate speech and harassment policies. Their decisions are used to train and improve our machine learning systems.

Fluent in dozens of languages all over the world and carefully evaluate flags

Training and judgement to understand nuanced meaning, in terms of context or language

Enforce our guidelines consistently, whatever the creator's background, political viewpoint, position, or affiliation

Remove content that violates policies

We penalize creators for posting hate and harassment content

1 **Notify creator** to let them know what we've done, and why we've done it

2 **Give a warning** for the first Community Guidelines violation

3 **Give a strike** for subsequent violations, along with no posting privileges for at least a week

4 **Terminate channels** if they get three strikes in 90 days, or more quickly for extreme offenses

Policy exceptions

We may make exceptions to our policies for content with a compelling educational, documentary, scientific, or artistic purpose, called "EDSA" content. For example, we might allow comedy or satire that uses slurs or other offensive language if that context is clear to the viewer.

Reduce spread of borderline content

Some content comes close to violating our hate speech and harassment policies – but doesn't quite cross the line. We reduce the spread of this borderline content by limiting recommendations on a viewer's homepage and "Up next" list.

What progress has YouTube made in removing hate speech and harassment?

With the right technology, expertise, and policy in place, we're removing videos, channels, and comments containing hate speech and harassment faster than ever before.

Hate speech*
97K+ videos removed
46M+ comments removed

Harassment*
77K+ videos removed
136M+ comments removed

77%
of videos removed for violating policies had less than 10 views*

youtube.com/howyoutubeworks

*Between Oct 2020 - Dec 2020