twitch.json

hatespeech

Page ID: 21

Source: https://help.twitch.tv/s/article/how-to-use-automod?language=en\_US

['discrimin']

On this page: How AutoMod Works How to Configure AutoMod Blocked and Permitted Terms and Phrases Frequently Asked Questions How AutoMod Works AutoMod is a moderation tool that blocks inappropriate or harassing chat with powerful moderator control.

AutoMod detects misspellings and evasive language automatically.

When a user sends a message that AutoMod flags as potentially inappropriate, the message is held and shown to moderators to Allow or Deny .

If allowed, the message will be sent to chat.

Broadcasters can configure AutoMod by selecting one of four levels.

These levels affect how aggressively AutoMod holds back messages under four categories: discrimination, sexual content, hostility, and profanity.

These settings can only be changed by the broadcaster, while the dictionary each setting contains is managed by Twitch.

AutoMod does not timeout, ban, or mute users from any channel ; it merely withholds comments that fall within the broadcaster's chosen moderation settings to be either approved or denied by the broadcaster's moderators.

Users with messages held by AutoMod will receive a notification that their message is being held for review, as well as a notification if the message was approved or denied.

If approved, the message will instantly appear in chat as the next message.

Broadcasters and their moderators will also still see all the same timeout and ban chatbox options on those messages, so even if the message was denied, there will still be the ability to timeout or ban the user.

How to Configure AutoMod We recommend you discuss with or inform your moderator team what settings you have used and perhaps which words or phrases you care most about filtering, in order to ensure they adhere to your preferences when they are prompted by AutoMod to review chat messages for approval or denial.

['discrimin', 'hate']

On this page, click Turn On AutoMod .

You will be defaulted to Level 1 , the lowest active setting.

Drag the slider bar to choose the setting most appropriate for you, from Level 0 (no filtering) to Level 4 (the strongest default setting).

To tailor AutoMod to your own custom settings, you can use the drop down menus to update each moderation category individually.

You can see how strictly AutoMod will filter based on how many shields are displayed in each moderation category, from zero (no moderation) to four (a lot of moderation).

AutoMod Moderation Categories Category Description Discrimination Race, religion, gender, etc.

Hate speech falls under this category.

Sexual Content Words or phrases referring to sexual acts and/or anatomy.

Hostility Provocation and bullying.

Profanity Expletives, curse words, and vulgarity.

This filter especially helps those who wish to keep their community family-friendly.

What AutoMod catches at each level Level Description Level 0

No filtering.

Words and terms you have added to your channel's blocked terms will still be blocked even if AutoMod is turned off.

This can prevent messages from being sent.

Level 1 Some filtering on discrimination only.

Level 2 Some filtering on discrimination and sexual content, more filtering on hostility.

Level 3 More filtering on discrimination, sexual content and hostility.

Level 4 More filtering on discrimination, sexual content and profanity, and most filtering on hostility.

Hate speech, bullying, and sexually explicit content is not allowed under Twitch’s community guidelines .

If you see it, please report it .

AutoMod and Channel Moderators Channel moderators can also access and update AutoMod settings by opening the Chat Settings in the channel via the settings cog on the bottom right hand corner.

From there, select Manage Moderation Settings .

This will open the moderation preferences page for the channel, where they can update the AutoMod level, create a custom level, and update terms in the blocked and permitted terms list.

Blocked and Permitted Terms and Phrases

In addition to AutoMod's machine learning and language processing, you can manually list terms to block or permit in your chat.

['hate']

The number of terms you can filter is in the tens of thousands, so you do not need to worry about reaching the cap.

You can add global and sub emotes to your blocked terms.

To block an emote from your channel, add the exact emote name to your blocked terms.

For example, adding Kappa to your blocked terms would prevent the Kappa emote from being posted in your chat.

When adding blocked terms, you can also use wildcards by adding a \* to your term.

For example, adding hate\* to your blocked terms will also cause haters and hateful to be blocked.

You can also use this feature to filter any variant of a url.

For example, someurl.com\* would filter http://someurl.com , http://someurl.com/somepage , http://www.someurl.com/some/other/page?abcdef , etc.

This can protect your chat from variations of links that are unwelcome.

If you're using AutoMod, denying a word or phrase held by AutoMod in chat will temporarily add that term to your Blocked Terms list.

Repeatedly denying a term will extend the amount of time it is blocked: from one hour, to a day, to seven days, to permanently.

You can remove an automatically added term from your blocked list at any time.