twitch.json

hatespeech

Page ID: 25

Source: https://safety.twitch.tv/s/article/Combating-Targeted-Attacks

['hate']

Preferred Language twitch.tv ↗ Combating Targeted Attacks Harassment of any kind, whether in the form of hate raids, malicious spam, or other targeted attacks, is against our Community Guidelines and counter to our community values.

We are continuing to build new features to prevent the harm caused by malicious spam, botting, and raiding, but in the meantime we wanted to highlight the tools that are currently available to help combat this kind of behavior.

We know many Creators are already using these strategies, and have been sharing tactics and tools with each other, and we want to ensure this information is readily available for others who may need it.

These settings are available via the Moderation Settings in Stream Manager.

Attack Prevention Strategies Appoint Moderators.

If you don’t already have a moderator, consider finding someone within your community to support your channel.

Moderators are the first line of defense against bad actors, and they play a number of other important roles such as welcoming and engaging with viewers, editing stream elements, managing Channel Points rewards, and more.

['hate', 'abus']

Turn on Followers-only Mode.

This will give you time to identify bots when they join the channel, and determine whether or not to ban them before they can chat.

The shorter the follow age you select, the easier it will be for bad actors to “wait out” the challenge.

You can always update the follow age based on the specific situation or if you prefer setting an even higher bar, try Sub-only mode.

Turn on AutoMod (Level 2 or Higher Recommended).

Automod uses machine learning to automatically hold messages that may be hateful or harassing.

The higher the level you select, the more types of messages will be held for review.

You can also adjust the Automod level for the specific type of language you want to prevent.

Enable Email and/or Phone Verification.

This will require select chatters to have a verified email and/or mobile phone number associated with their account before being allowed to chat.

You can require this for all chatters, first-time chatters, or based on account or follow age.

You can also exempt VIPs, moderators, and subscribers from verifying.

Manage Incoming Raids Settings.

While not all malicious attacks use the Raids product, you can protect yourself from potential abuse by managing your raids settings .

You can manage incoming raids while live by temporarily denying all incoming raids for up to an hour.

To do so, you will need to add the Stop Raids for 1 Hour quick action to your stream manager and enable raids in your raids settings.

In addition, you can manage who can raid your channel with viewer count, account age, and other customizable settings.

Actions to Take if You're Targeted (or Supporting Someone Who Is) Clear Chat.

If something malicious has been said in chat by several people or bots in a short span of time, the fastest way to remove all comments is to clear the chat.

To use this command, type and send /clear

['hate', 'abus']

That doesn’t mean that there aren’t good options out there, only that we don’t have the access or information we’d need to endorse their use.

We are confident however in the tools we offer, so we choose to focus on educating and promoting their use and how they can make your Twitch experience safer.

And of course, we’re always building new tools to add to what’s already available.

Why does it take so long to build new safety tools?

It takes time and rigorous testing to develop tools that are effective sitewide, and can scale globally as long term solutions across devices and regions – especially when bad actors are working around the clock to circumvent them.

Unfortunately, there is no single solution to most safety-related issues, particularly when it comes to hate raid attacks from malicious and highly-motivated actors.

When we launch a new safety tool it is essential that it work across devices at scale on Day 1, which requires thoughtful and thorough planning, testing, and deployment from a highly specialized team of experts who are working on this every day.

Even once a tool is launched, the work is never over—our team is always monitoring and performing quality checks on existing tools to ensure they’re functioning optimally, and catching abusive behaviors as intended.

Getting it right takes time.

Why don’t you enable Creators and their moderators to ban IP addresses or device IDs of known harassers?

We’ve found that IP bans are not always effective solutions to ban evasion because they’re imprecise and extremely easy for bad actors to circumvent.

It's easy for bot accounts to rotate IP addresses to evade detection, and large but legitimate organizations, like dorms for example, often share the same IP address, leading to a lot of false positive bans.

We face a similar problem with device IDs – there are ways for bad actors to circumvent them that limit their effectiveness, particularly against automated, targeted attacks.

Why don’t you limit the number of accounts users can make with one email address?

We actually do limit the number of accounts users can make through one email address to prevent mass botting or other abusive behavior.

However, we don’t limit it to just one account, as many users of our site have more than one account attached to their email for very legitimate reasons—such as developers who have made helpful chatbots for their communities.

There are still some ways for persistent bad-actors to circumvent our limits, and we’re constantly working hard to improve how we proactively detect and take action against them.

Why can’t you share more about what’s in your roadmap?

Malicious actors are agile and highly motivated, and they put a lot of time and energy into evading the safety hurdles that we’re constantly building and improving.

We know it’s frustrating when we can’t share details on upcoming safety tools, but the more information we offer about what we're doing to stop them–particularly before we've launched–the easier it is for them to navigate around them or prepare.

That said, safety is a vitally important topic that is foundational to every Creator’s experience on Twitch, and the community deserves to feel confident that we are working on this and feel actively informed on our progress.