twitch.json

misinformation

Page ID: 8

Source: https://safety.twitch.tv/s/article/H1-2022-NetzDG-Transparency-Report?language=en\_US

['spam']

Content moderation solutions that work for uploaded, video-based services do not work, or work differently, on Twitch.

Through experimentation and investment, we have learned that for Twitch user safety is best protected, and most scalable, when we employ a range of tools and processes, and when we partner with, and empower, our community members.

The result is a layered approach to safety—one that combines the efforts of both Twitch (through tooling and staffing) and members of the community, working together.

For more information regarding our layered approach to safety, please see the full Transparency Report , which includes a number of safety updates we have made recently.

1) Terms of Service and Community Guidelines Our safety approach starts with our Terms of Service and Community Guidelines, which jointly provide the framework for the relationship between Twitch and its users and define what expression and behavior are allowed on the service and what is not, which in many cases aligns with what is unlawful under German law.

Our Community Guidelines prohibit a wide range of behaviors, including incitement to violence, hateful conduct, harassment, spam, and sexually explicit content.

These guidelines are developed by a dedicated team, in consultation with the Twitch Community, NGOs, and academic partners.

Alongside these Community Guidelines, Twitch’s Terms of Service also governs permitted content, which in many cases also aligns with what is lawful under German law.

In our Terms of Service, sections 8 (User Content), 9 (Prohibited Conduct) and 14 (Termination) govern user-generated content, its acceptability, and the consequences of posting unacceptable user generated content.

Under section 8, the user warrants that they will not post certain types of content on Twitch that are unacceptable for the Twitch community, including content that infringes on third-party rights or that is defamatory.

While section 8 emphasizes that users are free to share their political opinions and participate in political activities, it obligates the users to comply with all applicable laws and regulations while doing so.

Under section 9, the user agrees, among other limitations, not to violate any law, contract, intellectual property, or other third-party right and that their user-generated content will not be inaccurate, unlawful, infringing, defamatory, obscene, pornographic, invasive of privacy or publicity rights, harassing, threatening, abusive, inflammatory, or otherwise objectionable.

['spam', 'trust']

2) Community-led Content Moderation Creators are expected to uphold the service-wide Community Guidelines in their channels, and are invited to adopt stricter standards for their channels if they choose.

We provide creators with tools to set, communicate, and enforce the standards of behavior in their channel (more info below).

We also provide viewer-level controls that enable viewers to control the content they see.

At the same time, Twitch applies various technologies to proactively detect and remove certain kinds of harmful content before users ever encounter it.

Twitch supports creators in building communities that are unique and personal, with the expectation that those communities are healthy and abide by the Twitch Community Guidelines.

To accomplish this, many Twitch creators ask trusted members of their communities to help moderate chat in the creator’s channel.

These channel moderators (“mods”) and moderation tools are the foundation of chat moderation in every creator’s Twitch channel.

To make this model work, we invest heavily to provide our creators and their mods with tools that are flexible and powerful enough to enforce both Twitch’s Community Guidelines and any stricter, channel-specific standards established by the creator.

Our suite of moderation tools supports two objectives: identifying potentially harmful content for moderator review, and scaling moderator controls to support fast-moving Twitch chat messages.

Creators and their mods can use tools—provided by Twitch—to manage who can chat in their channel and what content can be seen in chat.

To manage who is actively participating in their community, creators and their mods can remove bad actors from chat by issuing temporary and permanent bans.

These bans delete a chatter’s recent messages from the channel, and prevent them from sending further messages in the channel during the time they are suspended.

Twitch provides them with additional information to help guide decisions through Suspicious User Controls , a customizable setting that uses machine learning (see below for further information) to flag accounts that may have been created by a user who is trying to evade a previous channel-level ban through various account signals.

Over time, the tool will learn from the actions taken by the creators and mods and the accuracy of its predictions will continue to improve.

Creators and mods can also change certain settings to restrict who can chat to more trusted groups such as followers, subscribers, or those who have verified their account with an email or phone number.

Mods can send their own chat messages, which carry their green Moderator badge, to guide the tone of the chat.

To control what messages can be seen in chat, creators and mods utilize two core features: AutoMod and Blocked Terms.

When enabled, AutoMod pre-screens chat messages and holds messages that contain content detected as risky, preventing them from being visible on the channel unless they are approved by a mod.

Blocked Terms allow creators to further tailor AutoMod by adding custom terms or phrases that will be always blocked in their channel.

These features are best utilized through Mod View, a customizable channel interface that provides mods with a toolbelt of ‘widgets’ for moderation tasks like reviewing messages held by AutoMod, keeping tabs on actions taken by other mods on their team, changing moderation settings, and more.

It’s important to remember that actions taken by a creator and their moderator(s) can only affect a user’s access in that channel.

Channel bans, time-outs, and chat deletion only apply within a channel, and do not affect the user’s access to other channels or other parts of the Twitch service.

However, creators and moderators (or any Twitch user) can report conduct that violates our Community Guidelines through the Twitch reporting tool, which can then be actioned on a service-wide basis by Twitch moderation staff.

3) Proactive Content Moderation

In addition to our community-led moderation efforts, Twitch applies various technologies to proactively detect and remove certain kinds of harmful content before users ever encounter it.

We have machine learning models that we are continually updating to address offensive usernames, spam and scams, offensive emotes, and bot accounts, including globally banned words that disallow the usage of some words or phrases across Twitch.

We make sure to collect training data from a broad sampling of expert labeled data (either internal data labeled by our policy and Trust & Safety teams or by external vendors using criteria provided by our policy team) for quality assurance.

For models that have a higher risk of bias given the content type, such as offensive usernames, we have worked on specific bias monitoring and mitigation work.

We also perform quality assurance using model performance methods which are assessed via a "dark launch" period prior to full rollout (for example, this model actively collects data but does not act on it).

This allows us to analyze and test the results.

Once live, there are weekly team reviews of model performance where the metrics are reviewed by the team.

In terms of outside evaluation, we conduct internal research and have participated in third party assessments.

Our automated system also scans livestreams for content that is particularly objectionable and potentially harmful to our communities, including extreme violence, gore, and pornography.