twitch.json

misinformation

Page ID: 9

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

['trust']

Our Community Guidelines (CGs) lay a sitewide foundation for safety and prohibit harmful behaviors that violate our values as a global community.

Streamers and their volunteer moderators use customizable Twitch safety tools to ensure these standards are met within their channel, while further tailoring the versatile settings we provide according to their stream’s unique needs.

Viewers are also empowered to report users or behaviors that break our CGs, as well as set additional controls for the types of content they want to see.

Our global Safety Operations team works in tandem, 24/7/365, to respond swiftly to these user reports and help mitigate further harm.

Behind the scenes, we are always updating proactive detection filters as an important first line of defense.

That said, we prioritize having a human in the loop for every one of these aspects of safety at Twitch, in order to better stay ahead of bad actors and ensure Trust & Safety processes stay accurate and fair for our community members.

We call this safety infrastructure our “layered approach”.

We will discuss each of these pieces (from the bottom of the pyramid to the top) and how they fit together in the following section.

Community Guidelines Twitch’s Community Guidelines are the foundation of our safety ecosystem.

These guidelines set the guardrails for all user generated content and activity on the service.

Because the Community Guidelines communicate the expectations for behavior on Twitch, clarity is important.

Over time we have increased clarity by adding descriptions and specific examples of prohibited behavior and content (and specific exceptions) wherever possible.

['trust']

Over the last two years, we have implemented technologies that scan content on the service and then flag it for review by human specialists (we call this “machine detection” or “proactive detection”).

Examples of this type of content are nudity, sexual content, gore, and extreme violence.

Twitch is predominantly a live-streaming service, and most of the content that is streamed is not recorded or uploaded.

Because it’s ephemeral, live-streaming provides a more challenging environment for machine detection than other content types do.

Nevertheless, we have found ways to make machine detection viable and useful on Twitch, and we will continue to invest in these technologies to improve them.

User Reporting: Community reports are a crucial part of maintaining the safety and trust of our community and upholding our Community Guidelines.

We believe user reporting is particularly effective on Twitch because the vast majority of the content on Twitch — video and chat — is public.

We encourage streamers, human moderators, and viewers to report content that violates our Community Guidelines so we can take appropriate action on a service-wide basis.

User reports are sent to our team of content moderation professionals to review.

Review and Enforcement: Our highly trained and experienced professionals who review user reports and content that is flagged by our machine detection tools.

These content moderation professionals work across multiple locations, and support over 25 languages, in order to provide 24/7/365 capacity to review reports as they come in across the globe.

Reports are prioritized so that the most harmful behavior can be dealt with most quickly.

['disinfo', 'misinfo']

In addition to the tools that we provide streamers and their mods, we also want viewers to be able to customize the safety of their experience.

To enable that, we provide viewers with features — such as mature flags, chat filters, and blocking other users — that they can use to customize content they encounter and interactions they have across the service.

H1 2022 Safety Updates Community Guidelines: In the first half of 2022 we made revisions to our Community Guidelines, aimed at reducing the risk of severe harm to our community and clarifying existing policies.

Usernames : Usernames that are hateful or harassing have always been prohibited, but to help make Twitch a more welcoming place for our diverse global audience, we broke them out into a designated policy and additionally prohibited names that reference hard drugs and sexual content.

We also launched a new machine learning model to catch violative usernames and automatically enforce against them or prompt them to reset, depending on the severity of the violation (Announced Feb 2022, implemented March 2022)

Harmful Misinformation Actors : We now prohibit on Twitch those whose online presence is dedicated to persistently sharing widely-disproven harmful misinformation topics, such as conspiracies that promote violence.

We defined these criteria and took this actor-focused approach in collaboration with third-party experts including the Global Disinformation Index, who advised that because misinformation doesn’t go viral on Twitch, the greatest opportunity for harmful disinformation would be if disinformation actors were able to form and grow disinformation communities on the site.

(Announced March 2022, implemented March 2022)

Sexual Content : As part of our commitment to improving the clarity of our Community Guidelines, we updated the language of our sexual content policy and added more examples to more clearly explain our definition of sexually explicit content.

We also clarified what we consider to be “sexually explicit” content.

While sexual content has always been prohibited on Twitch, its categorization under “Sexual Violence and Exploitation” didn’t feel appropriate or accurate, as there is a distinct difference between content that is sexually explicit and content that is violent or exploitative.

(Announced April 2022, implemented April 2022) Self-Harm : We updated our guidelines regarding discussions of sensitive topics, such as self harm, to ensure that they do not glorify or encourage self-destructive behavior, and do not lead to additional harm.

We added more examples to illustrate what Twitch defines as self harm, and explicitly called out the glorification of eating disorders as a behavior we prohibit.

['trust']

In the hours after the shooting, we monitored our service to prevent any re-streams of that video or any related content, and we collaborated with other tech companies through the Global Internet Forum to Counter Terrorism to share relevant information and limit the spread of this footage online.

Since the event, Twitch has been working closely with several law enforcement agencies (such as the FBI, Department of Homeland Security, and NYPD Cyber Intelligence Unit) to ensure they have access to any and all information that will aid in the investigation and prosecution of the crime.

Twitch has a zero-tolerance policy against hateful conduct and violence of any kind, and we condemn it strongly .

Moderation in Channels: Coverage, Removals, and Enforcements Overview On Twitch, we empower streamers to build communities that are unique and personal.

That empowerment is paired with the expectation that those communities must be healthy and abide by the Twitch Community Guidelines.

Streamers are ultimately responsible for their streams and the communities they build, but they often enlist the help of trusted community members to help them uphold the norms and expectations they set.

These channel moderators (“mods”) perform a number of roles from welcoming new viewers, to answering questions, managing stream elements, and moderating chat.

Moderators and the suite of moderation tools they employ are essential to the health of chat in every streamer’s Twitch channel.

To make this model work, we invest heavily to provide tools that are flexible and powerful enough to enforce those community standards within their channel.

We focus our moderator support tooling in two main areas: identifying potentially harmful content for moderator review, and scaling moderator controls to support fast-moving Twitch chat.

Streamers and their mods can use tooling 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, they can remove bad actors from chat by issuing temporary (“timeouts”) and/or 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.

Streamers and mods can also change settings to restrict who can chat from everyone to more trusted groups such as followers or subscribers only.

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, streamers and mods utilize two core features: AutoMod and Blocked Terms.

When enabled, AutoMod pre-screens chat messages in 17 languages and holds messages that contain content detected as risky, preventing them from being visible in chat unless they are approved by a mod.

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

These features are best utilized through Mod View , a customizable channel interface that provides mods with a set 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 streamer and their moderators can only affect a user’s access in that channel.

['spam']

In H1 2022, there was a decline in both the number of chat messages manually removed by moderators, and the number of messages proactively withheld from chat via moderation tools such as AutoMod and channel Blocked Terms.

The number of chat messages manually removed decreased from 24.8M to 21.3M (-14.0% HoH), while the number of chat messages proactively withheld decreased from 103.7M to 75.7M (-27.0% HoH).

Overall, the number of deleted chat messages per 1,000 messages decreased from 4.0 to 3.2 (-20.5% HoH).

We believe that one factor affecting these decreases was a collection of enhancements of our proactive detection and safety tooling released in H2 2021.

These enhancements prevented violative messages from showing-up in channels thus driving down the number of harmful messages that must subsequently be removed.

Specifically: (1) In August 2021, we improved our proactive detection models to identify and withhold spam messages that were previously caught and subsequently removed via moderation, (2) In September 2021 we released Phone-verified Chat, which empowers streamers to set stricter standards for who can chat in their channel, preventing some potentially unwelcome chatters from being able to send messages, and (3) In November 2021 we released Ban Evasion Detection, which uses machine learning to flag accounts that may be trying to evade a previous channel-level ban so streamers and moderators can take action proactively.

Channel Enforcement Actions In addition to deleting messages, channel moderators can choose to remove harmful and disruptive users from a channel, either using a timeout for a customizable length of time, or a ban, to prevent any future harm that a disruptive user might cause in the channel.

Banned users may submit an appeal called an Unban Request should they wish to return to the channel.

Unban decisions are made at the discretion of the channel owner and their moderators.

Twitch offers a number of customizable chat “modes” that dictate the types of accounts that can participate in chat.

Examples include followers-only mode, and subscriber-only mode, as well as phone and email verified only.

These proactive modes can help reduce unwanted messages such as spam, and prevent malicious activity, reducing the need for reactive channel enforcements in these cases.

Total channel enforcement actions also declined from 7.8M to 6.6M (-15% HoH).

Enforcements per channel also decreased from 0.46 to 0.40 (-13% HoH).

We believe a number of factors may have contributed to this decline, including enhanced proactive detection and prevention of harmful behaviors and accounts due to new technologies and tooling.

Reports and Enforcements It’s worth noting again that Twitch is a live-streaming service, thus the vast majority of the content on Twitch is ephemeral.

For this reason, we do not consider “content removal” as the primary means of enforcing streamer adherence to our Community Guidelines.

Content is flagged by either machine detection or via user-submitted reports, and our team of content moderation professionals is responsible for reviewing these reports and issuing the appropriate “enforcements” for verified violations.

['spam', 'impersonat', 'integrity']

Adult Nudity, Pornography, and Sexual Conduct We limit community exposure to content that is not appropriate for a diverse audience.

This includes restricting content that involves nudity, insufficient coverage of the body, or inappropriate attire, as well as content that is overtly sexual in nature.

Enforcements for Adult Nudity, Pornography, and Sexual Content increased from 33K in H2 2021 to 40.0K in H1 2022 (+22.8%).

Enforcements / KHW also increased from 0.00292 to 0.00359 (+23.0%).

Once again, it seems that the new Username Policy accounted for most of this increase, as non-username-related enforcements for sexual content and nudity actually decreased slightly.

Spam and Other Community Guidelines Violations Twitch prohibits disruptive activities such as spamming, because these types of activities violate the integrity of Twitch services, and detract from users’ experiences on Twitch.

We also do not allow other dishonest or inappropriate behaviors such as impersonation, broadcasting others against their wishes, ban evasion, misuse of Twitch tools, intentionally miscategorizing a stream, cheating in a game or playing a prohibited game, inappropriate usernames, and underaged user accounts.

Total enforcements issued in this category decreased significantly from 497K to 164K between H2 2021 and H1 2022 (-67.1%).

Enforcements/KHW decreased from .045 to .014 (-68.9%).

This continues a trend of steady decreases since H2 2020.

Of note, spam reports also decreased in H1 2022, from 4M in H2 2021 to 3.5M (-13%).

Contributing factors of this most recent decline likely include proactive measures we introduced in H2 2021 to proactively curb spam and bot abuse, as well as the H2 2021 introduction of new channel-level safety tools that help prevent lower-quality bot and spam accounts from chatting.

In addition to the enforcement actions listed, Twitch programmatically identifies large numbers of bot accounts.

These accounts, which are typically used to artificially inflate view counts, are identified and terminated.

In H1 2022, Twitch took programmatic bulk actions resulting in the issuance of an additional 970K enforcements.

These enforcement actions are not included in the figures listed above because they do not stem from reports or machine detection of harmful content.

Law Enforcement and Government Requests Overview Twitch’s Law Enforcement Response (LER) team is responsible for handling all cases related to any harm against a minor, escalation of violent threats or terrorist acts to appropriate authorities, other legally required reporting to law enforcement, and responding to requests for user data from law enforcement agencies.

Cases of these types are escalated to the Law Enforcement Response team from our content moderation team.