twitch.json

hatespeech

Page ID: 12

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

['violen']

We also believe that clear, relevant Community Guidelines are critical to establishing consistency in our enforcement actions.

Service-Level Safety Service-level safety consists of technology and operations used to uphold the Community Guidelines across Twitch.

Service-level safety consists of three parts: machine detection, user reporting, and review and enforcement.

Machine Detection:

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.

['violen', 'hate']

These tools allow streamers and mods to manage who can participate in chat, automatically filter unwanted messages before they are displayed, give users “time outs” (lock them out of chat for a period of time), or permanently block them from participating in the channel.

Viewer-Level Safety

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.

(Announced June 2022, implemented June 2022)

Products Appeals Portal : If community members believe they incorrectly received an account enforcement, they can submit an appeal via the new appeals portal .

The appeals portal provides visibility into the enforcements that are eligible for appeal, and also displays the status and outcome of ongoing and prior requests, respectively.

(Released March 2022)

['violen', 'hate']

We added a series of educational videos featuring Twitch streamers in order to visually present safety information in a more easily consumable and engaging format.

(Released June 2021) Suspended channel page messaging: We added more detail to the messaging on unavailable channels’ landing pages in order to give viewers more insight into why the channel they’re visiting is unavailable.

(Implemented January 2022)

Important Safety Incidents on Twitch : On Saturday May 14, 2022, a gunman opened fire in a Buffalo, New York, supermarket and live-streamed it on Twitch.

The gunman simultaneously released a white-supremacist manifesto on another content-sharing service.

Twitch used several mechanisms to detect, escalate, and remove the stream and user from our service less than two minutes after the violence began.

The live-stream was not recorded through Twitch features, but was recorded by a viewer — and that video was re-posted on other services.

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.

['hate']

Reports Made on Twitch Overall, user reports have decreased consistently since an atypical spike in H1 2021, when pandemic shutdowns helped drive an influx of new users.

From H2 2021 to H1 2022, the total number of reports submitted by users decreased from 7.8M to 7.4M (-4.8% HoH).

When evaluating user report volume, we also use Reports per Thousand Hours Watched (reports/KHW) because it shows whether reporting volumes are increasing or decreasing relative to the overall amount of Twitch usage.

Reports / KHW also decreased from 0.70 to 0.64 from H2 2021 to H1 2022 (-7.7% HoH).

We believe this decrease is due at least in part to new technologies and tooling we implemented throughout H2 2021, which are successfully deterring more reportable behavior before it occurs.

User reports/KHW decreased by 10-15% in all categories except for Hateful Conduct & Harassment, which saw an increase from 2.1M to 2.3M (+8.0%), and Terrorism, which increased more substantially from 112K to 273K (+144%).

The increase in terrorism reports appears to be largely attributed to the conflict between Russia and Ukraine.

Weekly reporting for Terrorism increased from 4.5K reports/week prior to the beginning of the conflict (24-Feb 2022) to 12.1K reports/week thereafter (+166%) .

Enforcements User reports are prioritized based on a number of factors, including the classification and severity of the reported behavior, and whether or not the behavior constitutes an illegal act.

Each report, regardless of priority, is sent to our content moderation team for review.

If the reviewer agrees that the report demonstrates a violation of the Community Guidelines, the reviewer will issue an enforcement action against the violator’s account.

The enforcement issued depends on the nature of the violation, and can range from a warning, a temporary suspension (1-30 days), or for the most serious offenses, an indefinite suspension from Twitch.

['discrimin', 'violen', 'hate', 'abus']

We also maintain an appeals process, so that if a user believes an enforcement is incorrect, unwarranted or unfair, they can appeal the enforcement.

For more information on account enforcement, see our Account Enforcements page.

In H1 2022, the total number of enforcement actions increased from 886K to 1.1M (+23% HoH) and enforcements / KHW increased from 0.079 to 0.095 (+20% HoH).

This could seem surprising, given that user reports decreased from H2 2021 to H1 2022, but one factor that helps explain this divergence is the release of our new user reporting tool in March 2022, as it helped improve the accuracy of user reports.

In the sections below, we provide data and analysis on the various types of enforcement actions that Twitch has issued between H1 2020 and H2 2022.

Hateful Conduct, Sexual Harassment, and Harassment We do not tolerate conduct or speech that is hateful or harassing, or that encourages or incites others to engage in hateful or harassing conduct.

This includes unwanted sexual advances and solicitations, inciting targeted community abuse, and expressions of hatred based on an identity-based protected characteristic.

Hateful conduct is any content or activity that promotes or encourages discrimination, denigration, harassment, or violence based on the following protected characteristics: race, ethnicity, color, caste, national origin, immigration status, religion, sex, gender, gender identity, sexual orientation, disability, serious medical condition, and veteran status.

We also provide certain protections for age.

Enforcements for Hateful Conduct, Sexual Harassment, and Harassment increased from 171K to 183K (+6.7%), while enforcements/KHW increased from 0.0153 to 0.0159 (+3.5%).

We believe this increase was predominantly due to our new Username Policy, which in February 2022 introduced a stricter standard for usernames as well as a machine learning tool that, after a grace period, automatically enforced against violative channels.

Hateful conduct, sexual harassment, and harassment enforcements remained flat when calculated without these username-specific enforcements.

Extreme Violence, Gore, and Other Obscene Conduct In an effort to limit community exposure to content that may be illegal, jarring or damaging, we prohibit media and conduct that focuses on extreme gore or violence, sexual violence, violent threats, self-harm behaviors, animal cruelty, dangerous or distracted driving, and other illegal, disturbing, or frightening content/conduct.

Enforcements for Extreme Violence, Gore, and Other Obscene Conduct increased from 7.5K in H2 2021, to 8.6K in H1 2022 (+10.8%).

Enforcements/KHW increased from 0.00069 to 0.00075 (+8.7%), even though the number of user reports for Extreme Violence, Gore, and Other Obscene Conduct decreased across the same periods.

While fluctuation is normal for a category with lower reporting and enforcement rates, the March 2022 username policy update, which took a stronger stance against usernames with drug, violence, and self-harm references, drove the increase in enforcements for this category in H1 2022.

Terrorism, Terrorist Propaganda, and Recruitment Twitch does not allow content that depicts, glorifies, encourages, or supports terrorism, or violent extremist actors or acts.

This includes threatening to or encouraging others to commit acts that would result in serious physical harm to groups of people or significant property destruction.

This also includes displaying or linking to terrorist or extremist propaganda, even for the purposes of denouncing such content.

As mentioned in previous reports, we receive low volumes of reports and enforcements for terrorism on our service.

While enforcements in H1 2022 still remained low compared to industry norms, there was a notable increase in terrorism enforcements, from 34 in H2 2021 to 171 (+403% HoH).

Of note, 114 (67%) of these enforcements were for “Terrorism via username”, indicating that the March 2022 introduction of our new Username Policy likely played a significant role in this increase.

The remaining 57 enforcements in H1 2022 still represent an increase from H1 2021, though given the very low number of enforcements, this fluctuation is within expectations.

['violen', 'hate', 'abus']

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.

The LER team is also responsible for investigating reports of violations of Twitch’s Off-Service Conduct Policy, which launched in April 2021 and addresses serious offenses that pose a physical safety risk to the Twitch community or Twitch staff, even if these offenses occur entirely off Twitch.

The policy covers offenses such as extreme or deadly violence, acts of terrorism, leadership or membership in a known hate group, sexual assault, the sexual exploitation of minors, and other acts or credible threats that endanger members of the Twitch community.

These reports are made directly to the LER team through a dedicated reporting alias, or are referred to the LER team from the content moderation team if the offense is reported through Twitch’s user reporting tool.

NCMEC Reporting; Global Cooperation Twitch has zero tolerance for sexual exploitation of minors.

When we are made aware of media depicting sexual exploitation of a minor, or grooming behavior, we remove the content, investigate, and report to authorities via the National Center for Missing & Exploited Children (NCMEC).

We also work directly with aligned organizations throughout the world, such as INHOPE and ICMEC, to address and prevent exploitation media and grooming of minors from occurring on Twitch.

From H2 2021 to H1 2022, The number of NCMEC Cyber Tips submitted by Twitch increased from 4K to 6.7K (+68% HoH) while NCMEC CyberTips / KHW increased from 0.00036 to 0.00058 (+62% HoH).

We monitor any activity related to CSAM extremely closely on Twitch, including through machine learning models to detect potential predators, and through expert deep-dive investigations into bad actors and their networks.

While it’s unfortunately true that NCMEC reports are rising industry-wide, we believe part of the increase on Twitch in H2 2022 was driven by detection improvements throughout 2021, including significant investments in staffing, proactive detection, and user reporting.

This has enabled us to take action in a higher number of cases as Twitch’s audience grows.

Escalations to Law Enforcement When Twitch identifies credible threats of violence, a member of our Law Enforcement Response team will proactively send, or “escalate,” user data to appropriate law enforcement agencies.

Escalations to law enforcement increased from 32 to 52 (+63% HoH) in H1 2022, while escalations / KHW increased from 0.0000029 to 0.0000045 (+55% HoH).

Data requests processed by Twitch from law enforcement and governmental agencies increased from 96 to 137 (+43% HoH) in H1 2022, while requests processed / KHW increased from 0.0000086 to 0.0000119 (+38% HoH).

This is within the expected variance for the number of valid subpoenas and reservation hold requests, and aligns with volumes we have received from law enforcement in the past.

Past Transparency Reports July-December (H2) 2021 January-June (H1) 2021

Full Year 2020 Safety at Twitch H1 2022 Safety Updates Moderation in Channels: Coverage, Removals, and Enforcements Reports Made on Twitch Enforcements Law Enforcement and Government Requests Past Transparency Reports Company About Legal Blog Help twitch.tv ↗ Newsroom Safety News Press Terms of Service Privacy Policy Ad Choices Cookie Policy Partners Affiliates

© 2023 Twitch Interactive, Inc.