twitch.json

hatespeech

Page ID: 11

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

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

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.

Section 9 also reserves Twitch’s right to remove, screen, or edit user-generated content where such content violates the Terms of Service or applicable law, and warns that any use of Twitch in violation of the Terms of Service may result in termination or suspension of the user’s rights to use Twitch.

Finally, Section 14 states Twitch’s right to terminate a user’s relationship with Twitch and to block their further access to Twitch (to the extent permitted by applicable law) if, among other cases, the user violates the Terms of Service, which includes the foregoing mentioned provisions on user-generated content.

It is of particular importance to Twitch to clearly communicate to users what kind of content may be distributed and how.

As set out above, our Terms of Service and Community Guidelines therefore explain what Twitch offers for their users to express themselves, which conduct is prohibited, and how we enforce these rules, in a clear and transparent manner.

In the Community Guidelines, we provide a further comprehensive overview on the key concerns of content distribution, such as harassment, sexual content, and copyrighted content.

Help articles provide additional examples on the permissibility of the distribution of content as well as the concrete enforcement steps that can be expected from Twitch, and the venues for appeal that are open to users that believe that they and their content were affected in error.

['violen']

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.

When the system identifies content that may potentially fall into these categories, the system generates an urgent report that is escalated to determine the appropriate enforcement action, which may include removing the content.

For credible user reports of high-harm content, Twitch’s automated systems also temporarily downgrade the visibility of potentially problematic content until a full moderator review can be completed, including hiding the reported channel from discoverability, meaning the channel cannot be found through the home page, search bar, or other discovery tools, until the moderator review is completed.

4) Submission of Reports about Allegedly Unlawful Content We also empower users to report harmful or inappropriate behavior to Twitch, with an easy, direct reporting mechanism from the content.

These reports are reviewed and acted on by a team of skilled and trained professionals who can apply service-wide enforcement actions.

Content that violates our Community Guidelines will be removed globally (if applicable) and a suspension will be issued to the offending user, up to an indefinite suspension for the first infraction depending on the nature of the violation.

For content that is identified as unlawful under the German Criminal Code provisions listed in NetzDG but does not violate our Community Guidelines, we will geo-block the content from Germany.

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

There is an initial page with broad categories of violations; some of those categories have secondary pages to help provide more detailed reporting reasons.

This helps walk the user through the reporting process, making it easier to select the most relevant category.

For logged-in users, the violation categories will include those covered under NetzDG criminal statutes as well as categories that only violate Twitch’s terms of service; logged-in users are permitted to file a report for either.

An open text field allows users to submit additional text evidence if desired.

If a category relevant to NetzDG is selected, a final interface before submission asks the user to verify the truthfulness of the information and submit an email contact for timely updates.

Entry point for reporting: triple dot in the lower right of video player Initial reporting categories Additional reporting categories under “Bullying or Harassment” Additional categories under “Hateful Conduct” Additional categories under “Search” Open text field for text-based evidence submission Final interface before NetzDG report submission Users who are not logged in can report content in an identical manner as a logged-in user.

The access to our reporting interface is in the same locations for a not logged-in user as it is for a logged-in user.

The only distinction between the two experiences is that not logged-in users will only see violation categories related to NetzDG, both in the initial reporting categories list as well as on the more detailed reporting reasons pages.

Initial reporting screen for not logged-in users Additional reporting categories under “Hate speech, harassment, or abuse” for not logged-in users Additional reporting categories under “Inappropriate or obscene content” for not logged-in users Additional categories under “Give me more options…” for not logged-in users To make it easy for users to submit a report, Twitch used existing reporting reasons that map to one or more of the relevant criminal code offenses and created new reporting reasons for statutes that were not otherwise covered.

The following 12 categories each match to one or several of the relevant Criminal Code offenses.

Since Twitch used existing reporting reasons where applicable, some offenses can be found under multiple categories: Defamation or Insult (accessed under ‘Hateful Conduct’ in the initial reporting form)

§ 166 StGB: Defamation of religions, religious and ideological associations § 185 StGB:

Insult § 186 StGB: Defamation § 187 StGB: Intentional defamation § 189 StGB: Defiling memory of dead Extreme Violence or Gore (accessed under ‘Violence or Gore’ in the initial reporting form)

§ 131 StGB: Dissemination of depictions of violence Forming Criminal Organizations (accessed under ‘Search’)

§ 129 StGB: Forming criminal organizations Harassing or bullying someone (accessed under ‘Bullying or Harassment’)

§ 185 StGB: Insult § 186 StGB: Defamation § 187 StGB: Intentional defamation § 189 StGB: Defiling memory of dead Hate speech, hateful symbols, or other hateful conduct (accessed under ‘Hateful Conduct’)

§ 130 StGB: Incitement to hatred § 166 StGB: Defamation of religions, religious and ideological associations Rewarding or approving of offenses (accessed under ‘Search’)

§ 140 StGB: Rewarding and approving of offenses Sexual conduct involving violence, children, or animals (accessed under ‘‘Search’)

§ 184b StGB: Distribution, acquisition and possession of child pornography Sharing private photos of someone else (accessed under ‘Search’)

§ 201a StGB: Violation of intimate privacy by taking photographs Terrorism or Acts of Mass Violence (accessed under ‘Terrorism’)

§ 89a StGB: Preparation of a serious violent offense endangering the state § 91 StGB:

Encouraging the commission of a serious violent offense endangering the state § 129a StGB: Forming terrorist organizations § 129b StGB: Criminal and terrorist organizations abroad; extended confiscation and deprivation Threats, harm or endangerment of someone (accessed under ‘Hateful Conduct’)

§ 111 StGB: Public incitement to crime § 126 StGB: Breach of the public peace by threatening to commit offenses § 241 StGB: Threatening the commission of a felony Treasonous forgery or forgery of data (accessed under ‘Search’)

§ 100a StGB: Treasonous forgery § 269 StGB: Forgery of data intended to provide proof Unconstitutional content (accessed under ‘Search’)

§ 86 StGB: Dissemination of propaganda material of unconstitutional organizations § 86a StGB: Using symbols of unconstitutional organizations Once an individual has completed the reporting form, their report automatically reaches us through an internal review tool—no additional action is required by the reporting party.

We then concurrently review whether the reported content violates our Community Guidelines and/or the relevant provisions of the German Criminal Code listed in NetzDG (see next section for more details).

5) Handling of Reports In general, Twitch prioritizes user reports based on a number of factors, including the classification and severity of the reported behavior, and whether or not the behavior is illegal.

When we receive a user report, it kicks off two simultaneous workstreams: (1) a review against Twitch’s Community Guidelines, and (2) a review against the German Criminal Code offenses listed in NetzDG.

['violen']

For particularly complex reports where the legality of the reported content is still unclear, the report is then escalated to our in-house lawyers.

If our in-house lawyers have any questions about the legality of content identified in a report, they obtain a legal assessment from outside counsel in Germany.

If the reported content is deemed unlawful under the relevant provisions of the German Criminal Code listed in NetzDG, we disable access to that content in Germany.

As most content on Twitch is gone the moment it’s created, often Twitch cannot delete/block the offending content as it is no longer available at the time of review since the live broadcast is over.

If the content was saved or is in the form of a Video-On-Demand (VOD) or a user’s Clip—the latter being a snippet of a live stream that is 90 seconds or less—Twitch will restrict access or delete the VOD if the content is deemed violative.

In addition to the above, Twitch also maintains a Law Enforcement Response team, which is responsible for handling all cases related to any harm against a child, escalation of violent threats, or terrorist acts to appropriate authorities, any other legally required reporting to law enforcement, and requests for user data from law enforcement agencies.

Our content moderation team escalates these types of cases to our Law Enforcement Response team.

For further information, please see our Global Transparency Report .

6) Resources / Review Teams Across our various content moderation locations and offices, there are at any one time approximately 15 content moderation employees with specialized, dedicated training on NetzDG.

This number can fluctuate by small degrees depending on the volume of cases that require review and is optimized to ensure rigorous standards for Review Time are met (see Part 2, section 4).

Twitch will expand or reduce the number of these employees depending on case numbers, in order to ensure adherence to NetzDG requirements.

Content moderation professionals come from a variety of professional and academic backgrounds to ensure both competency in functional job roles but also diversity of thought—we do not believe there is a single profile that makes a good content moderator, nor would we feel this would be the best way to safeguard our community, who are similarly diverse.

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

The first tool offered to both creators and viewers is the reporting function, as described in detail above.

Once a report has been filed, the user will also be presented with an option to block the account.

As well as the reporting function, Twitch also provides creators with the ability to build a team of community moderators who moderate chat in the creator’s channel.

Twitch provides both creators and their moderators with a powerful suite of tools such as AutoMod, Chat Modes , account verification , Suspicious User Detection, and Mod View to make their roles as easy and intuitive as possible.

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

Viewers are also provided with various tools including chat filters for Discrimination, Sexually Explicit Language, Hostility, and Profanity.

Part 2: Statistics on Reports The following tables show the number of reports received between January 1 and June 30 (H1 2022).

1) Report Volumes Table 1: Reports by Reporter Type

The table below includes all reports submitted under report reasons related to NetzDG, broken down by reporter type.

These reports include potential violations of both Twitch’s Community Guidelines/Terms of Service and the German Criminal Code provisions listed under NetzDG.

Note that Twitch cannot verify whether reports allegedly submitted by an association were actually submitted by an association; we rely on user self-identification in the reporting flow.

Time Period Reports from Associations Reports from Other Individuals Total Reports H1 2022 6,060 38,239 44,299 Table 2: Reports by Report Reason

This table breaks down the number of reports included in the table above according to the reason cited by the reporting party.

Report Reason # of Reports (H1 2022)

Defamation or Insult 9,060

Extreme Violence or Gore 1,233 Forming Criminal Organizations 317 Harassing or bullying someone 7,736 Hate speech, hateful symbols, or other hateful conduct 16,564 Rewarding or approving of offenses 425 Sexual conduct involving violence, children, or animals 739 Sharing private photos of someone else 515 Terrorism or Acts of Mass Violence 1,797 Threats, harm or endangerment of someone 3,681 Treasonous forgery or forgery of data 975 Unconstitutional content 1,257 2) Number of Reports Forwarded for External Review In H1 2022 (January 01 - June 30), Twitch did not forward any cases to self-regulatory bodies or external counsel for review.

3) Removal/Blocking Volumes Table 3: Deletion/Blocking Volumes by Reporter Type

The following table outlines the number of reports that led to the blocking or removal of content, broken down by reporter type.

The table includes all reports submitted under report reasons related to NetzDG, and includes both violations of Twitch’s Community Guidelines/Terms of Service and the German Criminal Codes listed under NetzDG.

Time Period Reports from Associations Reports from Other Individuals Total H1 2022 361 2,533 2,894 Table 4: Deletion/Blocking Volumes by Report Reason

The following table outlines the number of reports that led to the blocking or removal of content, broken down by the reason cited by the reporting party at the time of report (which may not be the same as the actual reason for removal or blocking).

Again, this table includes reports that led to removal under Twitch’s applicable Community Guidelines/Terms of Service and the German Criminal Code provisions listed under NetzDG.

Report Reason # of Reports (H1 2022)

Defamation or Insult 624

Extreme Violence or Gore 48 Forming Criminal Organizations 16 Harassing or bullying someone 309 Hate speech, hateful symbols, or other hateful conduct 1,214 Rewarding or approving of offenses 7 Sexual conduct involving violence, children, or animals 47 Sharing private photos of someone else 83 Terrorism or Acts of Mass Violence 87 Threats, harm or endangerment of someone 188 Treasonous forgery or forgery of data 159 Unconstitutional content 112 Table 5: Deletion/Blocking Volumes by Justification

The following table outlines the number of reports that led to global removal of content due to a Terms of Service (ToS) or Community Guidelines (CG) violation (ToS/CG Violation) or local blocking due to a violation of a provision of the German Criminal Code listed in the NetzDG (NetzDG Violation).

The first column shows the number of reports that led to content removal as a result of violations of a provision of the German Criminal Code listed in the NetzDG that were also violations of our ToS/CG—content was therefore removed globally.

The second column shows the number of reports that led to content removal as a result of violations of a provision of the German Criminal Code listed in the NetzDG that were not violations of our ToS/CG—content was therefore blocked only in Germany.

As discussed above, all reports are reviewed simultaneously for violations of both ToS/CG and a provision of the German Criminal Code listed in the NetzDG.

Time Period # of ToS/CG Violations # of NetzDG Violations H1 2022 627 2,267 Table 6: Deletion/Blocking Volumes by Step in Review Process

The following table outlines the number of reports that led to the removal or blocking of content at each step in the review process.

['violen', 'hate']

Reports from Other Individuals Time Period 24 Hours 48 Hours 7 Days > 7 Days H1 2022 2,492 22 11 4 Total Time Period 24 Hours 48 Hours 7 Days > 7 Days H1 2022 2,850 25 11 4 For the four reports that exceeded seven days, this was due to our QA teams reviewing closed reports, reopening them, and correcting errors in enforcement.

Table 8: Review Time by Report Reason

The following table outlines the number of reports that led to the blocking or removal of content, broken down by review time and the reason cited by the reporting party at the time of report (which may not be the same as the actual reason for removal or blocking).

Number of Reports (H1 2022)

Report Reason 24 Hours 48 Hours 7 Days >7 Days Defamation or Insult 617 5 2 0

Extreme Violence or Gore 48 0 0 0

Forming Criminal Organizations 16 0 0 0

Harassing or bullying someone 304 3 3 0 Hate speech, hateful symbols, or other hateful conduct 1,189 0 5 2 Rewarding or approving of offenses 6 0 0 1 Sexual conduct involving violence, children, or animals 47 0 1 0

Sharing private photos of someone else 83 14 0 0

Terrorism or Acts of Mass Violence 86 0 0 0

Threats, harm or endangerment of someone 185 0 0 1 Treasonous forgery or forgery of data 159 0 0 0

Unconstitutional content 110 2 0 0 5) Appeals

The following tables outline the number of appeals received by Twitch.

Table 9 outlines the number of appeals in the context of illegal content.

The first table separates by number of relevant appeals submitted by the complainant, and by the affected person.

The second table separates the number of times Twitch reversed the decision due to appeal by the complainant, and by the affected person.