Australian Code of Practice on Disinformation and Misinformation
Twitch
Baseline Transparency Report - May 2024

Summary

Twitch is a live streaming service, where creators engage in a wide variety of different activities, including video games, music, cooking, and creating creative content.

At Twitch, we strive to create a space that supports and sustains streamers' ability to express themselves, and provides a welcoming and entertaining environment for viewers, free of illegal and harmful interactions. This starts with Twitch's Community Guidelines, which seek to balance user expression with community safety, and set the rules for the behaviour of everyone on Twitch. Our Community Guidelines are developed by a dedicated team of policy professionals in consultation with external safety, human rights, and policy experts, and we review and update them regularly to respond to the community's evolving needs.

We identify potential violations of our Community Guidelines using a combination of machine detection, proactive human review, and user reporting. Our global safety operations team works to quickly review content and accounts flagged by users and by our machine detection models. The speed at which we can respond to user reports is critical given the live nature of Twitch, and in H2 2023, we responded to 95% of reports in under 1 hour and 99.95% of reports in under 24 hours. We prioritise having a human in the review process to ensure that decisions are accurate and fair for our community members.

We take pride in how Twitch fosters community and brings people together, and we believe that individuals who use online services to spread false, harmful information do not have a place in our community. This is why we have a Harmful Misinformation Actor policy. Harmful misinformation actors account for a disproportionate amount of damaging, widely debunked misinformation online. These actors share three characteristics: their online presence—whether on or off Twitch—is dedicated to (1) persistently sharing (2) widely disproven and broadly shared (3) harmful misinformation topics, such as conspiracies that promote violence. We prohibit harmful misinformation actors who meet all three of these criteria since taken together they create the highest risk of harm, including inciting real world harm.

Even if someone is not a Harmful Misinformation Actor, Twitch prohibits and enforces against misinformation that targets specific communities under our Hateful Conduct & Harassment policies, and we take action on content that encourages others to engage in physically harmful behaviour under our Self-Destructive Behaviour policy.

In addition to misinformation, Twitch invests significant resources to ban bots, spammers, impersonators, and other types of bad actors to combat inauthenticity on our service. We have proactive detection working alongside our reporting system to programmatically remove bots, known bad actors, and those who are trying to evade a suspension or ban.

While misinformation is not currently prevalent on Twitch, we recognize the harm that this content can cause, particularly when it is related to an election. We are always evolving our approach to safety in accordance with expert guidance and trends in our community. We understand that the prevalence of harmful misinformation can change, and we will continue to engage with industry, academia, and civil society to adapt our approach as necessary to ensure its continuing effectiveness. We participate in a variety of industry knowledge-sharing initiatives—including the EU Code of Practice on Disinformation, the New Zealand Code of Practice for Online Safety and Harms (which also addresses disinformation), the EU Hate Speech Code, the EU Internet Forum, and the Global Internet Forum to Counter Terrorism (GIFCT)—to stay abreast of industry trends and risks.

We are proud to contribute to the goals and commitments of the Australian Voluntary Code of Practice on Disinformation and Misinformation (ACPDM).

Commitments under the Code

Twitch has committed to the following six Objectives and related Outcomes.

Objective 1 - Provide safeguards against harms that may arise from disinformation and misinformation	
1a	Signatories contribute to reducing the risk of harms that may arise from the propagation of disinformation and misinformation on digital platforms by adopting a range of scalable measures.
1b	Users will be informed about the types of behaviours and types of content that will be prohibited and/or managed by Signatories under this Code.
1c	Users can report content or behaviours to Signatories that violate their policies under section 5.10 through publicly available and accessible reporting tools.
1d	Users will be able to access general information about Signatories' actions in response to reports made under 5.11.
1e	Users will be able to access general information about Signatories' use of recommender systems and have options relating to content suggested by recommender systems.
Objective 2 - Disrupt advertising and monetisation incentives for disinformation	
2	Advertising and/or monetisation incentives for disinformation and misinformation are

	reduced.
Objective 3 - Work to ensure the integrity and security of services and products delivered by digital platforms.	
3	The risk that Inauthentic User Behaviours undermine the integrity and security of services and products is reduced.
Objective 4 - Empower consumers to make better informed choices of digital content.	
4	Users are enabled to make more informed choices about the source of news and factual content accessed via digital platforms and are better equipped to identify misinformation.
Objective 6 - Strengthen public understanding of disinformation and misinformation through support of strategic research	
6	Signatories support the efforts of independent researchers to improve public understanding of disinformation and misinformation.
Objective 7 - Signatories will publicise the measures they take to combat disinformation and misinformation.	
7	The public can access information about the measures Signatories have taken to combat disinformation and misinformation.

Twitch did not subscribe to Objective 5 (Improve public awareness of the source of political advertising carried on digital platforms) as we do not permit political ads.

Reporting against commitments

Outcome 1a: Reducing harm by adopting scalable measures

In order to reduce harm to our community and the public without undermining our streamers' open dialogue with their audiences, we prohibit Harmful Misinformation Actors who persistently share misinformation on or off of Twitch. We suspend users whose online presence is dedicated to (1) persistently sharing (2) widely disproven and broadly shared (3) harmful misinformation topics.

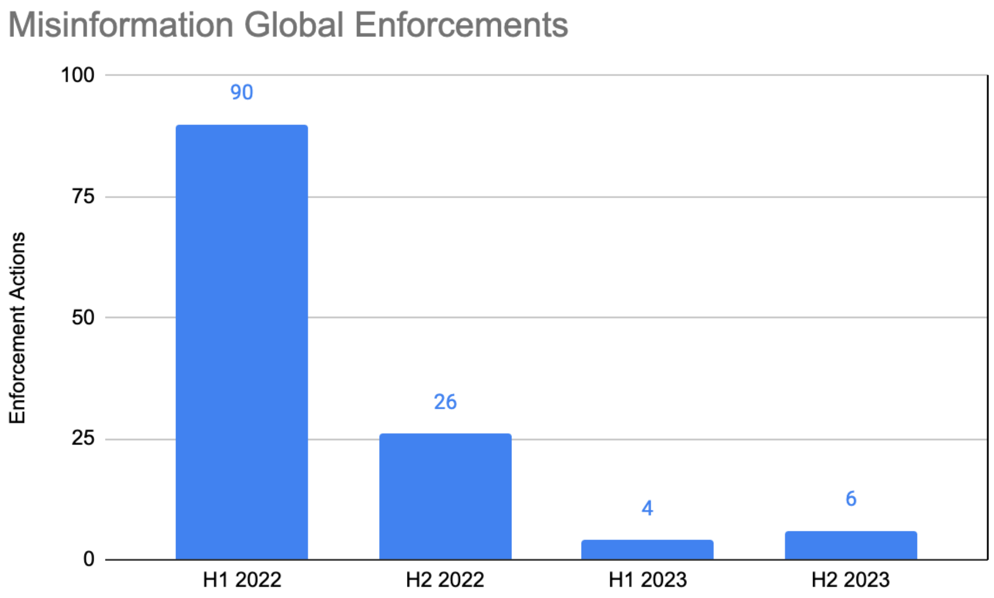
This policy is focused on Twitch users who persistently share harmful misinformation, including AI-generated misinformation. This focus represents a refinement of the ACPDM's definition of misinformation. Twitch's policy will not be applied to users based upon individual statements or discussions that occur on the channel. We evaluate whether a user violates the policy by assessing both their on-service behaviour as well as their off-service behaviour.

Under our Harmful Misinformation Actor Policy, we cover the following topic areas, and will continue to update this list as new trends emerge:

- Misinformation that targets protected groups, which is already prohibited under our Hateful Conduct & Harassment Policy

- Harmful health misinformation and wide-spread conspiracy theories related to dangerous treatments, COVID-19, and COVID-19 vaccine misinformation
    - Discussions of treatments that are known to be harmful without noting the dangers of such treatments
    - For COVID-19—and any other WHO-declared Public Health Emergency of International Concern (PHEIC)—misinformation that causes imminent physical harm or is part of a broad conspiracy
- Misinformation promoted by conspiracy networks tied to violence and/or promoting violence
- Civic misinformation that undermines the integrity of a civic or political process
    - Promotion of verifiably false claims related to the outcome of a fully vetted political process, including election rigging, ballot tampering, vote tallying, or election fraud
- In instances of public emergencies (e.g., wildfires, earthquakes, active shootings), we may also act on misinformation that may impact public safety)

In H2 2023, we indefinitely suspended 6 accounts globally for violating our Harmful Misinformation Actor Policy. Historical enforcement information is included in the chart below (we introduced our Harmful Misinformation Actor policy in H1 2022).



Our enforcement numbers are relatively low due to several factors. (i) The mechanics of Twitch are not conducive to spreading misinformation or investing in large-scale disinformation campaigns. It is extremely difficult for a new streamer to garner large numbers of concurrent viewers; it takes time to grow an audience on Twitch. Most Twitch content is also long-form and ephemeral. Since this means that most content is gone the moment it is created, it is not shared and does not go viral in the same way that it does on other UGC video and social media services, where videos are uploaded and can be viewed and shared by users on demand. (ii) Our targeted policy only applies to those who persistently share harmful misinformation. Due to the long-form nature of Twitch's content, we are focused on a streamer's aggregated content rather than a specific, isolated statement within a longer piece of content. (iii) When we launched our Harmful Misinformation Actor policy, we took swift action against accounts that posed a threat to our community. We believe enforcement of our policy—particularly upon its adoption in H1 2022—has been an effective deterrent to harmful misinformation actors; we have not seen large numbers of them attempt to join our service.

Even if someone is not a Harmful Misinformation Actor, Twitch enforces on misinformation that targets specific communities under our Hateful Conduct & Harassment policies, and we take action on content that encourages others to engage in physically harmful behaviour under our Self-Destructive Behaviour policy. More information on enforcements under these policies can be found in Twitch's Safety Transparency Report.

Outcome 1b: Inform users about what content is targeted

Twitch's Harmful Misinformation Actor policy is outlined in our Community Guidelines (CGs). Our aim is to have CGs that are clear and easy to follow but also thorough (with examples of prohibited behavior) to help Twitch users understand the boundaries we have set so they can feel confident expressing themselves within those boundaries.

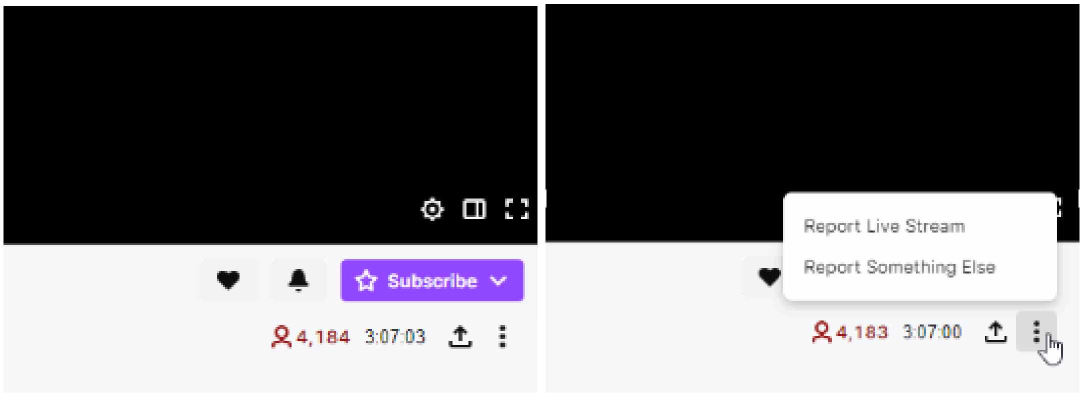
When we launched our Harmful Misinformation Actor policy in March 2022, we also published a blog post to flag the change for our community and provide more context on the policy.

Additionally, when a user violates our Harmful Misinformation Actor Policy—or any of our policies—they receive a detailed email notification. The notification includes the action taken, whether a suspension is permanent or temporary, the reason for the suspension, examples of violating content, a link to the Community Guidelines to learn more about the policy, where the violation occurred, and a link to the Appeals Portal if they disagree with the decision.

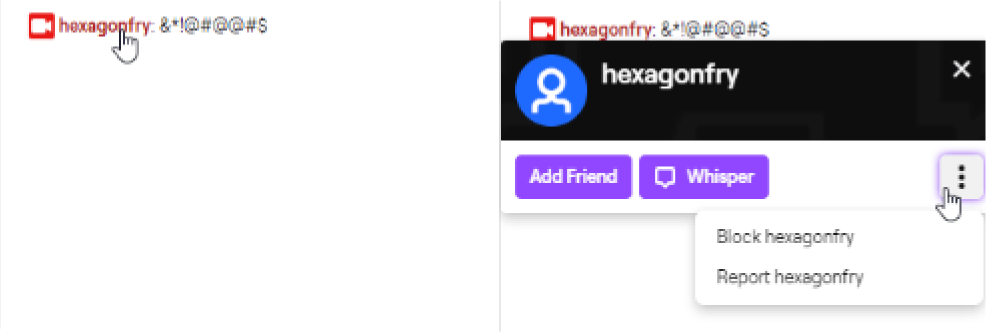
Outcome 1c: Users can easily report offending content

Users are able to report potential harmful misinformation actors with an easy, direct reporting mechanism. Users can submit a report by clicking on the three vertical dots icon, which is shown in the bottom right below the video player on the channel page; on the bottom right of a clip, highlight or past broadcast; or on the bottom right when you click on a username to report. An option to “Report Live Stream” or “Report [username]” will appear for the user to enter our reporting interface.

Reporting a Channel

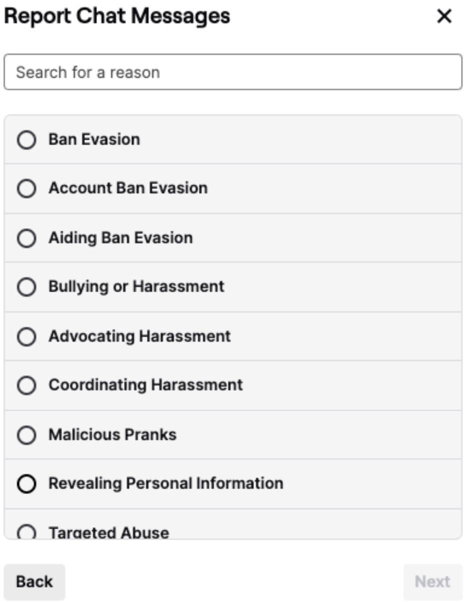


Reporting a User in Chat


hexagonfry: &*!@#@@#$

The interface then prompts the user to select the most relevant category for the violation, which in this case is "Misinformation." Alternatively, a user can search for the appropriate reporting reason.

Reporting Interface


Report Chat Messages ✕

Search for a reason

○ Ban Evasion
○ Account Ban Evasion
○ Aiding Ban Evasion
○ Bullying or Harassment
○ Advocating Harassment
○ Coordinating Harassment
○ Malicious Pranks
○ Revealing Personal Information
○ Targeted Abuse

Back Next

Users can also submit a report to the specialised off-service investigations team through the team's email alias OSIT@twitch.tv.

Learn more about how to file a report on Twitch.

Outcome 1d: Information about reported content available
Twice-a-year, we publish a report outlining how we enforce our Community Guidelines, including our Harmful Misinformation Actor policy. This report is publicly available on our website. We also provide a publicly-available annual transparency report under the EU Code of Practice on Disinformation.

Outcome 1e: Information about recommender engines
Twitch has published a detailed summary of our various recommendation systems, outlining the main parameters used by each system. The page also provides information for how users can influence our recommendations and control what they see on Twitch.

When browsing, viewers can sort by 'Recommended for You' or by other channel attributes. Users can also customise their recommendations on Twitch by letting Twitch know if they are "not interested" in a streamer or content category that is recommended to them. At any time, users can navigate to their settings page and review what they have marked as "not interested" and then edit those selections. Users can learn more about this feature by visiting the Help Article.

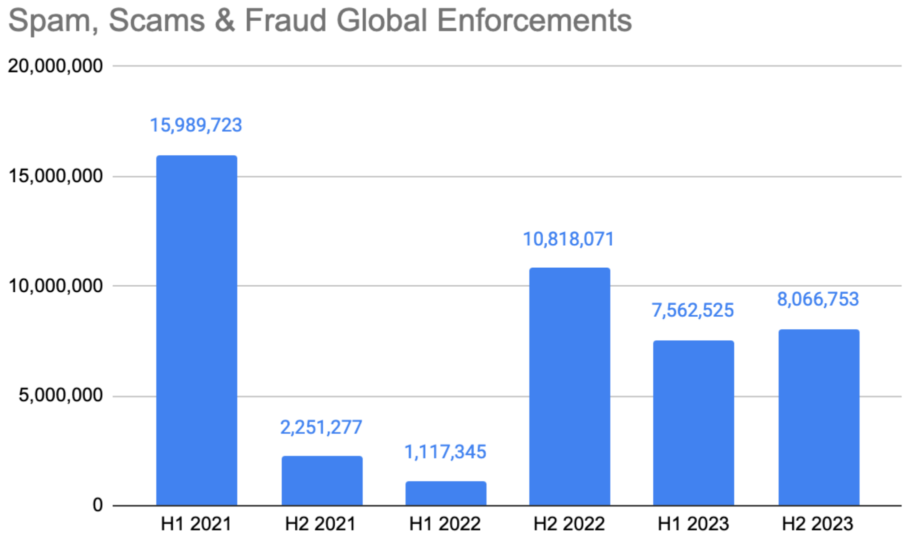
Objective 2: Disrupt advertising and monetisation incentives for disinformation.
Actors that systematically provide harmful misinformation are prohibited from the service, and are therefore not eligible for monetization. Additionally, Twitch's ads policy prohibits ads that contain deceptive, false, or misleading content as well as political content, such as campaigns for or against a politician, political party or related to an election, and/or content related to issues of public debate.

Objective 3: Work to ensure the integrity and security of services and products delivered by digital platforms.
Twitch's policies state that "Any content or activity that disrupts, interrupts, harms, or otherwise violates the integrity of Twitch services or another user's experience or devices is prohibited." This includes the creation of inauthentic and malicious bots, impersonation, engaging in viewership tampering (such as artificially inflating follow or live viewer stats), and selling or sharing user accounts, services, or features.

We use historical enforcement data to proactively identify patterns associated with bots and spammers. Depending on the level of confidence, we can take several actions against a suspected bot account, including requesting that the account verify a mobile phone, auto-reporting the account to be reviewed by our operations team, and adding client-side friction that increases the cost of automation.

Most cases of impersonation on Twitch are phishing attempts, where a fraudulent channel is trying to get a user to click on a malicious link. We scan the text on our channel pages for these malicious URLs and then report the channel for review by our operations team. We also actively monitor channels for viewership tampering, using a combination of handcrafted filters based on ASN and IP reputation, as well as a machine learning model based on past examples.

In H2 2023, we issued 8.1M account enforcements for spam, scams, and fraud globally; 11,207 of these were for accounts based in Australia. Spam can be both automated (published by bots or scripts) or coordinated (when an actor uses multiple accounts to spread deceptive content). Due to its automated and coordinated nature, spam is generally Twitch's largest category of enforcement and we often see significant fluctuations in enforcement between reporting periods. This is consistent with a general trend in the industry.



Objective 4: Empower consumers to make better informed choices of digital content.
Twitch mitigates the risk that users are exposed to harmful misinformation on the site through the measures discussed previously. We are also committed to providing users with information about how our recommendation systems work and options to customise their recommendations as discussed under Outcome 1e above.

Twitch has also invested in a media literacy campaign to empower users to think critically about what information they consume. Twitch collaborated with media literacy expert MediaWise to develop an array of educational materials that teach Twitch streamers and viewers how to better identify, and avoid spreading, misinformation and disinformation online. These materials are hosted on the Twitch Safety Center.

Objective 6: Strengthen public understanding of Disinformation and Misinformation through support of strategic research.
Twitch remains open to supporting independent research if approached. At this time, Twitch does not directly support any third-party research.

Objective 7: Signatories will publicise the measures they take to combat Disinformation.
In Outcome 1b and 1e of this report, we provide details—and links to the corresponding materials—regarding publicly available information on the measures we take to combat misinformation.

Concluding remarks

As a signatory to the Australian Voluntary Code of Practice on Disinformation and Misinformation, Twitch is committed to combating misinformation on our service in an effective yet targeted manner that balances freedom of expression with keeping our communities safe. We recognize that harmful misinformation, and its prevalence on our platform, may evolve and we will continue to evaluate and adapt the measures we have put in place to protect our users and the integrity of our service.