Aotearoa New Zealand Code of Practice for Online Safety and Harms

Annual Update Report (covering July 2023 to June 2024)

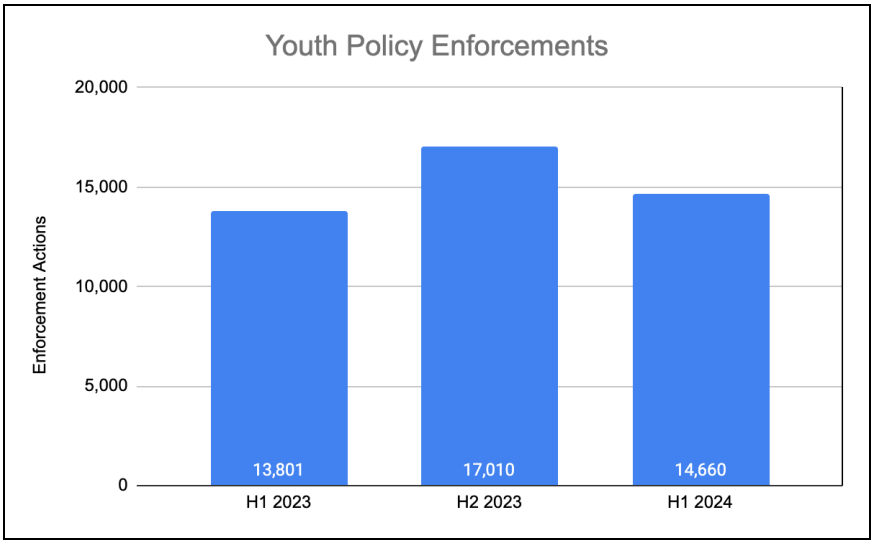
[*In this latest report, Twitch has incorporated suggestions from the Guidelines for Signatories, including specifying the 12-month reporting period, adding relevant information for users in Aotearoa New Zealand along with global metrics, and presenting trended data in clear, easy-to-read charts*.]

Signatory:	**Twitch Interactive, Inc.** 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. Streamers typically build a community over time by streaming for multi-hour sessions over a sustained period. Some streamers with large audiences eventually stream on Twitch as a full-time job, although even small or mid-sized streamers have the option to monetize so long as they meet our minimum requirements. All streamers and viewers must remain in compliance with our policies at all times.

4.1 Reduce the prevalence of harmful content online

Signatories commit to implementing policies, processes, products and/or programs that would promote safety and mitigate risks that may arise from the propagation of harmful content online, as it relates to the themes identified in section 1.4.

Outcome 1. Provide safeguards to reduce the risk of harm arising from online **child sexual exploitation & abuse (CSEA)**
Twitch strictly prohibits and maintains a zero-tolerance policy toward any content or behavior involving illegal child sexual abuse material, child exploitation, grooming, or other forms of child sexual misconduct. Violations of our Youth Safety policy result in immediate and indefinite suspension. (Measures 1-4) In H2 2023, we issued 17,010 enforcements globally for violations of our Youth Safety policy, followed by 14,660 enforcements in H1 2024. We issued 48 account enforcements for violations of our Youth Safety policy based on reports from users in New Zealand in H2 2023, followed by 112 enforcements based on New Zealand user reports in H1 2024. These actions addressed both illegal child sexual exploitation and abuse (CSEA) material, as well as content that, while not illegal, violated our Community Guidelines by endangering minors.



(Twitch Safety Center, H1 2024 Transparency Report)

We continue to report all illegal content and activity to the National Center for Missing and Exploited Children (NCMEC). We submitted 3.3K NCMEC CyberTips in H2 2023, and 1.5K H1 2024. The 53.5% reduction in CyberTips from H2 2023 to H1 2024 reflects a change in categorization to ensure we accurately report illegal content. We continue to enforce against any content that may endanger youth under our Youth Safety policy. We also continue to enhance our proactive detection tools to better identify under-13 users and potential predators. (Measure 4)

Twitch works closely with other social media services and industry organisations to share learnings, and make sure we are staying ahead of emerging risks in this space. We are an active participant in the Technology Coalition, which is a global alliance of tech companies who are working together to protect children from online sexual exploitation and abuse through tech innovation, research, and information sharing. Twitch is also a member of the Family Online Safety Institute (FOSI) and supports the Digital Wellness Lab at Boston Children's Hospital. We are also funding members of INHOPE, which is a global network of 50 hotlines in 46 countries (including New Zealand) that provide the public with a way to anonymously report CSAM content. (Measure 5)

Outcome 2: Provide safeguards to reduce the risk of harm arising from online **bullying or harassment**

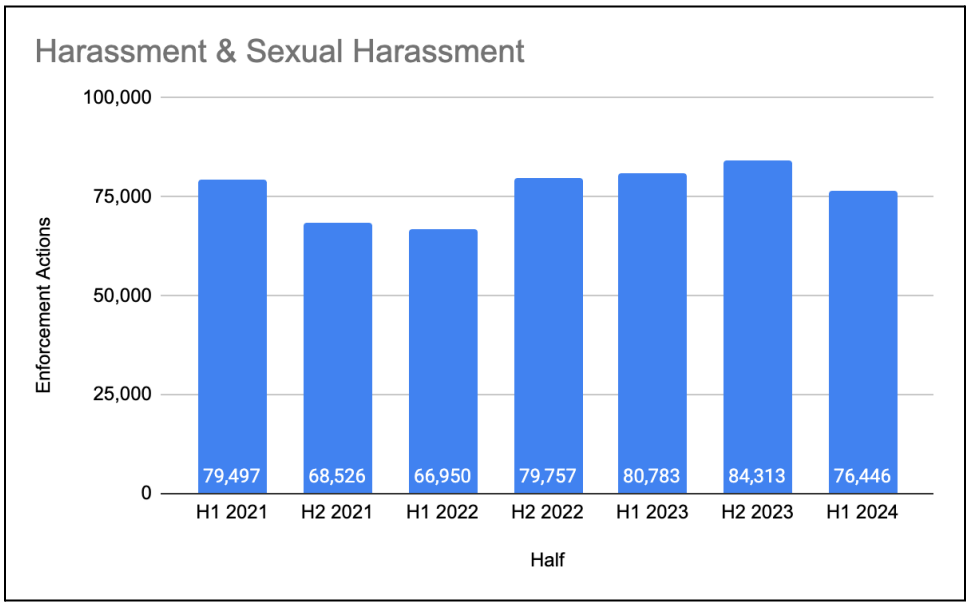
Twitch does not tolerate conduct or speech that is harassing, or that encourages or incites others to engage in harassing conduct.

While Twitch strives to create a safe experience across the site, we also empower streamers to define the personality and norms of their own community, and provide them tools to create a positive and welcoming environment. Our Baseline Report outlines the entire suite of site-wide policies and processes, as well as channel-level tooling, that help prevent and reduce online bullying and harassment, including AutoMod, Channel Moderators, Blocked Terms, Phone-verified Chat, Shared Ban Info, Shield Mode, and much more. We continue to make updates to these tools; in September 2023, we updated Mod View by integrating the Shield Mode tool, a feature to protect against sudden chat abuse. This update makes it easier

for streamers and moderators to activate heightened protection measures when necessary to curb abusive behaviour. (Measures 6 & 7)

Since our 2023 report, we also expanded our Off-Service Conduct Policy to cover doxxing, which involves sharing someone's private information without their consent, and swatting, the act of making false emergency reports to provoke a police response. This update ensures that users who engage in these severe forms of harassment, even if entirely off of Twitch, can no longer remain on the service. (Measures 6, 7, 8) Additionally, in July 2024, we clarified our Sexual Harassment Policy to remove ambiguity, making it easier for users to understand what behaviors are prohibited.

Twitch issued 84,313 harassment-related enforcements globally in H1 2023, and 76,446 in H1 2024. We issued 338 harassment-related enforcements based on reports from users in New Zealand in H1 2023, and 784 based on New Zealand user reports in H1 2024.



(Twitch Safety Center, H1 2024 Transparency Report)

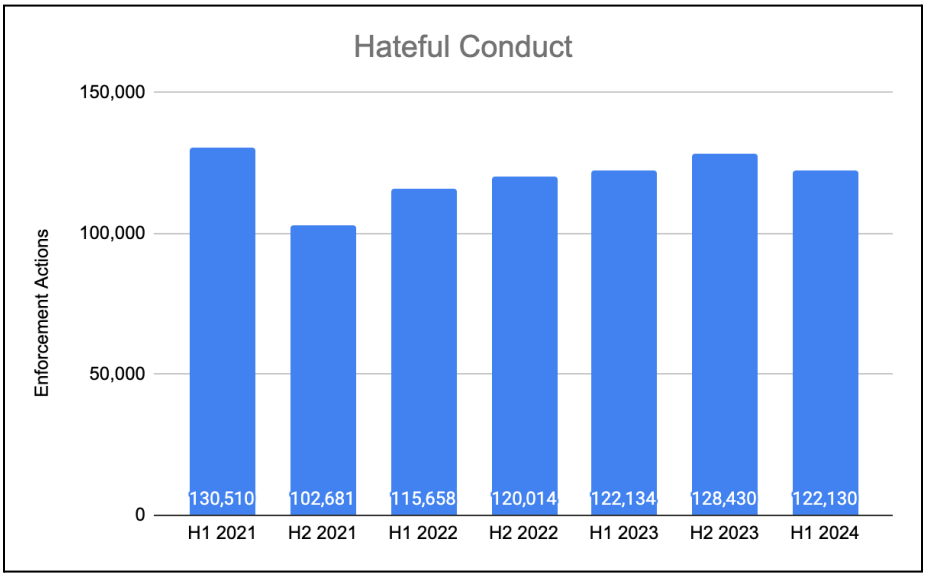
Outcome 3: Provide safeguards to reduce the risk of harm arising from online **hate speech**

Twitch has a zero-tolerance policy for hateful conduct, meaning that we act on every reported instance of hateful conduct that violates our policy. Twitch does not permit behaviour that is motivated by hatred, prejudice or intolerance, including behaviour that promotes or encourages discrimination, denigration, harassment, or violence based on the following protected characteristics: race, ethnicity, colour, 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. We afford every user equal protection under this policy, regardless of their particular characteristics.

Since our last report, Twitch has continued to invest in safeguards that reduce the risk of harm from online hate speech through several key updates. We've strengthened our Suspension Evasion Policy, making it harder for users suspended for hate speech to return

under new accounts, which reduces the likelihood of recurring harmful behaviour. (Measure 10) We've also introduced Follower Verification, adding another layer of protection to prevent malicious actors, including those who spread hate speech, from easily engaging with streamers and their communities. (Measure 11)

Twitch issued 128,430 enforcements globally for hateful conduct in H2 2023, and 122,130 enforcements in H1 2024. We issued 497 enforcements based on reports from users in New Zealand for hateful conduct in H2 2023, and 449 enforcements based on New Zealand user reports in H1 2024.

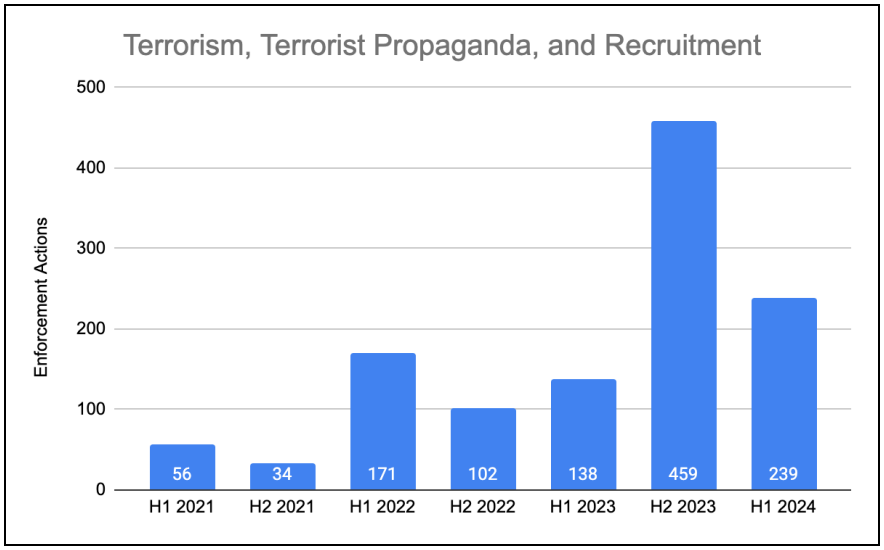


(Twitch Safety Center, H1 2024 Transparency Report)

Outcome 4: Provide safeguards to reduce the risk of harm arising from online **incitement of violence**

As outlined in our Baseline Report, Twitch prohibits the incitement of violence, including threatening to or encouraging others to commit acts that would result in serious physical harm to groups of people or significant property destruction. Additionally, Twitch's Terrorism and Violent Extremism Policy prohibits content that depicts, glorifies, encourages, or supports terrorism, or violent extremist actors or acts, and users may not display or link to terrorist or extremist propaganda, including graphic pictures or footage of terrorist or extremist violence, even for the purposes of denouncing such content. (Measure 15)

Terrorism and violent extremist-related enforcements decreased globally from 459 in H2 2023 to 239 in H1 2024, as we saw reduced impact of the Israel-Hamas conflict. We did not have extremist-related enforcements in New Zealand for H2 2023 or H1 2024.



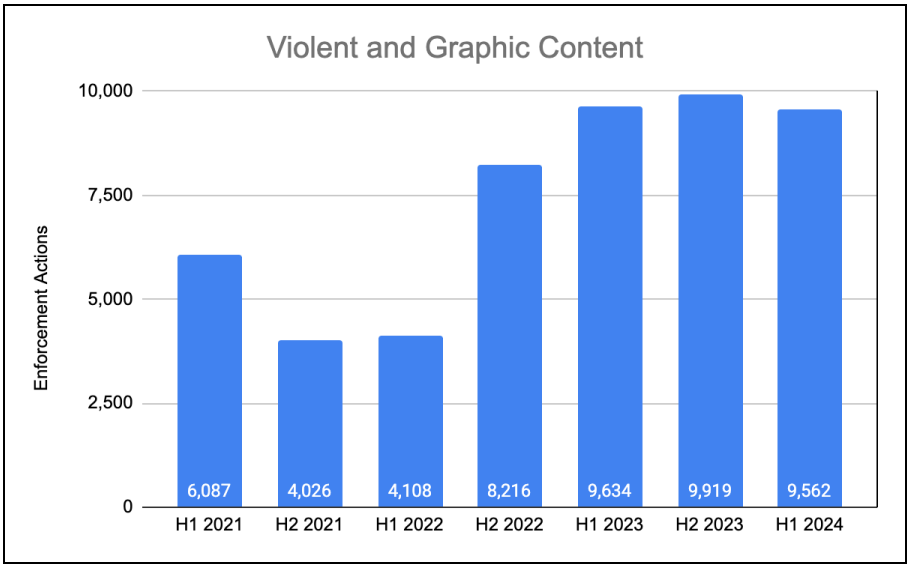
(Twitch Safety Center, H1 2024 Transparency Report)

Twitch continues to engage with industry, policymakers, academia, and civil society to combat terrorist content and violent extremism as a member of the European Union Internet Forum (EUIF). The EUIF regularly hosts tabletop exercises and workshops to facilitate information sharing and to test relevant notification and response protocols. (Measure 19)

Outcome 5: Provide safeguards to reduce the risk of harm arising from online **violent or graphic content**

Twitch has a zero-tolerance policy for acts and threats of violence. This includes, but is not limited to: attempts or threats to physically harm or kill others; attempts or threats to hack, DDOS, or SWAT others; and use of weapons to physically threaten, intimidate, harm, or kill others. Additionally, content that includes extreme or gratuitous gore and violence is prohibited.

Twitch issued 9,919 enforcements globally for violent and graphic content in H2 2023, and 9,562 enforcements in H1 2024. We issued 50 enforcements based on reports from users in New Zealand for violent and graphic content in H2 2023, followed by 62 enforcements based on New Zealand user reports in H1 2024.



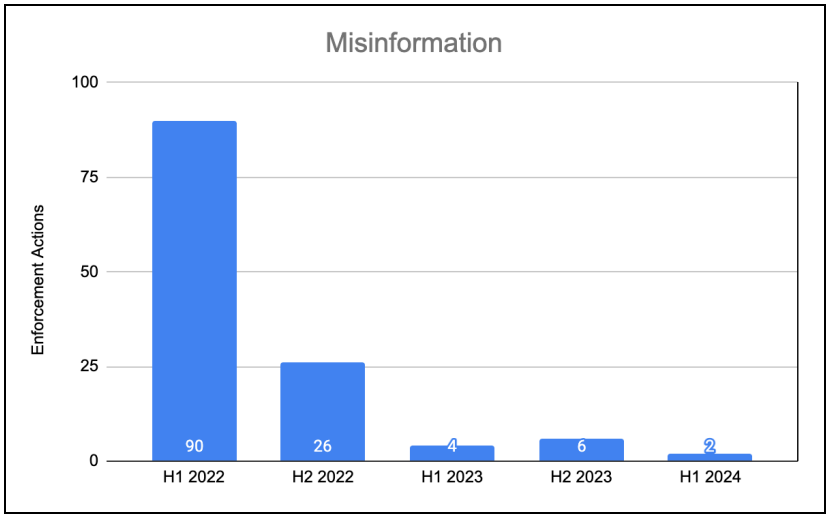
(Twitch Safety Center, H1 2024 Transparency Report)

As highlighted under Outcome 2, the expansion of our Off-Service Conduct Policy to include doxxing and swatting helps reduce the risk of these dangerous actions escalating into real-world violence, further safeguarding our community both on and off the service. By holding users accountable for their conduct beyond Twitch, we aim to create a more secure environment for all. (Measure 20)

Outcome 6: Provide safeguards to reduce the risk of harm arising from online **misinformation**

Misinformation is less prevalent on Twitch compared to other services as the mechanics of Twitch are not conducive to spreading misinformation. Most Twitch content is long form, which means it takes a lot of time (usually multiple hours) to do a live stream, especially relative to creating a post. It is extremely difficult for a new streamer to garner large numbers of concurrent viewers; it takes time to grow an audience on Twitch. The vast majority of Twitch content is also 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 user generated content (UGC) video and social media services.

Even though mis/disinformation is not prevalent on Twitch, we recognized the harm that this content could cause, particularly when it is related to an election, which is why we preemptively launched a misinformation policy in March 2022. As a result of the factors above, enforcement actions under this policy remain relatively low. Twitch issued 6 enforcement actions for misinformation in H2 2023, and just 2 in H1 2024. We did not have misinformation-related enforcement actions based on reports from New Zealand users in H2 2023 or H1 2024.



(Twitch Safety Center, H1 2024 Transparency Report)

In preparation for the record number of elections worldwide in 2024, at the start of the year, we assembled an internal cross-functional working group—across the Product, Policy, Operations, Legal, Risk Management, and Content teams—to ensure Twitch was prepared and could stay ahead of potential election-related harms. This group is empowered to conduct research, advise on policy and process, and evaluate existing measures for effectiveness.

Twitch also continues to engage in knowledge-sharing initiatives with industry and civil society through reporting under the EU Code of Practice on Disinformation and the Australian Voluntary Code of Practice on Disinformation and Misinformation (ACPDM). Additionally, Twitch participates in the EU Hate Speech Code, the EU Internet Forum, and the Global Internet Forum to Counter Terrorism (GIFCT). Through these initiatives, we learn more about malicious usage patterns and new tactics that the industry is seeing arise from misinformation actors, and then use this information to inform the continuous review of our policy and operational guidance. (Measure 27)

Outcome 7: Provide safeguards to reduce the risk of harm arising from online **disinformation**

Twitch does not make a distinction between misinformation and disinformation, and the mechanics of Twitch make it extremely difficult to invest in large-scale disinformation campaigns. Please refer to our Baseline Report and Outcome 6 above for more information on Twitch's misinformation policy.

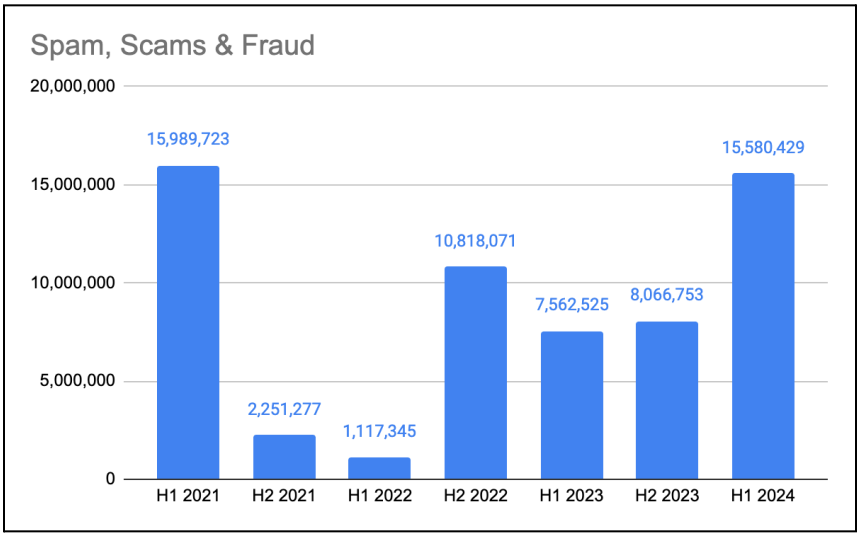
To ensure the integrity of the service, Twitch also prohibits "Any content or activity that disrupts, interrupts, harms, or otherwise violates the integrity of Twitch services or another user's experience or devices." This includes the creation of inauthentic and malicious bots, impersonation, engaging in viewership tampering (such as artificially inflating follow or live viewer stats), and selling or sharing user accounts, services, or features. (Measures 28, 33)

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.

Twitch issued 8.1M account enforcements for spam, scams, and fraud globally in H2 2023, and 15.6M in H1 2024. We issued 8 account enforcements for spam, scams and fraud based on reports from users in New Zealand in H2 2023, and another 8 enforcements based on reports from users in New Zealand in H1 2024. (The majority of Twitch's spam-related enforcements are based on proactive detection, not user reports, which is why the New Zealand-specific numbers are a small fraction of the global enforcement rates.) 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 general industry trends.



(Twitch Safety Center, H1 2024 Transparency Report)

4.2 Empower users to have more control and make informed choices

Signatories recognise that users have different needs, tolerances, and sensitivities that inform their experiences and interactions online. Content or behaviour that may be appropriate for some will not be appropriate for others, and a single baseline may not adequately satisfy or protect all users. Signatories will therefore empower users to have control and to make informed choices over the content they see and/or their experiences and interactions online. Signatories will also provide tools, programs, resources and/or services that will help users stay safe online.

Outcome 8. Users are empowered to **make informed decisions** about the content they see on the platform
Over the past year, Twitch has introduced several updates to give users more control over the content they encounter, empowering them to make informed decisions. Following the launch of Content Classification Labels (CCLs) in June 2023, we have continued to build on this feature by adding the ability to filter content from a CCL category. Introduced in May 2024, Content Display Preferences allows users to filter out content labelled with sensitive or explicit tags (Sexual Themes, Drugs Intoxication or Excessive Tobacco Use, Gambling, Violent and Graphic Depictions, Significant Profanity or Vulgarity, and Mature-Rated Games). The Content Display Preferences update also gives users the ability to blur thumbnails for content flagged with Sexual Themes CCLs. These feature updates help give users more autonomy over their viewing experience. (Measure 34) Recent research conducted by Ofcom, the UK online safety regulator, utilised the Twitch API to evaluate the impact of Twitch's CCLs and found significant improvements in content labelling accuracy. The probability of mature streams being correctly labelled increased substantially, helping users better understand the content before viewing it. The research supports the effectiveness of self-attested CCLs in enhancing user awareness and safety.
Outcome 9. Users are **empowered with control** over the content they see and/or their experiences and interactions online
As mentioned under Outcome 2, Twitch provides a number of tools to allow streamers to customise their channel and define the personality and norms of their community. Read more about channel-level customization tools. We have made a number of updates to these tools in the past year. In June, we launched the Chat Warnings feature, which gives streamers and moderators the ability to issue warnings to users before resorting to more severe actions like timeouts or bans. This approach empowers both users and moderators, encouraging self-correction and maintaining community harmony with a lighter touch. (Measure 37) Additionally, the introduction of Shared Mod Comments enables moderators to collaborate more effectively by sharing information, leading to more consistent and informed moderator decisions. In October 2023, we introduced our AutoMod Smart Detection tool in English, and by August 2024, we expanded its capabilities to include 13 additional languages. Smart Detection enhances AutoMod by learning from the unique moderation actions taken in each channel, allowing it to adapt to the specific preferences of streamers and their moderators. This feature detects and filters unwanted messages, including spam, more accurately by tailoring its behaviour based on the moderation patterns in each channel. (Measure 37) Finally, with the AutoMod testing feature, mods can now test specific messages against the moderation tool, giving them more control over how their AutoMod settings respond in different scenarios.

4.3 Enhance transparency of policies, processes and systems

Transparency helps build trust and facilitates accountability. Signatories will provide transparency of their policies, processes and systems for online safety and content moderation and their effectiveness to mitigate risks to users. Signatories, however, recognise that there is a

need to balance public transparency of measures taken under the Code with risks that may outweigh the benefit of transparency, such as protecting people's privacy, protecting trade secrets and not providing threat actors with information that may expose how they may circumvent or bypass enforcement protocols or systems.

Outcome 10. Transparency of policies, systems, processes and programs that aim to reduce the risk of online harms
This past year we have taken meaningful steps to clarify and be more transparent about how we enforce our rules. We introduced a clearer path to reinstatement, updated our appeals eligibility from 60 to 180 days, and launched education courses we believe will help prevent repeat violations of our policies. In July 2024, Twitch added four courses to the Safety Education Program (SEPs) for our partners for violations of our Gambling, Sexual Content, Sexual Harassment, and General Harassment policies. We have also introduced Enforcement Notes, which offer additional clarifications and examples within our Community Guidelines to make Twitch's rules easier to follow. For instance, under the Harassment policy, enforcement notes specify that repeated insults targeting an individual's identity will result in immediate suspension, providing clearer guidance on what constitutes actionable behaviour. Similarly, under the Spam, Scams, and Other Malicious Content policy, enforcement notes indicate that using automated bots to promote external websites or products during streams is prohibited and may result in immediate suspension. The goal is to outline how rules apply to trends we see on the service and eliminate community confusion.
Outcome 11. Publication of regular **transparency reports** on efforts to reduce the spread and prevalence of harmful content and related KPIs/metrics
Twitch has launched a new transparency report format designed to enhance the visibility of our efforts to reduce harmful content on the service. The new format aims to put the data front and centre, with a clear analysis of recent trends and helpful context around any significant changes relative to the previous report. New metrics include enforcements identified through proactive detection models, more granular categories such as hateful conduct and suspension evasion, accepted appeals by category and the total number of user reports in the 20 countries with the largest volume of reports to give a sense of geographical distribution. By regularly publishing these comprehensive transparency reports, Twitch aims to provide users with clear insights into our efforts and metrics related to harmful content, reinforcing our commitment to fostering a safer online environment.

4.4 Support independent research and evaluation

Independent local, regional or global research by academics and other experts to understand the impact of safety interventions and harmful content on society, as well as research on new content moderation and other technologies that may enhance safety and reduce harmful content online, are important for continuous improvement of safeguarding the digital ecosystem. Signatories will seek to support or participate in these research efforts.

Signatories may also seek to support independent evaluation of the systems, policies and processes they have implemented under the commitments of the Code. This may include

broader initiatives undertaken at the regional or global level, such as independent evaluations of Signatories' systems.

Outcome 12. Independent research to understand the impact of safety interventions and harmful content on society and/or research on new technologies to enhance safety or reduce harmful content online.
Twitch is committed to using data-driven research to enhance safety on its service. Through internal experiments, like the "nudges" study we conducted in August 2024, Twitch assesses the impact of safety interventions on reducing toxic behaviour. In this study, nudges—messages prompting users before posting potentially harmful content—led to a 9.26% reduction in toxic chat messages without significantly affecting user engagement. Most users (98.23%) who received nudges returned to chat within 48 hours. Currently, nudges remain active for 10% of users as we refine our findings. Twitch also provides open access to its API, enabling retrieval of most publicly available channel information, such as content classification labels (*see reference to Ofcom study under Outcome 8*), stream tags, and moderation settings. Accessing non-public data or making channel changes requires the channel owner or a user with the appropriate role on the channel to give permission. The API is available globally. Users must create a client ID, accept the Developer Services Agreement, and obtain approval for research use as determined case-by-case. Finally, our collaboration with organisations such as the Digital Trust and Safety Partnership (DTSP), Tech Coalition, and the Global Internet Forum for Counter Terrorism (GIFCT), among others, allows us to exchange perspectives on safety interventions and enhance our understanding of best practices.
Outcome 13. Support independent evaluation of the systems, policies and processes that have been implemented in relation to the Code.
Twitch remains committed to submitting annual compliance reports, in addition to our twice-a-year global transparency report, and working with the selected independent third-party organisation to review the report.