Aotearoa New Zealand Code of Practice for Online Safety and Harms

Annual Update Report – September 2023

Signatory:	**Twitch Interactive, Inc.** Twitch is a service for sharing live, interactive long-form video content. Streamers on Twitch primarily stream themselves playing video games, but a small segment of streamers broadcast in a number of different categories of 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)**
As detailed in our Baseline Report, Twitch prohibits and has zero tolerance for any content or activity that features or promotes illegal child sexual abuse material, child sexual exploitation, grooming, or other child sexual misconduct. The consequence is immediate and indefinite suspension. (Measures 1-4) Preventing and combating any form of CSEA continues to be a top priority for Twitch. Over the past year, we made further organizational changes to increase the number of staff who are able to respond to child safety escalations, monitor incoming volumes, and identify new and yet to be discovered trends that seek to harm children. We also made targeted improvements to our internal tools, which helped streamline the process of investigating, verifying and reporting CSEA, reducing CyberTip reporting time by approximately 50%. Twitch audited its internal guidance on identifying and handling CSEA content and, in March of 2023, launched its overhauled enforcement guidelines. This update makes it easier for the reviewing teams to identify content that presents the highest risk of harm, as well as evolving forms of exploitative content like generative AI-enabled CSAM/CSEA. In addition, we have expanded the policy coverage to include precursor behaviors, such as instructing minors to loosen their safety settings or asking them to move the conversation to platforms with less oversight or moderation features. (Measure 4) In conjunction with these policy updates,

Twitch iterated on our machine learning mechanisms to identify and remove users under the age of 13 and users who have been previously banned and created new accounts on Twitch.

Finally, in addition to a regular cadence of industry collaboration through the Technology Coalition, Thorn, and InHope, earlier this year Twitch presented at the Crimes Against Children Conference in Dallas, Texas on how to combat grooming in live streaming. (Measure 5).

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

Twitch prohibits harassment. We do this for many reasons, including: harassment deters the growth of vibrant and diverse communities, it prevents people from feeling safe on Twitch, it leads to streamer burn-out, and it creates a gateway for more severe forms of harm and abuse. Our Baseline Report outlines the entire suite of policies, processes and 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.

Our investment in combating online harassment does not stop there. For example, we have been investing in expanding the Off-Service Misconduct policy to cover doxxing and swatting that happen off our service. This would mean that if a Twitch user is confirmed to have engaged in severe doxxing or swatting (including threats of swatting), Twitch would suspend their Twitch account even if the incident happened on another service. The updates are planned to be launched later in 2023. (Measure 6, 7, 8)

In December 2022, Twitch partnered with the Cyberbullying Research Center (CRC) to understand the user experiences of harassment across different demographics such as countries, gender, age, and ethnicity. Throughout this year, Twitch and the CRC worked to develop a survey of over 40,000 users in the US, France, Japan - we are currently in the process of analyzing the data. The final report will provide the direction of policy updates that will be made in 2024. (Measure 9)

Twitch is also partnering with the Connected Learning Lab (CLL) from the University of California, Irvine, to study pro-social behavior (i.e. positive or “good” online behaviors). As part of this collaboration, the CLL research team is conducting a study of Twitch, industry peers, and academic research to provide tailored recommendations on how to build a pro-social environment online. As part of this study, Twitch has held workshop sessions, given expert interviews, and shared internal knowledge of Twitch systems. This research will conclude in December 2023.

In H1 2023, Twitch received 1.04M harassment-related reports, and issued 81K harassment-related enforcements.

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. 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. We afford every user equal protection under this policy, regardless of their particular characteristics.

Our range of proactive and reactive measures are described in detail in our Baseline Report. In it, we also highlight the importance of user reporting and the way in which our professional moderation team reviews and enforces reports that indicate violations of our Community Guidelines. The speed at which we can respond to user reports is particularly critical given the live nature of Twitch. However, we must balance speed with quality and accuracy of reviews, which is why we prioritize having a human in the loop for the review process to ensure it is accurate and fair for our community members. All content moderation professionals at Twitch undergo rigorous and extensive training, both prior to starting work and continuous learning throughout their tenure to address areas for improvement and stay on top of updates to our policies.

Twitch resolves the vast majority of reports within 24 hours, with 99.83% handled in this time period. The very small percentage of cases where Twitch takes longer than 24 hours to respond to a report are what we consider “Investigations" - where we have confirmed that no immediate harm is occurring but our content moderation teams are taking the time to thoroughly investigate and ensure we are taking the right actions for our community. The vast majority of reports are resolved much more quickly, with 9 out of every 10 reports resolved within an hour. In H1 2023 we have continued our investments in infrastructure and operations that are focused on response timelines (10 minutes) as we understand this is critical for a live-streaming service.

For hateful conduct specifically, Twitch received 1.42M reports and issued 122K enforcement actions H1 2023.

Twitch also added new terms to our internal Hateful Slurs list, Currently, Twitch is in the process of reviewing the English, Spanish, Japanese, and Chinese Slurs List to ensure that the list includes appropriate information and is up to date. At least three additional languages will be reviewed by December 2023. (Measure 10)

Finally, Twitch will be expanding the reach of its training program designed to counteract hate speech - developed with the Anti-Defamation League (ADL). By November 2023, all partnered streamers who violate our hateful conduct policy will be required to complete this training program. (Measure 13)

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

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. Twitch also does not allow 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.

In November 2022, Twitch implemented internal updates aimed at curtailing glorification of tragedies. Twitch also increased its cooperation with GIFCT as an official member and continues to receive information through GIFCT on violent or tragic events for our internal teams to monitor. (Measure 15). Additionally, Twitch plays an active role in the Extremism and Gaming Research Network to better understand emerging trends related to terrorism, extremism, and radicalization in gaming. (Measure 19).

In H1 2023, Twitch received 501K user reports related to violence, gore, or threats, and 180K reports related to terrorism, terrorist propaganda, or recruitment. In the same time period, Twitch issued 7.6K enforcements related to violence, gore, or threats, and 138 enforcements related to terrorism, terrorist propaganda, or recruitment.

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.

To better inform users of the content that they are consuming, Twitch launched Content Classification Labels (CCLs) in June 2023 for streamers and viewers (s*ee details below*). Streamers are now required to apply CCLs before they stream, to inform viewers that the creator's stream will contain mature themes related to: Mature-rated games; Sexual themes; Drugs, Intoxication, or Excessive Tobacco Use; Violent and Graphic Depictions; Significant Profanity or Vulgarity; and/or Gambling.

We now require fictional realistic depictions of violence in media or video games that feature extreme blood and gore to be labeled for viewers. Twitch still prohibits acts and threats of real-world violence, but labeling streams that include fictional blood and gore provides additional information that could be relevant for viewers who may not be comfortable with that type of content. More detailed guidelines can be found here. (Measure 20, 21, 22)

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

As described in the Baseline Report, we launched our first dedicated Harmful Misinformation Actor Policy in 2022 (Measure 23-24). In H2 2022 and H1 2023, we issued a combined 30 enforcements against harmful misinformation actors. Twitch misinformation 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. Most Twitch content is uniquely long-form and ephemeral – not optimized for virality. This means that most content is gone the moment it is created, so it is not shared and does not go viral in the same way that it does on other UGC video-streaming and social media services; (ii) our targeted policy only applies to those who persistently share harmful misinformation topics. This makes sense for Twitch—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 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 intentional misinformation actors and we have not seen large numbers of misinformation actors attempt to join our service.

Twitch also continues to be a signatory of, and contributor to, the EU Code of Practice on Disinformation and some of its subgroups, in order to keep abreast of misinformation trends and threats. And we invested in a livestream focused specifically on media literacy with non-profit MediaWise, a part of the Poynter institute, covering deepfakes, how to search for facts, and misinformation indicators (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. Please refer to our Baseline Report and Outcome 6 above for more information on Twitch's misinformation policy.

Besides our ongoing investments in approaches and tooling to combat inauthentic or bot-like behaviour and spam, Twitch continues to engage with the Global Disinformation Index (GDI) for analyses, reports, and briefings on the latest misinformation trends that may affect Twitch - for example, the integrity of elections and other civic processes, far-right extremism, antisemitism, anti-LGBTQIA+, Ukraine-Russia, and tragic events (e.g. mass shootings). GDI provides Twitch with actionable insights to quickly and confidently remove misinformation actors from our service if they are found to be in violation of our misinformation actor policy. (Measure 29, 33)

While Twitch continues to prohibit political advertising (Measure 31), earlier in the year we launched a new branded content disclosure tool that further increases the transparency of ads and sponsorships on our service (*see more details below under Outcome 9*).

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 behavior 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

To give viewers more control over the content they watch, Twitch launched Content Classification Labels (CCL) in June 2023. These labels replaced the binary Mature Content toggle previously offered, and must be applied by the streamer before streaming any content that includes:

- Mature-Rated Games,
- Sexual Themes,
- Drugs, Intoxication, or Excessive Tobacco Use,
- Violent and Graphic Depictions,
- Significant Profanity or Vulgarity, or
- Gambling

When one or more of these labels is applied, an interstitial is shown to viewers informing them of the type(s) of mature content present in the stream. Once the viewer provides consent, they can watch as they normally would. If they do not consent, the stream will not be shown.

To promote accurate application of these labels, Twitch also launched Content Classification Guidelines, which provide streamers with detailed descriptions and examples of content requiring a label. If a streamer fails to accurately label their content, they receive a warning via email. Repeated failure to utilize labels results in the label being automatically applied to the creator's streams for a set duration of time. (Measure 34)

In the first half of 2023, Twitch updated two policies to address sharing of synthetic and non-synthetic Non-Consensual Exploitative Images (NCEI), and partnered with the UK Revenge Porn Hotline to produce a livestream seeking to educate community members about their rights, abuse prevention, and resources available to victims. (Measure 36)

Finally, Twitch rolled out a self-service tool that allows users to request a customized report of their account and site usage data on Twitch. The tool can be accessed from users' Security & Privacy Settings. (Measure 36)

Outcome 9. Users are **empowered with control** over the content they see and/or their experiences and interactions online

Twitch's Batch Reporting Tool was launched to make it easier for streamers and moderators to submit reports to Twitch. Using this tool, they can see a list of recently banned or timed-out users and submit reports to Twitch in bulk. This feature was built in response to feedback that a primary reason for not submitting reports is that it is too difficult to submit reports while a stream is live. This feature allows streamers and moderators to focus on their roles during the stream, and easily revisit reporting when they have a break. (Measure 37)

In May 2023, Twitch launched a new Branded Content Disclosure tool to make it easy and straightforward for streamers to inform viewers that a stream includes branded content. When enabled, an automated disclosure message appears on the stream, and reappears once an hour. This standardizes the way viewers are informed of these commercial relationships, which increases clarity and transparency about why that content appears on stream. (Measure 38)

Twitch also updated its guidance on the types of products and services that may be promoted on Twitch. Some products and services aren't suitable for promotion on Twitch because they pose an unacceptable amount or type of risk to members of our community. We do not allow promotion of products and services prohibited by our Community Guidelines such as:

- Hateful Products or Services
- Illegal Products and Services
- Risky Gambling Products
- Unauthorized Sharing of Private Information
- Spam, Scams, and Other Malicious Conduct

Additionally, we do not allow promotion of the following products and services:

- Weapons
- Adult-oriented products or services
- Tobacco and tobacco related products
- Certain financial products and services
- Medical facilities and products
- Political content
- Cannabis-related products (Measure 38)

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
We continuously iterate and improve our Community Guidelines, which are easily accessible for all users to read. For many of our policies, we also include real examples of the type of prohibited behavior so users can better understand what our policies look like in practice. For our more complex policies, we include FAQs that provide even more detail. (Measure 39) Additionally, we regularly update and publish content to our Safety Center. We have blog posts, educational articles, guidelines, and additional resources for our users who are interested in how we enforce, update, and think about our policies. Users can also read about our approach to safety at the service, channel, and viewer levels on the Safety Center. (Measure 40)
Outcome 11. Publication of regular **transparency reports** on efforts to reduce the spread and prevalence of harmful content and related KPIs/metrics
Twitch continues to publish its Transparency Report twice a year. We detail policy changes, product improvements, and how we remain compliant with international regulation. Additionally, we publish separate reports for the Global Alliance for Responsible Media

(GARM, a cross-industry initiative established by the World Federation of Advertisers to address the challenge of harmful content on digital media platforms and its monetization via advertising) and NetzDG (the German Network Enforcement Act, which aims to combat online hate speech and misinformation), which are both publicly available. (Measure 41). At the end of last year, we also added Twitch's first transparency report under the EU's Terrorist Content Online Regulation, which presents information about actions taken in relation to the identification and removal of terrorist content.

In terms of metrics and KPIs, we share our total number of enforcements, reports, and law enforcement requests. These are further broken down by policy areas, per thousand hours watched, and compared to the previous half-year. And as mentioned above, we now provide specific information about response times for user reports. To view our latest Transparency Report, please follow this link.

Twitch remains committed to submitting annual compliance reports to the NZ Code Administrator, laying out the progress made in relation to our commitments under the code. (Measure 42).

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.
In addition to our new research partnerships with the Cyberbullying Research Center (CRC) and the Connected Learning Lab (CLL) from the University of California, Irvine, as described above, Twitch completed its independent Human Rights Impact Assessment (HRIA), which was managed by BSR, a non-affiliated, third-party global sustainable business network and consultancy that works with over 300 companies ranging from tech and entertainment companies to Fortune 500, multinational corporations. The full assessment, which used methodologies based upon the UN Guiding Principles on Business and Human Rights (UNGPs), including a consideration of the various human rights principles, standards, and methodologies upon which the UNGPs were built, can be found here. Twitch presented its own action plan in response to the recommendations, which can be found here.

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 organization to review the report.