Independent Review:

Aotearoa New Zealand Code of Practice for Online Safety and Harms Transparency Reports

January 2024

Dr. Philippa Smith

TABLE OF CONTENTS

1. INTRODUCTION

This is the first independent review of the transparency reports provided by signatories to the Aotearoa New Zealand Code of Practice for Online Safety and Harms in 2023. Five tech companies – Meta, Google, Twitch, TikTok and X (formerly Twitter) – voluntarily signed up to the Code when it was established in July 2022. Each signatory produced a baseline transparency report in 2022 and are subsequently required to submit annual reports with updated information so that their compliance to the commitments, outcomes, and measures – identified as relevant to their products/services in terms of reducing the prevalence of harmful content – can be evaluated.

Harmful content in the Code is classified under the following seven themes:

- child sexual exploitation and abuse
- bullying or harassment
- hate speech
- incitement of violence
- violent or graphic content
- misinformation
- disinformation

At the request of the Code administrator NZTech, the independent reviewer is required under Section K of The Code Terms of Reference to:

(a) **review** the annual compliance reports submitted by signatories.

(b) **evaluate** the level of progress made against the Commitments, Outcomes and Measures in The Code, as well as the commitments made by signatories in their participation forms (see Appendix 2 of The Code). This includes:

  - verification of claims in each report as to whether the signatories have published and implemented policies and processes that comply with their obligations
  - verification that those initiatives are accessible to Aotearoa New Zealand internet users
  - identification of any claims that cannot be attested

(c) **produce an analysis** of the signatories' reports and their progress within 90 days.

The latest transparency reports were submitted by October 2023 using a standard template (see Appendix 3 of the Code). These reports were reviewed taking into account the signatories' participant forms and their baseline reports from 2022.

The Code's four commitments are:

1. Reduce the prevalence of harmful content online
2. Empower users to have more control and make informed choices
3. Enhance transparency of policies, processes and systems
4. Support independent research and evaluation

All reports and participant forms, as well as the full list of the Code's commitments, outcomes and measures are accessible via the Code website at www.thecode.org.nz

Verification of claims involved the checking of information through links provided by signatories in their reports. In the absence of these, an online search to identify publication and implementation of policies and processes was conducted first before contacting the signatory for more information. Any claims that could not be attested were referred to the Oversight Committee.

While the reviewer's specific tasks have been outlined above, this report also includes recommendations for best practice guidelines to the Oversight Committee that may enhance future reports in terms of content and clarity. These feature at the conclusion of this report.

In presenting this report the reviewer acknowledges the Code's nine guiding principles (see pages 5-7 of the Code) that "ensure that the nature and benefits of the internet, as well as international human rights principles, best practices, and standards, are taken into account" and which include recognition of Te Tiriti o Waitangi/ Treaty of Waitangi and Value te ao Māori, namely:

- Mahi tahi | Solidarity
- Kauhanganuitanga | Balance
- Mana tangata | Dignity
- Mana | Respect

2. EVALUATION[1]

META	Facebook, Instagram
Code Commitments (opted in)	All

Review

Highlights

- Meta's 42-page report outlined a range of both new initiatives and the updating of existing processes to comply with the Code commitments, outcomes, and measures. This included multistakeholder engagement events such as the Meta Summit on Youth Safety and Wellbeing, workshops, and events, developing partnerships with NGOs and community groups, the development of a Climate InfoFinder tool and engaging with its Oversight Board for independent judgement on its actions.
- Contribution to New Zealand-specific initiatives were evident e.g., supporting local events, promoting education, and developing online safety and media literacy micro-learning modules for New Zealand schools.
- Claims were supported with url links to policy statements, and publication of announcements about its activities published on the Newsroom page of Meta's website.
- Screenshots provided useful visuals of what Meta users see on their devices when it came to new initiatives such as fact checking labels or managing recommendations.
- Both global and New Zealand-specific metrics were provided, though these were limited to the January – December 2022 period.
- Good use of case studies that demonstrated Meta's response to Co-ordinated Inauthentic Behaviour and adversarial threats.

Limitations

- Reporting period not stated.
- Length of report impacted with extraneous information affecting readability.
- Metrics confined to 2022.

Recommendations for future reports

- Expand reporting of metrics to include the most recent data available within the year of reporting.
- Add commentary about metrics to identify and explain trends e.g., impact of events or introduction of new initiatives.
- Reduce information already included in the baseline report, e.g., definitions, reference to removal of content during the pandemic, to enable greater focus on Meta's actions in the latest reporting period.

[1] Please note that these evaluations highlight some, but not all, of the initiatives of signatories to indicate the range of actions taken in the reporting year. More detailed information of initiatives can be accessed in the transparency reports available on https://thecode.org.nz

GOOGLE	YouTube
Code Commitments (opted in)	All

Review

Highlights

- Google's 20-page report detailed its activities and initiatives that align with the Code's commitments, outcomes, and measures. This included New Zealand-focused opportunities such as partnerships to run journalism digital skills training camps (one specifically for Maori cadet journalists), the launch of topical contexts in YouTube videos' information panels to help identify misinformation, and sponsorship of the APAC trusted media conference. The launch of an ads transparency centre on Google's website where users can check verifiable advertisers also included material relevant to New Zealand.
- Metrics for global enforcements relating to harmful content and violations were presented in tables showing quarterly across a one-year period from Q3 2022 to Q2 2023 (i.e., July 2022-June 2023). These tables enabled useful comparison of enforcement actions for various policies e.g., child safety, harassment, and cyberbullying and hateful or abusive content. A new initiative shifted certain violation categories into misinformation metrics e.g., impersonation and technically manipulated content.
- New Zealand metrics covered the period H1 (January-June) 2023.
- In-text url links were provided for verification.

Limitations

- Reporting period not stated.
- The wording of the timing of some initiatives was sometimes unclear. Phrases such as "we started" and users "can now access" did not make it obvious whether these occurred before or during the reporting period.

Recommendations for future reports

- Include commentary to explain enforcement metrics and any visible trends.
- Use more specific wording to identify whether initiatives described were new and occurred within the reporting period. Include dates for announcements if possible.
- Publish New Zealand metrics on Google website.

TWITCH

Code Commitments (opted in)	All except for measures 26, 31, 35, 44

Review

Highlights

- Twitch's 9-page report provided updates to its Code commitments, all of which were opted into apart from measures 26, 31, 35 and 44.
- Progress was seen in initiatives such as adding severe doxing or swatting to its Off-Service Misconduct policy (which suspends accounts even if these incidents happened on another service), the introduction of content classification label requirements for streamers and viewers and providing updated metrics for enforcements against harmful misinformation actors (H2 2022 and H1 2023).
- While Twitch had opted out of Measure 44 to run multi-stakeholder events, good efforts were still made to engage with users through educative webinars involving experts, e.g., the UK Revenge Porn Hotline, and research partnerships, e.g., Cyberbullying Research Centre and Connected Learning Lab.

Limitations

- Reporting period not stated.
- Minimal metrics were described in the report (mainly referencing H1 2023, apart from harmful misinformation as indicated above), even though Twitch's transparency report on its website provides more information with good explanation of its data and graphs.
- No New Zealand-specific metrics were included.
- More url links could have been provided to make verification easier.

Recommendations for future reports

- Specify reporting period.
- Include discussion of metrics and how these compare to previous collected data accessible on Twitch's website on the safety centre. Graphs from the website could be extracted and incorporated into the Code transparency report to enhance the clarity and comprehension of the information presented.
- Include New Zealand-specific data. If this is unavailable, this should be stated.
- Provide links for all claims in the report including metrics, livestreams, partnerships, updates etc for verification purposes rather than just referring the reader to see Twitch's Safety Center or Community Guidelines.

TIKTOK

Code Commitments (opted in)	All except for measure 31, measure 38

Review

Highlights

- TikTok's 24-page report presented information on its compliance to the Code which included a range of new or improved initiatives such as: partnering with online educators to run conference and summit workshops and/or webinars on online harm and on digital media literacy, launching a tool that informs users as to why they have been recommended a video, and working with the New Zealand Electoral Commission to deliver public service announcements. Proactive moderation during public events such as Matariki and the taking down of accounts associated with violent criminal organisations were enacted in 2023.
- Update of Community Guidelines included prohibiting use of synthetic or manipulated media that contains the likeness of any private figure, or of public figures used for endorsements or which violates its other policies.
- A new tool for academic researchers to access API information is available in the United States and Europe, with plans to introduce this to other countries in the future, including New Zealand.
- Comparative data for New Zealand was provided across two quarters - Q4 2022 and Q1 2023.
- In-text url links provided, but not for metrics.

Limitations

- Reporting period not stated.
- Enforcement statistics for violated community guidelines globally in the report limited to Q1 2023 period.
- A large amount of information in the report repeated community guidelines and processes that already appeared in TikTok's baseline report.

Recommendations for future reports

- Specify reporting period.
- Include metrics for global enforcement statistics across the other quarters of the reporting period to enable comparison. Support metrics with explanation or commentary. (Earlier metrics are accessible on TikTok's transparency newsroom webpage and can easily be extracted for this purpose.)
- The graphs from the transparency reports on TikTok's website are clear and enable useful comparisons. Some of these would be excellent inclusions in its Code transparency report.
- Make New Zealand data publicly accessible if possible.
- Consider including url links in the report to covert influence operations from their community guidelines enforcement reports within the reporting period. These would provide useful case studies to demonstrate effectiveness or any issues that occurred that impacted on guidelines.
- Focus more on progress made in the year rather than repeating information from the baseline report. Use url links to cross reference if necessary.

X

Code Commitments (opted in)	All

Review

Highlights

- New actions and initiatives tabled in X's 20-page report included introducing public access to parts of X's source code (e.g. the recommendations algorithm that controls the posts users see on the For You timeline) through GitHub in 2023, and Expanded Verification of X accounts to include New Zealand (requiring a premium subscription) where accounts can be validated.
- Identification of X's introduction of a new enforcement policy approach," Freedom of Speech, Not Reach", as part of its ownership transformation in 2023. Implementation of initiatives (some new, some evolving) included visibility filtering (restricting the reach of posts that violates its policies), and Community Notes (a form of crowd sourced moderation requiring sign up by users).
- Evidence to support publication of its commitments included links to public announcements about updates, policies, or new partnerships through X's own posts or on its website, e.g., Help Centre.
- Explanation and screenshots were usefully applied in some sections e.g., to show how users would view policies or instructions on their devices.
- Url links provided in footnotes.

Limitations

- Reporting period not stated.
- Subscription requirements for specific tiers of the new X API system available to independent researchers could have been made clearer.
- Metrics provided related to X's content moderation to each of its policies were limited to the first half of 2022 (January-June 30). Misinformation was not included, though this had featured in X's baseline report.
- No New Zealand-specific enforcement data was included, though X's baseline report for 2022 (covering July – December 2021) identified relevant New Zealand metrics for violations of their policies on child sexual exploitation, abusive behaviour, suicide and self-harm, hateful conduct, terrorism and violent extremism, sensitive media, misleading information, and impersonation.
- No mention was made of X's participation in online safety forums or summits (Measure 44). (NB. Participation has subsequently been verified.)

Recommendations for future reports

- Specify reporting period.
- Include more detailed metrics and graphs (with links to verify publication) that align with the reporting period to enable better assessment of progress with regards to the Code. Details such as the number of contributors signing up to Community Notes, including New Zealanders, and data on the average time frame it took for notes to appear on misinformation posts, for example, would aid identification of the efficacy of public response to an initiative and its value.
- Make New Zealand-specific metrics available to enable comparison with previous years.
- Given the transformation during the reporting period, clarifying what information in the baseline report was still valid or where it changed would be useful, e.g., level of human moderators.

3. Conclusion and Recommendations for Best Practice Transparency Reporting

This inaugural review of signatory transparency reports marks an important step following the introduction of the Aoteraoa New Zealand Code or Practice for Online Safety and Harms. As these are the first reports to be submitted by the five signatories following their 2022 baseline reports, it provides an opportunity to assess their progress a year on. In addition, the review aims to provide constructive feedback that identifies areas for improvement in reporting processes that will serve to ensure consistency and clarity of information. Equally, this will enable more effective comparison of reports as progress is tracked over time as well as assisting the Oversight Committee in its reviews of The Code.

Overall, the signatories responded to their commitments by presenting details such as updates to their policies, processes and resources and the introduction of new initiatives. (Inclusion of screenshots and other visual material were useful particularly when it came to demonstrating what users might experience on their devices with various initiatives.) Most reports included global and New Zealand-specific metrics that related to enforcement when policies were violated. The level of detail of metrics varied with some providing enforcement rates, proactive detection rates and response times and others adding useful comparisons with earlier data and explanations of observed trends. Signatories also indicated their efforts to engage with stakeholders through seminars and webinars as well as supporting educational resources to improve users' digital literacy skills.

Evaluation of each signatory's report showed variation in the ways in which they responded to the outcomes and measures, though this often reflected the type of service provided and the size of the organisation. However, it is also acknowledged that such reporting can be a complicated process in collating a substantial amount of material into an easy-to-read report that includes commentary for public consumption. Nevertheless, there are areas in the reports that would benefit from fine tuning to enhance future production as the signatories become more familiar with requirements and expectations.

Recommendations

This report concludes with the following recommendations to the Code's Oversight Committee for best practice guidelines for future annual transparency reports.

- **Reporting period**

  The annual reporting period was not stated by each signatory and requires clarification. All reports were identified as 2023, but either stated differing months on their cover sheets – September, October, or November – or not all. It is possible that these months align with the effective date of the Signatory Participation Form. However, on page 19 of the Code it states that the "first annual report will be submitted 45 days following 12 months (365 days) from the commencement date of the Code". The Oversight Committee may wish to specify a standard annual reporting period (month to month) to avoid confusion.

- **Demonstrating Progress**

  - **Metrics**
    The inclusion of KPIs/metrics is a requirement of the Code (Outcome 11) and presents an opportunity for signatories to quantitatively demonstrate progress in their efforts to reduce the spread and prevalence of harmful content. Comparison of metrics will become more significant over time with the delivery of each annual report. In some of the 2023 transparency reports metrics provided were minimal. More detailed metrics should include both numerator

and denominator details e.g. in some cases, numerical figures, or percentages of take downs, were given without an indication of total numbers of posts or videos to provide quantitative objectivity. Accompanying meaningful analysis of the data and commentary would provide insights such as whether an increase in the amount of content removed indicated the posting of more objectionable material that was identified, or whether the platform was more effective in removing material due to changes in policy or detection. The inclusion of a few simple graphs to illustrate trends or compare data over time should be encouraged. (Some of this material is already accessible on signatories' websites and could easily be reproduced in these reports, rather than just through the provision of links.)

- **A New Zealand-specific focus**
  More attention as to how the signatories' commitments affect New Zealanders is warranted in some of the reports given that the signatories are adhering to an Aotearoa New Zealand Code of Practice. While it is acknowledged that policies and practices have a global reach, a greater emphasis that frames these in a New Zealand context would be beneficial. This might include detailing when certain products and services will be delivered to end-users in New Zealand, or relating how signatories have pro-actively responded to a New Zealand-organised event to reduce harm – noting however that some signatories did prepare to respond to issues such as proactive moderation during Matariki. In addition, more detailed New Zealand-specific metrics and commentary would assist in identifying and tracking local trends across time. Making such data publicly accessible would aid and inform government and civil society – communities, researchers, NGOs, etc – with their own planning and online safety projects. If New Zealand metrics are not collected by a signatory, this should be notified in the report. The Oversight Committee may also wish to take note of signatories' responses to international codes and regulations and request signatories to comment on how these might impact New Zealand users.

- **Links**
  Url links for verification need to be checked before final submission of reports. Some links in reports were found to be inoperable or inaccurate and alternate ways were required to verify claims. The Oversight Committee should consider giving signatories opportunities to correct invalid urls or other errors or omissions in their reports that are uploaded to the Code website.

- **Alternative links**
  Offering alternative links to relevant information other than signatories' websites for independent verification and impact should be encouraged e.g., relevant news articles or the websites of conferences they have sponsored or supported. In addition, including brief case studies in the report of an initiative's direct or long-term impact, especially in New Zealand, would add further evidence of a signatory's efforts to meet its commitment to the Code and/or how it responds to new challenges as they arise, such as detection of AI deepfakes.

- **Report Content**

Signatories used the Code template as a framework for their reports. Some reports however, were too long due to the inclusion of extraneous material involving detailed descriptions and definitions already provided in the baseline reports. As a result, an overabundance of both repetitive information and promotional language obscured the more important aspects of these reports. Content – while informative – should be kept concise, reader-friendly and relevant for public consumption. The inclusion of cross references to the baseline reports through url links would assist those readers if more detail was deemed necessary.

Reader usability of reports needs to be considered taking note that they will be accessed by the public, researchers, and journalists. A balance of text, visuals and metrics should be sought.

Signatories should reference the measure they opted into alongside the evidence for each commitment/outcome. If a signatory has not been able to comply with a specific measure during the reporting period, then an explanation should be included.

Some signatories signalled where initiatives were a ‘work in progress’ or identified as planned for the future. This is useful to get a sense of on-going commitment to the Code, but unless evidence is provided the expectation is that this will be reported on in future transparency reports.