INDEPENDENT REVIEW

Aotearoa New Zealand Code of Practice for Online Safety and Harms 2024 Compliance Reports

January 2025

Dr. Philippa Smith

TABLE OF CONTENTS

1. INTRODUCTION

This independent review evaluates the 2024 compliance reports submitted by signatories to the Aotearoa New Zealand Code of Practice for Online Safety and Harms. Established in July 2022 and administered by NZTech, the Code is designed to hold tech companies accountable for reducing online harm and improving user safety.

Five signatories—Meta, Google, Twitch, TikTok and X (formerly Twitter)—voluntarily committed to the Code's principles in its inaugural year and continue to maintain this status. These companies provided baseline reports in 2022, followed by annual compliance reports submitted in October 2023 and 2024.

The compliance reports follow a standard template (outlined in Appendix 3 of the Code) and detail signatories measures taken over a 12-month period to address four key areas:

1. Reducing the prevalence of harmful content online.
2. Empowering users to exercise greater control and make informed decisions.
3. Enhancing transparency in policies, processes, and systems.
4. Supporting independent research and evaluation.

The Code focuses on seven critical areas of online harm:

- Child sexual exploitation and abuse
- Bullying and harassment
- Hate speech
- Incitement of violence
- Violent or graphic content
- Misinformation
- Disinformation

All compliance reports are publicly accessible via the Code website, ensuring transparency and accountability.

Scope of the Reviewer's Role

The independent reviewer is required to evaluate the compliance reports, highlighting strengths and identifying any gaps in signatories' reporting of the ways in which they foster safer online environments in Aotearoa New Zealand. More specifically, as laid out in Section K of the Code's Terms of Reference, the independent reviewer's responsibilities focus on three key areas:

1. **Assessment of Compliance Reports:**
   Reviewing the annual compliance reports submitted by signatories to the Code.

2. **Evaluation of Progress:**
   Measuring the progress signatories have made against the Commitments, Outcomes, and Measures outlined in the Code, as well as any additional commitments detailed in their participation forms (see Appendix 2 of the Code).

   This includes:

   - Verifying claims regarding the publication and implementation of policies and processes in line with the Code's obligations.
   - Confirming that these initiatives are accessible to internet users in Aotearoa New Zealand.
   - Referring any claims that cannot be substantiated to the Code's Oversight Committee.

3. **Assessment of Enforcement Metrics:**
   Evaluating the progression of processes, policies, and practices since preceding reports, and determining whether the metrics provided adequately reflect enforcement actions undertaken.

In presenting this report I acknowledge the four guiding principles that are provided to signatories, the Administrator and the Oversight Committee in the implementation and management of the Code to "ensure that the nature and benefits of the internet, as well as international human rights principles, best practices, and standards, are taken into account". These principles listed below are sourced in te ao Maori and underpin the Code's recognition of the constitutional significance of Te Tiriti o Waitangi/Treaty of Waitangi to Aotearoa New Zealand.

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

Dr. Philippa Smith
Independent Reviewer

2. EVALUATIONS[1]

META	Facebook, Instagram
Code Commitments (opted in)	All
Reporting Period	July 2023–June 2024
Metrics Period	January–December 2023

Review

Meta's 2024 report outlined a wide range of new and updated safety measures, including:

- **Automatic account disabling**: Implemented to limit suspicious adults from interacting with teens, alongside updated detection methods for child safety-related terms, phrases, and emojis.
- **New protective tools**: Tested features to combat sextortion scams and introduced a nudity protection tool in Instagram DMs, which blurs images containing nudity and prompts users to reconsider sharing them.
- **Facebook feed control**: As AI is now used to deliver material on Facebook feed, users can now control what they see e.g. hide posts, customise their feed, or report content
- **Enhanced encryption**: Rolled out end-to-end encryption for personal messages and calls on Messenger and Facebook for improved safety and security.
- **New collaborations**: Established AI Alliance with IBM to foster innovation in responsible AI development. Meta became a founding member of the Lantern programme, collaborating with partners to track potential predators by sharing critical information.
- **Community partnerships in New Zealand**: Worked with child safeguarding organizations, NGOs, the Ministry of Education, and the Australian Associated Press, supporting online media literacy programs and educational tools. A comprehensive strategy to combat misinformation was implemented during the New Zealand Election.

Metrics
Global and New Zealand metrics are provided for enforcement rates, and proactive enforcement percentages, (indicating which actions were taken without user reporting). Metrics expanded this year to include:

- Global and New Zealand pieces of content receiving warning labels on the accuracy of information.
- New Zealand metrics now divided into two categories – child nudity and physical abuse, and child exploitation – under policy for safeguarding against online child exploitation and abuse.
- Metrics on the reach of the NZ election trusted information campaign through Facebook and Instagram.

Meta's metric time period does not align with its reporting period, therefore no 2024 data is presented in this report.

Recommendations:

- Minimise extraneous information to enable greater focus on initiatives and metrics.
- Incorporate more trended data across years and provide commentary on any observable fluctuations, e.g. the increase in action taken on pieces of violent and graphic content on Instagram in Oct-Dec 2023.
- Review metric time frames to align them more closely with the reporting period.

[1] To keep this review concise, only selected initiatives as examples of progress are listed. See signatories' compliance reports published on the Code website for full details.

GOOGLE	YouTube
Code Commitments (opted in)	All
Reporting Period	July 2023–June 2024
Metrics Period	July 2023–June 2024

Review

Google is consistent in the presentation of its reports providing online safety and harms updates relating to its YouTube platform. New initiatives this year include:

- **Responsible AI innovation:** Users must disclose uploading of altered/synthetic media content including AI generated. Labelling to be introduced. This also applied to verified election advertisers.
- **New media literacy resources:** Three videos on evaluating the credibility of sources and responsible sharing of information added to the "Hit Pause" video campaign.
- **Community Posts Channel Strikes**: Channels strikes limiting creator activity can now be applied to community posts that violate Community Guidelines.
- **Keeping creators safe:** Channels detected as being possibly hijacked without the creator's knowledge may be automatically set to private until authenticity is verified
- **Keeping young people safe:** Limitations applied to repeated recommendations of problematic material that may affect the well-being of young people e.g. social aggression, idealised body types.
- **Updated firearms policy:** Content showing removal of certain safety devices on firearms is now prohibited.
- **New Zealand support:** ongoing support of media literacy resources for New Zealanders; greater visibility of local third-party crisis hotline information; taking measures prior to the 2023 NZ general election that made trusted sites on New Zealand election information easily discoverable and details about election advertisements run by verified advertisers accessible.

Metrics

Google's metric tables of quarterly enforcement metrics for the removal of videos and channels that violated policies across its reporting period remain steady though some fluctuations occur. A graph of trended data since 2017 estimating the number of views of violative content before removal - updated in each report - effectively indicates ongoing progress.

Metrics relating to the removal of videos uploaded from IP addresses in New Zealand that violated You Tube policies were expanded to cover a year (divided into quartiles), rather than a six-month total as offered previously.

Recommendations

- Reflect on trends and give explanations about fluctuations in metrics between reports to provide context.
- Maintain consistency with the expanded New Zealand removal metrics covering 12 months to enhance comparisons with future reports.
- Minimise extraneous information repeated in previous reports and focus on new and updated initiatives to show progress.

TWITCH	
Code Commitments (opted in)	All except for measures 26, 31, 35, 44
Reporting Period	July 2023–June 2024
Metrics	July 2023–June 2024

Review

Twitch's 2024 compliance report saw significant improvements this year with a broader range of global metrics, the inclusion of New Zealand data, informative graphs and links for verification and further information. Key examples of Twitch's proactive safety efforts include:

- **Chat warnings feature:** Streamers and moderators can now issue warnings to 'chatters' violating policies in their communities.
- **Shared mod comments function**: Moderators allowed to share information fostering consistency in their decision-making.
- **Follower verification**: Required verification settings introduced to protect against malicious actors following a channel.
- **Content filtering tools:** Users can filter content labelled with sensitive or explicit tags; an auto Mod smart detection tool enabling detection and filtering of unwanted messages based on moderator patterns is available in 13 languages.
- **Enhanced Safety Education**: New courses in safety education and rehabilitation available to violators to prevent repeated violations of policies.
- **Election preparation:** An internal cross-functional working group worked to address potential election-related harms with forthcoming worldwide elections in 2024 by conducting research, advising on policy and processes and evaluating effectiveness of measures.
- **Researcher support:** External research supported by open access to Twitch's API was demonstrated by referencing an Ofcom study analysing changes in 2023 to the platform's content classification labelling system.

Metrics

Global enforcement metrics for policies during the reporting year were accompanied by well-presented and informative graphs showing trended enforcement data covering a three-year period 2021-2024. This usefully demonstrated progress across commitments.

New Zealand metrics included for the first time in Twitch's reports indicating the volume of enforcements based on user reports from New Zealand across two half-year periods, i.e. H2 2023 and H1 2024.

Links directed readers to Twitch's website transparency report for more detailed safety metrics, commentary on observable trends, and a range of informative graphs.

Recommendations

- Include some of the images already available on Twitch's website of safety features such as Shield Mode tool to demonstrate functionality.
- Add commentary available in Twitch's transparency reports to explain fluctuations in data and trends.

TIKTOK	
Code Commitments (opted in)	All except for measures 31and 38
Reporting Period	1 July 2023 to 30 June 2024
Metrics	1 July 2023 to 30 June 2024

Review

TikTok presented a more focused report this year, aided by cross-referencing and links that avoided unnecessary repetition of information.

Updates on several key global initiatives included:

- a dedicated webpage on online bullying prevention.
- an automatic labelling system for content identified as AI-generated.
- the expansion of hate speech and hateful behaviour policies, particularly in response to global events such as the war in the Middle East.

A clearer engagement with New Zealand was demonstrated, with various initiatives such as:

- **Educational material:** A New Zealand-specific version of the Guardian's Guide (online and hardcopy) to assist parents and guardians navigate TikTok's safety tools and controls.
- **Training sessions:** Conducted with the Department of Internal Affairs and the New Zealand Police's digital and online child exploitation teams to provide insights and foster collaborations.
- **Promotion of local help services:** Offered contact points for New Zealand users needing support or redirection for issues related to sexual abuse.
- **Election strategies:** Promotion of public service announcements during the 2023 New Zealand election directing users to official resources for election-related information.
- **Partnerships:** Collaboration with various internet safety groups and NGOs in New Zealand to support conferences and events; engaged with an external specialised language-support provider to review Māori-related content on the platform.

Metrics

- Improved visibility of New Zealand enforcement statistics using tables was noted highlighting annual total number of videos removed, percentages of removal before user report and within 24 hours of posting. Expanding the metric time period to 12 months provided more comprehensive data, but the shift to annual totals rather than quartiles as in 2023 made comparative assessment more challenging.
- Global enforcement statistics, previously published previously in the compliance reports, now made available via links to TikTok's website where useful and interesting graphs were located.
- Percentages of the number of views before content removal featured in 2023, were omitted this year.

Recommendations:

- Include global enforcement metrics, trended data and a selection of interesting graphs accessible from TikTok's website in the next report to demonstrate progress.
- Establish consistency in the presentation of categories and time frames of metrics – including New Zealand data - in future reports to aid comparison.
- Explain exclusion of data which had previously been included.

X	
Code Commitments (opted in)	All
Reporting Period	1 October 2023–30 September 2024
Metrics	January–June 2024

Review

X's 2024 report details its approach to online safety policies and processes in line with its change in ownership and focus on a *Freedom of Speech, Not Reach* philosophy. While enforcement actions can be initiated, X's overall aim is to make content less discoverable as an alternative to removal. Initiatives include:

- **Updating of policies:** Url links to X's help centre highlighted updates in March 2024 to policies on themes such as abuse and harassment, hateful conduct, and private content.
- **Community notes development:** Designated contributors can now be alerted to add notes to misleading posts or advertisements if enough users submit requests for a note on a specific post. New Zealand is one of the 197 countries that incorporates Community Note contributions.
- **New detection methods of child sexual abuse:** Videos and GIFs posted on X are evaluated for child sexual abuse material using hash matching technology. Any account (real or computer generated) that engages with child sexual exploitation is removed.
- **Biannual transparency reports:** online transparency reports will now be published twice a year rather than annually.
- **Case study examples:** Provided link to example of its crisis protocol, taking action over violative content during evolving crisis or conflict situations.

Metrics

Global metrics included:

- number of contributors who report violations or provide helpful context through features like Community Notes.
- the volume of Community Note requests received.
- number of suspensions or content removed (automated and human moderation) featured for child safety violations only, while metrics on enforcement practices (automated and human) for other policies must be accessed through a link to X's transparency report (January to June 2024).
- number of users reports of violations for individual policies.

New Zealand enforcement metrics were limited to accounts suspended and content removed mainly for violations of violent content policy.

There is variation in the metrics provided compared with the baseline and 2023 reports which makes evaluation of progress challenging. More comprehensive New Zealand data, for example, is covered in the baseline report.

Recommendations:

- Transfer more enforcement metrics from X's biannual transparency reports into the compliance report and align more fully with the reporting period.
- Include more trended data to aid comparison.
- Provide more comprehensive New Zealand data, ideally to align with the Baseline report categories.
- Reduce background and repetitive content from previous reports while offering more detail on policy updates rather than just providing urls.

3. Conclusion and Recommendations

The 2024 compliance reports demonstrate signatories' efforts to adhere to the Code's commitments by enhancing online safety and mitigating harms on their platforms. External impacts such as global events, political instability, elections, conflicts, public health crises, and the rapid spread of misinformation can exacerbate negative online behaviours that platforms need to address. However, it is important to note that any changes to online safety and harms policies, processes, or other activities involving signatories after October 2024 fall outside the scope of this review and will be addressed in the 2025 compliance reports as applicable.

In these latest reports signatories responded to the recommendations made in the last independent review, as well as guidelines provided to support the preparation of their reports, albeit to varying extents. Some notable improvements were observed, such as the inclusion of more url links to access additional information or provide verification, and visual materials - images and graphs - helped illustrate the implementation and impact of initiatives. Greater efforts to incorporate New Zealand-specific information were also evident in some instances, particularly when it came to supporting local organisations focused on online safety events or media literacy education, directing users to local support networks, and countering misinformation during the general election. Rehabilitation and training education for violators of policies, offered by some signatories, indicated attempts to foster accountability, encourage behavioural change, and reduce repeat offenses.

While no unsubstantiated claims were referred to the Code's Oversight Committee, a key concern that arose from this review is a lack of consistency in the categories and reporting periods of metrics and KPIs incorporated by some signatories across their reports since the Code's establishment in 2022. These inconsistencies limit the ability to track progress and meaningful trends. This challenge is discussed further in the Measuring Progress section below.

The following sections provide key observations on the content and progress of the 2024 compliance reports, highlighting achievements and areas requiring attention, plus offering recommendations for the preparation of future reports.

CONTENT

All signatories clearly identified their reporting periods this time, including the months covered. Evidence was presented of ongoing efforts whether updates of initiatives or introduction of new features, such as the automated detection of offending images and tropes that violate their policies. Additionally, measures to empower users, such as personal content filtering or comment notification, were reported.

While progress across annual compliance reports is incremental, changes implemented by signatories often follow a phased approach as they respond to new safety challenges. This allows companies to test, evaluate and refine the effectiveness of their initiatives and it is appreciated when signatories allude to these processes to indicate their efforts. The use of case studies or examples in some reports were also useful ways to demonstrate the impact of measures.

Signatories acknowledged their use of AI for detection purposes and demonstrated efforts to keep pace with advancing digital technologies, particularly in identification and labelling of AI-generated content. However, no mention was made of any platform's own AI generative tools or profiles and any associated risks for users.

Although New Zealand is a small market for online services compared to many other countries, its diverse internet user communities face digital harms similar to those seen elsewhere. The increase in New Zealand-specific content in the 2024 reports is a welcome development, particularly where it highlights engagement in local campaigns, the creation of educational resources to enhance media literacy, support for safety organisations and their initiatives, and the promotion of counselling hotlines and support services. Additionally, most signatories took proactive steps ahead of the 2023 New Zealand elections to combat misinformation and help citizens access reliable election-related information.

Recommendations for signatories:

- Continue to work towards achieving a balance in report content that, while concise, provides necessary detail for clarification and contextualisation, sufficient information for verification purposes, and utilises url links to direct readers to webpages or transparency reports for more in-depth information.
- Include New Zealand-specific information to demonstrate local impact.
- Ensure clarity, consistency, and accessibility of information in line with the reporting period.
- Link activities with specific Code measures to reinforce impact.
- Provide comment on the safety aspects of any AI generated features that have been introduced.

MEASURING PROGRESS

The inclusion of metrics and KPIs is a crucial tool for measuring progress, with the Code emphasising the importance of these for facilitating comparisons. The compliance report template directs signatories to "provide metrics, if any, that demonstrate efforts related to the overall outcome". While all signatories provided some level of data, this is an area requiring attention in the improvement of future reports.

1. **Range of metrics:** Most signatories provided global enforcement metrics, particularly those related to proactive measures. However, additional data that is relevant to their commitments would enhance the insights available. For example, user notification rates and subsequent enforcement actions, the speed of content removal, and the number of views before content is removed, would offer a more comprehensive understanding of the effectiveness of initiatives. Although some signatories included these metrics, there is room for greater adoption. The Code (pages 32–34) also

suggests incorporating metrics such as participation numbers in education or media/digital literacy programs, which can serve as meaningful indicators of efforts to achieve outcomes.

Encouragingly, some signatories have improved their provision of New Zealand-specific metrics this year. These reports addressed key areas such as actions taken (content removal, account suspensions, or warnings) against material originating from New Zealand IP addresses or as a result of New Zealand user reports. In some cases the percentage of proactive content detection that violated signatories' policies was presented. Despite these commendable developments, significant variation remains in the depth and scope of data provided by signatories, highlighting the need for consistency within individual reports.

2. **Trended data:** Presenting trended data across multiple years with easy-to-read graphs or tables is an effective way used by some signatories to demonstrate the impact of policies and processes over time, making the reports more transparent and sparing readers the effort of cross-referencing previous reports. Commentary that explains trends and fluctuations, the impact of significant events, or shifts in user behaviour, also assists in contextualising the data.

3. **Consistency of metrics:**
   Inconsistencies in the delivery of some metrics and KPIs across the three reports each signatory has submitted since the establishment of the Code hinders the effective tracking of progress. These inconsistencies include the omission of previously reported categories and variations in reporting periods and time frames. While some signatories provided links to their transparency reports where broader global metrics and visual data such as graphs can be found, compliance reports should aid readers by presenting key statistics and the impact of initiatives upfront. Links can still serve as a useful resource for exploring finer details.

   Any changes in metrics presented due to updates in policies, processes, data availability, or alterations to categories should be accompanied by clear explanations to ensure transparency and accountability.

Recommendations for signatories:

- Include more comprehensive metrics/KPIs that go beyond enforcement data.
- Ensure consistency in reporting metrics across content and time periods to enhance transparency and comparability and offer explanations on any notable changes.
- Explain any exclusions of previously reported metrics/KPIs or changes to the categorisation of data.
- Provide more trended data that includes data points from previous reports. Graphs are an effective way to demonstrate trends.
- Include reflection on how reported actions and outcomes contribute to or respond to broader societal changes, events, or behavioural trends.

REPORT PRESENTATION

Overall there was a noticeable effort in reducing levels of extraneous information, prioritising of concise and relevant content and incorporating links for verification or for additional resources. These actions are commendable particularly given the complexity of preparing reports that address the range of themes of harmful online content under the Code.

In some cases, however, promotional language still dominated which made it harder to distinguish between general messaging and substantive updates on safety initiatives. Ambiguity in distinguishing between existing, new, updated, or superseded initiatives occurred at times. For example, updates to policies were announced giving a link to a webpage or a post for information, but the specific details of what aspects had actually changed were sometimes unclear.

The inclusion of images of mobile or screen interfaces effectively highlighted new safety features, such as labels, notes, or warnings. In several instances, metrics were presented more clearly through well-designed graphs and tables, which were particularly useful for displaying trends in the data.

Recommendations for signatories:

- Differentiate between existing, new, updated, or replaced initiatives. Using precise verbs such as "update," "expand," and "add" can assist to clarify change.
- Remember to spell out acronyms at their first mention to ensure understanding for all readers e.g. Child Sexual Abuse (CSE) and National Center for Missing & Exploited Children (NCMEC).
- Include page numbers in reports.
- Avoid promotional language or extraneous and repetitive information.

* * * *

Concluding Comment:

The Code plays an important role in holding technology companies accountable for improving the experiences and well-being of internet users in Aotearoa New Zealand. With three compliance reports submitted since 2022, the forthcoming review of the Code is timely and presents an opportunity for dialogue between signatories and the Oversight Committee to address any challenges in reporting, particularly in response to the evolving digital landscape. Notably, inconsistencies in some reports, as highlighted in this review, need to be addressed. Key considerations should include assessing whether the reporting template remains fit for purpose and if baseline reports require revision to keep pace with changes in content and metrics.