Executive Summary

Background

For the past year, Netsafe has been leading the development of a voluntary industry code - the **Aotearoa New Zealand Code of Practice for Online Safety and Harms** - that brings industry together under a set of principles and commitments, as well as provides a best practice self-regulatory framework aimed at enhancing people's safety and reducing harmful content online.

The intention and development of the Code is encapsulated by four key Māori principles of *mahi tahi* (solidarity), *kauhanganuitanga* (balance), *mana tangata* (humanity) and *mana* (respect), which are critical to serving the diverse user communities in New Zealand and realising the purpose and aspirations of the Code.

A broad range of digital platforms - including Meta (Facebook and Instagram), Google (YouTube), TikTok, Twitch, Twitter - have been involved in the initial drafting of the Code. The Code aims to provide best practices for a broad range of products and services, serving diverse and different user communities with different use cases and concerns. As such, it provides flexibility for potential Signatories to innovate and respond to online safety and harmful content concerns in a way that best matches their risk profiles, as well as recalibrate and shift tactics in order to iterate, improve and address evolving threats online in real-time.

The Code is not intended to replace or address obligations pertaining to existing law or other voluntary regulatory frameworks but instead focuses on the Signatories' architecture of systems, policies, processes, products and tools established to reduce the spread of potentially harmful content.

Unique features of the Code

The Code is an evolution of existing industry principles and standards that aims to broaden efforts, transparency and accountability for online safety and harm. It is built on existing practices in Aotearoa New Zealand and codes of practice in other parts of the world, mainly the EU Code of Practice on Disinformation,[1] the EU Code of Conduct on Countering Illegal Hate Speech Online,[2] the Australian Code of Practice on Disinformation and Misinformation[3] and the Digital Trust & Safety Partnership Best Practice Framework.[4] Most of the digital platforms who have been involved in the development of the Code are already signatories to or members of these other codes.

The Code is unique in that it provides a governance framework that aims to enable the Administrator, a multitude of relevant stakeholders, as well as the public to hold Signatories

[1] https://digital-strategy.ec.europa.eu/en/policies/code-practice-disinformation

[2] https://ec.europa.eu/info/policies/justice-and-fundamental-rights/combatting-discrimination/racism-and-xenophobia/eu-code-conduct-countering-illegal-hate-speech-online_en

[3] https://digi.org.au/disinformation-code/

[4] https://dtspartnership.org/best-practices/

to their commitments. Although voluntary, digital platforms who become Signatories commit to being held accountable. For this purpose, the Code introduces oversight powers for the Administrator and a multi-stakeholder Oversight Committee. The Oversight Committee may recommend to the Administrator the termination of a Signatory's membership or the public naming of a Signatory for failing to meet its commitments; while the Administrator may make binding decisions A complaints mechanism, allowing users to report on Signatories' non-compliance with Code commitments will also be established.

The Oversight Committee is responsible for monitoring Signatories' progress under the Code; evaluating their annual compliance reports; and reviewing and making recommendations for improvement of the Code. The Code, therefore, is intended to be a living document in that it is required to be regularly reviewed based on feedback from stakeholders and through operational learnings.

The Administrator and Signatories will have six months after the launch of the Code to establish the governance framework. The Code aims to be broadly inclusive, providing flexibility for a diverse range of digital platforms (both large and small) to participate and allows signatories' responses to the Code to evolve over time. Additional Signatories are welcome to join the Code and participate in its future development.

The Code commits Signatories to a set of guiding principles, commitments, outcomes and measures that are focused on seven safety and harmful content themes - 1) child sexual exploitation and abuse; 2) bullying or harassment; 3) hate speech; 4) incitement of violence; 5) violent or graphic content; 6) misinformation; and 7) disinformation - which Netsafe and the Signatories believe are of concern for Aotearoa New Zealand internet users. This makes the Code much broader than other existing industry codes. It also commits signatories to provide transparency about their policies, processes and systems.

Signatories are also required to provide annual compliance reports outlining actions and measures taken in relation to their commitments under the Code, which will be made public and open for scrutiny.

The code is unique not only in the elements described above, but also in its collation into one place, existing principles and frameworks drawn from international and local sources. These, along with online safety and digital platform expertise from Netsafe and the Signatories, and Māori advisory input, provided the basis for the development of the Code.

Developing the Code

The development of the Code from conceptualisation to the publication of the first draft for public feedback, took approximately nine months. Netsafe first announced the initiative early April 2021[5] and the Code was made public for feedback in December.

A first draft, which was modeled after the EU and Australian codes, was presented by Netsafe to the core group of industry participants in late July, which initiated discussions on

[5] https://www.netsafe.org.nz/onlinesafetycodeofpractice/

the text of the Code and how it could be made more impactful and operable for New Zealand.

The seven safety and harmful content themes were prioritised based on research[6] conducted by Netsafe on content people in New Zealand have viewed that have negatively affected their lives, as well as top trends of harmful digital communication content being reported to Netsafe.

Netsafe and the industry group consulted with Māori cultural advisors on the Code to better understand a Māori perspective with respect to the framing and governance. Revisions to the Code were made to incorporate some of their recommendations.

During this time, Netsafe and industry participants had their own discussions with a broad range of stakeholders regarding the Code and brought the feedback of those discussions back to the development and drafting of the Code.

In November and early December, prior to the publication of the draft Code for public feedback, Netsafe and the industry group briefed government and civil society stakeholders . Due to initial feedback from those sessions, the public consultation was extended from six weeks to ten weeks, running from 2 December 2021 to 2 February 2022.

As a result of the extended public submission period, Netsafe received feedback and insights from a wide range of stakeholders. Feedback included a mix of supportive notes and constructive criticism. In total, we reviewed 34 unique responses and 4767 template letters endorsing one of the unique letters. This feedback has shaped a revised version of the Code.

Kia kotahi te hoe (paddling in unison)

Te Rangapu Whakatutuki - the Administrator - and Signatories have different backgrounds and operating models with a diverse range of products and services, but are united in their desire to make a meaningful contribution to online safety in Aotearoa New Zealand.

As the Māori proverb *'Kia kotahi te hoe'* goes, we all come from our own *waka*, but when we come together in the same *waka*, our paddles must work in unison.

[6] https://www.netsafe.org.nz/advice/research

Aotearoa New Zealand Code of Practice for Online Safety and Harms

Table of Contents

1. Preamble

1.1 Purpose

The purpose of the Aotearoa New Zealand Code of Practice for Online Safety and Harms ("Code") is to enhance people's safety and contribute to reducing harmful content online. The Code lays out a set of commitments that signatories agree to meet.

The Code is voluntary. However, digital platforms who become Signatories commit to being held accountable to the commitments that are relevant to their products or services.

The Administrator and Signatories have different backgrounds and operating models with a diverse range of products and services, but are united in their desire to make a meaningful contribution to online safety in Aotearoa New Zealand. We all have different paddles, but are in the same *waka*:

> ***Kia kotahi te hoe****:* We all come from our own *waka*, but when we come together in the same *waka*, our paddles must work in unison.

1.2 Background

"Digital services are increasingly central to our daily lives, facilitating social discourse, economic activity, and much more. These services provide powerful tools for users across the globe to engage in a wide range of valuable online activity. But like any tool, they can also be misused to facilitate harmful behavior and content…Given the diversity of digital services, it is important to define an overall framework and set of aims for what constitutes a responsible approach, to which digital services can then map their specific practices." -- Digital Trust & Safety Partnership[7]

Given the growing number of people who rely on the internet, it is critical that online safety is purposefully factored into the product and policy design of digital platforms to improve the experiences and wellbeing of internet users. Recent research[8] by Netsafe shows one in five adults and twice as many young people in Aotearoa New Zealand received a digital communication that negatively impacted their life in 2020. To put that in a practical context, as 2021 has progressed, Netsafe is continuing to record a new "high" in the number of reports related to harmful digital communication. Experiences like this, directly and indirectly, can cause physical, financial, and psychological harm; decrease user confidence; and undermine investment in the digital economy and society.

The rapidly evolving and global nature of the internet requires an equally flexible and agile approach to safety. Effective online safety requires leveraging specialist knowledge and partnerships, and is underpinned by transparency.

---

[7] Digital Trust & Safety Partnership (DTSP). The DTSP is an industry-led initiative focused on promoting a safer and more trustworthy internet. Members of the DTSP are committed to developing best practices and assessments in the field of Trust and Safety. Such industry initiatives have been crucial to maturing and organizing other tech disciplines like cybersecurity.

[8] https://www.netsafe.org.nz/advice/research

The Code has been developed by Netsafe -- an independent, non-profit online safety organisation, that provides online safety support, expertise and education to people in Aotearoa New Zealand -- in collaboration with industry and consultation with Māori advisors and after broad feedback with Government, civil society and the public. Supported by a broad range of digital platforms, the Code has been designed for organisations providing online services to people in Aotearoa New Zealand.

1.3 *Mahi tahi* (solidarity), *kauhanganuitanga* (balance), *mana tangata* (humanity), and *mana* (respect)

The intention and development of the Code is encapsulated by four key Māori principles of *mahi tahi* (solidarity), *kauhanganuitanga* (balance), *mana tangata* (humanity) and *mana* (respect), which are critical and necessary to realising the purpose and aspirations of the Code.

- ***Mahi tahi*** -- meaning working together, sharing responsibility, collaboration, cooperation, teamwork -- is crucial for the success of the Code. Safeguarding the digital information ecosystem for users requires whole-of-society collaboration and cooperation, globally and locally. Achieving the Code's purpose will depend on a wide range of digital platforms working together and sharing responsibility with other stakeholders in order to have positive relationships and impacts on Aotearoa New Zealand internet users.
- ***Kauhanganuitanga*** -- meaning balance -- is necessary to balance the different interests, values, needs and concerns of digital society. It also speaks to the balancing of a complex global and open internet with local conditions and concerns. To achieve this, the Code needs to enable the broad range of digital platforms to responsibly balance safety, privacy, freedom of expression and other fundamental values while also addressing safety risks in ways that are most relevant and suitable to their products, services and user communities.
- ***Mana tangata*** -- meaning showing respect, generosity and care for others -- is central to the Code. Care and safety of people should be a priority, while also respecting their right to free speech and expression. The Code focuses on enhancing the safety and protecting the rights of users, through collaboration and mutual respect between the Signatories, Administrator and other involved stakeholders. To achieve this also requires openness and transparency.
- ***Mana*** -- meaning integrity and respect -- underpins all three principles of *mahi tahi*, *kauhanganuitanga* and *mana tangata*. The relationship the Signatories and Administrator have with their user community, government, civil society and other relevant stakeholders, as well as with each other should be underpinned by integrity and respect. All potentially affected parties by the Code, directly and indirectly, should be given due consideration in efforts to enhance safety and reduce harmful content online.

1.4 Subject matter

Signatories' efforts in relation to the Code will focus on harmful content online that falls under the themes listed below, which the Administrator and the Signatories believe are of great concern for Aotearoa New Zealand internet users (see section 4 for related Commitments).

1) Child sexual exploitation and abuse
2) Cyberbullying or harassment
3) Hate speech
4) Incitement of violence
5) Violent or graphic content
6) Misinformation
7) Disinformation

The themes may be revised or updated upon agreement by all Signatories.

1.5 Signatories

Digital platforms may become Signatories to the Code and may join at any time. Digital platforms who have signed the Code recognise their role as important actors within the Aotearoa New Zealand information ecosystem and will encourage other platforms to join the Code or use it as a best practice guide.

2. Guiding Principles

The open and global nature of the internet has helped deliver tremendous economic and social benefits, facilitating innovation, boosting productivity and economic growth, and creating new social and educational opportunities around the world. A particular strength of the Internet is that it supports diversity and the open exchange of opinion, speech, information, research, debate and conversation as well as creative expression. It also enables people, who may be isolated and struggling, to connect with others. At the same time, the internet can also be used to spread harmful content.

The Guiding Principles provide a set of values to guide Signatories and the Administrator of the Code. The Principles aim 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.

2.1 Promote safety

It is important that people can discover and engage safely on digital platforms. To freely exchange ideas, collaborate and entertain, people must feel safe to express their views.

2.2 Respect freedom of speech and expression and other fundamental human rights

The internet has dramatically increased the power of people around the world to express themselves freely. Any efforts to address safety and harmful content online should respect freedom of speech and expression and other fundamental human rights.

2.3 Protect user privacy

The privacy of users should be respected. Any actions taken by digital platforms to address the propagation of harmful content online should not contravene commitments that have been made to respect the privacy of users, including in terms and conditions, published policies and voluntary codes of conduct as well as by applicable laws. This includes respect for users' expectations of privacy when using digital products and services. Additionally, any access to data for research purposes must protect user privacy and may only be used in accordance with applicable law.

2.4 Recognise the transnational or global nature of the internet

Digital platforms are transnational or global by nature and need to be able to operate on a transnational or global scale. Digital communication transcends borders, and therefore, efforts to effectively address safety and harmful content online should recognise the value of cross-border communications and the need for scalability.

2.5 Broad applicability and participation

A broad range of products and services, serving different and diverse user communities, make up the digital information ecosystem. Efforts to systemically address safety and harmful content issues across the entire digital ecosystem need to be inclusive and flexible, encouraging a variety of current and future digital technologies to participate or be guided by the Code. Digital platforms should have sufficient flexibility to innovate and respond in a way that best matches their risk profiles, as well as recalibrate and shift tactics in order to iterate, improve and address evolving threats online in real-time.

2.6 Systems-based best practice standards

Best practice standards should focus on systems, policies and processes that enable digital platforms to responsibly balance safety, privacy, freedom of expression and other fundamental values. This provides an incentive for digital platforms to invest in policies, products, tools and programs that empower users to make informed decisions and have control over their experiences and interactions online. It also provides greater flexibility for digital platforms to respond and adapt quickly and appropriately to ever-changing risks of online harm.

2.7 Proportionality and necessity

The types of user behaviours, platform abuse, safety issues and harmful content online will vary greatly in incidence, risk-level and impact amongst the diverse range of products and

services offered by digital platforms. Digital platforms should be incentivised to take action and respond in a way that is both risk-based and proportionate and necessary to the level of harm. Furthermore, the severity and prevalence of harmful content online, its status in law, and other efforts already underway (such as global standard-setting bodies and frameworks) should be taken into account.

2.8 Whole-of-society collaboration and cooperation

Addressing complex online safety challenges requires a whole-of-society effort, locally and globally. A wide range of relevant stakeholders have roles and responsibilities in dealing with online safety and harmful content, including public authorities, civil society, news organisations, political parties, academia, educators, parents/caregivers, content creators, and people who use the internet each day. Empowering users, promoting safety and reducing harmful content online is a shared responsibility that will require concerted efforts and collaboration by and among various stakeholder groups, including digital platforms. Whole-of-society collaboration helps improve understanding through transparency of the safety activities of digital platforms, which in turn increases opportunities for meaningful and effective collaboration.

3. Scope, application, commencement and termination

3.1 Relevant products and services subject to the Code

The Administrator encourages all digital platforms, including non-signatories, to adhere to the Guiding Principles and Commitments contained within this Code. Recognising that the types of user activity and content that is subject to the Code will vary greatly in incidence, risk level, and impact amongst the diverse range of services and products offered by different digital platforms, Signatories are required to specify which commitments, outcomes and measures (see section 4) -- at the company, product or service-level -- that are most relevant to them for the purposes of the Code in the ‘Signatory Participation Form’ (see Appendix 2).

This form must be submitted upon the Signatory’s sign-on date to the Code, and will include, for each measure, either an initial assessment of practices being undertaken or an explanation as to why specific measures are not being implemented.

Products and services relevant to the code are those that facilitate user-generated content (including sponsored and shared) and are delivered to end-users based in Aotearoa New Zealand. The code, however, may also apply to other digital products or services where the spread or prevalence of harmful content online, as described in section 1.4 above, are a concern.

Given the range of products and services and varying capabilities of digital platforms, Signatories may take a phased approach to its Code commitments in order to build capacity. Signatories may amend their Signatory Participation Form, provided that they provide 90 days notice to the Administrator of any amendments.

3.2 Application of existing laws

The Code focuses on the Signatories' systems, policies, processes, products and tools established to prevent or reduce the spread of potentially harmful content. The Code recognises that there are a range of laws or regulatory arrangements that already exist in Aotearoa New Zealand to address harmful content online, which may overlap with some of the commitments covered by the Code.[9] To the extent of any conflict with the Code, those laws and regulations have primacy.

3.3 Changes and amendments

Any changes or amendments to the Code must be agreed by all current Signatories and the Administrator before coming into effect.

3.4 Commencement and duration

The Code commences on 25 July 2022. The commitments made by each Signatory apply to it from the effective date of the Signatory Participation Form (see Annex 2).

3.5 Termination

The Code and Signatories' commitments will continue indefinitely, subject to the following conditions for termination:

3.5.1 Termination for non-compliance

Signatories that repeatedly fail to comply with their commitments under the Code may be terminated by the Administrator, under advisement by the Oversight Committee, as outlined in section 6.2.2 and 6.2.3.

3.5.2 Termination by withdrawal

A Signatory may withdraw from the Code or a particular commitment under the Code by 90-day written notice to the Administrator. The written notice would be published on the website with all other materials related to the Code.

The Administrator may withdraw from the Code by 90-day written notice to the Signatories and Oversight Committee and will continue its role and responsibilities, as outlined in section 6, until a new Administrator is identified and commences the role.

[9] For example, the Harmful Digital Communications Act 2015 (HDCA), the hate speech provisions of the Human Rights Act, freedom of expression under the Bill of Rights Act, and The Films, Videos, and Publications Classification Act 1993 (see table of existing laws [to be inserted]).

3.5.3 Termination for regulation

The Code may be terminated if, in the view of the Administrator and Signatories, subsequent legislation has made the Code superfluous.

3.5.4 Termination by agreement

The Code may be terminated, if agreed upon by the Administrator and Signatories, for any other reason, aside from the ones outlined above.

4. Commitments, outcomes and measures

This section outlines a set of four (4) commitments and corresponding outcomes and measures aimed at addressing concerns related to safety and the spread of harmful content online. These commitments, outcomes and measures are informed by the Guiding Principles in section 2. They focus on Signatories' architecture of systems, policies, processes, products and tools established to enhance user safety and prevent or reduce the spread of potentially harmful content.

Signatories will make best efforts in relation to the following commitments:

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

Signatories are required to specify which commitments, outcomes and measures -- at the company, product or service-level -- that are most relevant to them for the purposes of the Code in the 'Signatory Participation Form' (see Appendix 2).

This form must be submitted upon the Signatory's sign-on date to the Code, which will include, for each measure, either an initial assessment of practices being undertaken or an explanation as to why specific measures are not being implemented.

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.

	Outcomes	Measures
4.1	**Outcome 1.** Provide safeguards to reduce the risk of harm arising from online **child sexual exploitation & abuse (CSEA)**	**Measure 1.** Implement, enforce and/or maintain policies, processes, products, and/or programs that seek to prevent known child sexual abuse material from being made available to users or accessible on their platforms and services
4.2		**Measure 2.** Implement, enforce and/or maintain policies, processes, products, and/or programs that

	Outcomes	Measures
		seek to prevent search results from surfacing child sexual abuse material
4.3		**Measure 3.** Implement, enforce and/or maintain policies, processes, products, and/or programs that seek to adopt enhanced safety measures to protect children online from peers or adults seeking to engage in harmful sexual activity with children (e.g. online grooming and predatory behaviour)
4.4		**Measure 4.** Implement, enforce and/or maintain policies, processes, products, and/or programs that seek to reduce new and ongoing opportunities for the sexual abuse or exploitation of children
4.5		**Measure 5.** Work to collaborate across industry and with other relevant stakeholders to respond to evolving threats
4.6	**Outcome 2:** Provide safeguards to reduce the risk of harm arising from online **bullying or harassment**	**Measure 6.** Implement, enforce and/or maintain policies and processes that seek to reduce the risk to individuals (both minors and adults) or groups from being the target of online bullying or harassment.
4.7		**Measure 7.** Implement and maintain products and/or tools that seek to mitigate the risk of individuals or groups from being the target of online bullying or harassment.
4.8		**Measure 8.** Implement, maintain and raise awareness of product or service related policies and tools for users to report online bullying or harassment content.
4.9		**Measure 9.** Support or maintain programs, initiatives or features that seek to educate and raise awareness on how to reduce or stop online bullying or harassment.
4.10	**Outcome 3:** Provide safeguards to reduce the risk of harm arising from online **hate speech**	**Measure 10.** Implement, enforce and/or maintain policies and processes that seek to prohibit or reduce the prevalence of hate speech.
4.11		**Measure 11.** Implement and maintain products and tools that seek to prohibit or reduce the prevalence of hate speech.

	Outcomes	Measures
4.12		**Measure 12.** Implement, maintain and raise awareness of product or service related policies and tools for users to report potential hate speech.
4.13		**Measure 13.** Support or maintain programs and initiatives that seek to encourage critical thinking and educate users on how to reduce or stop the spread of online hate speech.
4.14		**Measure 14.** Work to collaborate across industry and with other relevant stakeholders to support efforts to respond to evolving harms arising from online hate speech.
4.15	**Outcome 4:** Provide safeguards to reduce the risk of harm arising from online **incitement of violence**	**Measure 15.** Implement, enforce and/or maintain policies and processes that seek to prohibit or reduce the prevalence of content that potentially incites violence.
4.16		**Measure 16.** Implement and maintain products and tools that seek to prohibit or reduce the prevalence of content that potentially incites violence.
4.17		**Measure 17.** Implement, maintain and raise awareness of product or service related policies and tools for users to report content that potentially incites violence.
4.18		**Measure 18.** Support or maintain programs and initiatives that seek to educate users on how to reduce or stop the spread of online content that incites violence.
4.19		**Measure 19.** Work to collaborate across industry and with other relevant stakeholders to support efforts to respond to evolving harms arising from online content that incites violence.
4.20	**Outcome 5:** Provide safeguards to reduce the risk of harm arising from online **violent or graphic content**	**Measure 20.** Implement, enforce and/or maintain policies and processes that seek to prohibit and/or reduce the spread of violent or graphic content online.
4.21		**Measure 21.** Implement and maintain products and tools that seek to prohibit and/or reduce the spread of violent or graphic content.

	Outcomes	Measures
4.22		**Measure 22.** Implement, maintain and raise awareness of product or service related policies and tools for users to report potential violent and graphic content.
4.23	**Outcome 6:** Provide safeguards to reduce the risk of harm arising from online **misinformation**	**Measure 23.** Implement, enforce and/or maintain policies, processes and/or products that seek to reduce the spread of online misinformation.
4.24		**Measure 24.** Implement, enforce and/or maintain policies and processes that seek to penalise users who repeatedly post or share misinformation that violates related policies.
4.25		**Measure 25.** Support or maintain media literacy programs and initiatives that seek to encourage critical thinking and educate users on how to reduce or stop the spread of misinformation.
4.26		**Measure 26.** Support or maintain programs and/or initiatives that seek to support civil society, fact-checking bodies and/or other relevant organisations working to combat misinformation.
4.27		**Measure 27.** Work to collaborate across industry and with other relevant stakeholders to support efforts to respond to evolving harms arising from misinformation.
4.28	**Outcome 7:** Provide safeguards to reduce the risk of harm arising from online **disinformation**	**Measure 28.** Implement, enforce and/or maintain policies, processes and/or products that seek to suspend, remove, disable, or penalise the use of fake accounts that are misleading, deceptive and/or may cause harm.
4.29		**Measure 29.** Implement, enforce and/or maintain policies, processes and/or products that seek to remove accounts, (including profiles, pages, handles, channels, etc) that repeatedly spread disinformation.
4.30		**Measure 30.** Implement, enforce and/or maintain policies, processes and/or products that seek to provide information on public accounts (including profiles, pages, handles, channels, etc) that empower users to make informed decisions (e.g. date a public profile was

	Outcomes	Measures
		created, date of changes to primary account information, number of followers).
4.31		**Measure 31.** Implement, enforce and/or maintain policies, processes and/or products that seek to provide transparency on paid political content (e.g. advertising or sponsored content) and give users more context and information (e.g. paid political or electoral ad labels or who paid for the ad).
4.32		**Measure 32.** Implement, enforce and/or maintain policies, processes and/or products that seek to disrupt advertising and/or reduce economic incentives for users who profit from disinformation.
4.33		**Measure 33.** Work to collaborate across industry and with other relevant stakeholders to support efforts to respond to evolving harms arising from disinformation.

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.

	Outcomes	Measures
4.34	**Outcome 8.** Users are empowered to make informed decisions about the content they see on the platform	**Measure 34.** Implement, enforce and/or maintain policies, processes, products and/or programs that helps users make more informed decisions on the content they see
4.35		**Measure 35.** Implement, enforce and/or maintain policies, processes, products and/or programs that seek to promote accurate and credible information about highly significant issues of societal importance and of relevance to the digital platform's user community (e.g. public health, climate change, elections)

4.36		**Measure 36.** Support programs and/or initiatives that educate or raise awareness on disinformation, misinformation and other harms, such as via media/digital literacy campaigns
4.37	**Outcome 9.** Users are empowered with control over the content they see and/or their experiences and interactions online	**Measure 37.** Implement, enforce and/or maintain policies, processes, products and/or programs that seek to provide users with appropriate control over the content they see, the character of their feed and/or their community online.
4.38		**Measure 38.** Launch and maintain products that provide users with controls over the appropriateness of the ads they see.

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.

	Outcomes	**Measures**
4.39	**Outcome 10.** Transparency of policies, systems, processes and programs that aim to reduce the risk of online harms	**Measure 39.** Publish and make accessible for users Signatories' safety and harms-related policies and terms of service.
4.40		**Measure 40.** Publish and make accessible information (such as via blog posts, press releases and/or media articles) on relevant policies, processes, and products that aim to reduce the spread and prevalence of harmful content online.
4.41	**Outcome 11.** Publication of regular transparency reports on efforts to reduce the spread and prevalence of harmful content and related KPIs/metrics	**Measure 41.** Publish periodic transparency reports with KPIs/metrics showing actions taken based on policies, processes and products to reduce the spread or prevalence of harmful content (e.g. periodic transparency reports on global removal of policy-violating content).
4.42		**Measure 42.** Submit to the Administrator an annual compliance report, as required in section

		5.4, that set out the measures in place and progress made in relation to Signatories' commitments under the Code.

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.

	Outcomes	**Measures**
4.43	**Outcome 12.** Independent research that helps build understanding of the impact of safety interventions and harmful content on society and/or research on new technologies to enhance safety or reduce harmful content online.	**Measure 43.** Support or participate, where appropriate, in programs and initiatives undertaken by researchers, civil society and other relevant organisations (such as fact-checking bodies). This may include broader regional or global research initiatives undertaken by the Signatory which may also benefit Aotearoa New Zealand.
4.44		**Measure 44.** Support or convene at least one event per year to foster multi-stakeholder dialogue, particularly with the research community, regarding one of the key themes of online safety and harmful content, as outlined in section 4. This may include broader regional or global events undertaken by the Signatory which involve Aotearoa New Zealand.
4.45	**Outcome 13.** Support independent evaluation of the systems, policies and processes that have been implemented in relation to the Code.	**Measure 45.** Commit to selecting an independent third-party organization to review the annual compliance reports submitted by Signatories, and evaluate the level of progress made against the Commitments, Outcomes and Measures, as outlined in section 4, as well as commitments made by Signatories in their Participation Form (see Appendix 2).

5. Governance, complaints and compliance

The Code aims to provide a governance framework that allows the Administrator, government, civil society and other relevant stakeholders, as well as the public to hold Signatories to their commitments. The Administrator and Signatories agree to establish, within six (6) months of the commencement of this Code, a governance framework setting out the powers, structure and mechanisms for oversight. The governance of the Code will include, at minimum, the following elements:

1) An Administrator;
2) An Oversight Committee to provide monitoring and oversight of Signatories' commitments and review of the Code;
3) A Complaints Mechanism that enables the public to report breaches by Signatories of their Code commitments;
4) Annual compliance reporting by Signatories on their Code commitments; and
5) Biennial review of the Code to assess if it is meeting its Purpose (as outlined in section 1.1) and objectives.

5.1 Administrator

The Administrator is an organisation agreed upon and appointed by the Signatories to oversee the administration of the Code. The Administrator would be able to demonstrate relevant experience in relation to its role and responsibilities, as outlined in section 6 and Annex 4 of the Code.

5.2 Oversight Committee

Reflecting the need for whole-of-society collaboration to protect the integrity of the digital information ecosystem from abuse, the Oversight Committee will be comprised of a range of stakeholders, including representatives from the Signatories, Maori cultural partners, civil society and other relevant and agreed-upon stakeholders (such as Government and academics), who will meet annually, at minimum, to review how Signatories are meeting their commitments under the Code. This includes assessing Signatories' annual compliance reports; complaints submitted through the Complaints Mechanism; and progress of the Code.

5.3 Public Complaints Mechanism

The Administrator will work with Signatories to establish a complaints policy, mechanism, violation definitions/criteria, and eligibility of complainants (referred to as the 'Complaints Mechanism') for addressing non-compliance by Signatories to the commitments they have made under the Code. The Complaints Mechanism will allow people residing in Aotearoa New Zealand to submit complaints against Signatories who they believe are in breach of the Code, as it pertains to commitments, outcomes and measures detailed in section 4.

The Administrator will only receive complaints as it relates to the commitments in the Code. It will not receive complaints against Signatories regarding content on their platforms, including whether specific items of content should be retained or removed.

The Administrator will produce and publish an annual transparency report (which may be part of the Administrator's analysis of the Signatories' annual compliance reports, as outlined in section 5.4) on the complaints received and responded to.

5.3.1 Redress for non-compliance

The Administrator, as part of the Complaints Mechanism, will work with Signatories to establish the criteria for determining non-compliance and appropriate redress mechanism(s) for Signatories to respond to complaints. Signatories will be given a reasonable opportunity and time to consider and respond to the complaint(s).

Signatories that repeatedly fail to comply with their commitments under the Code may be terminated, as outlined in section 3.5.1.

5.4 Annual Compliance Reporting

Signatories will each provide an annual report to the Administrator setting out the measures implemented and the progress they have made in relation to the expected outcomes, as outlined in section 4 of the Code. Reports should follow the template as provided in Appendix 3 of the Code.

A first report outlining the current state of platforms' efforts to address Online Safety and harms concerns will be submitted to the Administrator within ninety (90) days of the commencement of the Code.

The first annual report will be submitted 45 days following 12 months (365 days) from the commencement date of the Code, and followed by yearly reports thereafter.

The Annual Reports will be published on a publicly accessible website maintained by the Administrator.

The Administrator will also publish and make publicly available an analysis of Signatories' reports and their progress within 90 days after the reports were submitted.

5.5 Review of Code

The Code will be reviewed by the Oversight Committee after it has been in operation for twelve (12) months, and thereafter at two-yearly intervals. The reviews will be based on the input of the Signatories, and other relevant and agreed-upon stakeholders (such as civil society organizations, academics, and government bodies).

Any changes or amendments to the Code, resulting from the Review, must be agreed by the Administrator and all Signatories before coming into effect.

6. Code administration

The daily administration of the Code will be handled by the Administrator, which would be appointed upon agreement of all Signatories. The Administrator may delegate or outsource its administrative functions, where appropriate, to a secretariat.

6.1 *Te Rangapu Whakatutuki* (the Administrator)

The Administrator plays an important role in the governance and administration of the Code - in ensuring that the evolution of the Code continues to be guided by the Māori principles of *mahi tahi* (solidarity), *kauhanganuitanga* (balance), *mana tangata* (humanity) and *mana* (respect), as outlined in section 1.3, as well as ensuring that Signatories are held to their commitments.

As in the Māori concept of ***Rangapu Whakatutuki*** -- meaning the agency responsible for achievement -- the Administrator is responsible for steering the diverse group of Signatories and stakeholders toward *Kia kotahi te hoe* (paddling in unison) and toward the greater purpose of enhancing people's safety and reducing harmful content online for Aotearoa New Zealanders.

The Administrator will perform the following functions:

1) Maintain a multi-stakeholder Oversight Committee and facilitate regular meetings, as outlined in section 5.

2) Establish and facilitate the Complaints Mechanism, as outlined in section 5.3, as well as produce and publish an annual transparency report (which may be part of the Administrator's analysis of the Signatories' annual compliance reports, as outlined in section 5.4) on the complaints received and responded to.

3) Facilitate the Signatory annual compliance reporting process; collect and publish the reports on a publicly accessible website maintained by the Administrator; and facilitate an analysis of Signatories' reports and their progress within 90 days after the reports were submitted (as outlined in section 5.4).

4) Facilitate the Code review process; related Oversight Committee meetings; and the drafting and approval of amendments, if any (as outlined in section 5.5).

5) Maintain a publicly accessible website ('the Code Website') that will house materials related to the Code (such as Signatories' Annual Compliance Reports and the Administrator's analysis; the complaints policies and reporting system; public comments about Signatories or the Code; governance policies; etc).

6) Engage and onboard new signatories for the Code, based on recommendations from the Oversight Committee and with agreement from current signatories.

The Administrator's duties and responsibilities are further documented in Appendix 4.

6.2 Administrator's powers

The Administrator may exercise the following powers to ensure compliance with the Code:

1. Make binding decisions on the admission of new signatories (with agreement from signatories);
2. Make binding decisions on the termination of a signatory, based on repeated failures to comply with the commitments of the Code (based on recommendations from the Oversight Committee). The Administrator will consult with the relevant Signatory before a public announcement of termination is made.
3. Make public comments about the Code and may name individual Signatories for positive or negative progress, where there is a proper basis to do so. The Administrator must provide reasonable notice and consult with the relevant Signatory before making public comments about the Signatory with respect to the Code.
4. Make amendments to the Code amendments, based on recommendations from the Oversight Committee, and with agreement from signatories.

6.3 Term & Performance Review

The Administrator will start its duties on the commencement date of the Code and will continue indefinitely. The Oversight Committee may appoint an external reviewer to conduct an independent performance review of the Administrator every two (2) years from the date of commencement.

6.4 Funding of *Te Rangapu Whakatutuki* (the Administrator)

Signatories agree to contribute annually to the funding of the Administrator for the management, oversight and administration of the Code, and for the responsibilities as outlined in this section and Appendix 4.

Signatories agree to provide sufficient funding to the Administrator to ensure it can fulfil its role and responsibilities.

Signatories shall respect and uphold the independence of the Administrator, irrespective of their financial contributions.

The Administrator will publish an annual financial report accounting for the funding received and how it was spent.

7. Glossary

The glossary provides general references and information on some of the key terms used in this Code. It is not intended to provide definitions as Signatories may have different definitions as it relates to their individual policies and practices.

Bullying and Harassment. Online bullying (also known as cyberbullying) is when a person uses digital technology to send, post or publish content with the intent to harm another

person or a group. This behaviour is often aggressive, is repeated and involves some kind of power imbalance between the people involved.

Child Sexual Exploitation and Abuse (CSEA). CSEA includes the sharing of child sexual abuse material, the livestreaming of child sexual abuse and the online grooming of children.

Complaints. Concerns raised by the public about possible breaches of the Code by signatories as contemplated by section 4, noting that this definition excludes individual complaints of signatories' decisions regarding content on their platforms.

Disinformation. (i) Digital content that is verifiably false or misleading or deceptive; (ii) propagated amongst users of digital platforms via inauthentic behaviours; and (iii) the dissemination of which is reasonably likely to cause harm.

Hate Speech. Digital content that promotes hatred or violence against individuals or a group on the basis of age, race, ethnicity, national origin, disability, religious affiliation, caste, sexual orientation, sex, gender identity and expression, immigration status, veteran status and victims of a major violent event and their kin.

Harm. Actors, behaviours and/or content online that pose an imminent and serious threat to the safety of users and/or the integrity of the digital information ecosystem, and of which may lead to real world harm. The types of harm, for the purpose of this Code, are outlined in section 1.4.

Incitement of Violence. Digital content that incites or facilitates serious violence, where there is a genuine risk of physical harm or direct threats to public safety.

Misinformation. (i) Digital Content (often legal) that is verifiably false or misleading or deceptive; (ii) propagated by users of digital platforms; and (iii) disseminated to (reasonably likely, but may not be clearly intended to) cause harm.

User-Generated Content (UGC). Alternatively known as user-created content (UCC), is any form of digital content, such as images, videos, text, and audio, that has been posted by users on online platforms.

Violent or Graphic Content. Real-world imagery that is gratuitously shocking, graphic, sadistic, or gruesome or that promotes, normalizes, or glorifies extreme violence or suffering.

Appendix 1: Table of Signatories

	Signatory	Date Code Signed
1	Meta (Facebook and Instagram)	26 July 2022
2	Google (YouTube)	26 July 2022
3	TikTok	26 July 2022
4	Twitch	26 July 2022
5	Twitter	26 July 2022
6		
7		
8		
9		
10		
11		
12		

Appendix 2: Signatory Participation Form

Given the range of products and services, supporting different and diverse user communities, with varying capabilities of digital platforms, Signatories may make commitments to the Code that best matches their risk profiles, either for the company or for specific products/services.

Signatories are required to specify which commitments, outcomes and measures -- at the company, product or service-level -- that are most relevant to them for the purposes of the Code in the 'Signatory Participation Form' (see Appendix 2).

This form must be submitted upon the Signatory's sign-on date to the Code, and will include, for each measure, either an initial assessment of practices being undertaken or an explanation as to why specific measures are not being implemented.

Signatories may amend this form, provided that they provide 90 days notice to the Administrator of any amendments.

Signatory:	

Date Effective:	

***If applicable*: Relevant Products / Services:**	*[To be completed if the signatory is applying different commitments, outcomes and measures for different products & services. The last column of the table below may be expanded so that the signatory may indicate which commitments, outcomes and measures apply to different products and services.]*

4.1 Reduce the prevalence of harmful content online

Signatories will indicate below which commitments, outcomes and measures are relevant for their company or for their products/services as it relates to reducing the prevalence of harmful content online.

	Outcomes	**Measures**	**Relevant to company, product or service**	
4.1	**Outcome 1.** Provide safeguards to reduce the risk of harm arising from online **child sexual**	**Measure 1.** Implement, enforce and/or maintain policies, processes, products, and/or programs that seek to prevent known child sexual abuse material from being made available to	Y/N	[Comment]

	exploitation & abuse (CSEA)	users or accessible on their platforms and services		
4.2		**Measure 2.** Implement, enforce and/or maintain policies, processes, products, and/or programs that seek to prevent search results from surfacing child sexual abuse material	Y/N	[Comment]
4.3		**Measure 3.** Implement, enforce and/or maintain policies, processes, products, and/or programs that seek to adopt enhanced safety measures to protect children online from peers or adults seeking to engage in harmful sexual activity with children (e.g. online grooming and predatory behaviour)	Y/N	[Comment]
4.4		**Measure 4.** Implement, enforce and/or maintain policies, processes, products, and/or programs that seek to reduce new and ongoing opportunities for the sexual abuse or exploitation of children	Y/N	[Comment]
4.5		**Measure 5.** Work to collaborate across industry and with other relevant stakeholders to respond to evolving threats	Y/N	[Comment]
4.6	**Outcome 2:** Provide safeguards to reduce the risk of harm arising from online **bullying or harassment**	**Measure 6.** Implement, enforce and/or maintain policies and processes that seek to reduce the risk to individuals (both minors and adults) or groups from being the target of online bullying or harassment.	Y/N	[Comment]
4.7		**Measure 7.** Implement and maintain products and/or tools that seek to mitigate the risk of individuals or groups from being the target of online bullying or harassment.	Y/N	[Comment]
4.8		**Measure 8.** Implement, maintain and raise awareness of product or service related policies and tools for users to report online bullying or harassment content.	Y/N	[Comment]
4.9		**Measure 9.** Support or maintain programs, initiatives or features that	Y/N	[Comment]

		seek to educate and raise awareness on how to reduce or stop online bullying or harassment.		
4.10	**Outcome 3:** Provide safeguards to reduce the risk of harm arising from online **hate speech**	**Measure 10.** Implement, enforce and/or maintain policies and processes that seek to prohibit or reduce the prevalence of hate speech.	Y/N	[Comment]
4.11		**Measure 11.** Implement and maintain products and tools that seek to prohibit or reduce the prevalence of hate speech.	Y/N	[Comment]
4.12		**Measure 12.** Implement, maintain and raise awareness of product or service related policies and tools for users to report potential hate speech.	Y/N	[Comment]
4.13		**Measure 13.** Support or maintain programs and initiatives that seek to encourage critical thinking and educate users on how to reduce or stop the spread of online hate speech.	Y/N	[Comment]
4.14		**Measure 14.** Work to collaborate across industry and with other relevant stakeholders to support efforts to respond to evolving harms arising from online hate speech.	Y/N	[Comment]
4.15	**Outcome 4:** Provide safeguards to reduce the risk of harm arising from online **incitement of violence**	**Measure 15.** Implement, enforce and/or maintain policies and processes that seek to prohibit or reduce the prevalence of content that potentially incites violence.	Y/N	[Comment]
4.16		**Measure 16.** Implement and maintain products and tools that seek to prohibit or reduce the prevalence of content that potentially incites violence.	Y/N	[Comment]
4.17		**Measure 17.** Implement, maintain and raise awareness of product or service related policies and tools for users to report content that potentially incites violence.	Y/N	[Comment]

4.18		**Measure 18.** Support or maintain programs and initiatives that seek to educate users on how to reduce or stop the spread of online content that incites violence.	Y/N	[Comment]
4.19		**Measure 19.** Work to collaborate across industry and with other relevant stakeholders to support efforts to respond to evolving harms arising from online content that incites violence.	Y/N	[Comment]
4.20	**Outcome 5:** Provide safeguards to reduce the risk of harm arising from online **violent or graphic content**	**Measure 20.** Implement, enforce and/or maintain policies and processes that seek to prohibit and/or reduce the spread of violent or graphic content online.	Y/N	[Comment]
4.21		**Measure 21.** Implement and maintain products and tools that seek to and/or reduce the spread of violent or graphic content.	Y/N	[Comment]
4.22		**Measure 22.** Implement, maintain and raise awareness of product or service related policies and tools for users to report potential violent and graphic content.	Y/N	[Comment]
4.23	**Outcome 6:** Provide safeguards to reduce the risk of harm arising from online **misinformation**	**Measure 23.** Implement, enforce and/or maintain policies, processes and/or products that seek to reduce the spread of online misinformation.	Y/N	[Comment]
4.24		**Measure 24.** Implement, enforce and/or maintain policies and processes that seek to penalise users who repeatedly post or share misinformation that violates related policies.	Y/N	[Comment]
4.25		**Measure 25.** Support or maintain media literacy programs and initiatives that seek to encourage critical thinking and educate users on how to reduce or stop the spread of misinformation.	Y/N	[Comment]
4.26		**Measure 26.** Support or maintain programs and/or initiatives that seek to support civil society, fact-checking bodies and/or other relevant	Y/N	[Comment]

		organisations working to combat misinformation.		
4.27		**Measure 27.** Work to collaborate across industry and with other relevant stakeholders to support efforts to respond to evolving harms arising from misinformation.	Y/N	[Comment]
4.28	**Outcome 7:** Provide safeguards to reduce the risk of harm arising from online **disinformation**	**Measure 28.** Implement, enforce and/or maintain policies, processes and/or products that seek to suspend, remove, disable, or penalise the use of fake accounts that are misleading, deceptive and/or may cause harm.	Y/N	[Comment]
4.29		**Measure 29.** Implement, enforce and/or maintain policies, processes and/or products that seek to remove accounts, (including profiles, pages, handles, channels, etc) that repeatedly spread disinformation.	Y/N	[Comment]
4.30		**Measure 30.** Implement, enforce and/or maintain policies, processes and/or products that seek to provide information on public accounts (including profiles, pages, handles, channels, etc) that empower users to make informed decisions (e.g. date a public profile was created, date of changes to primary account information, number of followers).	Y/N	[Comment]
4.31		**Measure 31.** Implement, enforce and/or maintain policies, processes and/or products that seek to provide transparency on paid political content (e.g. advertising or sponsored content) and give users more context and information (e.g. paid political or electoral ad labels or who paid for the ad).	Y/N	[Comment]
4.32		**Measure 32.** Implement, enforce and/or maintain policies, processes and/or products that seek to disrupt advertising and/or reduce economic incentives for users who profit from disinformation.	Y/N	[Comment]

4.33		**Measure 33.** Work to collaborate across industry and with other relevant stakeholders to support efforts to respond to evolving harms arising from disinformation.	Y/N	[Comment]

4.2 Empower users to have more control and make informed choices

Signatories will indicate below which commitments, outcomes and measures are relevant for their company or for their products/services as it relates to empowering users to have more control and make informed choices.

	Outcomes	**Measures**	**Relevant to company, product or service**	
4.34	**Outcome 8.** Users are empowered to make informed decisions about the content they see on the platform	**Measure 34.** Implement, enforce and/or maintain policies, processes, products and/or programs that helps users make more informed decisions on the content they see	Y/N	[Comment]
4.35		**Measure 35.** Implement, enforce and/or maintain policies, processes, products and/or programs that seek to promote accurate and credible information about highly significant issues of societal importance and of relevance to the digital platform's user community (e.g. public health, climate change, elections)	Y/N	[Comment]
4.36		**Measure 36.** Launch programs and/or initiatives that educate or raise awareness on disinformation, misinformation and other harms, such as via media/digital literacy campaigns	Y/N	[Comment]
4.37	**Outcome 9.** Users are empowered with control over the content they see and/or their experiences and interactions online	**Measure 37.** Implement, enforce and/or maintain policies, processes, products and/or programs that seek to provide users with appropriate control over the content they see, the character of their feed and/or their community online.	Y/N	[Comment]
4.38		**Measure 38.** Launch and maintain products that provide users with controls over the appropriateness of the ads they see.	Y/N	[Comment]

4.3 Enhance transparency of policies, processes and systems

Signatories will indicate below which commitments, outcomes and measures are relevant for their company or for their products/services as it relates to enhancing transparency of policies, processes and systems.

	Outcomes	Measures	Relevant to company, product or service	
4.39	**Outcome 10.** Transparency of policies, systems, processes and programs that aim to reduce the risk of online harms	**Measure 39.** Publish and make accessible for users Signatories' safety and harms-related policies and terms of service.	Y/N	[Comment]
4.40		**Measure 40.** Publish and make accessible information (such as via blog posts, press releases and/or media articles) on relevant policies, processes, and products that aim to reduce the spread and prevalence of harmful content online.	Y/N	[Comment]
4.41	**Outcome 11.** Publication of regular transparency reports on efforts to reduce the spread and prevalence of harmful content and related KPIs/metrics	**Measure 41.** Publish periodic transparency reports with KPIs/metrics showing actions taken based on policies, processes and products to reduce the spread or prevalence of harmful content (e.g. periodic transparency reports on removal of policy-violating content).	Y/N	[Comment]
4.42		**Measure 42.** Submit to the Administrator an annual compliance report, as required in section 5.4, that set out the measures in place and progress made in relation to Signatories' commitments under the Code.	Y/N	[Comment]

4.4 Support independent research and evaluation

Signatories will indicate below which commitments, outcomes and measures are relevant for their company or for their products/services as it relates to supporting independent research and evaluation.

	Outcomes	Measures	Relevant to company, product or service	
4.43	**Outcome 12.** Independent research that helps build understanding of the impact of safety interventions and harmful content on society and/or research on new technologies to enhance safety or reduce harmful content online.	**Measure 43.** Support or participate, where appropriate, in programs and initiatives undertaken by researchers, civil society and other relevant organisations (such as fact-checking bodies). This may include broader regional or global research initiatives undertaken by the Signatory which may also benefit Aotearoa New Zealand.	Y/N	[Comment]
4.44		**Measure 44.** Support or convene at least one event per year to foster multi-stakeholder dialogue, particularly with the research community, regarding one of the key themes of online safety and harmful content, as outlined in section 4. This may include broader regional or global events undertaken by the Signatory which involve Aotearoa New Zealand.	Y/N	[Comment]
4.45	**Outcome 13:** Support independent evaluation of the systems, policies and processes that have been implemented in relation to the Code.	**Measure 45.** Commit to selecting an independent third-party organization to review the annual compliance reports submitted by Signatories, and evaluate the level of progress made against the Commitments, Outcomes and Measures, as outlined in section 4, as well as commitments made by Signatories in their Participation Form (see Appendix 2).	Y/N	[Comment]

Appendix 3: Report Template

Signatory:	

***If applicable*:** **Relevant Products / Services:**	

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)**
[Provide details of measures implemented in relation to signatories measures 1-5 of the Code] [Provide metrics, if any, that demonstrate efforts related to the overall *outcome*, e.g. pieces of content removed for violating CSEA policies, number of people who participated in education programs]
Outcome 2: Provide safeguards to reduce the risk of harm arising from online **bullying or harassment**

[Provide details of measures implemented in relation to signatories measures 6-9 of the Code] [Provide metrics, if any, that demonstrate efforts related to the overall *outcome*, e.g. pieces of content removed for violating bullying and/or harassment policies, number of people who participated in education programs]
Outcome 3: Provide safeguards to reduce the risk of harm arising from online **hate speech**
[Provide details of measures implemented in relation to signatories measures 10-14 of the code] [Provide metrics, if any, that demonstrate efforts related to the overall *outcome*, e.g. pieces of content removed for violating hate speech policies, number of people who participated in education programs]
Outcome 4: Provide safeguards to reduce the risk of harm arising from online **incitement of violence**
[Provide details of measures implemented in relation to signatories measures 15-19 of the code] [Provide metrics, if any, that demonstrate efforts related to the overall *outcome*, e.g. pieces of content removed for violating incitement of violence policies, number of people who participated in education programs]
Outcome 5: Provide safeguards to reduce the risk of harm arising from online **violent or graphic content**
[Provide details of measures implemented in relation to signatories measures 20-22 of the code] [Provide metrics, if any, that demonstrate efforts related to the overall *outcome*, e.g. pieces of content removed for violating graphic or violent content policies]

Outcome 6: Provide safeguards to reduce the risk of harm arising from online **misinformation**
[Provide details of measures implemented in relation to signatories measures 23-27 of the code] [Provide metrics, if any, that demonstrate efforts related to the overall *outcome*, e.g. pieces of content removed for violating misinformation policies, number of people who participated in media/digital literacy education programs]
Outcome 7: Provide safeguards to reduce the risk of harm arising from online **disinformation**
[Provide details of measures implemented in relation to signatories measures 28-33 of the code] [Provide metrics, if any, that demonstrate efforts related to the overall *outcome*, e.g. pieces of accounts or pages removed for violating disinformation policies, number of people who participated in media/digital literacy education programs]

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
[Provide details of measures implemented in relation to signatories measures 34-36 of the Code]

[Provide metrics, if any, that demonstrate efforts related to the overall outcome]
Outcome 9. Users are **empowered with control** over the content they see and/or their experiences and interactions online
[Provide details of measures implemented in relation to signatories measures 37-38 of the Code] [Provide metrics, if any, that demonstrate efforts related to the overall outcome]

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
[Provide details of measures implemented in relation to signatories measures 39-40 of the Code]
Outcome 11. Publication of regular **transparency reports** on efforts to reduce the spread and prevalence of harmful content and related KPIs/metrics
[Provide details of measures implemented in relation to signatories measures 41-42 of the Code]

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.
[Provide details of measures implemented in relation to signatories measures 43-44 of the Code]
Outcome 13. Support independent evaluation of the systems, policies and processes that have been implemented in relation to the Code.
[Provide details of measures implemented in relation to signatories measure 45 of the Code]

Appendix 4: Administrator Duties & Responsibilities

Appendix 4 outlines in more detail what is expected of the appointed Administrator.

I. Administrator

The Signatories of the Aotearoa New Zealand Code of Practice for Online Safety and Harms have recommended XXXXX as the Administrator. The Administrator's Duties will come into effect upon finalization of the Code.

II. Duties of the Administrator

The Administrator will perform its duties to the best of its ability as follows (collectively, the 'Administrator's Duties'):

1. **Oversight & Evaluation**
    a. Establish an Oversight Committee, as outlined in section 5.2, that will be comprised of representatives from the Signatories, Maori partners, civil society and other relevant and agreed-upon stakeholders (such as Government and academics), facilitate annual meetings, at minimum, for the purpose of reviewing how Signatories are meeting their commitments as well as progress of the Code.
    b. Monitor Signatories' progress and address non-compliance concerns, as outlined in section 6.2
    c. Produce and publish an analysis of Signatories' annual compliance reports and their progress within 90 days after the reports are submitted, which may include the annual transparency report on the complaints received and responded to via the Complaints Mechanism, as outlined in section 5.4 ('Administrator's Annual Report').
    d. Publish and make publicly available Signatories' annual compliance reports, as well as the Administrator's Annual Report, on the Code Website.
2. **Public Events and Stakeholder Engagements**
    a. Organize at least one public event per year, with relevant stakeholders, for discussion about matters rising from the Code and the Administrator's Annual Report.
    b. Liaise with Government, civil society and other relevant stakeholders on the Code.

c. Engage and onboard new signatories for the Code, based on recommendations from the Oversight Committee and with agreement from current signatories.

3. **Governance and Complaints Mechanism**

    a. Establish and facilitate the complaints mechanism, as outlined in section 6 of the Code. This includes:

        i. Publishing a complaints policy and procedure for people to report potential non-compliance with the Code.

        ii. Review and screen complaints reported; assess validity of complaints; compile and submit to relevant Signatory for response.

        iii. Produce and publish the communication (e.g. in the form of a report, blog post or other format) regarding eligible complaints received and accepted for response via the Complaints Mechanism.

        iv. Produce and publish an annual transparency report (which may be part of the Administrator's analysis of the Signatories' annual compliance reports, as outlined in section 5.4) on the complaints received and responded to.

4. **Communication & Website Maintenance**

    a. Manage and facilitate regular communications on the progress of the Code.

    b. Maintain a publicly accessible website ('the Code Website') that will house materials related to the Code, such as (but not limited to) the Annual Compliance Reports submitted by Signatories, the complaints reporting system, public comments, the Administrator's Annual Report, and industry relevant legislation.

III. Funding

The Administrator is funded by the Signatories to exercise the duties outlined in section II of this Appendix. The Administrator will assess the required level of funding, for the following year, no less than 60 days prior to the end of the existing year and communicate this to the Signatories. The Signatories agree to provide sufficient funding to the Administrator to ensure the Administrator can fulfill the Administrator's Duties.

The Administrator will use funding for the following purposes:

1) Administrative staff, who will execute and carry out the Administrator's Duties

2) All reasonable costs / expenses in relation to fulfilling the Administrator's Duties, including development and maintenance of the Code Website and organising the annual event.