Aotearoa New Zealand Code of Practice for Online Safety and Harms

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 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:	**Twitter**

Date Effective:	**29 July 2022**

***If applicable*: Relevant Products / Services:**	Twitter's mission is to serve the public conversation. Transparency is fundamental to our work in achieving that mission. This annual report outlines our commitments and progress under the Australian Code of Practice on Disinformation and Misinformation (the Code) to demonstrate Twitter's efforts to protect the public conversation and uphold the integrity of our service. We are committed to providing meaningful transparency reporting to the public. We do this both in partnership with government, academics, and civil society under schemes like the Code and through existing, proactive self reporting initiatives like the biannual Twitter Transparency Report.

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 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	Y	Twitter
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	Twitter
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	Twitter
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	Twitter
4.5		**Measure 5.** Work to collaborate across industry and with other relevant stakeholders to respond to evolving threats	Y	Twitter
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	Twitter
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	Twitter
4.8		**Measure 8.** Implement, maintain and raise awareness of product or service	Y	Twitter

		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.	Y	Twitter
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	Twitter
4.11		**Measure 11.** Implement and maintain products and tools that seek to prohibit or reduce the prevalence of hate speech.	Y	Twitter
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	Twitter
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	Twitter
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	Twitter
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	Twitter
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	Twitter
4.17		**Measure 17.** Implement, maintain and raise awareness of product or service related policies and tools for users to	Y	Twitter

		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.	Y	Twitter
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	Twitter
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	Twitter
4.21		**Measure 21.** Implement and maintain products and tools that seek to and/or reduce the spread of violent or graphic content.	Y	Twitter
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	Twitter
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	Twitter
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	Twitter
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	Twitter
4.26		**Measure 26.** Support or maintain programs and/or initiatives that seek to support civil society, fact-checking	Y	Twitter

		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.	Y	Twitter
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	Twitter
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	Twitter
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	Twitter
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	Twitter
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	Twitter
4.33		**Measure 33.** Work to collaborate across industry and with other relevant	Y	Twitter

		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 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	Twitter
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	Twitter
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	Twitter
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	Twitter
4.38		**Measure 38.** Launch and maintain products that provide users with controls over the appropriateness of the ads they see.	Y	Twitter

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	Twitter
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	Twitter
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	Twitter
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	Twitter

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	Twitter
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	Twitter
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	Twitter