1 December 2021

Netsafe and industry Signatories to the Code
thecode@netsafe.org.nz

To whom it may concern

Draft Aotearoa New Zealand Code of Practice for Online Safety and Harms

Brainbox has prepared two detailed public policy analyses in 2021 on regulation of social media and content moderation. We have drawn on this work in preparing this comment. Consistent with our findings in our 2021 reports, we broadly welcome the decision to pioneer a self regulatory model based on transparency reporting. We take this opportunity also to thank Martin Cocker for his work at Netsafe during his tenure and wish him all the best.

The Code is not backed by democratic institutions or by legislation. Its legitimacy therefore rests almost exclusively on public trust and confidence in two things: the process followed to develop it; and the transparency and legitimacy arrangements for its implementation. I note industry representatives on the call emphasised that civil society input into the Committee and into the Administrator's work would be an essential part of the Code's success.

We express our support for the comments made by other attendees about the limited time and resourcing available for engaging in meaningful consultation on the Code. We observe that, in other situations, industry representatives have expressed similar frustrations with the limitations of Government consultation processes. We record our concern that the depth and quality of consultation achieved in the meeting today risks being misrepresented in public statements about the Code. We welcome the decision today to expand consultation.

We draw your attention to key principles under the Code, including: the importance of freedom of expression; the value of transparency; the need for collaboration and a whole of sector approach; the importance of civil society's role and input.

Brainbox will be sending this comment to meeting attendees immediately with a view to supporting civil society response. Given the draft is confidential, Brainbox will also publish this feedback once it receives approval from Netsafe and the Signatories to do so.

For the Code to succeed, it must anticipate and resolve disputes about how it operates. It must also provide a means of resolving significant trade-offs touching on the interests, rights and responsibilities of the Administrator, the public, and Signatories to the Code. If these cannot be resolved in a respectable manner, they will undermine trust and confidence in the scheme and in State confidence of its legitimacy and effectiveness. As a result, the scheme will fail and legislation will be introduced to resolve those trade-offs.

Thank you for the opportunity to provide input into the Code (see table attached).

Yours faithfully

Tom Barraclough

Ref	Topic	Explanation
Consultation process, transparency, freedom of expression		
1	Code confidentiality	When the Code was distributed, it was done so confidentially. Freedom of expression and transparency are core principles under the Code. 1. Who decided that the Code would be distributed confidentially: Netsafe, Signatories, or a particular Signatory? 2. What are some of the factors influencing the decision to distribute the Code confidentially?
2	Origins of Code	The stated intent of the Code during the meeting was to influence the course of regulation in New Zealand. 1. Which agency/company first proposed the drafting of a Code? 2. Which agency/company broadly spent the most time on the Code's drafting and scope?
3	Consultation	As a voluntary Code, its credibility and support will rest heavily on public perceptions of the Code's own integrity, including the processes followed in its development. Brainbox would be concerned to see the meeting today referred to in any way that suggests it fostered meaningful feedback.
4	Suggestion that consultation will be ongoing throughout implementation and that implementation will resolve issues	Attendees raised serious concerns about whether consultation was meaningful. In response, a company representative emphasised that consultation and scrutiny would be ongoing as part of the Code's operation. An agency representative emphasised that some issues with the Code could only be worked out and resolved through implementation. While these are important insights, I do not accept they address the concerns being raised. 1. The Code has been drafted to reflect existing practice by the Signatory companies. There is no suggestion that existing Signatories will be unable to comply with the Code. 2. The core implementation issues raised in this submission primarily relate to matters where the Code is silent, or where there is no procedure for resolving the issue in the Code itself. The Code will therefore be unable to support resolution of these issues through implementation. 3. If the success of the Code is intended to rely on the input of civil society groups, the Code must create a mechanism to fund and support civil society to play this role adequately beyond the core role of the Committee.
5	Government departments input into consultation	Brainbox supports the investigation of non-government regulatory initiatives because of the risk States can pose to human rights. Transparency is a key principle of the Code and a measure adopted by Signatories to prevent abuse of transparency arrangements by States generally. 1. Trust and confidence in the Code would be enhanced by providing a list of government agencies who have had

		input into the Code. 2. Trust and confidence would be enhanced by providing any written materials provided by a government agency in the course of the Code's development.
6	Comparative Codes	The Code itself, and comments made during the call, explicitly claim that the Code reflects existing practice and insights from other self-regulatory approaches. It is impossible to assess the accuracy of this claim without seeing comparative Codes and the work done to analyse those Codes. Civil society capacity is extremely limited for investigating these alternative Codes and commentary about those Codes. 1. To enhance trust and confidence and facilitate input from civil society groups, the Signatories and Netsafe should collate and provide a bundle of relevant information they have relied on to support the proposition that the Code: reflects existing best practice; and incorporates insights from other similar regulatory initiatives. This should include the Codes themselves and any commentary on those Codes, or critical analysis of them by commentators. 2. To enhance trust and confidence, facilitate input from civil society groups, and enable meaningful consultation, the Administrator and the Signatories should provide any internal analyses they have relied upon in drafting the Code that illustrates how the Code does or does not correspond with other examples of similar initiatives.
Perception of the Code as a regulatory initiative once implemented		
7	Funding under the Code	The Signatories will fund the Administrator. Initiatives such as the Oversight Board (Meta/Facebook) secure the Board's funding to prevent perceived or actual influence by regulatee over regulator. 1. What security of funding will be provided to the Administrator? 2. What processes will be put in place to provide public assurance that funding is not (intentionally or not) influencing the Administrator's conduct? 3. What dispute resolution processes will be put in place to enable the Administrator to dispute the Signatories' funding decisions? 4. What dispute resolution processes will be put in place to resolve disagreement between a Signatory and the Administrator about the acceptability of findings made by the Administrator and the Administrator's compliance with the Code?
8	Netsafe's funding	The Code will provide Netsafe with access to a significant funding stream and a significant array of responsibilities. This may create a perception that Netsafe's interests and institutional incentives are to ensure the Code persists as a self-regulatory measure, and to avoid disputes with Signatories that could lead to Signatories withdrawing from the Code or terminating it "for any other reason" pursuant to the Code. ● Funding decisions under the Code should be transparent and subject to measures that guarantee the independence of the Administrator and its ability to make decisions

		without fear of consequences if the Signatories disagree.
9	Netsafe's other roles in a State regulatory system	Netsafe has responsibilities under the HDCA. It will also have responsibilities under this industry self-regulatory Code, where it is required to reach conclusions on the adequacy of platforms' systems. 1. What analysis has been done to assess whether perceived or actual conflicts might arise between its role in relation to the HDCA and its role in relation to the industry Code? 2. Will any such analysis be shared publicly? 3. If any real or perceived conflict exists, what processes will be put in place to manage the conflict?
10	Complaints by government actors	Civil society commentators have expressed concern about both: trusted flagging systems by government entities; and the transparency of government interactions with platforms and regulatory mechanisms governing platform actions. On the call, it was stated that the Administrator would receive complaints by government agencies. This is not apparent from the Code itself. It creates risks of perception that could undermine the Code or otherwise undermine trust and confidence in the platform, the Administrator or the Code. 1. If the Code is to permit complaints by government agencies, it should explicitly say so. 2. If the Code is to permit complaints by government agencies, it should require companies and the Administrator to report on these complaints, the specifics of the complaints, and the outcomes taken. 3. It should be noted that while State regulation creates some kinds of risks, proceeding without the benefit of legislative protections for companies and for the Administrator creates other kinds of risks, including pressure or misuse by State actors.
Interpretation of the Code		
11	Principles and values as interpretive guides	The Code includes definitions and key principles, as well as values. The adequacy of Signatories' conduct under the Code will rest heavily on the way the language of the Code is interpreted. 1. If the principles and values in the Code are intended to be taken into account in interpreting the Code, the Code should directly say this. 2. The Code should anticipate a process for public complainants, for Signatories, and for the Administrator to seek independent adjudication on matters of interpretation. This need not be a Court to be effective.
12	Scope of the Code	The Code purportedly takes a systems approach based on Netsafe's existing experience. It excludes complaints about specific content moderation decisions. While we understand the intent here is to avoid framing the Code as an appeal mechanism against company content moderation decisions, we do not think the Code can practically exclude receiving complaints from users about individual content moderation decisions. Further, we think this scoping mechanism risks being used as a way of dismissing complaints that are important for otherwise assessing the quality

		of companies' content moderation systems. 1. The drafters should consider how complaints under this Code that complain about individual content moderation decisions will be registered by the Administrator for systemic assessment purposes. 2. The Code should acknowledge that complaints about individual content moderation decisions are a crucial method of enabling the public to complain about the integrity and reliability of platform content moderation systems. In particular, individual complaints will be essential for assessing whether platform false positive/false negative rates are consistent with public expectations and the values and principles under the Code.
Comment on specific points		
13	Framing: "safeguarding" the "digital ecosystem" against "abuse"	Digital products have embedded norms in them that facilitate or inhibit certain kinds of behaviours. Some digital products may have embedded normativity that unintentionally fosters, incentivises or affords harmful conduct as defined by the Code. Therefore, it may be that bare use of "the digital ecosystem" may be abusive, without any "abuse" of that system. 1. To frame individual conduct as "abuse of the digital ecosystem" is to exclude the possibility that some products (perhaps of future Signatories) may themselves be harmful or abusive. 2. It would be preferable to frame "abuse of the digital ecosystem" as instead being "abusive use of the digital ecosystem" or similar that emphasises actions by individuals. 3. The issue of whether digital ecosystems are themselves harmful or abusive can be dealt with through the wider Code, which directs attention to architecture, policies and processes, and matters of design.
14	Signatory commitment to limited outcomes and measures	A Signatory is able to restrict its commitment to specific outcomes or measures (appendix 2). As currently drafted, a Signatory could commit to the least onerous or meaningful measures, but nevertheless publicly describe themselves as a "Signatory" in the same way as a Signatory who commits to all outcomes and measures. This may undermine Signatory collaboration. It may also undermine the value of being a Signatory to the Code. 1. The Code should consider how to limit the ability of Signatories to misrepresent the extent of their compliance with the Code. 2. The Code should create a method for raising and resolving disputes between Signatories who may disagree with claims being made by other Signatories.
15	The role of smaller companies	The Code has been drafted by established industry companies with well established market positions and revenue streams. Civil society commentators have expressed concern that regulatory measures may present barriers to entry to under resourced market entrants. Other self-regulatory mechanisms such as Tech Against Terrorism focus on sharing best practice with new

		entrants. 1. The Code should explicitly incorporate the principle of proportionality into the Outcomes and Measures to account for the resourcing of smaller companies and prevent undue barriers to entry into established markets, as well as preventing the Code from becoming a tool for use by Signatories to exclude new entrants to markets. 2. The Code should account for the role of established market players in sharing and facilitating the adoption of good practice by market entrants, in the same way as Tech Against Terrorism does.
16	Machine readability of transparency reporting and evaluation information	Capacity by the Administrator and by Civil Society is limited. Analysis of reporting information can be enhanced by ensuring that transparency reporting information, data sets, and other relevant information for evaluative purposes is published in machine readable formats. This is reflected in the Santa Clara principles and in some existing practice by technology companies. The Code should explicitly require that transparency, accountability and evaluative information is machine readable in a way that is intended to enhance independent scrutiny.
17	Definition of harm includes “integrity of online ecosystem”	The current definition of “harm” includes both threat to the safety of users, as well as [threat to] the integrity of the digital information ecosystem (presumably “… which may lead to real world harm” is a modifier to “integrity of the digital information ecosystem”). This definition should be revised and made clearer. We infer that this definition is intended to account for “harms” caused by misinformation or disinformation which are more diffuse. This is a worthwhile feature, however including it within the definition of “harm” makes references elsewhere to “physical harm” unnecessarily complex. 1. The definition of harm should not be limited to “users” and should include non-users. 2. The definition of harm should be revised. It is ambiguous and risks becoming circular.[1] It also undermines the integrity of the wider Code because of the frequency of the word “harm” in other parts. 3. If harm to the integrity of the digital information ecosystem” (which may or may not include “real world harm”) is to be referenced in the Code, it should sit under a specific definition. 4. The notion of “harm to the integrity of the digital information ecosystem that may lead to real world harm” risks being overly broad and therefore unworkable or unable to be operationalised precisely.
18	Proportionality and necessity of harm - p 8 (2.7)	Mention was made during the call of the way that different people suffer different harms differently depending on contextual factors and their sociopolitical position. 2.7 refers to proportionality and taking a risk-based approach proportionate to the level of harm. 1. A human rights approach protects minority groups, defined

[1] Poses an imminent and serious threat to the integrity of the digital information ecosystem, and of which may lead to real world imminent and serious threat to the safety of users.

		broadly to include people in an ethnic, political, or other minority. Harms to this minority risk being assessed as being of lesser weight than harms to a wider majority. 2. 2.7 should include an explicit but broad reference to human rights instruments. We do not think it is adequate to leave these considerations to be "read in" to 2.7 because of their importance. 3. 2.7 should include specific reference to other special protections for minority groups or disadvantaged groups as part of the risk-based proportionality assessment. In a New Zealand context, there should be an explicit decision made as to whether Signatories will commit to te Tiriti o Waitangi / the Treaty of Waitangi.
19	Encryption	Mention was made on the call of platforms like Telegram. The Code does not mention encryption specifically. A commitment to encryption could be read into the commitment to privacy, to privacy laws, and to terms of service. The Code should include a specific statement on expectations about encrypted communications. Encryption is too important and too controversial to be left to implication and risks undermining public trust and confidence in the Code, transparency of enforcement, and agreement on what the Code requires in relation to encryption.
Dispute resolution		
20	Dispute resolution under the Code	The Code does not include any direction as to how disagreements about the application or interpretation of the Code should be resolved. Dispute resolution is essential for two reasons: (1) the effective operation of the Code between Administrator, Signatories, the Committee, and for public complainants; (2) public perception about the legitimacy of the Code and that any disputes are resolved according to the values, principles, and specific terms of the Code itself. We anticipate potential disputes covering: ● Administrator behaviour, including findings and compliance with notice and consultation requirements when making findings against a Signatory ● Administrator application and interpretation of the Code ● Adequacy of standards of assessment and measurement adopted under the Code ● Quality and completeness of reports ● Whether a Signatory's systems are appropriate given the level of the harm involved (for example, whether false positive/negative rates have been set in ways consistent with the principles and values of the Code) ● Adequacy of transparency reporting ● Rights, obligations and disagreements between Signatories, including representations about Signatories' compliance with the Code ● Specific matters, such as whether something is "appropriate" (example, whether "independent evaluation" is "appropriate" (outcome 13, measure 45)). ● Whether Signatories have complied with their agreement

		to “provide sufficient funding to the Administrator to ensure it can fulfil its role and responsibilities” (6.4), including whether funding is being restricted to control the Administrator’s effectiveness or behaviour In light of these comments: 1. The Code should set out a transparent dispute resolution procedure that accounts for the above. 2. To the extent that disputes of this kind are to be resolved through reporting or input by civil society groups, the Code ought to provide for funding and support to enable effective input.