2.

TE MANA WHAKAATU
Classification Office

16 February 2020

Hon Jan Tinetti
Minister of Internal Affairs
Parliament Buildings
WELLINGTON

Dear Minister Tinetti

Background Note for Meeting 17 February 2021 – taking an Integrated Approach to Online Harm

Purpose

1. This paper sets out some of the key developments in the global landscape and regulation of online harm to help inform our discussions on opportunities and improvements for New Zealand's regulation.

Background

2. In our meeting on 17 December, I advised you that my Office was considering the Royal Commission's recommendations in their report on the March 15 attacks, in particular the recommendations for amendments to the Films, Videos and Publications Classification Act 1993 ("**the Act**"), taking into account the related recommendations for changes to hate speech legislation.

3. At our meeting I also referenced what we saw as the increasingly important topic of dangerous disinformation, and outlined the work we were commencing on a national survey on mis/disinformation in New Zealand.

4. Since our meeting we have witnessed the rioting and violence that took place in Washington on 6 January. My Office was on high alert from the commencement of the riots, as we were aware of the greatly escalated risk of violence and death occurring and being filmed, and subsequently distributed online – a risk that sadly came to pass.

5. Also since our meeting you have introduced a new Bill to the House; intended to counter violent extremism online – the Films, Videos, and Publications Classification (Urgent Interim Classification of Publications and Prevention of Online Harm) Amendment Bill ("**the CVE Bill**").

6. Much of the discussion precipitated by the introduction of that Bill has echoed that which has occurred in other jurisdictions as various governments have struggled with the challenge of regulating potentially harmful or illegal content online, and have considered where the balance lies between protecting the vital

freedoms of speech and access to information, while also providing necessary protections and legal boundaries for people online.

7. While the recommended amendment of the definition of 'objectionable' by the Royal Commission and the CVE Bill hold the promise of updating both *what* might be defined as objectionable or illegal content and the *tools* available to regulators to sanction such content respectively, we think there is also now an opportunity to think about *how* harmful content can be distributed, amplified and monetised in the digital age. The existing Act essentially reflects a pre-internet view of the world, and there is a limit as to how effective amendments built on this outdated base and regulatory structures will be. We think some international developments may provide signposts towards what a new digital-ready regulatory platform could look like for New Zealanders, serving to protect our essential democratic freedoms while also guarding against corrosive hatred and terror.

Initiatives in New Zealand that intersect with Online Harm

8. Many in New Zealand have had first-hand experience of how internet platforms can be weaponised by terrorists, having been exposed to the horrific livestream video from the March 15 attacks as it went viral on the Internet. Work on the changes proposed in the **CVE Bill** commenced soon after this event, and are intended to address actual or perceived shortcomings in the existing legislative regime as it relates to illegal terrorist content of this kind. We have reservations around the practical effect of some of the changes proposed in the bill, and have in particular sought engagement with the Department of Internal Affairs to express our views on the importance of prescribing limits, ensuring transparency and protecting human rights in primary legislation, rather than delegating these to regulation. We will be raising these points in our submission on the bill.

9. One of the significant challenges inherent in addressing online harm is the speed at which technology driven change occurs. The **Christchurch Call** launched in May 2019 is a hugely significant multinational and multi-industry accord capturing a range of voluntary commitments to 'eliminate terrorist and violent extremist content online'. Associated work by the Global Internet Forum to Combat Terrorism (**GIFCT**) and in particular the introduction of initiatives such as its Content Incident Protocol (**CIP**) has resulted in a significantly improved and more responsive internet when it comes to perpetrator-filmed terrorist content (as seen by this Office in industry responses to attempts to livestream attacks in Halle, Germany, in 2019 and in Glendale, Arizona in 2020).

10. Two weeks after the 15 March attacks Justice Minister Andrew Little announced a review of New Zealand's existing **hate speech legislation**, and subsequently the **Royal Commission of Inquiry** focused four out of its 44 recommendations on hate speech and hate crime – releasing a companion report dealing with these topics. The Prime Minister released a statement following release of the Royal Commission's report confirming that consultation would be undertaken with community groups and parties from right across Parliament to test these proposals before bringing forward legislative change.

11. Increasingly, attention is growing around the potential impacts and harms of disinformation, with the twin global drivers of Covid-19 disinformation and anti-vax rhetoric potentially generating public health risks on one hand, while on the other the contribution of conspiracy theories to violent extremism has been highlighted by events such as those that transpired in Washington in January. The Classification Office

formed the view last year that building an evidence base around disinformation in this country[1], and around New Zealander's experience of, and views on disinformation could provide an invaluable base from which to inform policy decisions. Work on a **nationwide survey on disinformation** is now well advanced (we are currently in the pilot phase with Colmar Brunton). We understand that **Cabinet** is also looking at advice on how to approach issues concerning disinformation.

12. There is currently no provision for the Classification Office to make a publication illegal simply for being false or misleading, and nor do we think there should be. Our interest in disinformation and conspiracy theories increased when we noticed that certain conspiracies around the origins of Covid-19 were resulting in publications depicting or promoting criminal acts (such as attacks on cell phone towers). That category of publication is or may be classifiable. Internationally, we have seen this type of issue has increased attention to categories of content that may be seriously misleading and even harmful, but which fall short of illegality. What to do about content that may be harmful but which is not illegal, and which may be subject to propagation or amplification by platform algorithms, is one of a number of issues that would be able to be explored in a fundamental review of New Zealand's media regulation. An intention to undertake a broad review of the **Media Content Regulatory System** that could include issues around social media was signalled by Minister Martin under the last government, but no substantive steps have been made in this direction.

13. We think there is an opportunity now to take stock of the various initiatives underway and to look to the latest thinking overseas to determine whether we can be more ambitious on behalf of the people of Aotearoa to both ensure that they have the protections they are entitled to expect, and also to reinforce their freedoms of speech and expression. We anticipate that any changes to hate speech legislation, or proposals that could filter content on the internet, will be strongly debated and tested. As they should be.

14. Given that fact, and the existing commitment to consultation and engagement on these issues, there may be an opportunity to broaden the discussion from the definitions and boundaries of illegal speech, and the enforcement tools that might be applied to such content, to also include:
    - the potential for tools and information, support and education;
    - broader engagement (including industry), on the role of and expectations of digital providers - ranging from digital industry behemoths such as Facebook and Google down to small, specialist forums. Key questions include what responsibilities should these providers have, and what expectations of transparency should be applied to them?

15. The latter areas strike us as particularly important in light of the Royal Commission's focus on **social cohesion** as a key area of social value that needs to be invested in and built up as part of a longer term strategy to build resilience against violent extremism. Evidence appears to be mounting that ignoring harmful and corrosive content online can be highly damaging to social cohesion. We are not aware of any change initiative being undertaken in New Zealand that squarely addresses this issue – and yet it is emerging as an area of very significant attention in a number of other jurisdictions that we have been speaking with, as outlined below.

16. Overall, we see opportunities to take a more joined-up approach to the various intersecting proposals and work impacting on harmful content, which could both improve engagement and value from consultation

[1] Supplementing work underway or already undertaken by academics and other agencies such as NetSafe.

(which currently risks going through iterative, piecemeal processes), and also to introduce relevant thinking and approaches from international initiatives in this area.

International Content Regulation – The Themes

17. Since the events of March 15, the Classification Office has greatly expanded its engagement with international policy makers, regulators, think-tanks and academics who are engaged in work on initiatives to deal with the pressing issue of addressing hate speech and terror online and the closely related issue of dangerous disinformation. Many of those we have spoken with are wrestling with the challenge of regulating the massive and opaque digital industry, where platforms with massive reach to a national population might not even have any physical presence in the country. Uppermost in the minds of many is also the critical need for balance – it is a simple reality that digital technologies, innovation and social media offer users huge benefits and utility, and regulation must not inhibit these aspects unnecessarily.

18. Many of those we have spoken with in jurisdictions such as Canada, Ireland, Australia, France, the UK and the EU have either recently announced or are close to announcing very significant and wide-ranging regulatory reform packages – underlining the degree to which content and digital regulation is widely seen as a global priority. Immediately before Christmas last year, the UK announced its response to the Online Harms White paper representing the culmination of several year's work, and the EU also announced its major Digital Services Act initiative.

19. While none of the initiatives we are seeing developed internationally are identical, even without undertaking any stringent or comprehensive comparative analysis, we can see that there are key common themes and strategies emerging from this international body of work. Attached at **Appendix A** is a comparative table illustrating some common features.

20. There appears to be a level of consensus around the key principles or pillars that should underpin thinking in this area – the French Mission's report "Creating a French framework to make social media platforms more accountable" detailing a framework for regulating social media platforms is one example of work that draws out a number of common themes or 'pillars':

    - **First pillar:** A public regulatory policy guaranteeing individual freedoms and platforms' entrepreneurial freedom.
    - **Second pillar**: A prescriptive regulation focusing on the accountability of social networks, implemented by an independent administrative authority and based on three obligations for the platforms:
        - Obligation of transparency of the function of ordering content,
        - Obligation of transparency of the function which implements the Terms of Service and the moderation of content,
        - Obligation to defend the integrity of its users
    - **Third pillar:** Informed political dialogue between the operators, the government, the legislature and civil society.
    - **Fourth pillar:** An independent administrative authority, acting in partnership with other branches of the state, and open to civil society.
    - **Fifth pillar:** A European cooperation, reinforcing Member States' capacity to act against global platforms and reducing the risks related to implementation in each Member State.

21. The obligation of transparency referenced in the framework above is an extremely common one – we have seen many similar statements on the importance of transparency from around the world, for example by requiring regular reporting from digital platforms, and empowering regulators to require disclosure of aggregate data, trends and information that may be important to understand risks of harm to users within the regulator state. An important subset of this is the growing awareness of the need for **algorithmic transparency**, so that harm and unintended consequences of the use of AI-assisted algorithms by such platforms can be identified and addressed before they become critical. This information can be helpful for government, industry and civil society.

22. The challenge of the highly diverse, variegated and vast range of different sizes and types of digital platforms and distributors of information and content online appears increasingly to be engaged with in a pragmatic way by many of these proposals by setting out a stepped regime – very large platforms, with associated greater resources (and also greater scope for harm) have the highest standards and expectations imposed, with mid and small operators having commensurately simpler and more appropriate requirements imposed on them.

23. Accordingly we see emerging from international work in this space a set of relatively pragmatic approaches to the key questions of *when* governments should consider regulation of the internet, and *how* this might be done in a way that protects citizen's rights and freedoms.

Summary

24. New Zealand currently has a range of work being undertaken to address various aspects of the very significant issue of content comprising hate speech, violent extremism and disinformation.

25. All of the Government's work in this area will attract significant attention and debate, and will also merit a significant commitment to engagement right across New Zealand, including with civil society, industry and the communities who may be both most affected by hate speech, and who may also be vulnerable to any proposals that may result in increased surveillance or enforcement. Given that, there may be scope to consider widening the engagement to also incorporate some of the key challenges brought about by the internet; the digital distribution model, the responsibility of digital platforms and providers, and the need for greater transparency in this space.

26. There could be scope for a coordinated approach to be taken; one that is informed by the latest international thinking in this area, as well as by engagement with those who are particularly impacted by this type of harm, such as the Muslim and Jewish communities, Māori, migrants, LGBT+ and young people.

27. We would be interested in discussing with you how such an approach might address the realities of the current digital industrial distribution model, provide opportunity to team up with industry but also to hold them to account, while also paying attention to the opportunities beyond regulation – the scope for education, tools and information, research and evidence, community support and the building of social cohesion.

Appendix A
International Comparison Table

Features	UK	EU	Australia	Canada	Ireland	France	Germany
Key legislation	**Online Harms legislation** Bill to be introduced in 2021. Government response to White Paper available here.	**Digital Services Act [proposal] and Digital Markets Act** *[include date and link].* The focus of the DSA is to create a safer digital space in which the fundamental rights of all users of digital services are protected	**Enhancing Online Safety Act 2015** [Include date and link – also note Australian legislators have been busy with recent amendments and specialist leg inc. the Criminal Code Amendment (Sharing of Abhorrent Violent Material) Act 2019	s9(2)(f)(iv) and s9(2)(g)(i)	**Online Safety and Media Regulation Bill 2020**	**Lutte contre la haine sur internet 2020**	**Network Enforcement Act (NetzDG)** "to improve law enforcement in social media"
Independent media regulator for online harm	**Yes** Ofcom [quango] – arm's length from government, independent from industry.	**Yes** New European Board for Digital Services. The European Commission will also have new powers in enforcement and monitoring – applies to very large platforms.	**Yes** E-safety Commissioner - arm's length from government, independent from industry		**Yes** Online safety commissioner	**Yes** Conseil Superieur de l'audiovisuel	TBC
Graduated / tiered system of regulation for companies	**Yes**	**Yes**	TBC		No set 'tiers' = but size of company/platform is a factor in 'designation' for regulation	TBC	**Yes** Regulation applies to companies over a certain size.
Transparency requirements	**Yes**	**Yes**	TBC		**Yes**	**Yes**	**Yes**

Note: There has been an enormous range of activity across the globe in response to issues around harmful online content, violent extremism and disinformation – the below chart simply provides a high level set of current examples. We see consistent themes around setting responsibilities on online providers to have the right mechanisms in place to prevent and/or address a variety of harms, along with increased transparency requirements. Future-proofing any moves in this space is also critical.

TE MANA WHAKAATU
Classification Office

3 March 2021

Hon Jan Tinetti
Minister of Internal Affairs
Parliament Buildings
WELLINGTON

Dear Minister Tinetti

Classification Office Research Project - National Survey on Misinformation/Disinformation

Purpose

1. Further to our meeting on 17 February 2021, this paper sets out a summary of the major research project that the Classification Office is undertaking this year.

Summary of topic

2. The Classification Office plans to carry out a nationally representative survey of the New Zealand public on the topic of misinformation, disinformation, and the relationship to conspiracy theories, extremism and real-world harms. The final survey questions are attached in Appendix A.

Who is conducting the survey?

3. We have commissioned Colmar Brunton to conduct the survey.

Who will be surveyed?

4. A nationwide online survey will be conducted with 2,000 people aged 18 years and over, with a booster sample of 300 people, aged 16 to 17 years old (involvement subject to parental/caregiver consent). Participants will be sourced using online panels (Colmar Brunton and Dynata). For the main sample of 2,000 people, we will set quotas based on the 2018 Census population characteristics for age by gender, region, and household income by household size (as a proxy for socio-economic status).

Who was consulted during the survey design?

5. The Classification Office has consulted with academics and researchers (both internationally and in New Zealand), government agencies with an interest in CVE and/or mis/disinformation, and NGOs (a full list is included in Appendix B). This consultation was preceded by a literature review of the existing studies on mis/disinformation, and was informed by reviewing how similar studies were constructed overseas.

When?

6. We have completed our research design, consultation, cognitive testing and pilot phases of the survey with Colmar Brunton. This is our current timeline for the remaining steps:

Task	Current dates
Main fieldwork	18 February – 17 March
Topline findings (in the form of data tables)	22 March
Draft report from Colmar Brunton	12 April
Updated data tables and raw data	13 April
Publication and launch of research	Likely to be June

7. The Classification Office will draft and finalise the research report once we have the final data and initial analysis report from Colmar Brunton.

Why this topic?

8. There are increasing concerns in New Zealand, and globally, about the prevalence of misinformation and disinformation – especially online – and the potential for this to cause harm, contribute to extremism, and impact on important matters such as democracy, public health and public safety. Through our work, we have seen how the spread of false, and sometimes hostile, disinformation and conspiracy theories continue to impact on communities during the Covid-19 pandemic, and how extremist theories and ideology can contribute to real-world violence such as the March 15 terrorist attacks in Christchurch.

9. As with other complex issues relating to online content – such as the impact of pornography on young people – dealing with the spread of false information requires a joined up approach that looks at regulation, information and education. To achieve this, we see a need for robust, up-to-date evidence about the scope of these issues in New Zealand, with the understanding that what needs to be researched and evaluated is growing. To date, there appears to be few nationally representative studies (in New Zealand or internationally) that cover the range of interconnecting issues relating to mis/disinformation and how this may lead to real world harms and a general loss of trust.

10. The spread of mis/disinformation encompasses a broad set of issues that no single agency has responsibility for. We believe that responding to these complex and cross-cutting issues requires a connected-up approach amongst government agencies with regulatory oversight, alongside NGOs, educators, mainstream media organisations and community groups, while at the same time supporting and engaging the public.

What will happen with the findings from this research?

11. Data will help us unpack how mis/disinformation and conspiracy theories operate in relation to New Zealanders' online experiences and the relationship to large social media platforms and other digital ecosystems, where this information is often shared. It will inform our classification approach to publications that are related to conspiracy theories or appear to be inspired by mis/disinformation to endorse criminal or violent acts. It will also add to the evidence base for understanding how disinformation actors, extremist and fringe groups spread false information – and the negative impacts this may have on wider society, in particular for marginalised communities.

12. The findings from this research will be published and promoted with the aim of reaching a wide audience including parents, teachers, government agencies, and the general public. We will continue to engage

with stakeholders to make the best use of this research. We are developing a communications plan to assist with this process, and will keep you advised on this.

13. Findings from the research may inform the development of:
    - Cross-government work on potential policy and regulatory responses,
    - Information and resources for the public, including support for schools and students in media literacy and critical thinking,
    - Frameworks and strategies for the handling and treatment of related extremist promotional material.

Our role: The Classification Office

14. The Classification Office Te Mana Whakaatu is an independent Crown entity responsible for classifying publications that may need to be restricted or banned. The legal definition of a 'publication' covers a wide range of mediums such as films, videos, music recordings, books, magazines, video games and online content. We conduct research and produce evidence-based resources to promote media literacy and enable New Zealanders to make informed choices about what they, and their tamariki watch.

15. The Classification Office has no mandate to restrict or ban content simply on the basis of fairness, balance or accuracy. However, we do have a mandate to restrict material that could encourage behaviour that poses a risk of self-harm or harm to others, and material that is promotional of criminal, terrorist or violent acts.

16. Worldwide, evidence seems to be accumulating of conspiracy theories contributing to violent or criminal acts. In the NZ context, we have seen an increase in attacks on cell towers which appear to be correlated with the appearance of 5G conspiracy theories. Our Office did an internal assessment of a video of an attack on a cell-tower in New Zealand that was circulating on social media during lockdown. While we did not consider that it reached the threshold for "objectionable" under our legislation, it appears to be evidence of a global conspiracy theory inspiring real world, harmful acts. Another local example is the Christchurch Mosque terrorist attacker's "manifesto" document, *The Great Replacement*. This is based on a conspiracy theory of elitist complicity in systematic racial replacement of white people through (amongst other things) mass migration, demographic growth and a drop in European birth rates.

17. Budget 2020 enabled the Classification Office to establish a dedicated team working on countering violent extremism, with a focus on online content promoting terrorism and violence. The new CVE team has a specialist classification role, as well as including a research, education and outreach function and we're proactively engaging with New Zealand and overseas government agencies, academics, and experts at the forefront of countering violent extremism, to share insights and identify solutions. The initial work of this team has highlighted the linkages between mis/disinformation and extremist material, and identified this as an area that we need to understand better.

Ngā mihi nui,

David Shanks

Chief Censor

Classification Office Te Mana Whakaatu

Appendix A: Final Survey Questions

Refer below and attached to the email with this briefing (dated 25 February 2021)

Questionnaire 11 Feb FINAL.pdf

Appendix A withheld in full, pursuant to s18(d). Soon to be made public.

Appendix B: Stakeholders Consulted

Academics and Researchers
We consulted with the research group from Te Pūnaha Matatini who are working on misinformation emerging from the Pandemic. The research group includes experts in conspiracy theories, survey design, discourse analysis and sociology from Waikato University, University of Auckland, Auckland University of Technology, and University of Canterbury. We also separately consulted with academics and researchers with relevant expertise from Massey University, Otago University, the London School of Economics and Political Science, the Institute of Environmental Science, Boise State University, University of Auckland, Auckland University of Technology, and Victoria University of Wellington.
Government Agencies
Department of Prime Minister and Cabinet, Department of Internal Affairs, Ministry of Education, Ministry of Health, Broadcasting Standards Authority, Department of Conservation, Public Services Commission, NZ Police, and the Ministry of Justice.
Other
Transparency International New Zealand
Le Va