DEPARTMENT OF THE
PRIME MINISTER AND **CABINET**
TE TARI O TE PIRIMIA ME TE KOMITI MATUA

16 May 2024

[redacted]

Reference: OIA-2023/24-0683

Dear [redacted]

Official Information Act request relating to security briefings related to AI

Thank you for your Official Information Act 1982 (the Act) request received on 29 March 2024. You requested:

> *"Specifically, I am looking to obtain any information, intelligence reports or briefing notes pertaining to identified risks and/or opportunities in the area of Artificial Intelligence (AI) over the last ten years, which may have any direct relevance to Counter-Terrorism and national security in New Zealand."*

On 29 March 2024, we requested a refinement from you. You responded with the following refinement:

> *For the part seeking: any information, intelligence reports or briefing notes*
> - *Keeping my OIA request to just official written reports/briefings and advice provided for the Prime Minister or Minister for National Security and Intelligence would be appreciated. Additionally, official reports produced by DPMC for the public (or possibly to other government departments) would also be helpful if available.*
>
> *For the part seeking: the last ten years*
> - *Can I amend this to the last five years please.*
>
> *For the part seeking: Counter-Terrorism and national security in New Zealand*
> - *If possible, I am looking to obtain AI-relevant information which falls within the context of counter-terrorism, counter-extremism and public safety. These are the areas of AI and national security I am interested in exploring further please.*
> - *In essence, I am looking to research the potential risks, opportunities and threats posed by developing AI technologies, from specifically a Counter-Terrorism and public safety (New Zealand) perspective - this may include cyber, physical and political security areas.*

Information being released

One document has been identified as relevant to your request but is being withheld in full under s 6(a) of the Act.

Also identified as relevant to your request are some briefings provided by the Department of the Prime Minister and Cabinet's (DPMC) Policy Advisory Group to the Prime Minister. These briefings are provided to the Prime Minister in confidence to support him in his role as leader of the Government and chair of Cabinet. These briefings are withheld in their entirety under the following sections of the Act:

- section 6(a), to protect the security or defence of New Zealand or the international relations of New Zealand.
- section 9(2)(f)(ii), to maintain collective and individual ministerial responsibility

- section 9(2)(f)(iv), to maintain the confidentiality of advice tendered by or to Ministers and officials
- section 9(2)(g)(i), to maintain the effective conduct of public affairs through the free and frank expression of opinion.

The following documents have been identified as relevant to your request and we are releasing them, subject to information being withheld, under the following sections of the Act:

- section 6(a) to protect the security or defence of New Zealand or the international relations of the Government of New Zealand
- section 6(b) to protect the entrusting of information to the Government of New Zealand
- section 9(2)(a), to protect the privacy of individuals
- section 9(2)(ba)(i), to prevent damage to the public interest
- section 9(2)(g)(i), to maintain the effective conduct of public affairs through the free and frank expression of opinion
- section 9(2)(j), to enable negotiations to be carried on without prejudice or disadvantage.

Item	Date	Document Description/Subject
1.	8/04/2022	Briefing Update on Algorithmic workstreams
2.	12/08/2022	Excerpt from PM's weekly update
3.	13/07/2021	Speech notes for NSCAI conference

This topic covers several aspects of AI and machine learning (ML), and there are elements of this within the Christchurch Call commitments and work programmes. As the Call was being developed and implemented, a range of aspects of AI and ML were considered and worked on. Information on these can be found on the Christchurch Call website.

Information publicly available
The following information is also covered by your request and is publicly available on DPMC's website and the Christchurch Call website (www.christchurchcall.com):

- Proactive Release: Christchurch Call Research Partnership (dpmc.govt.nz)
- Proactive Release: Update on the Christchurch Call to Eliminate Terrorist and Violent Extremist Content Online - September 2023 - Department of the Prime Minister and Cabinet (dpmc.govt.nz)
- Christchurch-Call-Leaders-Summit-2023-Supporting-papers.pdf (christchurchcall.com)
- Christchurch-Call-Leaders-Summit-2023-Joint-Statement-ENG.pdf (christchurchcall.com)
- Christchurch-Call-Leaders-Summit-2023-Supporting-papers.pdf (christchurchcall.com)
- Christchurch-Call-2022-Community-Update.pdf (christchurchcall.com)
- Christchurch-Call-Joint-Statement-2022-English-version.pdf (christchurchcall.com)
- Christchurch-Call-2nd-Anniversary-Summit-Co-chair-Statement-2021.pdf (christchurchcall.com)
- Algorithms-and-Positive-Interventions-WorkPlan.pdf (christchurchcall.com)

Accordingly, I have refused your request for the documents listed above under section 18(d) of the Act – the information requested is publicly available.

In making my decision, I have considered the public interest considerations in section 9(1) of the Act. No public interest has been identified that would be sufficient to override the reasons for withholding information.

You have the right to ask the Ombudsman to investigate and review my decision under section 28(3) of the Act.

This response will be published on DPMC's website during our regular publication cycle. Typically, information is released monthly, or as otherwise determined. Your personal information including name and contact details will be removed for publication.

Yours sincerely

[redacted]

Janine Smith
Deputy Chief Executive, Policy

DEPARTMENT OF THE PRIME MINISTER AND CABINET
TE TARI O TE PIRIMIA ME TE KOMITI MATUA

CHRISTCHURCH CALL

Briefing

UPDATE ON ALGORITHMIC WORKSTREAM

To Rt Hon Jacinda Ardern, Prime Minister			
Date	8/04/2022	**Priority**	Routine
Deadline	6/05/2022	**Briefing Number**	DPMC-2021/22-1916

Purpose

To update you on the Christchurch Call algorithmic work, highlight some of the problems the Community is working on, describe a sample of approaches from researchers, tech companies, and policymakers, s 6(a) [redacted]

Recommendations

1. **Note** that notwithstanding platforms' sensitivities, the Call Unit has made some progress identifying blockages, finding ways forward, and concentrating effort across the Community; and

2. s 6(a) [redacted] **YES / NO**

s 9(2)(a) [redacted]	Rt Hon Jacinda Ardern **Prime Minister**
8 / 4 / 2022	......./......./......

Contact for telephone discussion if required:

Name	Position	Telephone	1st contact
s 9(2)(a) [redacted]	Christchurch Call Coordinator	s 9(2)(a) [redacted]	✓
s 9(2)(a) [redacted]	Chief Advisor		

Minister's office comments:

- □ Noted
- □ Seen
- □ Approved
- □ Needs change
- □ Withdrawn
- □ Not seen by Minister
- □ Overtaken by events
- □ Referred to

UPDATE ON THE CHRISTCHURCH CALL'S ALGORITHMIC WORKSTREAM

Summary

The Call's algorithmic commitments are the most complex and sensitive for the tech sector. In addition to the technical complexity of this work, tech leaders are juggling proprietary interests, competitive dynamics, and regulatory uncertainty. s 9(2)(g)(i) [redacted]

While the Global Internet Forum to Counter Terrorism (GIFCT) has tried to play a bridging role behind the scenes, it has faced obstacles 9(2)(g)(i) and 9(2)(ba)(i) [redacted] s 9(2)(g)(i) [redacted] In that context, the Unit has sought to create a series of options that provide multiple possible means of advancing this work.

Your intervention with tech leaders in the lead up to the 2022 Christchurch Call Leaders' Summit could be very helpful in cementing a few targeted 'asks'. We have suggested some options that can be developed further, subject to your feedback:

- Granular and localised prevalence data for various content types.
- A shared list of best practises, drawing on the Integrity Institute's recommendations.
- A pilot study on personalisation features and/or radicalisation
- A structure and funding to commission new tools for moderation and intervention.

Purpose

To update you on the Christchurch Call algorithmic work stream, highlight some of the problems the community is working on, describe a sample of approaches from researchers, tech companies, and policymakers, s 6(a) [redacted]

Report

1. As you know, the Call contains several commitments related to the important role algorithmic processes play in the management of terrorist and violent extremist content (TVEC) online:

Online Service Providers (OSPs)	OSPs and Governments together
6. Take measures to prevent the upload and dissemination of TVEC, including through cooperative technology development 11. Review the operation of algorithms that may drive users towards TVEC, and implement changes, and mechanisms for reporting	11 and 14. Facilitate the development and deployment of interventions and the redirection of users 15. Accelerate technical development of tools for detection and removal. 18. Support smaller platforms through sharing of technical solutions.

2. There has been progress against these commitments, including the continuous improvements made to the GIFCT hash database, the broadening of hash sharing members, and the deepening of the taxonomy. There have also been continuous improvements made to platforms' algorithmic moderation systems. As a result, the reported prevalence of TVEC on major OSPs is now extremely low.

3. This has meant a shift in focus, from the issue of algorithms helping to drive users towards TVEC, to the broader question of how the user environment facilitates radicalisation and amplification of non-violative and legal content. As part of the mandates developed for the May 2021 Leaders' Summit, the Community prioritised work to assess data and information needs of researchers to understand user journeys, and the impact of platform features on **radicalisation** and **amplification**.

4. Behind all this sit two core policy questions:

   a) Do content recommendation systems based on artificial intelligence and machine learning (AI/ML) have the potential to 'amplify' or drive users towards TVEC, or towards radicalising ecosystems or networks of users that connect with TVEC off-platform?

   b) Could recommender systems increase the exposure of 'at risk' individuals to content and user networks which supply the ideological justification and means/strategies for achieving violent ends?

5. s 9(2)(ba)(i) work with New Zealand as 'co-leads' of a regular Christchurch Call algorithmic working group meeting with around 70 people from across the Call Community, plus a few outside experts. This group helps oversee work taking place across the Community and advises where efforts are needed to realise our work plan objectives.

Getting answers is difficult, because of the nature of the systems being tested

6. AI/ML recommender systems are a **moving target** because they adapt over time from their interactions with users, as well as through iterative 'training' carried out by their owners. A system consists of multiple AI/ML algorithms. For example, one group of algorithms might identify particular characteristics (e.g. offensive language, graphic violence, manipulated imagery) and another might decide what to do with them based on a combination of identified characteristics (e.g. promote, demote, remove, send for human moderation). Multiple sets run in parallel serving different objectives (e.g. elevate high quality content, remove spam, delete harmful content etc.) Each of them is constantly adapting in response to new inputs including those that result from other algorithms running alongside. That means that the conclusions that can be drawn about a system at a particular point in time aren't necessarily reflective of the system as it will configure itself down the track.

7. AI/ML systems operate with **uncertainty**, which means they may be more or less certain about something they've been trained to identify and will make decisions based on what is deemed an acceptable level of uncertainty. False positives and false negatives are a feature of that system, as are potential biases and perverse consequences. There are real trade-offs to consider in managing this, as reducing uncertainty may come at the cost of bias (i.e. an algorithm can be more certain about content relating to a particular racial or language category than another). This can vary between content types. For instance, AI/ML systems are very good at classifying nudity, or finding copyright music, but much worse at assessments that involve context, e.g., bullying, satire, or irony.

8. Due to **personalisation features,** the outputs of AI/ML recommender systems can vary depending on the characteristics and behaviour of the individual user. This can include who they are friends with, their previous activity, or other characteristics the system has learned to associate, which even the programmer may not be aware of. This makes repeatable simulation difficult and creates a range of potential privacy law and data protection concerns as well as ethical questions for researchers around testing of users.

9. It is also difficult to test theories around amplification of TVEC on major platforms because **TVEC isn't supposed to be there**. Social media platforms participating in the Christchurch Call have measures to prevent upload and to remove TVEC. Content that sneaks through these systems can't easily be tested because it's hard to identify, and once identified is immediately removed.

The Call Community has developed some ideas around how to get past these difficulties

…Suppressing borderline content …

10. A number of stakeholders have been working on the basis that the indirect route to address amplification and radicalisation is through identifying and suppressing so-called 'borderline content' or 'grey-zone content.' That content is not illegal and does not violate companies' terms of service but can be identified in various ways as being close to the 'policy line'. Facebook founder Mark Zuckerberg in his 2020 testimony to Congress suggested that engagement-based systems drive users closer to the policy line, which could be a 'gateway' towards more extreme ideas and beliefs.

11. There have been some empirical studies on Facebook and YouTube to test this theory, with a range of conflicting and inconclusive observations. A key challenge is how to measure 'extremeness' (i.e. agreed metrics). In meetings of the Christchurch Call algorithmic working group, s 9(2)(ba)(i) have both called for further work on taxonomies and metrics for harm arising from borderline content which could be used to conduct studies. At present, the OSPs who proactively identify 'borderline content' - notably Meta and Google - tend to rate it against a range of quality metrics (e.g. whether information cites credible sources or makes qualified claims), rather than assessing harm.

12. The EU Internet Forum is commissioning a large study into the impacts of borderline content that could help to inform this. s 9(2)(g)(i) Previous attempts by the global advertising industry, through their 'Global Alliance for Responsible Media' (GARM), to develop uniform harm and prevalence metrics have proved extremely difficult s 9(2)(g)(i)

13. Members of the Christchurch Call Advisory Network have emphasised in these sessions that risks and harms are most effectively identified by the communities most impacted by them, and that improved transparency could help deliver a dynamic picture of how or whether system improvements are helping reduce those impacts.

14. The Call Unit has helped facilitate a study through the Global Partnership on Artificial Intelligence, involving s 2(ba)(i), in which community panels will be composed to help develop nuanced and locally contextualised harm metrics for Aotearoa New Zealand. This will help s 9(2)(ba)(i) to improve its content ranking system and to address a range of other AI/ML-specific challenges such as how to distinguish reclaimed language from offensive language and slurs.

15. The United Kingdom is considering approaches in its Online Safety Bill that could restrict certain categories of harmful but legal content. This has been widely criticised as government-mandated censorship. There are fine distinctions within a 'duty of care' or 'risk-management' system in terms of how much of the burden rests on definitions of content that originate with government vs platforms managing identified risks or harms through their own definitions. There is also a 'third way' in which governments and/or platforms facilitate conversations with community groups about harm definitions – as the Call Unit is seeking to do with s 2(ba)(i)

... changing the objectives of recommender systems

16. The Integrity Institute – a non-governmental organisation composed of former and current trust and safety employees at social media firms – has made a range of suggestions about recommender systems. They suggest the best approach is to move away altogether from an engagement-based model which risks promoting highly engaging adversarial narratives, whether or not these happen to be captured by an agreed definition of 'borderline'. Firms should instead optimise their recommender systems for 'quality' with the relevant metrics defined according to their corporate values system. These would need to be simpler and more values-based than existing terms of service or community standards.

17. YouTube takes this approach in its 'quality rater' guidelines which combine human and AI systems to identify good and bad quality content and either prioritise it for recommendation or suppress it in various ways. The lowest quality rating is attached to content which is potentially harmful as well as misleading or disreputable. Similar rankings are applied to users and groups. This approach has been phased in progressively since December 2019, in part informed by discussions around the Christchurch Call s 2(g)(i) [redacted]

18. However, this model is cost intensive, and the results are highly dependent on where selective investments are made to improve quality (e.g. in appointing panels of medical experts to rate information on COVID-19). It can also be circumvented e.g. if a user chooses to subscribe to channels that host borderline content, and we understand optimisation for user engagement continues to be used alongside the quality-based rankings. There are well publicised-issues, for instance with poor quality metrics for non-English language content, adding a further layer to the idea of a digital divide.

...audits and reporting mechanisms

19. The only way to reliably understand whether changes to AI/ML-based systems are helping reduce the risk from amplification and radicalisation is through independently verified research. As part of a package of measures in the draft EU Digital Services Act, certain AI/ML systems including recommenders operating on large social media platforms would need to be opened for 'audit' by authorities or independent researchers. If retained in the final version of the Act, and once in force (i.e. in several years), this could provide the means to identify and potentially correct harmful biases.

20. Such audits are easier prescribed than done. Separating the outputs of an AI/ML based system from the inputs of human users and human moderators who are fully integrated within these systems requires some innovative approaches to experimental design, as well as the willing cooperation of the platform concerned. While we are some way away from having a universal approach to multistakeholder reporting mechanisms across the whole Christchurch Call Community, there have been promising developments. The Call's Algorithms working group has narrowed down some of the questions and information needs in this area.

21. The GIFCT is working towards identifying viable and credible experimental approaches as part of its technical approaches working group. A/B testing is a method that can measure the impact of algorithmic tweaks across large samples of users. s 9(2)(j) [redacted]

22. We also continue to work with the Global Partnership on AI, s 9(2)(ba)(i) and an international consortium led by University of Otago researcher s 9(2)(a) on a proposal to test s 9(2)(ba)(i) recommendation system and assess how it treats different categories of TVEC-adjacent borderline content relative to a reverse chronological feed. The primary issue at this stage is around creating a robust metric to identify TVEC-adjacent content which remains a difficult work in progress.

There are two to three emerging regulatory approaches

23. There are a range of emerging regulatory approaches in this area. The EU approach is well elaborated and in late-stage negotiations, while the emerging features of any US approach are likely to be somewhat complementary in their focus on managing systemic risks with oversight. There is also an emerging Chinese approach which sets centrally the objectives and morals that should govern recommender systems.

24. In anticipation of the EU Digital Services Act, France has lifted provisions from the draft EU bill introducing a 'duty of care' principle and a range of transparency provisions into French law. That will mean that France's regulator will set some of the early technical principles around how platforms provide reasonable access for external assessment of AI/ML recommender systems and put in place mitigations for systemic risks that they identify.

25. In the US, several proposals have been tabled, including a bipartisan proposal from the Senate that would allow vetted independent researchers access to data to research biases and risks, with oversight from the Federal Trade Commission (FTC). A range of House Democrats have tabled proposals to amend intermediary liability protections for social media platforms, make them dependent on transparency and other conditions, or remove those protections altogether in cases where content is amplified. One draft bill proposes to outlaw social media algorithms that discriminate based on protected characteristics. It is likely that some of the common elements of these draft bills will gain momentum over time, particularly around strengthened FTC regulatory oversight, and there appears to be a high level of commonality with the features of the EU's Digital Services Act.

26. China has passed a law on social media algorithms which sets out a range of requirements, including transparency for users about recommendations and the ability to 'opt out' of recommendation features. It requires algorithmic systems to follow a range of ethical guidelines including to "prioritise mainstream values' and "create positive energy". In effect, this also creates categories of legal 'undesirable content' which should be removed or de-prioritised.

27. These three main 'regulatory blocs' are likely to be the most influential in shaping algorithmic regulation in the future. Others, including the UK, Canada, Ireland and Australia are at different stages of developing legislation in this area, with a range of common features reflecting the desire to manage 'harms' and harmful behaviour that forms part of the user environment but which doesn't map directly onto illegal content.

28. In New Zealand, our current regulatory system has no specific requirements on transparency, nor algorithmic risk management or oversight. The Department of Internal Affairs' content regulatory systems review presents an opportunity for New Zealand to innovate in this area, and OSPs have indicated a willingness to do more at a local level including through the voluntary code of practice being developed with NetSafe. The Unit would be supportive of more fully leveraging our collective expertise, and that of the New Zealand technical community (e.g. Internet New Zealand) in the development of innovative regulatory options.

We can improve on this work through greater company engagement

29. As reported in our briefing on the GIFCT of 17 December 2021 and aide-memoire of 9 February 2022, there have been issues with enabling s 9(2)(g)(i) s 9(2)(g)(i) the outstanding issues identified by the Christchurch Call algorithms group. For instance:

- s 9(2)(g)(i) and s 9(2)(j) As a result, pilot studies are generally taking place in areas and on topics where companies' trust and safety teams are most willing to welcome specific researchers in, rather than areas that are most pressing.

- s 9(2)(g)(i) and s 9(2)(j) Its technical approaches working group has identified a large 'priority list' including for example systems to classify audio content which would be useful for a range of smaller GIFCT members such as s 9(2)(b)(ii) Some NGOs are calling for an 'innovation fund' to help deliver these tools.

30. s 9(2)(g)(i)

This is linked to issues around mis-, dis-, and mal-information

31. Algorithmic amplification and the role of online user experiences in radicalisation are both important questions in the context of work on dis-/mis-/mal-information. Any outcomes that provide researchers and campaigners with better tools to understand and counter harmful disinformation narratives will also be useful in this space, hence it is helpful to maintain a close dialogue between the Call Community and any structures that are built to help address disinformation.

s 6(a) and s 9(2)(j)

[redacted] [redacted]

a) **Provide granular or local prevalence data,** s 9(2)(j)

The first draft of the NetSafe code of practice on online safety suggested that major social media platforms would be willing to provide regular transparency reporting on New Zealand content actioned under a range of categories. This would be a good start. However, a far richer conversation about algorithmic impacts would be possible if certain platforms were prepared to provide granular prevalence information at a local level.

"Prevalence" reporting involves providing information about how many people saw a piece of content before it was identified and removed (or actioned in other ways). This kind of reporting is often based on sampling and can be difficult to compile, thus it goes beyond the current practise in transparency reports.

Adding **'granularity'** i.e. a richer set of information about that content could be really helpful in assessing the improvements being made to AI/ML content moderation e.g. to understand whether certain types of users were being targeted. This would be helpful across a range of categories e.g. hate content, disinformation content and identity-linked conspiracy theories, and other forms of content that play a role in violent extremism activity online.

Such reporting would help New Zealand-based researchers and practitioners to reach better-informed conclusions about the online ecosystem and the risks that arise from terrorist and violent extremist groups, and to develop locally-tailored positive interventions.

An ideal version of this would see regular interaction with communities coming together with regulators and platforms to discuss the data in time series, as well as their direct experiences to help fill out a dynamic view of the evolution of harms, and the impact of system improvements and interventions.

b) **Develop best practices** for more actively managing risks around radicalisation and amplification without requiring new content definitions.

s 6(a) [redacted] it is difficult and at times problematic when governments seek to regulate by defining new content categories for legal and non-violating content. There is a clear preference across the Call Community for interventions that don't require this.

A list of agreed best practices could be helpful in highlighting what tools are already available and could be implemented more widely. The Integrity Institute has developed some ideas which serve as a useful starting point, including:

- Limiting the ability for new users, or users who post low quality or violating content to reach large audiences.
- Optimising AI/ML recommender systems for content quality.
- Introducing 'nudges' for users to reduce adversarial behaviour online.
- Doing integrity testing, drawing on the expertise of ex-radicalised people and/or impacted communities.

c) Help the Call community to scope out a **study on personalisation features**, potentially with reference to former or ex-radicalised people, or by looking at how user interactions lead to the consumption of TVEC off-platform (e.g. via out links).

We have a major knowledge gap around personalisation features and the impacts they may have on at-risk users being exposed to radicalising content. Conducting a study might help to build confidence that this gap is being addressed proactively.

While this isn't the only viable approach, it may be possible to look at the online 'stepping stones' taken by ex-radicalised people and to use this as a basis for study.

One option s 9(2)g(i) [redacted] would be to use "A/B testing" to effectively test two different versions of a recommender system at a large scale. This could be a good test case for the GIFCT's work on methodological principles.

d) Work with the Call Community to put in place a multistakeholder structure and additional funding – either in GIFCT or elsewhere – to develop and make available **algorithmic tools** for risk management, intervention, and identification and removal of TVEC.

At present the GIFCT hash sharing database remains a highly effective system for blocking photo and video content, s 6(b) [redacted]

There remain a range of gaps, including: audio classifier algorithms; hash-matching for terrorist audio content; and workflow systems to help input open-source information about terrorist web content shared through multiple platforms.

There is also a lack of knowledge sharing on systems for intervention, and in the availability of such tools for smaller platforms who aren't able to shoulder the development costs themselves.

We understand that some companies may be sensitive around this, particularly as they look to commercialise their own tools for sale to small platforms. However, this would fulfil commitments made in both the Call text, and the companies' 9-point plan from 2019 (see Attachment A).

Next Steps

33. The Call Unit welcomes your feedback on the options outlined in paragraph 32 above and subject to your views, can prioritise this list s 9(2)(j) [redacted] Achieving even one or two of these asks would mark substantial forward progress in the Call work plan and a big contribution by New Zealand to a global community looking for a safer and more civil internet. The Call Unit would be available to discuss these with you prior to your visit, if helpful.

Attachments:	Classification:	Title:
Attachment A:	Unclassified	Table of Relevant Christchurch Call Commitments and Work Plan Objectives

Attachment A:

Relevant Call Commitments and Work Plan objectives

Pillar 1: Recommender algorithms and User Journeys

Christchurch Call Commitments		2021 Work Plan Objectives
11	**Review the operation of algorithms and other processes** that may drive users towards and/or amplify terrorist and violent extremist content to better understand possible intervention points and to **implement changes** where this occurs. […] This may include **building** appropriate **mechanisms for reporting**, designed in a multi-stakeholder process and without compromising trade secrets or the effectiveness of service providers' practices through unnecessary disclosure.	The Call Community will devote effort and resources to better understanding the "user journey" and the role this may play in the broader radicalisation process. **We will design a multi-stakeholder process** to establish what methods can safely be used and what information is needed - without compromising trade secrets or the effectiveness of Online Service Providers' practises through unnecessary disclosure- to allow stakeholders to better understand the outcomes of algorithmic processes and their potential to amplify terrorist and violent extremist content.

Pillar 2: Detection and Removal of Terrorist & Violent Extremist Content

Christchurch Call Commitments		2021 Work Plan Objectives
6	**Take transparent, specific measures** seeking to prevent the upload of terrorist and violent extremist content and to prevent its dissemination on social media and similar content-sharing services, including its immediate and permanent removal, without prejudice to law enforcement and user appeals requirements, in a manner consistent with human rights and fundamental freedoms. Cooperative measures to achieve these outcomes may include technology development, the expansion and use of shared databases of hashes and URLs, and effective notice and takedown procedures.	This year the Community will host an inclusive discussion on developing a framework to continuously review and improve the efficiency of [complaints and appeals processes] and support greater transparency and explainability in this area.
		2019 9-Point Plan announced by Amazon, Google, Microsoft, Facebook, and Twitter
15	**Accelerate research into and development of technical solutions** to prevent the upload of and to detect and immediately remove terrorist and violent extremist content online, and share these solutions through open channels, drawing on expertise from academia, researchers, and civil society.	"We commit to working collaboratively across industry, governments, educational institutions, and NGOs to develop a shared understanding of the contexts in which terrorist and violent extremist content is published and to improve technology to detect and remove terrorist and violent extremist content more effectively and efficiently. This will include: • Work to create robust shared data sets to accelerate machine learning and AI and sharing insights and learnings from the data. • Development of open source or other shared tools to detect and remove terrorist or violent extremist content. • Enablement of all companies, large and small, to contribute to the collective effort and to better address detection and removal of this content on their platforms and services.
18	**Support smaller platforms as they build capacity to remove terrorist and violent extremist content,** including through sharing technical solutions and relevant databases of hashes or other relevant material, such as the GIFCT shared database.	

Pillar 3: Positive Interventions

	Christchurch Call Commitments	2021 Work Plan Objectives
11	**Review the operation of algorithms and other processes** […] This may include **using algorithms and other processes to redirect users** from such content or the promotion of credible, positive alternatives or counter-narratives.	[We will...] **Empower a new generation of community-driven online interventions** This year the Call Community working with the GIFCT will seek to identify and empower the next generation of digital interventions against radicalisation, working to build a consistent framework for comparative evaluation… Governments will work in an open multi-stakeholder context to identify information that could be shared to assist with positive interventions.
14	**Develop effective interventions**, based on trusted information sharing about the effects of algorithmic and other processes, to redirect users from terrorist and violent extremist content.	

OIA Response: PM Weekly Report 12/8/2022

Weekly Update

To: Prime Minister (Rt Hon Jacinda Ardern)					
Date	12/08/2022	Report number	DPMC-4598189	Priority	Routine

2022 Christchurch Call Leaders' Summit

Overview

1. This report provides an overview of preparations for the 2022 Christchurch Call Leaders' Summit scheduled for 20 September 15h30-17h30 EDT in New York. We will provide you with weekly updates in the lead up to the Summit, and bespoke advice on particular issues as required.

Remainder of briefing is out of scope…

10. We anticipate four main policy outputs from the 2022 Summit:

a) A new **technological tool for studying algorithmic impacts:** developed through a partnership of the New Zealand government with the United States, Twitter, Microsoft, and OpenMined (an open-source research consortium and software developer):

- This initiative follows directly on from your discussions with Twitter in San Francisco and Microsoft in Redmond, and subsequent work with both firms and other partners. It also contributes to the commitment in the joint statement with President Biden to announcing new measures in this area.

- If successful, this approach will enable independent researchers to conduct research on algorithmic outcomes and impacts without compromising user privacy, in a manner that is scalable and allows for the study of impacts across multiple platforms.

- It should lead to a better understanding of what drives online radicalisation, how terrorist content spreads across platforms, and what social media companies, community organisations, and governments can do to make the environment safer and more user-friendly.

- It could have application and a positive legacy beyond the Christchurch Call, enabling investigations into a wider range of possible impacts and biases and opening a new field of research into social media algorithms and the information environment.

- The Unit is working with partners on how to communicate this outcome to the public. We will prioritise giving it a distinct Christchurch Call brand identity, and ensuring we give credit to the effort and resources invested by the four main partners, including New Zealand.

- We are also working on some other, related lines of effort. These include work under the Global Partnership on Artificial Intelligence (GPAI), s 6(a). The GIFCT continues to work on algorithmic issues; we have emphasised the need for this work to continue, while s 6(a)

Remainder of briefing is out of scope…

Draft PM Remarks
National Security Commission on AI: Global Emerging Technology Summit
13 July 2021

[Mihi]

- It is a great pleasure to be a part of this important conversation hosted by the National Security Council on AI. Thank you to the Commission, to Chairman Eric Schmidt for the invitation and to all of the contributors who are here today.

- Open democratic societies like ours must show that they can thrive and hold true to their values even as they lead the adoption of new technologies.

- We need to show that our model of participatory democracy, and transparent decision-making – in addition to the inherent advantages it confers – offers us a better way of managing the dialogue with citizens about the benefits and trade-offs of new technologies, and allows us to draw on a thriving community and private sector that often leads the change.

- New Zealand brings a unique context – as a society founded on a partnership between indigenous people and more recent arrivals, blending diversity with unique cultural foundations. We are geographically a long way from our neighbours. Technology connects us with both our history and the outside world.

- Even as we're excited by the potential of artificial intelligence, we are committed to ensuring it is developed in an ethical and human-centric framework.

- It's important for New Zealand as we develop our tech industry and find ourselves increasingly woven into global networks with likeminded partners that we can find an expression of those values in our use of technology. That's why we are pleased to participate in this process, and also why we're involved in related initiatives such as the Global Partnership on AI.

- An ethical and human centred approach is important across everything we do.

- I've heard it said that the internet is necessarily just a mirror that reflects society – but rather a prism that refracts it. The machine learning tools that help us organise and comprehend information, can also amplify and distort it in unexpected ways. Social ills that existed before can find a more virulent expression online.

- For New Zealand that became startlingly clear on 15 March 2019, when a terrorist steeped in online conspiracy theories, in white supremacy, and islamophobia, consciously built an atrocity 'for the internet' – committed a racially motivated act of mass murder and broadcast it around the world across all of the major social media platforms. 51 people from our Muslim community were killed, 40 injured, among them children and infants and the elderly.

- This atrocity exposed some sinister aspects of the online environment that none of us had anticipated. The viral content spread rapidly and widely, re-victimising families and communities and inciting other attacks around the world.

- The Christchurch Call represents New Zealand's effort to use the moral obligation we had to effect change globally – mobilising governments and online service providers to work with civil society to take action against terrorism and violent extremism online. It committed us to do so while upholding human rights including freedom of expression and defending a free open and secure internet.

- The Christchurch Call is an emblematic example of the strength of open societies working together.

- We are delighted with the announcement that the US is becoming a part of our shared effort. They will be joining more than 50 other countries, a global network of civil society organisations, and many of the world's largest tech companies.

- The US brings a huge amount of technical knowledge and practical experience to the table. We are very pleased to have you as a partner.

- One of the most pressing issues we will look at together is the role of the AI in radicalisation. What does an online 'user journey' look like for a person who is on the pathway towards violence, and what role to machine learning processes play in helping to guide them there? What can we do collectively to the user interface to help prevent radicalisation to violence and mitigate national security threats?

- This is difficult work, but it goes to the core of what it means to have ethical and responsible AI

- A big priority for us will be to bring more people into the conversation. Marginalised communities, victims of terrorism, people from a range of cultures and backgrounds who can speak to different online experiences and help guide the necessary change.

- Diverse and marginalised voices are essential as part of our conversation about AI, data, and emerging technology issues generally.

- To return to the theme with which I began, technology needs to serve and improve the lives of all our citizens. That is the best way to ensure the adoption and development of emerging technologies occurs in ways that protect our national security.

- This in turn requires open and inclusive approaches to developing solutions.

- This is true for the work of the Christchurch Call.

- It is equally important for a wide range of other critical issues in emergent technology. These include the challenges of tackling a growing tide of misinformation and disinformation, dealing with the growing cyber security threat, and managing the impacts of cybercrime.

- For New Zealand, we welcome this conversation – and the desire to build real, enduring partnerships to tackle these difficult problems together.

~