REPORT OF THE AUSTRALIAN TASKFORCE TO COMBAT TERRORIST AND EXTREME VIOLENT MATERIAL ONLINE

21 June 2019

Disclaimer

The material in this report is of a general nature and should not be regarded as legal advice or relied on for assistance in any particular circumstance or emergency situation. In any important matter, you should seek appropriate independent professional advice in relation to your own circumstances.

The Commonwealth accepts no responsibility or liability for any damage, loss or expense incurred as a result of the reliance on information contained in this report.

Copyright

Enquiries about this paper should be directed to vctaskforcesecretariat@communications.gov.au

Attribution

Use of all or part of this report must include the following attribution:

Using the Commonwealth Coat of Arms

The terms of use for the Coat of Arms are available from the Department of the Prime Minister and Cabinet website (see www.dpmc.gov.au/government/commonwealth-coat-arms).

Executive summary

The terrorist attacks that took place in Christchurch on 15 March 2019 shocked the world. Citizens and governments alike questioned how the mass murder of 51 men, women and children could take place in a peaceful, democratic country like New Zealand. In the hours and days that followed, it became apparent that the internet was exploited to amplify the crimes. The alleged perpetrator live-streamed the murders on Facebook and from there, the video quickly spread, with individuals attempting to upload copies on mainstream and smaller digital platforms and websites.

Condemnation of the acts, and of the use of online platforms in disseminating this content, was swift. On 26 March 2019, the Prime Minister, the Hon Scott Morrison MP, chaired a Summit in Brisbane to discuss Australian Government and industry responses to the sharing of content related to the Christchurch terrorist attack. The Summit brought together representatives from the major digital platforms, Australian Internet Service Providers (ISPs), the heads of relevant Government agencies, along with the Attorney-General, the then Minister for Communications and the Arts and the Minister for Home Affairs.

In this forum, the Government made clear that the community expected more from the digital platforms and that it wanted to see industry bring forward concrete measures to prevent extreme violent content from being disseminated so readily on their services. A key outcome of the Summit was the establishment of the Taskforce to Combat Terrorist and Extreme Violent Material Online (the Taskforce). Comprising government and industry representatives, the objective of the Taskforce was to provide advice to Government on practical, tangible and effective measures and commitments to combat the upload and dissemination of terrorist and extreme violent material.

This report provides that advice. It identifies actions and recommendations that fall into one of five streams: prevention; detection and removal; transparency; deterrence; and capacity building. These actions and recommendations build on and extend the commitments already made by industry and Government following the attacks, including changes by individual firms to the operation of their services. They are also consistent with principles contained within the Christchurch Call to Action.

Ultimately, the Government will assess whether the actions detailed in this report represent a sufficient step forward in terms of ensuring the safety of Australians online. The Government has made clear that it is willing to consider regulatory options where the voluntary commitments put forward by industry fall short of the mark. This reflects a growing consensus internationally that the internet should not be a forum or tool for the proliferation of harmful content, and that more needs to be done — particularly by the larger and well-resourced digital platforms — to make the internet a safer place.

A number of countries, including Australia, have strongly advocated for international cooperation on this issue through various multilateral fora. Consistent with the Christchurch Call to Action, the Australian Government is working internationally and through the actions agreed, including through this taskforce, to drive concrete initiatives on preventing terrorist and violent extremist exploitation of the internet.

There is clearly an appetite, in Australia and overseas, for tangible, concrete measures to tackle the upload and spread of terrorist and extreme violent material online. Whether these fora and the parallel commitments to action from industry can deliver on this outcome will become clearer over the coming months.

Actions and recommendations

Action 1 — Proactive technical intervention

Recommendation 1.1 — Digital platforms to:

a) continue to develop and report to the Australian Government, as per the requirements of Recommendations 7.1 and 7.2, on the ongoing development of technical solutions that seek to prevent terrorist and extreme violent material from being uploaded onto their services, consistent with commitments made in support of the Christchurch Call to Action; and

b) implement these technical solutions in a manner appropriate to the relevant service.

Recommendation 1.2 — Digital platforms to work with other members of the Global Internet Forum to Counter Terrorism (GIFCT) to strengthen:

a) the hash-sharing database with the aim that it:
   i. enables digital platforms to identify all hashed material and block re-upload; and
   ii. facilitates member companies to systematically add newly identified terrorist content.

b) the URL-sharing consortium with the aim that it:
   i. where appropriate, is utilised as broadly as possible by GIFCT members; and
   ii. prevents users from sharing URL links to known or identified terrorist material hosted on external addresses.

Recommendation 1.3 — Digital platforms to work with other members of the GIFCT to explore the capacity to expand the hash-sharing database and URL-sharing consortium to incorporate extreme violent content. The intent would be to align, to the extent possible, with the categories of violent content prohibited by platforms under their respective community standards and terms of service, such as graphic violence, violent content or gore.

Recommendation 1.4 — Digital platforms to review the operation of algorithms and other processes that may drive users towards (or amplify) terrorist and extreme violent material to better understand possible intervention points, and to implement changes where this occurs. This may include using algorithms and other processes to redirect users from such content, or the promotion of credible, positive alternatives or counter-narratives.

Recommendation 1.5 — Digital platforms to have in place clear, efficient appeals mechanisms that provide users with the ability to challenge moderation decisions regarding terrorist and extreme violent material.

Action 2 — Enhanced moderation

Recommendation 2.1 — Digital platforms to:

a) continue to develop and report to the Australian Government, as per the requirements of Recommendations 7.1 and 7.2, on the ongoing development and implementation of technical solutions that seek to identify terrorist and extreme violent material on their respective services; and

b) where feasible and appropriate, use technology to expeditiously remove or disable access to such content without disrupting legitimate use of the services.

Recommendation 2.2 — Digital platforms to:

a) inform users about their reporting options and processes;

b) implement visible and intuitive user reporting mechanisms and minimise friction for users in reporting problematic content; and

c) articulate to users the benefits and outcomes of reporting and the importance of acting where they are exposed to content or behaviour of concern.

Recommendation 2.3 — Digital platforms to:

a) assign the highest level of priority (similar to that for other abhorrent content such as child abuse) to the triaging and moderation of terrorist and extreme violent material; and

b) implement systems and processes to review such material, when flagged or identified, expeditiously and in compliance with relevant Australian laws.

Action 3 — Live-streaming controls

Recommendation 3.1 — Digital platforms that provide live-streaming services to identify, fast-track and report to the Australian Government, as per the requirements of Recommendations 7.1 and 7.2, on the implementation of appropriate checks on live-streaming aimed at reducing the risk of users disseminating terrorist and extreme violent material online. Depending on the service in question, these checks may include:

a) strengthening account validation processes, particularly for account creation;

b) limiting the ability of new users to live-stream until they have established a pattern of behaviour that adheres to community standards or terms of service, for example:
   i. ‘cooling off periods’ before a new user can live-stream (such as 24 hours);
   ii. limiting audience size or the capacity to monetise live-streamed content for new users;
   iii. implementing streamer ratings or scores; and / or
   iv. monitoring account activity.

Recommendation 3.2 — Digital platforms that provide live-streaming services to implement policies that suspend the ability of users to live-stream where they have serially and / or materially breached community standards or terms of service.

Recommendation 3.3 — Digital platforms that provide live-streaming services to ensure they have in place priority or accelerated review of flags of terrorist or extreme violent content on these services, noting that Google (YouTube) currently has such processes in place.

Action 4 — Industry-Government collaboration

Recommendation 4.1 — Overseen and managed by the Australia-New Zealand Counter-Terrorism Committee, digital platforms and relevant Australian Government agencies to convene a 'testing event' in 2019-20 simulating a scenario which will allow all parties to gauge whether industry tools, and Government processes, are working as intended, particularly as they mature in response to technology and increased investment in content moderation. The details of this 'testing event' would be developed collaboratively between relevant agencies and platforms. Pending the success of the initial event, this process could be repeated on a regular basis to measure the ongoing effectiveness of systems and system improvements.

Recommendation 4.2 — Digital platforms to fund (including via the GIFCT, as appropriate), with the support of the Australian Government, research and academic efforts to better understand, prevent and counter terrorist and extreme violent material online, including both the offline and online impacts of this activity, and use this knowledge to develop and promote positive alternatives and counter-messaging online.

Recommendation 4.3 — Relevant Australian Government agencies, academia, researchers, and civil society bodies that monitor and review terrorist and extremist organisations to share with digital platforms (where legally and operationally feasible) indicators of terrorism, terrorist products and depictions of violent crimes.

Action 5 — Content blocking

Immediate

Recommendation 5.1 — The eSafety Commissioner to consider utilising subsection 581(2A) of the *Telecommunications Act 1997* to direct the ISPs currently blocking domains hosting the footage of the Christchurch attacks and the alleged perpetrator's manifesto to maintain these blocks while the feasibility of longer-term arrangements is assessed, as per recommendation 5.3. The Government to host a landing page for blocked domains in the event that one or more notices are issued by the eSafety Commissioner.

Recommendation 5.2 — The eSafety Commissioner, in consultation with Communications Alliance, to develop a protocol to govern the interim use of subsection 581(2A) of the *Telecommunications Act 1997* in the circumstances of an online crisis event. This protocol would set out the arrangements and process for implementing blocks of websites hosting offending content, including:

a) the means of determining which ISPs would be subject to blocking orders (and the reporting obligations specified under recommendations 7.1 and 7.2), the length of time that the ISPs will be required to implement the blocks, and the process for removing the blocks;

b) the process to be used to determine whether the terrorist or extreme violent material is sufficiently serious to warrant blocking action, and to identify the domains that are hosting the material;

c) guidance on the circumstances in which it is anticipated that this power may be used by the eSafety Commissioner;

d) the landing page for the blocks and the method of communicating the notice; and

e) to the extent possible, ensure automated notification processes are used to their fullest extent and are as efficient as possible.

Statutory reform

Recommendation 5.3 — The Australian Government to pursue legislative amendments to establish a content blocking framework for terrorist and extreme violent material online in crisis events. This new framework should:

a) incorporate the matters stipulated in the interim protocol (Recommendation 5.2); and

b) address additional factors including indemnity, notification processes for content hosts and the automation of blocks.

Action 6 — Emergency response network

Recommendation 6.1 — Consistent with the Christchurch Call, and the GIFCT members' response, the Government will amend the Australian Government Crisis Management Framework to include a new online crisis response protocol for terrorist and extreme violent material online.

Recommendation 6.2 — Australian Government agencies, ISPs and digital platforms to work collaboratively to develop the new online crisis response protocol that:

a) utilises the 24/7 capability of Crisis Coordination Centre to notify relevant government agencies of online crisis events involving terrorist and extreme violent material;

b) provides for the eSafety Commissioner to undertake the initial assessment of any content flagged in response to an online crisis event;

c) maximises automation in the flow of information in all directions (to and from industry members and Government agencies); and

d) incorporates dedicated contact and action points that integrate, as far as possible, existing industry and agency arrangements.

Recommendation 6.3 — The ACMA to consider the participation of media companies within the online crisis response protocol.

Action 7 — Periodic reporting

Progress reports on actions

Recommendation 7.1 — In the absence of an international standard-setting and monitoring process, within three months of the finalisation of this report, digital platforms, ISPs and Australian Government agencies responsible for implementing the recommendations outlined in the Taskforce report to provide a detailed implementation plan to the Secretary of the Department of Communications and the Arts, outlining the response to recommendations and the timeframes for the implementation of these measures.

Recommendation 7.2 — Digital platforms, ISPs and Australian Government agencies responsible for implementing the recommendations outlined in the Taskforce report to submit annual progress reports (aligned with financial year) to the Secretary of the Department of Communications and the Arts to support Government consideration of progress in implementing the recommendations and actions. The fulfilment of this reporting obligation could be achieved through existing statutory reporting requirements.

Parliamentary oversight

Recommendation 7.3 — The Australian Government to consider the extent to which the Parliamentary Committee structure could be expanded to regularly assess and report on online harms relevant to the Australian community, including those associated with terrorist and extreme violent material online.

Transparency reports

Recommendation 7.4 — Digital platforms to publish reports (at least half yearly) outlining their efforts to detect and remove terrorist and extreme violent material on their services. These reports are intended to demonstrate the nature and extent of actions being taken by platforms which could include:

a) the number of items flagged by users for potential violations of policies against the promotion of terrorism or extreme violent content;

b) the total number of items removed by the digital platform including;

i. how the items were detected (the platforms' systems, user flags, trusted flaggers or other external experts); and

ii. the number of items detected by platform systems that were removed before users engaged with the content;

c) the number and entity type (e.g. video, channel) of items of terrorist content and extreme violent content removed by the platform;

d) examples of content flagged for promotion of terrorism or extreme violence that did and did not violate the platform's guidelines;

e) the number of items of terrorist content and extreme violent content that were flagged or identified by the platforms' systems, including:

i. the source of identification (AI or hashing database);

ii. whether the content was on-service, or flagged or identified at the point of attempted upload; and

iii. the action taken (automatically blocked or referred for moderation);

f) the total number of items of terrorist content and extreme violent content that were subject to moderation, broken down by those that were flagged by users, systems, other sources, and the total volume of content removed; and

g) the average time taken to review and action flagged items of terrorist content and extreme violent content, or the number of times flagged terrorist content or extreme violent content was viewed by users before action was taken.

Action 8 — Account management

Recommendation 8.1 — Digital platforms to ensure their account management practices and policies can be enforced against users who upload and share terrorist and extreme violent content, recognising that there will be legitimate public interest reasons for the dissemination of this content in some cases, such as to condemn the acts depicted.

Recommendation 8.2 — Digital platforms to investigate and continue to develop systems that seek to prevent the automated creation of accounts and the circumnavigation of account suspensions, recognising that persistent bad actors may find a way to circumvent these systems.

Recommendation 8.3 — Digital platforms to ensure that appeals processes are accessible to users who consider that they have been wrongly suspended or subject to other penalties incorrectly.

Action 9 — Capacity building

Recommendation 9.1 — The GIFCT members of the Taskforce and the Australian Government to support the examination of reforms to the governance and structural arrangements of the GIFCT, through existing collaborative fora or other means, that could include:

a) establishing the GIFCT as a stand-alone, industry funded, independent body, with dedicated resources and full-time staff;

b) expanding the GIFCT membership, with a particular focus on smaller companies; and

c) supporting the work of the GIFCT with annual programs of planned activity and progress reporting.

Recommendation 9.2 — The GIFCT members of the Taskforce to advocate for the GIFCT to:

a) establish a central repository of technical tools to enable them to more effectively prevent, detect and respond to online terrorist and extreme violent and actively make these solutions accessible to smaller online services;

b) support the deployment of prevention, detection and response mechanisms, particularly for smaller online services that may lack the expertise to deploy technical solutions; and

c) deepen partnerships with training and knowledge-sharing initiatives like Tech Against Terror and support organisations to roll-out additional seminars and workshops globally.

Recommendation 9.3 — ISPs and their industry body, Communications Alliance, to:

a) determine the applicability of the Taskforce recommendations to Australian ISPs beyond those participating in the Taskforce; and

b) work with smaller ISPs to develop technical and logistical capacity for content blocking and emergency response networks.

Definitions and terms

Terrorist and extreme violent material

Violence is — to varying degrees — a pervasive feature of society and its impacts and harms (beyond those imposed on the victims) will depend on context, intent and type. The Taskforce has developed the following definition of "terrorist and extreme violent material" for its work:

Terrorist and extreme violent material is audio, visual or audio-visual material that:

- depicts an actual terrorist act targeting civilians (as opposed to animated or fictionalized);
- depicts actual (as opposed to animated or fictionalized) violent crime; or
- promotes, advocates, encourages or instructs a terrorist, terrorist group or terrorist act, or a person to commit actual (as opposed to animated or fictionalized) violent crime.

Terrorist Act

As per section 100.1 of the *Criminal Code Act 1995*.

Violent crime

Murder; attempted murder; torture; rape; and violent kidnapping (as per the definition of abhorrent violent conduct in the AVM Act, excluding terrorist acts, as these are addressed directly).

Violent crime may also include the categories of violent content and material prohibited by the digital platforms as part of their respective community standards and terms of service, such as graphic violence, violent content, or gore.

Terrorist group and terrorist

An organisation listed by the Government as a terrorist organisation under the *Criminal Code Act 1995*. A 'terrorist' is a member of such a group.

Exclusions

Discussions with Taskforce members have highlighted the importance of establishing appropriate exceptions to the definition of terrorist and extreme violent material.
These exclusions include content that is:

- taken and distributed by innocent bystanders (not in any way associated with the alleged perpetrator);
- produced by news organisations or journalists and distributed for reasonable and legitimate journalistic purposes;

- intended to raise awareness about human rights abuses, discrimination or acts of terrorism;
- a reasonable expression of protest, political dissent or contemporary social commentary; and
- produced, uploaded or shared for legitimate academic, artistic, law enforcement, Government, research, satirical, educational or documentary purposes.

Criminal Code Amendment (Sharing of Abhorrent Violent Material) Act 2019

The recently passed *Criminal Code Amendment (Sharing of Abhorrent Violent Material) Act 2019* (the AVM Act) is targeted at removing the most repugnant of terrorist and extreme violent material that can be accessed online in Australia. For these reasons, the AVM Act has a more limited definition and refers to 'abhorrent violent material'. The Act creates new offences in the *Criminal Code Act 1995* (Cth) that are aimed at reducing the use of online platforms to share abhorrent violent material.

'Abhorrent violent material', is defined as audio, visual or audio-visual material recorded of extreme violent acts. Exhaustively, these violent acts are:

- a terrorist act (involving serious physical harm or death, and otherwise within the meaning of 100.1 of the Criminal Code);
- the murder of another person;
- the attempted murder of another person;
- the torture of another person;
- the rape of another person; or
- kidnapping involving violence.

Importantly, 'abhorrent violent material' is restricted to material recorded or streamed by the perpetrator or their accomplice. The Taskforce considered that it was important to distinguish between this definition, which relates to specific violent conduct, and the broader definition of 'terrorist and extreme violent material' used to guide and inform the work of the Taskforce.

Context

While the definition and exclusions outlined above seek to provide clarity as to what is and isn't terrorist and extreme violent material, the Taskforce members have discussed the fact that context is critical in assessing the measures that seek to combat the dissemination of such content online. There are circumstances where the upload and sharing of otherwise harmful content online will be legitimate and motivated by a desire to address social issues and raise awareness of illegal activity or unconscionable conduct, rather than to cause harm or fear in the community.

A well-known example concerns the use of Facebook Live to broadcast the aftermath of the shooting of Philando Castile.[1] While graphic, the content enhanced transparency of the incident and informed public discussion about the use of lethal force by police. In another example, videos and images of the victims of Mexico's ongoing drug war on social media have raised awareness of cartel violence in an environment where journalists are at risk.[2] However, similar content uploaded by cartel members to extort families, or warn off rivals, is clearly unacceptable.[3]

Assessing these contextual factors is difficult, particularly as events are occurring. Nonetheless, these are judgements that digital platforms and other providers of online content are frequently required to make to ensure that their users are not exposed to graphic or confronting content. In instances where there are public interest justifications for the content, it will be important to ensure that minors are provided with appropriate protections, and that content is covered with interstitial material that alerts users to graphic content and requires them to acknowledge this to gain access. Users also have a responsibility to mark graphic or confronting content appropriately at upload.

Additional definitions and terms are included in the following table.

Acronym/Term	Definition/Outline
ACMA	Australian Communications and Media Authority
AI	Artificial Intelligence
Algorithms	The automated interpretation or calculation of data that is used to serve particular content to a user
Appeal mechanisms	Avenues for users to lodge a request for a platform to review a decision made regarding material that has been flagged or reported.
AVM Act	Criminal Code Amendment (Sharing of Abhorrent Violent Material) Act 2019
Digital platform	For the purposes of this report, digital platform means a provider of social media or similar public-facing, content-sharing services.
GIFCT	Global Internet Forum to Counter Terrorism
Hashing	The process of creating a 'digital fingerprint' of content, which can then be used to identify and block content with an identical fingerprint from being uploaded onto a platform.

1 Brad Parks and Holly Yan (2017) From FB Live to witness stand: Philando Castile's girlfriend testifies, Available at: https://edition.cnn.com/2017/06/06/us/philando-castile-officer-trial-testimony/index.html (Accessed: 12 June 2019).

2 Reporters Without Borders (2018) Worldwide Round-Up. CNN [Online]. Available at: http://cdn.cnn.com/cnn/2018/images/12/19/worldwilde_round-up.pdf (Accessed: 12 June 2019).

3 Rebecca Plevin (2019) 'We're Going to Find You.' Mexican Cartels Turn Social Media Into Tools for Extortion, Threats, and Violence, Available at: https://pulitzercenter.org/reporting/were-going-find-you-mexican-cartels-turn-social-media-tools-extortion-threats-and-violence (Accessed: 12 June 2019).

Acronym/Term	Definition/Outline
ISPs	Internet Service Providers
Industry	Digital platforms and ISPs
Live-streaming	The sharing of live audio-visual material via the internet
Online crisis event	An event that involves terrorist or extreme violent material being disseminated online in a manner likely to cause significant harm to the Australian community, and that warrants a rapid, coordinated and decisive response by industry and relevant government agencies.
Tech Against Terrorism	A public-private partnership launched by the United Nations to support smaller tech platforms, share best practice and tools for the identification and removal of terrorist propaganda.
Trusted flaggers	Individuals or organisations who have been deemed a reliable 'flagger' of offending material on a platform. These flaggers are often granted special flagging privileges/abilities, and/or have prioritised flag reviews (varies depending on the platform).
User reporting	Tools that allow platform users to flag content that potentially violates the guidelines and policies of that platform.