Response to the Parliamentary Joint Committee on Law Enforcement

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

OCTOBER 2021

FACEBOOK

Executive summary

Facebook welcomes the opportunity to participate in the Parliamentary Joint Committee on Law Enforcement's review of the *Criminal Code Amendment (Sharing of Abhorrent Violent Material Act 2019)* (the AVM law).

Combatting terrorism and extreme violence is a continuous responsibility, and governments, experts, and digital platforms all have roles to play. We take responsibility for detecting and removing this content from Facebook's services, and this submission outlines the significant work and investment we have undertaken.

Facebook now has more than 40,000 people working on safety and security at Facebook, and we've invested more than US$13 billion (~AU$17.5 billion) on safety and security since 2016. We expect that we will spend more than US$5 billion (~AU$6.6 billion) on safety and security in 2021.

As well as our longstanding work to combat terrorist and extreme violent content, we have also taken the lessons learned from the 2019 Christchurch attacks. We have continued to make meaningful progress to ensure that online services cannot be abused in this way again, including as part of our commitments as a signatory to the Christchurch Call. The improvements we have made over the two years span our policies, enforcement, partnerships and research. Some highlights include:

1. **Policies**. Under our Community Standards, we have developed a number of policies that prohibit terrorist and extreme violent material on our services. We have developed or enhanced a number of policies since 2019, including:
    - prohibiting white nationalism (alongside our longstanding prohibition on white supremacy)
    - developing new policies to prohibit militarised social movements and violence-inducing conspiracy theories
    - strengthening our hate speech policies, including: (1) introducing a new 'hateful stereotypes' policy; (2) prohibiting claims that deny or distort the Holocaust; (3) disallowing ads that claim a group with "protected characteristics" is a threat to the safety, health or survival of others; (4) expanding our ads policies to better protect immigrants, migrants, refugees and asylum seekers from hateful claims; (5) removing attacks on "concepts" linked to protected characteristics (such as religions) if the attacks are likely to contribute to imminent physical harm, intimidation or discrimination against the people associated with that protected characteristic.

2. **Enforcement.** We are continually taking steps to improve our ability to proactively detect hate and extremism on our services. We have banned more than 250 white supremacist organisations globally (a number which includes groups in Australia) and we have removed nearly 900 militarised social movements from our platform.[1]

   We have made a number of product changes to livestreaming since 2019, including as part of our commitments as a signatory to the Christchurch Call. Some of these product changes include restricting more users from products like Facebook Live and increasing our capability to respond rapidly to livestreams.

   We recognise that we can always improve our enforcement, so we make data available to allow for scrutiny and accountability of the enforcement of our policies. We are increasingly identifying and removing violating content via artificial intelligence, so we don't need to rely on users seeing and reporting the content. In our last Community Standards Enforcement Report, we indicated that
   - 99.7 per cent of the terrorist content we took action against was detected proactively
   - 97.8 per cent of the organised hate content we took action against was detected proactively
   - 97.6 per cent of hate speech we took action against was detected proactively.[2]

3. **Partnerships**. While we have made significant progress as a company in combatting terrorist and extreme violent content, our work is supported by a multi-faceted and collaborative effort between a range of stakeholders, including companies, civil society organisations, experts, and governments. For this reason, we have prioritised partnerships with these groups.

   Some of our most important partnerships include:
   - the cross-industry group the Global Internet Forum to Counter Terrorism (GIFCT), of which we are a founding member. Since 2019, the GIFCT has now transitioned to become an independent organisation, with the appointment of an Executive Director, Nicholas Rasmussen. The GIFCT's database of shared digital "hashes" (fingerprints) and

[1] Facebook, 'An update to how we address movements and organizations tied to violence', *Facebook Newsroom*, blog post updated 19 January 2021, https://about.fb.com/news/2020/08/addressing-movements-and-organizations-tied-to-violence/.

[2] Facebook, *Community Standards Enforcement Q2 2021*, https://transparency.fb.com/data/community-standards-enforcement/dangerous-organizations/facebook/

agreed protocols for responding to a live terrorist incident improve our ability to enforce our policies. The GIFCT Hash Sharing Database now contains more than 320,000 hashes.[3] The infrastructure established around the GIFCT - including a content incident protocol - improve the capacity of all GIFCT members to take a coordinated industry-wide response to a crisis.

- working with civil society groups to understand developments on the ground and to deploy programs to counter violent extremism. Initiatives like our Search Redirect Program or support for counterspeech initiatives help to combat radicalisation and push back against hate. We have also established an Australia-specific Combatting Online Hate Advisory Group, to ensure Australia civil society groups and experts have a direct channel to give us advice or feedback about how to better combat online hate, before it manifests into terrorist or extreme violent activity.
- a significant amount of work in collaboration with governments and law enforcement and we contact law enforcement when we encounter credible threats of harm.

4. **Research**. We fund a significant amount of research to contribute to our own understanding of hate and extremism online, and to provide insights that contribute to the broader community of practice. We fund research on extremism via GNET, the research arm of the GIFCT. Between June 2020 and July 2021, GNET published 198 insights from 245 authors based in twenty-four countries around the world.

   We also fund research unilaterally. In particular, we have commissioned two pieces of research specific to Australia that have been publicly released in 2021: (1) hate speech experienced by Aboriginal and Torres Strait Islander people online; and (2) how LGBTQI+ Australians use our services, including how they combat online hate. We are continuing to fund more research which will be released in 2022.

To provide governments and the Australian public with confidence around our considerable investment and work to combat terrorist and extreme violent content, we support regulation that holds digital platforms accountable by creating incentives for companies to responsibly balance values like safety, privacy, and freedom of expression, and fosters trust through meaningful transparency. Facebook has been

[3] GIFCT, *GIFCT Transparency Report July 2021*, https://gifct.org/wp-content/uploads/2021/07/GIFCT-TransparencyReport2021.pdf.

calling for new rules for the internet - including content regulation - around the world for many years.[4]

We wholeheartedly support the objectives of the AVM law: governments should set in place regulatory frameworks to ensure swift action is taken in relation to terrorist and extreme violent content online. Indeed, since the passage of the AVM law, we have seen laws and codes to combat online terrorist content developed in jurisdictions like New Zealand, the United Kingdom and the European Union[5], and we have supported these efforts. Our aim to be a constructive partner is in line with our global commitment to working with governments to develop content regulation that is proportionate, workable, and creates appropriate incentives for digital platforms.

A review of the AVM law is timely. As you would expect for any legislation drafted very quickly and passed through Parliament within five days, there are some opportunities for improvement. Our submission makes constructive suggestions to ensure the law is working effectively and as intended. Specifically, we have made recommendations to clarify:

- the definition of AVM content
- whether the law represents a proactive monitoring obligation
- how defences may operate and apply.

We welcome the opportunity to work with Australian policymakers on combatting terrorist and violent content online, including the members of the Parliamentary Joint Committee on Law Enforcement, and would be pleased to discuss any of these suggestions.

[4] M Bickert, *Charting a way forward: online content regulation*, white paper released February 2020, https://about.fb.com/wp-content/uploads/2020/02/Charting-A-Way-Forward_Online-Content-Regulation-White-Paper-1.pdf.

[5] These include the Online Safety Code of Practice (NZ); the interim code of practice on terrorist content and activity online (UK); and the regulation addressing the dissemination of terrorist content online (EU).

Table of contents

Comments on the AVM law

To provide governments and the Australian public with confidence around our considerable investment and work to combat terrorist and extreme violent content, we support regulation that holds digital platforms accountable by creating incentives for companies to responsibly balance values like safety, privacy, and freedom of expression, and fosters trust through meaningful transparency.

We wholeheartedly support the objectives of the AVM law and our aim to be a constructive partner is in line with our global commitment to working with governments to develop content regulation that is proportionate, workable, and creates appropriate incentives for digital platforms.[6]

We have worked hard to enhance our relationships with the eSafety Commissioner's Office, the Australian Federal Police and others (building on existing relationships), to ensure the legislation operates effectively. We've worked with eSafety on a small number of incidents: our efforts to work with the Australian Government have been much greater than just in relation to receiving notices. We have established a working relationship of informally briefing eSafety (at a minimum) whenever we see possible extreme violent, terrorist or crisis content on our services that may be of interest to them. We have also proactively notified the AFP of a number of instances where we have seen content on our services that could potentially constitute AVM.

A review of the AVM law is timely. As you would expect for any legislation drafted very quickly and passed through Parliament within five days, there are some opportunities for improvement - and our submission makes constructive suggestions about how the drafting could be amended to ensure the law is working effectively and as intended by the Australian Government.

There are some areas of drafting and questions that may benefit from further consideration in how they work in practice. For example, the penalties in the law (s474.34) are very severe, and include a criminal component (including potential imprisonment) directed towards the individuals responsible for content moderation. These penalties are less likely to be effective incentives for companies to build best-practice systems-based approaches to content moderation.

The remainder of this submission is in two parts. Firstly, we have outlined the work that we have undertaken in order to combat terrorist and extreme violent content on our services, particularly since 2019. This is intended to assist the Committee and

[6] M Bickert, *Charting a way forward: online content regulation*, white paper released February 2020, https://about.fb.com/wp-content/uploads/2020/02/Charting-A-Way-Forward_Online-Content-Regulation-White-Paper-1.pdf.

Australian policymakers in their work to identify appropriate regulatory responses to combat online terrorist and extreme violent content. The second part provides some recommendations for amendments to the AVM law. It draws from our lived experience of putting the law into practice, as well as making a number of recommendations, to ensure the legislation is operating as intended by the Australian Government.

Facebook's work in combatting terrorist and extreme violent content

Facebook has made significant commitments and investments to combat terrorist and extremist content on our platform. In particular, we now have more than 40,000 people working on safety and security within Facebook, and we've invested more than US$13 billion (~AU$17.5 billion) in this area since 2016.

In this section, for the benefit of the Parliamentary Joint Committee on Law Enforcement, we explain the approach that Facebook takes to combatting terrorism and graphic violent content. Our strategy comprises:

1. Policies
2. Enforcement
3. Partnerships
4. Research.

Policies

The policies that outline what is and is not allowed on Facebook are called our Community Standards.[7] Our policies are based on feedback from our community and the advice of experts in fields such as technology, public safety and human rights. Our Community Standards are also not static: we amend them regularly in response to feedback or developments.

A number of parts of our Community Standards are material to this inquiry, including our policies on dangerous organisations, militarised social movements and violence-inducing conspiracy theories, hate speech, and violence and incitement.

Dangerous individuals and organisations

[7] Facebook, *Community Standards*, https://www.facebook.com/communitystandards/.

Facebook's Community Standards prohibit any organisation or individual that proclaims a violent mission or are engaged in violence from having a presence on Facebook. Specifically, we do not allow on our platform:

- terrorist organisations and terrorists
- hate organisations, and their leaders and prominent members
- mass / multiple murderers (including attempted murderers).

As well as removing these groups, we do not allow content that praises, supports or represents them.

Defining "terrorism" is a significant challenge. There is much debate among experts and policymakers about a definition of terrorism. It is a highly contested term, and most governments or inter-governmental fora do not have an agreed definition of terrorism.

However, as part of our industry-leading work to combat terrorist content, Facebook has developed our definition of terrorism (which we use in assessing content on our platform).[8] We define a terrorist organisation as:

> "Any non-governmental organisation that engages in premeditated acts of violence against persons or property to intimidate a civilian population, government, or international organisation in order to achieve a political, religious, or ideological aim."

Our definition is agnostic to the ideology or political goals of a group, which means it includes everything from religious extremists and violent separatists to white supremacists and militant environmental groups. It's about whether they use violence to pursue those goals. This is a definition that is applied across the more than 3 billion people who use Facebook around the world.

Militarised social movements and violence-inducing conspiracy theories

In August 2020, we expanded our dangerous organisations policy to capture "militarised social movements" and content relating to "violence-inducing conspiracy theories".

Implementation of these policies began with Pages, Groups, Events, and Instagram accounts dedicated to militarised social movements and violence-inducing conspiracy theories. Some examples of content that may be captured under this policy includes

[8] Facebook, 'Combating hate and extremism', *Facebook Newsroom*, 17 September 2019, https://about.fb.com/news/2019/09/combating-hate-and-extremism/.

content relating to militarised social movements like the Oathkeepers and a violence-inducing conspiracy theory like QAnon.[9]

Hate speech

We have provided information below about our policies relating to hate speech, given it can be a precursor for terrorist or extreme violent activity.

We don't allow hate speech on Facebook. It creates an environment of intimidation and exclusion, may promote offline violence, and can inhibit people from using their voice and feeling safe to connect freely.

We define hate speech as a direct attack against people on the basis of what we call protected characteristics. We have currently listed the following as protected characteristics:

- race
- ethnicity
- national origin
- disability
- religious affiliation
- caste
- sexual orientation
- sex
- gender identity
- serious disease.

We define attacks as violent or dehumanising speech, harmful stereotypes, statements of inferiority, expressions of contempt, disgust or dismissal, cursing, and calls for exclusion or segregation.

To ensure our policies are relevant and effective, we regularly review them, in consultation with experts and academics, including from Australia. We have made a number of changes over the last 12 months to expand our hate speech policies in our Community Standards. These include:

- the development of a new hateful stereotypes policy, which will, in the first instance, prohibit content depicting blackface and stereotypes that Jewish

[9] Facebook, 'An update to how we address movements and organizations tied to violence', *Facebook Newsroom*, blog post updated 19 January 2021, https://about.fb.com/news/2020/08/addressing-movements-and-organizations-tied-to-violence/.

people run the world.[10] We continue to consult on possible expansions to this policy to capture other hateful stereotypes.

- expansions in our ads policies to better protect immigrants, migrants, refugees and asylum seekers from hateful claims[11]
- expansions in our ads policies to prohibit claims that a group is a threat to the safety, health or survival of others on the basis of that group's race, ethnicity, national origin, religious affiliation, sexual orientation, gender, gender identity, serious disease or disability.[12]
- removing any claims that deny or distort the Holocaust, on the basis of expert consultation and research.[13]

Violence and incitement

We aim to prevent potential offline harm that may be related to content on Facebook. While we understand that people commonly express disdain or disagreement by threatening or calling for violence in non-serious ways, we remove language that incites or facilitates serious violence. We remove content, disable accounts, and work with law enforcement when we believe there is a genuine risk of physical harm or direct threats to public safety. We also try to consider the language and context in order to distinguish casual statements from content that constitutes a credible threat to public or personal safety.

This policy means we are able to take action against content that is calling for violence or incitement, even if the author has not yet been designated by us as a dangerous organisation or individual.[14]

Enforcement

Enforcing our policies against terrorist and extreme violent organisations is a constant challenge: just as terrorist groups have been resilient to counterterrorism efforts in the offline world, we are in an adversarial situation in detecting and removing these groups. We need to continuously improve in order to help keep our community on Facebook safe.

---

[10] G Rosen, 'Community Standards Enforcement Report August 2020, *Facebook Newsroom*, 11 August 2020, https://about.fb.com/news/2020/08/community-standards-enforcement-report-aug-2020/.
[11] Facebook, 'Meeting the unique challenges of the 2020 elections', *Facebook Newsroom*, 26 June 2020, https://about.fb.com/news/2020/06/meeting-unique-elections-challenges/
[12] Ibid.
[13] M Bickert, 'Removing Holocaust denial content', *Facebook Newsroom*, 12 October 2020, https://about.fb.com/news/2020/10/removing-holocaust-denial-content//
[14] As an example, see our work in relation to boogaloo content last year: Facebook, 'Banning a violent network in the US', *Facebook Newsroom*, 30 June 2020, https://about.fb.com/news/2020/06/banning-a-violent-network-in-the-us/.

Although our enforcement will not always be perfect, we have made significant progress in detecting and removing terrorist and extremist groups on our services. We have banned more than 250 white supremacist organisations globally and we have removed nearly 900 militarised social movements from our platform. Some of the individuals and organisations designated in Australia include Blair Cottrell, Neil Erickson, Tom Sewell, the Lads Society, the United Patriots Front, True Blue Crew and the Antipodean Resistance.

We detect dangerous organisations and terrorist content via a playbook and a series of automated techniques, which were first developed three years ago to detect content related to terrorist organisations such as ISIS, al Qaeda and their affiliates. We've since expanded these techniques substantially:

- We're now able to detect text embedded in images and videos in order to understand its full context.
- We've built media matching technology to find content that's identical or near-identical to photos, videos, text and even audio that we've already removed.
- We've now expanded to detect more groups tied to different hate-based and violent extremist ideologies and using different languages.
- We have learned from the techniques we currently use in the cyber security space to develop a new tactic that targets a banned group's presence across our apps. We do this by identifying signals that indicate a banned organisation has a presence, and then proactively investigating associated accounts, Pages and Groups before removing them all at once. Once we remove their presence, we work to identify attempts by the group to come back on our platform.
- We're also studying how dangerous organisations initially bypassed our detection, as well as how they attempt to return to Facebook after we remove their accounts, in order to strengthen our enforcement and create new barriers to keep them off our apps.
- We've increased our capability to rapidly respond to livestreams that may breach our Community Standards, including by reviewing all livestreams in an area that may involve footage of an attack and increasing our 24/7 capacity to respond to livestream reports.

In addition to the changes outlined above, there are additional significant steps we have taken in relation to Facebook Live. We have:

- improved our response times to user reports of Live and recently Live videos.
- announced restrictions on who can use Facebook Live in honour of the Christchurch Call. We now apply a 'one strike' policy to Live in connection with a broader range of offenses. Anyone who violates our most serious policies are restricted from using Live for set periods of time – for example 30 days –

> starting on their first offense. For instance, someone who shares a link to a statement from a terrorist group with no context is now immediately blocked from using Live for a set period of time. These restrictions are a meaningful change that would have added friction to the Christchurch attacker.

We are always looking at ways to improve our detection and enforcement, using advancements in technology and partnerships. We have been working to collect camera footage from law enforcement partners in the US and UK from their firearms training programs - providing a valuable source of data to train our systems. This helps improve our detection of real-world, first-person shooter footage of violent events and avoid incorrectly detecting other types of footage. We have been collecting and ingesting that data from existing partners and hope to expand this collaboration to law enforcement agencies in other countries soon.[15]

In addition to building new tools, we've also employed new strategies, such as leveraging off-platform signals to identify dangerous content on Facebook, and implementing procedures to audit the accuracy of our artificial intelligence's decisions over time.

Measuring the effectiveness of enforcement

We make data publicly available regularly to assist in assessing and measuring the effectiveness of our enforcement approaches.

Our progress can be primarily measured through our transparent quarterly Community Standards Enforcement Report. We have long reported on the amount of terrorist content we have removed from our services, but for some time the reporting only covered content relating to Al Qaeda, ISIS and their affiliates. In 2019, we expanded our reporting to *all* terrorist organisations; and, in 2020, we updated these metrics to report on content that propagates organised hate (such as white supremacy) separate to terrorism content.

According to the last Community Standards report (August 2021)[16], in the period April to June 2021, on Facebook, we took action against:

- 7.1 million pieces of content for terrorism
- 6.2 million pieces of content for organised hate
- 31.5 million pieces of content for hate speech.

---

[15] Facebook, 'Combating hate and extremism', *Facebook Newsroom*, 17 September 2019, https://about.fb.com/news/2019/09/combating-hate-and-extremism/.

[16] Facebook, *Community Standards Enforcement Q2 2021*, https://transparency.fb.com/data/community-standards-enforcement/dangerous-organizations/facebook/

For each category, we also reported on the percentage of content that was detected proactively by us using artificial intelligence (compared to the percentage brought to our attention from a user report). Our ambition is to increasingly detect and remove content proactively, before users even see it, and so we have been investing significantly in artificial intelligence that helps us proactively detect this content. In the last reporting period:

- 99.7 per cent of the terrorist content we took action against was detected proactively
- 97.8 per cent of the organised hate content we took action against was detected proactively
- 97.6 per cent of hate speech we took action against was detected proactively.

We have also developed a metric called *prevalence*, where we estimate how prevalent violating content is on Facebook. We think of this metric as how many views of violating content our enforcement approach did not identify - either because people saw the content before we could take action, or because we missed the violation altogether.[17] We hold ourselves accountable to these numbers. In the last report:

- 0.05 per cent of views of content on Facebook contained hate speech. This means, for every 10,000 views of content on Facebook, 5 contained hate speech. This metric has been *halved* over the last 12 months.
- For terrorist or organised hate content, there are insufficient views to precisely estimate prevalence for these types of content. Because it is so infrequent, we estimate the upper limit for prevalence. For these types of content, it is 0.07 per cent of content views. This means that out of every 10,000 views of content on Facebook, we estimate no more than 7 of those views contained content that violated the policy.

Our enforcement approach has been scrutinised externally. For example, in 2020, a third-party independent test run by the European Commission annually found that Facebook assessed 95.7% and Instagram assessed 91.8% of hate speech notifications in less than 24 hours, compared to 81.5% for YouTube and 76.6% for Twitter.[18] The European Commission also stated that "only Facebook informs users systematically; all the other platforms have to make improvements."

We are undertaking an independent, third-party audit - starting this year - to validate the numbers we publish in our Community Standards Enforcement Report.[19]

---

[17] A Kantor, 'Measuring our progress combating hate speech', *Facebook Newsroom*, 19 November 2020, https://about.fb.com/news/2020/11/measuring-progress-combating-hate-speech/.

[18] G Rosen, 'New EU report finds progress fighting hate speech', *Facebook Newsroom*, 23 June 2020, https://about.fb.com/news/2020/06/progress-fighting-hate-speech/.

[19] V Sarang, 'Independent audit of Community Standards Enforcement Report metrics', *Facebook Newsroom*, 11 August 2020, https://about.fb.com/news/2020/08/independent-audit-of-enforcement-report-metrics/.

Facebook is the only technology company undertaking an independent, third-party audit for its transparency metrics.

This audit is a follow up to the work by the Data Transparency Advisory Group (DTAG), which was set up in 2018, to provide an independent, public assessment of whether the metrics we share in the Community Standards Enforcement Report provide accurate and meaningful measures of Facebook's content moderation challenges and our work to address them. DTAG is an independent body made up of international experts in measurement, statistics, criminology and governance. Initial findings by the group. In their first report, the DTAG noted that they found our approach and methodology sound and reasonable, but highlighted areas where we could be more open in order to build more accountability and responsiveness to the people who use our platform.[20] These important insights help inform our work.

Partnerships

While we have made significant progress as a company in combatting online hate and violence, we also enter into partnerships with other companies, civil society organisations, experts, and governments. Some of these partnerships are outlined below.

Cross-industry partnerships

Cross-industry partnerships are vital in countering online terrorism and extremism, because these groups generally work across multiple digital platforms and services to achieve their aims.

Facebook is one of four founding members of a cross-industry partnership called the Global Internet Forum to Counter Terrorism (GIFCT). It is a partnership that allows for collaboration and information-sharing to counter terrorism and extremism online, and works closely with governments, civil society and academia as well.

In 2020, the GIFCT transitioned to an independent organisation, appointed an inaugural and highly-respected Executive Director in Nicholas Rasmussen, and advanced significantly in the cooperative efforts implemented by its members. The GIFCT has also established an Independent Advisory Committee (which includes a NGO representative from Australia) and now includes a number of industry members.

[20] Data Transparency Advisory Group, *Findings of the Data Transparency Advisory Group,* https://law.yale.edu/sites/default/files/area/center/justice/document/dtag_report_5.22.2019.pdf.

The GIFCT has created a cross-industry database of "hashes" (unique digital fingerprints) of known violent terrorism imagery or propaganda. To date, the Hash Sharing Consortium has reached 320,000 unique hashes in the database - the result of approximately 270,000 visually distinct images and approximately 51,000 visually distinct videos having been added.[21] This helps to improve each company's ability to quickly detect and remove content involving a hash in the database.

The GIFCT has also developed a Content Incident Protocol - an agreed process for how companies will react if a real-world terrorist event triggers the sharing of online content. It was developed in response to the 2019 attacks in Christchurch.

Civil society partnerships

Working with civil society organisations is critical to combatting hate and extremism. We regularly work with civil society organisations to hear feedback on our policies and enforcement, to understand trends and developments on the ground, and to reach memberships of the community at risk of radicalisation.

Some examples of our global partnerships include:

- Creation of a Search Redirect program. Search Redirect helps combat extremism by redirecting hate-related search terms on Facebook towards resources, education, and outreach groups. In 2019, we extended this program to Australia via a partnership with Exit Australia, a local organisation that helps people leave violent extremism and terrorism.
    - On International Holocaust Remembrance Day 2021, we launched a new Search Redirect module related to the Holocaust.[22] Anyone who searches on our platform for terms associated with either the Holocaust or Holocaust denial, will see a message from Facebook encouraging them to connect to the site www.aboutHolocaust.org which was created by the World Jewish Congress with the support of UNESCO (the United Nations Educational, Scientific and Cultural Organization) with the goal of providing people with essential information about the history of the Holocaust and its legacy.
    - We have also developed a Redirect initiative for QAnon. When someone searches for terms related to QAnon on Facebook and Instagram, we will redirect them to credible resources from the Global Network on Extremism and Technology (GNET), the academic research network of

[21] GIFCT, *GIFCT Transparency Report July 2021*, https://gifct.org/wp-content/uploads/2021/07/GIFCT-TransparencyReport2021.pdf.

[22] G Rosen, 'Connecting people to credible information about the Holocaust off Facebook', *Facebook Newsroom*, 27 January 2021, https://fb.workplace.com/groups/waitwhataskpr/permalink/5051911028190805/.

the GIFCT. These resources help inform people of the realities of QAnon and its ties to violence and real world harm.[23]
    - We have launched a similar Redirect Initiative for when people search for QAnon-adjacent terms related to child sex trafficking. When searching for "save the children", our prompt redirects users to the website of the actual NGO Save The Children.
    - Our Search Redirect initiative has been evaluated by Moonshot CVE as part of our commitment to ensuring the effectiveness of our program initiatives.[24]
- Counterspeech initiatives. One of the best methods for pushing back on hate speech is counterspeech: standing up to call out hate. Facebook works with NGOs around the world to support them in undertaking effective counterspeech, and we have created a hub[25] with resources and support specifically for NGOs.

Over the last twelve months, we have continued building partnerships with Australia-based organisations. This engagement has taken a variety of forms, including

- undertaking concerted engagement with representatives from the Australian Jewish and Muslim communities to seek feedback on what they are seeing in relation to anti-Semitism and Islamophobia
- establishing an Australia-specific Combatting Online Hate Advisory Group in October 2020. The Advisory Group contains representatives of marginalised communities, and experts in different forms of online hate such as white supremacy. The Advisory Group meets quarterly, to provide a forum to discuss how industry and civil society can work together more closely to combat online hate in Australia.

This builds on existing partnerships we have had within Australia, including a long-standing nine year partnership with PROJECT ROCKIT to help equip Australian school students with the skills required to engage online safely and push back on online hate.[26]

Working with government and law enforcement

We also work closely with the Australian Government and other governments around the world on combatting terrorist and extreme violent material. We have close

[23] Facebook, 'An update on our enforcement against QAnon', *Facebook Newsroom*, 21 October 2020, https://about.fb.com/news/2020/08/addressing-movements-and-organizations-tied-to-violence/.
[24] Moonshot CVE, *Facebook Redirect Programme: Moonshot Evaluation*, https://moonshotcve.com/facebook-redirect-programme-evaluation-report/
[25] Available at counterspeech.fb.com
[26] R Thomas, 'Young people at the centre', *Facebook Australia Blog*, 8 February 2021, https://australia.fb.com/post/young-people-at-the-centre.

ongoing engagement with law enforcement and security agencies. We have also held sessions with Australian law enforcement and security agencies to discuss high-level trends on the terrorism threat environment within Australia.

As outlined earlier, Facebook is a founding signatory to the Christchurch Call, which was a ground-breaking commitment by governments, industry and civil society to eliminate terrorist and violent extremist content online, led by the New Zealand and French Governments.[27] We signed up to the voluntary nine-point industry plan, which contained a number of commitments to improve our effectiveness in combatting terrorist and extreme violent content. In recognition of the Call, we have been making meaningful progress against the Call's commitments.

We were also a member of the Australian Government Taskforce to Combat Terrorist and Extreme Violent Material Online, and we have been regularly reporting to the Australian Government on the Taskforce commitments since. This has included providing feedback to the Home Affairs Department in developing an Online Crisis Event Arrangement and participating in an Online Crisis Event simulation convened by the Department in October 2020.

Internationally, we have been working with the Australian Government (and other governments) in international fora like the Organisation for Economic Cooperation and Development (OECD). There is significant work underway through the OECD on Voluntary Transparency Reporting Protocols, which was announced and sponsored by the Australian Government.[28] Facebook is the only company to co-lead one of the working groups under this project; we co-lead a working group with the Australian Department of Home Affairs (previously the eSafety Commissioner's Office). We intend to continue to play an industry leadership role to support this important work through the OECD.

Research

In order to ensure our policies and enforcement approach reflects the latest research, we also partner with academics and experts.

Via the GIFCT, we have funded the Global Research Network on Terrorism and Technology (GRNTT) to develop research and provide policy recommendations around terrorists' and extremists' use of the internet. A total of 13 papers were

[27] Facebook, 'Facebook joins other tech companies to support the Christchurch Call to Action', *Facebook Newsroom*, 15 May 2019, https://about.fb.com/news/2019/05/christchurch-call-to-action/.
[28] S Morrison, *More action to prevent online terror*, media release 26 August 2019, https://www.pm.gov.au/media/more-action-prevent-online-terror.

produced and shared in openly accessible formats from the first phase of GRNTT's research.[29]

The second phase of GIFCT's research was led by the International Centre for the Study of Radicalisation (ICSR), based at King's College London. ICSR has established the Global Network on Extremism and Technology (GNET) and brings together an international consortium of leading academic institutions and experts with core institutional partnerships from the US, UK, Australia (The Lowy Institute), Germany and Singapore to study and share findings on combating terrorist and violent extremist use of digital platforms. The next phase of reports GIFCT has funded via GNET are in the process of being released.

These research reports are in addition to the insights reports that GNET publishes multiple times a week, which inform the work of GIFCT members.[30]

Facebook has also funded our own research round on misinformation and polarisation. 25 winners were announced in August 2020 and include two Australian proposals. A number of the successful proposals are examining polarisation (including how it can lead to extremism).[31]

We have also commissioned, funded or otherwise been involved with a number of other research reports relating to terrorism and extremism, including:

- The Centre for Analysis of the Radical Right has undertaken a report on A Guide to Online Radical-Right Symbols, Slogans and Slurs.[32] This includes symbols, slogans and slurs used by Australian members of the radical right.
- The Centre for Analysis of the Radical Right have also provided us with a report on The Many Faces of the Radical Right and How to Counter Their Threat.[33]
- HOPE Not Hate have undertaken a report on the far right on Facebook[34]
- The Henry Jackson Society have delivered the report Free to Be Extreme[35]

---

[29] Global Network on Extremism and Technology, *Reports*, https://gnet-research.org/resources/reports/

[30] Global Network on Extremism and Technology, *Insights*, https://gnet-research.org/resources/insights/

[31] A Leavitt and K Grant, 'Announcing the winners of Facebook's request for proposals on misinformation and polarization', *Facebook Research Blog*, 7 August 2020, https://research.fb.com/blog/2020/08/announcing-the-winners-of-facebooks-request-for-proposals-on-misinformation-and-polarization/

[32] Centre for Analysis of the Radical Right, *A Guide to Online Radical-Right Symbols, Slogans and Slurs*, https://usercontent.one/wp/www.radicalrightanalysis.com/wp-content/uploads/2020/05/CARR-A-Guide-to-Online-Radical-Right-Symbols-Slogan-and-Slurs.pdf.

[33] Centre for Analysis of the Radical Right, *The Many Faces of the Radical Right and How to Counter Their Threat*, https://www.radicalrightanalysis.com/wp-content/uploads/2020/08/CARR-report-oD.pdf

[34] Hope Not Hate, *The Far Right on Facebook: a practical investigation into right-wing hate content on the platform*'.

[35] N Malik for the Henry Jackson Society, *Free to be extreme*, https://henryjacksonsociety.org/wp-content/uploads/2020/01/HJS-Free-to-be-Extreme-Report-FINAL-web.pdf

- Moonshot CVE has evaluated in a report the effectiveness of the Facebook Search Redirect program.[36]

We have also commissioned Australia-specific research to understand the experience of online hate from the perspective of two sets of potentially vulnerable groups:

- Aboriginal and Torres Strait Islander people. Research was conducted by Dr Tristan Kennedy at Macquarie University.
- LGBTQI+ Australians. Research is being conducted by Dr Ben Hanckel from Western Sydney University.
- Asian Australians. The Online Hate Prevention Institute is reviewing and assessing whether Asian Australians have experienced more online hate since the COVID-19 pandemic.

We look forward to continuing to expand our efforts to fund research on hate and extremism in Australia and globally in 2021.

Recommended amendments to the AVM law

Clarifying the definition of AVM content

The legislation defines AVM (s474.2) as online material depicting abhorrent violent conduct, specifically a perpetrator or accomplice:

- engaging in a terrorist act
- murdering another person
- attempting to murder another person
- torturing another person
- raping another person
- kidnapping another person (under threat of violence).

These are reasonable categories of content and behaviour to include in the definition of AVM, and we agree that criminal conduct has no place on our services. The challenge is that quickly identifying this content, and determining whether it constitutes an offence under the Criminal Code is not always straightforward. In some instances, there can be challenges in putting these categories into operation and identifying this content in practice.

Digital platforms have an impetus to respond as quickly as possible, in real time and potentially as a crisis is unfolding, to determine whether a piece of online content falls into one of these categories. In many cases, contextual information will not yet be

[36] Moonshot CVE, *Facebook Redirect Programme: Moonshot Evaluation*, https://moonshotcve.com/facebook-redirect-programme-evaluation-report/.

available, as law enforcement or other stakeholders are still in the middle of responding to the situation.

There are a range of questions that the first responders in a digital platform are required to consider when confronted with a possible piece of AVM. For example:

- Has it been filmed by a bystander? (in which case it would fall outside the offence)
- What is the intention of the perpetrator? (eg. attempted murder is classified as AVM but manslaughter is not)
- Would the footage be subject to a potential defence? (eg. would footage depicting the treatment of those within Don Dale Correctional Facility potentially violate the AVM law, notwithstanding the very high public value of ensuring that important footage like that remains available?)

To assist the Committee, we have provided two anonymised, real-life examples of our operation under the law to demonstrate how unclear the definition can be in the midst of a crisis situation.

Case study 1: Definition of crimes caught by the AVM law
We received an informal report that claimed a parent had taken their own children, for whom they may not have had custody rights, and was livestreaming on Facebook Live. We located the livestream and, on review, it appeared the parent was raising awareness of perceived injustices relating to their children and there was no evidence of threats or harm to the children. Law enforcement could not confirm for us whether it was a kidnapping (ie. whether the parent had custodial rights or not). This case demonstrates that it can be challenging to determine whether a livestream constitutes a kidnapping, as it requires the digital platform to have sufficient knowledge of the surrounding circumstances. However, the AVM law could have resulted in Facebook being deemed "reckless", even though we quickly responded to an informal report and would have had no way of knowing whether or not this instance constituted kidnapping. There would be similar challenges in identifying the motive of an individual in real time for other categories of AVM, such as attempted murder.

Case study 2: Determining the providence of material
We received a formal report (albeit not a notice) that there was a copy of the Christchurch attack video on our services. We located the content and quickly determined it was not the Christchurch attack; a user had shared non-graphic footage from an attack on a mosque in South Asia and it was not clear whether it had been taken by a perpetrator or accomplice, or a bystander. The user appeared to have shared this footage to raise awareness about the risk of violence that they face in their mosque and clearly condemned the violence. This presents significant uncertainty on whether the AVM legislation would consider this content to be AVM.

For those tasked with responding to a potential piece of AVM within a digital platform, they are put in a situation of responding incredibly quickly to a single piece of content, often with no context or background, and with very high stakes. The very high penalties associated with AVM are likely to encourage platforms to err on the side of over-enforcement, which means there could be potentially very valuable or important content that is removed, even if it is not clearly AVM under Australian law.

In order to provide greater clarity and ask digital platforms to make decisions with confidence when they are responding in real time, we recommend an amendment to the legislation that would define AVM as content that is *clearly* or *reasonably identifiable as AVM* depicting one of the nominated categories of abhorrent violent conduct.

Clarifying whether the law sets a proactive monitoring obligation

There has been significant debate following the passage of the law on whether it sets a proactive monitoring obligation for digital platforms to search for terrorist content. The Attorney-General's Department has released a fact sheet that indicates the intention of the Government is not to "criminalise ignorance".[37] The fact sheet also says:

> *The Act does not require providers to take steps to make themselves aware of abhorrent violent material accessible on their platforms and does not require that providers monitor all content on their platforms.*

We appreciate the assistance of the Department in clarifying the Government's intended operation of the law. However, as explained below, it appears that the law may in fact establish a threshold that would inadvertently operate differently.

To be clear, Facebook does proactively detect terrorist and violent content on our services. Even on end-to-end encrypted services like WhatsApp, we look for behavioural signals that could represent terrorist groups and we monitor unencrypted surfaces for content that violates our policies.

We support content regulation frameworks that encourage companies to build robust and risk-based systems for moderating content, including systems that are proactively detecting harmful content.[38]

---

[37] Australian Government Attorney-General's Department, *Sharing of Abhorrent Violent Material Act Fact Sheet*, https://www.ag.gov.au/sites/default/files/2021-09/AVM%20_Fact_Sheet%282021%29.PDF.

[38] M Bickert, *Charting a way forward: online content regulation*, white paper released February 2020,

However, it is important that content regulation does not inadvertently inhibit proactive detection, or penalise those companies who look for harmful content proactively. Proactive detection and enforcement will always be imperfect: artificial intelligence has limits, and technology is simply not able to account for every possible human behaviour in advance. An obligation to proactively detect all AVM without fail would be impossible, given the vast amount of content on the internet.

It is also possible that content may not be immediately apparent as violating the AVM law (especially if it falls in the ambiguous grey areas of the definition of AVM). And so, laws that potentially penalise or hold companies liable for proactively detecting a piece of content and not actioning it immediately and with perfect accuracy, could pose a deterrence or disincentive for companies to proactively monitor in the first place.

The provision of the law that establishes whether a person or company is held to be 'reckless' (s474.34) indicates they are not liable if they are 'not aware' of the content. Liability for failing to expeditiously remove AVM is not expressly tied to awareness of the AVM but rather an awareness of the "risk" that the service can be used to access AVM.

Courts could take an expansive view of what constitutes awareness: it is possible that awareness could include, for example, the existence of a single media report (even if there is no URL or identifiable way of finding the content), an informal text message from a stakeholder, or artificial intelligence proactively detecting a piece of content but not actioning it (even for understandable reasons).

In this way, the current draft of the law appears to be out of alignment with the Government's stated intention and understanding of the law.

We recommend amending the legislation to provide greater clarity on whether it represents a proactive monitoring obligation (and ensuring it does not deter those companies who choose to proactively monitor for harmful content), given this would reflect the Government's intention as stated in guidance from the Attorney-General's Department. This could be addressed in a number of ways. The legislation could specifically define 'awareness' as after a company has received a formal notice from the eSafety Commissioner; or it could define awareness as both instances where a person is aware of the content *and aware that it constitutes AVM*.

---

https://about.fb.com/wp-content/uploads/2020/02/Charting-A-Way-Forward_Online-Content-Regulation-White-Paper-1.pdf.

We recognise that ensuring that the law appropriately incentivises industry and holds it accountable with respect to TVEC is important to our society, consequently, we stand ready to assist with any further insights that can assist with drafting amendments to improve and clarify the law.

Clarifying when defences apply

Similar to the definition of AVM content, the current drafting of defences (s474.37) is seemingly so narrow as to provide uncertainty for a digital platform in the moment of responding to a real-time crisis.

The defences are particularly challenging around excluding news content. We saw significant sharing of the Christchurch attack video on our services by news organisations in particular, so they can play an active role in distributing AVM for news purposes.

The current defence is limited to content published by someone in a professional capacity as a journalist (as well as relating to a news or current affairs report, and being in the public interest). Under the current drafting of the law (and absent any further guidance), digital platforms are required to quickly ascertain whether the journalism defence applies. Digital platforms are in no position to understand whether the person posting a piece of AVM content from a news page is working in a professional capacity as a journalist or not, and it would be challenging to assess whether the content is “in the public interest”. Without clearer direction, the defence is very challenging for a digital platform to apply.

We suggest the eSafety Commissioner could be responsible for making assessments and determinations about whether a defence would apply - if a digital platform indicates they see a potential case to consider a defence. While the material is under consideration by the eSafety Commissioner, the service should not be liable for any offence for failing to expeditiously remove content.