12 February 2021

Committee Secretary
Parliamentary Joint Committee on Intelligence and Security
PO Box 6021
Parliament House
Canberra ACT 2600

By email: pjcis@aph.gov.au

Dear Committee,

Thank you for the opportunity to provide a written submission to the Australian Parliamentary Joint Committee on Intelligence and Security (PJCIS) regarding the inquiry into extremist movements and radicalism in Australia.

Communities around the world have been impacted by incidents of mass violence, terrorism, and violent extremism with tragic frequency in recent years. These events demand a robust public policy response from every sector, and the inclusion of leaders from affected communities, domain experts, and Governments to prevent future attacks while supporting healthy and cohesive societies. We also know technology companies play a critical role in addressing these issues and that content removal online cannot alone solve these issues and challenges in the near or long term.

Twitter has been at the forefront of responding to the evolving challenge of preventing violent extremism and terrorist exploitation of the Internet. Our approaches combine a clear, consistent policy philosophy with flexible, evolving capabilities that enable us to move quickly to take action in these instances while staying true to our commitment to an open and safe internet.

Our rules exist to ensure all people can participate in the public conversation freely and safely on Twitter. Our work will never be complete, as the threats we face constantly evolve.

Twitter is committed to working with the Australian Government, as well as industry, academia, and vital civil society as we continue to build our shared understanding of these issues, and to find optimal ways to approach them together.

Kind regards,

[redacted]

Kara Hinesley
Director of Public Policy
Australia and New Zealand

[redacted]

Kathleen Reen
Senior Director of Public Policy
Asia Pacific

Overview

Twitter is a place where people from around the world come together in an open and free exchange of ideas and conversations. As those factors and the features of our service develop, our policies and enforcement options evolve continuously as well to address new and emerging behaviours online.

Tackling terrorism, violent extremism, and preventing violent attacks requires a whole of society response, including from technology companies. It has long been a priority of Twitter to remove this content from the service. Twitter has no incentive to keep terrorist and violent extremist content available on our platform. It is against our rules. Such content does not serve our business interests, and is fundamentally antithetical to our values.

This submission focuses on providing insights into Twitter's work to protect the health of the public conversation and our approaches to these challenging issues. In this submission, we will cover our efforts to combat terrorism, violent extremist groups, hateful conduct, platform manipulation, and our global and local partnerships and broader societal engagement.

Twitter policies and protecting the health of the conversation

Twitter's purpose is to serve the public conversation, and all individuals accessing or using Twitter's services must adhere to the policies set forth in the Twitter Rules.[1] Accounts under investigation or that have been detected as sharing content in violation of the Twitter Rules may be required to remove content, or in serious cases, will see their account permanently suspended.[2]

The Twitter Rules prohibit violent threats and the promotion of terrorism and violent extremism.[3]

Specifically, we do not allow users to make specific threats of violence against an individual or group of people, or threaten or promote violent extremism or terrorism. There is no place on Twitter for violent organisations, including terrorist organisations, violent extremist groups, or individuals who affiliate with and promote their illicit activities.

Additionally using a behaviour-led approach along with a combination of machine learning and human review, we are able to prioritise reports and take action quickly. While content is important, we also leverage behaviour-based signals that look at how accounts behave before we look at the content they are posting. This is how we seek to scale our efforts globally and accelerate our action.

Policy on Terrorism

[1] Twitter, 2021. The Twitter Rules. [online] Help.twitter.com. Available at: <https://help.twitter.com/en/rules-and-policies/twitter-rules> [Accessed 12 February 2021].
[2] Twitter, T., 2021. *The Twitter Rules*. [online] Help.twitter.com. Available at: <https://help.twitter.com/en/rules-and-policies/enforcement-options> [Accessed 12 February 2021].
[3] Twitter, T., 2021. The Twitter Rules. [online] Help.twitter.com. Available at: <https://help.twitter.com/en/rules-and-policies/violent-groups> [Accessed 12 February 2021].

Individuals are prohibited from making specific threats of violence, or to wish for the serious physical harm, death, or disease of an individual or group of people. This includes, but is not limited to, threatening or promoting terrorism.

Since 2015 we have suspended more than 1.7 million accounts for violations related to promotion of terrorism.[4] In our latest reporting period from January to June 2020, action was taken on 90,684 unique accounts under this policy, with 94% of those accounts proactively identified and actioned. Our current methods of surfacing potentially violating content for review also include leveraging the shared industry hash database supported by the Global Internet Forum to Counter Terrorism (GIFCT).[5]

We continue to see a sharp downward trend in the number of accounts actioned in recent years. This likely reflects the changing behaviour patterns of bad actors, and significant improvements in our defences, for example, making it much harder for bad actors to compromise accounts, which was a commonly used technique by some terrorist organisations.

Due to our zero-tolerance policy enforcement, we have been able to take swift action on account ban evaders (i.e. accounts permanently suspended that seek to reestablish a presence on Twitter) and other identified forms of behaviour used by terrorist entities and their affiliates. In the majority of cases, we take action at the account creation stage before the account even Tweets.

Twitter has also taken concrete steps to reduce the risk of livestreaming being abused by terrorists, while recognising that during a crisis these tools can also be used by news organisations, citizens and governments and via a range of channels and platforms. We have automated prioritisation of live video reports and made improvements to our hash technology for uploaded images, videos, GIFs and also better leverage hashes shared by industry partners to improve content detection and policy enforcement efforts. We also developed proprietary technology to proactively flag violent and gory content in media for human review and actioning, which guards against viral proliferation and copycat behaviour, while ensuring we do not remove media that is documenting violence and potential human rights abuses.

We are reassured by the progress we have made, including recognition by independent experts. For example, Dublin City University Professor Maura Conway found in a detailed study that *"ISIS's previously strong and vibrant Twitter community is now...virtually non-existent."*[6]

We also know that the challenges we face are not static, nor are bad actors homogenous from one country to the next in how they evolve and behave. Our approach therefore combines flexibility with a clear, consistent policy philosophy, enabling us to move quickly while establishing clear norms of what is unacceptable behaviour on Twitter.

---

[4] Twitter, 2021. Transparency Centre. [online] Transparency.twitter.com. Available at: <https://transparency.twitter.com/en.html> [Accessed 12 February 2021].
[5] Global Internet Forum to Counter Terrorism, 2021. Joint Tech Innovation. [online] GIFCT.org. Available at: <https://gifct.org/joint-tech-innovation/#row-hash> [Accessed 12 February 2021].
[6] Conway, M., 2021. Disrupting Daesh: Measuring Takedown of Online Terrorist Material and Its Impacts. [online] Taylor & Francis. Available at: <https://www.tandfonline.com/doi/full/10.1080/1057610X.2018.1513984> [Accessed 12 February 2021].

Policy on Hateful Conduct

People on Twitter are not permitted to promote violence against or directly attack or threaten other people on the basis of race, ethnicity, national origin, sexual orientation, gender, gender identity, religious affiliation, age, disability, or serious disease. We also do not allow accounts whose primary purpose is inciting harm toward others on the basis of these categories.[7]

We do not allow individuals to use hateful images or symbols in their profile image or profile header. Individuals on the platform are not allowed to use the username, display name, or profile bio to engage in abusive behaviour, such as targeted harassment or expressing hate toward a person, group, or protected category.

Under this policy, we take action against behaviour that targets individuals or an entire protected category with hateful conduct. Targeting can happen in a number of ways; for example, mentions, including a photo of an individual, or referring to someone by their full name.

When determining the penalty for violating this policy, we consider a number of factors including, but not limited to, the severity of the violation and an individual's previous record of rule violations. For example, we may ask someone to remove the violating content and serve a period of time in read-only mode before they can Tweet again. Subsequent violations will lead to longer read-only periods and may eventually result in permanent account suspension. If an account is engaging primarily in abusive behaviour, or is deemed to have shared a violent threat, we will permanently suspend the account upon any initial review.

Manipulation of the Public Conversation

As stated above, our policies regarding terrorism, violent extremist groups, and hateful conduct are strictly enforced, as are all our policies. Another area that fills out our wider approach to these challenges is our expanding work to address manipulation; we take additional steps to safeguard the public conversation on Twitter from manipulation.[8]

As a uniquely open, public service, we proactively enforce policies and use technology to halt the spread of content propagated through manipulative tactics on Twitter, such as automation, or attempting to deliberately game trending topics on Twitter.

Our Site Integrity team is dedicated to identifying and investigating suspected platform manipulation on Twitter, including activity associated with coordinated malicious activity that we are able to reliably associate with state-affiliated actors. A partnership of teams across the company, employ a range of both open-source and proprietary signals and tools to identify when attempted coordinated manipulation may be taking place, as well as the actors responsible for it.

---

[7] Twitter, 2021. The Twitter Rules. [online] Help.twitter.com. Available at: <https://help.twitter.com/en/rules-and-policies/hateful-conduct-policy> [Accessed 12 February 2021].
[8] Twitter, 2021. The Twitter Rules. [online] Help.twitter.com. Available at: <https://help.twitter.com/en/rules-and-policies/platform-manipulation> [Accessed 12 February 2021].

We challenged 135.7 million accounts in our last public reporting period in our Twitter Transparency Report for the above behaviours. We did so by requesting additional details, like email addresses and phone numbers, in order to authenticate the account.[9]

We partner closely with governments, law enforcement agencies, academia, expert researchers, including from civil society organisations, and our peer companies to improve our understanding of the actors involved in information operations and develop a holistic strategy for addressing them.

In line with our principles of transparency and to improve public understanding of manipulation and inauthentic influence campaigns, Twitter also makes publicly available archives of Tweets and media that we believe resulted from state-linked information operations on our service.[10] This archive is a unique resource. It is the only database of its kind produced by industry. It is our fundamental belief that these accounts should be made public and searchable so that members of the public, governments, and researchers can investigate, learn, and build media literacy capacities for the future. Since 2018, we have expanded this dataset considerably with several separate updates over the past couple of years. We're the only company to offer this level of granularity and transparency.

For our part, we are continuously learning, evolving, and building a technological and personnel-driven approach to combating inauthentic influence campaigns. We hope that holistic, transparent disclosures such as this can help us all learn and build the necessary societal defenses and capacities to protect public conversation.

Partnerships and societal engagement

We work with law enforcement and numerous public safety agencies and organisations in Australia, as well as around the world. As our partnerships deepen, we are able to better respond to the changing and emerging threats we all face together.

Our wider efforts on countering violent extremism going back to 2015 have focused on bolstering the voices of non-governmental organisations and credible outside groups to use our uniquely open service to spread positive and affirmative campaigns that seek to offer an alternative to narratives of hate and extremism.

We have partnered with organisations delivering counter and alternative narrative initiatives across the globe, and we encourage the Committee to consider the role of the Australian Government in supporting the work of the credible and critical messengers in this space at home and abroad.

The GIFCT, Industry and Academic Collaborations

[9] Twitter, 2021. Transparency Centre. [online] Transparency.twitter.com. Available at: <https://transparency.twitter.com/en/reports/platform-manipulation.html#2020-jan-jun> [Accessed 12 February 2021].
[10] Twitter, 2021. Transparency Centre. [online] Transparency.twitter.com. Available at: <https://transparency.twitter.com/en/reports/information-operations.html> [Accessed 12 February 2021].

We know collaboration with our industry peers and civil society is critically important to addressing common threats from terrorism globally.

In June 2017, we launched the Global Internet Forum to Counter Terrorism (GIFCT) in a partnership with YouTube, Facebook, and Microsoft. The GIFCT facilitates, among other things, vital information sharing, technical cooperation, and research collaboration, including with academic institutions.

In September 2019 during the UN General Assembly, the GIFCT announced it would be reorganised as an independent Non-Governmental Organisation (NGO). This came with an updated mission statement, that includes directly addressing violent extremism, which broadened the scope of focus in line with the Christchurch Call to Action and the multistakeholder work together. The announcement also included plans to reorganise the GIFCT goals, while maintaining its output and impact to date. The GIFCT is now an independent 501(c)(3) registered in the United States and it has an independent Executive Director and staff.

As a founding member of the GIFCT, Twitter has assisted in this evolution of the organisation led by an dedicated Executive Director and supported by technology, counterterrorism, and operations teams, as well as, an inaugural Independent Advisory Committee (IAC). As the Operating Board Chair for 2021, Twitter is continuing to dedicatedly support the work of the GIFCT as it deploys annual programs, training, and expands on its transparency efforts.

One of the most important and impactful recent developments of the GIFCT is the real-time content incident protocol (CIP), established to enable us to respond to violent acts involving an online component quickly. This is to ensure that companies are able share valuable information across industry to limit the spread of terrorist and violent extremist content. The GIFCT's CIP now ensures even deeper collaboration between industry following an incident. This protocol was first deployed following the October 2019 terrorist attack in Halle, Germany.[11] These processes continue to evolve over time. They sustain and deepen industry collaboration and capacity, while incorporating the advice of key civil society and government stakeholders.

As part of that industry collaboration, Twitter also initiatied partnerships with 13 companies to share URLs of content that have been shared by accounts suspended by Twiter for the promotion of terrorism. These might be links to other services where people could access files, longer videos, PDFs, and other materials. More than 13,000 unique URLs have been shared to date.[12]

This information sharing ensures the hosting companies can monitor and track similar behaviour, taking enforcement action pursuant with their individual policies. This is not a high-tech approach, but it is simple and effective, recognising the resource constraints of smaller companies in the digital ecosystem, and the wider impact that removing content at its source has in disrupting the spread of terrorist content. We are eager to partner with additional companies to expand this project, and we look forward to building on our existing partnerships in the future.

---

[11] Global Internet Forum to Counter Terrorism, 2021. Statement Halle Shooting. [online] GIFCT.org. Available at: <https://gifct.org/press/gifct-statement-halle-shooting/> [Accessed 12 February 2021].
[12] Global Internet Forum to Counter Terrorism, 2021. About GIFCT. [online] GIFCT.org. Available at: <https://gifct.org/about/> [Accessed 12 February 2021].

Our commitments to practical and vital partnerships towards lasting solutions include an array of critical efforts: Twitter is a participant in the Aqaba Process.[13] We are a signatory to the Christchurch Call to Action[14] and the Australian Taskforce to Combat Terrorism and Extreme Violent Material Online.[15] We are invested in the Global Research Network on Terrorism and Technology (GRNTT) to develop research and policy recommendations designed to prevent terrorist exploitation of technology.[16] We are a member of the Online Hate Observatory working on developing a better understanding of the mechanics behind online hate to build better answers in cooperation with non-government organisations (NGOs), researchers, and relevant governments.[17] Twitter is also actively involved in the development and consultation of the joint Australia-New Zealand government funded Organisation of Economic and Cooperative Development (OECD) Voluntary Transparency Reporting Protocols.

We are committed to continuing to update our products and policies to address online behaviours while growing these critical partnerships with industry peers, expanding our wider mentoring efforts, strengthening our new crisis protocol arrangements, and supporting the expansion of these organisational and coalition memberships across industry into the future.

Beyond these, our wider efforts on countering violent extremism support many different non-governmental organisations and groups around the world. These organisations and groups can use our uniquely open service to spread positive and affirmative campaigns that seek to offer an alternative to narratives of hate. In terms of local engagement, our teams have worked on online campaigns, virtual programs, best practice training sessions, and provided support through our pro-bono advertising grant program, Ads for Good, to a number of organisations. Programs have also included the DIGI Engage[18] conference for the past three years in partnership with the Australian Department of Home Affairs, which brought together young people who participated in training focused on overcoming combat hate, division, and extremism within our online and real-world communities.[19]

Ideologies can only be successfully and authentically countered by those who have the credibility to take on the core messages being propagated. If these core messages go unchallenged then we know the removal of content will always be an incomplete and unsuccessful response to the larger challenges. These groups and organisations do critical work and policy makers should

[13] UN OCT, 2021. [online] Available at: <https://twitter.com/UN_OCT/status/1176257320311054339?s=20> [Accessed 12 February 2021].
[14] Christchurchcall.com. 2019. Christchurch Call | to eliminate terrorist and violent extremist content online. [online] Available at: <https://www.christchurchcall.com/> [Accessed 12 February 2021].
[15] Australian Prime Minister & Cabinet. 2019. [online] Available at: <https://www.pmc.gov.au/resource-centre/national-security/report-australian-taskforce-combat-terrorist-and-extreme-violent-material-online> [Accessed 12 February 2021].
[16] RUSI. 2021. The Global Research Network on Terrorism and Technology. [online] Available at: <https://rusi.org/projects/global-research-network-terrorism-and-technology> [Accessed 12 February 2021].
[17] Crif - Conseil Représentatif des Institutions Juives de France. 2021. *Crif - Crif publishes first results of its Online Hate Speech Observatory*. [online] Available at: <http://www.crif.org/en/actualites/crif-crif-publishes-first-results-its-online-hate-speech-observatory> [Accessed 12 February 2021].
[18]Digi Engage. 2021. Home. [online] Available at: <https://digiengage.live/> [Accessed 12 February 2021].
[19]Junkee. 2020. DIGI Engage Is The Wholesome & Inspiring Online Event We Need Right Now. [online] Available at: <https://junkee.com/digi-engage-2020/253879> [Accessed 12 February 2021].

continue to find ways to broaden support for these vital, preventative, mitigating and societal building efforts.

The Christchurch Call to Action and Australian Taskforce to Combat Terrorism and Extreme Violent Material Online

Because terrorism cannot be solved by the tech industry alone, the Christchurch Call to Action has proven to be a landmark global multi-stakeholder initiative, one that convenes and unites governments, industry, and civil society in novel ways behind mutual commitments to eliminate terrorist and violent extremist content online, while upholding the principles of freedom of expression and an Open Internet.[20]

It is important to recognise the role of society and leaders within or a part of online communities and the roles they continue to vitally play in the essential work to build safe, welcome and healthy communities. As a uniquely open service, at Twitter we see regular examples around the world of the people who use our service, communities, and groups successfully challenging hate and division, particularly following violent public acts. As the world began to comprehend the horror of what took place in Christchurch in March 2019, some may have sought to promote hate, but there was another conversation taking place, one that reached many, many more people, one with more resonance as well. For example, The hashtag #HelloBrother saw people around the world recognising the brave act of one victim and rejecting the terrorist's narrative, while hundreds of thousands of Tweets expressed similar sentiments in their own way. The hashtag #TheyAreUs, which emerged the day after the attacks, expressed a message of solidarity generating over 37,000 Tweets alone in the initial 24-hour period after the attacks. This shows the potential of open public conversation and what it can empower — a global platform for the best of society to challenge violence and hatred.

In fulfilling our commitments in the Christchurch Call, we have taken a wide range of actions and continue to work to meet the Call. Twitter invests in technology to prioritise signals, including user reports, to ensure we can respond as quickly as possible to a potential incident, building on the work we have done to harness proprietary technology to detect, and disrupt, bad actors proactively.

As part of our commitment to educate users about our rules and to further prohibit the promotion of terrorism or violent extremist groups, we have also updated our rules and associated materials to be clearer where these policies apply. These are accompanied by more data being provided in our Transparency Report, published approximately every six months, allowing public consideration of the actions we are taking under our rules, as well as how much content is detected by our proactive efforts.[21]

In furtherance of this important work, we have also partnered with academics at the University of Otago's National Centre for Peace and Conflict Studies (NCPACS) through our #DataforGood program to use Twitter data to study the ways online conversations can be used to promote

[20] Christchurchcall.com. 2019. Christchurch Call | to eliminate terrorist and violent extremist content online. [online] Available at: <https://www.christchurchcall.com/> [Accessed 12 February 2021].
[21] Twitter, 2021. Transparency Centre. [online] Transparency.twitter.com. Available at: <https://transparency.twitter.com/> [Accessed 12 February 2021].

tolerance and inclusion instead of division and exclusion.[22] Preliminary research just a week after the Christchurch massacre analysed data generated from tens of thousands of public Tweets anchored to the violence, highlighting a local and global outpouring of support for victims, solidarity with the citizens of New Zealand, the affirmation of democratic ideals, pushback against terrorism, and unequivocal condemnation of the perpetrator.

Additionally, Twitter is one of the original member companies to the Australian Taskforce to Combat Violent Terrorist and Extreme Material Online. Twitter is committed to working together with industry and the Australian Government to help keep Australians safe as per the recommendations identified in the 2019 Consensus Report.[23] Namely, the Consensus Report identified actions that are focused on 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 Christchurch attacks and are also consistent with principles contained within the Christchurch Call to Action.

A whole of society response

The challenges we face as a society are complex, varied, and constantly evolving. These challenges are reflected and often magnified by technology. The push and pull factors influencing individuals vary widely, and there is no one solution to prevent an individual turning to violence. This is a long-term problem requiring a long-term response, not just the removal of content.

While we strictly enforce our policies, removing all discussion of particular viewpoints, no matter how uncomfortable our customers may find them, does not eliminate the ideology underpinning them. Quite often, it moves these views into darker corners of the Internet where they cannot be challenged and held to account. As companies improve in their efforts, this content continues to migrate to less-governed platforms and services. We are committed to learning and improving, but every part of the online ecosystem has a part to play.

We have a critical role, but tech companies and online content removal will not solve these issues in isolation. They are systemic and societal and so they require a whole of society approach. We welcome the opportunity to continue to work with the Australian Government, our industry peers, academics, and civil society to find the right solutions.

Conclusion

Our goal is to protect the health of the public conversation and to take immediate action on those who seek to spread messages of terror and violent extremism. However, no solution is perfect, and no technology is capable of detecting every potential threat or protecting societies and communities from extremism and violent threats on their own. We know that the challenges we

[22] Hinesley, K., 2021. Our #DataForGood partnership with New Zealand's NCPACS. [online] Blog.twitter.com. Available at: <https://blog.twitter.com/en_us/topics/company/2020/christchurch-otago-nspacs.html> [Accessed 12 February 2021].
[23] Prime Minister & Cabinet. 2019. [online] Available at: <https://www.pmc.gov.au/resource-centre/national-security/report-australian-taskforce-combat-terrorist-and-extreme-violent-material-online> [Accessed 12 February 2021].

face are not static, nor are bad actors homogenous from one country to the next in how they evolve, behave, or the tactics they deploy to evade detection.

Twitter's efforts in Australia and around the globe to support civil society voices and promote positive messages have seen Twitter employees train groups on five continents and provide pro-bono advertising to groups to enable their messages to reach millions of people. When we talk about the health of the public conversation at Twitter, we see the principles of civility, empathy, and mutual respect as foundational to our work. We know evolving our policies, products and partnerships are critical to the evolving challenges. We will not solve problems by removing content alone. We should not underestimate the power of open conversation to change minds, perspectives, and behaviours.

We stand ready to assist the Committee in its important work and will continue to work on ways Internet companies can stop the spread of terrorist and violent extremist content on our services.