22 July 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 Chair,

Thank you for the opportunity to provide responses to questions taken on notice following Twitter's appearance as part of the Australian Parliamentary Joint Committee on Intelligence and Security (PJCIS) regarding the inquiry into extremist movements and radicalism in Australia.

We have endeavoured to provide answers to the best of our ability for questions outlined by the Committee below.

Twitter is committed to working with the Australian Government, our industry partners, non-government organisations and wider civil society as we build our shared understanding of the issues and find optimal ways to approach these together.

Please don't hesitate to let us know if there is any additional information we can provide to assist the Committee. Thank you again for the opportunity to provide input as part of this important legislative process.

Kind regards,

[redacted]

Kara Hinesley
Director of Public Policy
Australia and New Zealand

[redacted]

Kathleen Reen
Senior Director of Public Policy
Asia Pacific

Australian Parliamentary Joint Committee on Intelligence and Security

Inquiry into extremist movements and radicalism in Australia

Answers to Questions on Notice
Twitter Inc.

1. **Have you seen any evidence from your platforms that the banning of content actually reinforces pre-existing prejudices and views and actually makes it harder to combat the risks of extremism? — Mr Tim Wilson, MP**

Our mission is to increase healthy public conversation, and we continually evaluate the incentives created by our product and policies to adjust for the mission.

At Twitter, our research has largely been focused on the value of counterspeech and framing efforts to reduce hate and its impacts online. In partnership with community groups and social cohesion organisations, we found that if someone is at a vulnerable point in their life and susceptible to dogmatic views espoused by charismatic actors, the sharing of a counter-narrative can be the offer of another path; one that is hopeful for their future, takes an optimistic view of life, and reminds them of all they can't afford to leave behind – family, friends, and community.[1]

At Harvard University's Berkman Klein Center, research led by Susan Benesch along with Daniel Jones, examines how practitioners of "counterspeech" can use social media to battle hateful, bigoted messaging and ideology, and how broadcasting hateful comments to a larger audience – which may seem counterintuitive when the focus is frequently on deleting content – may actually draw the attention of a larger audience to a particular piece of speech, which can serve as an educational tool.[2]

Through programs like Radicalisation Awareness Network's (RAN) #ExitHate campaign in Europe[3] or our DIGI Engage conferences that have been running annually in Australia since 2016,[4] Twitter has consistently been a part of the broader discussion about how effective online campaigns can be vehicles for such social change and what it means to create counter-narratives.

Twitter, by virtue of its public and distributed nature, can showcase and share powerful alternative narratives. Through trainings and workshops we have run around the world, we have highlighted examples like #IllRideWithYou,[5] an organic response to the fear some Muslims had of wearing religious clothing on public transport after the Sydney Lindt Cafe siege; and #NotInMyName[6] when British Muslims decried the acts of Daesh to illustrate how the extremists had strayed from a peaceful, mainstream Muslim community. We saw the #StopIslam[7] movement trend, borne of prejudice after the attacks on Brussels, but which was quickly met with overwhelming criticism by Twitter users who cited the hashtag and rejected the intolerant content of the minority.

Recently using Twitter data, the University of Otago and National Centre for Peace and Conflict Studies (NCPACS) undertook research regarding the conversation on Twitter around the Christchurch terror attacks. By using data generated from tens of thousands of public Tweets anchored to the violence, the research discovered 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. For example, the hashtag #TheyAreUs emerged, offering a message of solidarity and generating 37,000 Tweets in 24 hours alone.[8]

These examples offered an uplifting view of humanity at times when it was most needed, proving that the Twitter community is capable of great empathy and solidarity, much like the world it reflects.

---

[1] https://blog.twitter.com/en_us/a/2016/twitter-recently-joined-young-people-in-vienna-to-talk-about-alternative-narratives-changing
[2] https://cyber.harvard.edu/story/2019-08/combating-hate-speech-through-counterspeech
[3] https://blog.twitter.com/en_us/a/2016/twitter-supports-radicalisation-awareness-network-campaign-encouraging-europeans-to-exithate
[4] https://digi.org.au/events/
[5] https://twitter.com/hashtag/IllRideWithYou
[6] https://twitter.com/hashtag/NotInMyName
[7] https://twitter.com/hashtag/StopIslam
[8] https://blog.twitter.com/en_us/topics/company/2020/christchurch-otago-nspacs

2. How long have you been using machine learning algorithms to target advertisements and target users of your services? — Ms Celia Hammond, MP

Like many peer companies, Twitter's business is largely based upon advertising, but we have some key differences.

In order to explain how targeting works for advertisements on Twitter, it's important to understand Twitter's Promoted Products and how that data is used. While Twitter may collect data, the type of data and what it is linked to may be significantly different than our peers, depending upon how people use Twitter. In general, rather than focusing on who you are, our data is more about what a person is interested in; for example, what a user Retweets, Likes, Follows – all indicators of which are public by default.

Twitter's promoted products include Promoted Ads, Follower Ads, and Trend Takeover. They are clearly marked with a "Promoted" icon, and users can interact with most promoted content in much the same way as organic content. For promoted content, a person's activity on Twitter, the information they provide when they create an account, and our relationships with ad partners all help tailor relevant promoted content.

Twitter Privacy Policy

We work hard to make our global Privacy Policy clear and easy to understand.[9] We transparently outline the types of data people share with us, how we use that data, and highlight the meaningful controls people have over both. Key updates to our Privacy Policy include:

- We've added more explanations about the information an advertiser may get when a person engages with an advertisement on Twitter and about how we process their data at Twitter when they use our products and services.
- We've made it clearer that when a person Tweets, their Tweets are publicly shared with all of the people who use Twitter's services, including the developers who make use of our APIs[10] – unless the Tweets are from a protected account.[11]
- We've renamed "Personalising across your devices" to "Personalising based on your inferred identity" to clarify for people what this setting controls.[12]

Based on feedback we receive, we continue to refine our Privacy Policy to make sure it's as clear and understandable as possible.

Promoted Products on Twitter

Twitter does not currently offer a programmatic advertising program for publishers like some other companies. However, in addition to our policies, Twitter has invested in a suite of solutions aimed at ensuring a safe advertising experience for everyone who uses the platform. Recently, Twitter has used Machine Learning (ML) to ensure ads are not served near objectionable content and to predict the value of ad requests.[13]

Safeguarding Advertising Experiences on Twitter

In addition to the controls available to everyone on Twitter and brand safety controls for advertisers, Twitter leverages a combination of Machine Learning (ML) and human review to ensure that ads do not serve around objectionable content.

Twitter prevents ad placement adjacent to Tweets that have been deemed "Sensitive Content" by our Twitter Service Team or by the Tweets' authors.[14] People on Twitter are also able to self-classify their Tweets as sensitive.

Every Tweet from our partners goes through two levels of scrutiny before it becomes monetisable through our Amplify Pre-Roll program[15]:

- Algorithmic check which scans Tweet text for any unsafe language;

[9] https://twitter.com/en/privacy
[10] https://help.twitter.com/en/rules-and-policies/twitter-api
[11] https://help.twitter.com/en/safety-and-security/public-and-protected-tweets
[12] https://help.twitter.com/en/about-personalization-across-your-devices
[13] https://blog.twitter.com/engineering/en_us/topics/insights/2020/using-machine-learning-to-predict-the-value-of-ad-requests
[14] https://help.twitter.com/en/rules-and-policies/media-policy
[15] https://business.twitter.com/en/advertising/campaign-types/amplify-pre-roll.html

- Manual human review of every single monetised video to ensure that they meet our Brand Safety standards; and
- We also hold regular proactive educational sessions with our partners to help them successfully monetise their content on Twitter within our brand safety standards.

Brand Safety Controls for Ads on Profiles

Every time a profile is updated, our ML model searches the content of the profile page with the goal of ensuring that content is brand safe, according to our Brand Safety policies, before a Promoted Ad is served. We only serve ads on profiles that we deem to be safe for ads. We may also block ads from serving on individual customer profiles based on the content or behavior of the account and lack of alignment with our Brand Safety policies.

3. **If you were legally liable for potentially defamatory comments left on your platform, what would you do to protect yourself from that? What steps would you have to take to protect yourself financially and as a corporate entity? — Chair**

In general, Twitter is not in a position to arbitrate the veracity of allegations and requires valid legal process to remove material that is alleged to be defamatory.

It is important to recognise that Internet intermediaries are often not in the same position as an originator to assess whether content is defamatory. Internet intermediaries are not armed with the background information, evidence, or requisite knowledge in order to make this assessment. To reflect the practicalities of dissemination of defamatory matter through Internet infrastructure, intermediaries should not be considered publishers unless and until the intermediary is on notice that the content has been deemed defamatory by a court of law.

Our Twitter Rules make clear that users are responsible for the content they post, and there are a wide-ranging set of rules that Twitter enforces when content is posted that infringes those rules.[16]

An individual can make a report to Twitter directly from a Tweet, List, Direct Message, or Profile for certain violations of our Twitter Rules or Terms of Service. Additionally, our teams are able to undertake a review of content when a report is filed via our Legal Request Submissions Site with valid legal process attached.[17]

4. **Why shouldn't we make recommendations to do things to your algorithms to prevent you from allowing people to go down into the echo chamber? — Mr Julian Lesser, MP**

Since Twitter's creation over a decade ago, we have offered people many ways to control their experience on the platform. For instance, Twitter has invested considerable resources into providing people with a greater level of control and choice over how algorithms affect the content they might see on the service.

Twitter uses a range of behavioural signals to determine how Tweets are organised and presented in the Home Timeline, Conversations, and Search based on relevance. Twitter relies on behavioural signals – such as how accounts behave and react to one another – to identify accounts that detract from a healthy public conversation, such as spam and abuse.

In the Home Timeline, for example, Twitter has allowed people to choose between viewing the top Tweets first (i.e. the Top Tweets are based on accounts and content the user interacts with the most), or seeing the most recent Tweets first (i.e. Tweets are displayed in reverse chronological order with the most recent Tweet first) since introducing an algorithmic timeline in 2016.[18] This option can be easily identified and executed by tapping the 'sparkle icon' in the main app interface, and these options are actively communicated in official public communication channels, like company blogs, official accounts, and the Twitter Help Centre.[19]

Additionally, the simple use of a hashtag can display a diversity of opinions across Twitter. A hashtag—written with a # symbol—is used to index keywords or topics on Twitter.[20] This function was created on Twitter, and allows people to easily search for conversations or follow topics across the entire platform. If a person Tweets with a

[16] https://help.twitter.com/en/rules-and-policies#twitter-rules
[17] https://help.twitter.com/en/rules-and-policies/twitter-legal-faqs
[18] Blog.twitter.com. Never Miss Important Tweets From People You Follow. [online] Available at: https://blog.twitter.com/official/en_us/a/2016/never-miss-important-tweets-from-people-you-follow.html <Accessed May 2021>
[19] Twitter Support, 2018, Available at: https://twitter.com/TwitterSupport/status/1075506036818104320 <Accessed May 2021>; Twitter Help Centre, 2021. Available at: https://help.twitter.com/en/using-twitter/twitter-timeline <Accessed May 2021>.
[20] https://help.twitter.com/en/using-twitter/how-to-use-hashtags

hashtag from a public account, anyone who does a Search for that hashtag may find that Tweet, enabling them to discover all different types of content.

Further to these public facing mechanisms, we have engaged with rapidly developing, critical subject areas such as algorithmic transparency, choice, and ethics. In April 2021, we announced the establishment of Twitter's Responsible Machine Learning Initiative, led by our internal Machine Learning Ethics, Transparency and Accountability (META) team. This team's responsibilities include driving transparency about our machine learning decisions, how we arrived at them, and vitally, enabling agency and algorithmic choice for people on Twitter.[21] We believe it is important to understand the agency held by the individual when using the Twitter service and the choices available to them regarding how algorithms might affect what they see.

Twitter believes algorithmic transparency and choice provides important competition for digital platforms. In its regulatory approach, Australia should encourage frameworks that encourage formats that give Australians flexibility in terms of content, conversations, and choice.

[21] Blog.twitter.com. Introducing our Responsible Machine Learning Initiative, https://blog.twitter.com/en_us/topics/company/2021/introducing-responsible-machine-learning-initiative.html, <Accessed July 2021>.