14 February 2021

Online Safety Branch, Content Division
Department of Infrastructure, Transport, Regional Development and Communications
GPO Box 594
Canberra ACT 2600

By email: OnlineSafety@infrastructure.gov.au

Dear Online Safety Branch,

Thank you for the opportunity to provide a written submission to the Australian Department of Infrastructure, Transport, Regional Development and Communications regarding the *Online Safety Bill 2020* (**the Bill**).

We share the Australian Government's desire to promote online safety, and Twitter remains focused on helping people feel safe, secure, and empowered to participate in the public conversation every day.

As we continue to iterate and strengthen our approach to meet evolving contours and challenges surrounding online behaviours, we're moving with urgency, purpose, and commitment to develop and enforce a range of policy, procedural, and product changes to help people feel safe, welcome, and to control their experience on Twitter. We support smart regulation, and our focus is on working with governments to ensure that regulation of the digital industry is practical, effective, feasible to implement while remaining inclusive and keeping core democratic values intact while promoting tech innovation, including Twitter's core commitment to an Open Internet worldwide.

Twitter is committed to working with the Australian Government, our industry partners, non-government organisations and wider civil society as we continue to build our shared understanding of the issues and find optimal ways to approach these together. We trust this written submission will be a useful input to the Department's consultation process. Working with the broader community we will continue to test, learn, and improve quickly so that our platform remains open, accessible, effective, and safe for everyone.

Thank you again for the opportunity to input to this important process and landmark legislation.

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 shares the Australian Government's view that online safety is a shared responsibility and that social media service providers, relevant electronic services, and designated internet service providers play an important role in protecting the community from harmful content online.

We emphasise at the outset that online content regulation requires a proportionate approach to balance protections from harm with human rights and other vital interests, including freedom of expression, privacy, and procedural fairness. This balance ensures that companies and regulators alike have clearly delineated responsibilities regarding protections for users' rights, as well as a shared commitment to foster a diverse public square consistent with community expectations.

This submission will address key areas outlined in the Bill, as well as detail corresponding to advancements where Twitter has introduced changes to our policies, processes, and product to achieve a healthier, safer platform.

Basic online safety expectations

Part 4 of the Bill outlines that the Minister may determine basic online safety expectations (BOSE) for social media services, relevant electronic services, and designated internet services, and that those specified services may be required to give the eSafety Commissioner reports about compliance with the applicable BOSE.[1]

Delivering the highest safety standards requires investment, something that a range of companies are committed to and have a track record of doing. Companies should also be incentivised to prioritise such resourcing on illegal and harmful content where the interventions are well-tested with proven effectiveness and empirical data. Any requirements, like the recommended BOSE, should seek to outline best practices rather than prescriptive requirements, and should avoid placing requirements across the digital ecosystem that only an established subset of companies can attain.

In order to continue to foster digital growth and innovation in the Australian economy, and to ensure reasonable and fair competition, it is critically important to avoid placing requirements across the digital ecosystem that only large, mature companies can reasonably comply with. In addition to size and scale, an effective approach for responding to a specific safety issue on Twitter, for instance, may be very different for another platform given inherent differences in products and services across companies.

Moreover, before these expectations are introduced, we would strongly encourage additional independent research on the part of the Australian Government to understand online harms. We recommend that research be as inclusive, representative, and as publicly available as possible. As the issues rapidly evolve, we, along with researchers and third sector organisations, are still learning what

[1] Department of Infrastructure, Transport, Regional Development and Communications 2021. [online] Available at: <https://www.communications.gov.au/have-your-say/consultation-bill-new-online-safety-act> [Accessed 12 February 2021].

the most effective interventions are to meet those rapidly moving changes. Recognising the vital ongoing research, analysis, and academic debate underway, it is also critical that neither the government nor a regulatory authority are overly prescriptive in the setting of any basic expectations and ensure any legislated requirements are technically feasible.

Reporting provisions

Division 3 of Part 4 of the Bill focuses on periodic reporting about compliance with BOSE and periodic reporting determinations made by the eSafety Commissioner. Regarding recommendations around additional reporting, we strongly believe in providing public, straightforward data, and information that provides useful insights into the types of requests we receive from governments, law enforcement, and others around the world, which we make public in our biannual Twitter Transparency Report.[2]

Meaningful transparency between companies, regulators, civil society, and the general public is fundamental to the work we do at Twitter. This transparency is a key tenet of our efforts to preserve and protect the Open Internet. We also understand that genuine, open collaboration between industry, government, and the third sector is required to address these issues. In line with this philosophy, for the past eight years our biannual Twitter Transparency Report has highlighted trends in requests made to Twitter from around the globe.[3]

Over time, we have significantly expanded the information we disclose in our reports, adding metrics on platform manipulation, Twitter Rules enforcement, including abuse, child sexual exploitation, hateful conduct, private information, sensitive media, and violent threats. We also report our proactive efforts to eradicate child sexual exploitation, the promotion of terrorism and violent extremism from our service.

Recognising that the public as well as policy makers and regulators want to be better informed of our enforcement processes, we also launched our new Twitter Transparency Centre in 2020 to make our data easier to understand and analyse for those who access our Transparency Reports.[4] We now include sections covering information requests, removal requests, copyright notices, trademark notices, email security, Twitter Rules enforcement, platform manipulation, and state-backed information operations.

We also upload legal requests for content removal directly into the Lumen database, a long-term partnership with the Berkman Klein Center at Harvard University.[5] We encourage the public to explore these resources to better understand the day-to-day requests Twitter receives from governments, law enforcement, and other entities around the world.

---

[2]Twitter, 2021. Transparency Centre. [online] Transparency.twitter.com. Available at: <https://transparency.twitter.com/en.html> [Accessed 12 February 2021].
[3]*Ibid.*
[4]Blog.twitter.com. 2021. Insights from the 17th Twitter Transparency Report. [online] Available at: <https://blog.twitter.com/en_us/topics/company/2020/ttr-17.html> [Accessed 12 February 2021].
[5]Lumendatabase.org. 2021. Search :: Lumen. [online] Available at: <https://lumendatabase.org/twitter> [Accessed 12 February 2021].

In this context, we would encourage the Australian Government to implement a single transparency reporting framework consistent with global standards with regular reporting intervals that would allow for proper allocation of resources instead of setting up a potentially onerous and ad hoc system of transparency reporting. For example, the Santa Clara Principles on Transparency and Accountability in Content Moderation provide a global minimum standard for acceptable transparency reporting in content moderation.[6] These principles set out the basic requirements of legitimate content moderation systems, including requirements to publish regular aggregate figures, provide notice to affected individuals, and establish accountable review processes for content moderation decisions.

We hope these types of global protocols will prove instructive in the development or adaptation of the Australian Government's transparency reporting guidelines within the context of this Bill, and we would strongly recommend that government efforts to improve industry transparency reporting practices use these principles as a starting point in order to promote industry compliance with global best practice norms.

Additionally in line with our commitment to provide greater transparency, Twitter is the only major service to proactively make public conversation data available via an application programming interface (API) for the purposes of research.[7] By harnessing the power of the Twitter API, partners are able to tap into the public conversation and study collective issues facing global communities to bring about new insights to universal issues, devise fresh approaches to problems, and foster social good.

Building on this work, this year Twitter launched the Academic Research product track in order to enable academic researchers to access increased data from the public conversation to study topics that are as diverse as the conversations on Twitter. This track provides qualified academics the opportunity to access new endpoints, including the full history of public conversation data, a higher volume of Tweets, and more precise filtering capabilities.[8]

In line with our principles of transparency and to improve public understanding of inauthentic influence campaigns, Twitter has also published public archives of Tweets and media that we believe resulted from state-backed information operations.[9] We have proactively expanded these datasets with several separate updates over the past couple of years, and we're the only company to offer this level of granularity and transparency.

We believe the open exchange of information can have a positive global impact, and thus through our efforts to provide meaningful transparency, endeavour to earn public trust, and enable accountability. Our transparency reports reflect not only the evolution of the public conversation on our service, but the work we do every day to protect and support the people who use Twitter. We understand that

---

[6]Santa Clara Principles. 2021. Santa Clara Principles on Transparency and Accountability in Content Moderation. [online] Available at: <https://santaclaraprinciples.org/> [Accessed 12 February 2021].
[7]Developer.twitter.com. 2021. Advancing Academic Research with Twitter Data. [online] Available at: <https://developer.twitter.com/en/solutions/academic-research> [Accessed 12 February 2021].
[8]Blog.twitter.com. 2021. Enabling the future of academic research with the Twitter API. [online] Available at: <https://blog.twitter.com/developer/en_us/topics/tools/2021/enabling-the-future-of-academic-research-with-the-twitter-api.html> [Accessed 12 February 2021].
[9]Twitter, 2021. Transparency Centre. [online] Transparency.twitter.com. Available at: <https://transparency.twitter.com/en/reports/information-operations.html> [Accessed 12 February 2021].

these issues require genuine and deep collaboration between industry, regulators, academics, the media, and civil society. To that end, we will continue to make strides in creating a healthier service as we look for opportunities to provide more transparency while balancing privacy considerations.

Harmonisation of online content regulatory schemes

We are supportive of the Bill's overarching purpose to harmonise and achieve better consistency across the five different content schemes under the eSafety Commissioner's purview including: the existing Cyberbullying Scheme for children in Part 5; the existing Image-based Abuse Scheme in Part 6; the new proposed Cyber-abuse Scheme for adults in Part 7; the new proposed Abhorrent Violent Material Blocking Scheme in Part 8; and the new Online Content Scheme in Part 9 originally under schedules 5 and 7 of the *Broadcasting Services Act 1992* (BSA).

We also believe that the current sequencing of the complaints process outlined in the Cyber-bullying Scheme for children and Cyber-abuse Scheme for adults, which directs people to file reports with a service provider in the first instance, is the most expedient way to respond to potential violations of our rules. This reflects the policies and protections that service providers have in place to protect users and the eSafety Commissioner's role as a safety net for Australians. This process is the most efficient way to enable companies to remove violative or harmful material and help people as quickly as possible.

However, the sequencing for the complaints process is different for the Image-based Abuse Scheme and Online Content Scheme, where an individual can complain to the eSafety Commissioner in the first instance before alerting the service provider that violative or harmful content may be on their service.

People are able to make reports directly to Twitter when they are concerned there has been a violation of the Twitter Rules or local law.[10] We are then able to prioritise the review of reports based on the potential severity and immediateness of harm so we can act expeditiously to respond and remove violative content from our service. We would encourage the Government to adapt the legislative framework so that all the schemes adjudicated by the eSafety Commissioner follow the same, consistent process of filing reports with the relevant service provider in the first instance and help platforms tackle violative behaviour in all of its manifestations as quickly as possible.

24 hour removal timeframe

Over the years, Twitter has worked closely with the Office of the eSafety Commissioner since its creation through the *Enhancing Online Safety Act 2015* (EOSA). Under the EOSA, Twitter signed up as a Tier 1 service, and participates in the scheme on an ongoing cooperative basis.[11] We actively engaged in the consultation process for developing the image-based abuse civil penalties regime, and

[10]Twitter, 2021. Help Centre. [online] Help.twitter.com. Available at: < https://help.twitter.com/forms> [Accessed 12 February 2021].
[11]Office of the eSafety Commissioner 2021. [online] Available at: <https://www.esafety.gov.au/about-us/consultation-cooperation/working-with-social-media> [Accessed 12 February 2021].

the consultation process for the Safety by Design (SbD) principles, an important initiative by the Office of the eSafety Commissioner to both mitigate and address the wide range of safety challenges that occur through digital products and services. We provided submissions outlining our work in detail and participated in workshops hosted by the eSafety Office and the Department.

We also worked with the Office of the eSafety Commissioner through the Online Safety Consultative Working Group (OSCWG) and act as a member of the newly formed eSafety Advisory Committee (EAC). The eSafety Commissioner has visited our headquarters in San Francisco, and we have facilitated high level meetings with company executives in the U.S., Singapore and Australia. We are in regular communication with the eSafety Office and cooperate on the small number of removal requests that have been escalated to the eSafety investigations team for resolution.

Twitter has promptly responded to requests under the EOSA law over the past five years, and have in place rapid response protocols with the eSafety Office; thus, we would appreciate further clarification regarding the perceived need to reduce the turnaround time from 48 to 24 hours for the cyber-bullying and image-based abuse schemes.

In the Department's Discussion Paper published last year, both the French illicit content law and the German *Network Enforcement Act* (NetzDG) were used as examples of international frameworks that justify why response times should be shortened to 24 hours. What the discussion paper failed to surface was that the legislative regimes referenced in the paper are narrowly focused and tailored primarily towards removal of hate speech rather than broad content schemes, such as that drafted in the Bill.

Given the proposal to shorten turnaround times to 24 hours would apply to vast types of content covered under the Bill, there may be frequent factors that necessitate a longer review period. The shortened time frame will make it difficult to accommodate procedural checks on possible errors in reports, the removal of legitimate speech, and providing necessary user notices. We believe the Government would want these procedural protections in place as a matter of consumer and citizen protections. These also guard against potential overreach and protect freedom of expression for Australians. We believe it's possible to act quickly, safely as needed, and still achieve these balanced outcomes.

Most importantly, the Discussion Paper released last year states that online service providers already work closely with the eSafety Commissioner in the administration of the current content schemes, and the Office already experiences prompt removal from online service providers when they are issued with a report. Therefore, it's unclear why it is necessary to further reduce and codify the turnaround time from 48 to 24 hours when the eSafety Commissioner routinely states that companies remove content promptly.[12]

We also believe the Bill would benefit from additional clarifications regarding intermediary liability in exceptional cases where investigation of a complaint may take more than 24 hours. With edge cases

[12]Department of Infrastructure, Transport, Regional Development and Communications 2021. [online] Available at: <https://www.communications.gov.au/have-your-say/consultation-new-online-safety-act> [Accessed 12 February 2021].

that involve additional investigation, the current response time of 48 hours can allow for reasonable timeframes that provide our teams with the opportunity to work closely with in-house experts and eSafety investigators to fully review a report and ensure that we are able to put people first while also operating a fair system of enforcement in line with our obligations.

The Department's Discussion Paper published last year also noted that voluntary arrangements and liaison processes that have been put in place by social media and service providers for the eSafety Office have "been successful in the vast majority of cases in relation to Australian-hosted internet content, cyberbullying material and intimate images." In the context in which these current processes are working well and voluntary compliance is the norm, we would caution against imposing new legal obligations or introducing new legal powers for regulators above and beyond what is required to adjudicate the five online content schemes. To the extent that new powers for regulators and administrative officials are introduced, we strongly urge that they are accompanied by a high standard of transparency and accountability for how they are exercised.

Additionally, we believe that there is room for clarification for the current wording of the Bill. For example, terms such as "serious" are used to quantify the level of harm that needs to occur in order for content to be within scope of the scheme. However, these determinant terms are not properly defined within the Bill, and the corresponding vagueness around these determinant terms adds ambiguity when conducting content moderation where specific context is key. This introduces another layer of complexity, and when compounded by a shorter removal timeframe, adds friction to the overall content moderation process. If the language for determinant terms can be more explicitly scoped, it reduces ambiguity in the interpretation of the Bill and will facilitate content review.

Accountability and oversight

The exposure draft of the Bill introduces significant expansion of the eSafety Commisioner's information gathering, investigative, and disclosure powers. Part 13 sets out a new power for the eSafety Commissioner to obtain information about the contact details or identity of an end-user from a social media service, relevant electronic service, or designated internet service if necessary.

As digital platforms are not primary publishers of content, we believe enabling the eSafety Commissioner to make end-user information requests and levy civil penalties, where appropriate, could help deter individual behaviours that lead to the propagation of harmful and illegal content online.

However, Part 14 of the Online Safety Bill deals with how the eSafety Commissioner conducts investigations, and mirrors existing powers currently in Part 13 of the BSA. Under this section, the eSafety Commissioner will have the power to summon a person by written notice to appear before the eSafety Commissioner to produce documents or to answer questions; or to provide documents or other information to the eSafety Commissioner relevant to the subject matter of the investigation.

Twitter works closely with federal and state law enforcement agencies in Australia in the course of their investigations. As a global company, Twitter exercises due diligence to respect local laws in jurisdictions around the world, and duly reviews all legal processes. We also maintain dedicated contact channels for law enforcement and respond to legal processes issued in compliance with

applicable law.[13]

As currently drafted, the Bill essentially confers quasi-judicial and law enforcement powers to the eSafety Commissioner without any accompanying guidelines or guardrails around which cases would constitute grounds for the Commissioner to exercise these powers other than the very broad 'serious harm' definition. In a law enforcement setting, such investigative powers are typically accompanied through a well-established legal process, like obtaining a warrant. Section 224 allows for the eSafety Commissioner to refer matters to law enforcement agencies at the Commissioner's discretion; however, we believe that the process as currently drafted lacks appropriate oversight and due process (i.e. independent judicial authorisation and the establishment of probable cause). Any obligation on any social media service provider, relevant electronic service, or designated internet service provider to either disclose user information as well as remove content, must be balanced with procedural fairness.

Thus, the expansion of the eSafety Commissioner's powers that are currently proposed under the Bill should be coupled with concomitant levels of oversight. While section 220 of the Bill allows for an application to the Administrative Appeals Tribunal (AAT) to review a specific decision by the eSafety Commissioner, there are no regular or routine avenues for recourse, oversight processes, or transparency structures to counterbalance the extraordinary discretion vested within the eSafety Commissioner. The eSafety Office has a substantial public facing responsibility, voice and role in the community, and transparency and clarity in this area is fitting and necessary.

Therefore, as drafted, this process lacks that necessary structural oversight and accountability in the way that builds public trust and understanding. For example, there is a lack of counter-notice processes for people that are being accused under the various schemes, as well as a lack of consistent reporting obligations on the part of the Office of the eSafety Commissioner. It would be beneficial for the eSafety Commissioner to provide transparency and openly share data on the Office's metrics and methodologies that would allow for independent evaluation, research, and understanding about the actions and operations, especially as the eSafety Commissioner's remit and powers expand. This context would also allow for greater transparency and accountability for how the Office of the eSafety Commissioner adjudicates discretion under the relevant legislative schemes.

Accordingly, we would encourage the Government to include robust oversight processes in the Bill to ensure consistency of decision-making and documentation of actions taken so they can be open to scrutiny, review, and to build public confidence and trust in responses from both relevant companies and the eSafety Office alike.

Conclusion

Twitter is engaged in open dialogue with governments around the world as we seek to foster collaborative partnerships and continue to drive forward online safety solutions. Across all areas, the

[13]Twitter 2021. Help Centre. Guidelines for law enforcement. [online] Available at: <https://help.twitter.com/en/rules-and-policies/twitter-law-enforcement-support> [Accessed 12 February 2021].

investments Twitter has made to protect the health of the public conversation are now generating clear and tangible safety benefits for people who use our service.[14]

Our work will never be complete as the threats we face constantly evolve. Going forward, we look forward to continuing to work with the Australian Government, the Office of the eSafety Commissioner, civil society, nonprofits, and industry to address online safety and work to create lasting global solutions to build a safer and open Internet.

[14]Blog.twitter.com. 2021. A healthier Twitter: Progress and more to do. [online] Available at: <https://blog.twitter.com/en_us/topics/company/2019/health-update.html> [Accessed 12 February 2021].