12 November 2021

Director, Online Safety Reform and Research Section
Department of Infrastructure, Transport, Regional Development and Communications
[redacted]

By email: OnlineSafety@infrastructure.gov.au

Dear Director,

Thank you for the opportunity to provide a written submission and feedback on the draft *Online Safety (Basic Online Safety Expectations) Determination 2021*.

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 our commitments 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, inclusive, and keeps certain core democratic values intact while promoting tech innovation, including our core commitment to an Open Internet worldwide.

Twitter is committed to working with the Australian Government, our industry partners, non-government organisations, academics, 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 into this important process.

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 online service providers play an important role in protecting the community from harmful content online.

Our submission stands together with the respective submissions from the Digital Industry Group Inc. (DIGI) and Communications Alliance (Comms Alliance), both of which Twitter is a member. For clarity, and to complement and reinforce these statements, we've structured this submission to address the key issues contained in the BOSE as they pertain to Twitter operating in Australia, as well as detail advancements where we have introduced relevant changes to our policies, processes, and product to achieve a healthier, safer platform and protect the people that use Twitter.

We trust this written submission, together with Comms Alliance and DIGI's submissions, will be useful inputs to the Government's work. We urge and encourage the Department to consider this holistic feedback while undertaking the review of this draft legislative instrument.

Scope and definitional concerns

We understand that the purpose of the Basic Online Safety Expectations (**BOSE**) is to reflect the expectations that the Government has for online service providers and to set the benchmark for online services to take proactive and preventative steps to protect Australians from abusive conduct and harmful content online.

As drafted, the BOSE limit the provision of, or access to, certain online content (i.e. "material") that is unlawful or "may be harmful," and it also outlines actions that are required of service providers – ranging from reporting, record keeping, complaint handling – upon the identification of such material.

As currently drafted with such a broad scope, however, we would argue that the BOSE substantially expands the remit of the *Online Safety Act 2021* (Cth) (**OSA**). In reality, many of the expectations go well beyond any basic measures and international best practice.

For example, there is ambiguity within the BOSE with regards to definitions, as the draft instrument states that it applies to material that "may be harmful," which does not reflect the definitions contained within the OSA and may require service providers to make subjective decisions regarding the definition of "harm." The instrument also uses the term "safe" to refer to thresholds within multiple expectations, yet the term is not defined in the legislative instrument, nor is it defined in the OSA. Additionally, the BOSE does not clearly articulate or contain any public interest exemptions with the draft instrument.

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.

Expectation 6: provider will take reasonable steps to ensure safe use

Expectation 6(1) asks providers to "take reasonable steps to ensure that end-users are able to use the [respective] service in a safe manner." However, this expectation creates a standard of care regarding digital content that is far in excess of the standards that apply offline for content in the form of printed materials, broadcast materials, or films, which conflicts with the BOSE Consultation Paper that states that the key principle underlying the OSA is that "the rules and protections we enjoy offline should also apply online."[1]

What is also important to consider is the breadth of businesses subject to this requirement. As drafted, every

[1] https://www.infrastructure.gov.au/sites/default/files/documents/draft-online-safety-basic-online-safety-expectations-determination-2021.pdf

business with an online presence is required to comply, regardless of whether it is a public website or closed internet environment. As noted, key terms in the BOSE lack clear definitions. Thus, this expectation raises the following questions:

- How is "safe" defined in this context? The concept of safety is subjective – while one user may feel safe, another user (of the same objective category of vulnerability) may not share that feeling. Additionally, the term "safe" is undefined within the BOSE and the OSA; and
- How is "harmful" defined in this context? Providers are expected to minimise material or activity on the service that is or may be harmful, which is also subjective and remains undefined in the BOSE.

Lack of definitional clarity could lead to both conflicts in interpretation and confusion when it comes to operationalising the expectations. Given the central nature of the term "safe" in applying the BOSE, we propose it be defined for the purpose of this instrument to mean material that is not illegal, but with regards to which the eSafety Commissioner has powers under the OSA, i.e. cyber-bullying material, cyber abuse material, and image-based abuse material.

The lack of clarity of what is expected is even more concerning given that providers are expected to ensure that end-users are able to use the online service in a "safe" manner.

Furthermore, this unrealistic expectation of ensuring complete safety is coupled with a lack of a safe harbour approach that would give providers certainty that their efforts are indeed meeting the expectations, and that they would not incur civil liability for good faith actions taken to comply with the BOSE.

The fact that the BOSE does not impose a duty that is enforceable by proceedings in a court cannot detract from what seems to be an intention to substantially extend the OSA from a 'notice and take-down' regime to a "proactively detect, moderate, report and remove" regime.

While Expectation 6(3) lists a (non-exhaustive) number of "reasonable steps" that could be taken to "ensure" that services can be used in a "safe" manner, taking some or even all, of these steps does not seem to guarantee compliance with Core Expectation 6 as Core Expectation 7 indicates that the providers will need to consult with the eSafety Commissioner "in determining what are reasonable steps for the purposes of [Core Expectation 6]."

Not only do providers have to consult the Commissioner, they also have to take into account any relevant guidance material made available by the eSafety Office in determining what are reasonable steps (Additional Expectation 7(2)). This could create significant operational burdens as the BOSE does not set out any structure or regular cadence by which the eSafety Commissioner would inform companies of new or updated guidance, which could create impossible deadlines to keep up with for bespoke engineering, product, and policy developments.

We would recommend that the language in Expectation 6 be amended to focus on reducing the risk of harm arising as a direct consequence of use of services to "as reasonably practicable" since "safety" is a subjective and contentious term. Harms should also be clearly limited to those covered by OSA to avoid regulatory overlap.

Expectation 7: provider will consult with the eSafety Commissioner and refer to Commissioner's guidance in determining reasonable steps to ensure safe use

The Consultation Paper establishes that "service providers are best placed to identify these emerging forms of harmful end user conduct or material, and so the flexibility of this regime means that providers choose the best way to address them on their service in the most responsive way."

We agree that service providers are best placed as to what reasonable steps to take to enable a safe use of our respective services. However, we are alarmed by the substantial discretionary powers that Core Expectation 7 grants to the Office of the eSafety Commissioner. While the Consultation Paper suggests that the service providers' expectation to consult with the Commissioner could be discharged by "seeking the advice of the Commissioner" or "following guidance issued by the Commissioner," we are concerned what this will mean in

practice.

This is of particular importance as the eSafety Commissioner has been assigned the powers to judge whether a provider has contravened one or more BOSE in section 48 of the OSA. In effect, it gives the eSafety Commissioner unprecedented power to intervene in companies' business operations with no level of oversight of the Commissioner's engagements with companies and no ability for companies to contest the suitability of the "guidance" provided. The reporting provisions of the OSA in effect enable the Commissioner to audit if companies are following her guidance; however, the Commissioner is not required to publish a company's response to being named and shamed.

It is also not clear how a provider is expected to proceed in the scenario where a provider and the eSafety Commissioner have differing opinions and/or guidance as to what would constitute "reasonable steps" to "ensure" that end-users are able to use the service in a "safe" manner.

We are concerned that the advice that the Office of the eSafety Commissioner will be providing in direct communications to individual online service providers will lead to diverging advice and practice. This not only bears the risk of an inconsistent approach but may also impact competition, as different providers could receive different guidance on the obligations and constraints they are subject to, which could lead to an unfair advantage in certain situations.

We would recommend that the eSafety Commissioner be compelled to produce written guidance for specific sectors and types of business based on risk/cost benefits, similar to Australian Tax Office (**ATO**) rulings that are publicly made available. This will ensure greater transparency and consistency in the Commissioner's engagement with businesses and companies subject to the OSA.

Expectation 9: provider will take reasonable steps regarding anonymous accounts

The assumption that anonymous accounts are at higher risk of perpetuating abuse than those where the user's real name is revealed is flawed. Expectation 9 requests that providers take reasonable steps to prevent anonymous accounts from being used to deal with unlawful or harmful material. At the same time, the expectation proposes identify verification as one possible reasonable step.

Pseudonymous accounts

At Twitter, we are guided by our values, and never more so than when it comes to fundamental issues like identity.

To be clear, pseudonymity is not a shield against Terms of Service violations, and Twitter takes action against pseudonymous accounts that are in violation of the Twitter Rules. It is against the rules to have a fake account on Twitter, and when creating a new account on Twitter, our users have to provide a verified phone number or email address when signing up. Twitter also supports the removal of illegal content while balancing the need to protect free expression, and our teams duly review legal requests that we receive.

We believe, however, that everyone has the right to share their voice without requiring a government ID to do so. Our approach in this space has been developed in consultation with leading NGOs. While pseudonymity has been a vital tool for speaking out in oppressive regimes, it is no less critical in democratic societies like Australia.

Pseudonymity may be used to explore a person's identity, to find support as victims of crimes, or to highlight issues faced by vulnerable communities. Indeed, many of the first voices to speak out on societal wrongdoings, have done so behind some degree of pseudonymity. Once they do, their experience can encourage others to do the same, knowing they don't have to put their name to their experience if they're not comfortable doing so.

Perhaps most fundamentally of all, some of the communities who may lack access to government IDs are exactly those who we strive to give a voice to on Twitter. Additionally, empirical evidence overwhelmingly points

to anonymity bans as ineffective.[2] Currently, there is not conclusive evidence that requiring the display of names and identities will reliably reduce social problems, and many studies have documented the problems it creates, like posing real threats to vulnerable communities online that rely on anonymity — victims of domestic violence, members of the LGBTQIA+ community, political and human rights activists, journalists, whistleblowers and informants, to name a few.[3]

In international human rights law, the International Covenant on Civil and Political Rights (ICCPR) and Universal Declaration of Human Rights (UDHR) – both of which Australia is a signatory – enshrine the right to freedom of expression, including the freedom to "guard information and ideas anonymously."[4] Additionally, former UN Special Rapporteur on the Promotion and Protection of the Right to Freedom of Opinion and Expression Frank LaRue has said that "[i]n order for individuals to exercise their right to privacy in communications, they must be able to ensure that these remain private, secure and, if they choose, anonymous."[5] Thus, tools like encryption and anonymity empower individuals to circumvent barriers and access information, and such expression in the public interest is at the "core of the concept of democratic society."[6]

In other countries, for example, the right to communicate anonymously or under a pseudonym is protected through constitutional law or common law.

In the United States, before a service provider may be compelled to unmask an anonymous speaker: (1) a reasonable attempt to notify the user of the request and the lawsuit must be made; and (2) the plaintiff must make a prima facie showing of the elements of the asserted cause of action.[7] Moreover, the party seeking discovery must demonstrate a compelling need for the anonymous speaker's identity rooted in the idea that "obvious invasions of interests fundamental to personal autonomy must be supported by a compelling interest"[8] and that a litigant "does not have a generalized right to rummage at will through information that [another party] has limited from public view."[9]

In South Korea, the Government imposed a law in 2004 requiring users to provide their national identification numbers before posting on election-related websites. In 2007, the requirement to identify users was broadened to all sites with more than 300,000 daily visitors. Studies show that during the time the policy was in operation, there was no significant decrease in online abuse. In fact, the Korean Communications Commission found that 'hateful' comments decreased by less than 1% during the first year the policy was in force. Other studies found short-term decreases in online participation and the number of violent comments, but saw no long-term changes. The policy doesn't appear to have prevented the spread of misinformation or conspiracy theories, and later saw a massive hack in which 35 million South Koreans' national identification numbers were stolen.[10]

The internet is not a monoculture; it is a rich variety of subcultures which engage with anonymity and identity in diverse ways. Posting anonymously can allow people to protect themselves so they can openly discuss and deal with complex topics safely. It can also allow people to speak out about abuse and seek information.

For Twitter to endure, it needs to provide an environment for users to feel safe in communicating on the platform. To provide this environment, Twitter maintains the freedom for people to speak truth to power while also removing bad faith actors on the platform who intend to use it to divide, threaten, or manipulate.

Platform manipulation

---

[2] Michigan State University (Huang, G & Li, K 2016), 'The effect of anonymity on conformity to group norms in online contexts: A meta-analysis' International Journal of Communication, vol. 10, no. 1, pp. 398–415. <https:// joc.org/index.php/ijoc/article/view/4037>.
[3] Matias, J. N. (2017). Why Real Names Don't Fix Trolling. <https://guides.coralproject.net/real-names-dont-fix-trolling/>.
[4] https://www.ohchr.org/en/professionalinterest/pages/ccpr.aspx
[5] https://www.ohchr.org/Documents/Issues/Opinion/Communications/States/Selected_References_SR_Report.pdf
[6] *Ibid.*
[7] Krinsky v. Doe, 72 Cal. Rptr. 3d 231, 239, 244–46 (Cal. Ct. App. 2008); Glassdoor, Inc. v. Super. Ct., 9 Cal. App. 5th 623 (2017)
[8] Williams v. Superior Court, 3 Cal. 5th 531, 557 (2017).
[9] Dkt. 7; see also Fed. R. Civ. P. 26(b); Tompkins v. Detroit Metro. Airport, 278 F.R.D. 387, 388-89 (E.D. Mich. 2012
[10] https://www.aspistrategist.org.au/naming-names-wont-stop-abuse-on-social-media/

People are not permitted to use Twitter in a manner intended to artificially amplify, suppress information, or engage in behavior that manipulates or disrupts other people's experience on the service.

We prohibit the creation or use of fake accounts. We also do not allow spam or platform manipulation, such as bulk, aggressive, or deceptive activity that misleads others and disrupts their experience on Twitter.

Some of the factors that we take into account when determining whether an account is fake include the use of stock or stolen avatar photos; the use of stolen or copied profile bios; and the use of intentionally misleading profile information, including profile location.

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. This includes building new proprietary systems to identify and remove ban evaders at speed and scale.

We routinely identify suspicious account activity, such as exceptionally high-volume Tweeting with the same hashtag or mentioning the same @handle without a reply from the account being addressed. When we identify such activity, we require an individual using the service to confirm human control of the account or their identity.

We have increased our use of challenges intended to catch automated accounts, such as reCAPTCHAs (that require individuals to identify portions of an image or type words displayed on screen), and password reset requests that protect potentially compromised accounts.

In our most recent Transparency Report, we challenged over 143 million accounts for engaging in suspected spammy behaviour, including those engaged in suspected platform manipulation.[11]

Since 2018, we also introduced a registration process for developers requesting access to our application programming interfaces (APIs) to prevent the registration of spammy and low quality apps, and we are continuing to roll out improvements to our proactive enforcements against common policy violations.[12]

Automation and automated accounts

People often refer to bots when describing everything from automated account activity to individuals who would prefer to be anonymous for personal or safety reasons, or avoid a photo because they've got strong privacy concerns.

In sum, a bot is an automated account. With regards to automation, our rules specifically state that platform manipulation and spam are prohibited on Twitter. People cannot use Twitter's services in a manner intended to artificially amplify or suppress information or engage in behaviour that manipulates or disrupts people's experience on Twitter.

It's important to note, however, that not all forms of automation are violations of the Twitter Rules. We've seen innovative and creative uses of automation that enrich the Twitter experience. For example, accounts that track air quality, earthquakes, or general reminders to drink your water like @tinycarebot.

Automation can also be a powerful tool. For example, a conversational bot can help surface information about orders or voting information, such as the Twitter Direct Message Chatbot set up during the 2019 Australian Election that provided voting information from the Australian Electoral Commission.[13] This kind of innovative tooling has proved safe and efficient for a myriad of civic and corporate functions, especially at a time of social distancing when digital communications proved essential to public health contingency plans.

[11] https://transparency.twitter.com/en/reports/platform-manipulation.html#2020-jul-dec
[12] https://blog.twitter.com/developer/en_us/topics/tips/2018/automation-and-the-use-of-multiple-accounts
[13] https://blog.twitter.com/en_au/topics/company/2019/get--ausvotes2019-election-information-through-twitter-

Unfortunately, much of the non-peer reviewed and commercially-driven research we see making sweeping assessments about automated accounts is based on deeply flawed methodologies or technical understanding of automation.

This means when groups of bots or incidents of malicious automation are identified by researchers, they are unable to factor in defensive measures taken by Twitter. Our actions to limit the spread of spammy or automated content are not available to developers or researchers using the public APIs, and the presence of a bot account on Twitter is not an indication that content from that user is shown or distributed in the same way as organic content. The end user experience of someone using the Twitter app or website is not replicated by looking at an unfiltered stream of content obtained via our public API. Therefore, the actions we take – such as challenging, filtering, and removing accounts – are not reflected in research.

Abuse and harassment

We're committed to enabling safe and healthy conversations on the service. We work with safety advocates, academics, researchers, and community groups that support our work to prevent abuse, harassment, or attempts to intimidate or silence someone else's voice. By providing continuous feedback on our safety mechanisms, these partners help us maintain a safe environment.

As part of that commitment, we have introduced a number of recent updates and policies to reduce abuse and harassment, which have resulted in:

- Impressions for rule-violating Tweets: Our impressions metric captures the number of views a violative Tweet received prior to removal. From July - December 2020, Twitter removed rule-violating 3.5 million Tweets. Out of those Tweets, 77% received fewer than 100 impressions prior to removal.[14]
- In total, impressions on violative Tweets accounted for less than 0.1% of all impressions for all Tweets during that time period.[15]

We also stepped up the level of proactive enforcement across the service and invested in technological solutions to respond to ever-evolving malicious online activity. Today, by using technology, 65% of the abusive content we action is surfaced proactively for human review, instead of relying on reports from people using Twitter.[16]

Commitment to privacy

This also raises questions regarding rights to privacy, expression, and association. Anonymity and pseudonymity are critical to protect the privacy of the people who use online services. Restrictions on anonymity or pseudonymity online risk deeply undermining trust in public debate and conversation and putting vulnerable communities at heightened likelihood of targeted abuse and harm.

We have a range of ways for people to control their privacy experience on Twitter, from offering pseudonymous accounts to letting people control who sees their Tweets to providing a wide array of granular privacy controls. Our privacy efforts have enabled people around the world to use Twitter to protect their own data.

That same philosophy guides how we work to protect the data people share with Twitter. We empower the people who use our service to make informed decisions about the data they share with us. We believe individuals should know, and have meaningful control over, what data is being collected about them, how it is used, and when it is shared.

There is also a deeper question regarding rights to privacy with respect to privacy law in Australia. It should be noted that Australian Privacy Principle 2 requires that individuals must have the option of not identifying themselves, or using a pseudonym, when interacting with entities regulated by the *Privacy Act 1988*.

[14] https://transparency.twitter.com/en/reports/rules-enforcement.html#2020-jul-dec
[15] https://blog.twitter.com/en_us/topics/company/2021/an-update-to-the-twitter-transparency-center
[16] *Ibid.*

We recommend that this requirement clarify that companies taking action against accounts that contain harmful content (of the kinds regulated under the OSA) should apply to all such accounts, regardless of whether they are anonymous or pseudonymous, and these restrictions remain proportionate to the harm and conduct in accordance with company published policies. Anonymous and pseudonymous speech has been fundamental in the development of democratic societies, and we must continue to protect the ability for people to seek, receive, and impart information or ideas through these channels.

Expectation 10: provider will consult and cooperate with other service providers to promote safe use

The expectation that such a vast number of regulated companies with such diverse business models must consult and cooperate with other service providers is concerning and impractical. For example, the notion that providers could share volumetric data across platforms with relative ease (Additional Expectation 10(2)(a)) is fallacious. This could require organisations to disclose commercially sensitive, private, confidential, or proprietary information. Also it also begs the question – if someone potentially violates the rules on one platform, should private companies then decide that individuals should lose their ability to interact on other online forums across the internet without oversight, meaningful avenues of redress, or established legal avenues to challenge such action.

There are a number of existing multi-stakeholder organisations, forums, and initiatives in the technology space, and Twitter has a long history of involvement in a number of international initiatives to combat serious online threats. For example, we are members and signatories of many coalitions and organisations, including but not limited to, the Global Internet Forum to Counter Terrorism (**GIFCT**), the Aqaba Process, the Christchurch Call to Action, and the Australian Taskforce to Combat Terrorism and Extreme Violent Material Online. We are also invested in the Global Research Network on Terrorism and Technology (GRNTT) to develop research and policy recommendations designed to prevent terrorist exploitation of technology. As the GIFCT Operating Board chair for 2021, Twitter has endeavoured to support the work of the GIFCT as it deploys annual programs, training, and expands on its transparency efforts as an independent nonprofit.

We are also 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. Twitter is also actively involved in the development and consultation of the joint Australia-New Zealand government funded Organisation for Economic Co-operation and Development (OECD) Voluntary Transparency Reporting Protocols.

At an EU level, we have worked closely with the Radicalisation Awareness Network (RAN) and have been part of the Civil Society Empowerment Program, supporting organisations across the EU in countering violent extremism. In France, for example, we also continued our long-time work with NGOs, like the International League Against Racism and Anti-Semitism (LICRA) and Conseil Représentatif des Institutions juives de France (CRIF) through training, pro bono advertising grants, and donations.

Twitter also has a long-standing collaboration with the National Center for Missing and Exploited Children (NCMEC). We are active members of several coalitions, such as the Technology Coalition, the ICT Coalition, the WeProtect Global Alliance, INHOPE and the Fair Play Alliance, that bring companies and NGOs together to develop solutions that disrupt the exchange of child sexual abuse materials online and prevent the sexual exploitation of children.

We believe that wider efforts on promoting safe use of online services that are focused on bolstering the voices of non-governmental organisations and nonprofits would facilitate the desired consultation and cooperation in the private sector. Many of these nonprofit and non-governmental groups do critical work, and policy makers should continue to find ways to broaden support for these efforts and initiatives that promote best practices concerning the safe use of services, such as the Safety by Design initiative or the eSafety Online Safety Grant

program.[17]

We would recommend that this expectation be modified to bolster support for current cooperation and knowledge-sharing via various forums, similar to those mentioned above, and request that the expectation for detection of cross platform attacks be removed, as it poses major privacy concerns and would not be technically feasible across the vast products and services that are offered across the entirety of the Internet.

Expectation 12: access of children to class 2 materials must be consistent with other policy initiatives

Expectation 12 requires service providers to ensure that technological or other measures are in effect to prevent access by children to Class 2 material, which is online content that would be classified as unsuitable for minors under the Online Content Scheme in the OSA.

We have concerns that this expectation overlaps with other concurrent regulatory work-streams and duplicates legal thresholds that are set out in other areas of the OSA, as well as other pieces of pending legislation. Currently, there are four separate Federal Government initiatives under the OSA concerned with issues relating to the protection of children from online harms, including the Industry Codes, the BOSE, and the Restricted Access Systems (**RAS**), and the Age Verification (**AV**) roadmap, which deals broadly with pornography (which includes Class 2 material) and is also being driven by the Office of the eSafety Commissioner. Another initiative concerning online age verification is contained in the exposure draft of the *Privacy Legislation Amendment (Enhancing Online Privacy and Other Measures) Bill 2021* (**Online Privacy Bill**).[18] The draft RAS declaration and Online Privacy Bill have only recently been released on 25 October 2021, and the consultation on the AV roadmap is ongoing.

Earlier this year, the Federal Government also announced its Deregulation Agenda, which would take a whole-of-government approach to regulatory policy and focus on reducing barriers affecting Australia's productivity growth and competitiveness.[19] We believe that this expectation, as well as the other work-streams, are counterintuitive to the Government's priorities under this policy agenda, including regulator best practice and performance metrics, and would urge the Government to remove this expectation and relegate this work to the existing work-streams to ensure clarity and consistency.

Additionally, Expectation 12 (b) states that companies would need to conduct "child safety risk assessments," which as drafted is broad and unclear regarding what would be needed to comply with this expectation. As stated above, with the multiple overlapping workstreams, an ambiguous expectation to conduct child safety risk assessments will only cause further burden of uncertainty especially in a field where tremendous work is ongoing. Thus, we request that this expectation be removed.

Expectation 16: provider will make accessible information on how to complain to Commissioner

Expectation 16 requests that service providers ensure that information and guidance are made available to end-users on how to make a complaint to the eSafety Commissioner.

In the eSafety Commissioner's recent Position Paper, it states that "industry participants will handle reports and complaints about Class 1 and Class 2 material and codes compliance in the first instance. eSafety will act as a 'safety net' if resolution of a complaint is not satisfactory."[20] The OSA itself also stipulates a clear expectation that complaints first be made to the respective service providers.

It's important that users can easily identify the steps to take when reporting content to a service provider. Consequently, we recommend that this expectation be amended to make it clear that the expectation is not to make this information available at the same hierarchical level as information for Twitter's own reporting options,

[17] https://www.esafety.gov.au/about-us/newsroom/grants-support-young-people-navigate-online-world
[18] https://consultations.ag.gov.au/rights-and-protections/online-privacy-bill-exposure-draft
[19] https://deregulation.pmc.gov.au/
[20] eSafety Commissioner, Development of industry codes under the Online Safety Act Position Paper, p.5-6.

and that it ought to be made clear to end-users that complaints be directed to service providers in the first instance before reporting content to the eSafety Commissioner. This will help avoid any confusion that might result from users encountering multiple, separate reporting mechanisms when seeking information about lodging complaints with a service provider.

Expectation 19: provider will keep records regarding certain matters

Expectation 19 asks providers to keep records of reports and complaints about material for 5 years after the making of the report or complaint to which the record relates. However, in the absence of a preservation order, many providers do not keep account information or user communications after a certain amount of time when an account is suspended, a specific Tweet(s) is removed, or when the user deletes their accounts, irrespective of whether a complaint once may have been filed in relation to the user's communications.

This expectation would also run afoul of international privacy laws requiring the removal of account data after a specified amount of time. To hold data for 5 years would require substantial platform re-engineering and would involve significant costs, where it can be done at all. It would be useful to understand what exactly providers are expected to record and store and for what exact purpose, as well as what would constitute a reasonable standard inline with conflicting privacy law and international regimes.

Expectation 20: provider will provide requested information to the eSafety Commissioner

As drafted, Expectation 20 overlaps with the extensive powers the eSafety Commissioner already has to request reports under the OSA, which are enforceable by the Federal Court; thus, this expectation appears to create an unnecessary additional reporting requirement without a strong rationale.

As drafted, the expectation also requires responses within 30 days of the service provider receiving written notice of the request from the Commissioner. As a practical matter, depending on the extent and nature of the request, the time needed to gather the requested information could extend beyond 30 days, and we would ask for the time period to be designated as a "reasonable time" that is commensurate with the information requested.

Further, the expectation outlines that a provider must give the Commissioner a statement that sets out the number of complaints made to the provider during a specified period. As discussed previously, keeping records of user complaints is not a fair indication of the safety of an online service. Complaint numbers do not necessarily equate to safety levels, and we believe this metric is not a strong indicator of action taken to combat online harm when divorced from the holistic efforts that are taking place through proactive interventions and broader content moderation actions.

Additionally, the definition of "provider" needs to be updated to clarify that notices are: (1) received properly by the actual provider of the service, and (2) in a timely manner to ensure that receipt is confirmed before the clock starts ticking with respect to the notice period. Additionally, there is a lack of clarity under this expectation in regards to what would constitute a "removal notice," i.e. a removal ordered by the Commissioner or via a user report. The latter assumption must be removed if this is contingent on a user report, which is problematic if the content is lawful or outside the scope of the content governed by the OSA.

We would also recommend that the BOSE include provisions to protect confidentiality and commercially sensitive information. Under the *Consumer and Competition Act 2010* (Cth), confidentiality protections do not permit the Australian Consumer and Competition Commission (ACCC) to disclose confidential information to third parties, other than advisors or consultants engaged directly by the ACCC, without first providing a company with notice of its intention to do so, such as where it is compelled to do so by law.[21]

Given the breadth of the eSafety Commissioner's powers to compel information from companies under the OSA

[21] https://www.legislation.gov.au/Details/C2021C00010

and the BOSE, we would recommend that similar confidentiality protections are built into the reporting requirements to ensure protection of commercially sensitive information and guard against the creation of competition issues in the Australian digital market.

Expectation 21: provider will have designated contact person

Expectation 21 envisages that an individual who is an employee or agent of the provider be nominated as the contact person for the purposes of the OSA and be notified to the eSafety Commissioner.

We welcome clarification as to what is meant with 'for the purposes of the Act' in the context of 24-hour removal notices that may be received from the Commissioner. It needs to be clear that the nominated contact person will not be available 24/7 to receive removal notices and to act on those notices. These functions will be fulfilled by numerous personnel designated for that purpose and cannot be assigned to a single individual.

An alternative approach would be for the expectation to require a provider to have a single contact point (e.g. an email address or a contact number that leads to the appropriate department within the service provider and which can be staffed 24/7) instead of a single contact person. For example, Twitter has established dedicated reporting channels for the eSafety Commissioner to send reports through for review or to request information through our Legal Request Submissions Site. We would strongly recommend that such processes remain in place to ensure adequate coverage so that reports from the eSafety Office will be sent immediately and be reviewed by our global 24/7 specialised teams.

Additional questions regarding implementation timelines

The BOSE is set to commence immediately upon registration of the Determination. While the Determination in of itself does not impose a duty that is enforceable by proceedings in a court, the eSafety Commissioner has the power to "name and shame" providers that have, in the Commissioner's view, contravened one or more BOSE for the respective service that they supply.[22]

The OSA does not specify any timeframe that would be appropriate for the Commissioner to allow for implementation of the BOSE. Consequently, it remains at the Commissioner's discretion when she wishes to exercise her right to request, by written notice, periodic or non-periodic notices and to "name and shame" as a result of the information that has been provided to her – and it remains at her discretion to allow (or disallow as the case may be) a certain "grace period" for providers to implement what might be complex processes or technical measures required to fulfil the BOSE.

Given the extraordinary discretionary powers afforded to the Commissioner, the lack of clarity of what is required of providers in some respects and the far-reaching nature of other expectations, we request that the BOSE commence six months after registration, and thereby provide business and operational certainty to service providers in Australia.

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 while protecting vital public expression. Across all areas, the 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.[23]

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

---

[22] *Online Safety Act 2021* (Cth), s 45(2)-(4).
[23] Twitter Blog, 2021 [online] blog.twitter.com. Available at: <https://blog.twitter.com/en_us/topics/company/2019/health-update.html> [Accessed 7 November 2021].