Introduction

The draft Online Safety (Basic Online Safety Expectations) Determination 2021 provides unprecedented pressure and accountability to online providers to regulate their content. The expectations in their current form could result in discrimination, and physical/mental harm to activists, people of colour, marginalised communities, dissidents, and everyday people who use online platforms/providers to raise genuine concerns, share their voices. In the following sections, this submission will further examine the concerns and provides recommendations that will help protect privacy, safety and the right to freedom of expression of every Australian.

Furthermore, I would like to extend my support to the submission made by Digital Rights Watch on BOSE[1]. The expectation, additional expectations and reasonable steps are extracted directly from their submission [1].

Expectations, Concerns and Recommendation

Expectation: The provider will take reasonable steps regarding encrypted services.

Additional expectations: If the service uses encryption, the provider of the service will take reasonable steps to develop and implement processes to detect and address material or activity on the service that is or may be unlawful or harmful

Concerns

This expectation undermines the psychological safety of Australians, who primarily rely on encryption to protect the confidentiality of their data/information managed by the providers. Expectations and additional expectations push the providers to intercept, circumvent, or weaken the encryption mechanisms that could undermine Australians' confidentiality and become a national security threat. I am using the following examples to support my concerns.

Examples

1) In November 2019, The US Justice Department charged two former Twitter employees for spying for the Saudi government. Twitter is an online media platform that does not offer End to End encryption [2].

2) In July 2020, The accounts of Former US president Barack Obama, Elon Musk and Jeff Bezz accounts were hacked, which has now pushed Twitter to adapt End to end encryption on its platforms [3].

As the examples presented above, a platform that does not offer End to end encryption has been targeted by a nation-state and other malicious actors. If these expectations were to be met, encryption might be weakened or circumvented by the providers, resulting in a severe threat to the confidentiality of Australians.

Recommendation

Additional expectations should explicitly state that providers should not weaken, circumvent encryption threatening Australian's confidentiality, privacy rights and Australia's national security.

Additional expectation: Provider will take reasonable steps regarding anonymous accounts

● **Additional expectation** - (1) If the service permits the use of anonymous accounts, the provider of the service will take reasonable steps to prevent those accounts being used to deal with material, or for activity, that is or may be unlawful or harmful.
● **Reasonable steps that could be taken** - (2) Without limiting subsection (1), reasonable steps for the purposes of that subsection could include the following:
(a) having processes that prevent the same person from repeatedly using anonymous accounts to post material or to engage in activity that is unlawful or harmful.
(b) having processes that require verification of identity or ownership of accounts.

Concerns

As the famous batman movie quote, "The mask is not for you, it's to protect the people you care about" [4]. Anonymity is the mask for the people who want to raise their voices against human rights abuse, corruption, oppression and poverty without fear and retaliation. The expectations, additional expectations, reasonable steps described in BOSE will threaten individuals who fight against those highlighted issues.

To further support my concerns, the following examples from history, which supports and highlights the importance of anonymity.

Examples

1) As Ian Bell describes Jamal Khashoggi, "the Saudi Arabian journalist who fell foul of his country's ruling dynasty after moving abroad so he could criticise it more freely" [5]. Jamal Khashoggi openly criticised and voiced his opinions on Twitter against the malpractices of the Saudi Government and crown prince Mohammed bin Salman. However, this led to his demise. Khashoggi was 59 when targeted and killed by a group of Saudi Operatives after entering the Saudi consulate in Istanbul on 2 October 2018 [5].

2) The Watergate scandal brought to light abuse of power by former US President Richard Nixon, which resulted in his impeachment. It contributed to significant reforms on campaign finance, government ethics, intelligence oversight and the president's war powers.
The individual behind the watergate scandal was only known by his Psudoname until 2005. This event changed the course of US politics and ensured the further accountability of elected officials in power [6][7].

These examples describe the contrasting results of non-anonymity and anonymity. One resulted in being hunted down and killed. The other resulted in sweeping reforms (strengthening democracy) and protected the individual who was identified in 2005 as former Senior FBI Agent Mark Felt, who died of natural causes at the age of 95 [8].

The examples above highlight why people opt for anonymity on online platforms/providers to avoid retaliation for expressing their genuine opinions or blowing the whistle on corruption and abuse.

Recommendations

- Should remove the expectations, additional expectations, and reasonable steps as they weaken and threaten anonymity, resulting in suppression of genuine freedom of speech and may instil fear against whistleblowing.
- Should advise providers to strengthen their community guidelines and be more proactive in actioning abuse reports.

Expectations - Provider Will Take Reasonable Steps to Ensure Safe Use

Additional expectation - (2) The provider of the service will take reasonable steps to proactively minimise the extent to which material or activity on the service is or may be unlawful or harmful.

Reasonable Steps

1. developing and implementing processes to detect, moderate, report and remove (as applicable) material or activity on the service that is or may be unlawful or harmful;

Concerns

The amount of content digested by providers cannot be verified manually by humans, driving the providers to use automated processes powered by artificial intelligence. When providers implement automated processes, it is very likely to implement Artificial intelligent models that are highly biased, which could potentially lead to discrimination [9][10].

Without proper safeguards, reasonable step a) could result in catastrophic failure leading to incorrectly flagging legitimate material/activity. In addition, providers have been involved in unethical practices to obtain/retain training data that violates the privacy of the users [11][12].

Recommendation

- The proposal should clearly state that providers process should be non-discriminatory and should not violate freedom of expression and privacy laws.
- Any Artificial intelligence model/tool that may be used to assess a material/activity unlawful or harmful by the providers should be open-sourced. The training dataset and results of the efficiency of that tool/model should be made public.
- Providers should allow the users to appeal for incorrect removal of material or activity, and the appeal process should have complete transparency.
- Should recommend the providers not solely rely on the automated process, significant human oversight must be present when identifying unlawful or harmful material.

Bibliography

1. Consultation on a draft Online Safety (Basic Online Safety Expectations) Determination 2021 https://digitalrightswatch.org.au/wp-content/uploads/2021/11/Global-Partners-Digital-Digital-Rights-Watch-Joint-Submission.pdf
2. Former Twitter employees charged with spying for Saudi Arabia by digging into the accounts of kingdom critics https://www.washingtonpost.com/national-security/former-twitter-employees-charged-with-spying-for-saudi-arabia-by-digging-into-the-accounts-of-kingdom-critics/2019/11/06/2e9593da-00a0-11ea-8bab-0fc209e065a8_story.html
3. After This Week's Hack, It Is Past Time for Twitter to End-to-End Encrypt Direct Messages https://www.eff.org/deeplinks/2020/07/after-weeks-hack-it-past-time-twitter-end-end-encrypt-direct-messages.
4. Batman quote perfectly sums up why we should all be wearing a face-covering https://www.mirror.co.uk/film/batman-quote-perfectly-sums-up-22354301.
5. Jamal Khashoggi obituary https://www.theguardian.com/world/2018/oct/19/jamal-khashoggi-obituary
6. Watergate led to sweeping reforms. Here's what we'll need after Trump. https://www.washingtonpost.com/outlook/2019/11/15/watergate-led-sweeping-reforms-heres-what-well-need-after-trump/
7. Watergate scandal https://en.wikipedia.org/wiki/Watergate_scandal
8. Watergate's Deep Throat, Mark Felt, dies https://www.theguardian.com/world/2008/dec/19/watergate-deep-throat-dies
9. Research shows AI is often biased. Here is how to make algorithms work for all of us https://www.weforum.org/agenda/2021/07/ai-machine-learning-bias-discrimination/
10. San Francisco Bans Facial Recognition Technology https://www.nytimes.com/2019/05/14/us/facial-recognition-ban-san-francisco.html
11. How Photos of Your Kids Are Powering Surveillance Technology https://www.nytimes.com/interactive/2019/10/11/technology/flickr-facial-recognition.html
12. This face recognition company is causing havoc in Russia—and could come to the U.S. soonhttps://splinternews.com/this-face-recognition-company-is-causing-havoc-in-russi-1793856482
13. Saudis' Image Makers: A Troll Army and a Twitter Insider https://www.nytimes.com/2018/10/20/us/politics/saudi-image-campaign-twitter.html