Submission to the Department of Infrastructure, Transport, Regional Development and Communication

on the proposed

Online Safety Bill 2020

14 February 2021

DIGITAL RIGHTS WATCH

Overview

We welcome the opportunity to submit comments to the Department of Infrastructure, Transport, Regional Development and Communications concerning the Online Safety Bill 2020. We commend the overarching objectives of the Bill in creating safe and accountable online spaces, and protecting children and our communities from harm caused by malicious online activity.

Digital Rights Watch regularly engages in consultation with the local and federal government on issues relating to human rights in the digital era. We remain concerned about the lack of a federal level protection for rights and freedoms guaranteed by the Universal Declaration of Human Rights, particularly—as it relates to this draft Bill—the right to privacy, the right to freedom of opinion and expression, the right to work, and the right to education.

Some of our previous work relevant to the topics covered in this submission:

- UN Human Rights Council Australia Universal Periodic Review https://digitalrightswatch.org.au/2020/08/28/access-now-and-digital-rights-watch-joint-submission-to-the-un-human-rights-council/
- UN Inquiry into the Right to Privacy in a Digital Age https://digitalrightswatch.org.au/2018/04/10/submission-to-un-inquiry-into-the-right-to-privacy-in-a-digital-age/
- UN Inquiry into Freedom of Expression in Telcos and the Internet (includes website blocking) https://digitalrightswatch.org.au/2016/10/30/un-inquiry-into-freedom-of-expression-in-telcos-and-the-internet/

Digital Rights Watch

Digital Rights Watch is a charity organisation founded in 2016 whose mission is to ensure that people in Australia are equipped, empowered and enabled to uphold their digital rights. We stand for Privacy, Democracy, Fairness & Freedom in a digital age. We believe that digital rights are human rights which see their expression online. We educate, campaign, and advocate for a digital environment where individuals have the power to maintain their human rights.[1]

[1]Learn more about our work on our website: https://digitalrightswatch.org.au/

General remarks

At Digital Rights Watch, we welcomed the objectives of the draft Online Safety Bill to "improve and promote Australia's online safety."[2] Several of the powers proposed in the Bill create critical pathways of redress for children and adults suffering online bullying, abuse, and non-consensual sharing of intimate images. These powers are important as online issues can translate to significant real-life harms. While we believe there is some room for improvement in these areas, especially in creating reporting and redress mechanisms, we are extremely concerned that some of the powers in the Bill will undermine digital rights and exacerbate harm for vulnerable groups.

We have broken down our submission into 5 sections according to our key areas of concern, followed by a recommendations section where we outline the main changes this Bill should undergo to protect human rights online and safeguard freedom of expression.

1. Online Content Scheme

The Online Safety Bill relies heavily on the National Classification Code to determine which type of content may be issued with a removal notice by the eSafety Commissioner.[3] However, the classification system in Australia has been criticised for being outdated and overly broad. Further using it as the basis of broad and discretionary powers entrusted to an administrative official with no appropriate accountability mechanisms and no judicial oversight is a concerning development.

Class 1 aligns with content that would be deemed "Refused Classification" (RC). This includes content that deals with sex or "revolting or abhorrent phenomena" in a way that offends against the standards of "morality, decency and propriety generally accepted by resonable adults." Class 2 material includes content that is likely to be classified as X18+ or R18+. This includes non-violent sexual activity, or anything that is "unsuitable for a minor to see."

The adult cyber-abuse scheme in particular could be broadly overinterpreted and used to suppress and silence speech. **What constitutes content that is "offensive or malicious" (Section 7 and 8 of the draft Bill) is extremely subjective and could capture broad categories of protected speech, such as political expression**. While it can be offensive and often seen as malicious, satire or comedy form a vital part of the public discussion and allow the greater public to process key events, allowing us to "comfort the afflicted and afflict the comfortable."[4] It is our recommendation that the adult cyber-abuse scheme is removed from the draft Bill entirely because of its potential for overreach and existing legal avenues for adults to challenge defamatory, harmful or illegal content. If the scheme remains a part of

[2] https://www.communications.gov.au/have-your-say/consultation-bill-new-online-safety-act
[3] https://www.legislation.gov.au/Details/F2013C00006
[4] A Point of View: What's the Point of Satire? https://www.bbc.com/news/magazine-31442441

the Bill, we suggest that broad exemptions are created to protect political speech and safeguard freedom of expression.

Furthermore, taken together, Class 1 and 2 material captures any and all sexual content (violent in nature or not). The way this removal scheme is drafted in the Bill, it is likely to cause significant harm to those who work in the sex industry, including sex workers, pornography creators, online sex-positive educators, and activists. Recently, the pandemic forced many people to work online or at home, including sex workers, who would otherwise be providing sexual services lawfully in their places of work. This scheme risks undermining the livelihood and ultimately the safety and wellbeing of sex workers by putting their work at risk of sanction.[5] Moreover, we have already seen as a result of the controversial Stop Enabling Sex Traffickers Act (SESTA) and Allow States and Victims to Fight Online Sex Trafficking Act (FOSTA) legislation in the US that when sex workers are forced offline they are often pushed into unsafe working environments, in turn, creating direct harm.[6]

We are further concerned that this complaints mechanism could be abused by those seeking retribution or seeking to cause (material and mental) harm to sex workers or sex-positive educators by filing repeated complaints to the Commissioner. **The final text of the Bill should prohibit the abuse of the complaints mechanism and create penalties for those who do.** Additionally, the discretion given to the Commissioner under Division 5 Section 42 to investigate and search for Class 1 and 2 material online without the filing of a complaint seems disproportionate to the goal of creating and promoting online safety. It also places the Commissioner's office in the position of proactively determining what any Australian "reasonable adult" would or wouldn't not consider offensive. **We recommend that the Commissioner be limited to only acting on complaints with regard to Class 1 or Class 2 material.**

The scheme also does not contain an adequate appeals mechanism for individuals and companies who receive removal notices. While Section 220 of the Bill does provide a method for people to challenge decisions through the Administrative Appeals Tribunal (AAT), there should be additional opportunities for people to challenge take down notices, without having to go through the court system. By the time someone goes through the process with the AAT, the harm (and potential loss of income) associated with the removal has already occured. **The Commissioner must be able to receive appeals within a 24 hour window, and provide an effective remedy, including the ability to reinstate content.**

2. The Abhorrent Violent Material Blocking Scheme

Part 8 of the Bill gives the eSafety Commissioner the power to issue a blocking request or notice to Internet Service Providers (ISPs) to block domain names, URLs or IP addresses that provide access to such material. The Commissioner does not need to observe any requirements of procedural fairness for these requests. Under Section 100 of the Bill,

[5] https://www.theguardian.com/technology/2020/dec/23/everyone-and-their-mum-is-on-it-onlyfans-boomed-in-popularity-during-the-pandemic
[6] https://hackinghustling.org/erased-the-impact-of-fosta-sesta-2020/

blocking notices cannot be for longer than 3 months, however, there are no limitations to how many times the Commissioner can renew such a blocking notice.

While there is no doubt that we need mechanisms to deal with viral violent videos and content online and the harm they cause (and indeed already some exist internationally under the Global Internet Forum to Counter Terrorism), the proposed scheme is overly simplistic and overlooks complex underlying issues.[7]

There are some limits to this power under Section 104 of the Bill which includes some exempt material for conducting scientific, medical, academic or historical research, or relating to news reporting that is in the public interest. While we welcome these limitations, there remains a wide scope of discretion for the eSafety Commissioner to determine what is indeed in the public interest. **We recommend that decisions over website or content blocking in this category, or determinations that involve consideration of what constitutes the public interest remain with the judiciary and not at the discretion of the eSafety Commissioner.**

In some circumstances, violent acts captured and shared online can be of vital importance to hold those in power accountable, to shine the light on otherwise hidden human rights violations, and be the catalyst for social change. The virality of the video of the murder of George Floyd by a police officer in the US played a key role for the Black Lives Matter movement in 2020. Closer to home, a viral video of a NSW Police officer using excessive force against an Indigenous teenager prompted important discussions about racism in Australian law enforcement.[8] It is critical that such content remain archived for the public interest.

Furthermore, simply blocking people from seeing violent material does not solve the underlying issues causing the violence in the first place and it does not create justice or avenues of redress. It is essential that this scheme not be used to hide state use of violence and/or abuses of human rights.

We are also concerned that there are no safeguards or limitations in place under Section 100, with regard to the renewal of blocking notices. As documented by our friends at Access Now, internet blocking is a serious human rights issue that has been abused as a mechanism to suppress and limit dissent and democratic debate around the world.[9] We must tread very carefully when entering into this domain, to ensure that sites are only blocked in very limited circumstances, and never in a way that infringes upon the rights and freedoms guaranteed by international law.

[7] For more on how the GIFCT moderates content online and prevents the virality of content which incites or promotes violence, see reporting by Slate: https://slate.com/technology/2020/08/gifct-content-moderation-free-speech-online.html

[8] https://www.abc.net.au/news/2020-06-02/nsw-police-investigate-officer-over-arrest-of-indigenous-teen/12310758

[9] See Access Now #keepiton campaign for more: https://www.accessnow.org/keepiton/#problem

3. Basic Online Safety Expectations

Part 4 of the Bill gives the Minister power to determine basic online safety expectations (BOSE) for 'social media services', 'relevant electronic services', and 'designated internet services.'

Section 46 of the Bill requires the expectations to specify that the service should:

- Minimise cyber-bullying or abuse material targeted at a child or adult, non-consensual intimate images, Class 1 material, and abhorrent violent material,
- Take reasonable steps to prevent children from accessing class 2 material,
- Provide ways for people to make complaints about online content.

When drafted so broadly, these expectations incentivise proactive monitoring and removal of content that falls under Class 1 and 2. Given the scale of online content, digital platforms generally turn to automated processes (such as AI) to determine which content is or is not harmful, despite evidence that content moderation algorithms are not consistent in identifying content correctly.[10] This kind of content moderation has been shown to disproportionately remove some content over others, penalising Black, Indigneous, fat, and LGBTQ+ people.[11] As experience with the controversial SESTA/FOSTA in the US demonstrated, some platforms will default to blanket removal of all sexual content to avoid penalty rather than deal with the harder task of determining which content is actually harmful.

Automated processes have also not proven to be as effective for hate speech, making it more likely to be a visual-based scheme, and less effective at identifying specific forms of content like cyberbullying or abuse material. In 2018, Zuckerberg said it's "easier to detect a nipple than hate speech with AI."[12] **We recommend that any legislative regime avoid incentivising the use of automated solutions to identify and remove online content, regardless of content category.** If automated decision making is used for content moderation to comply with the provisions in this Bill, the Commissioner should take an active role in ensuring that these processes use open source tools, transparent standards, regular independent oversight and appropriate appeals mechanisms for cases of false positives.

The requirement under Section 46(d) of the Bill to take 'reasonable steps' to prevent children from accessing Class 2 content also raises concerns around the potential technological "solutions" that may come as a result. Not long ago the Department of Home Affairs

---

[10] https://www.eff.org/deeplinks/2020/10/facebooks-most-recent-transparency-report-demonstrates-pitfalls-automated-content

[11] The algorithms that detect hate speech online are biased against black people: https://www.vox.com/recode/2019/8/15/20806384/social-media-hate-speech-bias-black-african-american-facebook-twitter
Facebook repeatedly bans Indigenous activists: https://onlinecensorship.org/content/infographics
Instagram photo censorship: https://www.theguardian.com/technology/2020/oct/20/instagram-censored-one-of-these-photos-but-not-the-other-we-must-ask-why

[12] https://venturebeat.com/2018/04/25/zuckerberg-its-easier-to-detect-a-nipple-than-hate-speech-with-ai/

suggested the use of facial recognition technology for age verification to access porn sites.[13] This would create significant privacy and data protection issues for children and adults alike.

We are also concerned that BOSE, as currently defined, could be used to limit, restrict, or undermine encrypted services and communications. There are no provisions in the draft Bill to prevent this scope creep. As drafted, the BOSE could be made to compromise secure tools and technologies regardless of their overall merit if they somehow impede or prevent investigations by digital platforms into the content defined in Section 46. **We recommend that the Bill is amended to affirm the need for strong encryption and prohibit any interference of the powers prescribed with encrypted tools and technologies.**

4. Information Gathering Powers, Investigative Powers, and Encryption

Part 13 of the draft Bill provides that the Commissioner may obtain information about the identity of an end-user of a 'social media service', a 'relevant electronic service', or 'designated internet service.' Part 14 also provides the Commissioner with investigative powers, which includes a requirement that a person provides "any documents in the possession of the person that may contain information relevant."

Given that 'relevant electronic service' includes email, instant messaging, SMS and chat, without mentioning end-to-end encrypted messaging services, it is possible that the Commissioner's information gathering and investigative powers would extend to encrypted services. **The Bill needs additional clarification of the scope of these powers, and clear indication in Section 194 of the Bill that a provider is not expected to comply with a notice if it would require them to decrypt private communications channels or build systemic weaknesses to comply with the provisions of this Bill.**

The eSafety Commissioner has already publicly argued against end-to-end encryption, saying that it "will make investigations into online child sexual abuse more difficult."[14] While encryption may impede such investigations, it also provides everyone with digital security, and protects everyone from arbitrary surveillance by malicious actors and cybercrime (ie. identity theft). Further, it protects the privacy of victims of domestic violence, confidential sources of journalists, safety of political dissidents and all activists, lawyers, and reporters. Claiming that encryption exacerbates harm to children is unproven, and strengthens a regressive surveillance agenda at the expense of everyone's digital security. It is essential that compliance with this Bill does not create a way to compel providers to restrict or weaken their use and application of encryption across their platforms.

[13] https://www.abc.net.au/news/science/2020-03-05/age-verification-filter-for-online-porn-recommended-in-australia/12028870

[14] https://www.esafety.gov.au/about-us/blog/end-end-encryption-challenging-quest-for-balance

5. Additional Comments

The draft Bill prompts overarching questions regarding how much power and discretion should be entrusted to an administrative government official. Appointing the Commissioner as the arbiter of appropriate vs “offensive” content is an outdated and dangerous way to treat online content, just as it would be inappropriate for an official to search a library seeking out and censoring certain content or arbitrarily prohibiting people from accessing certain books.

Given the discretionary nature of many of these powers, the Commissioner should be subject to robust transparency reporting and a regular review of how the powers are used in practice. **We strongly recommend the creation of a multi-stakeholder oversight board for activity covered by the Bill.** While it might be appropriate to have an annual Parliamentary review and oversight included in the Bill, we believe that given the detrimental impact on specific communities, it would be appropriate to create community oversight for these powers.

While the goal of minimizing online harm for children is vital to our communities, we must acknowledge that policing the internet in such broad strokes will not guarantee us safety and potentially suppress protected speech and create extended damage to our rights and freedoms online.

Recommendations

- **Include a sunset clause.** Given the level of discretion which is given to the eSafety Commissioner under the Bill, there needs to be an opportunity to review whether these powers are working well, and decide if the legislation should be renewed or revisited. A sunset clause ensures such a process takes place at a given time.
- **Remove the adult cyber-abuse content scheme.** Due to an overlap with existing legal mechanisms for adults to seek a remedy in cases of defamatory, threatening or illegal content, this scheme only further removes accountability and creates a system ripe for abuse and suppression of freedom of expression online.
- **Establish a multi-stakeholder review board for activity covered by the Bill.** There is an international consensus that content moderation and take-downs require robust oversight and accountability to prevent abuse of power. The review board should be included in the Bill as a mechanism to review decisions made to remove and block content by the Commissioner. The Board should be made up of the groups most impacted by the proposed laws, including sex workers and activist, and meet regularly, at least annually, to closely examine how decisions are being made by the Commissioner’s office across a spectrum of complaints and investigations.
- **Require transparency reporting on complaints and take-downs.** There should be quarterly, or at least annual, reporting of across all the powers prescribed to the Commissioner by the Bill. This includes the categories of content take-downs, complaints received (vs actioned and escalated), and blocking notices issued,

including the reasoning. This will allow for public and Parliamentary scrutiny over the ultimate scope and impact of the Bill.

- **Articulate a meaningful and timely appeals process.** Individuals must retain their rights under the Bill which should include the ability to challenge removal notices in a timely manner, without having to seek an external judicial process to bring accountability to the Commissioner. Especially in cases where removal may directly impact income and livelihood, affected individuals should be able to seek remedy from the eSafety Commissioner's office if the removal is unjustified or arbitrary, including monetary damages as appropriate.
- **Include an explicit assurance that ISPs and/or digital platforms will not be expected to weaken or undermine encryption in any way to comply with any parts of this Bill.** Similar provisions prohibiting the requirement for the introduction of a systemic weakness exist in the Telecommunications and Other Amendments Bill (TOLA) the Assistance and Access Act.

Contact

Lucie Krahulcova | Programme Director | Digital Rights Watch | [redacted]