Submission to the Department of Infrastructure, Transport, Regional Development, Communications and the Arts

regarding the Online Safety (Basic Online Safety Expectations) Amendment Determination 2023

23 February 2024

DIGITAL RIGHTS WATCH

Who we are

Digital Rights Watch is a charity organisation founded in 2016 to promote and defend human rights as realised in the digital age. We stand for privacy, democracy, fairness and freedom. Digital Rights Watch educates, campaigns and advocates for a digital environment in which rights are respected, and connection and creativity can flourish. More information about our work is available on our website: www.digitalrightswatch.org.au

Acknowledgement of Country

Digital Rights Watch acknowledges the Traditional Owners of Country throughout Australia and their continuing connection to land and community. We acknowledge the Aboriginal and Torres Strait Islander peoples as the true custodians of this land that was never ceded and pay our respects to their cultures, and to elders past and present.

Contact

Samantha Floreani | Head of Policy | samantha@digitalrightswatch.org.au

General remarks

Digital Rights Watch welcomes the opportunity to provide feedback to the Department of Infrastructure, Transport, Regional Development, Communications and the Arts regarding the proposed *Online Safety (Basic Online Safety Expectations) Amendment Determination 2023* (BOSE).

As Australia's leading digital rights advocacy organisation, we are primarily concerned with ensuring an appropriate balance is struck with regard to the impact upon individuals' rights (including children's rights) and any adverse impacts upon privacy, digital security, and online safety of individuals and communities..

As always, we emphasise that privacy and digital security are essential to uphold safety. Questions of legitimacy, proportionality, and reasonableness also must be carefully considered in any rights-balancing activity when determining online safety policy interventions. Digital Rights Watch is contributing to this consultation in the spirit of seeking to ensure that Australia's approach to online safety does not end up disproportionately undermining safety in the quest to enhance it.

Over the years, Digital Rights Watch has actively participated in Australia's online safety policy space. Since the inception of the *Online Safety Act 2021* (OSA), we have consistently engaged with the Office of the eSafety Commissioner and other relevant government bodies and industry groups to provide a human rights-focused perspective to the consultation and policy development process.

Digital Rights Watch has made a range of submissions that are relevant to the BOSE, including:

- Submission to the eSafety Commissioner in response to the draft Designated Internet Services Standard and the draft Relevant Electronic Services Standard for class 1A and 1B material, December 2023
- Submission to the steering group of industry associations on the draft Consolidated Industry Codes of Practice for the Online Industry, Phase 1 (Class 1A and Class 1B material), October 2022.
- Submission to the Select Committee on Social Media and Online Safety for the Inquiry into Social Media and Online Safety, January 2022
- Submission to the Office of the eSafety Commissioner on the draft *Online Safety (Basic Online Safety Expectations) Determination 2021,* November 2021.
- Submission to the eSafety Commissioner on the draft Restricted Access Systems Declaration, November 2021
- Submission to the Department of Infrastructure, Transport, Regional Development and Communication on the proposed *Online Safety Bill 2020,* February 2021.

Much of our position in relation to the BOSE has not changed since our joint submission with Global Partners Digital in November 2021, in response to the original set of BOSE.[1]

A high-level summary of our original observations, which remain relevant, are as follows:

- automated processes to proactively monitor and remove content have been proven to result in the removal of lawful and legitimate content online and, as such, care should be taken when incentivising the use of such technologies and systems,
- encryption is essential to online safety and should be treated as such,
- online anonymity is essential to the safety, rights, and wellbeing of many individuals and must be protected.

Our additional observations and feedback specific to the proposed amendment to the BOSE are included in the following pages. We would welcome the opportunity to discuss this submission with the Department further.

Online safety reform and review timeline and conflicts

We note that the Statutory Review into the OSA is due to be imminently conducted, with the terms of reference released on 13 February 2024.[2] It is expected that this wide-ranging review will require extensive consultation over the coming months, with the final report due to the Government in the second half of 2024. We are frustrated that considerable amounts of time, effort, and resources have already been spent by both government and external stakeholders on amending the current BOSE, only for it to be subject to another review within the year.

We note that the Classification Code is also due for imminent review, in addition to the upcoming development of industry codes for Class 2 Material. We do not see how it is reasonable or sensible to move ahead with creating another set of industry codes, and indeed to review the Online Safety Act itself, without first dealing with the classification framework upon which the entire scheme relies. While we appreciate the urgency to improve Australia's online safety regime to meet evolving needs, threats, and emerging technologies, we also note that such an approach risks creating conflicts and inconsistencies, and in some cases may

[1] 'Consultation on a draft Online Safety (Basic Online Safety Expectations) Determination 2021,' *Digital Rights Watch and Global Partners Digital,* November 2021. https://www.infrastructure.gov.au/sites/default/files/documents/bose-global-partners-digital-and-digital-rights-watch-joint-submission.pdf

[2] 'Media Release: Ensuring our online safety laws keep Australians safe,' *The Hon Michelle Rowland MP, Minister for Communications,* 13 February 2024 https://minister.infrastructure.gov.au/rowland/media-release/ensuring-our-online-safety-laws-keep-australians-safe

ultimately lead to policy that makes people *less safe online* in attempts to improve online safety.

While we strongly advocate for robust public consultation and welcome the opportunity to participate to ensure that Australia's online safety scheme is rights-respecting and fit for purpose, the concurrent, overlapping, and at times conflicting pieces of policy makes providing meaningful input challenging. The way consultations are timed and the order in which they occur can result in placing an enormous burden upon civil society organisations, community and advocacy groups, and small or independent technology companies to be able to meaningfully engage throughout this process. Many of these groups freely provide expertise, experience, and essential community perspectives with little to no resourcing to do so and, in some cases, on a voluntary basis.

Digital Rights Watch strongly believes that civil society and community input into such essential tech policy—which ultimately stands to impact all Australians who participate in online life—is essential. Such participation helps to improve the quality of policy, can illuminate gaps, oversights or potential consequences that are otherwise overlooked, and plays an important role in developing public trust and the social contract necessary to the success of the scheme. We trust that the Department shares this belief and as such, we urge the Department and the eSafety Commissioner to be more mindful of the way that these consultations are timed, and the order in which they are conducted, to better respect the time, expertise, and contributions of community and civil society voices. We suggest a timeline and outline of how the reviews relate to each other be released to ensure transparency and accountability.

Generative AI and recommender systems

We generally support the inclusion of the additional expectations 8A and 8B that focus on generative AI and recommender systems.

The sudden influx of freely available and widely accessible generative AI tools has enabled their widespread and sometimes dangerous use at a scale and scope previously unseen. For example, AI generated deepfake "pornography"—or, more accurately, non-consensual sexual imagery—is running rampant.[3] Investigative reporting has shown recently that explicit imagery is being generated of Australian women without their consent.[4] AI generated alteration of women's bodies is not limited to shadowy forums or even limited to highly explicit material. The digitally altered image of Georgie Purcell MP earlier this year demonstrates how the use of AI tools can easily (and perhaps unwittingly to users) regurgitate

[3] 'A deepfake porn scandal has rocked the streaming community. Is Australian law on top of the issue?,' *SBS News*, 9 Februrary 2023 https://www.sbs.com.au/news/the-feed/article/a-streamer-was-caught-looking-at-ai-generated-porn-of-female-streamers-the-story-just-scratches-the-surface/vfb2936ml

[4] 'People are training AI on photos of Australian women to make explicit images without their consent,' *Crikey*, 11 January 2024. https://www.crikey.com.au/2024/01/11/ai-photos-australian-women-explicit-images-consent-civitai/

pre-existing sexist assumptions, in turn entrenching such perspectives in media coverage while also stripping the subject of digital bodily autonomy.[5]

We also share concerns regarding the dissemination of problematic material via recommender systems, especially when in combination with engagement algorithms and personalisation, which prioritise provocative, inflammatory, or otherwise emotive content in order to elicit the maximum amount of engagement. As an example, TikTok is particularly effective at using the amount of time that users linger on content to take them down algorithmic 'rabbit holes', progressively showing that person more and more extreme content.[6]

While we welcome the expectation placed on services to consider safety of users with regard to their recommender systems, as always, we emphasise that in order to deal effectively with the harms of recommender systems, attention must be given to the underlying business model of services and platforms that employ these tools. Critically, the data-extractive and privacy-invasive practices of these companies *enable* the ability to recommend content to users that targets their vulnerabilities.

Research in partnership between the University of Queensland, Monash University and VicHealth found that ads served to 16-25 year olds on Instagram and Facebook are dominated by highly targeted marketing for alcohol, gambling, and unhealthy food.[7] Data brokers routinely sell data related to potential vulnerabilities, for example, habitual gamblers are targeted online with advertising for gambling products.[8] Once users interact with content about body image, mental health, depression or self harm on TikTok—a platform centralised upon its recommender system—they are presented with more and more of the same content, with one investigation suggesting TikTok would show them problematic content every 39 seconds.[9]

The real power of recommender systems stems from the ability to leverage detailed user profiles based on data-extractive business models. Similarly, AI tools are (generally) built and trained upon huge amounts of personal information.

---

[5] 'An AI-generated image of a Victorian MP raises wider questions on digital ethics,' *ABC News,* 1 February 2024.
https://www.abc.net.au/news/2024-02-01/georgie-purcell-ai-image-nine-news-apology-digital-ethics/103408440

[6] 'Investigation: How TikTok's Algorithm Figures Out Your Deepest Desires,' *Wall Street Journal,* 21 July 2021.
https://www.wsj.com/video/series/inside-tiktoks-highly-secretive-algorithm/investigation-how-tiktok-algorithm-figures-out-your-deepest-desires/6C0C2040-FF25-4827-8528-2BD6612E3796

[7] 'Harmful industries' digital marketing to Australian children,' *VicHealth,* 25 November 2015.
https://www.vichealth.vic.gov.au/news-publications/research-publications/harmful-industries-digital-marketing-australian-children

[8] ''Heavy TAB gamblers' among groups targeted by online advertising database,' *The Guardian,* 15 August 2023.
https://www.theguardian.com/australia-news/2023/aug/15/tab-gamblers-betting-australia-targeted-microsoft-xandr-advertising-database

[9] 'Young TikTok users quickly encounter problematic posts, researchers say,' *New York Times,* 14 December 2022.
https://www.nytimes.com/2022/12/14/business/tiktok-safety-teens-eating-disorders-self-harm.html

Addressing these issues by simply removing or labelling the content overlooks the underlying data economy that enables such technologies, and as such will never address the underlying imperatives that result in the scale of harm caused by recommender systems and generative AI.

Ongoing complexities regarding "harmful" but not illegal content

We do note, however, that the inclusion of "harmful" in addition to "unlawful" in sections 8A(1) and 8B(2) as well as throughout the BOSE may create risks of over-capture of content.

As a subjective determinant, "harmful" could be taken to mean a broad category of material. For example, some people consider all pornographic or sexually explicit material to be "harmful", while others disagree. Some may consider content that depicts violence or drug use to be "harmful", while others disagree.

It is feasible to expect that s 8B(3)(c), which calls for ensuring that end-users are able to make complaints or enquiries about the presentation of material via recommender systems on a service that is "unlawful or harmful", could result in complaints being targeted towards groups that are politically or ideologically perceived to be "harmful" due to stigma, discrimination, hate, or other political or religious ideology. This may be, for example, by way of targeted malicious complaints being directed toward sex workers (even when posting material that is compliant with law and community guidelines), towards trans people, or towards people expressing political views that may cause offence to others.

We note this as a complicating factor with regard to "harmful" but not unlawful material, and that, depending on how such an expectation is implemented, it could lead to negative consequences for the safety of some groups and individuals, as well as unreasonable encroachment upon freedom of speech and expression.

In relation to generative AI, and to demonstrate the complexity here, we wish to draw the Department's attention to tools and groups such as 'DignifAI'—a group who claims to use Stable Diffusion to alter imagery of women who they have deemed to be "undignified" to put them into modest clothing.[10] While much attention has been directed towards the use of deepfakes to generate non-consensual sexual imagery, the same technology can also be used to non-consensually alter imagery in other politically motivated ways. At the heart of both uses of this technology is misogyny and efforts to intimidate, harass, and ultimately strip women of their agency through attempts to control their bodies. However, it is not clear that the designation of "harmful" would extend to meet

[10] 'DignifAI': 4chan is Editing Pictures to Clothe Women', *404 Media*, 5 February 2024. https://www.404media.co/dignifai-4chan-is-editing-pictures-to-clothe-women/

this use case. It is our experience that many common notions of "harmful" but not illegal content ("awful but lawful") over-emphasises sexual material.

Digital Rights Watch appreciates the intention to broaden the scope beyond material that is just unlawful, especially given the speed at which generative AI tools are developing, however it is essential that this category not become overly broad such that it inadvertently becomes a tool for motivated actors to weaponise against content they personally or politically find offensive, but does not meet the threshold of unlawful.

We note that while "serious harm" is defined in the *Online Safety Act 2021,* there is no specific definition of "harmful" as it would be applicable under the BOSE. Despite the lack of clarity on "harmful" within the Act, according to the 'Basic Online Safety Expectations Regulatory Guidance' issued by the eSafety Commissioner in September 2023, the designation of harmful depends on the scope of the Act:

> *"'Harmful material or activity is material or activity that may not be unlawful but is covered within the scope of the Act. It is also material or activity that should fall under a provider's terms of use, policies and procedures and standards of conduct for end-users (as outlined in Section 14 of the Determination)."*[11]

We further note that Division 10 of the OSA provides that the eSafety Commissioner can get the Classification Board to make a ruling determining whether material is Class 1 or Class 2, if necessary.

Due to this complexity, combined with the discretionary power afforded to the eSafety Commissioner, we remain troubled by the possibility of the designation of harmful being made based on political or personal ideological reasons. While the current eSafety Commissioner may not do this, there is little in the Act and corresponding documents that prevent this possibility under a different eSafety Commissioner in the future. While there are some transparency requirements placed upon the eSafety Commissioner, they do not go far enough. This issue should be revisited in the statutory review of the OSA.

We also strongly suggest that guidance on "harmful" should reflect and encompass the complexity and nuance of the concept, with an aim to minimise the risk of over-capture. Despite the recent advances in machine learning technology, there is currently no such thing as an automated system that can adequately parse this kind of nuance in a way that allows "harmless and lawful" material on a platform to continue unimpeded.

---

[11] 'Basic Online Safety Expectations Regulatory Guidance,' eSafety Commissioner, September 2023. https://www.esafety.gov.au/sites/default/files/2023-09/Basic-Online-Safety-Expectations-Regulatory-Guidance-updated-September-2023_0.pdf

Complaints mechanisms and transparency reporting

Notwithstanding the above section, Digital Rights Watch welcomes additional clarity and detail regarding the expectation that services must provide "clear and readily identifiable mechanisms" that enable them to report and make complaints about services' terms of use.

We do note, however, that it remains incredibly challenging for individuals to be able to seek redress from platform operators where they have, for example, have wrongfully had their account taken down or content removed, due to being incorrectly automatically flagged, or as a result of malicious complaints. As such, we suggest that the amendment to subsection 15(2) as listed in section 14 of the BOSE amendment be extended to also include appeals processes.

We are also pleased to note the additional requirements proposed for section 18 regarding the publication of transparency reports at regular intervals. We believe, however, that there is room to make the requirements of these transparency reports more clear and prescriptive, for example, as listed in the EU's Digital Services Act.

Digital Rights Watch would also strongly support the inclusion of measures for non-commercial researcher access to public interest data, so that essential research into platforms and their impact can be conducted without undue barriers.

Age assurance and age verification

Core expectation 12 requires providers to "take reasonable steps to prevent access by children to class 2 material".

Digital Rights Watch has provided extensive feedback through submissions, participation in roundtables, and discussions regarding the challenges of age assurance and age verification systems from the perspective of both privacy and digital security risks, as well as with regard to the challenges of effective technical implementation.

Our views on this issue have not changed since our previous submissions, and are summarised below for ease of reference:

1. Age verification is privacy-invasive, which can undermine the objective of reducing online harm. Most forms of age verification require the user to provide additional personal information beyond what is justifiable needed for proof-of-age in order to be effective. Incentivising companies and government agencies to collect, use, and store additional personal information in order to conduct age verification creates additional privacy and security risks which, in turn, can exacerbate online harms. There are

significant, if not insurmountable, challenges to implementing age verification in a way that is both effective, as well as minimising privacy and security risk.[12]

2. Age Verification and Restricted Access Systems have been considered in Australia and overseas in the past but have failed to be implemented due to their overreach, blunt approach, unreasonable impact upon individual's privacy, and the creation of adverse digital security risks.[13]

3. Other suggested "age assurance" processes such as age estimation based on face scanning also create privacy risks due to their use of biometric data, introduce the risk of inaccurate age classification, and are likely to have disproportionate negative impacts on marginalised communities. As precise biological age cannot be determined by an image of a human face, the use of this method guarantees that there will be adult individuals misclassified as children and prevented access to lawful material, as well as children misclassified as adults. Facial recognition technology has high rates of error across race and gender, which can result in significant differences in human age estimators across gender and race.[14]

4. Mandatory age verification is likely to act as a deterrent for many adults accessing legal content, and may prompt people of all ages toward less safe and secure internet services in order to circumnavigate providing personal information.[15]

5. The majority of suggested approaches to age verification are easily bypassed with the use of common technologies such as Virtual Private Networks (VPNs), significantly diminishing their effectiveness.[16]

The combination of these factors are likely to result in a system which is unduly invasive in data collection, creates new privacy and security risks by holding information on individuals, and yet is unlikely to be effective at preventing people under the age of 18 from accessing restricted content. If

[12] For further exploration of these issues, see for example, QUT Digital Media Research Centre submission in response to the calls for evidence regarding age verification and restricted access systems, Dr Zahra Stardust, Lucinda Nelson and Abdul Obeid. 14 September 2021. https://eprints.qut.edu.au/213887/1/2021_DMRC_and_ADMS_submission_re_age_verification_and_restricted_access_systems.pdf

[13] For example, the UK dropped its plan for online pornography age verification in 2019. While it has since then re-considered age verification, the reasons for abandoning it in the first place have not disappeared. https://www.theguardian.com/culture/2019/oct/16/uk-drops-plans-for-online-pornography-age-verification-system

[14] Guo, G., & Mu, G. (2010, June 13-18). Human age estimation: What is the influence across race and gender? IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops, San Francisco, CA, USA. https://doi.org/10.1109/CVPRW.2010.5543609.

[15] Blake, P. (2018). Age verification for online porn: more harm than good? Porn Studies, 6(2), 228– 237.

[16] Yar, M. (2019). Protecting children from internet pornography? A critical assessment of statutory age verification and its enforcement in the UK. Policing: An International Journal, 43(1), 183–197.

these underlying issues are not addressed, the outcome may be a system that is not simply ineffective but actively harmful.

We note that the Australian government decided against forcing sites to bring in age verification technology in 2023 following receiving the eSafety Commissioner's long-awaited roadmap for age verification for online pornography.[17]

The Government's response states:

> *"It is clear from the Roadmap that at present, each type of age verification or age assurance technology comes with its own privacy, security, effectiveness and implementation issues. For age assurance to be effective, it must:*
>
> - *work reliably without circumvention;*
> - *be comprehensively implemented, including where pornography is hosted outside of Australia's jurisdiction; and*
> - *balance privacy and security, without introducing risks to the personal information of adults who choose to access legal pornography.*
>
> *Age assurance technologies cannot yet meet all these requirements. While industry is taking steps to further develop these technologies, the Roadmap finds that the age assurance market is, at this time, immature. The Roadmap makes clear that a decision to mandate age assurance is not ready to be taken."*[18]

We are concerned that despite the government's own decision that age assurance is not an appropriate method for restricting access to content, these expectations continue to list it as an example of reasonable steps to meet s 12 of the BOSE, and in doing so continue to encourage the uptake of age assurance and age verification despite documented risks and failings.

Encrypted services

We remain concerned by section 8 of the BOSE, which requires services that use encryption to "take reasonable steps to develop and implement processes to detect and address material or activity on the service that is unlawful or harmful."

While subsection 8(2) does include exemptions to this, we note with alarm the recent draft industry standards for class 1A and 1B material, which included requirements for proactive detection. This would likely result in technology

[17] 'Australia will not force adult websites to bring in age verification due to privacy and security concerns,' *The Guardian,* 31 August 2023. https://www.theguardian.com/australia-news/2023/aug/31/roadmap-for-age-verification-online-pornographic-material-adult-websites-australia-law

[18] 'Government response to the Roadmap for Age Verification,' Department of Infrastructure, Transport, Regional Development, Communications and the Arts, August 2023. https://www.infrastructure.gov.au/sites/default/files/documents/government-response-to-the-roadmap-for-age-verification-august2023.pdf

providers implementing scanning technologies such as “client-side scanning”, whereby materials are scanned on a device before they are encrypted or after they are decrypted.

While such techniques technically may not “break” encryption, they do fundamentally undermine the promise and principle offered by encryption technologies. We remain concerned that the inclusion of section 8 in the BOSE, especially in combination with the proposed Industry Standards, unduly encourages companies to implement scanning technologies which would put the privacy and security of people’s communications at risk.

For more detail regarding the human rights risks and technical challenges of proactive detection in encrypted environments, we encourage the Department to read our latest submission to the eSafety Commissioner in response to the draft Industry Standards.[19]

Enforcement

Finally, and despite our concerns regarding some of the specifics included within the BOSE, we also note that there is a lack of strong enforcement mechanisms, rendering it a nice-to-have, but relatively toothless instrument. While the eSafety Commissioner may ask services for a report on how they are complying with the BOSE, and in turn may issue a Statement of Non-Compliance with one or more of the expectations, the subsequent pathways for enforcement through civil penalties have questionable levels of effectiveness. This was particularly brought into question as X/Twitter became the first platform to be issued a fine under the OSA (of the very small amount of $610,500) which it refused to pay.[20]

While Digital Rights Watch urges improvements in the substance of the BOSE (and indeed in the OSA as a whole), we also note that In order for the online safety regime to be effective, it will require more robust and meaningful enforcement mechanisms.

[19] ‘Submission to the eSafety Commissioner regarding the draft Designated Internet Services Standard and the draft Relevant Electronic Services Standard for class 1A and 1B material,’ *Digital Rights Watch,* 21 December 2023. https://digitalrightswatch.org.au/wp-content/uploads/2024/01/DRW-Submission-Draft-Online-Safety-Industry-Standards-Dec-2023.pdf

[20] ‘X fined $610,500 in Australia first for failing to crack down on child sexual abuse material,’ *The Guardian,* 15 October 2023 https://www.theguardian.com/technology/2023/oct/16/x-fined-610500-in-australia-first-for-failing-to-crack-down-on-child-sexual-abuse-material