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

21 December 2023

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 eSafety Commissioner regarding the draft Designated Internet Services (DIS) Standard and the draft Relevant Electronic Services (RES) Standard for class 1A and class 1B material ("the draft Standards").

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. As such, the majority of this submission focuses on the risks and challenges posed by scanning technologies and the proactive detection obligations contained in the draft standards. We recognise that there are a range of other requirements contained within the draft standards that are not addressed in this submission, but that we welcome, such as requirements for mechanisms to enable end-users to report and make complaints to services regarding illegal material.

Digital Rights Watch recognises the severity of harm caused by the dissemination of Child Sexual Abuse Material (CSAM) and other forms of illegal content, and that there are genuine challenges regarding regulating the creation, storage and distribution of unlawful material online. Yet Australia's policy response must be evidence-based, respect the human rights of all people (including children), and reflect software security best practices.

Due to the emotive nature of CSAM and "pro-terror" content, proposed solutions can sometimes risk becoming disproportionate or unreasonable. We recognise that the right to privacy, while important, is not absolute. However, we remain concerned that the dangers of widespread surveillance capability without adequate oversight and accountability can become obscured in discussions of such abhorrent conduct.

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 secure 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,* 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 current proposed Industry Standards, including:

- 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 Joint Committee on Law Enforcement regarding the Inquiry into law enforcement capabilities in relation to child exploitation, 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 Parliamentary Joint Committee on Law Enforcement in relation to the *Criminal Code Amendment (Sharing of Abhorrent Violent Material) Act 2019,* October 2021.
- Submission to the Department of Infrastructure, Transport, Regional Development and Communication on the proposed *Online Safety Bill 2020,* February 2021.

We welcome the opportunity to discuss this submission with the eSafety Commissioner further.

Proactive detection of CSAM and "pro-terror" material

Digital Rights Watch is particularly concerned with regard to the proactive detection provisions in both sets of draft standards, which include obligations upon services to detect and remove known CSAM and known pro-terror material and to disrupt and deter CSAM and pro-terror material (both known and first-generation[1]).

Taken together, these standards will apply to an extremely broad range of services that Australians use every day, including email, messaging, online dating and gaming, as well as personal file storage. Given that cyber threats are increasing, individuals, businesses, and governments rely on these services to have robust digital security in place. Any policy decision to weaken the security of such services must be carefully considered. Our key concerns are highlighted below.

[1] By "first-generation", we broadly mean material that is newly-generated, such that it is not contained within a particular dataset of confirmed material (hashed or otherwise) that would serve as the comparison point for matching purposes. When first-generation content is exposed to a machine learning classifier designed to determine the presence of particular material (such as CSAM), it means that the content has not previously been "seen" by the model; it was not part of the training or testing datasets.

Detection and removal of *known* material

We note that proactive detection as an approach to addressing the spread of illegal content was discussed at length in the consultation process conducted by the steering group of industry associations in the development of the initial Online Safety Industry Codes. At the time, we were pleased to note that the industry associations concluded that:

> *"the extension of proactive detection measures could have a negative impact on the privacy and security of end-users of private communications and file storage services, including services used by businesses and government enterprises."*[2]

We agreed, and continue to agree, that proactive detection of material in private and encrypted communications and file storage is an unreasonable invasion of privacy and creates additional security and safety risks for individuals, businesses, and governments.

We also note that the industry steering group sought to limit proactive detection requirements in the original codes in public and/or unencrypted online spaces to instances of *known* CSAM only, which we supported. While we understand the need to address the harms associated with the distribution of first-generation CSAM, the technology to detect material that has not been previously identified currently presents unreasonable challenges and risks (discussed further below).

Proactive detection will always carry with it some level of privacy and security risk. While Digital Rights Watch does not argue against the use of hash scanning in public and unencrypted platforms or services for known CSAM, we are deeply concerned by the lack of safeguards for privacy, security and encryption in the proposed draft standards. As currently drafted, the standards are likely to incentivise companies to minimise or undermine their application of encryption, or to implement technologies and processes that undermine the *aim* of encryption to establish private and secure services.

Disruption and deterrence of *any* CSAM and "pro-terror" material

While established processes exist to detect known CSAM — content that has been identified, investigated and confirmed as CSAM and is contained within one of the relevant databases — the technology to detect first-generation content, or material that has not previously been identified and confirmed, currently presents significant challenges and risks in relation to accuracy, over- or under- capture and privacy.

---

[2] Head terms, draft industry codes, as referenced in the Digital Rights Watch 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. https://digitalrightswatch.org.au/2022/10/11/submission-online-safety-draft-industry-codes/

High quality, automated detection of such content is extremely difficult, especially when the scope of target content is broad, not strictly defined, or difficult to assess without additional context. For example, a person may hold what might be flagged as "pro-terror" material for entirely legitimate purposes, such as for documentation of human rights abuses, political activism, investigative journalism, and so on. Such context cannot be easily assessed by scanning the image alone, and would require significantly more invasive methods to determine.

The automatic detection of "pro-terror" material is also particularly challenging given the changing notion of what does and does not count as "terrorism". As a most recent example, the developments between Israel and Palestine clearly demonstrate the ways in which the designation of "terrorism" can dynamically change based upon the political agenda of a given situation. Just last month, a woman was arrested for a WhatsApp status which was deemed as "terrorism" for expressing empathy for victims of conflict.[3]

Machine learning classifiers which seek to automatically flag possibly illegal content (as opposed to matching content to known, or previously identified content) may be improving, but they remain seriously flawed when it comes to classifying complex material at scale.

One issue is accuracy and the risk of both over- and under- capture of content. Over-capture—the mis-classification of legal material as prohibited—can lead to significant negative consequences for individuals holding the content that is wrongly flagged, such as removal of access to services[4] (which can lead to loss of income, support networks, and community), reputational damage, personal distress, and in the worst cases lead to wrongful accusation of criminal activity. On a larger scale, it can also result in the policing and censorship of legitimate material, including legal sexual material and sexual education and health material.[5]

---

[3] 'Palestinians arrested for 'terrorism' over WhatsApp status,' *Al Jazeera,* 9 November 2023. https://www.aljazeera.com/program/newsfeed/2023/11/9/palestinians-arrested-for-terrorism-over-whatsapp-status

[4] For example, a man was stripped of access to his Google accounts after he took a photo of his son's groin to send to a medical practitioner and it was wrongly flagged as CSAM. Experts said the case highlights the dangers of automated detection of CSAM. See 'Google refuses to reinstate man's account after he took medical images of son's groin,' *The Guardian,* 23 August 2022. https://www.theguardian.com/technology/2022/aug/22/google-csam-account-blocked

[5] Censorship and de-platforming of sexual health material already happens on mainstream digital platforms as a result of over-capture through automated content moderation, see for example: 'Deplatforming sex education on Meta: sex, power, and content moderation,' Joanna Williams, Media International Australia, 2023. https://journals.sagepub.com/doi/10.1177/1329878X231210612. It is also commonplace for sex

Another issue is that due to the lack of training data, classifier models are more likely to make mistakes related to marginalised groups, and in doing so further entrench pre-existing inequalities. Additionally, there remain ongoing challenges regarding the explainability of machine or deep learning classifiers. While this is an area of ongoing technical research and development, at this stage it may not be possible to explain or justify why some content is flagged by an automated machine learning content classification system.

A particular challenge in the detection of first-generation CSAM through machine learning classifiers and other artificial intelligence processes is that it needs to be able to reliably estimate the age of the person depicted in the content. This is technologically challenging and can result in high rates of false positives, especially for people who are within a few years of 18.[6]

One often-proposed mitigator for accuracy and explainability challenges in such automated systems is the inclusion of human review, oversight, or a "human-in-the-loop".[7] However, we caution against assuming such safeguard as sufficient, as recommendations or decisions made by AI or automated systems have been shown to influence human perception and decision making.[8] Further, research has shown a lack of scientific basis for assessing a person's age from attributes perceptible in an image, such as breast size or face shape. Specifically, it has been demonstrated that expert assessments performed by human paediatricians attempting to determine whether an individual appearing in

---

workers to be 'de-platformed', even when they follow community guidelines and are posting legal material. See: 'Automating whorephobia: sex, technology and the violence of deplatforming,' Danielle Blunt and Zahra Stardust, *Porn Studies,* 12 May 2021. https://www.tandfonline.com/doi/full/10.1080/23268743.2021.1947883?scroll=top&needAccess=true

[6] 'Proposal for a regulation laying down the rules to prevent and combat child sexual abuse: Complimentary impact assessment,' *European Parliamentary Research Service,* April 2023. https://www.europarl.europa.eu/RegData/etudes/STUD/2023/740248/EPRS_STU(2023)740248_EN.pdf

'Towards a Framework for Evaluating CSAM Prevention and Detection Tools in the Context of End-to-End Encryption Environments: a Case Study,' Peersman et al, *University of Bristol,* February 2023. https://research-information.bris.ac.uk/en/publications/towards-a-framework-for-evaluating-csam-prevention-and-detection-

[7] 'Explaining the technology for detecting child sexual abuse online,' *Child Rights International Network,* November 2023. https://home.crin.org/readlistenwatch/stories/explainer-detection-technologies-child-sexual-abuse-online

[8] 'Humans inherit artificial intelligence biases,' Lucia Vicente & Helena Matute, *Scientific Reports,* September 2023 https://www.nature.com/articles/s41598-023-42384-8

pornography is under the age of 18 resulted in the incorrect determination of adult women to be children two-thirds of the time.[9]

While we appreciate that the eSafety Commissioner wishes to promote investment into technologies (including machine learning and other forms of artificial intelligence) to disrupt and deter illegal conduct and material, Digital Rights Watch remains concerned that this is being incentivised without adequate safeguards in place. Given the immense potential for harmful consequences, it is essential that the technology is rigorously tested and considered for human rights implications before being implemented at scale.

We further note that when content is wrongfully removed from mainstream social media platforms it is currently extremely difficult for individuals to appeal the decision or seek redress. This compounds the harm of having had material incorrectly flagged or removed in the first place. In addition to requirements for reporting mechanisms for users to report or complain about prohibited material, there ought to also be requirements for meaningful appeal and redress processes such that individuals are able to take action in instances where they have been wrongfully targeted by detection, deterrence, removal, and disruption mechanisms. Such instances of wrongful or inaccurate classification ought to be contained in transparency reporting provided to the eSafety Commissioner and made public, such that it is possible to understand the efficacy of the requirements of the standards in relation to the unintended consequences.

Risks created by scanning technologies

There are currently two primary technologies used for image scanning:

- Perceptual hashing: in which specialised algorithms process images to produce a corresponding "fingerprint" or hash. This can then be compared to datasets containing the fingerprints of target content (in this case, known CSAM or "pro-terror" material).
- Machine learning: in which a model is trained on both innocuous and target content, to then scan and classify material. This can be used (with varying degrees of accuracy) on both known and first-generation material, and is more commonly used to filter video and text.

Both of these approaches require access to unencrypted data for matching or training, and both rely on proprietary tools developed from a corpus of target content, which may be controlled by a third party. Both methods can also be treated as a black box that inputs unencrypted content and outputs a determination of whether it is likely to contain targeted material.

[9] 'Inaccuracy of age assessment from images of post pubescent subjects in cases of alleged child pornography,' Arlan L Rosenbloom, *International Journal of legal medicine*, September 2012. https://pubmed.ncbi.nlm.nih.gov/22960879/

Currently, most tech companies' scanning processes run on their own servers—for instance, to detect malware and spam. However, such "server related" solutions cannot be implemented in end-to-end encrypted environments because the content must be decrypted on the server in order for the detection to run, in turn allowing third-party access to the content.[10] As such, there are increasingly frequent proposals for scanning to take place on the client-side, that is, the use of scanning software that runs directly on a user's device. While both processes happen outside of the individual's control, client-side scanning creates the capacity to scan files that might otherwise never leave a user's device, and in doing so extends the reach of surveillance into personal devices, pushing across the boundary between what is shared and what is private. As highlighted by security researchers: "because this privacy violation is performed at the scale of entire populations, it is a bulk surveillance technology."[11]

The specific privacy, security and implementation risks created by such systems vary, and can include the creation of new software security vulnerabilities, broadening of attack surfaces, additional privacy risks, and ongoing questions of efficacy in adversarial environments. We strongly recommend that the eSafety Commissioner engage with documentation of such risks including, but not limited to:

- 'Bugs in our Pockets: The Risks of Client-Side Scanning,' Abelson et al, *Cornell University,* October 2021.[12]
- 'Statement on encryption and mandatory Client-Side Scanning of Content,' *Internet Architecture Board,* December 2023.[13]
- 'Proposal for a regulation laying down the rules to prevent and combat child sexual abuse: Complimentary impact assessment', *European Parliamentary Research Service,* April 2023[14]

[10] 'Protecting Children in the Age of End-to-End Encryption,' Laura Draper, *Joint PIJIP/TLS Research Paper Series*. 80. https://digitalcommons.wcl.american.edu/research/80/

[11] 'Bugs in our Pockets: The Risks of Client-Side Scanning,' Abelson et al, *arXiv preprint arXiv:2110.07450*, 2021. https://arxiv.org/abs/2110.07450

[12] 'Bugs in our pockets' https://arxiv.org/abs/2110.07450

[13] 'IAB Statement on Encryption and Mandatory Client-Side Scanning of Content,' *Internet Architecture Board,* December 2023. https://www.iab.org/documents/correspondence-reports-documents/2023-2/iab-statement-on-encryption-and-mandatory-client-side-scanning-of-content/

[14] 'Proposal for a regulation laying down the rules to prevent and combat child sexual abuse: Complimentary impact assessment,' *European Parliamentary Research Service,* April 2023. https://www.europarl.europa.eu/RegData/etudes/STUD/2023/740248/EPRS_STU(2023)740248_EN.pdf

- ‘Deep perceptual hashing algorithms with hidden dual purpose: when client-side scanning does facial recognition,’ Jain et al, *Institute of Electrical and Electronics Engineers,* May 2023.[15]
- ‘Adversarial Detection Avoidance Attacks: Evaluating the robustness of perceptual hashing-based client-side scanning’ Shubham Jain, Ana-Maria Cretu, and Yves-Alexandre de Montjoye, *Imperial College London,* August 2022.[16]
- Open letter on the position of scientists and researchers on the EU’s proposed Sexual Abuse Regulation with over 450 signatures.[17]

Critically, it is not possible to design a client-side scanning system that can be technologically limited to *only* be used for one form of material. Once implemented, it is possible to expand or change the target content to cover any material that an influential minority, suitably motivated malicious actor, or specific jurisdiction may consider objectionable (for example, images depicting individuals of political interest such as whistleblowers, LGBTQ+ material, sex worker content, material related to reproductive health services, and so on). This may be achieved by adding additional material to the dataset, using a different dataset, or changing the training material.[18] It can also be achieved by altering the “fingerprint” of target images in order to trigger the flagging of innocuous material that the system reads as matching target content.[19]

---

[15] ‘Deep perceptual hashing algorithms with hidden dual purpose: when client-side scanning does facial recognition,’ Shubham Jain, Ana-Maria Cretu, Antoine Cully & Yves-Alexandre de Montjoye, *IEEE,* July 2023. https://ieeexplore.ieee.org/document/10179310

[16] ‘Adversarial Detection Avoidance Attacks: Evaluating the robustness of perceptual hashing-based client-side scanning,’ Shubham Jain, Ana-Maria Cretu, and Yves-Alexandre de Montjoye, *Imperial College London,* August 2022. https://www.usenix.org/conference/usenixsecurity22/presentation/jain

[17] ‘Joint statement of scientists and researchers on EU’s proposed Child Sexual Abuse Regulation,’ July 2023 https://docs.google.com/document/d/13Aeex72MtFBjKhExRTooVMWN9TC-pbH-5LEaAbMF91Y/edit

[18] ‘Why Adding Client-Side Scanning Breaks End-to-End Encryption,’ *Electronic Frontier Foundation,* November 2019. https://www.eff.org/deeplinks/2019/11/why-adding-client-side-scanning-breaks-end-end-encryption; see also ‘Bugs in our pockets’ https://arxiv.org/abs/2110.07450

[19] Perceptual hashing functions may function reasonably well in a non-adversarial context, given that it is not likely that two different images will just happen to have the same ‘fingerprint’ and generally speaking small modifications do not usually change that fingerprint. However, there are close to no guarantees in an adversarial setting. Researchers have demonstrated that it’s possible to deliberately modify images that are imperceptible to a human but completely change the fingerprint (see, for example, https://dl.acm.org/doi/pdf/10.1145/3460120.3484559), and that it’s possible to deliberately generate images that look different to the human eye but have the same fingerprint (see,

Additionally, due to the legal status and nature of the material (in this instance, CSAM and "pro-terror" content), auditing and assessing the training material or matching datasets for the purpose of transparency, accountability, and oversight is incredibly challenging. This is made more challenging given that hash databases are not inspectable to the human eye and require additional technical processes in order to conduct independent review.

Some have proposed homomorphic encryption as a solution that may work alongside end-to-end encryption, however such approaches currently take 10-15 seconds per image making them functionally impracticable, and it is still possible for malicious actors to leverage such a system to block or flag legitimate content for their own purposes.[20]

Encryption and "technical feasibility"

The eSafety Commissioner has publicly stated that it supports privacy and digital security, and does not advocate building in weaknesses or "backdoors" to undermine end-to-end encrypted services.

The eSafety Commissioner has publicly stated that it:

> *"...is not requiring companies to break end-to-end encryption through these standards nor do we expect companies to design systematic vulnerabilities or weaknesses into any of their end-to-end encrypted services."*[21]

However, there is nothing contained within the text of the standards which would ensure that they are interpreted in line with this stance. This expectation should be made clear in the body of the standards to ensure clarity of interpretation now and into the future.

In a statement on end-to-end encryption made by the eSafety Commissioner in October 2023, "proactive, device-based detection tools", also known as "client-side

---

for example, https://github.com/anishathalye/neural-hash-collider). For further explanation see 'Bugs in our pockets' https://arxiv.org/abs/2110.07450

[20] 'An Overview of Perceptual Hashing,' Hany Farid, *Journal of Online Trust and Safety,* October 2021 https://tsjournal.org/index.php/jots/article/view/24/14

'Protecting Children in the Age of End-to-End Encryption,' Laura Draper, *Joint PIJIP/TLS Research Paper Series*. 80. https://digitalcommons.wcl.american.edu/research/80/

[21] 'Australia releases new online safety standards to tackle terror and child sexual abuse content,' *The Guardian,* 20 November 2023. https://www.theguardian.com/australia-news/2023/nov/20/australia-esafety-standards-new-2023-targets-child-content-terrorism-detection

'Updated position statement on end-to-end encryption,' *eSafety Commissioner,* October 2023. https://www.esafety.gov.au/sites/default/files/2023-10/End-to-end-encryption-position-statement-oct2023.pdf

scanning" is listed as an existing method for proactively detecting CSAM "that can be utilised by online services alongside E2EE".[22] This — combined with specific references to "hash technologies" in the standards — demonstrates the eSafety Commissioner's position that client-side scanning can operate in harmony with end-to-end encryption.

While it may be technically true *in some cases* that client-side scanning can technically function without "breaking" encryption, by scanning materials before they are encrypted or after they are decrypted, client-side scanning undermines the promise and principle of private and secure communications offered by encryption. It means that the content of private messages, including those of at-risk users and vulnerable groups, would be subject to proactive surveillance.

As highlighted by security researchers, client-side scanning systems can "circumvent encryption completely by monitoring all content prior to the user encrypting and sending it, or after receiving it, or when backing it up to the cloud…it would thus replicate the behaviour of a law-enforcement wiretap."[23]

Further, there are encryption technologies that mask the lookup phase and period after the message has made passage through the network, as well as those that conceal metadata. These technologies are likely to come into conflict with the assertion that client-side scanning can work "alongside" encryption. It is essential that the eSafety Commissioner considers the role that encryption technologies play beyond just encryption in transit.

Becoming stuck in a debate over whether client-side scanning or other methods of proactive detection "break" encryption is not helpful, and allows for misunderstanding or confusion over semantics. For many people, hearing the eSafety Commissioner state that what is being proposed "will not break encryption" may be interpreted as "will protect privacy and security", but this is misleading.

The Child Rights International Network and defenddigitalme assessed a range of technologies used to identify CSAM in encrypted environments including client-side scanning, homomorphic encryption, wiretapping, and spyware/malware such as software like Pegasus[24]. They conclude that "encryption

[22] 'Bugs in our pockets' https://arxiv.org/abs/2110.07450

[23] 'Bugs in our pockets' https://arxiv.org/abs/2110.07450

[24] Pegasus is spyware developed by Israeli surveillance company, NSO Group. See: 'What is Pegasus Spyware and how does it hack phones?' *The Guardian,* 19 July 2021. https://www.theguardian.com/news/2021/jul/18/what-is-pegasus-spyware-and-how-does-it-hack-phones

can be broken in principle, if not in practice, if its aims are compromised."[25] They also highlight the importance of encryption in securing children's safety and rights, reject any generalised ban on encryption, and note that stakeholders "should recognise that technology can be repurposed to further a variety of policy goals, including surveillance and the identification of legitimate material."

It is essential to understand that encryption is the *means*, not the *end*. That is, we defend encryption as one of the few mechanisms available to practically ensure security and privacy of data, but the underlying goal remains privacy and security. Technological proposals such as client-side scanning undermine that underlying goal.

The European Parliamentary Research Service determined that "the detection of CSAM in end-to-end encrypted communications is possible but the solutions available are not sufficiently transparent and secure, and known detection mechanisms undermine the end-to-end protection offered by the encryption" and that detection of CSAM in end-to-end encrypted communications "enhance the vulnerabilities to attacks and abuse and would raise practical issues related to trust, accountability, and transparency."[26]

Digital Rights Watch has long advocated for robust digital security, including protecting the use of encryption. Encryption — and other digital security mechanisms — is essential for safety, national and individual security, and the protection of human rights.[27]

The UN Special Rapporteur on Freedom of Expression has referred to end-to-end encryption as "the most basic building block" for digital security on messaging apps. Because of its critical role, the Special Rapporteur further notes that: "the responsibility to safeguard freedom of expression and privacy may require companies to establish end-to-end encryption as a default setting in their messaging products". The Rapporteur also suggests that companies that offer messaging apps "should seek to provide the highest user privacy settings by default".[28]

---

[25] 'Privacy and Protection: A children's rights approach to encryption,' *Child Rights Network International,* January 2023.
https://home.crin.org/readlistenwatch/stories/privacy-and-protection

[26] 'Proposal for a regulation laying down the rules to prevent and combat child sexual abuse: Complimentary impact assessment,' *European Parliamentary Research Service,* April 2023.
https://www.europarl.europa.eu/RegData/etudes/STUD/2023/740248/EPRS_STU(2023)740248_EN.pdf

[27] See, for example, 'The Role of Encryption in Australia', *Access Now*, January 2018.
https://digitalrightswatch.org.au/2018/01/19/the-role-of-encryption-in-australia/

[28] 'Encryption and Anonymity follow-up report,' *United Nations Human Rights Special Procedures*, June 2018.

As a jurisdiction without a formal bill of rights, people in Australia are left with minimal protection of their human rights. There are very few limits on the powers afforded to law enforcement and intelligence agencies. We urge the eSafety Commissioner to consider how compelling services to weaken, undermine, bypass, or sidestep the privacy and security afforded by encryption threatens the safety of Australians at an individual, community, and national security level.

While we appreciate the extent to which encrypted communications can sometimes introduce friction into criminal investigations, we encourage much greater evidence-based discussion on these matters as a vital starting point for any consideration of such controversial and invasive technologies such as client-side scanning.

Technical feasibility provisions

While we appreciate that sections 20 and 21 include a provision stating that the requirements in these sections "[do] not require a provider to use a system, process or technology if it is not technically feasible for the provider to do so", we also note that the interpretation of "technical feasibility" under section 7 only takes into account the "expected financial cost to the provider" and whether it is reasonable for the provider to incur that cost.

There is no reference to the technical limitations that providers of encrypted services may face in taking action as set out by sections 20 and 21. Services that do not collect or store the requisite data in order to be able to meet proactive detection requirements, and services that use encryption such that the provider has no access to the material in order to perform proactive detection, may instead be forced or encouraged to begin gathering such data in order to comply with the proposed standards.

If the proposed eSafety standards are not amended to enshrine data minimisation principles, they run counter to the Australian Government's own stance, which has shifted towards responsible data governance as the ramifications from high-profile data breaches of Australian businesses in recent years have made the harms of data retention schemes apparent.[29] The recently released 2023-2030 Cyber Security Strategy notes that mishandled data "can cause grave damage to Australia's national interests" and pledges to review previous government-mandated data retention schemes to ensure they are "appropriately balanced" to prevent the types of "vulnerabilities that arise from

---

https://www.ohchr.org/Documents/Issues/Opinion/EncryptionAnonymityFollowUpReport.pdf

[29] 'Labor to reconsider mandatory data retention laws for companies in light of major hacks,' *The Guardian,* 22 November 2023. https://www.theguardian.com/australia-news/2023/nov/22/labor-mandatory-data-retention-laws-companies-hacks-cyber-security-strategy

entities holding significant volumes of data for longer than necessary."[30] Any business already taking a data minimisation approach should therefore be encouraged to continue, rather than to reverse this practice.

Alignment with TOLA

We also note that the interpretation of technical feasibility under the proposed standards does not align with that of the *Telecommunications and Other Legislation Amendment (Assistance and Access) Act 2018* (TOLA).

Under s 317W(7), when considering a Technical Capability Notice (TCN), the assessors must consider whether the requirements imposed are reasonable and proportionate, whether compliance is practicable and technically feasible, and whether it is the least intrusive measure that would be effective in achieving the legitimate objective.

The Department of Home Affairs explains that an assessment of practicability and technical feasibility considers resourcing (as does the interpretation of "technical feasibility" under s 7 of the draft industry standards) *as well as* the required technical procedures:

> *"An assistance instrument is technically feasible when the assistance sought relates to an* ***existing capability that is within the provider's power to utilise****"…"or where the new capability that is sought is one that the provider is able to build."*[31]
>
> *"The assessment of technical feasibility also denotes an assessment of what is technically feasible within the bounds of the legal safeguards in the legislation. For example, consider a situation where it is feasible to enable access to a targeted user's encrypted data carried over an end-to-end encrypted service, however doing so would create a material risk that unauthorised parties could access the data of another, non-targeted user. This activity* ***would not*** *be technically feasible, in a legal sense, within the parameters of the legislation because it would contravene the prohibition against systemic weaknesses."*[32]

---

[30] '2023-2030 Australian Cyber Security Strategy,' *Department of Home Affairs*, 2023, Section 9. https://www.homeaffairs.gov.au/cyber-security-subsite/files/2023-cyber-security-strategy.pdf

[31] 'Industry assistance under Part 15 of the *Telecommunications Act 1997 (Cth)*: Administrative guidance for agency engagement with designated communications providers,' *Department of Home Affairs*. https://www.homeaffairs.gov.au/nat-security/files/assistance-access-administrative-guidance.pdf

[32] 'Industry assistance under Part 15 of the *Telecommunications Act 1997 (Cth)*: Administrative guidance for agency engagement with designated communications

Unchecked executive power and misalignment of international and domestic policy

There is nothing within the enacting legislation, the *Online Safety Act 2021,* that creates or requires obligations upon service providers to proactively monitor communications over their networks. This was a key part of policy debates in the lead up to the passage of the Act. It is not acceptable for the eSafety Commissioner, by way of industry standards, to extend the obligations beyond the legislative intent reflected in the Act—this would represent a serious overreach of eSafety power.

We also note that the EU *Digital Services Act* has introduced a prohibition on general monitoring—in no small part due to concerns regarding some policymakers asking platforms to scan all communications to find particular content.

While international debate regarding proactive detection and client-side scanning is ongoing, some notable developments include:

- The European Data Protection Board put out a joint opinion in 2022 that warned against a proposal under discussion by lawmakers in Europe to scan for CSAM and noted that both client-side and server-side scanning were "fundamentally incompatible with the end-to-end encryption paradigm, since the communication channel, encrypted peer-to-peer, would need to be broken" and notes that even though the European proposal allows flexibility on choice of technologies (as is also the case in the Australian draft standards), the "structural incompatibility of some detection order with end-to-end encryption becomes in effect a strong disincentive to use end-to-end encryption."[33]
- A range of privacy, security and legal experts voiced opposition to the EU proposal for proactive detection of CSAM based on legality and proportionality, human rights, cyber security, and effectiveness.[34]

providers,' *Department of Home Affairs*.
https://www.homeaffairs.gov.au/nat-security/files/assistance-access-administrative-guidance.pdf

[33] 'EDPB-EDPS Joint Opinion 04/2022 on the Proposal for a Regulation of the European Parliament and of the Council laying down rules to prevent and combat child sexual abuse,' *European Data Protection Board,* 28 July 2022.
https://edpb.europa.eu/our-work-tools/our-documents/edpbedps-joint-opinion/edpb-edps-joint-opinion-042022-proposal_en

[34] 'Europe's CSAM-scanning plan is a tipping point for democratic rights, experts warn,' *TechCrunch,* 25 October 2023.
https://techcrunch.com/2023/10/24/eu-csam-scanning-edps-seminar/

- In September 2023 the UK Government conceded that it is not technically possible to scan for CSAM without violating the privacy rights of all users and undermining encryption.[35]

More research needed to support evidence-based approaches

To conclude, we wish to highlight the need for more robust research into the area of CSAM and evidence-informed interventions such that effective and rights-respecting approaches can be established.

A systematic review of criminal justice responses to CSAM conducted by the Australian Institute of Criminology revealed "scarce evaluation research which limited the ability to holistically address CSAM."[36] The report also notes:

> *"Overall, the existing intervention literature in the area of CSAM is largely descriptive, with potentially promising interventions evaluated with low-quality research designs that do not reliably establish effectiveness."*
>
> *"Without a rigorous evidence base, policymakers and practitioners are unable to make reliable decisions about what criminal justice responses are effective in addressing CSAM offending and, potentially, what may be harmful."*

Given the intrusive nature of regularly-proposed technological "solutions" to CSAM and the significant risks they pose to human rights of all Australians (including children), we consider it unreasonable to introduce new requirements that incentivise industry to develop controversial and invasive technologies. A determination as to whether limitations on certain rights — such as the right to privacy — are indeed reasonable or proportionate cannot be demonstrated in the absence of a robust evidence base. Otherwise, interventions may stand to undermine human rights and risk creating additional harm while at the same time lacking substantial assurance of efficacy at actually reducing the harm of CSAM and its dissemination.

---

[35] See, for example, 'Britain Admits Defeat in Controversial Fight to Break Encryption,' *Wired,* 6 September 2023. https://www.wired.co.uk/article/britain-admits-defeat-in-online-safety-bill-encryption

[36] 'Criminal justice response to child sexual abuse material offending: a systematic review and evidence and gap map,' *Australian Institute of Criminology,* April 2021.: https://www.aic.gov.au/sites/default/files/2021-03/ti623_criminal_justice_responses_to_csam_offending.pdf

Recommendations

1. Ensure any technology used to disrupt and deter illegal content and material is rigorously tested and considered for human rights implications before being implemented at scale.
2. Clarify that the standards do not encourage or require the use of client-side scanning, given the privacy, security and implementation risks and the capacity for such tools to facilitate bulk surveillance.
3. Ensure the industry standards clearly state that encrypted services are not required to weaken, undermine, bypass or side-step encryption. For example, this may be achieved by amending the "technical feasibility" provisions to align with the interpretation as consistent with TOLA.
4. Clarify that the standards will not force or encourage service providers to begin gathering user data that they would otherwise not collect in order to comply.
5. Support robust research and the development of evidence-based interventions to deal effectively with CSAM and other illegal content and material.
6. In addition to reporting mechanisms for users to report or complain about prohibited material to services, include requirements for robust and meaningful appeal and redress mechanisms in cases where an individual's content has been misclassified or wrongfully flagged.