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

regarding the statutory review of the

Online Safety Act 2021

21 June 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 (DRW) welcomes the opportunity to submit comments to the Department of Infrastructure, Transport, Regional Development, Communications and the Arts as part of the Statutory Review of the Online Safety Act 2021.

As Australia's leading digital rights organisation, DRW is primarily concerned with the human rights, safety and wellbeing of individuals and communities in the digital age. **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.

Digital Rights Watch actively participates in public consultations regarding the development of legislation and policy in relation to technology and human rights, and have consistently contributed to the public debate regarding Australia's online safety scheme. Our recent submissions relevant to online safety in Australia include:

- Submission to the initial Online Safety Legislative Reform Consultation (February 2020)
- Submission on the proposed *Online Safety Bill*, (February 2021)
- Submission in response to the Restricted Access Systems discussion paper (September 2021) and draft declaration (November 2021)
- Submission on the draft Basic Online Safety Expectations (November 2021) and the later proposed amendments (March 2024)
- Submission to the Inquiry into Online Safety and Social Media (January 2022)
- Submission on the draft online safety Industry Codes (October 2022) and the subsequent draft Industry Standards (January 2024)

Digital Rights Watch welcomes the opportunity to participate in public hearings or further consultations and to provide comment and feedback on future specific proposals.

Human rights must be at the centre of Australia's approach to online safety

Protecting, enhancing and upholding human rights is essential to the realisation of meaningful safety—both online and off.

As the issues paper identifies, there are important nuances to be considered when assessing human rights impacts. A human-rights approach can help ensure that rights are appropriately balanced where there might be conflicts present. Very few rights are absolute, and so a human-rights approach assesses which rights might be interfered with, and then determine if it is necessary and proportionate interference, for a legitimate purpose.

In Australia and internationally, online safety policy proposals often threaten to infringe upon human rights, for example by undermining end-to-end encryption (or mechanisms to side step it altogether), privacy-invasive methods for age verification, processes which reduce or threaten online anonymity, and the reliance on increased automated content moderation. This is made worse in Australia with a lack of a federal human rights charter.

It is our experience that policymaking with regard to children's safety in particular often does not give enough weight to the full suite of rights afforded to children, including their right to privacy, freedom of expression, access to information, rights to take part in cultural and creative activities, and so on. Proposals under the banner of 'online safety' such as to ban children completely from social media[1] ignore the potential negative impact upon these, and other, rights of children.

It is our view that additional safeguards are necessary to ensure the Act upholds fundamental human rights and supporting principles.

Digital Rights Watch strongly suggests that the Australian government prioritise the creation and enactment of a federal Human Rights Act. Doing so would:

- assist in the creation of a rights-respecting culture in Australia,
- ensure that human rights are proactively considered in any new legislation,
- create a powerful tool to challenge injustice, including where facilitated by technologies, and
- provide opportunities for people to take action and seek justice where their rights have been violated.

We also support the creation of a separate but complementary Charter of Digital Rights and Principles, which could specifically focus on the application of human rights to existing and emerging technologies.[2] For example, the European Union's Declaration on Digital Rights and Principles was designed to complement existing rights, and to provide

[1] See, for example: Government of South Australia, 'Nation-leading move to protect our children from social media,' Media Release, 12 May 2024 https://www.premier.sa.gov.au/media-releases/news-items/nation-leading-move-to-protect-our-children-from-social-media; Josh Taylor, 'Peter Dutton wants a social media ban for children. But would 'real life' rules work?', *The Guardian,* 14 June 2024 https://www.theguardian.com/australia-news/article/2024/jun/14/peter-dutton-wants-a-social-media-ban-for-children-but-would-real-life-rules-work; Karen Midleton, 'Albanese follows Dutton's lead with tougher position on children's social media ban,' *The Guardian,* 13 June 2024 https://www.theguardian.com/media/article/2024/jun/13/anthony-albanese-peter-dutton-social-media-ban-age-16

[2] We suggest that this could be modelled on the European Union's Declaration of Digital Rights and Principles. For more detail see: Digital Rights Watch Submission to The Parliamentary Joint Committee on Human Rights regarding the Inquiry into Australia's Human Rights Framework, 29 June 2023. Available at: https://digitalrightswatch.org.au/2023/07/11/efa-sub-human-rights/

guidance for the European Union and its member states as they pursue "human centric" and "sustainable" digital transformation.[3]

Given that Australian leaders and policymakers are interested in Australia being a world leader in online safety, and is embarking on numerous digital transformation projects and concurrent tech-related regulatory efforts, a guiding set of overarching digital rights and principles would be useful to ensure that Australia's digital future is grounded in human rights, safety and dignity of all Australians.

Protecting privacy is a key part of safety

The issues paper notes that:

> *"In 2022, the House of Representatives Select Committee on Social Media and Online Safety concluded that while privacy concerns are critical to the rights of all internet users, those issues did not 'outweigh the fundamental issue of ensuring safety in online environments.'*[4]

It is our view that protecting privacy is an essential component to ensuring both rights and safety in online environments, on both an individual and systemic level.

On an individual level, approaches to online safety that undermine individuals' privacy—for instance, by way of creating processes that rely on collecting and storing additional personal information, or approaches based upon monitoring and surveillance—can in turn put that person at higher risk of privacy-related harm such as harmful targeted marketing[5] and targeted extreme content and disinformation[6], data breaches and identity theft, and other flow on effects related to the data broker market[7]. This is counterproductive to the goals of increasing online safety.

On a larger scale, Digital Rights Watch's approach centres upon exposing and challenging the structures that give rise to online harms, which we see not as a standalone problem,

[3] European Digital Rights and Principles, *European Commission.* https://digital-strategy.ec.europa.eu/en/policies/digital-principles/

[4] Page 53 of the Issues paper. Parliament of the Commonwealth of Australia, House of Representatives Select Committee on Social Media and Online Safety, 'Social Media and Online Safety' March 2022, [5.75], Social Media and Online Safety – Parliament of Australia (aph.gov.au).

[5] Foundation for Alcohol Research & Education, 'New research shows kids are targeted with alcohol, gambling and junk food ads online,' 4 June 2024, https://fare.org.au/new-research-shows-kids-are-targeted-with-alcohol-gambling-and-junk-food-ads-online/

[6] Samantha Floreani and Lizzie O'Shea, 'We must target the root cause of misinformation. We cannot fact check our way out of this,' *The Guardian,* 16 April 2024, https://www.theguardian.com/commentisfree/2024/apr/26/australia-government-misinformation-bill-social-media-x

[7] ACCC, 'Consumers lack visibility and choice over data collection practices,' 21 May 2024, https://www.accc.gov.au/media-release/consumers-lack-visibility-and-choice-over-data-collection-practices

but as **a symptom of data-extractive business models of digital platforms and advertisers that dominate our digital ecosystem. Protecting privacy is an essential component to meaningfully challenging these business models.**

Legislative efforts that target the symptom (such as removal of content), can address aspects of the problem, but without further intervention, leave the broader structures unaddressed. Digital Rights Watch sees bold privacy law reform as a key way to target the problem at its source: the commercial access to and exploitation of personal information.

Any reform to the Online Safety Act would do well to recognise the essential role that privacy and data protection play in enhancing online safety for both individuals and society at large.

We note that the Privacy Act 1988 is currently under review. In our view, strong privacy reform that favours the rights of users over data extractive business models is central to the goals of enhancing online safety for all people. For too long, Australia's privacy laws have not adequately reflected public expectations, and the lack of enforceable personal privacy rights continues to be a glaring omission in the international context.

Recommendations

1. Prioritise robust reform to Australia's *Privacy Act.*
2. Implement greater restrictions on targeted advertising: prohibit targeted advertising from predatory industries, prohibit targeted advertising directed at children entirely.
3. Implement greater regulation of data brokers.
4. Ensure requirements placed upon digital platforms and other service providers relating to online safety include due regard to upholding privacy of users.
5. Reflect the importance of digital privacy throughout the Online Safety Act, for instance, by not requiring unreasonably privacy-invasive approaches by service providers.

Australia's approach to online safety must shift from content-first to systems-first

Australia's approach to online safety to-date has disproportionately focused on removal of certain online content as the key point of intervention, rather than the underlying systems on which these platforms are built.

When the focus of online safety is solely on the symptoms— namely online abuse and harassment, misinformation, defamation, and the dissemination of sexually explicit material—without also considering the underlying business models, technological realities, legislative landscape, and social norms, the government risks creating additional online harms in its pursuit of mitigation. This is made worse by failing to listen to, or refusing to incorporate, feedback from the most affected communities.

Many aspects of the Online Safety Act fall into the trap of seeking to address the symptoms of harmful business models, rather than getting to the root cause. For instance, the Online Content Scheme incentivises increased automated content moderation and proactive monitoring on digital platforms. In doing so, it seeks to address the harm caused by certain forms of content online, but promotes an approach that (1) increases harm to particular subsets of the community who are subject to over- or under- capture of content and (2) exacerbates, rather than challenges, the data-extractive and surveillance-driven models of Big Tech.

Powers of the eSafety Commissioner

Over the past several years, the role and remit of Australia's eSafety Commissioner has continued to expand. Many of the questions in the Issues Paper clearly frame the desired approach to reform as a matter of providing the eSafety Commissioner with additional or expanded powers. We do not accept such a framing, which assumes that expanding the remit of the eSafety Commissioner is the only or correct way to deal with such harms.

The role of eSafety Commissioner is a difficult and important one, however we are concerned that the Australian government’s default response to any emerging issue to do with the internet is to expand the mandate of office of the eSafety Commissioner. Not all challenges, risks or issues regarding the internet and emerging technologies are well suited to be within the remit of the eSafety Commissioner, and we caution against unreasonable expansion of the role.

The eSafety Commissioner already has extensive powers that aim to compel industry to mandate codes and take down content, though the scope of these has recently been very publicly tested.

Throughout the public debate and consultation process regarding the initial *Online Safety Bill*, Digital Rights Watch and other groups raised concerns regarding the broadly defined powers and their potential to impact the ability to share content that many Australians engage with regularly and consensually (such as sexual content), as well as content that might be used for political accountability, satire or documentation of human rights abuses. While some statements made by the eSafety Commissioner and the former Communications Minister, Paul Fletcher, indicated that the provisions of the Act would not be used in ways that human rights groups fear—that is, to repress freedom of expression, stifle political discourse or deplatform legal sex work—we hold to our initial position: laws that are made can be used, and critics of the bill have every right to assume as much.

It is short-sighted policy making to give unchecked executive power to any one person. The best protection against unintended consequences is for this legislation to be precise in outlining the powers that the office holds and how it may exercise them. This legislation must also be underpinned by proper oversight and accountability. Anything less will mean that the regime contained in the Online Safety Act has enormous potential to limit our public discussions, our capacity to hold the powerful to account and our right to use the web with autonomy. Regulatory creep must be kept in check, even if it has not yet had the chance to 'go wrong'.

Recent developments, such as X's refusal to comply with the eSafety Commissioner's request to remove content on a global scale should not be taken as an invitation to expand powers. They offer an opportunity to reflect on the nature of such powers and the appropriate limits that ought to be in place in a well functioning democracy.

This is an important exercise not least because if statutory powers are overly open ended, courts may well exercise their own jurisdiction to address this. We note that the judgement refusing to extend the injunction highlights the gravity of what was being asked for when the eSafety Commissioner attempted to extend her reach beyond Australia. It should serve as a reminder that regulators cannot be granted unlimited powers:

> *40. The policy questions underlying the parties' dispute are large. They have generated widespread and sometimes heated controversy. Apart from questions concerning freedom of expression in Australia,* ***there is widespread alarm at the prospect of a decision by an official of a national government restricting access to controversial material on the internet by people all over the world****. It has been said that if such capacity existed it might be used by a variety of regimes for a variety of purposes, not all of which would be benign. The task of the Court, at least at this stage of the analysis, is only to determine the legal meaning and effect of the removal notice. That is done by construing its language and the language of the Act under which it was issued. It is ultimately the words used by Parliament that determine how far the notice reaches.*[8]

Transparency reporting

Given that the eSafety Commissioner has a broad remit of power that may infringe upon people's human rights, it is essential that there be adequate transparency and accountability mechanisms in place, in order to maintain oversight of the regulator and also measure its efficacy at the task of mitigating online harms.

In our initial consultation with regard to the original Online Safety Bill, Digital Rights Watch advocated strongly for greater transparency and reporting requirements from the eSafety Office.

Transparency reporting by the eSafety Commissioner is largely performed by way of their joint annual reports with the ACMA, and an indexed list of the titles of all relevant files

[8] eSafety Commissioner v X Corp [2024] FCA 499 https://www.judgments.fedcourt.gov.au/judgments/Judgments/fca/single/2024/2024fca0499

created in a six-month period in the central office of that department or agency (noting that most of the files listed on this page are empty/nil).[9]

We note that the eSafety Commissioner prioritises working informally with service providers, as documented in their most recent annual report:

> *"Building trust with regulated entities increases our efficiency and effectiveness. Where appropriate, we work informally with service providers to resolve individual complaints about online content and behaviour. We also consider any systemic online safety problems these complaints may uncover."*[10]

It seems that upward of 70% of content removal through the various schemes are removed by way of informal requests.[11] While this may prove to be the most timely way to remove content, it also means that an enormous amount of discretionary content decisions are happening behind closed doors, requiring significant cooperation from industry.

Recommendations

6. **Create requirements for more in depth mandatory transparency and accountability reporting**. This ought to include both informal and formal processes, the categories of content take-downs, complaints received (vs actioned and escalated), and blocking notices issued. It should also include the reasoning. Additional detail beyond the current aggregate reporting will allow for public and Parliamentary scrutiny over the ultimate scope and impact of the powers contained in the Act.

[9] eSafety Commissioner, 'Accountability Reporting,' accessed 20 June 2024, https://www.esafety.gov.au/about-us/corporate-documents/accountability-reporting

[10] Australian Communications Media Authority and eSafety Commissioner, Annual Report 2022-23, page 180, https://www.esafety.gov.au/sites/default/files/2023-10/ACMA-and-eSafety-Commissioner-annual-report-2022-23.pdf?v=1718688486812

[11] According to the eSafety Commissioner's latest annual report, informal requests resulted in 87% of material removed under the image based abuse scheme, 77% removed through informal requests under the adult cyber abuse scheme, and 84% removed through informal requests under the child cyberbullying scheme. Australian Communications Media Authority and eSafety Commissioner, Annual Report 2022-23, https://www.esafety.gov.au/sites/default/files/2023-10/ACMA-and-eSafety-Commissioner-annual-report-2022-23.pdf?v=1718688486812

The mechanism for the development of Codes and Standards risks the abrogation of democratic policy making

We remain concerned about the approach set out under the Online Safety Act that allows for the creation of Codes, either by industry or by industry upon invitation of the eSafety Commissioner. Such an approach to regulation is highly resource intensive and risks the abrogation of democratic oversight over rule-making.

In an environment where the resources of civil society are constrained, and industry faces no such limitations, there is a real risk that this process can become a de facto form of government and industry co-regulation. Given the documented bad behaviour by digital platforms, and a public mandate to regulate them, we do not consider such a situation to be justifiable.

We appreciate that this model of regulatory rule making brings with it flexibility and responsiveness, which are both important in a field which is subject to rapid change and technological development. However, this ought not be prioritised above accountability.

Concerns regarding the development of the class 1A and 1B Industry Standards

A pertinent example of how such broadly defined powers can risk abrogation of democratic policy making is the development of the Class 1A and 1B Industry Standards.

Part 9, Division 7 of Online Safety Act provides for industry bodies to develop new codes to regulate 'class 1' and 'class 2' material upon request from the eSafety Commissioner. Following development, eSafety can either register the codes or develop industry standards should the proposed codes fail to meet the requirements.

From September 2022 onwards, the steering group of industry associations developed eight sets of codes to regulate access to Class 1A and 1B content under Australia's classification scheme. In June-September 2023, the eSafety Commissioner registered six out of the eight codes.[12] The eSafety Commissioner declined to register the *Relevant Electronic Services Code* and the *Designated Internet Services Code* and instead moved to develop mandatory and enforceable Industry Standards.

The eSafety Commissioner stated that the decision to decline the codes because "they did not meet the statutory requirements for registration because they did not contain appropriate safeguards for users in Australia."[13]

It was well understood throughout the consultation process for the codes that the eSafety Commissioner expected the inclusion of broad proactive detection provisions—often

[12] The codes registered were: (1) Social Media Services Code, (2) Apps Distribution Services Code, (3) Hosting Services Code, (4) Internet Carriage Services Code, (5) Equipment Code, and following further adjustments to accommodate for developments in generative AI, (6) Internet Search Engine Services Code.

[13] eSafety Commissioner, 'Industry codes and standards', accessed 20 June 2024 from: https://www.esafety.gov.au/industry/codes

referred to as “client-side scanning”.[14] The industry bodies opted not to include proactive detection requirements that would require services to undermine their end-to-end encryption, nor the proactive scanning of individual’s personal online file storage, which ultimately fell foul of the eSafety Commissioner’s position on what constitutes “appropriate safeguards”. This was made clear in the summary of reasons to decline to register the codes, which noted specifically that the key concern was that the codes did not require encrypted services to detect and remove certain material,[15] and that they did not require end-user managed hosting services (such as online file storage) to scan for and remove known child sexual abuse material and known pro-terror material.[16]

The following draft Industry Standards developed by the eSafety Commissioner then included broad proactive detection provisions. Forty local and international organisations and over 560 members of the public called upon the eSafety Commissioner in response, calling for the protection of privacy, digital security and end-to-end encryption.[17] The final Industry Standards are yet to be tabled.

We strongly encourage the committee to review our Submission to the eSafety Commissioner in response to the draft Industry Standards, which details the significant privacy, digital security and surveillance risks of proactive detection systems and why such an approach should not be taken.[18]

[14] 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.# 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.

For more details on proactive detection and client side scanning, we recommend reviewing our submission to the eSafety Commissioner in response to the draft industry standards for class 1A and 1B material, available here:
https://digitalrightswatch.org.au/2024/01/08/submission-online-safety-standards/

[15] eSafety Commissioner, ‘Summary of Reasons – Relevant Electronic Services Code,’ 31 May 2024, accessed 20 June 2024 from:
https://onlinesafety.org.au/wp-content/uploads/2023/06/eSafety_summary_Relevant_electronic_service_providers.pdf

[16] eSafety Commissioner, ‘Summary of Reasons – Designated Internet Services Code,’ 31 May 2024, accessed 20 June 2024 from:
https://onlinesafety.org.au/wp-content/uploads/2023/06/eSafety_summary_Designated_internet_service_providers.pdf

[17] Digital Rights Watch, ‘Local and international organisations urge Australia’s eSafety Commissioner against requiring the tech industry to scan users’ personal files and messages,’ Open Letter, 20 December 2023
https://digitalrightswatch.org.au/2023/12/20/esafety-joint-letter/

[18] Digital Rights Watch, ‘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
https://digitalrightswatch.org.au/2024/01/08/submission-online-safety-standards/

Crucially, 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 inappropriate for the eSafety Commissioner, by way of industry standards, to attempt to extend the obligations beyond the legislative intent reflected in the Act—this would represent a serious overreach of eSafety power without a clear public mandate to do so.

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.

Online content scheme

We note that the *Online Safety Act* relies heavily upon the National Classification Code in determining the classification of content that may be subject to removal notices, as well as the approaches expected or required by way of codes, standards and the Basic Online Safety Expectations.

We further note that the National Classification Code has long been criticised for being outdated and overly broad. We submit that it is not appropriate to force the Classification Code to be used in an online context. When developed, the Code referred to very different social and cultural contexts and forms of media.

Recommendation

7. Prioritise the review of the National Classification Code.

Statutory duty of care approach

In principle, we agree with the intent of introducing a statutory duty of care for platforms; to flip the burden of responsibility of maintaining safety and security from individuals to corporations who require a social licence to operate in Australia. While we do not see it as a total solution to the problems that the online safety regime is seeking to address, if implemented well it can contribute to improving the situation.

There is an enormous imbalance of power between large technology companies, and the people who use their products and services. Individuals are often compelled to hand over more personal information than is necessary for the business operations of a platform, they disproportionately suffer the consequences of data breaches, and are subject to manipulation tactics, to name a few ways in which this power dynamic materialises.

A systemic approach to online safety must address the problematic business model of online platforms, which is fundamentally designed to capture and monetise the attention and behavioural futures of individuals on such platforms. These businesses incentivise the proliferation of content, which makes a notice-and-takedown approach to platform regulation impractical at scale.

As such, insofar as the intent of a duty of care is to create a systemic approach to online platform regulation as opposed to a content-based approach, and to address the underlying systems upon which content is created and shared, we are supportive.

However, should the approach be simply to transform the existing Basic Online Safety Expectations into a mandatory set of requirements with penalties—as is suggested by the Issues paper—we are concerned that this may not deal with the complexity of the task at hand. For example, the challenge of demonstrating causation between a harm suffered by an alleged breach of the statutory duty of care is not insignificant, and will require careful thinking to ensure the mechanism is designed fairly and effectively, with respect to human rights. The subject of a statutory duty of care is subject to very active debate in the legal literature, with much contention regarding the efficacy and ability to operationalise such an approach. We suggest that the Australian government engage further with legal scholars before rushing to enhance the Basic Online Safety Expectations, which in their current state are a flawed foundation to build upon.

A singular duty of care is more favourable than several "duties of care" as implemented in the UK. A duty of care encourages a systemic approach to platform regulation, and away from playing "whack-a-mole" with a notice-and-takedown approach. We agree with the point made by Reset, that the implementation of several duties of care "moves the regulation away from a focus on the systems and back into specifying particular types of content."[19]

A duty of care must be carefully implemented if it is to be effective and avoid unintended consequences. It will inevitably involve placing even more reliance upon large, powerful and often monopolistic technology companies to proactively define, determine and identify what is and is not safe and appropriate. This ought to be done in ways that are transparent and rights-respecting. We think it is particularly important to avoid over-capture, increased marginalisation and censorship of particular people, communities and content.

Age assurance

Digital Rights Watch has long raised concerns regarding the use of age assurance technology—including both age verification and age estimation—to restrict access to particular content such as online pornography, or entire digital platforms or online services.

[19] Reset Australia, 'A duty of care in Australia's Online Safety Act,' Policy Briefing, April 2024, https://au.reset.tech/uploads/Duty-of-Care-Report-Reset.Tech.pdf

Our core concerns are summarised as follows:

(1) age verification creates significant privacy and digital security risks that represent a disproportionate human rights infringement when balanced against the purported benefits;

(2) age verification systems and programs have significant problems of implementation and workability, including issues of bias, accuracy and the ability to easily bypass systems by way of a VPN, or manipulate some age estimation tools with ageing filters.

Age verification is rife with significant privacy and digital security risks, as well as critical effectiveness and implementation issues. Age verification is privacy-invasive, which undermines the objective of reducing online harm. Most forms of age verification require the provision of additional personal information to be effective. Incentivising companies, third parties, and government agencies to collect, use and store additional personal information to conduct age verification creates additional privacy and security risk, which in turn can exacerbate online harms.

Recent research into age estimation tools—that is, tools that attempt to estimate or infer a user's age based on data inputs such as biometrics by way of a facial scan—shows that such technologies are unreliable, and have a racial and gender bias.[20] Other research has investigated the use of age estimation video surveillance in gambling establishments. When the developers of the age estimation tool were interviewed they admitted that it was of limited efficacy in detecting people under the age of 18. Researchers found that the age estimation tool was "performative in nature", ultimately relying upon humans to do the actual work of age verification.[21] Recent reporting has also documented the ease with which it is possible to bypass such tools.[22] In the eSafety Commissioner's own research from 2023, young people expressed concern regarding age assurance's limited efficacy, as well as privacy and security issues.[23]

We've had this debate before

Digital Rights Watch was pleased to see the sensible decision from the Australian government not to move ahead with the trial for age verification for online pornography in August 2023.[24] This decision was based on immaturity of the technology, privacy, digital

[20] Stardust, Z., Obeid, A., McKee, A., & Angus, D. (2024). Mandatory age verification for pornography access: Why it can't and won't 'save the children'. Big Data & Society, 11(2). https://doi.org/10.1177/20539517241252129

[21] O'Neill, C., Selwyn, N., Smith, G., Andrejevic, M., & Gu, X. (2022). The two faces of the child in facial recognition industry discourse: biometric capture between innocence and recalcitrance. Information, Communication & Society, 25(6), 752–767. https://doi.org/10.1080/1369118X.2022.2044501

[22] Cam Wilson, 'I tricked a selfie AI age-verification demo into letting a child 'buy' a knife,' *Crikey,* 14 June 2024, https://www.crikey.com.au/2024/06/14/selfie-ai-age-verification-tool-filter-trick/

[23] eSafety Commissioner, 'Questions, doubts and hopes: Young people's attitudes towards age assurance and the age-based restriction of access to online pornography,' September 2023, https://www.esafety.gov.au/sites/default/files/2023-08/Questions-Doubts-and-Hopes.pdf

[24] Australian Government, 'Australian Government response to the Roadmap for Age Verification,' *Department of Infrastructure, Transport, Regional Development, Communications and the Arts,* 30

security and implementation issues. However, this decision was later reversed in response to pressure to act on increased misogyny and violence against women.[25]

Using a human rights approach, the privacy invasion that comes with age verification may be justified, if it is a reasonable, necessary and proportionate means for a legitimate purpose.[26] However, to date, there is no compelling evidence that this is the case.

First, many young people access adult content on social media sites, rather than dedicated pornography sites. Mandating age assurance technology misses the mark, and to extend the proposal to apply to all social media sites would be a serious overreach. Further, the research is complex and at times conflicting when it comes to connecting mainstream pornography with gender-based violence.[27] Making policy decisions that impact human rights based on assumptions and unclear evidence about harm is not appropriate.

Second, in 2021 the Coalition government drafted the *Online Privacy Bill,* which would have required platforms to verify the age of their users and obtain parental consent for those under the age of 16.[28] Research conducted afterwards found that parents and carers were initially enthusiastic about the prospect of stronger laws to help protect their children, but this quickly deflated when they learned of the measures that would be needed to actually enforce it, such as providing identity documents to platforms or third parties, increased app tracking and monitoring, and ongoing age verification processes such as face scans.[29]

These issues with age verification make it an unviable mechanism and we argue it should not be mandatory on social media platforms or online pornography websites. Further

August 2023, https://www.infrastructure.gov.au/department/media/publications/australian-government-response-roadmap-age-verification

[25] Prime Minister of Australia, 'Tackling online harms,' Media Release, 1 May 2024, https://www.pm.gov.au/media/tackling-online-harms

[26] Lizzie O'Shea, 'Let's get this right and avoid knee-jerk decisions on misogyny and men's violence against women,' *Crikey,* 2 May 2024, https://www.crikey.com.au/2024/05/02/misogyny-violence-against-women-pornography-privacy-age-restriction/

[27] Lim, M.S.C., Carrotte, E.R. & Hellard, M.E. (2016). The impact of pornography on gender-based violence, sexual health and well-being: what do we know? *Journal of Epidemiol Community Health*, 70, 3–5. Mestre-Bach, G., Villena-Moya, A., & Chiclana-Actis, C. (2024). Pornography Use and Violence: A Systematic Review of the Last 20 Years. *Trauma, Violence, & Abuse*, *25*(2), 1088-1112. https://doi.org/10.1177/15248380231173619.

[28] Australian Government, 'Online Privacy Bill Exposure Draft,' Attorney-General's Department, 25 October 2021, https://consultations.ag.gov.au/rights-and-protections/online-privacy-bill-exposure-draft/

[29] University of Sydney, 'New study reveals teenagers' social media use and safety concerns,' 6 October 2023, https://www.sydney.edu.au/arts/news-and-events/news/2023/10/06/new-study-reveals-teenagers-social-media-use-and-safety-concerns.html

investment in prevention of gendered violence should address systemic misogyny embedded in society, which has been well-established by decades of research.

Recommendation

8. Do not create requirements for mandatory age verification.

Protecting end-to-end encryption

In the context of securing messaging via end-to-end encryption the issues paper states:

> *"In 2022, the House of Representatives Select Committee on Social Media and Online Safety concluded that while privacy concerns are critical to the rights of all internet users, those issues did not 'outweigh the fundamental issue of ensuring safety in online environments.'"*

DRW has repeatedly expressed concern with the framing of encryption as an inhibitor to safety, which runs counter to the established consensus in the cybersecurity industry who view encryption as vital to facilitate safety. Encryption is essential for all businesses, individuals, and digital security at a national level. Encryption facilitates the security of our online activities, protects data from potential cybercriminals, enables secure online transactions, and maintains the privacy and security of our online communications, including those of children. For example, encryption plays a crucial role in preventing malicious actors from accessing networked devices, including tapping into users' webcams or baby monitors.[30]

Undermining encryption is a threat to online safety.

While encryption may pose a challenge to some criminal investigations, it also provides ordinary Australians with digital security, and protects them from arbitrary surveillance by malicious actors and cybercrime (e.g. 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 strengthens a regressive surveillance agenda at the expense of everyone's digital security. It is essential that any reform to the Online Safety Act does not create a way to compel providers to restrict or weaken their use and application of encryption across their platforms—including methods that side-step it, such as 'client-side scanning'.

[30] Amy Wang, ''I'm in your baby's room': A hacker took over a baby monitor and broadcast threats, parents say,' *The Washington Post,* 20 December 2018, https://www.washingtonpost.com/technology/2018/12/20/nest-cam-baby-monitor-hacked-kidnap-threat-came-device-parents-say/

Any weakening of encryption would undermine the security of Australians' services, jeopardising the safety of the millions of people who rely on them each day. Prompting services to weaken, undermine, or otherwise bypass encryption threatens the digital security of Australia at a national level by introducing security weaknesses into everyday services used by Australians. Further, encryption is essential for the protection of vulnerable groups, including LGBTQ+ persons[31] and survivors of domestic violence, who rely on encryption to protect the sharing of information about safe relocation, the integrity of digital evidence, and to guard against unauthorised access to survivors' details or communications.[32]

Recommendations

9. Ensure the use of encryption (including end-to-end encryption) is actively protected in the *Online Safety Act.*
10. Re-frame the approach to encryption to reflect the importance of encryption to the safety and security of everyone. Services ought to be encouraged, not dissuaded, from having robust digital security measures in place.

Decentralised platforms are not the problem

We note that the issues paper calls out regulatory challenges posed by decentralised platforms: "Within the current regulatory framework, decentralisation makes it more difficult to hold users responsible for illegal or harmful content and conduct." Further, the issues paper states:

> *Decentralised services are typically created for the purpose of being censorship resistant. However, this raises concerns around the ability to moderate or regulate decentralised services or platforms, potentially increasing the vulnerability of marginalised individuals and groups or creating space for criminal activities or users who have been removed from mainstream services.*

This framing is misleading and dangerous. Decentralised platforms can be another form of social media that allows people to find community and provides opportunities for

[31] LGBT Tech, 'Encryption - Essential for the LGBTQ+ Community,' 18 October 2021, https://www.lgbttech.org/post/2019/11/22/lgbt-tech-release-encryption-one-sheet

[32] Internet Society, 'Understanding Encryption Fact Sheet: The Connections to Survivor Safety,' 18 December 2020 https://www.internetsociety.org/resources/doc/2020/understanding-encryption-the-connections-to-survivor-safety/

individual expression, often outside the harms of surveillance capitalism. They can also be an example of how online communities can self-manage and moderate their communities according to specific contextual and cultural rules and norms, rather than a top-down approach as on large mainstream platforms.

Further, as identified in the issues paper, decentralised platforms may provide people “with more power online by reducing reliance on mainstream, centralised servers” which can “provide greater control and information protection to users”.

Generative AI and online safety

The use of AI in relation to online safety is complex and currently lacks clear guidelines and regulation in Australia.

AI and machine learning are often touted as potential solutions to or remedies for online abuse. During the Women’s World Cup hosted in Australia in 2023, FIFA engaged ‘Threat Matrix’ to monitor the social media of players, coaches and staff for online abuse and harassment using AI.[33] However, the effectiveness of these automated ‘detect and delete’ systems has been called into question[34] as they are often opaque and lack a nuanced understanding of communicative norms. These approaches also only monitor publicly available comments and posts, meaning that direct messages and emails can only be monitored by allowing a third party to seriously invade the privacy of a particular user’s social media and online accounts.

Relatedly, and as discussed in our submission to the Online Safety Draft Industry Standards, “[h]igh quality, automated detection of [Child Sexual Abuse Material and “pro terror” content] is extremely difficult, especially when the scope of target content is broad, not strictly defined, or difficult to assess without additional context.” The AI and machine learning software currently used to automatically scan and detect content is often over-zealous and under-effective, making it unreliable and risky for a range of human rights abuses.[35]

At the same time, we are concerned with the rising use of generative AI to create “deepfakes”. Deepfakes are images, videos and audio files that look and sound real but have been either manipulated or fabricated using AI. Recent high-profile examples have included Joe Biden and Taylor Swift, but the risks for everyday Australians remains high

[33] FIFA, ‘Tackling online abuse at the FIFA Women's World Cup 2023,’ FIFA Social Media Protection Service, 9 December 2023, https://inside.fifa.com/social-impact/campaigns/no-discrimination/fifa-social-media-protection-service/fifa-womens-world-cup-australia-new-zealand-2023

[34] Gorwa, R., Binns, R., & Katzenbach, C. (2020). Algorithmic content moderation: Technical and political challenges in the automation of platform governance. *Big Data & Society*, *7*(1). https://doi.org/10.1177/2053951719897945

[35] For further information, please see our submission to the Online Safety Draft Industry Standards https://digitalrightswatch.org.au/2024/01/08/submission-online-safety-standards/

when deepfake sexually explicit material is an increasing social problem in Australia.[36] Further, deepfakes pose a risk to democracy and democratic processes in how images and audio can be manipulated and shared through easily-accessible online tools.[37] Our recommendation is that the Committee take seriously the threats posed by deepfakes while also keeping in mind Australians' rights to privacy and the limitations of legislation and technology in addressing the risks and harms of deepfakes.

The responses to address online safety and AI must be robust and long-term, tackling the social and cultural dimensions of inequity that leads to harmful, dangerous and inappropriate conduct using AI, and not simply relying on legal frameworks to address long-standing social issues.

The question of AI model training data provenance

Many of the harms that arise from AI and Automated Decision-Making (ADM) stem from inappropriate collection and use of personal information. The models that enable the creation of non-consensual deep fake sexually explicit material were trained using images of real people. The models that enable people to generate and disseminate harassing text and imagery were trained on immense amounts of personal information. Any discussion of attempting to mitigate the harms caused by generative AI must take seriously that the models used by these systems, and the companies that have built them, have done so off the back of decades of unbridled privacy-invasion and rampant generation and collection of data.

As such, robust privacy regulation can go a long way toward mitigating privacy-related harms caused by AI. While robust privacy regulation cannot solve or mitigate *all* risks raised by AI technologies, strengthening the Privacy Act will play a fundamental role in upholding people's right to privacy and the protection of their personal information as AI becomes increasingly commonplace.[38]

[36] See: Jeannie Marie Paterson,''Picture to burn': The law probably won't protect Taylor (or other women) from deepfakes,' The University of Melbourne, 8 February 2024, https://pursuit.unimelb.edu.au/articles/picture-to-burn-the-law-probably-won-t-protect-taylor-or-other-women-from-deepfakes; 'Police investigate fake nude photos of about 50 Bacchus Marsh Grammar students being circulated online,' *ABC News,* 11 June 2024, https://www.abc.net.au/news/2024-06-11/bacchus-marsh-grammar-explicit-images-ai-nude/103965298

[37] Andrew Ray, 'Disinformation, Deepfakes and Democracies: The Need for Legislative Reform,' *UNSW Law Journal,* 2021, https://www.unswlawjournal.unsw.edu.au/article/disinformation-deepfakes-and-democracies-the-need-for-legislative-reform

[38] For further detail on the ways the Privacy Act should be reformed, see Digital Rights Watch Submission to the Attorney-General's Department on the 2022 Report regarding the review of the *Privacy Act 1988*, 31 March 2023. Available at: https://digitalrightswatch.org.au/2023/04/03/submission-privacy-act-review-report/