Human Rights Law Centre

Submission to the Statutory Review of the *Online Safety Act 2021*

5 July 2024 **Alice Drury | David Mejia-Canales | Joel Harriss**

Human
Rights
Law
Centre

The Human Rights Law Centre
Level 17, 461 Bourke Street
Melbourne VIC 3000

T: + 61 3 8636 4450
F: + 61 3 8636 4455
[redacted]
W: www.hrlc.org.au

The Human Rights Law Centre

The Human Rights Law Centre uses strategic legal action, policy solutions and advocacy to support people and communities to eliminate inequality and injustice and build a fairer, more compassionate Australia. We work in coalition with key partners, including community organisations, law firms and barristers, academics and experts, and international and domestic human rights organisations.

The Human Rights Law Centre acknowledges the people of the Kulin and Eora Nations, the traditional owners of the unceded land on which our offices sit, and the ongoing work of Aboriginal and Torres Strait Islander peoples, communities and organisations to unravel the injustices imposed on First Nations people since colonisation. We support the self-determination of Aboriginal and Torres Strait Islander peoples.

Follow us at https://twitter.com/humanrightsHRLC

Join us at www.facebook.com/HumanRightsLawCentreHRLC/

Contents

1. Introduction

The Human Rights Law Centre thanks the Department of Infrastructure, Transport, Regional Development, Communication and the Arts (**the Department**) for the opportunity to make a submission to the Statutory Review of the *Online Safety Act 2021* (Cth) (**the Act**).

Australia has been a pioneer on digital platform regulation. Australia was the first country to legislate for online safety and to appoint an online safety commissioner. Additionally, Australia led the way in legislating negotiations between digital platforms and news outlets, and insights from the Australian Competition and Consumer Commission's *Digital Platforms Inquiry Final Report* continue to shape policy development locally and globally.[1]

However, despite this strong foundation, Australia risks being left behind if it does not continue to prioritise strong and effective regulation that can build safer online spaces for all.

The emergence of internet technologies has brought many human rights benefits, like platforming diverse voices, new freedoms of association, and unprecedented access to information. However, they have also facilitated the rapid spread of harmful content, hate speech, misinformation and disinformation which can distort public debate or harm electoral processes.

Large digital platforms have a profound influence on public discourse: they shape the information people encounter, influencing their decisions and beliefs. Their business models also prioritise serving users with polarising, emotive, often anger-inducing material which drives user engagement and therefore maximises platforms' profit.

Current regulations provide limited accountability for the systemic harm caused by digital platforms' products and business models. It is crucial for Australia to step up its regulatory efforts to address these challenges and safeguard the benefits of digital advancement for all.

The statutory review of the Act is an opportunity to implement a comprehensive regulatory framework for digital safety. A failure to do so will only create a void that private entities will exploit for their benefit. Allowing digital platforms to operate within a widening policy gap poses significant risks, as profit-driven platforms will inevitably prioritise products that are lucrative over those that promote societal well-being.

The Human Rights Law Centre supports a form of regulation that is not premised on industry self-regulation or co-regulation. Instead, we recommend a systemic and comprehensive regulatory regime modelled on the European Union's *Digital Services Act* (**DSA**). The DSA is designed to achieve digital platform transparency and accountability to address the significant human rights implications of their operations.

The Human Rights Law Centre has previously made submissions to the *Communications Legislation Amendment (Combatting Misinformation and Disinformation) Bill 2023* (Cth) and the Senate's Standing Committee on Economics' Inquiry into the Influence of Large Online Platforms. Those submissions, to varying extents, addressed similar matters that the Department is currently seeking feedback on. We include those submissions as part of our feedback into this process.

[1] Reset.Tech Australia, 'Digital Platform Regulation Green Paper', (Discussion paper, April 2024) 1 <https://au.reset.tech/uploads/Digital-Platform-Regulation-Green-Paper.pdf>.

2. Recommendations

The statutory review of the Act presents an opportunity to begin to harmonise the various regulations and policies aimed at ensuring online safety for all.

The Human Rights Law Centre recommends that the government adopts Reset.Tech Australia's risk-based approach to online safety regulation, grounded in human rights law. This approach should incorporate these five basic principles:

1. **A duty of care**: a legislated overarching duty of care which would place broad obligations on digital platforms to ensure user safety in systemic ways.

2. **Risk assessment**: platforms should be required to assess all of their systems and elements for risks, while also helping platforms realise their duty of care.

3. **Risk mitigation**: platforms must be required to implement reasonable steps to mitigate against each of the risks identified in their risk assessments.

4. **Transparency measures**: platforms should be required to independently audit their risk assessments and make them publicly available. Platforms should also be required to give vetted researchers and civil society organisations access to platforms' public-interest data.

5. **Accountability measures**: which prioritise changes to platform systems over content take-down and significant penalties for non-compliance.

These five principles are discussed further, below.

3. Australia's regulatory approach to online services, systems and processes

The Act currently features two schemes aimed at preventing online harm: The Online Content Scheme and the Basic Online Safety Expectations (**BOSE**).

3.1 The Online Content Scheme

Under the Online Content Scheme, the eSafety Commissioner (**Commissioner**) has the authority to register codes created by industry bodies or associations that represent various sectors of the industry. Once registered by the Commissioner, these codes and standards become mandatory and enforceable.

The focus of these standards is on illegal or restricted material as defined by the National Classification Code. The Commissioner is empowered to moderate content that involves, among other things, cyber-bullying of children, cyber abuse of adults, image-based abuse and abhorrent violent material such as acts of terrorism.

The Act gives the Commissioner certain powers with respect to content moderation and removal. The Commissioner is also empowered to investigate complaints made under public complaints schemes.

The Commissioner's content moderation powers primarily involve notice-and-take-down powers, whereby the Commissioner issues a removal notice to a platform,[2] which must be adhered to within 24 or 48 hours, otherwise a penalty may be imposed.[3]

3.2 The BOSE

The BOSE define the minimum safety standards that the Government expects online service providers to meet. They serve as a benchmark for digital services to proactively safeguard the Australian community from abusive conduct and harmful content online.[4]

Whereas content moderation is a reactive approach to minimising online harms, the BOSE are intended to address upstream risks before harms materialise. However, digital service providers are not required to implement the BOSE, similarly the Commissioner is not empowered to enforce these expectations, nor dictate what the expectations ought to be.

Ultimately, the BOSE are largely voluntary and unenforceable.[5]

---

[2] *Online Safety Act 2021* (Cth), Part 9.
[3] *Online Safety Act 2021* (Cth), Part 9.
[4] Commonwealth of Australia, *Online Safety (Basic Online Safety Expectations) Determination 2022,* 20 January 2022.
[5] Reset.Tech Australia, B*riefing: Can safety standards be enforceable?*, 5 February 2024, available at: https://au.reset.tech/news/briefing-can-safety-standards-be-enforceable/.

4. Centring of human rights

4.1 Human rights are being undermined online

Digital platforms have brought unprecedented opportunities for global communication and information sharing. However, alongside these benefits, there has been a concerning rise in online activities that undermine fundamental human rights.

For example, the amplification of incorrect or misleading information about elections or referenda can question the integrity of free and fair elections. Hate speech fuels violence and poses threats to people's lives and wellbeing. Children and young people are too often exposed to harmful, abusive and exploitative content. Moreover, coordinated disinformation campaigns have jeopardised the right to health, exacerbating challenges during the COVID-19 pandemic.

The growing influence of powerful algorithms, the proliferation of artificial intelligence technologies, extensive user data harvesting and the addictive design features of social media platforms and other digital products exacerbate these risks.

International human rights organisations have consistently raised alarms about the proliferation of harmful content online, emphasising the urgent need to apply a human rights framework to mitigate these dangers.[6] Similarly, governments worldwide are increasingly recognising the necessity of placing human rights at the core of their regulatory approaches.[7]

Harmful online content has the potential to impact upon the enjoyment of many established human rights, including but not limited to:[8]

(i) Freedom of expression;

(ii) Freedom of thought and conscience;

(iii) Right to information;

(iv) Right to participate in public affairs;

---

[6] See eg United Nations Human Rights Council Resolution (2013) *The safety of journalists* UN Doc A/HRC/RES/39/6 (27 September 2018) 7; UN Secretary-General's High-level Panel on Digital Cooperation (2019), The Age of Digital Interdependence; UN Office of the High Commissioner for Human Rights (2017), 'Freedom of Expression Monitors Issue Joint Declaration on 'Fake News', Disinformation and Propaganda', https://www.ohchr.org/EN/NewsEvents/Pages/DisplayNews.aspx?NewsID=21287&LangID=E; Kate Jones, Online Disinformation and Political Discourse: Applying a Human Rights Framework, Research Paper, November 2019 https://www.chathamhouse.org/sites/default/files/2019-11-05-Online-Disinformation-Human-Rights.pdf 27.

[7] See eg UN Human Rights Council, *Role of States in countering the negative impact of disinformation on the enjoyment and realization of human rights*, 30 March 2022, UN Doc A/HRC/49/l.31/Rev.1; UN Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression, *Disinformation and freedom of opinion and expression*, 13 April 2021, UN Doc A/HRC/47/25; UN Human Rights Committee, *General comment no. 34 – Article 19: Freedoms of opinion and expression*, 12 September 2011, UN Doc CCPR/C/GC/34; UN Committee on the Rights of the Child, *General Comment No. 25 (2021) on Children's Rights in Relation to the Digital Environment* (2 March 2021) UN Doc CRC/C/GC/257. See also: Access Now, Civil Liberties Union for Europe and EDRi, *Informing the disinfo debate: a policy guide for protecting human rights*, (December 2021) https://www.accessnow.org/cms/assets/uploads/2021/12/Informing-the-disinfo-debate-report.pdf; Amnesty International, *A human rights approach to tackle disinformation*, 14 April 2022, https://www.amnesty.org/en/wp-content/uploads/2022/04/IOR4054862022ENGLISH.pdf

[8] The Global Online Safety Regulators Network has recognised that human rights abuses can occur through the "production, distribution, and consumption of illegal and harmful online content": Global Online Safety Regulators Network, *Position Statement: Human Rights & Online Safety Regulation* September 2023, 2-3.

(v) Right to vote;

(vi) Right to privacy;

(vii) Right to health;

(viii) Right to live free from discrimination; and

(ix) Rights of the child to safety.

Suggestions by some stakeholders that digital platforms should not be regulated (or regulated to a minimal extent) due to concerns regarding the impacts to the right to freedom of expression, fundamentally mischaracterise this right and its interaction with other fundamental human rights.

A misguided interpretation of the right to free speech (including by free speech absolutists) has been weaponised to avoid accountability for the harms caused by abuses *of* the right to free speech. The Australian Government must not be dissuaded from pursuing a comprehensive regulatory regime by such arguments.

The right to free speech ought to be understood in relation to other fundamental rights, including the right to freedom of thought and conscience, right to information, right to participate in public affairs and the right to vote, among others.

4.2 Responsibility for upholding human rights

Australia, as a party to the major international human rights treaties,[9] is required to protect the human rights of all people in Australia's jurisdiction. These obligations do not diminish or disappear online. The United Nations' Human Rights Council has issued resolutions emphasising that the rights that people enjoy offline must also be protected online.[10]

In addition to the role of governments in upholding rights, industry also has a role to play in ensuring that human rights are upheld. The United Nations' *Guiding Principle on Business and Human Rights* provides a framework to ensure that the activities of all business, regardless of their operations, size, location or ownership structure, are compliant with human rights.[11]

These principles require businesses to respect human rights and remedy any adverse impacts that their operations may have created or contributed to.[12]

At a minimum, this should involve regular human rights impact assessments by businesses of their products, operations and policies and due diligence processes aimed at identifying, preventing or mitigating actual or potential adverse impacts on human rights caused by their activities.[13]

While business do not have the same legal duties as nation states, they still have obligations to respect human rights and abide by human rights regulation. This involves, for example, businesses recognising that

[9] Australia is a signatory to the International Covenant on Civil and Political Rights; International Covenant on Economic, Social and Cultural Rights; International Convention on the Elimination of All Forms of Racial Discrimination; Convention on the Elimination of Discrimination against Women; Convention against Torture and Other Cruel, Inhuman or Degrading Treatment or Punishment; Convention on the Rights of the Child; and Convention on the Rights of Persons with Disabilities.
[10] UN Human Rights Council Resolutions (2012-2018), *The promotion, protection and enjoyment of human rights on the Internet*, UN Doc A/HRC/RES/38/7 (5 July 2018), A/HRC/RES/32/13 (1 July 2016), A/HRC/RES/26/13 (26 June 2014), A/HRC/RES/20/8 (5 July 2012).
[11] United Nations Office of Human Rights, *Guiding Principles on Business and Human Rights*, HR/PUB/11/04, 2011.
[12] Ibid at 13
[13] UN Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression, *Disinformation and freedom of opinion and expression*, 13 April 2021, UN Doc A/HRC/47/25, 14-15.

Australian governments do have a duty to protect all individuals in Australia from human rights violations and therefore agreeing to be bound by any rules that governments enact to give effect to this duty.[14]

Respecting the human rights of people in Australia should be the price of doing business here.

4.3 Defining 'online harm'

While there is some conjecture as to how to define 'online harm', the World Economic Forum's *'Toolkit for Digital Safety Design Interventions and Innovations: Typology of Online Harms*'[15] (**Typology of Online Harms**) report is frequently cited as helping to define the general nature of online harms. The Typology of Online Harms is useful in providing a framework with which to identify and categorise online harms. This submission adopts this framework for understanding the concept of online harm.

The Typology of Online Harms recognises content, contact and conduct risks of online interactions.[16] Content harms include harms sustained in content production, distribution and consumption.[17] Contact harms are those harms that can occur as a result of online interactions with others, whereas conduct harms are harms incurred through an individual user's behaviour which is facilitated by technology and digital platforms.[18]

Harm in digital spaces is rampant,[19] with victims experiencing a range of significant and lasting impacts.[20] Certain groups, including children, women, Aboriginal and Torres Strait Islander peoples, individuals from culturally and linguistically diverse backgrounds and people who identify as LGBTQIA+, are more likely to experience harmful online content and conduct like harassment and hate speech.[21]

The Typology of Online Harms rightly acknowledges the role that users play in the production, distribution and consumption of content, but is also cognisant of the ways in which technology facilitates behaviour that is conducive to harm.[22]

Often, the design of myriad systems and processes adopted by digital platforms is what exposes users to harmful online content in the first place, not to mention their ability to amplify harmful material and content. It is imperative, therefore, that the responsibility for content, contact and conduct risks and harms must include the digital platforms.

It follows that the definition of 'online harm' must also include harms resulting from the actions or omissions of the digital platforms, including, for example: the use of addictive design features, recommender systems, data harvesting, insufficient systemic protections against harmful online material, ineffective content moderation systems and opaque mechanisms to report misconduct or abuse to the platforms themselves.

It is against this backdrop that this submission calls for a greater responsibility for online harm to be borne by those bodies that enable online harms to exist and thrive, particularly those with the ability and the capacity to make the relevant changes to mitigate them – the digital platforms.

---

[14] Kate Jones, Online Disinformation and Political Discourse: Applying a Human Rights Framework, Research Paper, November 2019 https://www.chathamhouse.org/sites/default/files/2019-11-05-Online-Disinformation-Human-Rights.pdf at 30.
[15] World Economic Forum, *Toolkit for Digital Safety Design Interventions and Innovations: Typology of Online Harms*, 2023.
[16] Ibid 4.
[17] Ibid 5.
[18] Ibid.
[19] Social Media and Online Safety, *'Social Media and Online Safety'*, Report of the House of Representatives Select Committee on Social Media and Online Safety, (March 2022), 11 - 12.
[20] Ibid.
[21] Ibid 29-44.
[22] World Economic Forum, *Toolkit for Digital Safety Design Interventions and Innovations: Typology of Online Harms*, 2023, 12.

5. Elements of a comprehensive digital regulatory framework

Consultation question 7: Should regulatory obligations depend on a service provider's risk or reach?

Australia's regulatory regime has thus far left us with strong regulation of the most egregious types of online content, but weak regulation of the systems and processes which enable the dissemination of harmful online content in the first place.

Best practice regulatory regimes use a risk-based approach to assess and classify technologies based on their potential harm. This means making digital platforms identify and prioritise the issues that pose the greatest potential harm to users and allocate their resources to manage or reduce those risks.

Best practice regulation also requires that platforms are transparent about the risks for harm that their products create and how they will reduce or eliminate those risks.

One such regulatory regime proposal is Reset.Tech Australia's "*Five Pillars for Addressing Systemic Risks*" (**Five Pillars**).[23]

The Five Pillars provide a structured and comprehensive approach to tackling the complexities of digital platform regulation. By focusing on systemic risks, the Five Pillars aims to establish a robust framework that promotes safety, transparency, and accountability, with a particular emphasis on protecting children and young people.

The "Five Pillars" model encompasses the following:

1. An overarching duty of care;
2. Comprehensive risk assessments;
3. Risk mitigation;
4. Transparency measures; and
5. Accountability measures.

This submission will examine the Five Pillars as a suggested approach to reforming the Act while highlighting their potential to enhance the efficacy of digital regulation and protect the rights and well-being of all.

5.1 An overarching duty of care

While current regulatory models largely focus on digital platform users, a duty of care imposes a proactive obligation on digital platforms to ensure their systems are designed to mitigate for the risks of online harm before it happens.

A "duty of care" is a legal concept which creates a responsibility on entities or individuals to ensure a reasonable standard of care to prevent reckless or avoidable harms to those who are owed the duty. A duty

[23] Reset.Tech Australia, 'Digital Platform Regulation Green Paper', (Discussion paper, April 2024) 1 <https://au.reset.tech/uploads/Digital-Platform-Regulation-Green-Paper.pdf>.

of care encompasses considerations for the type of harm, whether the harm was foreseeable, the role of the duty-bearer and the relationship with the harmed individual or entity.

In practice, a duty of care should involve the following elements:

5.1.1 Requirements for digital platforms to uphold and protect the human rights of their users

Protecting and upholding the human rights of digital platform users is paramount in any regulatory framework.

Digital platforms can have a profound impact on the lives of their users, influencing their access to information, the freedom of expression, their privacy, and security. A robust regulatory regime must prioritise the rights and freedoms of individuals, ensuring that platforms operate in a manner that respects and safeguards these fundamental rights.

This includes implementing measures that prevent the exploitation and abuse of users, protecting their personal data, and ensuring that they have the right to redress and accountability.

5.1.2 Requirements to undertake comprehensive risk assessments to identify and analyse risks stemming from the systems used by digital platforms

A duty of care would require platforms to assess and analyse their systems and processes to determine any risks of harm caused by virtue of their technical design.[24]

While the platforms do not produce the harmful content that appears online, the systems they employ and profit from contribute to the distribution, consumption and amplification of this content.

5.1.3 Requirements to address any identified risks though effective risk mitigation measures

A duty of care would require digital platforms to thoroughly scrutinise their systems and think critically about how these systems may be contributing to harm while incentivising them to take proactive risk mitigation strategies.[25]

For example, recommendation algorithms that amplify harmful content would need to be redesigned to prioritise safety and well-being of users, particularly for children and young people. Privacy settings and data collection practices would need to be more transparent and user-friendly to protect personal information. Additionally, systems for reporting and responding to harmful content would need to be made more accessible and efficient to ensure prompt action and support for affected users.

The imposition of a duty of care would also remove any defence of platform ignorance or non-involvement in the dissemination of harmful content. Accordingly, platforms would need to invest in preventative measures that, as far as foreseeable, minimise the platform's liability for any harm their products cause.

5.1.4 Requirements for platforms to open up their assessment and mitigation measures for scrutiny by third parties

A crucial element of a duty of care must be ensuring that risk assessments and mitigation measures are open to scrutiny by vetted third-party researchers and civil society organisations. Transparency measures ensure that a platforms' adherence to the relevant requirements can be tested and verified, otherwise a platform's compliance with their duty of care could not be guaranteed.

Such transparency would simultaneously provide an additional safeguard and give platforms the benefit of a wider, and more diverse range of perspectives on risk assessments and mitigation. In this respect, the

---

[24] Reset.Tech Australia, *A duty of care in Australia's Online Safety Act* (Policy Briefing), April 2024, 6.
[25] Ibid 9.

claims of platforms can be tested by entities with a wide range of interests, perspectives and experiences, inherently diversifying the regulatory system and enhancing effectiveness.

5.1.5 Mechanisms for redress for harm caused by a breach of the duty of care.

While the Act includes a public-facing complaint mechanism that allows users to report harmful content under certain conditions, this mechanism could be expanded to enable more users and communities to seek redress directly.[26]

For instance, the Act currently permits individuals to lodge complaints about cyberbullying, image-based abuse, and other specific types of harmful content. However, expanding the scope to include broader categories of harm, such as misinformation, hate speech, and coordinated digital harassment, could provide a more comprehensive avenue for redress.

Additionally, simplifying the complaint process and making it more accessible to groups more vulnerable to online harms like non-English speakers, and people with disabilities, would ensure that a wider range of affected individuals and communities can seek help and resolution.

From a resourcing perspective, imposing a duty of care is also a practical approach to ensuring that platforms invest in thorough risk assessment and mitigation strategies. Given that digital platforms, especially large corporations, have substantial resources, they can be required to deploy these resources proactively.

For example, platforms could establish dedicated teams to handle user complaints more efficiently, develop advanced tools to detect and address harmful content swiftly, and create support services for victims of online abuse. By mandating these investments, the duty of care would help minimise harm and provide robust mechanisms for users to obtain redress when breaches occur.[27]

5.2 Comprehensive risk assessments

Requiring digital platforms to assess all their systems and components for risks would incentivise systemic change and help them fulfil their duty of care.

The DSA provides a model risk assessment template for platforms to identify, analyse and assess any systemic risks stemming from their systems. Under the DSA, 'very large platforms' are required to undertake annual risk assessments to identify any significant systemic risks arising from the functioning and use of their services, including their algorithms, recommender systems, content moderation systems, user terms and conditions, advertising systems and data-related practices.[28]

In their risk assessments, very large platforms are required to consider the following systemic risks:[29]

- Actual or foreseeable negative impacts on a range of fundamental rights, including the right to dignity, to respect for private and family life, protection of personal data, freedom of expression and information, the prohibition on discrimination and the rights of the child;

[26] Ibid 10.
[27] Ibid 9.
[28] Regulation (EU) 2022/2065 of the European Parliament and of the Council of 19 October 2022 on a Single Market For Digital Services and amending Directive 2000/31/EC (***Digital Services Act***), Article 34. Under the DSA, escalating obligations and requirements apply to platforms with the greatest influence and resources. *Very Large Online Platforms* are currently defined as those with 45 million or more average monthly users in the EU. Violations of the DSA can result in fines of up to 6% of a platform's annual worldwide turnover for a financial year, representing meaningful incentives for firms with the greatest influence and resources.
[29] *Digital Services Act* (EU), Article 34. By requiring entities to carry out human rights due diligence of this kind, risk assessments can support platforms to meet the standards set out in the UN Guiding Principles on Business and Human Rights, which is the authoritative international standard on the corporate responsibility to respect human rights.

- Any foreseeable negative effects on civic discourse and electoral processes, and public security;
- Any actual or foreseeable negative effects in relation to gender-based violence, the protection of public health and minors, and serious negative consequences to a person's physical and mental well-being; and
- The dissemination of illegal content through their services.

Importantly, the risk assessment approach employed by the DSA extends to the impact of platforms' activities on a range of human rights, and also encompasses risks associated with content that is otherwise legal but which may pose serious societal risks.

One of the most significant threats to digital content regulation is the infringement upon the right to freedom of expression and opinion. In the DSA context, it is significant that risk assessments, which are submitted to and overseen by the European Commission, are required to consider the impact of moderation systems on the right to freedom of expression and opinion.

Where risk assessments undertaken by very large platforms identify systemic risks, these platforms are required to implement "reasonable, proportionate and effective mitigation measures".[30]

The risk assessments and mitigation measures implemented by digital platforms are also subject to annual independent audits at the expense of the platforms,[31] similar to any other regular auditing processes, like financial audits.

The risk-based framework in the DSA is designed to provide a multi-layered approach to ensure that systemic risks and shortfalls in addressing them can be identified and rectified. This model is simultaneously robust, and sufficiently flexible, to ensure online harms are mitigated.

Importantly, this model also ensures that the burden and primary obligation of avoiding and addressing harm is borne by the platforms themselves.

Requiring digital platforms operating in Australia to assess their respective systems for risks would enable them to also realise part of their duty of care. Risk assessments should identify the ways that platforms' systems allow for, or are indifferent to, the publication of harmful online content, like:[32]

1. Illegal materials, such as class 1A and 1B materials (as those materials are defined in the Act);
2. Harmful materials, such as those materials already included in the Act, namely image-based abuse, abhorrent violent materials, child cyberbullying and abuse;
3. Online scams;
4. Hate speech, violence or serious harm to individuals;
5. Risks to fundamental human rights;
6. Risks electoral processes and electoral integrity; and
7. Risks to the best interests of children, public health and the environment.

[30] Ibid, Article 35.
[31] Ibid, Article 37.
[32] Reset.Tech Australia, *A duty of care in Australia's Online Safety Act*, Policy Briefing, April 2024, 16.

Case study: Risk assessment approaches in practice

The University of Technology Sydney's Human Technology Institute's proposed model law for facial recognition technology adopts a risk-based approach to regulating the development and use of facial recognition technology.[33]

The model law intends to ensure responsible use of facial recognition technology and protect against risks it poses to human rights.[34] The model law includes comprehensive human rights risk assessments.

These assessments require a number of factors to be taken into consideration in relation to a facial recognition technology application, including: the spatial context in which the application is to be used; the functionality of the application; the performance of the application; whether the application produces an output that leads to a decision that has a legal or quasi-legal effect; and whether affected individuals are able to provide free and informed consent (or withhold consent) prior to the use of the application.[35]

These factors are used to evaluate specific human rights vulnerabilities posed by the application or technology to ultimately determine its overall risk level. This process is qualitative, having regard to both subjective and objective considerations.[36]

Where the application is found to impact human rights, consideration must be given to which human rights are being restricted, whether the restriction is legally justified and whether the restriction is 'reasonable, necessary and proportionate' in all the circumstances.[37]

Where there are better ways of achieving the particular aim without involving restrictions on human rights,[38] the non-restrictive avenues are to be pursued.

The model law would be overseen by an independent regulator that has expertise in human rights, especially the right to privacy, and can work constructively with a wide range of stakeholders.[39]

This example of human rights risk assessment is robust and is centred upon, as far as possible, protecting and promoting human rights while also being flexible enough to allow for technological innovation.

5.3 Risk mitigation

Risk mitigation is the necessary next step once risks have been identified by digital platforms.

That being said, it is imperative that the Australian Government is mindful of the human rights implications of any risk mitigation measures it imposes. In this respect, any regulatory efforts that are imposed on digital platforms ought to be themselves human rights compliant. An important aspect of this is ensuring that the freedom of expression is protected by not incentivising the removal of content or shadow-banning users as a default risk mitigation strategy.

Regulatory regimes that are reliant on content moderation can have unintended consequences for human rights. For example, penalising platforms for content-moderation failures incentivises platforms to err on the side of caution and moderate more (as opposed to less) content, resulting in needless restrictions on the freedom of expression and the right to access information in circumstances beyond what was initially contemplated by the regulatory model.

[33] See generally Nicholas Davis, Lauren Berry, Edward Santow, *Facial recognition technology towards a model law* (Report, September 2022).
[34] Ibid 5.
[35] Ibid 46.
[36] Ibid 55.
[37] Ibid 56.
[38] Ibid 56.
[39] Ibid 80.

Content moderation approaches can't be scaled to meet the contemporary risks of the digital world which is fuelled by masses of content - particularly bot-generated content - leaving digital platforms and regulators playing a perpetual game of 'whack-a-mole'.[40]

It is for these reasons that content moderation should only be seen as a small part of any comprehensive regulatory regime. Instead, the regulatory focus should be on transparency and accountability and risk mitigation.

Risk mitigation measures could include user empowerment and education measures to inform users about online safety, digital literacy, and the responsible use of their platforms. Platforms could enhance their machine learning capabilities to detect and flag potentially harmful content more effectively or prevent their algorithms throttling harmful content, hate speech or misinformation without resorting to blanket content removal or user bans.

These measures could not only address risks proactively but also contribute to a safer online environment while preserving user's rights to freedom of expression and access to information.

Case study: Risk mitigation frameworks in practice

The European Union's approach to promoting trustworthy AI, exemplified by the Ethics Guidelines for Trustworthy Intelligence and the *Artificial Intelligence Act* (EU) (**AI Act**), provides a robust framework for regulating AI systems based on their risk. The EU's framework classifies AI systems into four levels of risk.[41]

High-risk AI systems include those used in critical infrastructure that could endanger lives and health, educational or vocational training systems that impact access to education and career paths, and the safety components of products like those used in AI-assisted surgery. They also encompass AI systems used in employment, migration, asylum and border control management, justice, and democratic processes, among others.[42]

High-risk AI systems must meet strict obligations before entering the market, including: adequate risk assessment and mitigation; proof of high-quality datasets to minimise discrimination; detailed documentation for compliance assessment; and appropriate human oversight to identify and reduce risks.[43]

All remote biometric identification systems, like those that allow people to be identified based on their biometric characteristics from a distance through sensors or surveillance cameras, are deemed high-risk under the framework and must adhere to stringent requirements. For example, their use in public spaces for law enforcement is generally prohibited, with narrow, regulated exceptions, such as for finding missing children or preventing imminent terrorist threats. These uses require authorisation and are subject to time, geographic, and database constraints.[44]

Limited risk AI pertains to transparency-related risks in AI use and deployment. The AI Act mandates that AI developers inform users when they are interacting with AI. For example, users must be notified when communicating with chatbots. Additionally, AI-generated public information, like deep fakes in audio and video formats, must be clearly labelled.[45]

Minimal or no risk AI applications, such as AI-enabled video games or spam filters, are freely used.[46]

[40] Reset.Tech Australia, *A duty of care in Australia's Online Safety Act* (Policy Briefing), April 2024, 2.
[41] European Commission, '*Shaping Europe's Digital Future*, (online, 19 June 2024) <https://digital-strategy.ec.europa.eu/en/policies/regulatory-framework-ai>.
[42] Ibid.
[43] Ibid.
[44] Ibid.
[45] Ibid.
[46] Ibid.

The EU's risk-based approach to AI regulation recognises that AI systems present a range of risks to people's health, safety and human rights and it seeks to regulate these based on a risk mitigation approach to balance people's rights and interests with technological innovation while also not incentivising harm.

5.4 Transparency measures

The Australian public currently has few meaningful opportunities to understand how digital platforms shape the information environment that we are exposed to.

A lack of transparency prevents users from understanding the ways in which they are tracked and targeted, and how systems determine the information that is delivered to them. There is also little transparency as to how decisions are made by platforms regarding content moderation and removals, or what avenues are available to users to seek reviews of platforms' decisions.

Transparency requirements within an effective regulatory framework should mandate that digital platforms disclose their risk assessments and mitigation efforts. This includes detailing how they identify and address potential risks associated with their products. Additionally, platforms must provide comprehensive information on their advertising and recommender systems, as well as the user profiling conducted based on individual online activities.

Under the DSA, on request from the regulator, platforms are required to provide external researchers with access to data for the purposes of conducting research on detecting, identifying and understanding the systemic risks to which risk assessments apply, and assess the adequacy, efficiency and impacts of platforms' risk mitigation measures.[47] This data regime is a significant feature of the DSA, and ensures that platforms are subject to scrutiny from non-government third-parties.

When implemented appropriately, robust data access regimes can support civil society to operate as an 'early warning system' for emerging risks, as well as providing a further layer of accountability in relation to the platforms' risk-mitigation commitments.

Measures to enhance transparency should:[48]

- Require digital platforms to publish their annual risk assessments;
- Require digital platforms to undertake annual public transparency reports, which are prescriptive and standardised;
- Provide advertisement repositories whereby all platforms are required to make publicly available meta-data about paid-for ads on each platform, including rejected ads;
- Provide repositories whereby platforms are required to make data publicly available about their paid-for data harvesting arrangements; and
- Enable researchers to access public interest data (with such researchers vetted to ensure that no malevolent actors access this data).

While platforms have valid concerns with the implications of sharing their data with governments and other researchers - such as privacy concerns - these issues are surmountable and readily addressed through appropriate safeguards for protecting sensitive data.[49]

In the broader online harm context, these concerns should not outweigh the public good in ensuring transparency and access to such data.

[47] *Digital Services Act* (EU), Article 40(4).
[48] Reset.Tech Australia, *A duty of care in Australia's Online Safety Act* (Policy Briefing), April 2024, 16.
[49] See for eg *Digital Services Act* (EU), Article 40(5)(b).

5.5 Accountability measures

The business models of large digital platforms are driven by serving users emotive, divisive content and compiling detailed personal profiling generated from user data.[50]

In this context, self-regulation and co-regulation are ineffective measures in the face of strong commercial drivers to serve harmful content. Self-regulation and co-regulation are also inappropriate for such a powerful and high-risk sector, especially where the commercial interests and business models of digital platforms can conflict with human rights, community needs, and the public interest.

Instead of self-regulation or co-regulation, obligations imposed on digital platforms should be overseen and enforced by a well-resourced, independent regulator with a deep understanding of human rights law with the requisite powers to verify digital platforms' information, and hold them accountable for failures to report and act.

The ability to impose and enforce measures to curb risks is crucial to an effective regulator. Recently, we have seen instances of digital platforms flouting existing online safety rules and dismissing the imposition of even quite modest fines or penalties.[51]

In order to ensure effective regulation, regulators must be empowered and well-resourced. Strong enforcement powers should include:[52]

1. The ability to investigate platforms' systems and designs and identify risks to human rights;

2. The ability to compel platforms to address risks by changing platforms' systems and elements;

3. The ability to issue substantial penalties, ideally a percentage of a platform's annual global turnover, for breaches, of their duty of care for example;

4. Recourse to further repercussions to impose upon organisations where failures are persistent and all other avenues have been exhausted. For example, the ability to control the availability of certain services where penalties have been blatantly disregarded;

5. Enhanced capabilities to receive complaints from consumers, consumer groups and third-party researchers regarding systemic risks and breaches of the duty of care;

6. Robust investigative and information-gathering powers; and

7. Effective notice-and-take-down powers to intervene in relation to material that is particularly harmful, where required, and ideally as a last resort.

The DSA's supervised, risk-based approach to harm minimisation was borne of a recognition that self-regulatory and co-regulatory models were not effective.[53] The Australian Government should not leave digital platforms to regulate themselves or leave it open for them to opt in to purely voluntary regulatory schemes.

---

[50] Kate Jones, Online Disinformation and Political Discourse: Applying a Human Rights Framework, Research Paper, November 2019 https://www.chathamhouse.org/sites/default/files/2019-11-05-Online-Disinformation-Human-Rights.pdf at 52.
[51] eSafety Commissioner, *Statement on removal of extreme violent content*, 23 April 2024, accessible at https://www.esafety.gov.au/newsroom/media-releases/statement-on-removal-of-extreme-violent-content.
[52] Reset.Tech Australia, *A duty of care in Australia's Online Safety Act* (Policy Briefing), April 2024, 16.
[53] Reset.Tech Australia, *How outdated approaches to regulation harm children and young people and why Australia urgently needs to pivot* (Report, December 2022), accessible at: https://au.reset.tech/uploads/report_-co-regulation-fails-young-people-final-151222.pdf.

Regulators themselves should also be transparent, such that the broader community and the platforms are able to understand why a particular decision was made by a regulator and why the regulator has, or has not, taken particular action. Further, there ought to be sufficient oversight of the regulator – to ensure that regulator itself is human rights compliant, transparent and accountable.

5.6 A note on a uniform approach to overseeing the digital environment

Presently, there are a number of Australian government protections against online harms (in a broad sense). These protections include, among others:

- The eSafety Commissioner administering the Act;
- The Office of the Australian Information Commissioner upholding privacy and information access rights under the *Privacy Act 1988* and the *Freedom of Information Act 1982*;
- The Australian Communications and Media Authority which is responsible for regulating communications and media; and
- The Australian Competition and Consumer Commission which regulates Australian Consumer Law, including online transactions and combating online scams.

These bodies exist in tandem with a range of other agencies and government departments which, working together, regulate Australia's digital environment. This patchwork approach to digital regulation is fraught with danger.

While each body has a unique and important responsibility in this space, there is potential for these responsibilities to overlap and for multiple bodies to work on the same area. There is an inherent risk that allocating responsibilities to each of these disparate bodies is ineffective and causes unnecessary duplication while making it difficult for the platforms and the general public to understand which agency is responsible for what.

Consideration should be given to, at minimum, establishing a standing parliamentary committee dedicated to overseeing digital reforms and conducting relevant inquiries. Such a committee could prioritise coordinated reforms to Australia's digital regulation, leveraging its democratic legitimacy to ensure comprehensive and effective oversight.

This committee could also prevent overlapping efforts and promote unified responses to digital regulation challenges. In recent years, multiple parliamentary and departmental inquiries have been initiated into different, yet overlapping, aspects of digital regulation, highlighting the need for a more streamlined approach.