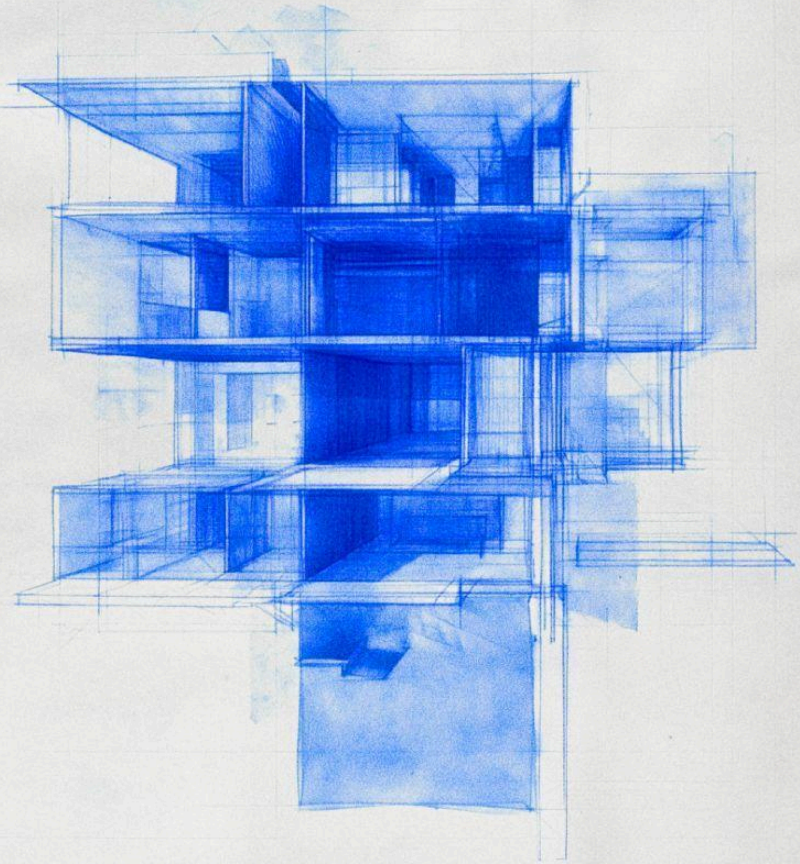
November 2024

Reset·Tech
AUSTRALIA

Five Necessary Elements
For Effective Digital Platform Regulation



Contents

Foreword

Professor Lorna Woods, OBE

Australia is no stranger to regulation in the name of online safety. Its pioneering role has been recognised, but the central focus of the enforceable provisions is content. While dealing with problematic content is and will remain an issue, a content-based approach overlooks the impact that the design of the services, particularly the objective of increasing user engagement and frictionless communication, has on users' communication choices. It is this latter focus that is designated 'systems regulation'.

Systems regulation ties in with behavioural science, concerns about addictive design, 'fast thinking', 'dark patterns' and nudges, as well as concerns around 'surveillance capitalism' and 'platform decay'. Service and interface design and business choices do not just have impacts once content has been created but also affect the choice to create and the engagement with networks of other users as well. Examples are videos defaulting to autoplay, curated playlists, data voids, hashtags and algorithmic promotion, as well as financial incentives for content creators and feedback loops created through metrification (receiving likes reinforces user behaviour). Emojis create a new shorthand for communication, and frictionless posting creates the conditions for virality, as well as targeted pile-ons.

Significantly, the decisions around design and the coding of a service are the results of choices that the service providers have made, as well as decisions around their terms of service or contracts with users – including revenue sharing and incentives for content creation – and the resources service providers deploy to enforce that. Service providers can legitimately be made to take responsibility for their choices in this regard or for ignoring information that indicates a risk of harm to users. The five elements of regulation outlined in this report return the cost of harms to those responsible for them – an application of the micro-economically efficient 'polluter pays' principle. This approach is risk-based and outcome-oriented. Given the failure of self-regulation, a regulator with sufficient powers would be necessary to ensure compliance by regulated entities.

While the justification for systems-based regulation is based fundamentally on the principle that those who create or exacerbate risk should have responsibility for mitigating it, there are other practical reasons for adopting this approach. One relates to the scale of the problem. As the UK communications regulator Ofcom noted, 'The sheer volume of text, audio and video generated or shared by online platforms is far beyond that available on broadcast television and radio', meaning that '[e]xisting frameworks could not be transferred wholesale to the online world'. We need a response that is preventative, not palliative.

This report provides the essential concepts and implementation principles for bringing systems regulation from a loose dream (as in the voluntary elements of Australia's Basic Online Safety Expectations) to enforceable reality. The framework here sets up a mechanism that is built to last and tailored to weather whatever technological storms are on the horizon. Such is the beauty of a systemic approach.

State of the Nation

The meaningful and comprehensive reform of digital platforms has become urgent. It is good news that the language of transparency, accountability, and systemic regulation have increasingly entered the Government lexicon. But what do these concepts mean in a digital platform context, and how do we ensure these crucial keywords do not become mere 'weasel words',[1] whittled down and stripped of their effectiveness? This report is a guide to the five essential, interlocking concepts making up modern online safety regulation and is designed to be used as a practical guide for those tasked with the hefty job of drafting legislation and implementing policy frameworks.

Australia has a proud history as a first mover and innovator in digital platform regulation. Australia was the first country to legislate on online safety and introduce an online safety commissioner,[2] as well as the first to legislate negotiations between digital platforms and news providers.[3] Analysis from the Australian Competition and Consumer Commission's (ACCC) Digital Platforms Inquiry Final Report[4] continues to influence cutting-edge policy thinking, locally and internationally.[5]

Australia's early-wave tech regulation is now met with early-wave enforcement challenges. The limitations of the *Online Safety Act 2021* have been brought into sharp view this year. For example, the limited enforceability of the Act's flagship online safety standards (the Basic Online Safety Expectations) was highlighted in the statutory review process.[6] Enforceability and enforcement challenges have also plagued those relying on the News Media Bargaining Code package. After Meta took the rather foreseeable path of simply walking out of its commitments, the Government has been left searching for an adequate 'stick'.

Meanwhile, digital threats are evolving and scaling up in ways that seemed almost unimaginable only a few years ago. New risks, driven by increasingly powerful algorithms and an explosion of data harvesting, have now surpassed the ability of existing digital regulatory frameworks to effectively manage them. Australia is not alone in facing these risks, but other countries are now making substantial progress – in particular, the EU[7] and the UK,[8] with emerging progress in Canada.[9] These jurisdictions have drawn upon the innovations and exemplars of Australian policy but introduced more comprehensive, preventative and muscular regulatory models. These models encourage platform conduct that ensures user safety and is commensurate with public expectations for digital regulation more broadly.

---

[1] Don Watson, *Watson's Dictionary of Weasel Words* (Penguin Books Australia, 2005).
[2] Via the *Enhancing Online Safety for Children Act 2015*, https://www.legislation.gov.au/C2015A00024/2017-06-23/text.
[3] Via the *Treasury Laws Amendment (News Media and Digital Platforms Mandatory Bargaining Code) Act 2021* , https://www.legislation.gov.au/C2021A00021/asmade/text.
[4] Australian Competition and Consumer Commission, *Digital Platforms Inquiry Final Report* (2019), https://www.accc.gov.au/about-us/publications/digital-platforms-inquiry-final-report.
[5] For example, see the Digital Markets, Competition and Consumers Bill 2024 (UK), https://bills.parliament.uk/publications/54208/documents/4421; Government of Canada, *Towards guiding principles — Diversity of content in the digital age* (2020), *https://www.canada.ca/en/canadian-heritage/services/diversity-content-digital-age/towards-guiding-principles.html*; Government of Canada, *News Media Canada* (2021), https://ised-isde.canada.ca/site/strategic-policy-sector/en/marketplace-framework-policy/copyright-policy/submissions-consultation-modern-copyright-framework-online-intermediaries/news-media-canada-nmc.
[6] 'Expectations do not impose a legally enforceable duty on service providers to implement the expectations', and 'there are no penalties for a service provider failing to comply with the expectations outlined in the Basic Online Safety Expectations Determination' (see Department of Infrastructure, Transport, Regional Development, Communications and the Arts, *Statutory Review of the Online Safety Act 2021*, https://www.infrastructure.gov.au/sites/default/files/documents/online-safety-act-2021-review-issues-paper-26-april-2024.pdf) . The point on enforceability is distinct from recent litigation outcomes between the regulator and X Corp, which pertain to non-compliance with transparency notices. While platforms may face enforcement actions for failing to comply with regulator requests for information, this does not equate to being held accountable for the substance of their safety measures.
[7] *Digital Services Act 2022* (European Union), https://eur-lex.europa.eu/eli/reg/2022/2065/oj.
[8] *Online Safety Act 2023* (UK), https://www.legislation.gov.uk/ukpga/2023/50/enacted.
[9] Online Harms Bill 2024 (Canada), https://www.parl.ca/LegisInfo/en/bill/44-1/c-63.

By contrast, Australia is still largely reliant on a hopeful but outdated desire for industry-led and largely self-regulated processes. These processes invite industry to shape the rules by which they are then supervised and are often described as co-regulation. Co-regulation is not a new concept, particularly in the broadcasting and telecommunications sectors, that adhere to the Telecommunications Consumer Protections Code and the Commercial Television Industry Code of Practice. However, there are structural and seismic differences between telecommunications providers, such as TV and radio, and foreign digital platform behemoths. The co-regulatory concept has surpassed the confined sectoral context in which it was designed to operate, and this is creating regulatory failures for consumers.

Harm happens as governments wait for self-regulation and co-regulation to fail. Nine years on from the first online safety legislation and five years on from the findings of the ACCC's *Digital Platform Services Inquiry,* Australia faces new digital challenges. A non-exhaustive list includes:

- Personalised and persistent scam calls, texts and advertisements linked to digital advertising business models, causing significant economic harm to Australians;[10]
- Ongoing risks of online harms for children,[11] including online exploitation;[12]
- Increasing cyber abuse directed at adults, especially women,[13] and hate speech directed at minorities;[14]
- Vast and invasive data breaches, exacerbated by Australia's weak privacy and data protection laws, widening existing holes in national and personal security;[15]
- Implementation challenges over the News Media Bargaining Code, with Meta's exit from the deals threatening a loss of over $100 million to the Australian news market;[16]
- A deteriorating information environment, with upticks in fringe and palpably false content, including a rise in AI-generated content with unclear provenance;[17]

---

[10] National Anti-Scam Centre, *National Anti-Scam Centre in Action Quarterly Update* (2023), https://www.accc.gov.au/about-us/publications/serial-publications/national-anti-scam-centre-quarterly-update/national-anti-scam-centre-quarterly-update-march-2024; Consumer Policy Research Centre, *Singled out* (2024), https://cprc.org.au/wp-content/uploads/2024/02/CPRC-Singled-Out-Final-Feb-2024.pdf; Reset.Tech Australia, *Any buyer accepted: Unregulated data markets create personal security risks* (2024), https://au.reset.tech/uploads/Reset-Australia-Report-Any-Buyer-Accepted-240926-V1-WEB-%281%29.pdf.

[11] These range from EdTech apps that breach students' privacy [see Human Rights Watch, *How dare they peep into my private life?* (2022), https://www.hrw.org/report/2022/05/25/how-dare-they-peep-my-private-life/childrens-rights-violations-governments], to algorithms that serve them pro-eating disorder content [Reset.Tech Australia, *Not just algorithms: assuring user safety online with systemic regulatory frameworks* (2024), https://au.reset.tech/news/report-not-just-algorithms/].

[12] Office of the eSafety Commissioner, *World-first report shows leading tech companies are not doing enough to tackle online child abuse* (2022), https://www.esafety.gov.au/newsroom/media-releases/world-first-report-shows-leading-tech-companies-are-not-doing-enough-to-tackle-online-child-abuse.

[13] Office of the eSafety Commissioner, *Women in the spotlight: How online abuse impacts women in their working lives* (2022), https://www.esafety.gov.au/research/how-online-abuse-impacts-women-working-lives.

[14] See, for example, the experience of Indigenous Australians during the Voice referendum in Jack Latimore, 'Meta rules online racism against Indigenous people meets community standards' (2023) *Sydney Morning Herald*, https://www.smh.com.au/national/meta-rules-online-racism-against-indigenous-people-meets-community-standards-20230815-p5dwqt.html.

[15] Office of the Australian Information Commissioner, *Notifiable data breaches report* (2024), http://www.oaic.gov.au/__data/assets/pdf_file/0021/156531/Notifiable-data-breaches-report-July-to-December-2023.pdf; Reset.Tech Australia, *Australians for sale targeted advertising, data brokering, and consumer manipulation* (2023), https://au.reset.tech/news/coming-soon-australians-for-sale-report/.

[16] Minister for Communications, *Press conference* (2024), https://minister.infrastructure.gov.au/rowland/interview/transcript-press-conference-sydney-0; Rod Sims, 'Australia's News Media Bargaining Code led the world. It's time to finish what we started' (2022) *The Conversation*, https://theconversation.com/australias-news-media-bargaining-code-led-the-world-its-time-to-finish-what-we-started-188586.

[17] 'Highest level of mis-and-disinformation we've seen online' *ABC Radio National* (Online, Australian Broadcasting Corporation, 2023) https://www.abc.net.au/listen/programs/radionational-breakfast/aec-on-referendum-education-campaign-and-misinformation-/102758190; Pranshu Verma, 'The rise of AI fake news is creating a "misinformation superspreader' (2023) *Washington Post*, https://www.washingtonpost.com/technology/2023/12/17/ai-fake-news-misinformation/.

- Governance challenges to DIGI's Australian Code of Practice on Misinformation and Disinformation, with X (formerly known as Twitter) exiting the Code after routine failures to respond to independent reports of serious breaches;[18] and
- Deepening national security threats of ideologically motivated extremism,[19] with intensifying links to content recommender systems or algorithms.[20]

Over the last year, governments at home and around the world have also learned that:

- Voluntary or, at best, co-regulatory schemes do not produce high-quality protections for Australians[21] and can simply be ignored by platforms. The reputational risk approach, once held as a sufficient incentive for voluntary public interest safeguards,[22] is simply not enough;
- Even legislation and fine regimes are vulnerable to dismissal by very large platforms if the fines are considered simple costs of doing business or if the enforcement regimes are considered evadable due to industry's jurisdictional arbitrage tactics or other manoeuvres; [23] and
- The state of transparency collapse across the tech industry is chilling independent research right when regulators need that know-how and evidence to do their job. Some platforms are going as far as pursuing 'lawfare' tactics, while others are shutting down vital tools and data sources.[24] Access to key types of platform data and rights to conduct public interest research must be legislated.[25]

[18] DIGI, *Media statement* (2023), https://digi.org.au/category/media-statement/.
[19] Australian Security Intelligence Organisation, *Director General's Annual Threat Assessment* (2022), https://www.asio.gov.au/resources/speeches-and-statements/director-generals-annual-threat-assessment-2022; Australian Security Intelligence Organisation, *Director General's Annual Threat Assessment* (2023), https://www.asio.gov.au/director-generals-annual-threat-assessment-2023; Australian Security Intelligence Organisation, *Director General's Annual Threat Assessment* (2024), https://www.oni.gov.au/asio-annual-threat-assessment-2024.
[20] Reset.Tech Australia, *Algorithms as a weapon against women: How YouTube lures boys and young men into the 'Manosphere'* (2022), https://au.reset.tech/news/algorithms-as-a-weapon-against-women-how-youtube-lures-boys-and-young-men-into-the-manosphere/; Manoel H Ribeiro *et al.*, *Auditing radicalization pathways on YouTube* (2019), https://www.researchgate.net/publication/335337464_Auditing_Radicalization_Pathways_on_YouTube.
[21] Reset.Tech Australia, *Does digital co-regulation function in children's best interests?* (2024), https://au.reset.tech/news/does-digital-co-regulation-function-in-children-s-best-interests/; Reset.Tech Australia, *How outdated approaches to regulation harm children and young people and why Australia urgently needs to pivot* (2022), https://au.reset.tech/uploads/report_-co-regulation-fails-young-people-final-151222.pdf.
[22] Tess Bennett, 'Social media giants "no longer fear reputation risks"' (2024) *AFR*, https://www.afr.com/technology/social-media-giants-no-longer-fear-reputation-risks-20240422-p5flls.
[23] See the facts from *X Corp v eSafety Commissioner* [2024] FCA 1159, where X attempted to avoid a fine on the basis of changes to the company.
[24] Justin Hendrix, "The demise of CrowdTangle and what it means for independent technology research' (2024) *Tech Policy Press*, https://www.techpolicy.press/the-demise-of-crowdtangle-and-what-it-means-for-independent-technology-research/
[25] Supplementary explanatory memorandum, Communications Legislation Amendment (Combatting Misinformation and Disinformation) Bill 2024 (Cth), https://parlinfo.aph.gov.au/parlInfo/search/display/display.w3p;query=Id%3A%22legislation%2Fems%2Fr7239_ems_66335a38-df6e-4b06-a8a3-e20d4d8670af%22

The Albanese Government has recognised that Australia needs a comprehensive regulatory model that addresses the underlying *systems* of digital platforms, rather than continuing to rely on content-based regulatory responses. In November 2024, the Minister for Communications announced an intention to introduce a legislated duty of care, citing a 'growing global effort' and an intent to 'deliver a more systemic and preventative approach to making online services safer and healthier'.[26]

The model the Albanese Government introduces to the Parliament should include *all five elements* required for systemic and preventative digital regulation, namely:

1. An overarching duty of care owed by digital platforms to Australian users;
2. Requirements for platforms to assess all their systems and elements for a defined set of risks;
3. Requirements for platforms to implement reasonable steps to mitigate each risk;
4. Five sources of transparency, including annual risk assessments, prescriptive public transparency reports, independent audits of risk assessments and transparency reports, data portals for ad repositories and content moderation decisions, and researcher access to public interest data; and
5. Enforceable regulations and empowered regulators to compel behavioural change.

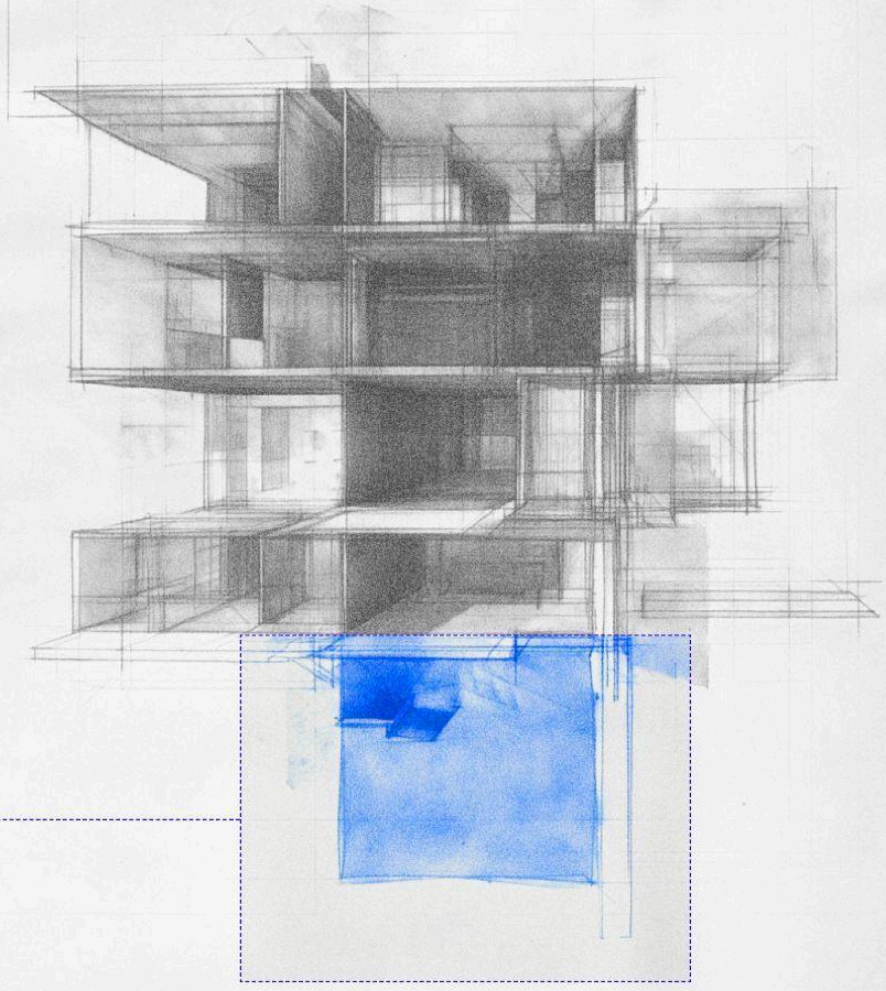
Note that these need to be implemented alongside a reformed and updated *Privacy Act* that protects Australians from predatory digital business practices. The proposals put forward in the *Privacy Act Review Report*[27] are strong and move in the right direction. These are needed to mitigate the personal and national security risks that social media platforms and other digital platforms routinely generate.

This report draws on previous thinking about the five elements of tech regulation that would be necessary to achieve effective, comprehensive digital regulation in Australia,[28] building on this to include implementation guidelines. As the discussion has advanced, we hope these implementation guidelines add clarity about how to move from slogans to meaningful change.

[26] Hon Michelle Rowland MP, 'The governance of digital platforms' (Speech, The Sydney Institute), 13th November 2024, https://minister.infrastructure.gov.au/rowland/speech/speech-sydney-institute-governance-digital-platforms.
[27] Attorney General's Department, *Government response to the Privacy Act Review Report* (2024), https://www.ag.gov.au/rights-and-protections/publications/government-response-privacy-act-review-report.
[28] Reset.Tech Australia, *A duty of care in Australia's online safety framework* (2024), https://au.reset.tech/uploads/Duty-of-Care-Report-Reset.Tech.pdf; Reset.Tech Australia, *Digital Platform Regulation Green Paper* (2024), https://au.reset.tech/uploads/Digital-Platform-Regulation-Green-Paper.pdf.

Five Necessary Elements
For Effective Digital Platform Regulation

Element One:
Duty of Care



November 2024

Reset.Tech Australia

Element One: A Duty of Care

What's This?

Ensuring that digital platforms play their part in reducing the risk architecture requires flipping the table from older models of regulation, where end users shoulder the bulk of the risk, to instead placing responsibilities onto digital platforms to keep end users safe. Based on learning from international models, placing a duty of care on digital platforms may help drive the systemic and preventative focus that is urgently needed in Australia.

A duty of care approach is a way to implement systemic regulation that moves the focus beyond the content layer of the digital world to the underlying systems – the environment where content is created, shared and promoted. The design of these underlying systems is entirely within a platform's control (to a lesser extent where content is generated by users). Focusing regulation on systems and processes requires platforms to assess whether there is a risk of harm to users arising from their technical systems, design and business models while still encouraging user expression.

Focusing on design and operation is important because, despite their name, platforms are not entirely neutral, passive transmitters when it comes to content. Intentionally or not, their choice of architecture impacts content. This includes the role of recommender and content moderation systems, for example, and how engagement features are designed to create social pressures or allow for anonymous accounts. Duty of care is a way to implement systemic regulation that can address these types of risks.

Duty of care is a familiar model for risk management in Australia, with established frameworks in workplace health and safety. An online statutory duty of care exists in the UK's *Online Safety Act 2023* (UK OSA)[29] and is contemplated in draft Canadian legislation, the Online Harms Bill 2024.[30] We note that proposals for a duty of care in Australia should be mindful of the British experience and avoid being watered down into pluralised duties of care. Introducing duties of care rather than a singular duty of care reduces the systemic focus and introduces content-focused confusions and limitations into regulation.[31]

---

[29] *Online Safety Act 2023* (UK), https://www.legislation.gov.uk/ukpga/2023/50/enacted.
[30] Online Harms Bill 2024 (Canada), https://www.parl.ca/LegisInfo/en/bill/44-1/c-63.
[31] Rys Farthing and Lorna Woods, 'The dangers of pluralisation: A singular duty of care in the Online Safety Act' (2024) *The Policymaker*, https://thepolicymaker.jmi.org.au/the-dangers-of-pluralisation-a-singular-duty-of-care-in-the-online-safety-act/.

Practical Implementation Guidelines

- The duty of care concept is gaining traction in Australian tech policy. It is essential that it is understood not as a standalone and all-encompassing measure but as a legal hook to activate a comprehensive policy framework. The scope and substance of a duty of care are clarified through risk assessments and mitigations (Elements 2 and 3) and realised through transparency and accountability measures (Elements 4 and 5).
- Industry typically resists a singular and broad duty on the basis of perceived regulatory uncertainty. However, to be effective, the duty must activate fulsome risk assessment and mitigation across all of a large platform's systems. Without a single, overarching duty, it is too easy to create content carve-outs and protection gaps.
- Indicative drafting language for a single, overarching duty is expressed in Figure 1.
- 'Reasonable steps' is a common standard in Australian jurisprudence. Learning from *eSafety Commissioner v X Corp,*[32] it is necessary that legislation defines the meaning and content of reasonable steps in the technical context of digital platforms and online services. Note that the reasonable steps standard, once clarified, may provide a template for expected risk mitigation methods. See Element 3 for drafting language on these.

Figure 1: Indicative drafting language for a single, overarching duty

The duty of care obligations of a large provider of a regulated online service are:

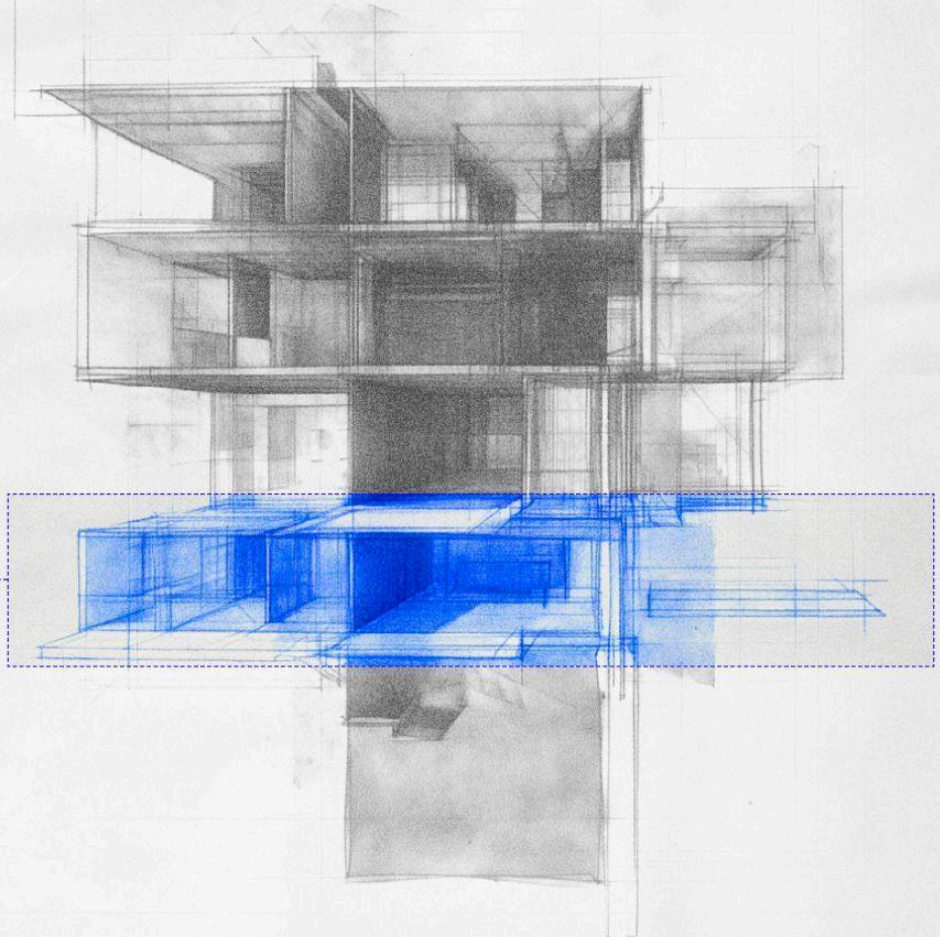
a) to take reasonable steps to conduct its business (including the design and operation of systems and processes relevant to providing the service) and to provide the service, with honesty, integrity, and with due skill, care and diligence; and

b) in conducting its business (including the design and operation of systems and processes relevant to providing the service) and providing the service, to take reasonable steps to prevent matters from arising that would (or would be likely to) cause harm or detriment to end-users of the service.

[32] *eSafety Commissioner v X Corp* [2024] FCA 499.

Five Necessary Elements
For Effective Digital Platform Regulation

Element Two:
Risk Assessment



November 2024

Reset.Tech Australia

Element Two: Risk Assessment

What's This?

Once responsibility has been placed onto digital platforms to safeguard end users, requirements to produce risk assessments can introduce a comprehensive focus into the regulatory framework. This approach has strong international precedents. Requirements to produce risk assessments for systemic risks on digital platforms exist in both the *Digital Services Act 2022* (DSA)[33] and the UK OSA.

Currently, risk assessments are part of the Australian Basic Online Safety Expectations (BOSE), although they are suggested as an example of a reasonable step to address specific risks covered by the BOSE. They are neither mandatory nor comprehensive. In addition, the Office of the eSafety Commissioner has created a world-leading Safety by Design assessment tool, which serves as guidance and advice for digital product developers.[34] While this tool has significant strengths, it is a self-assessment tool linked to a set of safety risks and was not designed to support regulatory enforcement.

Requirements to produce risk assessments may ensure that platforms adequately review and identify the risks that their systems and processes create. As the Centre on Regulation in Europe describes, risk assessment activities begin with a comprehensive mapping activity that identifies the ecosystem in which platforms operate, the roles and behaviours of users, business decisions made by platforms and how these interactions produce risks.[35] In other words, risk assessments have the capacity to encourage digital platforms to think comprehensively about how their platforms can create or amplify risks.

---

[33] *Digital Services Act 2022* (European Union), https://eur-lex.europa.eu/eli/reg/2022/2065/oj.
[34] Office of the eSafety Commissioner, *Assessment tools* (2023), https://www.esafety.gov.au/industry/safety-by-design/assessment-tools.
[35] Sally Broughton and Micova Andrea Calef, *Elements for effective systemic risk assessment under the DSA* (2022), https://cerre.eu/wp-content/uploads/2023/07/CERRE-DSA-Systemic-Risk-Report.pdf.

Practical Implementation Guidelines

- A risk assessment report is a critical document to set up the proactive and preventative component of systemic, risk-based regulation. Under the European model, risk assessment reports must meet a certain standard of diligence. The regulator's assessment of diligence allows regulators to flag potential issues before demonstrable harm has occurred. That is, regulators can enter into discussions with platforms about specific features or functionalities based on the diligence of a risk assessment without waiting for harm to happen.
- Outside the cadence of annual risk assessment reports lies an expectation for risk assessments to be produced each time a new product or feature is introduced to the market that may introduce new risks, creating accountability at an earlier stage in a product life cycle.
- Australia is not starting from a standstill in this regard. The eSafety Commissioner put forward a groundbreaking set of Safety by Design principles in 2018. The principles call for 'documented risk management and impact assessments to assess and remediate any potential online harms that could be enabled or facilitated by the product or service'. The Safety by Design principles' mitigation measures are covered in Element 3.

Figure 2: Indicative drafting language for risk assessments

A large provider of a regulated online service must undertake an assessment (a risk assessment) that identifies and assesses the risks associated with providing the service.

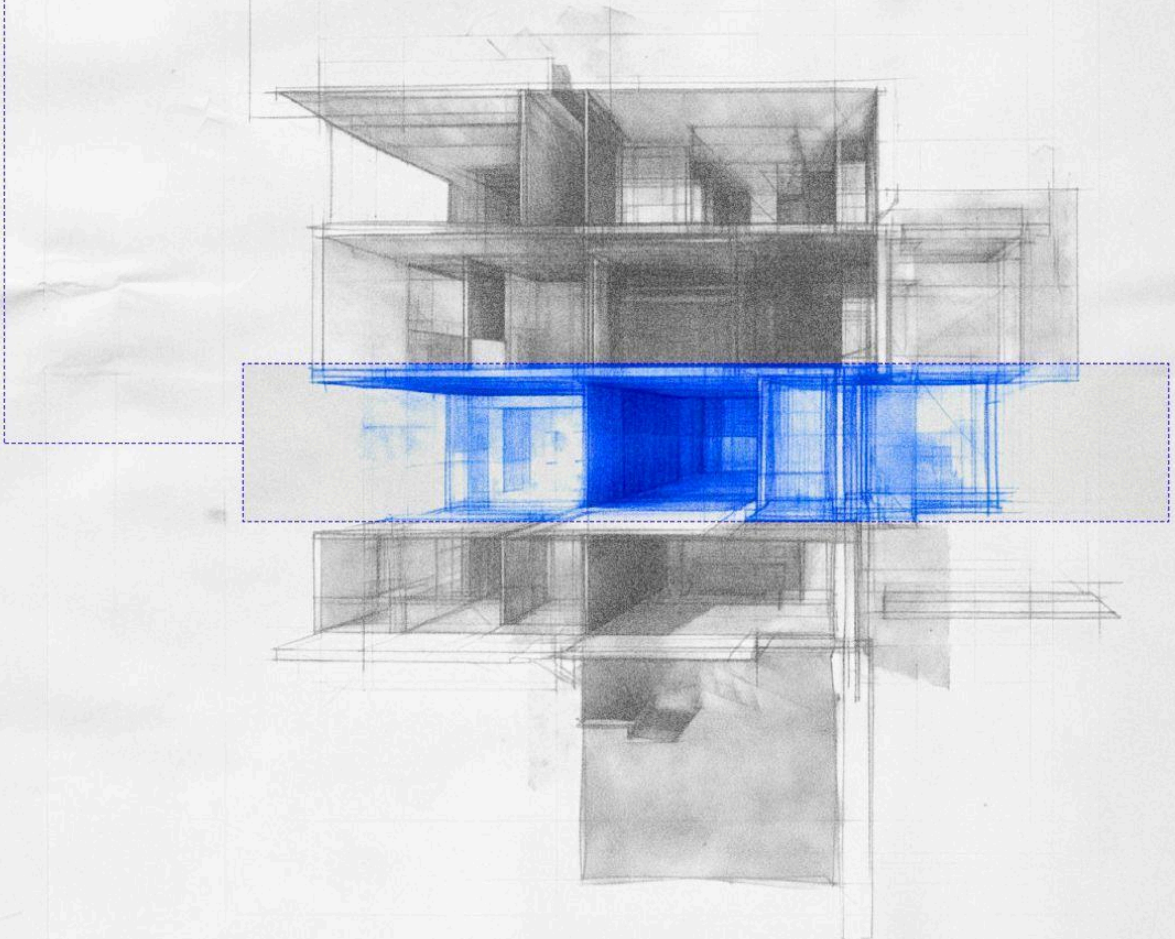
The provider must have regard to the following matters in undertaking a risk assessment:

a) Risks to young people;
b) Risks to mental wellbeing;
c) Risks of the instruction and promotion of harmful practices;
d) Risks of illegal content, conduct and activity; and
e) Risks to personal safety & security.

Risks to young people are to be determined based on a best interests test.

Five Necessary Elements
For Effective Digital Platform Regulation

Element Three:
Risk Mitigation



November 2024

Reset.Tech Australia

Element Three: Risk Mitigation

What's This?

The responsibility to identify a comprehensive, systemic set of risks can be preventative when digital platforms are required to actively mitigate and minimise the likelihood and severity of these risks. This way, platforms can be incentivised to implement changes that prevent harm from occurring in the first instance. In this sense, risk mitigation measures are the equivalent of 'placing a fence at the top of a cliff rather than ambulances at the bottom' – as the idiom goes.

Again, strong international precedents exist for risk mitigation requirements. The DSA[36] (and the UK OSA) places obligations on platforms to mitigate identified risks, and Canada's Online Harms Bill also imposes obligations on platforms to mitigate risks aligned with their duties. Currently, risk assessments that include risk mitigation measures are part of the Australian BOSE, although they are suggested as an example of a reasonable step in response to a range of risks covered by the BOSE and are not mandatory.

We have seen requirements for risk mitigation measures begin to bring about positive changes overseas. For example, the European Commission has opened formal proceedings against Meta for failing to adequately identify risk mitigation measures to curb harm to minors and for failing to adequately adopt mitigation measures regarding visibility around political content and flagging illegal content, among others.[37]

Practical Implementation Guidelines

Risk mitigation measures would involve making changes to systems and processes to bake in safety from the get-go. These include, for example:

- Changing the design, features or functioning of services, including online interfaces;
- Changing terms and conditions and their enforcement;
- Changing content moderation processes;
- Testing and changing algorithms, including recommender systems;
- Changing advertising systems, including the way advertisements are targeted at or presented to people;
- Improving internal business processes to maximise safety;
- Collaborating with other digital services;
- Taking targeted measures to improve child safety, such as age assurance or parental control tools; and
- Ensuring that evidence about potential illegal activities is stored and reported in ways helpful to law enforcement.

[36] See Article 35, *Digital Services Act 2022* (European Union), https://eur-lex.europa.eu/eli/reg/2022/2065/oj.
[37] European Commission, *Commission opens formal proceedings against Meta under the Digital Services Act related to the protection of minors on Facebook and Instagram* (2024) https://ec.europa.eu/commission/presscorner/detail/en/ip_24_2664

The DSA specifically outlines a set of mitigation measures that can be expected from digital platforms (See Figure 3). These are now considered international best practice. Australian expectations may harmonise with EU requirements to reduce the compliance burden on platforms. This would introduce a robust mechanism that encourages platforms to implement preventative measures and allows regulators to meaningfully interrogate proposed measures while they are still risks rather than actualised harms.

Figure 3: Extracts from Article 35 of the Digital Services Act

Mitigation of risks

(a) adapting the design, features or functioning of their services, including their online interfaces;

(b) adapting their terms and conditions and their enforcement;

(c) adapting content moderation processes, including the speed and quality of processing notices related to specific types of illegal content and, where appropriate, the expeditious removal of, or the disabling of access to, the content notified, in particular in respect of illegal hate speech or cyber violence, as well as adapting any relevant decision-making processes and dedicated resources for content moderation;

(d) testing and adapting their algorithmic systems, including their recommender systems;

(e) adapting their advertising systems and adopting targeted measures aimed at limiting or adjusting the presentation of advertisements in association with the service they provide;

(f) reinforcing the internal processes, resources, testing, documentation, or supervision of any of their activities in particular as regards detection of systemic risk;

(g) initiating or adjusting cooperation with trusted flaggers in accordance with Article 22 and the implementation of the decisions of out-of-court dispute settlement bodies pursuant to Article 21;

(h) initiating or adjusting cooperation with other providers of online platforms or of online search engines through the codes of conduct and the crisis protocols referred to in Articles 45 and 48 respectively;

(i) taking awareness-raising measures and adapting their online interface in order to give recipients of the service more information;

(j) taking targeted measures to protect the rights of the child, including age verification and parental control tools, tools aimed at helping minors signal abuse or obtain support, as appropriate;

(k) ensuring that an item of information, whether it constitutes a generated or manipulated image, audio or video that appreciably resembles existing persons, objects, places or other entities or events and falsely appears to a person to be authentic or truthful is distinguishable through prominent markings when presented on their online interfaces, and, in addition, providing an easy to use functionality which enables recipients of the service to indicate such information.[38]

The eSafety Commissioner's Safety by Design principles provide a useful framework that has been drawn upon in numerous jurisdictions to inform what best-practice risk mitigation should look like. These guidelines may be entrenched in legislation to be made effective and enforceable. Figure 4 proposes some minor amendments to integrate the spirit of these principles into draft legislation.

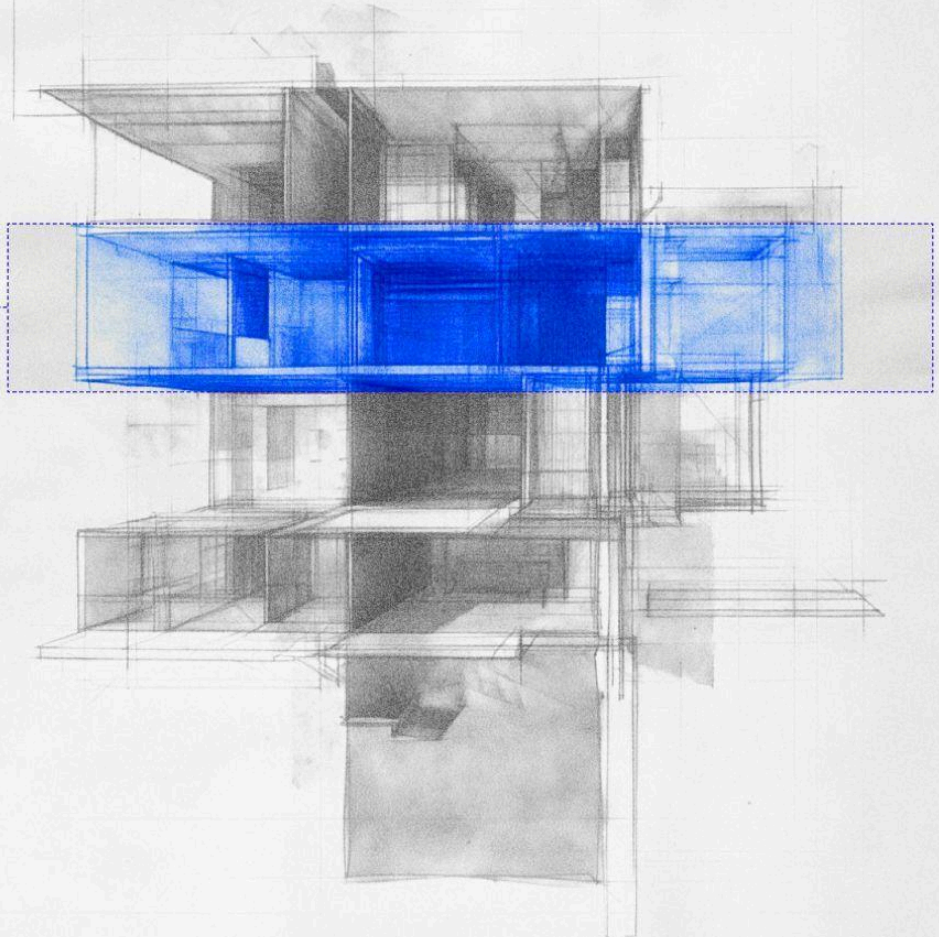
[38] See Article 35, *Digital Services Act 2022* (European Union), https://eur-lex.europa.eu/eli/reg/2022/2065/oj.

Figure 4: Potential language for risk mitigation measures

Drafting language for a reformed *Online Safety Act*	Safety By Design Principles on user empowerment and autonomy
Risk mitigation measures 1) Providers must take reasonable steps to mitigate risks outlined in their risk assessments, including by, but not limited to: a) Applying the highest available privacy and safety user settings by default, b) Reliable enforcement of terms of service, c) Providing appropriate transparency and due process (where necessary)) into user-facing decisions, including: i) Removal or de-amplification of user-generated content, ii) Rejection of advertisements, and; iii) Targeting of advertisements. d) Developing technical features to mitigate risk and harms, and ensuring these are meaningfully provided in the user journey, e) Providing timely, responsive support for user safety issues, ensuring that safety and security concerns are dealt with promptly and effectively, once flagged, f) Evaluating all design and function features to ensure that risk factors for all users – particularly for those with distinct characteristics and capabilities –have been mitigated before products or features are released to the public. *Civil penalty provision* Subclause (1) is a civil penalty provision.	1) Provide technical measures and tools that adequately allow users to manage their own safety, and that are set to the most secure privacy and safety levels by default. 2) Establish clear protocols and consequences for service violations that serve as meaningful deterrents and reflect the values and expectations of the users. 3) Leverage the use of technical features to mitigate risks and harms, which can be flagged to users at relevant points in the service, and which prompt and optimise safer interactions. 4) Provide built-in support functions and feedback loops for users that inform users on the status of their reports, what outcomes have been taken and offer an opportunity for appeal. 5) Evaluate all design and function features to ensure that risk factors for all users – particularly for those with distinct characteristics and capabilities –have been mitigated before products or features are released to the public.

Five Necessary Elements
For Effective Digital Platform Regulation

Element Four: Transparency Measures



November 2024

Reset.Tech Australia

Element Four: Transparency Measures

What's This?

Regulating for transparency helps address the power asymmetry of large digital platforms by making some of the information necessary for understanding online risks visible to the public and regulators. This enables individuals to make informed choices about platform use and allows regulators to take action. Current Australian measures for transparency in the online safety framework stem from requirements in the BOSE. Under the BOSE, the Office of the eSafety Commissioner has the power to request a range of information from platforms through periodic and non-periodic transparency notices.[39] While responses to these notices are sent directly to the Office of the eSafety Commissioner, the Commissioner is empowered to publish a statement regarding reports on their website, which serves a subsequent public transparency function.[40] The platforms have not always adequately responded to these requests.[41]

Internationally, transparency requirements are stronger in other markets with regulation. For example, the DSA introduces five key types of public transparency measures: annual risk assessments released in summary form to the public after a period; highly prescriptive annual transparency reports sharing detailed data about platform functioning; annual independent audits; data portals, including ad repositories and content moderation data; and researcher access to public interest data.[42] Similarly, the UK OSA introduces two key public transparency measures: annual risk assessments and annual transparency reports.[43] Learning from this, Australia may adopt a model of transparency that includes five key measures:

1. Requirements for summaries of risk assessments to be published;
2. Annual, prescriptive transparency reports;
3. Annual independent audits;
4. Data portals; and
5. Researcher access, including API initiatives.

[39] Office of the eSafety Commissioner, *Responses to transparency notices* (2024), https://www.esafety.gov.au/industry/basic-online-safety-expectations/responses-to-transparency-notices.
[40] *Online Safety Act 2021*(Cth) Division 3(A) 59 2, https://www.legislation.gov.au/C2021A00076/latest/text.
[41] See, for example, *X Corp v eSafety Commissioner* (VID956/2023). Status available at https://www.comcourts.gov.au/file/Federal/P/VID956/2023/actions.
[42] See, for example, Reset.Tech Australia, *Achieving digital platform public transparency in Australia* (2024), https://au.reset.tech/news/achieving-digital-platform-public-transparency-in-australia/.
[43] See, for example, Reset.Tech Australia, *Achieving digital platform public transparency in Australia* (2024), https://au.reset.tech/news/achieving-digital-platform-public-transparency-in-australia/.

Practical Implementation Guidelines

Implementation detail on the five measures is as follows:

1. Risk assessment reports

Summaries of risk assessment reports need to be made publicly available within a reasonable passage of time. Platforms need to be able to provide sufficient detail to regulators, including sensitive information, to allow regulators to adequately assess the diligence of risk identification and the effectiveness of any mitigation measures. We appreciate that this will take time and that the entirety of risk assessments may not be made public as a result. Risk assessments serve multiple purposes, and public accountability is one of the many purposes they can fulfil. Public summaries need to be shared in sufficient detail within a reasonable time window.

2. Transparency reports

Annual prescriptive transparency reports need to be made available. These reports need to be more than just a public relations document; they need to answer a number of key questions set out by regulators with up-to-date Australian data. Some potential metrics are the summaries in Figure 3.[44]

3. Audit reports

While platforms are rightly the authors of risk assessments and transparency reports, some sort of independent auditing process for both is needed. Similar requirements exist in the EU DSA, and a small industry of auditors and compliance software has innovated and emerged to serve these needs.

4. Data portals

The DSA compels timely data from platforms on advertising and user-generated content, referred to as 'ad repositories', and content moderation data. Ad repositories make visible and searchable what advertisements are running in the EU and who is paying for them.[45] Content moderation data are topline figures on the outcomes of platform decision-making in the EU, including their assessments of and responses to terms of service violations.[46] Similar data portals containing Australian data should be made available.

5. Researcher access, including API initiatives

Australia made its first formal sign of progress on researcher access to platform data in November 2024 by way of a government amendment to the Combatting Misinformation and Disinformation Bill.[47] The data access scheme outlined in these amendments provides for one type of researcher access: bespoke and unique requests between researchers and platforms, mediated by a regulator. Platforms should also be encouraged to make a public API available for research purposes, similar to Europe and largely the U.S. Researcher access is urgently required for adjacent matters of online risks and harms. Figure 6 provides an overview of what a relevant data access regime may look like. Note that this is distinct from API access, which is a request made to the platforms directly (i.e. not mediated by a regulator).

---

[44] For more information about what these might look like, see Reset.Tech Australia, *Achieving digital platform public transparency* (2024), https://au.reset.tech/news/achieving-digital-platform-public-transparency-in-australia/.
[45] See, for example, Amazon's 'Ad Library API' here: https://advertising.amazon.com/API/docs/en-us/ad-library.
[46] Data is submitted directly to the DSA Transparency Database. A public version is available at https://transparency.dsa.ec.europa.eu/dashboard.
[47] Government amendment [sheet ZC302], Combatting Misinformation and Disinformation Bill 2024 (Cth) https://parlinfo.aph.gov.au/parlInfo/download/legislation/amend/r7239_amend_2f37259b-dc40-4cd4-98cd-720a8ef3da91/upload_pdf/ZC302.pdf;fileType=application%2Fpdf.

Figure 5: Indicative transparency report metrics

- Metrics on the design, features, or functioning of services:
  - Data on internal safety tests consisting of features and systems conducted, including a description of tests and outcomes, and nature of adaptations made as a result that affect Australian end users
  - Changes to Community Guidelines and Terms of Service for Australian end users
  - Human resources dedicated to trust and safety, including information about the number located within Australia, the number dedicated to Australian safety issues and the safety of Australian end users, qualifications and training, and support

- Problematic use metrics:
  - Number of adult users demonstrating problematic overuse, and data about average and median use times
    - Number of push notifications sent to these users on average per day (in app and out of app)
  - Number of child users (under 18) demonstrating problematic overuse, and data about average and median use times
    - Number of push notifications sent to these users on average per day (in app and out of app)
  - Number of child users (under 18) accessing the platforms between 10pm and 6am in their time zone, and data about average overnight usage
  - Estimates of the number of users under the minimum age of use according to the terms of service and data about average detection and response to these accounts

- Child sexual exploitation and abuse metrics:
  - Numbers of adult users blocked for contact with minors and data about response times and previous reportings of users
  - Numbers of adult users reported by minors, and data about responses and response times
  - Number of CSAM reports made, and data about responses and response times

- Online scam metrics:
  - Number of online scam and scam posts reported on the platform, including data about the detection method (organic or user report), average engagement and responses including average response time

For a full list, including potential metrics for disinformation see the report, 'Achieving digital platform public transparency'[48]

[48] Reset.Tech Australia, *Achieving digital platform public transparency* (2024), https://au.reset.tech/news/achieving-digital-platform-public-transparency-in-australia/.

Figure 6: Elements of a data access regime for online safety issues

Building on the model under the DSA,[49] Australian researchers could have mandated access to platform data. In Australia, the requirements for an Australian vetted researcher could include:

- Affiliation with a research organisation, including academic and not-for-profit research organisations,
- Australian residency or citizenship for the researchers, or at least the lead researcher, and
- Non-commercial purpose limitations.

Suitable research projects should be provided with data. A suitable project proposal would include information demonstrating that:

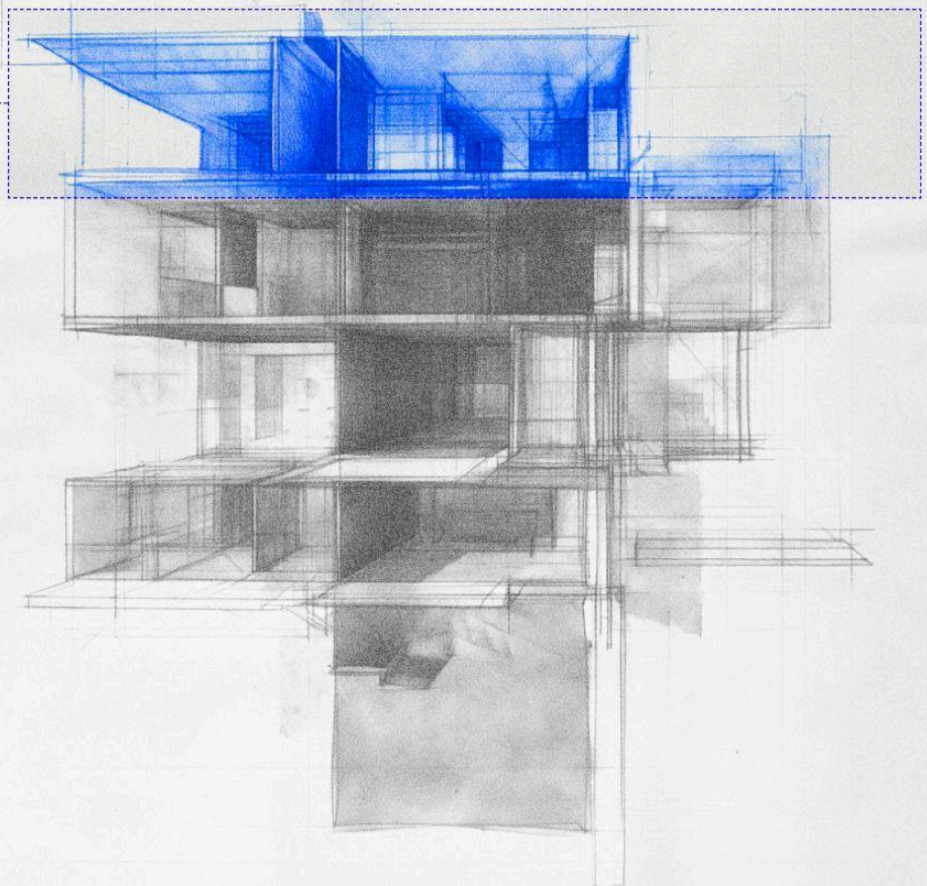
- The research aligns with the objectives of the relevant legislative instrument, such as the *Online Safety Act 2021*, and is broadly of public benefit. This excludes data concerning trade secrets;
- Funding for the research is fully disclosed;
- Access to the specific data requested and the indicated timeline is necessary and proportionate to the research purposes;
- Data security, confidentiality and personal data safety requirements will be met; and
- The research results will be publicly available free of charge within a reasonable period after completion.

The process for requesting data can be managed by the Australian Communications and Media Authority, the Office of the eSafety Commissioner or another appointed independent organisation. In addition, existing data and data tools, such as APIs, should be made available to Australian researchers free of charge, as they are in other markets.

[49] See Article 40 of *Digital Services Act 2022* (European Union), https://eur-lex.europa.eu/eli/reg/2022/2065/oj.

Five Necessary Elements
For Effective Digital Platform Regulation

Element Five: Accountability via Enforceability



November 2024

Reset.Tech Australia

Element Five: Accountability via Enforceability

What's This?

As seen with the current co-regulatory and mostly voluntary approach, where platforms have an outsized role in setting their own standards, the best interests of end users are not prioritised. Enforceability is key yet comparatively weak in Australia. International regulators possess a range of enforcement powers that are not currently available to the Office of the eSafety Commissioner to compel redress. Enforcement powers should include the ability to issue significant fines for failures to meet required improvements. Figure 6 highlights the scale of the fining regime available to comparable regulators.

Furthermore, strong last-resort measures are needed to prevent platforms from disregarding regulators' requests. International examples of last-resort measures include:

- Under the DSA, in cases of significant and persistent failures where attempts at engagement have failed, regulators can turn off services. Specifically, the DSA outlines that if an 'infringement has not been remedied or is continuing and is causing serious harm, and that infringement entails a criminal offence involving a threat to the life or safety of persons', regulators can work with domestic courts to order temporary restrictions of access.[50]
- Alternatively, under the UK OSA, with the agreement of the courts, Ofcom can require payment providers, advertisers and internet service providers to stop working with a site, preventing it from generating revenue or being accessed from the UK.[51]
- In extreme cases in the UK, criminal sanctions can be imposed on senior management if transparency measures are not met. The UK OSA requires companies to identify senior managers who are liable for responding to information notices. Failure to comply with an information notice request is a criminal offence.[52] These measures stand in stark contrast to Australian enforcement powers, where requests for information have been ignored and fines of $610,500 issued.[53]

[50] See Article 51(3) of *Digital Services Act 2022* (European Union), https://eur-lex.europa.eu/eli/reg/2022/2065/oj.
[51] UK Department for Science, Innovation and Technology, *Online Safety Act: Explainer* (2024), https://www.gov.uk/government/publications/online-safety-act-explainer/.
[52] For more information, see Department for Science, Innovation and Technology, *Online Safety Act: New criminal offences circular* (2024), https://www.gov.uk/government/publications/online-safety-act-new-criminal-offences-circular/online-safety-act-new-criminal-offences-circular.
[53] Georgie Hewson, 'Australia's eSafety commission fines Elon Musk's X $610,500 for failing to meet anti-child abuse standards' (2023) *ABC*, https://www.abc.net.au/news/202310-16/social-media-x-finedover-gaps-in-child-abuseprevention/102980590.

Figure 7: Scale of the fining regime available to comparable regulators

- Under the UK OSA, companies can be fined up to £18 million or 10% of their qualifying worldwide revenue, whichever is greater.
- Under the DSA, companies can be issued penalties of up to 6% of global annual turnover for failing to effectively mitigate risks and up to 1% of global annual turnover for supplying incomplete or misleading information as part of meeting transparency obligations.
- In Australia, regulators in adjacent domains of consumer protection and financial services have comparable fining abilities. For example, the ACCC can fine up to 10% of annual turnover for franchising violations,[54] and the ASIC can fine up to 10% of annual turnover, capped at $782.5 million, for violations of ASIC-administered legislation.[55]

Practical Implementation Guidelines

- To achieve accountability in Australia, online safety standards must be set by public institutions. Online safety standards continue to be effectively set by industry in Australia, which intentionally and significantly reduces the avenues for meaningful accountability. This probably eventuated as a legacy of co-regulation in telecommunications and broadcasting. Extending it to digital platform regulation is ineffective.
- The largely offshore nature of the worst digital offenders can make enforcement difficult. It is also a common legal technique for large digital platforms to claim that they do not do business in Australia. As Australia's online safety standards strengthen to become necessarily prescriptive in nature and consequential in enforcement powers, large platforms need to be effectively brought into the jurisdiction, most likely via an amendment to the *Online Safety Act 2021*. Additionally, to prove an effective deterrent, fining regimes and last-resort measures need to match the scale of the risk and the size of the company generating it.
- Examples of necessary accountability measures include:
    - Compelling redress and **changes** to platforms' systems and elements rather than just compelling transparency or takedown;
    - Issue penalties that match the scale of digital platforms' global profits;
    - Turn off services where failures are persistent and all other measures have been exhausted;
    - Enhance the public-facing complaints mechanism to include complaints from individuals and consumer groups regarding systemic risks and breaches of duty of care;
    - Have strong investigative and information-gathering powers; and
    - Have effective notice and takedown powers.

[54] Australian Competition and Consumer Commission, *Fines and penalties* (nd), https://www.accc.gov.au/business/compliance-and-enforcement/fines-and-penalties.
[55] Australian Securities and Investments Commission, *Fines and penalties* (2023), https://asic.gov.au/about-asic/asic-investigations-and-enforcement/fines-and-penalties/.

Conclusion

Australia urgently needs a comprehensive regulatory model that addresses the underlying systems of digital platforms rather than continuing to rely on content-based regulatory responses. What is needed is a regulatory model that includes *all five elements*, namely:

1. An overarching duty of care owed by digital platforms to Australian users. An overarching duty of care would place broad obligations on platforms to ensure user safety in systemic ways. Specific responsibilities may be enumerated by focusing on requirements for risk assessments.

2. Requirements for platforms to assess all their systems and elements for a defined set of risks. These could include, for example, risks related to; children and young people, and their best interests; end-uses mental wellbeing; the instruction and promotion of harmful practices such as suicide; risks of illegal content, conduct and activit distributing illegal materials, and; risks to personal safety and security, such as scams and data breaches. Such requirements would incentivise systemic change and help platforms realise their duty of care.

3. Requirements for platforms to mitigate each risk. As a corollary of risk assessments, platforms must be required to implement reasonable steps to mitigate each identified risk. These measures must be included in the assessments sent to regulators.

4. Five sources of transparency. These include annual risk assessments, prescriptive public transparency reports, independent audits of risk assessments and transparency reports, data portals for ad repositories and content moderation decisions, and researcher access to public interest data. These need to exist alongside strong investigative powers for regulators.

5. Enforceable regulations and empowered regulators to compel behavioural change. This means regulators are empowered and resourced to:
    - Compel redress and changes to platforms' systems and elements rather than just compel transparency or takedown;
    - Issue penalties that match the scale of digital platforms' global profits;
    - Turn off services where failures are persistent and all other measures have been exhausted;
    - Enhance the public-facing complaints mechanism to include complaints from individuals and consumer groups regarding systemic risks and breaches of duty of care;
    - Have strong investigative and information-gathering powers; and
    - Have effective notice and takedown powers.

Additionally, a reformed *Privacy Act* that protects Australians from predatory digital business practices is essential. Reforms must offer meaningful protections for personal data, including metadata, and impose strict requirements regarding fairness and reasonableness to justify data processing by digital platforms and social media. In particular, these requirements should address the market structure and dynamics of harmful digital business models, as highlighted in the recent ACCC report on data firms.[56] The proposals outlined in the *Privacy Act Review Report*[57] are robust and a step in the right direction. These are crucial to safeguard risks to personal and national security.

---

[56] Australian Competition and Consumer Commission, *Digital platform services inquiry interim report 8: Data products and services – How information is collected and used by data firms in Australia* (2024), https://www.accc.gov.au/system/files/Digital-platform-services-inquiry-March-2024-interim-report.pdf.
[57] Attorney General's Department, *Government response to the Privacy Act Review Report* (2023), https://www.ag.gov.au/rights-and-protections/publications/government-response-privacy-act-review-report.

November 2024

Reset·Tech
AUSTRALIA

Five Necessary Elements
For Effective Digital Platform Regulation