Reset.
AUSTRALIA

Response to the House Select Committee Inquiry into Social Media and Online Safety

Reset Australia
Jan 2022

Prepared by:
Dr Rys Farthing
Dhakshayini Sooriyakumaran
Elena Yi-Ching Ho

Executive summary

Five overarching policy directions should be considered to improve our regulatory framework to protect all Australians from the harms of social media platforms. This involves a focus on:

1. **Systems and processes:** Focus regulation on eliminating risks from systems and processes, expanding on our current focus on content moderation
2. **Community and societal risk:** Expand regulations to addresses community & societal risks, building on our comprehensive approach to Individual risks
3. **Platform accountability and transparency:** Ensure regulation creates accountability & transparency, rather than placing burden on individuals
4. **Comprehensive regulation:** Ensure the regulatory framework is comprehensive, by improving our current regulatory gaps and disjunctures
5. **Strong regulators and enforced regulation:** Ensure regulation is strong and enforced, by moving away from self- and co-regulation and resourcing and joining up regulators

This would create a more streamlined approach to regulation, replacing multiple disjointed obligations with more aligned upstream duties, reducing the regulatory burden on Australia's successful tech industry. It would also be interoperable with emerging international requirements, ensuring Australian industry could expand into international markets with minimal regulatory friction.

Contents

1. About Reset Australia & this submission

Reset Australia is an independent, non-partisan policy think tank committed to driving public policy advocacy, research, and civic engagement to strengthen our democracy within the context of technology. We are the Australian affiliate of Reset, a global initiative working to counter digital threats to democracy.

This submission has been prepared in response to the House Select Committee on Social Media and Online Safety's inquiry into Social Media and Online Safety announced in Dec 2021. It outlines Reset Australia's broader thinking around the types of legislation and regulation that the Australian government could consider, and responds to some of the Terms of Reference as published by the Committee.

2. Five policy directions for regulating Social Media in Australia

When powerful forces 'move fast and break things', they can leave a wake of destruction in their path. Social media platforms in Australia have too often done exactly that. These platforms have wreaked havoc on our public square, leaving people facing a myriad of risks from algorithmic bias to harmful content. Our institutions, such as Parliament and the press, have also been exposed to harmful mis/disinformation and arbitrary shut downs. The lack of precautionary planning and the speed of growth of social media companies has often destabilised the Australian community, with serious online and offline consequences.

Yet, this rapid change has also created a digital world ripe with opportunities, generating innovations that strengthen the economy and improve our lives. In 2021, the Australian tech sector contributed $167bn to the economy[1], and kept many families and children connected, working and learning during the pandemic. The digital world can be a force for good, and the impact of the technology sectors, if directed and regulated, can be transformative.

Reset Australia welcomes this inquiry. Now is the time for the Government to rethink regulation of the digital world, and how these could work better for all Australians. Australia was an early mover in online safety, with the appointment of the eSafety Commissioner and the landmark ACCC Digital Platforms Inquiry. Our approach is now ready for a refresh. We have a strong regulatory framework that addresses some harms of social media, in some sectors, that can be built on and further refined and developed.

We believe that five overarching policy directions should be considered to build a regulatory framework to protect all Australians, whilst continuing the growth of a vibrant Australian tech sector. These are described below.

[1] Tech Council & Accenture 2020 *The Economic Contributions of Australia's Tech Sector* https://techcouncil.com.au/wp-content/uploads/2021/08/TCA-Tech-sectors-economic-contribution-full-res.pdf

2.1 Eliminating risks from systems and processes

Regulation should pivot towards targeting risks created across the systems and processes developed by digital services. The aspects of systems and processes, and related risks, that regulation could address includes:

- **Algorithms.** These drive much of the content delivery in social media platforms, both in terms of content and advertising. For example, YouTube estimates that 70% of content viewed on their platform is as a result of their recommender algorithm and autoplay. These systems, designed by platforms, often using machine learning or other AI technologies, often promote risky or harmful content. Yet algorithms are not trained in ways that consider risks.

- **Platform design.** The user interface and user experiences of social media platforms are highly curated and engineered: each design element reflects a decision point made by a company. Platforms can be designed in ways that create risks. For example, many platforms design their user journey in ways that maximise data extraction from your device, social media apps and other internet based activity. For example, apps that nudge you to, or automatically connect with, your address book or track your GPS location. This data is used to preferences and interests, and personalise your ad experience (termed 'surveillance advertising'). This is a 'dark pattern' that maximises profits but does not consider the data risks it creates, in subtle and persuasive ways.

- **Specific features.** Specific features can also create risks. For example, features that enable the live broadcasting of locations, or photo filters that make people appear thinner. These features can combine in ways that amplify or create new risks. For example, video live streaming and the ability to receive messages from stranger's accounts creates unique risks for young users[2]. Features are developed and refined by platforms to meet identified priorities, such as maximising engagement, growing reach or extending the amount of time users stay on a platform. These priorities often do not consider risk; if 'minimising risk' was a systemic design aim many features would operate differently or be abandoned.

These sorts of systems and processes manufacture and amplify risks but none of them are inevitable. Social media platforms can change and improve their systems, and regulation can encourage them to do so.

Regulatory approaches that take a more narrow focus on content moderation (focusing on takedown/deletion of harmful or illegal content for example) are not systemic enough, nor are they commensurate with the scale of the problem at hand. They doom regulators to a perpetual game of content 'whack-a-mole' on an impossible scale.

Australia's existing *Online Safety Act* focuses largely on content, but through the Basic Online Safety Expectations and the industry codes developed as part of this may address some systemic risks. However, co-regulatory codes and guidance from regulators will not be adequate to create the scale of change needed to ensure safety. These risks are simply too important to leave up to industry to address — whose business models incentivise and

[2]See for example, The Times' investigation in grooming via YouTube livestreams. Harry Shukman 2018 'Predators coax children into exposing themselves' *The Times* www.thetimes.co.uk/article/predators-coax-children-into-exposing-themselves-lfws0fjdp

reward risky systems. Nor will the proposed Codes cover all of the systems and processes that need to be addressed. A more comprehensive approach is needed to ensure the regulatory framework is fit for newly emerged and emerging technologies.

Case study: How a risk focused, systemic approach worked to protect children

Regulations that remove risks from systems have already reduced risks for children and young people. Without regulating content, the UK's *Age Appropriate Design code* led to 'upstream' risk reductions such as:

- Defaulting children's accounts to private. In the 8 months leading up to the enforcement of the UK's code, TikTok announced that it was defaulting all users aged 13-15 to private accounts[3], Facebook announced that 'everyone who is under 16 years old (or under 18 in certain countries) will be defaulted into a private account when they join Instagram[4]' and Google announced that it would 'gradually start adjusting the default upload setting to the most private option available for users ages 13-17 on YouTube[5]'
- Reducing the ways advertisers themselves could micro-target commercial advertising at children. Google announced it was blocking microtargeting based on age, gender or interests of people under 18[6], and Meta limit the ability of advertisers to select children to target, allowing selected targeting based on age, gender and geography[7], and in-platform tracking to personalise ads[8]
- Turning off 'Autoplay' by default features which can see children 'nudged' into watching more content than they intended. For example, Google and subsidiary YouTube announced they would turn off Autoplay for those under 18[9]

[3] Eric Han 2021 'Strengthening privacy and safety for youth on TikTok' newsroom.tiktok.com/en-us/strengthening-privacy-and-safety-for-youth
[4] Instagram 2021 'Giving young people a safer, more private experience' about.instagram.com/blog/announcements/giving-young-people-a-safer-more-private-experience
[5] James Beser 2021 'New safety and digital wellbeing options for younger people on YouTube'' blog.youtube/news-and-events/new-safety-and-digital-wellbeing-options-younger-people-youtube-and-youtube-kids/
[6] James Beser 2021 'New safety and digital wellbeing options for younger people on YouTube' blog.youtube/news-and-events/new-safety-and-digital-wellbeing-options-younger-people-youtube-and-youtube-kids/
[7] Instagram 2021 'Giving young people a safer, more private experience' about.instagram.com/blog/announcements/giving-young-people-a-safer-more-private-experience
[8] Elena Yi-Ching Ho & Rys Farthing 2021 *How Facebook are still targeting teens with advertising* Reset Australia / Fairplay https://fairplayforkids.org/wp-content/uploads/2021/11/fbsurveillancereport.pdf
[9] James Beser 2021 'New safety and digital wellbeing options for younger people on YouTube' blog.youtube/news-and-events/new-safety-and-digital-wellbeing-options-younger-people-youtube-and-youtube-kids/

Eliminating risks from systems & processes: examples from other jurisdictions

The UK government described their draft Online Safety Act as 'a "systems and processes" bill — aimed at addressing systemic issues with online platforms rather than seeking to regulate individual content'. The final act is expected to focus on the 'content and activity' of platforms[10]. The bill achieves this by creating multiple and overlapping 'duties of care' for service providers, including[11]:

- Duties to reduce illegal content risks, such as:
  - Undertake an illegal content risk assessment
  - Taking proportionate steps to mitigate and effectively manage risks identified in illegal content risk assessments
  - A duty to operate using proportionate systems and processes designed to minimise the presence, duration of presence and dissemination of illegal content
- Duties to regard freedom of expression and privacy set, such as:
  - Impact assessment about free expression & privacy, when deciding safety policies & procedures
  - Likewise, impact assess existing policies & procedures
  - Duties to act on risks identified in these assessments
  - Keep impact assessments up to date, & be publicly available (or summaries)
- Duties about reporting and redress, with similar obligations
- Record-keeping and review duties, with similar obligations

In a similar vein, the EU's proposed Digital Services Act (DSA)will oblige platforms to[12]:

> *'assess the systemic risks stemming from the functioning and use of their service, as well as by potential misuses by the recipients of the service, and take appropriate mitigating measures'*

Recital 58 of the DSA goes on place requirements on platforms to diligently mitigate risks identified in the risk assessment, by for example:

> *'enhancing or otherwise adapting the design and functioning of their content moderation, algorithmic recommender systems and online interfaces, so that they discourage and limit the dissemination of illegal content, adapting their decision-making processes, or adapting their terms and conditions. They may also include corrective measures, such as discontinuing advertising revenue for specific content, or other actions, such as improving the visibility of authoritative information sources. ... They may also initiate or increase cooperation with trusted flaggers....'*

[10] Joint Committee on the Draft Online Safety Bill 2021 *Draft Online Safety Bill: Report of Sessions 2021-22* https://committees.parliament.uk/publications/8206/documents/84092/default/
[11] *Draft Online Safety Bill* 2021, UK https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/
[12] Recital 56 European Commission 2020 *Proposal For A Regulation of the European Parliament & of the Council on a Single Market For Digital Services (Digital Services Act) & Amending Directive* 2000/31/ https://oeil.secure.europarl.europa.eu/oeil/popups/ficheprocedure.do?lang=en&reference=2020/0361

Regulating content recommender algorithms: examples from other jurisdictions

In the US, multiple legislative changes have been proposed to regulate algorithms. The latest of these, the Justice Against Malicious Algorithms Act was put before the House in Oct 2021[13]. It proposes a specific reform that creates liability for platforms for any harms caused by content where that content was recommended by the platform. This holds them accountable where algorithmic recommendations have been knowingly reckless.

The Digital Services Act[14] requires platforms to assess and mitigate risks emerging from their algorithms in the first instance. Recital 62 outlines why this is important stating that:

> *'Recommender systems can have a significant impact on the ability of recipients to retrieve and interact with information online. They also play an important role in the amplification of certain messages, the viral dissemination of information and the stimulation of online behaviour'.*

Regulating ad delivery systems: examples from other jurisdictions

The DSA requires risk assessments and risk mitigation Advertising Systems[15], and requires transparency measures around promoted ads[16]. Recital 52 outlines why[17]:

> *'online advertisement can contribute to significant risks'* [a proposed amendment suggests noting that the risks are economic and political[18]] '*ranging from advertisement that is itself illegal content, to contributing to financial incentives for the publication or amplification of illegal or otherwise harmful content and activities online, or the discriminatory display of advertising'*

In the US, there is currently a petition with the FTC to ban surveillance advertising

---

[13] US House of Representatives *H.R.5596 - Justice Against Malicious Algorithms Act of 2021* www.congress.gov/bill/117th-congress/house-bill/5596

[14] European Commission 2020 *Proposal For A Regulation of the European Parliament & of the Council on a Single Market For Digital Services (Digital Services Act) & Amending Directive* 2000/31/ https://oeil.secure.europarl.europa.eu/oeil/popups/ficheprocedure.do?lang=en&reference=2020/0361

[15] It also places obligations on platforms to provide users with information and choice about the way ad delivery systems target them (with explicit consent needed to 'opt-in' to surveillance advertising). While informed choice and consent are necessary, they are not sufficient for a robust regulatory response

[16] Recital 63 European Commission 2020 *Proposal For A Regulation of the European Parliament & of the Council on a Single Market For Digital Services (Digital Services Act) & Amending Directive* 2000/31/ https://oeil.secure.europarl.europa.eu/oeil/popups/ficheprocedure.do?lang=en&reference=2020/0361

[17] European Commission 2020 *Proposal For A Regulation of the European Parliament & of the Council on a Single Market For Digital Services (Digital Services Act) & Amending Directive* 2000/31/ https://oeil.secure.europarl.europa.eu/oeil/popups/ficheprocedure.do?lang=en&reference=2020/0361

[18] European Parliament 2021 *Opinion of the Committee on Economic and Monetary Affairs for the Committee on the Internal Market and Consumer Protection on the Digital Services Act* www.europarl.europa.eu/doceo/document/ECON-AD-693929_EN.pdf

practices[19]. No specific legislation has been proposed.

2.2 Expand regulations to address community & societal risks

The risks addressed by existing legislation are too narrow, and this leaves Australians vulnerable to collective risks. Collective risks come in two interconnected forms.

Firstly, there are risks posed to specific communities, such as indigenous communities, migrant communities, people of colour, women and LGBTIQ+ people. These communities often suffer unique and disproportionate harms in the digital world. While some of the risks they face may be addressed by regulation around individual harms, an 'offensive-piece -of-content' by 'offensive-piece-of-content' approach can miss the collective nature of the problem. Disinformation and hate speech can affect particular communities in ways that differ from individual harm.

Secondly, platforms create societal risks. The scale and reach of social media platforms has the capacity to influence and affect Australian institutions, such as Parliament, the Press and healthcare systems, often with destabilising effects. For example, we have seen how social media platforms have been used to undermine public health messaging around vaccine roll out (often in ways with particular consequences for marginalised communities), and foreign bots engaged in Australian electoral discussions. This is not the stuff of 'conspiracy theories'; a 2021 Senate hearing revealed that Australia has been the target of a number of sophisticated foreign disinformation campaigns, including a network linked to marketing firms based in the UAE, Nigeria and Egypt, all enabled by platforms[20]. A further list of examples of electoral and other disinformation in Australia is provided in appendix one.

Expanding the definitions of harms (and risks) addressed in Australia's regulatory framework would better protect Australian communities and society at large. This means tackling mis and disinformation, and explicitly addressing hate speech. Currently mis/disinformation is covered by a co-regulatory Code that has been widely criticised as 'not meeting expectations' including by regulators[21].

Case study: Societal risks in election processes

A QUT study which examined around 54,000 accounts during and after the 2019 Australian Federal Election (looking at over 1 million tweets) revealed that 13% of accounts were 'very

[19] Federal Trade Commission 2021 *Filed Before the Federal Trade Commission Washington, D.C. 20580 Re: Petition For Rulemaking To Prohibit Surveillance Advertising 12/03/202*1 www.ftc.gov/system/files/attachments/other-applications-petitions-requests/r207005_-_petition_for_rule_to_prohibit_surveillance_advertising_0.pdf

[20] Select Committee on Foreign Interference Through Social Media, Senate, 30 July 2021

[21] Zoe Samios & Lisa Visentin 2020 'ACMA: Tech giants' code to handle fake news fails to meet expectations' *SMH* www.smh.com.au/politics/federal/acma-tech-giants-code-to-handle-fake-news-fails-to-meet-expectations-20201026-p568oq.html

likely' to be bots, with the majority originating from New York[22]. This is estimated to be more than double the rate of bot accounts in the US presidential election.

These can have big impacts: research into the US election by ANU indicated that the average bot was 2.5 times more influential than the average human, measured by success at attracting exposure via retweets[23].

Case study: Societal and community risks through mis/disinformation

Chinese Australians have faced misinformation in the past, often in what appear to be coordinated disinformation campaigns[24]. Social media platforms, such as WeChat, Weibo and Douyin have been found to serve targeted misinformation to Chinese language speakers in Australia.

In 2019, WeChat in particular was a site of much political campaigning in Mandarin which often included misinformation, and coordinated sharing that could be categorised as disinformation[25]. One MP described it as 'malicious false content[26].

Outside of Australia, coordinated misinformation campaigns were deployed in an apparent attempt to deter Chinese Americans from voting[27].

Addressing societal & community risks: examples from other jurisdictions

Canada, the EU and Germany are moving towards frameworks that address community and societal risks, as well as individual risks. Figure one documents these.

Recital 57 of the Digital Services Act explicitly describes the three types of systemic risks that platforms must assess and mitigate, which include community and societal risks[28]:

1. Risks associated with the misuse of their service through the dissemination of illegal content, such as CSAM, hate speech, and the conduct of illegal activities. This includes

[22] Felicity Caldwell 2019 *'Bots stormed Twitter in their thousands during the federal election'* SMH www.smh.com.au/politics/federal/bots-stormed-twitter-in-their-thousands-during-the-federal-election-20190719-p528s0.html
[23] Sherryn Groch 2018 '*Twitter bots more influential than people in US election: research'* SMH www.smh.com.au/national/twitter-bots-more-influential-than-people-in-us-election-research-20180913
[24] Lawson 2020 'WeChat the channel for China disinformation campaigns' *Canberra Times* www.canberratimes.com
[25] Lawson 2020 'WeChat the channel for China disinformation campaigns' *Canberra Times* www.canberratimes.com
[26] Paul Karp 2019 'Penny Wong Blast WeChat' *the Guardian* www.theguardian.com/australia-news/2019/may/07/penny-wong-blasts-malicious-wechat-campaign-spreading-fake-news-about-labor
[27] Joe Fitzgerald Rodriguez & Shannon Lin 2020 'Misinformation Image on WeChat' *ProPublica* www.propublica.org/article/misinformation-image-on-wechat-attempts-to-frighten-chinese-americans-out-of-voting
[28] Recital 57, European Commission 2020 *Proposal For A Regulation of the European Parliament & of the Council on a Single Market For Digital Services (Digital Services Act) & Amending Directive* 2000/31/ https://oeil.secure.europarl.europa.eu/oeil/popups/ficheprocedure.do?lang=en&reference=2020/0361(COD)

risks created where content may be amplified by platforms to an especially vast audience. These are largely, but not exclusively, individual risks
2. Risks that affect people's rights. This includes the design of the algorithmic systems and the misuse of their service through the submission of abusive notices or other methods for silencing speech or hampering competition. These would include community risks as we have described them
3. The use of a platform to share disinformation that has a foreseeable impact on health, civic discourse, electoral processes, public security and children's safety. This includes mitigating against fake accounts and bots. These would include societal risks as we have described them

Likewise, the proposals for Canada's new regulatory framework will be explicitly based on the premise that platforms are often 'used to share content depicting real-world acts of violence in an effort to incite violence, intimidate the public or segments of the public, and damage societal cohesion'[29].

The jurisdictions that explicitly address hate speech and disinformation in their legislation are documented in table one.

[29] See Module 1(A)e, Government of Canada 2021 *Technical paper: The Government's Proposed Approach to Addressing Harmful Content Online* www.canada.ca/en/canadian-heritage/campaigns/harmful-online-content/technical-paper.html#a2. Note: these were tabled before Canada's 2021 snap election

	EU	Canada	Germany	UK	Ireland	Australia
Key legislation addressing harms	Digital Services Act (in draft)	Online safety proposals (in draft)	NetzDG, and others (passed)	Online Safety Bill (in draft)	Online safety & Media Regulation (in draft)	Online Safety Act (passed)
Definition of Harm, Individual, Community or Societal	No set definition, the focus is on harms that violate rights. This will include societal harms, and community harm through hate speech	Individual (aligned to existing definitions of hate speech) Societal (damage to societal cohesion, vulnerable groups)	Based on existing criminal law. This includes Individual and some community harms through hate speech	Individual (Content having an adverse physical or psychological response on adults of children)	Individual (Illegal content, individually intimidating or threatening content, eating disorder, self harm & suicide content)	Individual (content that is "offensive" to adults or children, content that is refused classification etc)
Systems Vs Takedown	Systems + Takedown	Takedown	Takedown	Systems + Takedown	Systems + Takedown	Takedown (+ potentially some systems through co-regulatory Codes)
Content In Scope	Illegal + indirectly, legal Disinfo included indirectly Hate speech indirectly included	Illegal Disinfo out of scope Hate speech in scope	Illegal Disinfo out of scope Hate speech in scope	Illegal + legal List of harms to be added later but unclear whether disinfo & hate speech is in scope (could be in scope where content is harmful to adults)	Illegal + legal Disinfo out of scope Individual hate speech content could be in scope, where it intimidates, threatens, humiliates or persecutes	Illegal + legal Disinfo out of scope Individual hate speech content could be in scope, where it causes offence to an individual or would be considered menacing, harassing or offensive
Services In Scope	Intermediary services e.g. ISPs and online platforms Private messaging out of scope	Social media Private messaging out of scope	Social media	Services which host or facilitate UGC, apart from news media outlets. Private messaging in scope.	Broad range of platforms and services inc press publications which enable UGC Private messaging in for criminal content	Social media services, Relevant electronic service and ISPs (Tight definition of "social media")

Powers Of Regulator	Fines Information gathering powers Algorithmic audit mandatory	Information gathering powers Inspection powers No algorithmic audit	Fines	Fines Information gathering powers Language seems to allow algorithmic inspection	Fines Information gathering powers. No algorithmic audit	Fines Offers public facing complaint mechanisms, Investigation, Audit (not algorithmic)
Independence Of Regulator	Independent as well as EC oversight of large platforms	Independent Creates Digital Safety Commissioner and Digital Recourse Council of Canada,	Independent	Independent however OSB keeps provisions for political agenda setting	Independent Creates Online Safety Commissioners	Independent
Transparency	Six monthly transparency reports (publicly published) Data access for pre-vetted researchers	Transparency reporting inc data on takedown volumes and processes.		Annual transparency reports No data sharing provisions	Periodic transparency reporting	Transparency reporting

Figure one: Comparative approaches to types addressing harms through regulation

2.3 Ensure regulation creates accountability & transparency

There are multiple ways governments can regulate the digital world, but the most effective policies require accountability and transparency from tech platforms themselves. Regulations that identify the core risks as stemming from platforms themselves — and squarely place the burden of responsibility on digital services — should be prioritised.

Regulation can place duties on users in multiple ways, but these are often inappropriate or ineffective:

- Solutions that position individual users (especially children and parents) as key actors in the frontline of improving safety are often inappropriate and will fail to protect all Australians. The scale of the risks created by platforms exceed the capability of individuals to effectively manage in isolation, especially for children. The ability to 'change settings', 'effectively report content' or 'turn on safe search' will not be enough. User's informed choice around settings and options is necessary, but it is not sufficient to ensure safety, particularly for those lacking the capabilities or support to do so

- Solutions that pass responsibility on to users (as parents or consumers) to read 'the fine print' or consent to a risky system misrepresents the power asymmetry between users and tech companies. The nature of the global digital architecture, and its utility in everyday life, means that withdrawing consent is not a viable option for most Australians. For example, 75% of the world's most popular million websites have google analytics and trackers built into them[30]. A 'buyer beware' approach will fail where users have no viable alternatives

- Solutions that position individual users (be they 'trolls' or influencers) as the key actors responsible for harm undersells the role of platforms in creating the risky digital environments that enable and encourage toxic actors. Platforms manufacture and amplify harmful content; they hand trolls and other bad actors the tools they need to cause harm and provide incentives, including funding[31], to encourage their ongoing poor behaviour

Accountability means that platforms themselves should have responsibilities to mitigate risks, and should be held to account where they fail and harm occurs.

Accountability also requires transparency. Part of the problem of making social media safe is that legislators, regulators, researchers and civil society often do not know enough about the specific mechanics of how platforms work nor their consequences. Requiring transparency through, for example, algorithmic audits and impact statements could help remedy this.

---

[30] Steven Englehardt & Arvind Narayanan 2016 'Online Tracking: A 1-million-site Measurement & Analysis' *CCS '16: Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security* doi.org/10.1145/2976749.2978313

[31] Karen Hao 2021 'How Facebook and Google Fund Global Misinformation' *MIT Technology Review* www.technologyreview.com/2021/11/20/1039076/facebook-google-disinformation-clickbait/

Regulating for accountability: examples from other jurisdictions

International regulation is increasingly driving towards holding tech companies responsible for any risks that create, and accountable for any harms that occur. For example, the UK's Online Safety Bill uses the principle of a 'duty of care' to create accountability.

The proposal for a Duty of Care in tech regulation was extensively developed Professor Lorna Woods and Will Perrin at the Carnegie Trust, and has support from Australian academics such as Katharine Gelber at QUT[32]. Broadly speaking, the proposals suggest that service providers should be held responsible for their online spaces in the same way that property owners are responsible for physical spaces, and that service providers should have a duty of care to those using their platforms. Professor Woods' has argued that a statutory duty of care would be 'simple, broadly based and largely future-proof', much like long-enduring occupational health and safety regulations which have adopted this approach[33].

The emerging Online Safety Bill in the UK places multiple duties of care on regulated services to reduce the risks in their content and operations[34]. These include risk assessment and mitigation processes, as well as transparency and accountability requirements.. Combined, these duties oblige:

> *'service providers to do particular things, such as undertake risk assessments, to comply with safety duties in respect of illegal content, content that is harmful to children and content that is harmful to adults and other duties, for example in respect of journalistic content. ... They are things that providers are required to do to satisfy the regulator. They are not duties to people who use their platforms, and they are not designed to create new grounds for individuals to take providers to court.'*

Regulating for transparency: examples from other jurisdictions

There is an emerging consensus that regulation needs to require transparency from social media platforms. For example:

---

[32] Katharine Gelber 2021 'A better way to regulate online hate speech: require social media companies to bear a duty of care to users' *The Conversation* https://theconversation.com/a-better-way-to-regulate-online-hate-speech-require-social-media-companies-to-bear-a-duty-of-care-to-users-163808

[33] Lorna Woods & William Perrin UK 2019 *Online harm reduction – a statutory duty of care and regulator* Carnegie Trust *https://www.carnegieuktrust.org.uk/publications/online-harm-reduction-a-statutory-duty-of-care-and-regulator/*

[34] *Draft Online Safety Bill* 2021, UK https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/985033/Draft_Online_Safety_Bill_Bookmarked.pdf

- UK: The Online Safety Bill[35] proposes Increased information gathering powers for regulators (cl70) and investigations powers (cl75) including power to interview and enter and inspection, and annual transparency reports

- US: The proposed Platform Accountability and Consumer Transparency Act[36] suggests compelling social media platforms to share data with researchers (sec 4) and establish a regulatory Commission on transparency (sec 3)

- EU: The Digital Service Act [37] proposes that regulators and vetted academic researchers must be able to access data from large platforms, and that platforms must produce six-monthly transparency reports that will be publicly published (recital 60 and others)

2.4 Ensure the regulatory framework is comprehensive

The rapid growth of the technology has seen Australia's issue-by-issue (e.g. 'cyber bullying', 'image-based abuse' etc), sector-by-sector (e.g. 'social media platforms' 'messaging services' etc) regulatory framework struggle to keep pace. Many new and emergent technologies are missed, and innovative companies straddling the gaps between existing industry definitions are inappropriately regulated.

A. Gaps between industries and services

Australian regulation often takes a sector-by-sector approach, which can fail to adequately address the shared functionalities and integration between the social media sector and multiple other industries. The most obvious of these issues is the integration of traditional media and social media platforms, but equally complicated functionalities exist between social media platforms and data brokers, other online services, the advertising sector, the broader telecommunications industry, and increasingly emergency services and health and social care services as they become central to public messaging campaigns (among others).

Current legislation oftens fails to reflect these integrations and diverse functionalities. Using Roblox, an online kids game, as an example highlights the sorts of peculiarities this can lead to. The current definition of a 'social media' company (as laid out in the *Online Safety Act* and the proposed Enhancing Online Privacy Bill) would fail to cover Roblox. Roblox allows the creation of personal avatars; facilitates and encourages interaction and communication between users, and; allows users to create and share games for others to play. Because users do not post content *per se* — they 'post' games — they are unlikely to be considered a social media platform under the existing definition. Roblox is however covered under the *Online Safety Act* as a 'relevant electronic service' as it facilitates messaging and game play between users. But it would not be covered by the proposed Enhancing Online Privacy Act, unless 2.5

[35] *Draft Online Safety Bill* 2021, UK https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/
[36] *Platform Accountability and Consumer Transparency Act 2021, USA* www.congress.gov/bill/116th-congress/senate-bill/4066/all-info
[37] Recital 60, European Commission 2020 *Proposal For A Regulation of the European Parliament & of the Council on a Single Market For Digital Services (Digital Services Act) & Amending Directive* 2000/31/ https://oeil.secure.europarl.europa.eu/oeil/popups/ficheprocedure.do?lang=en&reference=2020/0361

million Australians logged on making it a 'large online platform' under the Bill. Roblox is a platform used mostly by kids, only 30% of their global audience is over 16[38]. There are a total of 3.6m 5-16 year olds in Australia[39], making the 2.5m threshold for coverage improbable. What does this mean? Because Roblox allows users to 'create and share games' rather than 'create and share content', kids may be protected against cyberbullying with regulation, but may not be protected from exploitative data practices or privacy incursions unless almost all Australia's younger kids join the platform, in which case they then may be protected with regulation again. This is regulatory bingo.

Scope limitations can also create regulatory discrepancies between the digital and non-digital world. For example, the ACCC's Digital Platform Inquiry explored the patchwork of regulations that applied to digital publishers compared to telecoms, radio comms and broadcast industries[40]. The ACCC found that despite providing comparable services, digital platforms often feel outside the scope of existing regulation, such as press statements of principles or legislation around advertising gambling and medicine.

Likewise, some exemptions in Australian regulation have not kept pace with the changing digital world. The use of turnover thresholds, such as exempting business above or below $3m or $10m annual turnover, are blunt and do not reflect the emerging nature of the tech industry. In particular, small tech startups can create significant risks for Australians but can often be overlooked. For example, the *Privacy Act* places obligations on businesses with an annual turnover of over $3m but exempts those under this threshold — even those handling significant amounts of data. Blunt exemptions often miss risks, and can leave large companies with safe practice facing a high regulatory burden, while extremely risky small companies continue to deliver harmful products and services.

B. Gaps in emergent technologies

Likewise, the issue-by-issue, sector-by-sector approach cannot anticipate risks created by innovations and emergent technologies. This has left recent innovations unregulated in Australia, including for example; surveillance advertising and ad delivery systems; AI; blockchain and its integration across systems, and neural technologies.

These gaps suggest that the current approach is unable to future-proof the regulatory framework, and that as technologies evolve, more and more gaps will emerge. Risk focused, systemic models may be more successful at future proofing themselves.

Australians use a wide range of digital services, and seamlessly move between technologies, sectors and companies of all sizes. Their safety should be ensured across their whole digital ecosystem. Gaps and exclusions within Australian regulation have often left Australians reliant on foreign legislation for protection.

---

[38] Jessica Clement 2021 *Distribution of Roblox games users worldwide as of September 2020, by age* www.statista.com/statistics/1190869/roblox-games-users-global-distribution-age/
[39] Australian Bureau of Statistics 2021 *National, state and territory population, Jun 2021* www.abs.gov.au/statistics/people/population/national-state-and-territory-population/latest-release#data-downloads-data-cubes
[40] Table 4.1, ACCC 2019 *Digital Platforms Inquiry: Final Report* www.accc.gov.au/system/files/Digital%20Platforms%20Inquiry%20-%20Final%20report%20-%20part%201.pdf

Case study: Protection children comprehensively through the UK's Age Appropriate Design Code

The UK's Age Appropriate Design Code 2020[41] aims to improve the collection and use of children's data and adopts a risk-based, and rights realising approach. The Code details 15 standards that services must meet (ranging from the best interests principle, to requiring impact assessments) and explicitly applies to "information society services likely to be accessed by children".

The definition of Information Society Service covers the vast majority of for-profit online services including; apps and programs; search engines; online messaging or VOIP[42] services; content streamers (eg video, music or gaming services); online games; news or educational websites; connected toys and other connected devices with electronic controllers, as well as; social media platforms.

Services "likely to be accessed" by children are in scope, even if they don't target children. This means unless they *effectively* exclude children, all digital services are covered.

The result is a Code that will protect children across the breadth of digital services they use in interoperable ways. Comparable regulations exist around the world that have taken different approaches:

- Australia's draft Enhancing Online Privacy Act targets social media platforms for provisions relating to young people, but excludes other large platforms, data brokers and a host of other services that may engage in risky practices[43]
- America's *Child Online Privacy and Protection Act* (COPPA) only applies to services that have actual knowledge of users being under 13[44]. As long as a service remains unaware of children using their platforms, COPPA does not apply. Only three years ago, in 2019 YouTube settled a complaint with the FTC about violations that may have emerged from this distinction[45]

2.5 Ensure regulation is strong and enforced

Big tech poses big risks and necessitates a robust regulatory response. However, because Australia has to date engaged self- and co-regulatory models by default, our regulatory framework has often failed to reduce risks as rigorously as they otherwise may have.

Future regulation needs to start from the premise that self- and co-regulation will not be sufficient for the social media sector. Reset Australia believes self- and co-regulation have a

[41] *Age Appropriate Design Code*, UK 2020 https://ico.org.uk/for-organisations/
[42] Voice over Internet Protocol
[43] See for example, Reset Australia 2021 *Response to the draft Enhancing Online Privacy Act 2021*
[44] Children's Online Privacy and Protection Act, US 1988 www.ftc.gov/enforcement/rules/rulemaking-regulatory-reform-proceedings/childrens-online-privacy-protection-rule
[45] FTC 2019 *Google & YouTube Will Pay Record $170m for Alleged Violations of Children's Privacy Law* www.ftc.gov/news-events/press-releases/2019/09/google-youtube-will-pay-record-170-million-alleged-violations

role to play in the Australian regulatory landscape at large, but that unfortunately the risks posed by the digital environment are:

- High impact, and include significant public health and community safety concerns
- Significant to the community, and the public has an appetite for the certainty of robust regulations
- Unable to be adequately dealt with by lighter touch regulations. The social media sector has demonstrated a track record of systemic compliance issues, including multiple breaches of existing legislation and a generally anemic response to self-regulation

This warrants a pivot towards primary and subordinate legislation and regulation for the sector. Appendix two documents our rationale for recommending an end- to self and co-regulation for the tech sector in more detail.

Alongside strengthening existing regulation, regulators need to be resourced and enabled to enforce this. This includes the ability to fully utilise existing regulation as well as any new legislation proposed.

Strong & enforced regulation: examples from other jurisdictions

- International developments indicate a shift away from self- and co-regulatory mechanisms towards 'black letter law'. For example, the DSA upgrades obligations from the voluntary Disinformation Code 2018 into binding legislation.

- Many are further empowering and enabling existing regulators:
  - In the EU, the Digital Services Act proposes new enforcement powers including the ability to order the cessation of infringements, levy fines of up to 6% of global annual turnover as well as periodic penalty payments of up to 5% of average global daily turnover, and accept binding commitments
  - In the UK, the draft Online Safety Bill proposes enforcement powers including directions for improvement, notices of non-compliance, and fiscal penalties like civil fines up to £18 million or 10% of worldwide revenue, and business disruption measures

- Some jurisdictions are establishing new regulators or regulatory functions. For example; Canada is proposing establishing both a new Digital Safety Commissioner and Digital Recourse Council (to handle complaints); Ireland is looking to establish an Online Safety Commission, as part of a broader Media Commission; a number of proposed regulations in the US suggest adding new divisions to the FTC, such as a Youth Privacy and Marketing Division as part of proposals to update COPPA[46], and; in the UK the Online Safety Act will hand over new powers to Ofcom. Just as gaps in regulations themselves need to be addressed, so to do gaps between regulators.

- The ability of regulators to enforce requirements depends on some extent to the level of resourcing they have available. Some Australian regulators are not funded to the same extent as their international counterparts.

[46] Proposed *Children and Teens Online Privacy Protection Act* 2021, US https://www.congress.gov/bill/117th-congress/senate-bill/1628/text?r=2&s=2

Approximate funding per person, in AUD, of different Information Commissioners	
$1.11pp	Office of the Australian Information Commissioner. Australia Based on an annual budget $28,487,000 for 2021-22, Australian population of 25,739,256 in 2021
$1.96pp	Information Commissioner's Office, UK Based on an annual budget £70,625,526 for 2021-22, UK population of 67,081,000 in 2020
$6.04pp	Data Protection Commission, Ireland Based on an annual budget €19,128,000 for 2021-22, Irish population of 5,011,500 in 2021 (Ireland also has EU wide data protection functions)

Reconsidering the powers and resourcing of regulators needs to be part of any attempt to ensure Australia's regulatory frameworks can adequately tackle the risks posed by Big Tech.

3. Response to the Committee's Terms of Reference

3.1 The range of harms that may be faced by Australians

There are two key harms that are currently not adequately addressed by Australia's regulatory framework. Firstly, some categories of harm are overlooked and secondly, the contextual causes of many harms are inadequately addressed.

Firstly, community and societal harms are not addressed directly in regulation.

- Communities, such as indigenous communities, migrant communities, people of colour, women and LGBTIQ+ people suffer unique and disproportionate risks in the digital world, but this is overlooked in Australia's current regulatory framework. Disinformation and hate speech can affect particular communities in ways that differ from individual harm. Current regulations will not adequately address this.
- Societal harms are currently overlooked in our regulatory framework. The scale and reach of social media platforms has the capacity to influence and affect Australian institutions, such as Parliament, the Press and healthcare systems. While regulations focus on individual harms, they will not adequately address the significant societal risks platforms create.

Secondly, regulation should move upstream of harms, and focus on reducing risks across all the systems and processes developed by digital services. This upstream approach goes beyond content, and creates a broader digital context of safety, and is essential to provide 'contextual safeguarding' against harm for all Australians.

Exploring what this expanded approach looks like for children and young people is a telling example, as a number of contextual risks are inadequately covered by existing frameworks. The child online safety sector has a commonly used typology that characterises the full breadth of online harms children face; the 4Cs[47]. Figure two contrasts the 4Cs with Australia's existing regulatory framework, highlighting gaps in protections. A focus on risks in systems and processes could better address these.

[47] Sonia Livingstone & Mariya Stoilova 2021 *The 4Cs: Classifying Online Risk to Children, CO:RE Short Report Series on Key Topics* doi.org/10.21241/ssoar.71817

Risk	Some of the current regulatory framework	Gaps in framework
Content — risk of exposure to inappropriate content. For example, risks of exposure to violent content, racist content, pornography, sexualised imagery and mis & disinformation	The *Online Safety Act 2021* is establishing frameworks and Codes around class 1 and 2 materials, as well as developing a Restrictive Access System to limit access to age inappropriate materials like pornography Violent online material may be addressed by the *Sharing Abhorrent Violent Material Act* 2019	Regulation focuses on individual pieces of content, and overlooks the role of platforms in promoting harmful content to children (via algorithms, for example Hate speech, mis & disinformation are not adequately addressed in the current framework, but can be extremely harmful
Contact — risks of making inappropriate contact with others. E.g. Risks of exposure to online grooming, stalking & extremist recruitment	A number of online laws exist that address contact risks, from the *Criminal Code Amendment (Protecting Minors Online) Act* 2017 to criminal laws around terrorist recruitment Some of the *Online Safety Act*'s co-regulatory codes around ensuring user safety may address ways platforms can reduce contact risks. These are as yet unpublished and will be authored by industry	Existing legislation remedies some harms but does not mitigate risks. While they may criminalise individuals who make inappropriate contact, they do not require platforms to stop recommending adult strangers as 'friends' or 'accounts to follow' or prevent platforms enabling adult accounts to message children's accounts for example
Contract / Commercial — risks arising from inappropriate commercial activities and contract exploitation. E.g. risks of identity theft, gambling, profiling bias, surveillance advertising, persuasive design	Children's data is protected as adult's data under the *Privacy Act 1988*, which may reduce the risk of identity fraud The Online Privacy Code may reduce commercial risks to children's data, but it is yet to be published and will most likely be authored by industry The Restrictive Access System may reduce access to gambling services (but may not address loot boxes in games)	The use of children's data poses significant risks, and it is unlikely that an industry drafted code — penned by a sector that funds itself through the commercial exploitation of data — will draft a code that puts children's best interests first There is no regulation in Australia that addresses persuasive design
Conduct — risks associated with inappropriate behaviour. E.g. bullying, trolling, shaming, peer to peer harassment, or with harmful groups (e.g anti-vax groups)	The *Online Safety Act* includes specific provisions around cyber-bullying for children under 18. This includes taking down content that is deemed to be cyber bullying, and where the perpetrator is a child, the regulator is able to require apologies	Engagement with harmful communities falls outside the scope of current regulatory frameworks

Figure two: Australia's regulatory framework mapped against the 4Cs on online harm for children

3.2 Impact of algorithms on harms

Algorithms drive a number of platforms systems and processes, and can cause a range of harms.

The amplification of content harms is a risky and inevitable consequence of unregulated algorithms. Algorithms drive the recommender systems of most platforms, which decide what content is recommended to views. However, recommended systems — and the algorithms that drive them — often promote risky or harmful content. These can cause:

- Societal harm. Facebook's use of 'engagement based ranking' to prioritise content in its algorithm, for example, amplifies harm. Engagement (be it a click through, a comment or a reshare) drives profits for platforms. However content which elicits an extreme reaction, be it inflammatory divisive[48] or misinformation[49], is more likely to encourage engagement. This means divisive content and disinformation are systematically over-promoted in people's feeds. This can have consequences for political discourse, fragmenting and polarising society, and hindering our capacity for genuine political conversations. This threat is particularly salient during elections when manipulation of the information ecosystem can sway votes.
- Community harm. Content recommender systems have been shown to consistently suffer from issues around racism[50] and sexism[51]. Likewise, search engines have been shown to amplify race and sex descrimnination[52]. A recent experiment in Australia found that it took TikTok's recommender algorithm only 7 hour and 42 minutes to 'learn' that an account was interested in content that promoted harmful gender stereotypes and began to recommend this content at such a frequency that it would take only 5-6 days of regular use before their social media feed was completely filled with this content[53].
- Individual harm, such as affecting people's mental health and wellbeing. For example, for children and young people, algorithmic amplification can have a maladaptive effect on young people's body image, and is associated with unrealistic body ideals[54], by for example recommending Pro-Anorexia content[55] and AnaCoaches as 'friends' to children's accounts[56]. This access to harmful content and communities can have the effect of

[48] Luke Munn 2020 'Angry by design: toxic communication and technical architectures' *Humanities and Social Sciences Communications* doi.org/10.1057/s41599-020-00550-7
[49] Peter Dizikes 2018 'On Twitter, false news travels faster than true stories' *MIT News* https://news.mit.edu/2018/study-twitter-false-news-travels-faster-true-stories-0308
[50] Derek O'Callaghan, Derek Greene, Maura Conway, Joe Carthy, Pádraig Cunningham 2014 'Down the (White) Rabbit Hole: The Extreme Right and Online Recommender Systems' *Social Science Computer Review* doi.org/10.1177/0894439314555329
[51] Masoud Mansoury, Himan Abdollahpouri, Jessie Smith *et al* 2020 'Investigating Potential Factors Associated with Gender Discrimination in Collaborative Recommender Systems' *Cornell University Computer Science* *arXiv:2002.07786*
[52] Safia Noble 2018 *Algorithms of Oppression* NYU Press
[53] Dylan Williams, Alex McIntosh & Rys Farthing 2021 *Surveilling young people online* Reset Australia au.reset.tech/uploads/resettechaustralia_policymemo_tiktok_final_online.pdf
[54] Grace Holland & Marika Tiggemann 2016 "A systematic review of the impact of the use of social networking sites on body image and disordered eating outcomes" *Body Image* 17, pp.110-110 doi.org/10.1016/j.bodyim.2016.02.008
[55] Ysabel Gerrard 2018 'Beyond the hashtag: Circumventing content moderation on social media' *New Media & Society* 20(12):4492-4511. doi:10.1177/1461444818776611
[56]Suku Sukunesan 2021 'Anorexia coach': sexual predators online are targeting teens wanting to lose weight. Platforms are looking the other way

'normalising' disordered eating and trigger the emulation of these destructive behaviours[57]. It can also normalise other mental health and wellbeing risks, such as self harm. For example, in the UK the Coroners Office is investigating the role of social media algorithms in the suicide of a 14 year old, after alogirthms fed her more and more extreme self harm materials[58]. For those over 18, while research exploring the impact of recommender systems on mental health and well being is scant, it has been argued that algorithms can dominantly influence individuals' self-control, self-esteem, and even self-determination[59].

In a similar vein, algorithms can also cause harm through advertising systems, where they seek maximise engagement over safety). For example, advertising algorithms can:

- Cause disinformation and hate speech to be funded. Advertising algorithms that focus entirely on maximising engagement provide equal access to financial incentives to both good actors to post great content, and to bad actors to post harmful content. Any content that gets traffic is monetizable through undiscerning advertising algorithms. This business model has been shown to contribute substantially to the funding model of disinformation[60] and hate speech[61].
- Harm individuals. Advertising algorithms can exacerbate risky behaviors, for example excessive alcohol consumption. The essential functionality of these algorithms seems straight-forward and logical: ads for alcohol will be increasingly targeted at those who are more likely to engage with alcohol ads, i.e. those who drink more. From the algorithm's perspective, this is a success. From a public health perspective, this is perverse[62]. While individual platforms may have policies suggesting that they do not micro-target alcohol ads aggressively, others do not and without transparency we cannot be sure that those that do are properly implemented.

Other systems and processes, beyond content recommender systems and advertising systems, use algorithms and produce similar risks; friend/account recommendations are driven by algorithms, pop-up notifications, suggestive text. Future regulation needs to get 'upstream' of all of these uses and their potential harms. Digital services could be required to reduce risks across all their systems and processes, including algorithms.

---

theconversation.com/anorexia-coach-sexual-predators-online-are-targeting-teens-wanting-to-lose-weight-platforms-are-looking-the-other-way-162938

[57] Giuseppe Logrieco, Maria Marchili, Marco Roversi, & Alberto Villani 2021 'The Paradox of Tik Tok Anti-Pro-Anorexia Videos: How Social Media Can Promote Non-Suicidal Self-Injury and Anorexia.' *International Journal Of Environmental Research And Public Healt*h, 18(3), 1041. doi:10.3390/ijerph1803104

[58] Tom Knowles 2021 'Molly Russel: Coroner Voices Alarm Over Delays to Inquest' *The Times* www.thetimes.co.uk/article/molly-russell-coroner-voices-alarm-delays-inquest-gmfmk7bwp

[59] Urbano Reviglio & Claudio Agosti 2020 'Thinking Outside the Black-Box: The Case for "Algorithmic Sovereignty" in Social Media' *Social Media + Society* doi:10.1177/2056305120915613

[60] Global Disinformation Index 2020 *Ad-funded Covid 19 disinformation* https://disinformationindex.org/wp-content/uploads/2020/07/GDI_Ad-funded-COVID-19-Disinformation-1.pdf

[61] Karen Hao 2021 'How Facebook and Google Fund Global Misinformation' *MIT Technology Review* www.technologyreview.com/2021/11/20/1039076/facebook-google-disinformation-clickbait/

[62] As University of Queensland researchers describe it, 'when platform users depict their own drinking practices' which we add, includes engaging with alcohol advertising 'they generate data that signal their interest in alcohol consumption and its relation to specific times, places and cultural interests.... This information can then be used to target them.' Nicholas Carah, Carla Meurk, Matthew Males & Jennifer Brown 2017 'Emerging social media 'platform' approaches to alcohol marketing: a comparative analysis of the activity of the top 20 Australian alcohol brands on Facebook (2012-2014)' *Critical Public Health* doi.org/10.1080/09581596.2017.1282154

3.3 Age verification and age assurance systems

> *A note on definitions and language used by Reset Australia*[63]
>
> Age verification (AV): A system that relies on hard identifiers and/or verified sources of identification, which provide a high degree of certainty in determining the age of a user. These are often government backed forms of ID.
>
> Age estimation (AE:) A process that establishes a user is likely to be of a certain age or fall within an age range.
>
> Age assurance: An umbrella term for both age verification and age estimation solutions.

There are many methods that platforms can use to assure the age of users. 5Rights Foundation recently released a categorisation of the techniques currently available which we have included in appendix three[64].

Age assurance can be used for multiple purposes, each with their own impact on children and young people's safety and rights. Clarity around the *intent* of the assurance is needed, to ensure that it can be mapped to the most appropriate *method*. A blunt approach to selecting methods could cause unintended harm. For example:

- Access to restricted services such as online gambling or pornography most likely necessitates the use of 'hard' age verification methods. Hard techniques, such as checking government backed IDs, aim to prevent 'false positives' and err on the side of assuming a user should not access content/services if in doubt
- Access to protective features and processes, such as 'safer' data handling protocols, turning off cookies by default or being excluded from advertising databases most likely necessitates age estimation techniques. The aim should be to prevent 'false negatives' an err on the side of assuming a user should access protections if in doubt

Age assurance techniques therefore need to be deployed in risk-based ways, where the best interests of children are the primary consideration in deciding which method to implement. This requires considering a broad range of children's rights, including:

- Rights to equal access. Some methods of assurance can create disproportionate barriers for young people from low socio-economic households. Birth certificates and passports — by and large the two forms of government backed ID available to under 16 year olds — are expensive documents to purchase, keep safe and replace. Likewise, not all young people have access to these documents, especially young people with irregular immigration status. These could create new forms of digital disadvantage
- Right to privacy and security. Sharing and processing highly sensitive ID documents creates privacy and security risks and biometric methods create other privacy concerns

[63] From 5Rights Foundation 2021 *But How do they Know it's a Child?* 5Rights Foundation https://5rightsfoundation.com/uploads/But_How_Do_They_Know_It_is_a_Child.pdf
[64] 5Rights Foundation 2021 *But How do they Know it's a Child?* 5Rights Foundation https://5rightsfoundation.com/uploads/But_How_Do_They_Know_It_is_a_Child.pdf

- Rights to access information: Where they create undue barriers and are unnecessarily restrictive, age assurance methods can affect children's right to access information and enjoy the digital world
- Rights to protection: If used in the right place at the right time, age assurance techniques can protect children from unnecessary content risks, and be used to create a better digital experience for young people in general

The UK is currently developing a risk focused statutory code providing guidance around age assurance[65]. This outlines 11 minimum requirements for age assurance systems, indicating that they must:

1. Protect the privacy of users in accordance with applicable laws
2. Be proportionate with regard to the risks arising from the product or service and to the purpose of the age assurance system
3. Offer functionality appropriate to the capacity and age of a child who might use the service
4. Be secure, does not expose users or their data to unauthorised disclosure or security breaches, and does not use data gathered for the purposes of the age assurance system for any other purpose
5. Provide appropriate mechanisms and remedies for users to challenge or change decisions if their age is wrongly identified
6. Be accessible and inclusive to users with protected characteristics
7. Not unduly restrict access of children to services to which they should reasonably have access, for example, news, health and education services
8. Provide sufficient and meaningful information for a user to understand its operation, in a format and language that they can be reasonably expected to understand, including if they are a child
9. Be effective in assuring the actual age or age range of a user as required
10. Not rely solely on users to provide accurate information
11. Be compatible with; the *Data Protection Act* 2018, the *Age Appropriate Design Code,* the *Human Rights Act* 1998, the *Equality Act* 2010, and the United Nations Convention on the Rights of the Child and General Comment No. 25 (2021) on Children's Rights in Relation to the Digital Environment.

Age assurance is an extremely important policy agenda for children and young people's safety, and needs to be explored in a 'joined up' and considered way. There are two concurrent policy proposals under consideration around age assurance and age verification, both with very different intents; firstly as it is needed for restrictive access services declaration[66] (with the eSafety Commissioner) and; secondly as it is needed to ensure improved privacy protections[67] (with the Attorney General's Office). These could be aligned with overarching risk-based regulation to ensure the best outcomes for children.

Ultimately, age assurance systems are just one of the systems and processes that needs to be addressed to create a robust regulatory framework.. Effective age assurance should allow platforms to provide age appropriate services, as well as more effectively police their minimum age limits, but more comprehensive risk reductions are also needed.

---

[65] Age Assurance Minimum Standards Bill 2021, UK
https://bills.parliament.uk/publications/41683/documents/325
[66] eSafety Commissioner 2021 *Restricted Access System*
www.esafety.gov.au/about-us/consultation-cooperation/restricted-access-system
[67] Attorney General's Office 2021 *Online Privacy Bill Exposure Draft*
https://consultations.ag.gov.au/rights-and-protections/online-privacy-bill-exposure-draft/

3.4 Effectiveness, take-up and impact of industry measures for child online safety

Industry measures and self regulation have largely failed children and young people. The abundance and prevalence of known risks in many platform's systems and processes indicates a lack of appetite to proactively prioritise children's safety. Regulations that address risks in systems and processes, place obligations on platforms and holds them accountable for harms that eventuate where they have not adequately mitigated risks, are now needed.

3.5 Effectiveness and impact of parental control tools

While parental control tools are important, we do not believe that placing responsibility onto parents to manage online safety is the most effective solution. Platforms should be developing systems that are safe for children in the first instance.

This is in keeping with existing norms around effective ways to reduce industrial hazards. The hierarchy of hazard controls — a globally used framework — outlines that the most effect interventions emerge from eliminating hazards[68] (see figure three). Tools that create protective barriers, such as safe searches, are the last line of defence because every instance of individual failure, either from the tool or the user, leaves people exposed to risk.

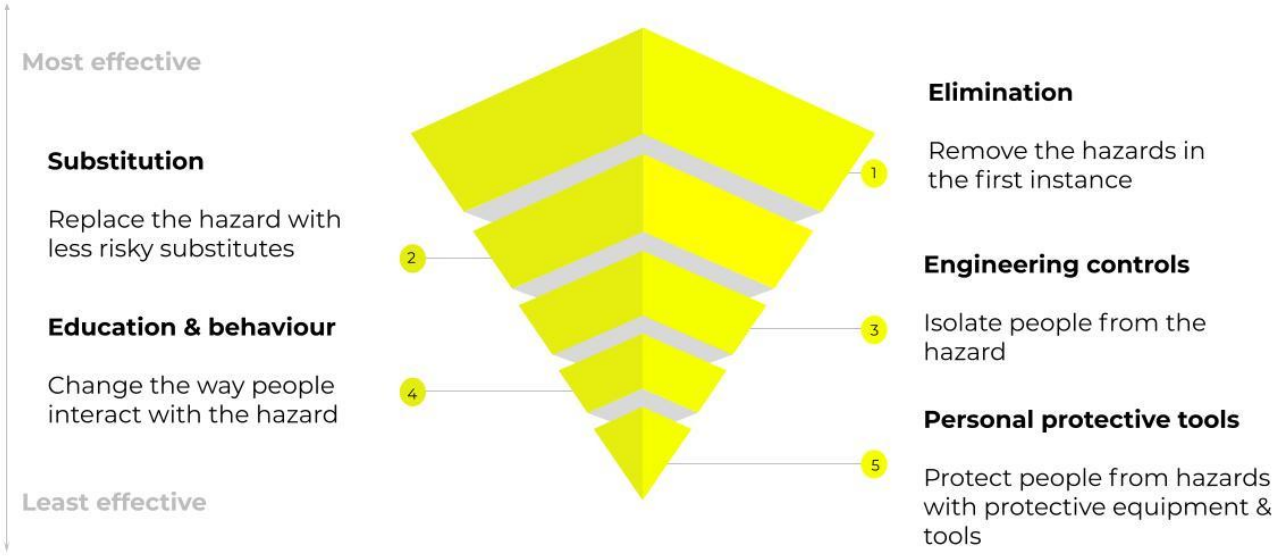

Figure three: The Hierarchy of Hazard Controls

[68] See for example WorkSafe Victoria 2021 *The Hierarchy of Controls* www.worksafe.vic.gov.au/hierarchy-control

3.6 Transparency and accountability of harms

Accountability means that regulations should identify the core risks as stemming from platforms themselves — and squarely place the burden of responsibility on digital service to mitigate them. This means that regulation needs to place duties on industry, and de-emphasise solutions that:

- Require users to manage immense risks themselves, such as education and training. These approaches are not commensurate with the risk at hand, nor will they be able to protect all users equally
- Require users to make more informed choices such as more clear terms and conditions. These 'buyer beware' approaches fail to appreciate the power asymmetry between users and industry
- Solutions that aim to address individual users behaviour, such as holding 'trolls' accountable for content. These approaches will not address the multiple ways social media companies enable and promote risky behaviour

These sorts of solutions are all necessary, but not sufficient. Tech companies should bear the burden of responsibility to be held accountable.

Transparency is the flipside of this coin, and is required to truly hold industry to account. Part of the issue in understanding how to best solve the problems of social media is that legislators, regulators, researchers and civil society simply do not know enough about the specific mechanics of how platforms work nor their consequences.

Accountability and transparency can be achieved a number of ways in regulations:

- Placing a duty/duties of care on platforms to reduce risks across their systems and processes. This includes duties to adequately assess and mitigate risks
    - Algorithmic audits would fall under this type of risk assessment
    - Data impact assessments are another tools that should be explored
- Requiring transparency about their risk assessment and mitigation processes, so that regulators and the public can understand the nature of the risks right across a platform, and evaluate if mitigation measures meet expectations / best practice
    - Allowing regulators and academic researchers access to platform data would greater ensure transparency around risks in an independent and *trustworthy* fashion

Ultimately, to hold platforms to account though regulations and regulators need to be strong and enabled. A focus on rigorous enforcement and empowered regulators should also be considered as part of the mix for future regulation.

3.7 Collection and use of data

Data is one of the key systems that drives the digital world and fuels social media platforms. Accordingly, Reset believes that the regulation of data should be focused on reducing risks, in comprehensive, accountable and transparent, and enforced ways.

The ongoing review of the *Privacy Act* 1988 is a welcome opportunity to ensure this. Reset will be responding to that review separately. The overarching direction of travel for data regulations could be aligned with the approaches outlined above, for example requiring:

- Obligations and duties on platforms to use data in less risky ways (in terms of data handling practices) and in ways that reduce risks for users (in terms of data uses). What is considered fair and reasonable by data processors should consider the risks that they create for Australians
- Ensuring data accountability and transparency. For example, data impact assessments and data audits should be required, and these should be made available to regulators and the public
- Ensuring the definitions and scope of what counts as 'personal data' matches the contemporary risks of the digital world, including for example 'look alike' accounts and pseudonymised data
- Likewise, ensuring that the scopes and definitions of industries covered by the *Privacy Act* covers all the industries and innovations that creates significant risks for Australians, regardless of their annual turnover
- An adequately resourced Office of the Australian Information Commissioner, with the ability to enforce any new obligations

We note the recently released draft *Enhancing Online Privacy Bill* that among other things proposes an upstream, systemic approach to protecting children's data using the 'best interests' principle. This is a welcome approach that could be replicated across other areas of legislation.

4. Recommendations

An overarching regulatory framework for Australia should be developed that:

- Creates a duty/duties of care to eliminate risks across a service's systems and processes — through, for example, requiring detailed risk assessments and audits of all relevant systems or processes including and beyond content
    - This includes algorithmic audits and data impact assessments
- Addresses community and societal risks associated with platforms — through for example, expanding the definitions of risks that need to be addressed to include community and societal risks. This will require a stronger approach to hate speech and disinformation
- Creates obligations for transparency and accountability — through for example, regulations that place responsibilities onto platforms themselves and including requirements to share and make public risk assessments and mitigations
- Comprehensively addresses the risks of the contemporary digital world — through for example, applying regulation to the broadest range of digital service providers with risk-based thresholds for additional obligations. Regulations for the social media sector should be aligned with regulations for other online services, data brokers and internet service providers for example
- Fostering a network of regulators that are more empowered, better joined up and resourced to effectively oversee this.

This would create a more streamlined approach to regulation, replacing multiple disjointed obligations with more aligned upstream duties and reducing the regulatory burden on Australia's successful tech industry. It would also be interoperable with emerging international requirements, ensuring Australian industry could expand into international markets with minimal regulatory friction.

This could involve for example:

- Building out and expanding our *Online Safety Act* over time to include:
    - A more systemic focus on duties to reduce risks in systems and processes (right across the service, and including for example algorithms and ad delivery systems);
    - Expanding the definition of risks to include community and societal risks, which necessitates an enhanced focus on mis/disinformation and hate speech
    - Requiring enhanced duties of care for accountability, and including requirements for transparency measures
    - Replacing voluntary and co-regulatory codes with upstream obligations in the Act
    - Ensuring that the broadest range of digital services remains covered, with risk based additional obligations. (The scope of the *Online Safety Act* is already very broad, and this provides a potential model for other regulations)
- Expanding our *Privacy Act* and Enhancing Online Privacy Act to:
    - Address a broader definition of personal data to cover metadata and other new forms of data fuelling the new digital world
    - Adopt a systemic focus on reducing the risks created through the processing of data
    - Apply to the broadest range of digital service providers with risk based additional obligations
    - Replace voluntary and co-regulatory codes with upstream obligations in the Act

Appendices

1. Australia examples of disinformation

Overview	Date	Description
Bots stormed Twitter in their thousands during the federal election[69] Evidence of inauthentic coordinated behaviour during the most recent election	2019	A QUT study which examined around 54,000 accounts out of more than 130,000 Twitter users active, during and after the 2019 Australian Federal Election (looking at over 1 million tweets) revealed that 13% of accounts were 'very likely' to be bots, with the majority originating from New York. This is estimated to be more than double the rate of bot accounts in the US presidential election. This was done through an AI program Botometer - which looks for signs such as tweeting frequently 24 hours a day, tweeting at regular intervals, usernames with lots of numbers and whether their followers also appeared to be bots. New accounts created during the election campaign were more likely to be bots.
Labor asks questions of WeChat over doctored accounts, 'fake news[70] Spread of fake news in WeChat shows potential for special interest groups to manipulate public sentiment to influence electoral outcomes in highly targeted groups	May 2019	Labor is losing the battle on influential Chinese social media site WeChat as a wave of fake news posts and doctored accounts target the Shorten campaign on issues such as Safe Schools, taxes and refugee policy. • While many of the posts are unauthorised, making it difficult to know who is responsible for them, one emerged on the weekend containing a doctored tweet purporting to come from Mr Shorten's personal account. The apparent tweet says: "Immigration of people from the Middle East is the future Australia needs." • It was found on multiple WeChat groups posted by Melbourne woman Jing (Jennifer) Li, who has previously identified herself as being a Liberal Party member. Neither Ms Li or the Coalition campaign office responded to questions about the post on Monday. • Another WeChat account - currently peddling a scare campaign on Labor's economic policies - has been traced back to former Liberal MP Michael Gidley, a member of the party's conservative faction, whose former Victorian state seat of Mount Waverley has high numbers of Chinese-Australian voters. • The account was registered in Mr Gidley's name in September 2017 before changing in April last year from "MichaelGidleyMP" to "Victoria Brief Talk". In March, four months after Mr Gidley lost his seat to

[69] Felicity Caldwell 2019 'Bots stormed Twitter in their thousands during the federal election' *Sydney Morning Herald* https://www.smh.com.au/politics/federal/bots-stormed-twitter-in-their-thousands-during-the-federal-election-20190719-p528s0.html
[70] Yan Zhuang & Farrah Tomazin 2019 'Labor asks questions of WeChat over doctored accounts, 'fake news' *Sydney Morning Herald* https://www.smh.com.au/national/labor-asks-questions-of-wechat-over-doctored-accounts-fake-news-20190506-p51kkj.html

		Labor in the state election, it changed again to "Australia Brief Talk". • The account remains active, with one post falsely claiming that under Labor's new tax policies, retirees whose main income is from share dividends will need to pay an additional $12,850 in taxes each year. It also claims Labor plans big personal tax increases and extra taxes on house sales of $30,000.
Facebook removed 'coordinated inauthentic behaviour' during Australian election[71] Facebook position on mis- and disinformation during Aus election	May 2019	Facebook's position - *"Facebook does not believe that it's an appropriate role for us to be the arbiter of truth over content shared by ordinary Australians or to referee political debates and prevent a politician's speech from reaching its audience and being subject to public debate and scrutiny."* • It also told the committee it removed 2.2bn fake accounts between January and March 2019, and "the majority of these accounts were caught within minutes of registration". • Guardian Australia revealed last month the ALP has used its post-election submission to the committee to call for an examination of whether Australian elections are vulnerable to influence by "malinformation" – a term invoked by the Australian Competition and Consumer Commission in its landmark digital platforms review. • In an interview with Guardian Australia in August, the ACCC chairman, Rod Sims, blasted Facebook's practices, and said the social media giant should have removed the bogus death tax claims given its own independent fact checking processes had found the material to be false.
Facebook videos, targeted texts and Clive Palmer memes: how digital advertising is shaping this election campaign[72] Ability of politicians and lobbying groups to spend on campaigning is unchecked	May 2019	Clive Palmer and United Australia Party + special interest groups • The most recent Nielsen figures put the cost of Palmer's ads since September at around A$30 million, though Palmer says himself he's spent at least A$50 million. • Despite the ubiquity of his ads, though, Palmer is still struggling to connect with most voters. This demonstrates a very important aspect to any advertising campaign: the actual brand still needs to be seen as offering real value to voters. The increasing influence of lobbying groups • One of the more interesting developments of this election so far is the increasing sophistication, knowledge and strategies of political lobbying groups, or Australia's equivalent to America's PACs. • GetUp! is one such group, collecting A$12.8 million in donations in the last 12 months alone. • The rise of these groups in Australian politics opens a Pandora's Box on just who can influence elections without even standing a single candidate – an issue that's becoming part of politics now in many Western democracies.

[71] Katharine Murpy 2019 ' Facebook removed 'coordinated inauthentic behaviour' during Australian election' *The Guardian* https://www.theguardian.com/australia-news/2019/oct/23/facebook-removed-coordinated-inauthentic-behaviour-during-australian-election
[72] Andrew Hughes 2019 'Facebook videos, targeted texts and Clive Palmer memes: how digital advertising is shaping this election campaign' *The Conversation* https://theconversation.com/facebook-videos-targeted-texts-and-clive-palmer-memes-how-digital-advertising-is-shaping-this-election-campaign-115629

Chinese media mocks Australia and Prime Minister in WeChat posts[73] Evidence of anti-liberal propaganda which has the potential to be chinese state interference	May 2019	Prime Minister Scott Morrison and the Coalition Government have been targeted by online propaganda coming from social media accounts affiliated with the Chinese Communist Party (CCP). • Across a period of five months from November 2018 to March 2019, the researchers analysed the Australian content on 47 of the most visited WeChat Official accounts in Mainland China, 29 of which were aligned with the CCP. • Propaganda researchers found that there was a clear "anti-Liberal story" coming from social media accounts, many of which have close affiliations to the Chinese Government. • The posts criticise Australia's involvement in the Five Eyes alliance. The researchers say there is little evidence of attacks on Bill Shorten and the Labor Party across their dataset, although this is happening elsewhere on WeChat.
Scammers from Kosovo manipulating Australian users to profit[74] Evidence of the types of divisive content that is used to generate engagement on the platform, whether that is for financial or ideological gain	2019	A network of Facebook pages run out of the Balkans profited from the manipulation of Australian public sentiment. Posts were designed to provoke outrage on hot button issues such as Islam, refugees and political correctness, driving clicks to stolen articles in order to earn revenue from Facebook's ad network. • The location information only recently became discoverable when Facebook flicked the switch to bring Australia into line with new advertising transparency measures that have been in place in the United States since mid-2018. • The Facebook pages have a combined fanbase of 130,000-plus, which has been built up over several years. The oldest and most popular page, "Australians against Sharia", has been publishing since June 2013. • The "Australians against Sharia" page, which has over 67,000 fans, has also reposted memes attacking Labor Party identities including Bill Shorten, Penny Wong and Julia Gillard, the Greens' Sarah Hanson-Young and the Liberal Party's Julie Bishop. • Facebook has now removed these pages, admitting that they violated their policies by engaging in "coordinated inauthentic behaviour".
Evidence shows the Internet Research Agency (IRA) targeting Australian politics between 2015 and	2015 and 2017	Twitter identified 3,841 accounts suspected of operating out of the Internet Research Agency in St Petersburg. A number of these same accounts Twitter identified as suspected of operating out of the Russian Internet Research Agency (IRA) targeted Australian politics in response to the downing of flight MH17, attempting to cultivate an audience through memes, hashtag games and Aussie cultural references. • Researchers from Clemson University in the US released 3 million tweets. Analysis of this data set

[73] Steve Cannane 2019 Chinese media mocks Australia and Prime Minister in WeChat posts *ABC News* https://www.google.com/url?q=https://www.abc.net.au/news/2019-05-09/pm-targeted-by-chinese-communist-party-related-wechat-accounts/11092238&sa=D&source=docs&ust=1641786278156843&usg=AOvVaw1mrFcD73xejHL-oxGD5dpC

[74] Michael Workman & Stephen Hutcheon 2019 'Facebook trolls and scammers from Kosovo are manipulating Australian users' *ABC News* https://www.abc.net.au/news/2019-03-15/trolls-from-kosovo-are-manipulating-australian-facebook-pages/10892680

2017[75] Evidence shows that Russia appears to have tested different tactics to manipulate the australian public		shows how these accounts targeted Australian politics – particularly in reaction to the Australian response to the downing of flight MH17. Some 5,000 tweets mention the terms "#auspol", "Australia" or "MH17" – with "Australia" the most common of the three. • A jump in activity focusing on MH17 correlates with the Australian government's response to the Russian missile attack on MH17, when Australia deployed fighter aircraft to operate in Syrian airspace where Russian aircraft were also operational. • A second spike in Feb 2017 actually has nothing to do with politics and instead refers to a hashtag game. These Russian accounts encouraged people to come up with Australian names for popular US television programs. While this may seem like innocent fun, it is also a technique of spy craft. "Assets", in this case, Australian citizens, are recruited on neutral, non-political terms before they are shifted towards political topics.
Minister urges skepticism as fake virus news spreads[76] Example of spread of disinformation, which is dangerous in elections	2020	Communications Minister Paul Fletcher has urged Australians to be skeptical of what they read online as misinformation on the coronavirus outbreak spreads rapidly. • Disinformation around the Coronavirus is spreading online, with posts including claims of how the virus can be caught, suggestions it was deliberately released as well directing people not to consume certain food or visit particular areas in Australia. • The rapid spread of disinformation is forcing Facebook and Google to ramp up efforts and use third-party fact-checkers to remove misleading information.
Bushfires, bots and arson claims: Australia flung in the global disinformation spotlight[77] Example of spread of disinformation, which is dangerous in elections	2020	Hashtag #ArsonEmergency became the focal point of a new online narrative surrounding the bushfire crisis in the beginning of 2020. • QUT social media analyst Timothy Graham studied 300 twitter accounts to identify any inauthentic behaviour driving the #ArsonEmergency hashtag which was used to push a narrative that the cause of the fires was arson. • Many of these accounts were found to be behaving 'suspiciously', compared to other hashtags trending including #AustraliaFire and #BushfireAustralia.

[75] Tom Sear & Micael Jensen 2018 'Russian trolls targeted Australian voters on Twitter via #auspol and #MH17' *The Conversation* https://theconversation.com/russian-trolls-targeted-australian-voters-on-twitter-via-auspol-and-mh17-101386
[76] Zoe Samios and Dana McCauley 2020 Minister urges scepticism as fake virus news spreads' Sydney Morning Herald https://www.smh.com.au/business/companies/minister-urges-scepticism-as-fake-virus-news-spreads-20200128-p53vjn.html
[77] Timothy Graham & Tobian Keller 2020 'Bushfires, bots and arson claims: Australia flung in the global disinformation spotlight' *The Conversation* https://theconversation.com/bushfires-bots-and-arson-claims-australia-flung-in-the-global-disinformation-spotlight-129556

2. Assessment of risk warranting escalation from self- and co-regulation to primary and subordinate legislation

Australia has developed a multi-path approach to industry regulation, using self-regulation, quasi-regulation, co-regulation and 'black letter law' (or explicit regulation by primary and subordinate legislation). Since the mid '90s, this multi-path approach has facilitated a range of regulatory responses, some 'light' and some 'hard touch' to different industries and issues.

While over a decade old, the *Best Practice Regulation Handbook*[78] outlined considerations to assess which path is appropriate for each industry/issue, that are still helpful prompts for reflection. The handbook suggests an evaluation of the options should consider:

- The level of risks and significance posed by the potential concern, noting that major public health and safety issues warrant explicit government regulation
- The community appetite for the certainty of legal sanctions, noting that self regulation is only feasible where there is no particular community interest
- The ability of the market to address the concern, noting that where there is "a systemic compliance problem with a history of intractable disputes and repeated or flagrant breaches of fair trading principles, and no possibility of effective sanctions being applied" explicit regulation is required.

We believe that social media has exceeded any reasonable threshold for explicit government regulation across all three considerations.

1. **The level of risks posed by social media platforms:** Social media platforms can create significant risks, including major public health risks. Taking Facebook and the pandemic as an example, Australia witnessed the enabling and promotion of harmful content and discussions. Both membership numbers and engagement among groups peddling 'anti-vaxx' and vaccine hesitant content grew across the pandemic in Australia[79]. We also saw the rise and promotion of 'anti vaxx' influencers, with more than 100 Instagram accounts promoting anti-vaxx content to more than 6 million users[80]. There were ,and will continue to be, deadly consequences while these risks continue.

2. **The community wants and expects the certainty provided by regulation:** There are now legitimate community expectations of explicit regulation of Big Tech in Australia. Last year, a Lowy Institute poll found that 90% of Australians think that the influence social media companies have is an important or critical threat to the vital interests of Australia[81]. And indeed, a poll by the Australian Financial Review in late 2020 found that

[78] Australian Government 2010 *Best Practice Regulation Handbook* Canberra
[79] Reset Tech Australia 2021 *Anti-vaccination & vaccine hesitant narratives intensify in Australian Facebook Groups* https://au.reset.tech/uploads/resetaustralia_social-listening_report_100521-1.pdf
[80] Jasper Jackson & Alexandra Heal 2020 'Instagraft: Covid conspiracy theorists selling silver spray and $50 seawater' *Bureau of Investigative Journalism* www.thebureauinvestigates.com/stories/2021-04-11/instagraft-covid-conspiracy-theorists-selling-silver-spray-and-50-seawater?mc_cid=faeeac9b83&mc_eid=ae64430abe
[81] Lowry Institute 2021 *Lowry Institute Poll* https://poll.lowyinstitute.org/charts/threats-australias-vital-interests/

77% of Australians felt that BigTech should face stronger Government regulations[82]. The scale and depth of the public's concerns warrants the strongest possible regulatory response.

3. **The social media sector has demonstrated systemic compliance problems:** While many sectors have worked hard to deserve the benefit of 'light touch' regulations, the social media sector has demonstrably not. For example:

    - YouTube settled a case for $170m USD with the FTC in 2019 for using children's data without necessary parental consent[83] and are currently facing a £2b 'class action' for unlawfully tracking and collecting children's data[84]. Google, their parent company, has also been fined for multiple breaches of existing regulation, including a €500m fine for acting in bad faith around EU copyright directives in France[85], €7m for failing to meet requirements around GDPR in Sweden[86] and €220m for anti competitive practices in their advertising systems in France again[87]. Earlier this year, the Texas Attorney General accused Google of deliberately stalling efforts to strengthen children's online privacy laws in the US, and documented Google executives 'bragging' about stalling EU attempts at improving consumer privacy[88].
    - Facebook has faced many fines, including a $5b USD penalty from the FTC for breaching consumer privacy regulations[89] and a $5m USD to settle civil rights lawsuits claiming the company's advertising system excluded people from seeing housing ads based on age, gender and race[90].
    - TikTok has also had its fair share of fines, settling a case for $5.7 m USD with the FTC in 2019 for using children's data without the necessary parental consent[91], and were fined €750k in the Netherlands over GDPR compliance[92]. They are currently facing a £1b plus lawsuit led by the UK's former Children's Commissioner for excessive data collection practices[93].

---

[82] Paul Smith 2020 'Big Tech on the Nose' *Australian Financial Review* www.afr.com/technology/big-tech-on-the-nose-as-aussies-demand-accountability-and-tougher-laws-20201030-p56a93

[83] FTC 2019 'Google and YouTube will Pay Record $170m for Alleged Violations of Children's Privacy Law' www.ftc.gov/news-events/press-releases/2019/09/google-youtube-will-pay-record-170-million-alleged-violations

[84] YouTube Data Claim 2020 'YouTube Data Claim' www.youtubedataclaim.co.uk/

[85] Ian Carlos Campbell 2021 'Google fined €500 million in France over bad faith negotiations with news outlets' *The Verge* www.theverge.com/2021/7/13/22575647/google-fine-500-million-french-authorities-news-showcase

[86] Vincent Manancourt 2020 'Google to appeal Swedish data watchdog' *Politico* www.politico.com/news/2020/03/11/google-to-appeal-swedish-data-watchdog-7m-fine-125460

[87] Simon Read 2021 'Google Fined €220m in France' *BBC* https://www.bbc.com/news/business-57383867

[88] Leah Nylen 2021 'Google sought feelow tech giants help is stalling kids privacy protections' *Politico* www.politico.com/news/2021/10/22/google-kids-privacy-protections-tech-giants-516834

[89] FTC 2019 *FTC imposes $5 Billion Penalty* www.ftc.gov/news-events/press-releases/2019/07/ftc-imposes-5-billion-penalty-sweeping-new-privacy

[90] Brakkton Booker 2019 'After Lawsuite, Facebook Announces Changes' *NPR* www.npr.org/2019/03/19/704831866/after-lawsuits-facebook-announces-changes-to-alleged-discriminatory-ad-targeting

[91] FTC 2019 *Video Social Networking App Settles* www.ftc.gov/news-events/press-releases/2019/02/video-social-networking-app-musically-agrees-settle-ft

[92] Dutch News 2021'Dutch Privacy Watchdog Fines TikTok' *Dutch News* www.dutchnews.nl/news/2021/07/dutch-privacy-watchdog-fines-tiktok-e750000-after-privacy-probe

[93] BBC 2021 'TikTok sued for billions over use of children's data' *BBC* www.bbc.co.uk/news/technology-56815480

Beyond compliance with existing regulations, at times the sector appears to actively resist 'doing the right thing'. For example, back in 2016, the Wall Street Journal found an internal Facebook presentation documenting that they know their platform was hosting a large number of extremist groups and promoting them to its users: "64% of all extremist group joins are due to our recommendation tools," the presentation said[94]. It was only in the wake of the insurrection in January 2021 that Mark Zuckerberg announced that the company will no longer recommend civic and political groups to its users.

This does not reflect a series of unrelated incidents. Most of these companies are publicly listed entities obligated to act in shareholder's best interests. Without legal requirements insisting that they prioritise user safety, they are bound to continue to prioritise shareholder profits.

3. Age assurance methods

5Rights Foundation[95] has described some existing age assurance methods, key details of which are below.

Method	Example	Considerations
Self-declaration. An age estimation technique	A user enters their date of birth when signing up to a platform	The most commonly used technique at the moment. This could also include measures which discourage false declarations of age, eg: • If a user enters a date of birth that indicates they are below the minimum age, platforms could block repeated attempts from the same IP address • Using language that elicits a more truthful age declaration, for example, "enter your date of birth" rather than "confirm that you are over 13" • Checking a user's date of birth twice. i.e when a user logs in the second time, ask them to confirm their date of birth. Children who gave a false date of birth on registration may not remember the date of birth they gave, which could flag them for moderation These place the burden on the child to self-declare their age correctly, and are highly spoofable
Hard-identifiers such as government backed IDs collected directly by a platform. An age verification technique	A user is asked to upload a copy of their passport or Medicare number to check against official records when they open an account. This is checked to verify age	Hard identifiers are most commonly used for age assurance by services that are restricted to users over 18. The emphasis is on proving users are adults The use of hard identifiers offers a high level of assurance but presents risks of privacy violations and potential exclusion ID documents can be reviewed by the platform or a third party provider for the platform (see below)

[94] Jeff Horwitz & Deepa Seetharaman 2020 'Facebook Executives Shut Down Efforts to Make the Site Less Divisive' *Wall Street Journal* www.wsj.com/articles/facebook-knows-it-encourages-division-top-executives-nixed-solutions-11590507

[95] 5Rights Foundation 2021 *But How do they Know it's a Child?* 5Rights Foundation https://5rightsfoundation.com/uploads/But_How_Do_They_Know_It_is_a_Child.pdf

Biometrics, such as facial analysis. An age estimation technique	A user has their photo run through an AI system to estimate their age	Facial analysis is a widely used form of biometric estimation for age and does not — in principle — recognise nor identify the individual Facial analysis compares the user's facial features against large datasets that have been used to train the technology through machine learning to estimate their age range Facial analysis is inclusive of those who may not be able to present a valid ID document. It can also be used in privacy preserving ways if services discard the facial image once it has estimated a user's age Caution is needed to ensure that the data of facial features is created in privacy preserving and inclusive ways, and that images scanned are truly discarded
Profiling and inference models, such as noting that those who watch unboxing videos may be children. An age estimation technique	Already collected user data, such as what each user 'liked', who their friends are etc., is scanned and inferences about their age calculated	Profiling and inference are already commonly used in commercial settings, including to estimate the age range of users This creates significant tension for children's right to privacy, and there is a need to ensure that inferences are inclusive and accurate
Capacity testing, such as asking a user to complete a puzzle. An age estimation technique	A user may be asked to complete a language test, solve a puzzle or undertake a task that gives an indication of their age range	Capacity testing allows a service to estimate a user's age based on an assessment of their aptitude or capacity Services can use capacity testing to assure age without collecting personal data from children These are not commonly in use, can be easily spoofed, and capacity is not always aligned with age
Cross-account authorisation, where a child uses an existing account to gain access to a new product or service. Can be either an age estimation or age verification technique	A user may be asked to 'log in' to their Apple account to download an age restricted app	Authorising accounts are often with large companies such as Apple, Facebook, Google or Twitter The method is dependent on the age assurance used by the authentication account providers (e.g Apple or Facebook), and it is unclear if these providers are able to assure a range of ages Raises concerns around data sharing practices, and entrenching the role of large companies into the architecture of the digital world
Account holder confirmation, such as asking a parent to confirm the age of a child. An age estimation technique	When a child's profile is created (e.g. on Disney+), the account holder who pays for the service is asked to input the child's age	A child's age or age range can be confirmed by an adult, often a parent or carer What is accessible to children in 'children's profiles' is decided by the device provider Requires children to have a caregiver to provide confirmation, which can be a barrier for children in alternative care arrangements These place the burden on the parents to self-declare their age correctly, and are spoofable
Device/operating system controls, which uses a device's 'parental control	When a device is formatted and connected to a family account, (e.g.	A child's age or age range can be confirmed by an adult, often a parent or carer What is accessible to children on 'children's devices' is decided

settings' to assure age. An age estimation technique	Google Family), the account holder whose pays for the service is asked to input the child's age	by the device provider Requires children to have a caregiver to provide confirmation, which can be a barrier for children in alternative care arrangements These place the burden on the parents to self-declare their age correctly, and are spoofable
Flagging, where other users are enabled to 'flag' accounts that seem to be underage for platform moderators to review. An age estimation technique	When any user comes across an account they believe to be 'too young' for the platform, they are able to flag this account to moderators to review	Places the onus on platform users to identify and report underage accounts Only works after an underage child has created an account It is unclear how moderators assure the age of the child once their account is flagged
Digital Identities through third party providers. Third party providers collect ID documents or credentials (e.g. passports, or facial scans) and store them as digital 'wallets'. Can be age estimation, but often Age Verification	A user creates a digital identify with a known provider (e.g Yoti), and then uses their Yoti ID to prove their identity to a platform	Can allow users to share only the attributes required to prove their identity or age The use of digital identities can reduce the need for users to repeatedly provide documents or other official sources of information. It has the potential to minimise data sharing whilst providing a robust measure of age These depend on third party companies that also create privacy and security risks The level of assurance depends on what is collected by the third party
Age tokens through third party providers. Third party providers collect ID documents or credentials (e.g. passports, or facial scans) and create digital age tokens. Can be age estimation, but often Age Verification	A user creates an account with a known provider, and then uses their account to prove their age to a platform	Age tokens contain only the information relating to the specific age or age range of a user, allowing platforms to establish if a user meets their age requirements without collecting other personal information Age tokens may not give a user's actual age and only provide confirmation that a user has passed or failed the service's required age (e.g are they 16 or over). Age tokens can be generated from a digital ID These depend on third party companies though that also create privacy and security risks The level of assurance depends on what is collected by the third party
B2B age assurance, through third party providers. Third party organisations check the age or identity of users. Can be age estimation, but often Age Verification	A user creates an account on a platform, triggering a third party company to do a background age check	These often follow the same process to the hard identifier scheme, but undertaken by a third party provider. Many commercial entities offer background identity or age checks It is unclear how often users know a third party is involved in the assurance process. This also creates data sharing risks