Reset. AUSTRALIA

Letter

Online Safety Bill 2021

2nd of March 2021

To the Senate Environment and Communications Legislation Committee,

Whilst we welcome the passage of the Online Safety Bill 2021 through Parliament, we encourage the careful consideration of this legislation and the implications it has for our broader community.

Our key concern arises from the short period of time between consultation close and the introduction of this Bill within the House. Additionally, we are worried that the short period of time that this Committee has to deliberate, seek evidence and engage with the Australian public before having to prepare its report will not allow for a robust discussion on a law that will have massive implications on our digital environment.

We **strongly urge** the Committee to seek evidence from civil society, academia, industry professionals and the public, via public hearings or other more substantive processes.

We also want to reiterate our commitment to working with the Committee and members of the broader community to arriving at a sensible solution. I have attached our latest submission to the draft Bill made to the Online Safety Branch of the Department of Infrastructure, Transport, Regional Development and Communications made earlier this year.

We look forward to engaging with the Senate Committee as they deliberate on this pivotal piece of legislation.

Regards,

Reset AUSTRALIA

Submission

Online Safety Bill 2020

WHO WE ARE

Reset Australia is an independent, non-partisan organisation 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, the global initiative working to counter digital threats to democracy. As the Australian partner in Reset's international network, we bring a diversity of new ideas home and provide Australian thought-leaders access to a global stage.

EXECUTIVE SUMMARY

The focus of this submission is to encourage a revision of the Act that expands the focus to more comprehensively and more systematically address the harms faced by Australians online.

Reset Australia offers conditional support (see Section 3.0) of the proposed approach of the Act concerning content takedown and moderation, which includes the new powers to deal with:

- Cyberbullying
- Cyber Abuse of Adults
- Non-consensual Sharing of Intimate Images
- Online Content Scheme
- Blocking Measures for Terrorist and Extreme Violent Materials Online

We also note that the context of online harms cannot be understood or adequately addressed without acknowledging and addressing the business model of digital platforms: the attention economy. The attention economy and the associated economic incentives it creates drive a digital ecology where individual and social harms are 'rewarded' and amplified.

Our recommendations to the exposure draft of the Online Safety Act is as follows. We recommend:

- A greater focus on the attention economy, through measures including:
  - Expanding the scope of online harms to reflect impacts to societies and democracy
  - Broader investigative powers that shift from end-user investigation to algorithmic audits
  - Transparency and oversight measures
  - An underpinning framework of comprehensive privacy and data rights
- Specific protections that include:
  - An enforced disinformation code
  - An Australian Democracy Action Plan
  - A robust Privacy Code for children that ensures the maximum levels of protection according to the best interest principle
- Reviewing provisions for the Bill's implementation which include:
  - A proportionate, risk-based approach
  - Proper enforcement measures
  - Meaningful mechanisms for appeal and recourse
  - Cooperation with international best practice

- Finally, we note that the Attorney General is currently reviewing our Privacy Act, which we have submitted to and stressed the importance of data privacy and protection, particularly for children. The Online Safety Act must align with this Code, and between them they must provide a truly robust, child-centred data processing regime that addresses all known online harms children face. We are seeking a commitment to harmonising the Online Safety Act and the Online Privacy Code when it is drafted.

1.0 CONTEXT

In order to design a systematic, adaptive and appropriate policy approach that will address both existing and emerging online harms, an understanding of the underlying drivers must be made clear. This will allow us to tackle the full spectrum of online harms -- from cyberbullying, abuse and violent materials (which this Bill covers well), to deeper manifestations such as threats to democracy and children.

1.1 The attention economy

The attention economy is primarily a 21st century phenomenon that has arisen from the commodification of user attention, driven largely by the digital platforms at first, using huge datasets of user information.

The business models of the digital platforms have a single objective - to capture and maintain user attention in order to maximise advertisements served and profits generated. As such, the algorithms which dictate the content and information we consume are optimised to fulfil this objective, resulting in an attention economy. To feed this machine, the platforms have built a sophisticated system of unfettered personal data collection, building comprehensive profiles of their users that encapsulate their interests, vices, political leanings, triggers and vulnerabilities. This data is then used to predict our engagement behaviour, constantly calculating what content has the greatest potential for keeping us engaged. This content has been shown to lean towards the extreme and sensational, as it is more likely to earn higher engagement[1,2].

This has resulted in the explosion of a data economy that has been facilitated through the commoditisation of personal information. This model, termed 'surveillance capitalism' by Shoshanna Zuboff,[3] is predicated on the extraction and exploitation of personal data for the primary purpose of predicting and changing individual behaviour. This emerging model (spearheaded by Google and later Facebook) sets a dangerous precedent for adoption by other industries, and flies against Australian ideals of autonomy, public safety and privacy.

The algorithms built by these companies dictate all of the content and information we consume. The use of services provided by the major digital platforms have become ubiquitous to the Australian way of life. With over 85% of Australians using social media 'most days',[4] the role that the digital platforms such as Facebook (including Instagram and WhatsApp), Twitter, Snapchat, TikTok and Google (including YouTube) play in our society has become fundamental to how we live, work and entertain ourselves.

From the acceleration in the breakdown of public trust in institutions, democracy and civic debate evidenced through the 2016 US Presidential Election and Brexit, to the public health risks associated with Covid-19 and anti-vaccination disinformation, and harmful content pushed to children, we are

---

[1] Vosoughi et al. (2018), 'The spread of true and false news online', *Science* found at https://science.sciencemag.org/content/359/6380/1146

[2] Nicas (2 Feb 2018), 'How YouTube Drives People to the Internet's Darkest Corners', *Wall Street Journal* found at https://www.wsj.com/articles/how-youtube-drives-viewers-to-the-internets-darkest-corners-1518020478

[3] Zuboff S (2019), 'The Age of Surveillance Capitalism,' Profile Books, London

[4] Yellow Social Media Report (2020) Part One: Consumers. Found at: https://2k5zke3drtv7fuwec1mzuxgv-wpengine.netdna-ssl.com/wp-content/uploads/2020/07/Yellow_Social_Media_Report_2020_Consumer.pdf

starting to experience the spectrum of harms that have arisen from this relationship. Most importantly, as the way online harms manifest are only just emerging and as more industries seek to capitalise on user data and the 'attention economy', the scale and scope of online harms will surely increase.

1.2 How the attention economy causes harm

From foreign interference in our democracy, the amplification of disinformation and extremist voices that drive division, to threats to the safety of our children, the societal harms caused by this unfettered and unregulated system have begun to emerge in earnest. In particular, the capacity for granular targeting down to specific communities and even individuals gives rise to a completely unprecedented landscape. Whilst these harms sometimes fall outside of what is considered illegal, their negative effects on an Australian way of life are clearly evident.

Harm to Society	Example of Harms
	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[5].
Foreign Interference	A number of the 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[6].
Amplification of Disinformation and Extremist Voices	Datasets were collected from six public anti-vaccination Facebook pages across Australia and the US, with it appearing that although anti-vaccination networks on Facebook are large and global in scope, the comment activity sub-networks appear to be 'small world'. This suggests that social media may have a role in spreading anti-vaccination ideas and making the movement durable on a global scale[7].
Safety of Children	A leaked Facebook document prepared by Facebook Australian executives outlines to advertisers their capability to target vulnerable teenagers as young as 14 who feel 'worthless', 'insecure' and 'defeated' by pinpointing the "moments when young people need a confidence boost" through monitoring posts, pictures, interaction and internet

[5] "Bots stormed Twitter in their thousands during the federal election" by Felicity Caldwell, The Sydney Morning Herald (July 20, 2019)

[6] "Russian trolls targeted Australian voters on Twitter via #auspol and #MH17" by Tom Sear, Michael Jensen, The Conversation (Aug 22, 2018)

[7] "Mapping the anti-vaccination movement on Facebook" by Naomi Smith & Tim Graham, Information, Communication & Society

activity in real-time[8].

Table 1: A selection of examples of societal harms caused by an unregulated attention economy

1.21 The attention economy and democracy

The data collection systems and resultant 'attention economy' has left us extremely vulnerable to many different forms of manipulation by foreign and malicious actors who wish to threaten the Australian democractic process, exploit our declining trust in our public institutions or generally divide Australian society at large.

The effects of this manipulation have already begun to be seen in Western democracies around the world, weaponising our personal information to drive division and interfere for geopolitical or financial gain. In particular, the capacity for micro-targeting on the digital platforms is completely unprecedented, exacerbating the effect of mis/disinformation whilst also making it much harder to regulate. Additionally, divisive, sensationalist clickbait has been shown to spread faster online, allowing foreign actors to be able to 'game' this system and peddle mass amounts of content with the intention of driving polarisation.

> *'unlike heritage media, digital and social... can be done in the "dark," so your opponents may not even be aware of the message you are pushing out'.*[9]

As clearly documented in the Australian Strategic Policy Institute's Hacking Democracies report[10], the issue of foreign entities utilising the digital platforms to interfere in democracies is pervasive and global. In particular:

- the intentional Russian interference in the 2016 US Presidential election, with bought ads designed to exploit division in society for political gain[11,12] and,
- the Cambridge Analytica scandal which leveraged user data to serve curated Brexit messaging[13,14]
- 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[15]

---

[8] "Facebook targets 'insecure' young people" by Darren Davidson, The Australian (May 1, 2017)

[9] Hughes (2 May 2019), 'Facebook videos, targeted texts and Clive Palmer memes: how digital advertising is shaping this election campaign', *The Conversation* found at: https://theconversation.com/facebook-videos-targeted-texts-and-clive-palmer-memes-how-digital-advertising-is-shaping-this-election-campaign-115629

[10] Hanson F et al. (2019) 'Hacking Democracies; cataloguing cyber-enabled attacks on elections', *ASPI Policy Brief* found at: https://s3-ap-southeast-2.amazonaws.com/ad-aspi/2019-05/Hacking%20democracies_0.pdf?.RKLLc8uKm1wobfWH1VvC.C88xGWYY29

[11] Kelly et al. (22 Aug 2018), 'This is what filter bubbles actually look like', *MIT Media Review* found at: https://www.technologyreview.com/s/611807/this-is-what-filter-bubbles-actually-look-like/

[12] Shane (1 Nov 2017), 'These are the ads Russians bought on Facebook in 2016', *New York Times* found at: https://www.nytimes.com/2017/11/01/us/politics/russia-2016-election-facebook.html

[13] Scott (30 July 2019), 'Cambridge Analytica did work for Brexit groups, says ex-staffer', *Politico* found at: https://www.politico.eu/article/cambridge-analytica-leave-eu-ukip-brexit-facebook/

[14] BBC News (26 July 2018), 'Vote Leave's targeted Brexit ads released by Facebook', https://www.bbc.com/news/uk-politics-44966969

[15] Workman M, Hutcheon S (March 16 2019), 'Facebook trolls and scammers from Kosovo are manipulating Australian users', *ABC News*

How these platforms facilitate broader harm isn't a theoretical possibility anymore, but tangible threat to our liberal democracy and cohesive society.

Case study: Bots Storm 2019 Federal Election[16]

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.
- Research into the US election by ANU indicated that the average bot was 2.5 times more influential than the average human. This was measured by their tweets and increased success at attracting exposure via retweets.
- Dr Graham said he was still examining the data to see what the Australian bots were tweeting about and whether they were partisan and it was still unknown who created them.
- "From a national perspective, the working hypothesis could be that if these are indeed bots, then they're being deployed by interested parties," he said.

1.22 The attention economy and children

The attention economy has particular consequences for children. The drive for limitless data collection has created a generation of children that are 'datafied from birth'[17]. From devices collecting data in utero[18], to connected toys and devices like a Barbie that analyses children's voices[19], to education and health care data routinely collected as part of childhood, the amount of data collected by third parties about children is truly staggering. On top of this, it is estimated that parents will post 1,300 photos and videos of their children online by the time they are 13[20]. Before reaching any sort of age of consent, masses of data has already been collected and processed about children.

This is troubling because data has a particular problem with permanence, and children have a long time to live. Once collected data does not degrade or erode without specific action. Without regulatory protections there is no way to be sure where data collected about children will be processed, or when, or indeed knowledge about if it is being used to harm children, or may harm them in the future. This is a violation of their right to privacy.

The precautionary principle is not exercised in the attention economy: even though we do not know the consequences of this extensive data collection or where or how this data may be used in the future, it is still collected and stored en mass. This encapsulates the huge power and information

[16] Housego (21 April 2019), 'Australian election targeted by Twitter bots', AFR found at: https://www.afr.com/politics/federal/australian-election-targeted-by-twitter-bots-20190426-p51hkc

[17] Childrens' Commissioner for England and Wales 2018 'Who knows what about me?'

[18] Barassi, V. 2017 'BabyVeillance' Social Media and Society

[19] "Privacy fears over smart barbie that can listen to your kids" by Samuel Gibbs, The Guardian (March 13, 2015)

[20] Childrens' Commissioner for England and Wales 2018 'Who knows what about me?'

asymmetry that fuels the attention economy. Children have no way of knowing how much is known about them or by whom, and importantly how this information will be deployed in their lives. At its heart, information about their personal life and experiences has become proprietary data in a business model that may not have their best interests at heart.

Case study: Recommendation systems in YouTube and children's online safety[21]

YouTube is the most popular video streaming service in Australia, reaching around 16.2 million adult Australian's each month[22]. In January 2020, 799,000 Australian children aged 2-18 accessed YouTube, and watched an average of 18.86 hours of content over the month[23].

Much of the content they consume will have been served to them by YouTube's recommendation system and autoplay. In 2018 YouTube's Chief Product Officer stated that 70% of viewing time was guided by their AI assisted recommendation system[24].

Not everything YouTube recommends will be safe, and there are many examples where YouTube content has caused harm to children. There are known cases of; far-right extremist groups using YouTube to recruit children as young as 12[25]; ten-year-olds being served negative body image content after searching for tap dancing videos[26]; of YouTube hosting content that is designed to appeal to children but has deeply age inappropriate themes[27], and; of YouTube promoting extremist and radicalising right-wing influencers to young people[28] . (And on the other side of the screen, YouTube's algorithm has been known to recommend 'family videos' of young children to adults who appear to have a sexual interest in children[29]). YouTube's recommendation system can either play a key role in either serving up harmful content, or restricting and reducing its spread.

Given that the majority of content Australian young people access through YouTube comes from their recommendation algorithm, it would make sense for the regulator to ensure algorithmic accountability for this system.

This is a direct example of the potential harms of the attention economy in action. While there is limited public data available about the inner workings of their recommendation system, a research paper published by Google, YouTube's parent company, outlined that their recommendation system was trained to increase watch time[30]. Their algorithm takes into account personal data (such as your previous viewing history and location) and content specific data (such as popularity of content and 'freshness') to rank and decide content to recommend and 'play next'. At no stage was any correction for harm nor considerations about the age-appropriateness factored into the recommendation

---

[21] While YouTube offers a specific service for the under 13s 'YouTube Kids', there is significant evidence that under 13 year olds regularly and frequently access YouTube's main service (Pew Research Centre, Many Turn to YouTue for Children's Content 2018)

[22] Nielsen Digital Landscape Jan 2020

[23] Calculated from Nielsen Digital Landscape Jan 2020

[24] "YouTube's Product Chief on Online Radicalization and Algorithmic Rabbit Holes" by Kevin Roose New York Times (March 29, 2019)

[25] "Far right recruiting children on YouTube" by Tom Knowles, The Times (Oct 6, 2020)

[26] Mozilla Foundation, YouTube Regrets 2019

[27] "The disturbing YouTube videos that are tricking children" BBC News (26 March 2017)

[28] "Alternative Influence: Broadcasting the Reactionary Right on YouTube" by Rebecca Lewis Data and Society 2018

[29] "On YouTubes digital playground, an open gate for paedophiles" by Max Fisher and Amanda Taub New York times (June 3, 2019)

[30] Paul Covington et al, 'Deep Neural Networks for YouTube Recommendations' 2016 *Proceedings of the 10th ACM Conference on Recommendation Systems*, ACM, New York, NY, USA

algorithm. The sole focus on maximising watch time could allow individual and social harms to flourish unfettered.

2. THE POLICY APPROACH REFLECTED IN THE ACT

Many of the reforms proposed in the Online Safety Act are extremely welcome. Australia has played a leading global role in online safety, specifically around content reporting and take down, and responding to bullying and abuse for children. It is right that these are being strengthened and extended to adults too.

However, as the broader discussion around the Attention Economy highlights, there are many other aspects of online harms that would benefit from an equal focus in the Act. There are some places where the requirements of the Act would be improved if they: shifted focus to be 'upstream' of harms; focussed on the role of digital service providers in creating a safe online experience in the first instance, and; expanded focus to recognise the breadth of harms Australians face online.

2.1 Shift the focus upstream of harms, before content moderation and take down are necessary

The policies and the new powers proposed in the *Online Safety Act* around the following are vitally important to ensure the safety of all Australians online:

- Cyberbullying and content takedown
- Cyber abuse of adults
- Non-consensual sharing of intimate images
- Determination and takedown of seriously harmful material
- Blocking measures for terrorist and abbohernt violent material online

Reset Australia has previously expressed support for, and continues to support, these proposed new powers. The requirement to reduce the time for takedown and civil penalties for perpetrators of cyber abuse are especially welcome, as they will provide pathways for recourse for victims and more robust mechanisms to ensure illegal content is eliminated online.

However, there are two key issues with this. Firstly, this focus is 'downstream' of harms, and requires them to occur before any actions happen. While downstream measures are a necessity in creating a safer digital world and preventing ongoing harm, these must be coupled with 'upstream' systemic interventions that prevent harm in the first place.

Secondly, it fails to address the deeper structural causes which drive the creation and promotion of harms through the attention economy. Without an explicit focus on the digital services and their designs, which lean toward the extreme and sensational, a moderation and take down approach will only ever be playing catch up.

Furthermore, content takedown and moderation policies are not adaptive enough to the types of content - such as unduly polarising, hateful or misinformative content - that is currently legally allowed but nevertheless is the cause of significant harm through inciting hate and violence or intentionally misinforming the public on important issues.

While content moderation policies are an important avenue to mitigate some of the worst of these harms, they are ill-equipped to regulate the profit model of these platforms that exploit user attention and drive vast profit through serving harmful disinformation.

Current Focus:

Content takedown/ moderation

The problem is seen to be caused by malicious actors, whether they be terrorists, cyberbullies or perpetrators of hate speech

The scope is content which is illegal (black & white)

The solution is seen the be policies which enforce platforms to deploy more robust content moderation practices (take down)

Future Focus:

The attention economy

The problem is seen to be the exploitation of user data & algorithms to maintain user attention, resulting in the amplification of extremist and sensational content

The scope becomes design & practices which cause societal harm and division

The solution is policies that promote transparency, regulate algorithmic amplification, and protect data rights and privacy

This focus is curious given that the eSafety Commission is a leading global advocate of systemic, upstream interventions with their Safety by Design principles. The Online Safety Act however does not address this and leaves any mention of this broader focus to a 'yet to be determined' BOSE.

It appears that the upstream focus is left to the Basic Online Safety Expectations, which are yet to be agreed and may not have the same regulatory force at the Act. We worry this could create lopsided requirements, with too little focus on prevention and too much focus on take down and moderation.

2.2 Embrace a more systemic focus on the role of digital service providers in creating a safe digital experience

We note that the exposure draft was is introduced with commentary that compared the issues of online safety with the 'small number of human interactions that go wrong offline'[31]. However, the problems of online safety are systemic for children and adults. Nothing in the digital world has 'come to be' by pure accident, and all the services that will be covered by the Act are designed and curated within an attention economy -- the Act needs to expand its focus to adequately address this.

[31] DITRDC Fact Sheet, Online Safety Bill 2020

The focus of the Cyber-bullying, Adult cyber-abuse, Image-based Abuse schemes position digital service providers as an 'go-between' between individual users who generate content, and individual users who consume content. When a user complains, these schemes enable regulation of other user's generated *content*. They do not provide enough provision to regulate the *service* digital platforms provide (bar their take down mechanisms). The Act is a timely opportunity for regulation to keep pace with innovation, and to enable regulators to 'lift the hood' and look at how these 'mere' go-betweens provide their service.

A focus on the service and design of the service is needed to prevent another lopsided policy focus. This includes, for example, the algorithms that companies created to search, refine and serve up content. This focus would also improve efficacy against offensive online content and abhorrent violent material.

2.3 Expand the scope to cover all harms

The Act addresses a limited set of risks. We agree with the Government's definition of 'online safety' as the harms that can affect people through exposure to illegal or inappropriate online content or harmful conduct. But we strongly believe that this definition and subsequent policy focus must be expanded to include the ways the digital platforms enable harms not just to individuals but to our communities, democracy and society.

As shown, our current digital architecture has been built to incentivise the propagation of disinformation and division within our communities, resulting in demonstrable harm not just to individuals and communities but Australia as a democratic sovereign state.

It is not just illegal or inappropriate online content or harmful conduct that is causing harm to our society.

For children specifically, this focus on illegal and inappropriate content at an individual level means that many commonly known pathways to harm remain unaddressed. The most current categorisation of online risks for children is the 4Cs, recently updated by Sonia Livingstone for Children Online: Research and Evidence. This framework highlights the many ways the attention economy is harming children, from misinformation and polarisation to data privacy risks.

The Cyberbullying Scheme presents some excellent remedies for Content and Conduct risks, and the Online Content and AVMB scheme addresses Content risks in world leading ways. However they miss the opportunity to comprehensively tackle contact and contract risks, as well as many cross cutting risks. These require a more systemic focus, embedded in the attention economy and its economic imperatives, to successfully address. Table 2 below overlays these risks with the focus of the Act, where the grey cells represent the risks where the Act could be broadened to more sufficiently addressed.

	Content Risks	**Conduct Risk**	**Contact Risks**	**Contract Risks**
Aggressive	Abhorrent violent material Violent, gory,	Bullying Hateful or hostile	Harassment, stalking, hateful behaviour, unwanted surveillance	Gambling, scams, identify theft, fraud, phishing, security risks

	graphic, hateful and extreme materials	peer behaviour, e.g. trolling, shaming, exclusion		
Sexual	Class 1 and Class 2 materials Pornography, sexualisation of culture, body norms	Sexual harassment, non consentual sexual messaging Sexual pressure	Sexual harassment, grooming, generating or sharing CSAM (NB: The take down provisions of the Act could reduce sharing CSAM, but do not address contact mechanisms, grooming nor self generation risks specifically)	Sextortion, streaming CSEA (NB: The take down provisions of the Act could reduce CSEA streaming, but do not address sextortion specifically)
Values	Mis/disinformation, age inappropriate content	Potentially harmful user communities, e.g. anti-vaccine, peer pressure	Ideological persuasion, radicalisation and extremist recruitment	Information filtering, profiling bias, polarisation, persuasive design
Cross cutting	Privacy & data protection abuses, physical and mental health risks, discrimination			

Table 2: The Four Cs of child online safety [32]

3.0 COMMENTS ON SPECIFIC PROVISIONS

This section highlights some key concerns we have with this proposed Bill. Please find the associated recommendation in Section 4.0.

Section 42 + Section 132 and 133 - Basic Online Safety Expectations and Industry Codes/Standards

Currently, this Act has taken a broad and expansive view on setting safety expectations. We commend the intention of this approach which demonstrates the Department's appetite to both create future-proof policy levers to deal with emerging online harms as well as recognising the seriousness of this issue through some of the enforcement measures detailed. As the rest of the Bill is concerned with content moderation, we see these provisions under the Basic Online Safety Expectations and the powers to set future Industry Codes and Standards represent a key pathway to understand and mitigate the harms that have arisen from the attention economy. Much of the functioning of the Act will depend on the content and enforcement of the BOSE code. Given this, the BOSE must include robust requirements to adequately safeguard children. While we expect to engage in the many consultations around these industry codes, the starting point for the development of the BOSE must be Safety by Design.

However, broad legislative approaches (especially within issues that are only beginning to emerge) must be tempered with appropriate checks on power, have clear mandates to consult and incorporate guidance from academics, civil sector actors and the general public and be built from a rights/principles based framework.

[32] CORE Updating the 4Cs of online risk 2020

As such many of our recommendations seek to incorporate some of these checks and balances. In particular:

- Transparent reporting and rationale of when powers are used/not used (4.14)
- Diverse and multi stakeholder oversight (4.14)
- An aligned framework of privacy and data rights (4.15)
- A proportionate risk-based approach (4.31)

Should be incorporated to not just provide accountability, but enhance the trust, perceived legitimacy and ultimate intention and function of this Bill to keep all Australians safe online.

Various - Removal Notices, Link Deletion Notices, App Removal Notices, Blocking Requests

These content moderation and takedown powers must work within a system of transparency and oversight (4.14) and meaningful appeals processes (4.33). Additionally, this in turn must be built of individual user rights to data and privacy (4.15). These safeguards will ensure an added layer of accountability that will improve the functionality of this Bill.

Section 43 - Commissioner may refuse to investigate certain matters

In line with our transparency recommendations below (4.14), if the Commissioner refuses to investigate certain matters, ensure that there is a transparent reporting on the reasoning for this decision.

Section 193 and 197 - Information-Gathering Powers and Investigative Powers

Whilst we strongly recommend that these powers are guided by stringing individual data and privacy rights (4.15), the shift in investigative focus must be turned to the platforms that have engendered and facilitated these harms. As such we recommend that these powers detail clearly defined scopes and empower the Commission to undertake algorithmic audits (4.12).

4.0 RECOMMENDATIONS

4.1 Addressing the attention economy

In order to unpack how the attention economy causes harm, this Bill should focus on prevention and limiting the spread of harmful content, not just taking it down after it has already caused harm. *This means that the regulator needs to be given more powers to investigate and comprehensively regulate recommendation algorithms and the underlying data extraction practices which enable this.* Deranking reported offending content is a good start, but this needs to be extended further. Content recommendation algorithms also need to 'derank' misinformation and other content that harms. Especially for children, we know that the majority of the content young people consume is served to them by recommendation algorithms so it is essential that these deliver a safe experience.

4.11 Expanding the Scope

Expand the current definition of serious harm and online safety to encompass the full range of modern online harms. As detailed, this must be expanded to reference the harm caused to communities, societies and democracy, and remain adaptive in order to capture new and emerging

technologies and innovation. In line with the UK Online Harms White Paper, this might be explicitly stated as online content or activity that:

- harms individual users, particularly children
- threatens our way of life in Australia either by
    - undermining national security OR
    - reducing trust and undermining our shared rights, responsibilities and opportunities to foster integration

This expansive framework of conceptualising online harms must be integrated in the language and approach of this Bill in order to adequately address this issue.

Recommendation
Expand the definition of online safety/online harms to be able to capture the harms caused to communities, society and democracy. Ensure that this expanded definition adequately addresses the 'values risks' faced by children.

This expanded scope must align with the ongoing Privacy Act review.

Recommendation
Commit to a review of the Online Safety Act and its associated Codes as the Privacy Act is reviewed, to ensure that all legislation is aligned and provide strong protection for children.

4.12 Investigative Powers - shifting from end-user investigation to algorithmic audits

The focus of this Bill to investigate end-users must incorporate principles of privacy, appeal/recourse and purpose limitation (detailed below) however are short-sighted when it comes to properly unpacking how these harms manifest - we must understand the 'black box' algorithms which facilitate them.

As such, an independent regulator must be given mandatory investigative powers via algorithmic audits.

The harms caused by the digital platforms, ranging from foreign interference to disinformation, needs a holistic approach and the remit of this authority should expand to provide insights into bigger questions - such as how platform curation algorithms open up risk and create harm to the public. Importantly, this isn't at the exclusion of platform/publisher content visibility issues remedied by this Bill, merely an expansion that might provide a systematic legislative approach, rather than one focussing on a specific sector

The systematic impacts of algorithmic amplification - that is the promotion/demotion of content that is currently dictated by the digital platform's internal algorithmic processes - is an issue that goes far beyond traffic and advertising revenue, and requires an expansive remit to address. Unilateral algorithmic curation and amplification has an outsized harmful impact on the Australian public and our democracy.

Information on these harms is held solely by the digital platforms, who do not make it available for transparent independent review under any circumstances. It seems extraordinary that the digital

platform companies have all the data and tools needed to track, measure and evaluate these harms - indeed these tools are a core part of their business, but they make nothing practical available for public oversight, even as they avoid all but the most basic interventions to protect the public from harm.

Without mandated access, regulators are forced to rely on the companies to police themselves through ineffective codes of conduct. This failed approach has been seen overseas and yet is still being tried here in Australia.

This is not an impossible suggestion as the digital platforms might make you believe. Algorithmic audits have been specifically proposed in the EU Digital Services Act (DSA), and represent a clear model to emulate here in Australia. Our legislative approach must be as flexible and encompassing as the harms we seek to address.

Algorithmic Audits

An algorithmic audit is a review process by which the outputs of algorithmic systems (in this case the curation systems of the digital platforms which display content) can be assessed for unfavourable, unwanted and/or harmful results. In addition to assessing if design decisions within the digital platform algorithms are actively anti-competitive, this process can also be used to assess numerous online harms to wider society and democracy such as disinformation and foreign interference.

How would an audit authority work?
The authority must have the ability to carry out an algorithm inspection with the consent of the digital platform company; or if the company does not provide consent, and there are reasonable grounds to suspect they are failing to comply with requirements, to use compulsory audit powers. It must be resourced (financially and technically) to carry out these actions, but they should also have the power to instruct independent experts to undertake an audit on their behalf. Examples for how this might be structured can be seen in multiple industries from aviation to drug therapeutics.

Recommendation
Institute an audit authority under an independent regulator empowered to investigate/audit the impact of algorithmic amplification on Australian society

4.14 Transparency and Oversight

We welcome the Acts' provisions on reporting, in particular the power for the Commission to request periodic reports. However, transparency must be embedded into all aspects of this Bill, including:

- Reporting when and why certain powers are enacted. This should be provided in an accessible, timely and (if required) redacted format. This includes (but isn't limited to):
  - Content moderation (removal notice, blocking request, link deletion, app removal) provisions contained in the Abhorrent Violent Material Scheme, the Online Content Scheme and the Cyber Abuse/Bullying sections
  - Decisions around complaints -- and whether the Commission is taking them on or not
  - The use of information and investigative powers

  - Processes that companies have in place for reporting illegal and harmful content and behaviour, the number of reports received and how many of those reports led to action.
- The rationale for the determination of BOSE and other industry codes/standards.
- Reporting on the progress and implementation of this Bill in achieving a safe online environment for all Australians

Recommendation
Transparent reporting on decision making processes in relation to the Bill's powers

Additionally, in order to build the needed legitimacy, trust and effectiveness of this regime, we recommend instituting an independent public-private-citizen multi stakeholder oversight board to oversee, give advice and provide accountability for the various powers laid out in this Act and our submission. This should include reviewing:

- Content takedown provisions detailed in the Online Content Scheme and Abhorrent Violent Materials sections
- When the Commission begins BOSE and Industry Code/Standard formation
- Appeals and recourse pathways detailed in Section 4.33 of this submission

Recommendation
Ensure academic, civil and general public engagement and oversight by instituting an external, independent advisory board that will provide advice and accountability for the appropriate provisions in this Act

4.15 Privacy and Data Rights

The proposed powers granted within this Act, in particular powers that are directed at end-users (see Part 13 & 14 Information-Gathering Powers and Investigative Powers) and the additional powers we have recommended in this submission must operate within a broader framework of privacy and data rights.

This should be done through the current review of the Privacy Act and incorporate elements of the European experience, in particular a rights-based approach with regard to their data subjects, can help ensure proper protection of Australians' privacy.[33] This framework is essential in relation to the Online Safety Act for two primary reasons.

1) It provides users with a mechanism to regain control over their personal data and provide a pathway to unpack the attention economy
2) It establishes a framework for checks and balances against some of the expansive powers proposed in this Bill and submission

Recommendation
In particular, we support the incorporation of the following rights. We recognise that this would more appropriately sit under an updated Privacy Act - however it is fundamental that the powers of this Act align.

- Right to Erasure, as in Article 17 GDPR

[33] Regulation (EU) 2016/679 2016, Chapter 3

The data subject shall have the right to obtain from the controller the erasure of personal data concerning him or her without undue delay and the controller shall have the obligation to erase personal data without undue delay

This is especially important for children, who must have an ensured access to this right so that they may delete all data held about them easily.

- Right to Data Portability, as in Article 20 GDPR
  The data subject shall have the right to receive the personal data concerning him or her, which he or she has provided to a controller, in a structured, commonly used and machine-readable format and have the right to transmit those data to another controller without hindrance from the controller to which the personal data have been provided

- Right to Object, Article 21 GDPR
  The data subject shall have the right to object, on grounds relating to his or her particular situation, at any time to processing of personal data concerning him or her

- Automated individual decision-making, including profiling, Article 22 GDPR
  The data subject shall have the right not to be subject to a decision based solely on automated processing, including profiling, which produces legal effects concerning him or her or similarly significantly affects him or her

4.2 Ensuring specific protections against harms

4.21 An Enforced Disinformation Code and Democracy Action Plan

Immediate action must be taken to understand and tackle the digital platform's role in facilitating disinformation, hate speech and polarisation. Whilst we are aware and engaged with developments within the ACMA regarding a voluntary code of conduct on disinformation, we would like to reiterate that self-regulatory pathways will not work. As mentioned in their Assessment of the EU Code of Practice on Disinformation[34]:

> *At present, it remains difficult to precisely assess the timeliness, comprehensiveness and impact of the platforms' actions, as the Commission and public authorities are still very much reliant on the willingness of platforms to share information and data. The lack of access to data ... (along with) the absence of meaningful KPIs to assess the effectiveness of platform's policies to counter the phenomenon, is a fundamental shortcoming of the current Code.*

Recommendation
Develop a mandatory and enforceable Code of Practice on Disinformation.

This Code should align with developments made in the EU and within the current ACMA process, specifically creating provisions that will:

[34] European Commission (2020), 'Staff Working Document: Assessment of the Code of Practice on Disinformation - Achievements and areas for further improvement'. Found at: https://ec.europa.eu/digital-single-market/en/news/assessment-code-practice-disinformation-achievements-and-areas-further-improvement

- disrupt monetisation and advertising incentives for disinformation
- provide avenues for meaningful data access for academic researchers, civil sector actors, think tanks and public regulators to undertake the requisite research on disinformation, as to increase public understanding of these harms

For more information, please refer to our submission to the Australian Code of Practice on Disinformation.

The threats to democracy are much broader, complex and intricate, encompassing harms that are only beginning to emerge. In order for our country to be resilient and adaptive to emerging harms, we must begin to develop frameworks that better allow us to counter digital threats to democracy.

This process should take guidance from the development of the European Democracy Action Plan,[35] which centres on three main pillars:

- Promoting free and fair elections
- Strengthen media freedom and pluralism
- Countering disinformation

Using this as a framework, the Australian Government should embark on a consultative process for our own framework.

Recommendation
Develop an Australian Democracy Action Plan

4.22 Harms to Children - Maximum Privacy Protections

We note that the Attorney General is currently exploring this in the review of the Privacy Act, which will address data privacy and protection for young people. The Online Safety Act must align with this, and between them they must provide a truly robust, child-centred data processing regime. Building on the UK's *Age Appropriate Design Code* (2020) and Ireland's *Fundamentals to a Child Oriented Approach to Data Processing* (2020), this Code must ensure that children's data is processed in ways that prevent commercial and contract harms. By taking a child-centred, 'best interest' approach these international codes provide a framework for regulatory requirements that safeguard children and their data from commercial and other online harms. Unless our Privacy Act review or the Online Safety Act truly create a fit-for-purpose protection mechanism, many of the online harms children face will slip through Australia's regulatory net. *A commitment to harmonising the Online Safety Act and our Privacy Act to create a child-centred, best interest data processing framework must be made.*

Recommendation
Implement a robust Privacy Code that governs the processing of children's data in accordance with the best interest principle.

The *Act (or the Code that stems from it) must include a requirement for service providers to proactively consider children's safety, through proactive child impact assessments.* While the

[35] European Commission (2020), 'European Democracy Action Plan', Found at: https://ec.europa.eu/commission/presscorner/detail/en/ip_20_2250

requirements for periodic and non periodic reports about compliance with the Basic Online Safety Expectations code are welcome, these requirements could be shaped to be even more upstream and proactive. There should be a requirement for 'child impact assessments' to be conducted by any service that falls under the scope of the Act, before they offer or update any service for children. This would ensure that companies do not see child online protection as an afterthought, and indeed that they act on issues identified before the eSafety Commission needs to intervene.

4.3 Act Implementation

4.31 Proportionate Risk-Based Approach

The digital world is an ecosystem of providers and services that work together. To create a safer ecosystem, the *Act must apply to all digital services and employ a risk-based approach for assessing potential harm, in particular services which are likely to be accessed by Australian children.* Given the global scale and numbers of users who can generate harmful content or conduct, it makes sense to focus as well on the role some companies play in promoting this content in the first place.

This approach should align with the development of the UK's Online Harms White Paper which states:

There would be a new statutory duty of care to make companies take more responsibility for the safety of their users. This duty would be risk-based and proportionate and focused on systems and processes, not individual pieces of content. Important principles would apply to the regulatory framework including users' rights to freedom of expression and privacy, innovation and protecting small and medium- sized enterprises.

This approach is inherently adaptive, avoiding a 'one size fits all approach' to companies and harms. It should also be guided by:

- a particular emphasis on protecting children
- ensuring a pro-innovation approach
- protecting freedom of expression online

Recommendation
Establish a proportionate and risk-based approach to defining digital platform obligations and approach to online harms reduction

This might look like:

- More powers to comprehensively investigate and require modifications to content recommendation systems and algorithms
- Requiring service providers to proactively consider children's safety, through proactive child impact assessments

4.32 Enforcement

We welcome the Bill's approach to enforcement of providing a spectrum of differing enforcement mechanisms. Our two key recommendations look at:

- Ensuring that equally proportionate enforcement measures are applied to BOSE and industry code/standards determinations as we believe this is the most viable way under the current Bill to mitigate attention economy harms

- Ensuring that civil penalties that would potentially be levied are proportionately disincentivizing to their scale

Additionally, to ensure that this Act is enforced with the cooperation of industry, service providers must incorporate the guidances issued by the regulator in its code of practice into relevant terms and conditions.

Recommendation
Ensure that civil penalties imposed onto the platforms are proportionate to the magnitude of the organisation so that they are properly disincentivising e.g. 10% of global annual turnover

Recommendation
Ensure that the BOSE and Industry Code/Standards determinations are backed up by a spectrum of proportionate enforcement measures

Recommendation
Ensure effective enforcement by exploring how digital platforms' own relevant terms and conditions incorporate guidance issued by the regulator in its codes of practice

4.33 Procedure, Appeal and Recourse

This Bill details expansive powers related to content takedown, link deletion, user blocking, account deletion and app removal. This is particularly in reference to powers detailed in the Abhorrent Violent Material Blocking Scheme and the Online Content Scheme (Part 8 and 9 respectively). Whilst we agree that that dangerous, violent and hateful material must be taken down, and that an independent regulator must be empowered to do so - proper transparent and procedural mechanisms must also be in place so that the original intention of these provisions may be upheld.

As such, we recommend several key changes be incorporated into the administration and implementation of these powers, including:

- Clearly publishing the impetus, reasoning and decision-making processes behind decisions to remove content, block and/or delete users/accounts and app removals
- Ensure users/companies have an effective route for appeal that includes independent evaluation, due process, timely and effective complaints management infrastructure, escalation procedures and transparent decision-making processes
- Consider instituting an independent public-private-citizen multi stakeholder oversight board to review escalated matters

Recommendation
Institute clear procedural infrastructure for appeal and recourse for individuals and companies

4.34 Miscellaneous

Recommendation
Where relevant, evidence of cooperation with UK law enforcement and other relevant government agencies, regulatory bodies and public agencies.

5.0 CONCLUSION

This is a bold legislative agenda that sets an intention for Australia to join the forefront of the policy movement tackling online harms. However, we would like to stress that this intention must turn into tangible actions that shift our focus beyond content moderation and takedown, and toward tackling the harms of the attention economy. Only by this approach, might we develop a framework that is encompassing and adaptive enough to actually keep Australians safe online.

We look forward to working with the Department and other stakeholders as we collectively work towards a better online environment.