Reset.
AUSTRALIA

Response to the *Amending Online Safety (Basic Online Safety Expectations) Determination* 2022 consultation

Reset.Tech Australia
Feb 2024

Reset.
AUSTRALIA

Summary

Reset.Tech Australia warmly welcomes the expansion of the Basic Online Safety Expectations (BOSE) as proposed in the amendments, and in particular:

- The increased focus on covering more systems;
- Improving transparency and accountability, and;
- The introduction of the children's best interests principle.

The ambition and direction of travel of these proposals is both necessary and timely. They will help reposition Australia as—once again—world leaders in the ambition to create a safe and secure digital world.

This submission outlines key areas where the overall amendments could be specifically strengthened, and areas where additional responsibilities could be effective to help realise this ambition.

Our proposals are informed by original research undertaken in Australia and comparative policy analysis, drawing on examples of best practice policy emerging around the world. Cognisant that the review of the BOSE is a precursor to a review of the *Online Safety Act*, and the interconnections between the two, some potential amendments to the *Online Safety Act* are also briefly outlined, to highlight how the BOSE could be made more enforceable and effective.

Specifically, Reset.Tech Australia recommends that the BOSE:

For subsection 6(2A):
- Include an additional requirement that 'best interests assessments' are to be undertaken and publicly released.
- Consult with children and young people around the development of elements of the BOSE that affect them, including 'best interests' requirements.

For subsection 6(5&6):
- Include an additional requirement that providers take reasonable steps to avoid deploying dark patterns on end-users.
- Include requirements that providers make all decision making regarding user controls must consider children's best interests as a primary consideration.
- Include 'allowing users to turn off recommender systems' as an example of a reasonable step.
- Include 'ensuring independent audits of the user control systems' as an example of a reasonable step.

For subsection 8A:
- Include 'retraining generative artificial intelligence that has been trained on illegal material' as an example of a reasonable step.
- Include 'ensuring that training materials for generative artificial intelligence capabilities and models comply with the APPs' as an example of a reasonable step.
- Include requirements that providers make all decision making regarding generative AI must consider children's best interests as a primary consideration.
- Include 'ensuring independent audits of the function of AI systems' as an example of a reasonable step.

For subsection 8B:
- Clarify the breadth of recommender systems covered.
- Include 'ensuring independent audits of the function of recommender systems' as an example of a reasonable step.
- Include requirements that providers make all decision making regarding recommender systems must consider children's best interests as a primary consideration.

For subsection 14 & 15:

- Include requirements around *adequately* responding to user reports.
- Ensure that data is made available regarding enforcement against end-users and violative content as part of routine transparency reports, and that this is subject to external scrutiny.
- Include requirements that providers make all decision making regarding enforcement of terms of use must consider children's best interests as a primary consideration.
- Data made available regarding enforcement of terms of use must be subject to independent audits.

For subsection 18A:

- Include additional requirements to report information on a broader range of metrics.
- All data provided to meet requirements regarding systems and enforcement of terms (subsections 18(A)1a-d, etc) is subject to independent oversight and analysis.

Include additional expectations that:

- Service providers take reasonable steps regarding content moderation systems.
- Service providers take reasonable steps regarding advertising approval systems.
- Service providers take reasonable steps regarding advertising management systems.
- Service providers take reasonable steps regarding all systems and elements involved in the operation of their service. This could be an initial expectation, with specific systems and processes listed beneath it.
- Service providers ensure researcher access to public interest data.

For consideration for the terms of reference for the *Online Safety Act* review:

- Introduce an overarching, enforceable duty of care.
- Create a public facing complaints system for BOSE violations.
- Create a presumption that all examples of reasonable steps outlined in the BOSE will be adopted, where they are relevant to a service.
- Increased civil penalties for non-compliance.

Reset.
AUSTRALIA

Contents

Reset.
AUSTRALIA

About Reset.Tech Australia & this submission

We welcome the opportunity to respond to the Department of Infrastructure, Transport, Regional Development, Communications and the Art's consultation regarding proposed amendments to the Online Safety (Basic Online Safety Expectations) Determinations of 2022. Reset.Tech Australia is an Australian policy development and research organisation. We specialise in independent and original research into the social impacts of technology, including social media companies. We are the Australian affiliate of Reset.Tech, a global initiative working to counter digital harms and threats. Our networked structure opens up strong comparative possibilities with other jurisdictions, such as in the EU, where the *Digital Services Act* is in operation, the UK, which has just passed an *Online Safety Act* and Canada, where an Online Safety Bill is under discussion.

We welcome the expansion of the Basic Online Safety Expectations (BOSE) in the proposed amendments. In particular, the inclusion of more systems in the BOSE, the increased focus on transparency and accountability, and the introduction of the children's best interests principle. These proposed amendments help position Australia's online safety regulatory regime towards a systemic, risk-focused model that is more suited to the complexities of the contemporary digital world. The ambitions of these proposals are necessary and welcome, and has the capacity to position Australia as a world leader in comprehensively addressing online harms.

In this submission, we respond to the specific proposals put forward, and outline key areas where the overall amendments could be specifically strengthened. Our proposals are informed by original research undertaken in Australia, as well as comparative policy analysis, where we draw on examples of best practice policy emerging around the world.

We have structured our response to highlight:

1. **Enhancements and improvements regarding systems and processes**, including:
    - Suggestions to improve safety from proposed expectations regarding AI, recommender systems and user-controls, enforcement of terms of service and complaints and reporting systems.
    - Suggestions to improve safety by including expectations around additional systems and elements, including content moderation systems, ad approval systems, ad management systems and a broader catch all for all 'systems and elements'.

2. **Suggestions to increasing accountability and transparency**, including:
    - Suggestions to enhance transparency, including additional oversight and scrutiny of public periodic transparency reports and ensuring researcher access to public interest data.
    - Suggestions to increase accountability, which may require changes to the *Online Safety Act*. For example, introducing an overarching duty of care, challenging assumptions that listed reasonable steps are voluntary and increasing civil penalties. These should be considered in the terms of reference for the upcoming review.

3. **Children's Best Interests**, including suggestions to enhance transparency and accountability for decisions that affect children.

4. **Hate speech**, where we note the need for a more systemic approach to effectively protect communities.

5. **Conclusions & recommendations.**

1. Strengthening and enhancing the focus on systems and processes

The proposed amendments' focus on systems and process of online services is a welcome evolution. As we have documented, a focus on the systems and processes that create and amplify risks enables an 'upstream' approach with the capacity to prevent harms. This approach is a proven and effective way to protect end users.[1]

The case for change

The proposals to place additional expectations on generative AI systems, recommender systems and user controls is extremely welcome. Research has shown that each of these systems can create risks for Australian users. For example:

- *Risks from AI:* Generative AI can multiply the prevalence and virality of illegal and harmful content. For example, generative AI is being used to create photorealistic CSAM[2], and advocates for children's wellbeing, such as the Molly Rose Foundation, have expressed concern that generative models could be used to trigger a damaging wave of harmful content that is accessible to children.[3]

- *Risk from recommender systems:* Content recommender systems often promote content that risks mental or physical injury, such as age-inappropriate violent, extremist content,[4] eating disorder content[5] or misogynistic content.[6] Friend or follower recommender systems often promote connections between children and adult's accounts that create grooming risks for children.[7] The harms of this can be significant.

- *Risks from user-controls:* Current user-controls are often set to default to 'lower levels' of protection for Australian users, including children, and these can create risks.[8] For example, at one stage Meta found that 75% of all 'inappropriate adult-minor contact'—or as it is more commonly called, grooming—on Facebook was a result of their 'People You May Know' friend recommender system.[9] The PYMK feature did/does not function in this way when accounts are private.

- *Risks from failures to enforce terms of use:* Platforms' terms of use, such as community guidelines and terms of service, frequently have robust policies around removing and demoting harmful

[1]Reset.Tech 2022 *The future of digital regulation in Australia: Five policy principles for a safer digital world* https://au.reset.tech/uploads/the-future-of-digital-regulations-in-australia.pdf
[2]Internet Watch Foundation 2023 *How AI is being abused to create child sexual abuse imagery* https://www.iwf.org.uk/media/q4zll2ya/iwf-ai-csam-report_public-oct23v1.pdf
[3]See for example Jim Norton 2023 'AI could 'trigger a damaging new wave' of the extreme content' *Daily Mail* https://www.dailymail.co.uk/news/article-12686383/ai-trigger-damaging-extreme-online-content-schoolgirl-molly-russell-suicide.html
[4]Ralph Housego & Rys Farthing 2022 'Social Grooming' AQ Magazine https://www.jstor.org/stable/27161413
[5]Reset.Tech 2022 *Designing for Disorder* https://au.reset.tech/news/designing-for-disorder-instagram-s-pro-eating-disorder-bubble-in-australia/
[6]Reset.Tech & IDS 2022 *Algorithms as a weapon against women: How YouTube lures boys and young men into the 'Manosphere'* https://au.reset.tech/news/algorithms-as-a-weapon-against-women-how-youtube-lures-boys-and-young-men-into-the-manosphere/
[7]See for example, Australian Child Rights Taskforce 2023 *Letter to the eSafety Commissioner* https://childrightstaskforce.org.au/wp-content/uploads/2023/01/Online-Safety-Codes_-ACRT-letter-to-eSafety.pdf
[8]Fairplay 2022 *Design discriminations on social media platforms* https://fairplayforkids.org/wp-content/uploads/2022/07/design-discriminations.pdf
[9]As made public in *Alexis Spence et al. v. Meta,* U.S. District Court for the Northern District of California, Case No. 3:22-cv-03294 (filed June 6, 2022) p. 11-12, *Growth, Friending + PYMK, and Downstream Integrity Problems.* https://pugetstaffing.filevineapp.com/s/9eb2BZcUfhdTxkxIfV45CJnIivYHhdWcRRuQVwSMz120RVs7ATmxn9r5

content such as pro-eating disorder content or pro-suicide content. However, these are often not enforced and content and content creators known to push harmful content remains available on, and promoted by, online service providers. For example, providers often fail to remove pro-anorexia coaches[10] or pro-anorexia content when they are reported, and indeed keep promoting this content to Australian users.[11]

- *Risks from failures of complaints and reporting systems:* The current system for making complaints and reporting illegal and harmful content places the burden on those who have been harmed to file reports.[12] However, there is often a lack of clarity for those who report content or abuse around what happens, and worse, often a lack of action and remedy.[13]

Expanding BOSE to cover these systems is a welcome and necessary step to reduce risks for Australian users. To adequately protect Australian users, two key amendments to these proposals are necessary. Firstly, minor amendments to the proposals regarding generative AI capabilities, recommender systems and user-controls are needed. Secondly, additional systems and elements need to be explicitly named and expectations for safety extended to cover all relevant systems and elements. We discuss these below.

[10]Suku Sukunesan 2021 ''Anorexia coach': sexual predators online are targeting teens wanting to lose weight. Platforms are looking the other way' *The Conversation* https://theconversation.com/anorexia-coach-sexual-predators-online-are-targeting-teens-wanting-to-lose-weight-platforms-are-looking-the-other-way-162938
[11]Reset.Tech Australia 2022 *Designing for Disorder* https://au.reset.tech/news/designing-for-disorder-instagram-s-pro-eating-disorder-bubble-in-australia/
[12]Michael Salter, Delanie Woodlock, Tim Wong 2023 'The sexual politics of technology industry responses to online child sexual exploitation during COVID-19: "This pernicious elitism"' *Child Abuse & Neglect* https://doi.org/10.1016/j.chiabu.2023.106559
[13]House of Commons Science and Technology Committee (UK) 2019 I*mpact of social media and screen-use on young people's health* https://publications.parliament.uk/pa/cm201719/cmselect/cmsctech/822/822.pdf, Pg 54

A. Amendments to proposals regarding Generative AI capabilities, recommender systems, user controls, enforcement of terms of use and complaints and reporting systems.

Strengthening expectations around Generative AI capabilities (subsection 8A)

- We note that the proposals regarding generative AI are relatively modest, and that broader proposals regarding AI are under consideration. We would support a broader response to AI regulation, that includes different rules for different levels of risk, across all AI models. Our response to these proposals regarding Generative AI in the BOSE are limited to these proposals, and we look forward to broader discussions regarding AI regulation in the near future.

- **Include 'retraining generative artificial intelligence that has been trained on illegal material' as an example of a reasonable step.** The proposed additional expectations note that providers should improve training quality data by removing unlawful and harmful materials. This may address training models going forward, but it does not address existing LLMs and MfMs that have already been trained to include this data. 8A(3)(C) needs to be revised to include reference to existing AI systems that may have been—and frequently have been—trained using data sets that include class 1A and 1B unlawful materials.[14] These need to be audited and retrained. We appreciate that this is not a small ask for fully operational AI services, so limit this suggestion to training out class 1A and 1B materials rather than 'harmful' materials. Without this step, any future ban on training models on CSAM or pro-terror material will be unsuccessful; existing models and models built from them will continue to be able to produce generative CSAM or pro-terror content.

- **Include 'ensuring that training materials for generative artificial intelligence capabilities and models comply with the APPs' as an example of a reasonable step.** The proposed additional expectations do not include consideration of the privacy of individuals in the development of training materials. While we understand that the *Online Safety Act* focuses on online harm, we would argue that privacy harms are also cognisable by regulators and indeed very real.[15] Privacy risks create online harms, and failing to enhance the BOSE in a way that aligns with the ambitions of the *Privacy Act Review*[16] creates gaps in protections. A new clause should be added to section 8A(3) to ensure that training materials for generative artificial intelligence capabilities and models comply with the Australian Privacy Principles, especially regarding informed consent.

- **Include 'ensuring independent audits of the function of AI systems' as an example of a reasonable step.** Data provided by online services for the purposes of transparency needs to be subject to independent oversight and auditing.

[14]See for example Davey Alba & Rachel Metz 2023 'Large AI Dataset Has Over 1,000 Child Abuse Images, Researchers Find' *Bloomberg* https://www.bloomberg.com/news/articles/2023-12-20/large-ai-dataset-has-over-1-000-child-abuse-images-researchers-find?leadSource=uverify%20wall
[15]See for example, Danielle Citron and Daniel Solove 2022 'Privacy Harms' *Boston University Law Review* 793 http://dx.doi.org/10.2139/ssrn.3782222
[16]Attorney General's Department 2022 *Privacy Act Review Report* https://www.ag.gov.au/sites/default/files/2023-02/privacy-act-review-report_0.pdf.

Strengthening expectations around Recommender systems (subsection 8B)

- **Clarify the breadth of recommender systems covered**. Recommender systems can recommend a wide range of 'content', all of which can create risks. For the avoidance of doubt, it should be made clear that section 8B addresses the full breadth of recommender systems, including content recommender systems, as well as friend or followers recommender systems, targeted advertising systems, search recommender systems and other recommender systems. It would be an oversight if the amendments were to protect end-users from illegal content delivered in social media posts but not the recommendations of illegal accounts, illegal advertising or illegal information.

  We note that there is emerging Australian precedent for this broad interpretation. For example, the *Privacy Act Review* proposed a comprehensive definition of targeting that includes a description of tailoring services. It defines targeting as the *"capture the collection, use or disclosure of information which relates to an individual including personal information, deidentified information, and unidentified information (internet history/tracking etc.) for tailoring services, content, information, advertisements or offers provided to or withheld from an individual (either on their own, or as a member of some group or class)."*[17] (emphasis added). The BOSE requirements around safety in recommender systems should likewise cover systems involved in recommending content, information, advertisements or offers.

  This would not necessarily require any changes to section 8B, but could be addressed via explanatory material used to interpret the Act.

- **Include 'ensuring independent audits of the function of recommender systems' as an example of a reasonable step.** Data provided by online services for the purposes of transparency needs to be subject to independent oversight and auditing.

Strengthening expectations around user-controls (subsection 6(5&6))

A word of caution is necessary on the capacity of user-controls to drive systemic changes. User controls are one part of the system of ensuring safety in the digital environment. An upstream focus that requires digital services to be safe, and places the responsibility of ensuring this on platform developers, is both more effective and appropriate. This is in keeping with existing norms around effective ways to reduce industrial hazards. The hierarchy of hazard controls —a globally used framework, including in Australian occupational safety standards—outlines that the most effective interventions to keep hazards from causing harm focus primarily on eliminating hazards in the first instance, and training users to protect themselves or providing protective tools last.[18] In the digital world, for example, user control tools such as 'private settings' or 'safe searches' must be considered the last line of defence because every instance of individual failure, either from the tool or the user, leaves users exposed to risk. There is evidence that this approach has limited effect.[19] Moreover, we do not consider it appropriate to place responsibility on individual users (or for younger users, their parents), to keep themselves safe from systemic harms created by negligent design or dangerous systems.

[17]Proposal 20.1(B), Attorney General's Department 2022 *Privacy Act Review Report* https://www.ag.gov.au/sites/default/files/2023-02/privacy-act-review-report_0.pdf. This proposed definition was agreed in principle by the Government in Australian Government 2023 *Government Response \ Privacy Act Review Report* https://www.ag.gov.au/sites/default/files/2023-09/government-response-privacy-act-review-report.PDF
[18]See for example WorkSafe Victoria 2021 *The Hierarchy of Controls* www.worksafe.vic.gov.au/hierarchy-control
[19]Mariya Stoilova, Monica Bulger & Sonia Livingstone 2024 'Do parental control tools fulfil family expectations for child protection? A rapid evidence review of the contexts and outcomes of use' *Journal of Children and Media*, https://doi.org/10.1080/17482798.2023.2265512

- **Requiring providers to take reasonable steps to avoid deploying dark patterns on end-users.** Deploying dark patterns against end-users is the opposite of empowering users, and to adequately enable users to protect their own best interests and provide autonomy, the BOSE should make clear that services should not deploy them. An additional subsection after 6(5) could require the provider of a service to take reasonable steps to avoid deploying dark patterns, either on purpose or in effect, on end-users.

  Dark patterns are deliberate design features that 'nudge' users away from actions that align with their best interests and toward actions that are in the platform's interest,[20] and are a type of consumer manipulation routinely deployed in the digital environment.[21] The phrase 'Dark Pattern' was originally coined by Brignull to describe a type of user interface that "has been carefully crafted to trick users into doing things", in ways that involve "a solid understanding of human psychology, and they do not have the user's interests in mind".[22] The European Commission describes dark patterns as "practices that materially distort or impair, either on purpose or in effect, the ability of recipients of the service to make autonomous and informed choices or decisions. Those practices can be used to persuade the recipients of the service to engage in unwanted behaviours or into undesired decisions which have negative consequences for them".[23]

  Australian end-users are routinely deceived by dark patterns. Research undertaken by Reset.Tech exploring the privacy policies and procedures used by ten apps popular with young people in Australia noted that eight out of ten deployed dark patterns regarding data and privacy policies, all of which had the capacity to actively "trick" young people into agreeing to sharing more personal data than is necessary.[24] Recent EU research undertaken by Reset.Tech demonstrates that these dark patterns are still prevalent in Very Large Online Platforms.[25] Dark patterns are frequently deployed in children's apps too, which encourage users to share more information than is necessary.[26]

  There is global precedent for addressing dark patterns in online harm frameworks. See, for example:

  - The EU's *Digital Services Act* (DSA) quite clearly prohibits dark patterns. It states that platforms are "*prohibited from deceiving or nudging recipients of the service and from distorting or impairing the autonomy, decision-making, or choice of the recipients of the service via the structure, design or functionalities of an online interface or a part thereof."*[27]

  - In the USA, the Federal Trade Commission (FTC) increasingly considers dark patterns as a unique type of consumer deception, meaning dark patterns fall under their jurisdiction without needing to demonstrate 'further harm' to consumers.[28] Deception is inherently harmful

[20]Arunesh Mathur *et al.* 2019 'Dark Patterns at Scale: Findings from a Crawl of 11K Shopping Websites' *Proceedings of the ACM on Human-Computer Interaction November*, pp. 81. https://dl.acm.org/doi/10.1145/3359183
[21]FTC 2022 *Bringing Dark Patterns to Light | Staff Report* https://www.ftc.gov/reports/bringing-dark-patterns-light
[22]Although Brignall has more recently shifted to using the broader language of deceptive design. See Harry Brignull 2010 'What is Deceptive Design' https://www.deceptive.design/
[23]Recital 67, EU 2022 *Digital Services Act*
https://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=CELEX:32022R2065
[24]Reset.Tech 2021 *Did We Really Consent to This?*
https://au.reset.tech/uploads/l01_resettechaustralia_policymemo_t_c_report_final-july.pdf.
[25]Reset Tech 2023 *Risks to Minors* https://www.reset.tech/resources/risktominors/
[26]Jenny Radesky, Alexis Hiniker A & Caroline McLaren 2022 'Prevalence and Characteristics of Manipulative Design in Mobile Applications Used by Children' *JAMA Netw Open.* 2022;5(6):e2217641. doi:10.1001/jamanetworkopen.2022.17641
[27]Recital 67, EU 2022 *Digital Services Act*
https://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=CELEX:32022R2065
[28]FTC 2022 *Bringing Dark Patterns to Light | Staff Report* https://www.ftc.gov/reports/bringing-dark-patterns-light. . Under Section 5 of the FTC Act, "*A representation, omission, or practice is deceptive under Section 5 if it is likely to mislead consumers acting reasonably under the circumstances and is material to consumers—that is, it would likely affect the consumer's conduct or decision with regard to a product or service.*"

because it deprives consumers of the ability to make free and informed choices about products and services.

- **Include 'allowing users to turn off recommender systems' as an example of a reasonable step**. Suggestions around reasonable steps (6(6)) should include an additional step of providing users with the choice to turn off recommender systems. Providing users with the ability to 'turn off' recommender systems is potentially the only way users as individuals can protect themselves from being recommended harmful or illegal content. It is more effective than allowing individuals to 'mute' particular users or unfollow particular accounts. It ensures that they would, for example, only see content their selected accounts post in reverse chronological order. Given the research that suggests inflammatory or otherwise harmful content is routinely prioritised by content recommender systems,[29] this would provide users with a meaningful choice around reducing their risk of encountering it. There is strong precedent for this, and Australia is becoming a global outlier:

    - The DSA provides European users with the right to turn off content recommender systems on Very Large Online Platforms. Article 27 states: "*for recommender systems that determine the relative order of information presented to recipients of the service, providers of online platforms providers of online platforms shall also make available a functionality that allows the recipient of the service to select and to modify at any time their preferred option. That functionality shall be directly and easily accessible from the specific section of the online platform's online interface where the information is being prioritised."*[30] This includes content recommender systems, friend recommender systems, search recommender systems and advertising recommender systems.

    - In the USA, 20 percent of the population have the right to opt-out of, or turn off, ad recommender systems. Various states—red and blue—including California, Colorado, Texas and Montana, have passed laws providing end-users the right to opt-out of targeted advertising systems.[31]

  Again, this could help harmonise the requirements under the BOSE with emerging requirements in a revised *Privacy Act*, where the right to opt-out (or turn off) targeting is being discussed.

- **Include 'ensuring independent audits of the user control systems' as an example of a reasonable step.** Data provided by online services for the purposes of transparency needs to be subject to independent oversight and auditing.

Strengthening expectations around enforcement of terms of use (subsections 14& 15)

- **Include requirements around *adequately* responding to user reports.** Subsections 15(2) requires online service providers to make mechanisms available to report violative content and issues, while 14(3-5) will create new and much needed requirements for services to respond to user reports in a timely fashion. These are necessary steps, but miss an obvious step; online service providers must be required to respond adequately to user reports in a manner consistent with their terms of use. Currently, for example, services frequently enable users to report content and provide automated updates to users about the outcome of their report, but the outcome of these reports often does not align with their terms of use. We appreciate that this is indeed the aim of these subsections, but it would be useful to make this a clearer expectation; 14(3)C could read 'ensure the response is

[29]See for example Reset Tech 2023 *X and Risks to Minors* https://www.reset.tech/resources/risktominors/x-and-risks-to-minors/
[30]EU 2022 *Digital Services Act* https://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=CELEX:32022R2065
[31]Via the *California Consumer Privacy Act, the Colorado Privacy Act, the Texas Data Privacy and Security Act, and the Montana Consumer Data Privacy Act*

consistent with stated terms of use'. This would be enhanced by requirements regarding the accuracy of content moderation systems, as described below, creating a belt and braces approach.

- **Ensure that data is made available regarding enforcement against end-users and violative content as part of routine transparency reports, and that this is subject to external scrutiny.** While the proposals, rightly, emphasise the need for the provision of information and data about compliance regarding the enforcement of terms of use, the expectations must be clear and this data must be subject to independent oversight. Amendments to section 18(1)A requires online service providers to report on 'metrics on the prevalence of harms, reports and complaints, and the service's responsiveness', but this must include metrics regarding the consistency of the response with the platform's terms of use. This would require an independent assessment or audit of a sample of responses.

- **Data made available regarding enforcement of terms of use must be subject to independent audits.** Data provided by online services for the purposes of transparency needs to be subject to independent oversight and auditing in general. There is precedent for requiring independent audits of platform's compliance, with the DSA requiring independent audits as one of their key five transparency measures.[32]

[32]See for example, Reset.Tech Australia 2024 *Accountability, the Online Safety Act and the Basic Online Safety Expectations: Can safety standards be enforceable?*
https://au.reset.tech/news/briefing-can-safety-standards-be-enforceable/

B. Amendments to include more systems and elements in the BOSE.

The proposals to create additional expectations for the systems described above are welcome. They should be more comprehensive in nature. Other systems are also innately involved in creating risks for end-users online, and require comparable additional expectations. Specifically;

Including expectations around content moderation systems

- Content moderation systems are an integral part of ensuring safety on any service that hosts user-generated content. They are the systems that ensure online service providers are able to detect, classify and respond to content that is hosted on their platform which breaches guidelines. For example, a platform may detect illegal content and remove it, or it may be made aware of harmful content that breaches their guidelines when a user reports it, and may classify it as violating guidelines and then respond to it by removing it, applying a label or sensitivity filter to this content, or demoting or 'shadow' banning it.[33] Content moderation systems rely heavily on algorithms—but distinctly different algorithms from recommender systems—although they often also have human moderators in the loop. These algorithms also warrant scrutiny. As DP-REG note,[34] algorithms play two crucial roles in content moderation systems:
    - Detection: They are used to automate the detection of content that might violate a provider's community guidelines, and;
    - Classifiers: Classifying content as violative or not, which automates the provider's response to each piece of content classified as violative.

The case for including content moderation systems

Research undertaken by Reset.Tech in the EU demonstrates that content moderation systems routinely fail to minimise the presence and promotion of content that is harmful or illegal. For example, we tested the efficacy of content moderation systems on TikTok, Instagram and Twitter in minimising the presence of pro-suicide and pro-self harm material, and pro-eating disorder material, all of which had been independently verified as harmful to users by a clinical psychologist and which violate platform guidelines.[35] All of these systems failed to safely and adequately moderate this content in two ways:

1. Detection: Firstly, online service providers did not adequately detect this content. We followed a sample of harmful content (as confirmed by a psychologist) that violated providers' guidelines and analysed how much of this content was detected and responded to over a week. Alarmingly, little harmful content was detected and responded to by numerous platforms content moderation systems (either by removing or labelling) (see figure 1).
2. Classifiers: Secondly, even when online service providers were made aware of the content, they failed to classify it and therefore respond appropriately. We reported a sample of harmful material and found that one week later, the vast majority of this content was not removed (see figure 2). We also monitored for labelling or demotion, and found no evidence of either of those responses from the platform's content moderation system. Similar evaluations have uncovered these failings also occurring with Australian content moderation systems when it comes to risks associated with

[33]In the UK, these are part of 'reporting and redress' systems
[34]Digital Platforms Regulators Forum 2023 *Literature summary: Harms and risks of algorithms* https://dp-reg.gov.au/sites/default/files/documents/2023-11/Working%20paper%201%20Literature%20Summary%20-%20Harms%20and%20risks%20of%20algorithms.pdf
[35]Reset Tech 2023 *Risks to Minors* https://www.reset.tech/resources/risktominors/

electoral integrity.[36] It appears that despite having reporting systems to allow users to report harmful or illegal material, and channels bringing this content to platforms' attention, moderation systems still fail to adequately respond. (see figure 2).

	% of pro-suicide and pro-self harm content removed without reporting	% of pro-eating disorder content removed without reporting
TikTok	0% (n=79)	5.61% (n=107)
Instagram	0% (n=119)	0% (n=125)
X	6.25% (n=96)	2.70% (n=111)

Figure 1: The amount of harmful, violative content removed by online service provider before reporting, by content nature.

	% of pro-suicide and pro-self harm content removed after reporting	% of pro-eating disorder content removed after reporting
TikTok	1.27% (n=79)	6.27% (n=107)
Instagram	29.41% (n=119)	10.40% (n=125)
X	7.08% (n=96)	3.78% (n=111)

Figure 2: The amount of harmful, violative content removed by online service provider after reporting, by content nature.

Without requirements to take reasonable steps to ensure content moderation systems work to ensure user-safety, services will fail to respond to illegal or harmful content when they become aware of it.

Content moderation systems and processes also include the use of 'trusted flaggers', including third party fact checkers. We note that there has been recent controversy about the level of transparency around trusted flaggers, and how online service providers engage with them.[37] Greater oversight and transparency over the content moderation system would help promote user trust and ensure safety.

- **Additional expectations that providers take reasonable steps regarding content moderation systems should be included**. This could be modelled on Sections 8A and 8B to note that if a service moderates its content, either using proactive detection or by enabling user-reporting, the provider of the service should take reasonable steps to consider end-user safety and incorporate safety measures in the design, implementation and maintenance of the moderation systems on the service. These must be designed in a way that minimises the extent to which illegal or harmful content is distributed on the platform. Reasonable steps could include; ensuring that risks

[36]Reset.Tech 2023 *How do platforms respond to user-reports of electoral process misinformation? An experimental evaluation from the lead-up to Australia's referendum*
https://au.reset.tech/uploads/Electoral-Process-Misinformation-September.pdf
[37]See for example IPA 2023 *The Arbiters Of Truth – Analysis Of Fact Checking Organisations During The 2023 Voice Referendum*
https://ipa.org.au/publications-ipa/research-papers/the-arbiters-of-truth-analysis-of-fact-checking-organisations-during-the-2023-voice-referendum

assessments of the impact are undertaken; introducing independent audits and evaluations of the efficacy of these systems, and; ensuring end-users can make complaints about the functioning of content moderation systems.

Including expectations around advertising approval systems (managing the content of ads)

- Ad approval systems are used by services to determine which ads can run on their service, and which are denied in accordance with their advertising policies. These systems are often automated, with technology used to detect ads that potentially breach their guidelines. Others claim to be human moderated or use a combination of the two ('human in the loop').[38]

The case for including advertising approval systems

Research undertaken by Reset.Tech in Australia has demonstrated how ads containing harmful content are easily 'approved' by ad approval systems, and how they fail to protect end-users from harmful or illegal material. For example, we experimented to see if we could get ads purporting to contain 'Spicy cocktails recipes using only what you can find in your 'rents (parents') liquor cabinet', or ads to help girls 'find your gentleman now (money emoji)', or to win prizes by gambling. These ads were approved on Instagram to a target audience of 13-17 year olds.[39] This experiment was repeated internationally, with colleagues finding ads for 'skittles parties' (drug parties) and 'ana tips' (pro-anorexia tips) were all likewise approved.[40]

Further suggesting systemic failings, in 2023 we tested the ad approval systems on Facebook, TikTok and X to see if we could get approval to run ads containing electoral process misinformation around the Voice referendum, such as ads suggesting the referendum was being held on Nov 31st (a non-existent date), or that the referendum had been cancelled, or was voluntary or postal. The vast majority of ads were approved, at a rate of between 70-100%, depending on the platform.

The European Commission likewise notes that "*online advertising can contribute to significant risks, ranging from advertisements that are themselves illegal content, to contributing to financial incentives for the publication or amplification of illegal or otherwise harmful content and activities online*".[41]

- **Additional expectations that providers take reasonable steps regarding advertising approval systems should be included**. This could be modelled on Sections 8A and 8B to note that if a service uses an advertising approval system, the provider of the service should take reasonable steps to consider end-user safety and incorporate safety measures in the design, implementation and maintenance of advertising approval systems on the service. These must be designed in a way that minimises the extent to which illegal or harmful content is distributed on the platform. Reasonable steps could include; ensuring that risks assessments are undertaken; introducing independent audits and evaluations of the efficacy of these systems, and; ensuring end-users can make complaints about the functioning of ad approval systems.

[38]For a description of TikTok's, Facebook's and X's ad approval systems see Reset.Tech 2023 *How do platforms handle electoral misinformation in paid-for advertising? An experimental evaluation using the Voice referendum* https://au.reset.tech/news/report-misinformation-in-paid-for-advertising/
[39]Reset.Tech 2021 *Profiling Children for Advertising* http://au.reset.tech/news/profiling-children-for-advertising-facebooks-monetisation-of-young-peoples-personal-data/
[40]Tech Transparency Project 2021 *Facebook's Repeat Fail on Harmful Teen Ads* https://www.techtransparencyproject.org/articles/facebooks-repeat-fail-harmful-teen-ads
[41]Recital 68, EU 2022 *Digital Services Act* https://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=CELEX:32022R2065

- There is precedent for including requirements around advertising approval systems. The EU's DSA includes a number of requirements around advertising approval systems, including requirements around risk assessments, and, where advertising is targeted, the ability to turn off targeting.[42]

Including expectations around advertising management systems (managing the targeting of ads)

- Ad management systems manage the process of targeting users with ads, which involves the collection and analysis of data, the creation of a 'profile' of a user and the process of subsequently matching an ad with the appropriate profile. Ad management systems include Meta Ads Manager, Google Ad Campaign Manager and ads.tiktok.com for instance.

The case for including advertising management systems

Research has highlighted some worrying features within ad management systems, from the use of dark patterns in their design,[43] to their ability to target ads at under 18 year olds in countries where it is illegal to do so,[44] to the ability to target unsafe ads at vulnerable demographics.[45] In this latter experiment, Meta's Ad Manager system was found to allow advertisers to target children they had profiled in 'vulnerable' categories such as 13-17 year olds interested in 'Gambling', 'Alcohol', 'Extreme weight loss', 'E cigarettes', etc. More alarmingly, ads were approved to run to each of these vulnerable profiles containing content that posed unique harms, such as ads calling for beach body ready looks to children interested in extreme weight loss, or ads containing recipes for cocktails made from booze stolen from your parents' liquor cabinet to children interested in alcohol.[46] Further, research has shown how extensive 'vulnerable' advertising profiles are across Australia, with profiles being created that allow for the targeting of 'heavy gamblers', 'problematic alcohol users', families in 'financial distress' and children and young people based on their geolocation.[47] None of these profiles are covered by existing regulations, and there are no restrictions on how they are used.

The European Commission notes how targeted advertising can create risk: "*when recipients of the service are presented with advertisements based on targeting techniques optimised to match their interests and potentially appeal to their vulnerabilities, this can have particularly serious negative effects. In certain cases, manipulative techniques can negatively impact entire groups and amplify societal harms, for example by contributing to disinformation campaigns or by discriminating*

[42]See for example, recital 68 & 69, EU 2022 *Digital Services Act*
https://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=CELEX:32022R2065
[43]Reset. Tech Global 2023 *Ad management systems and targeting minors*
https://www.reset.tech/resources/risktominors/ad-manager-systems-and-targeting-minors/
[44]Reset. Tech Global 2023 *Ad management systems and targeting minors*
https://www.reset.tech/resources/risktominors/ad-manager-systems-and-targeting-minors/
[45]Reset.Tech Australia 2020 *Profiling Children for Advertising*
https://au.reset.tech/news/profiling-children-for-advertising-facebooks-monetisation-of-young-peoples-personal-data/
[46] This research was subsequently repeated in the US, see Tech Transparency Project 2021 *Pills, Cocktails, and Anorexia: Facebook Allows Harmful Ads to Target Teens*
https://www.techtransparencyproject.org/articles/pills-cocktails-and-anorexia-facebook-allows-harmful-ads-target-teens. It and drew condemnation from the US Senate Commerce Committee See transcript at Rev.com 2021 *Facebook Head of Safety Testimony on Mental Health Effects: Full Senate Hearing Transcript*
[47]Reset.Tech 2023 *Australians for Sale: Targeted Advertising, Data Brokering and Consumer Manipulation*
https://au.reset.tech/news/coming-soon-australians-for-sale-report/

against certain groups. Online platforms are particularly sensitive environments for such practices and they present a higher societal risk."[48]

Online service providers often do not demonstrate willingness nor transparency around improved ad management systems, which suggests a strong need for transparency, independent scrutiny and accountability regarding these systems. In preparation for regulatory requirements in the UK's *Age Appropriate Design Code* and the EU's DSA, Facebook announced that it was stopping the practice of targeting under 18 year olds in July 2021.[49] In September 2021, Meta's Head of Safety, Antigone Davis, told the US Senate Commerce Committee that Meta had changed, and now "have very limited advertising to young people. You can only actually now target a young person based on their gender, age, or location."[50] Subsequent research from Reset.Tech Australia & Fairplay in November 2021 suggested that Meta's statements were misleading, and children were still being targeted based on profiling derived from their online activity.[51] Meta responded by claiming the research was flawed, ignoring the main claims of the study by stating they did not use "data from our advertisers' and partners' websites and apps to personalise ads to people under 18."[52] Under pressure from the US Senate Commerce Committee on Commerce, Science, and Transportation to clarify, Adam Mosseri, Head of Instagram, when giving evidence in December 2021 made it more clear that in fact "the (ad delivery) system also uses activity that teens use within the app."[53] It turns out, the research was not so flawed, and the changes Meta instigated were cosmetic and did not address the core issue of profiling teens in unsafe ways that targeted their vulnerabilities. In February 2023, Meta claimed to have implemented the changes necessary under the DSA and UK's *Age Appropriate Design Code*, claiming that ""engagement on our apps — like following certain Instagram posts or Facebook pages — won't inform the types of ads they see...Age and location will be the only information about a teen that we'll use to show them ads.'"[54] This is eight months after they made the initial claim and two testimonies in front of the US Senate Commerce Committee later, and only in the face of strong, enforceable regulatory action.

- **Additional expectations that providers take reasonable steps regarding advertising management systems should be included**. This could be modelled on Sections 8A and 8B to note that if a service delivered targeted advertising using an ad management system, the provider of the service should take reasonable steps to consider end-user safety and incorporate safety measures in the design, implementation and maintenance of the ad management system. These must be designed in a way that minimises the vulnerability of end users. Reasonable steps could include; ensuring that risk assessments of the impact of targeting are undertaken; introducing independent audits and evaluations of the efficacy of these systems; ensuring end-users can make complaints about the functioning of ad management systems; providing information to end-users about how they are being profiled; and crucially; the option to turn off or opt-out of receiving targeted ads (but

[48]Recital 69, EU 2022 *Digital Services Act*
https://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=CELEX:32022R2065
[49]Meta 2021 *Giving Young People a Safer, More Private Experience on Instagram*
https://about.fb.com/news/2021/07/instagram-safe-and-private-for-young-people/
[50]See transcript at Rev.com 2021 *Facebook Head of Safety Testimony on Mental Health Effects: Full Senate Hearing Transcript*
https://www.rev.com/blog/transcripts/facebook-head-of-safety-testimony-on-mental-health-effects-full-senate-hearing-transcript
[51]Reset.Tech Australia & Fairplay 2021 *Facebook still misusing young people's data*
https://au.reset.tech/news/facebook-caught-red-handed-harvesting-teens-data/
[52]WTSP 2021 'Facebook using teens data' *WTSP News*
http://www.wtsp.com/article/tech/facebook-report-teen-data/67-457034e6-f374-40e9-9239-49fc9e5a89
[53]See transcript on C-SPAN, C-SPAN 2021 *Senate Hearing on Online Protections for Children*
https://www.c-span.org/video/?516470-1/senate-hearing-online-protections-children
[54]Meta 2023 *Continuing to Create Age-Appropriate Ad Experiences for Teens*
https://about.fb.com/news/2023/01/age-appropriate-ads-for-teens/

still allowing advertising to be served to them). To be clear, this does not mean end-users would have an ad free experience, just a *targeted* ad free experience.

- There is precedent for including requirements around advertising approval systems. Including:
    - The EU's DSA includes a number of requirements around advertising management systems, including requirements around risk assessments, the requirement to provide users with information about profiling, the right not to be profiled based on special category date ('sensitive' data) and the right to object, i.e. the right to turn off targeting for advertising.[55]
    - Twenty percent of the American population enjoys the right to turn off targeting for advertising, and instead receive contextual advertising. Various states (both red and blue), including California, Colorado, Texas and Montana, have passed regulation ensuring end-users can decline targeting for advertising.[56] The EU and US are large markets, it is both technically and financially feasible to offer these protections in large markets. Australian end-users could also enjoy these protections.

- There is currently widespread support among the Australian public[57] and civil society[58] for the right to turn off the receipt of targeted advertising.

Including expectations around all systems and elements that give rise to risks

- To set expectations and help create a culture of compliance, it is important to specifically designate a number of systems that need to be considered by services under the BOSE (including AI capabilities, recommender systems, user control systems, content moderation systems, ad approval systems and ad manager systems). However given the pace of technological change, and the complexity of each individual service, a list of named systems could never be comprehensive, warranting the inclusion of a catch-all provision to encourage a future-proofed and truly systemic approach.

- **Additional expectations that providers take reasonable steps regarding all systems and elements involved in the operation of their service should be considered**. Expectations that providers take reasonable steps to safeguard end-users regarding all systems and elements that contribute to risks should be included in the BOSE. This could require identification of risk-creating systems and elements, and requirements to consider end-user safety in the design,

[55]See for example, recital 68 & 69, EU 2022 *Digital Services Act* https://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=CELEX:32022R2065.
Regarding the adequacy of information provided to end-users, the DSA requires that, "*providers of online platforms should therefore be required to ensure that the recipients of the service have certain individualised information necessary for them to understand when and on whose behalf the advertisement is presented. They should ensure that the information is salient, including through standardised visual or audio marks, clearly identifiable and unambiguous for the average recipient of the service, and should be adapted to the nature of the individual service's online interface. In addition, recipients of the service should have information directly accessible from the online interface where the advertisement is presented, on the main parameters used for determining that a specific advertisement is presented to them, providing meaningful explanations of the logic used to that end, including when this is based on profiling."* Regarding the right to object, EU end-users have "*the right to object, automated individual decision-making, including profiling, and specifically the need to obtain consent of the data subject prior to the processing of personal data for targeted advertising*" reinforced by the DSA and the EU's *General Data Protection Regulation* (GDPR). (See EU 2016 *General Data Protection Regulation* https://eur-lex.europa.eu/eli/reg/2016/679/oj)
[56] Via the *California Consumer Privacy Act, the Colorado Privacy Act, the Texas Data Privacy and Security Act, and the Montana Consumer Data Privacy Act*
[57]Reset.Tech Australia 2023 Intrusive and Unhelpful https://au.reset.tech/news/report-intrusive-and-unhelpful-targeted-advertising-in-australia/
[58]Reset.Tech Australia 2023 *Targeted advertising: Are we going far enough?* https://au.reset.tech/news/briefing-targeted-advertising-and-profiling-in-the-privacy-act-review-are-we-going-far-enough/

implementation and maintenance of these systems. Examples of reasonable steps could include risk assessments, independent evaluations and enabling complaint mechanisms.

- There is precedent for both listing specific systems and noting that 'all systems and elements' are subject to safety requirements, emerging from (see figure 3):
    - The EU's DSA. The regulation states that very large online service providers *"should focus on the systems or other elements that may contribute to the risks, including all the algorithmic systems that may be relevant, in particular their recommender systems and advertising systems... (and) assess whether their terms and conditions and the enforcement thereof are appropriate... their content moderation processes, technical tools and allocated resources"*[59] (emphasis added). The DSA does not limit requirements to only those systems and elements listed.
    - The UK's *Online Safety Act 2023* ('UK OSA').[60] The UK OSA places specific duties of care on services regarding:
        - Illegal content
        - For services likely to be accessed by children, child safety
        - Rights to freedom of expression and privacy
        - Reporting and redress
        - Record-keeping and review duties

      These duties are not limited to particular systems or processes, rather, they place broad obligations across the operations of a service.

- There is also strong public support for covering all systems and processes. In January 2024, Reset.Tech commissioned YouGov to poll 1,005 Australian adults. We found overwhelming support for including expectations regarding more systems—such as advertising systems and content moderation systems—and all systems in general (see figure 4).

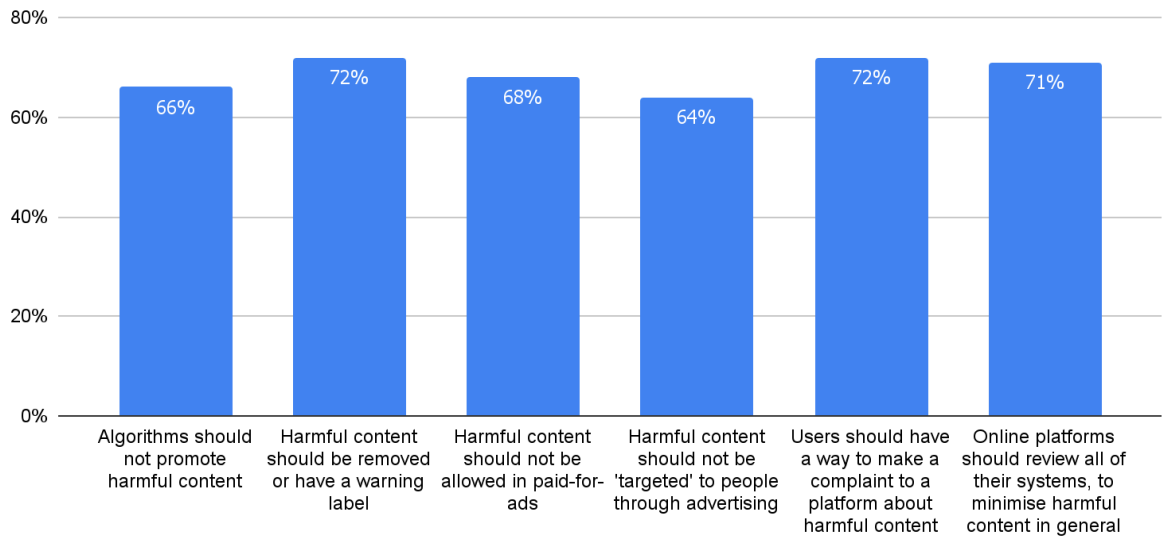

Figure 4: responses to the question 'which of the following do you think online safety regulations should require?' n=1,005.

[59] Recital 84, EU 2022 *Digital Services Act* https://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=CELEX:32022R2065
[60] UK 2023 *Online Safety Act* https://www.legislation.gov.uk/ukpga/2023/50/enacted

Systems 'designated' in the DSA as subject to risk assessment criteria for Very Large Online Platforms	Systems 'designated' in the UK OSA as requiring measures to ensure duties of care are met across Platforms	Systems 'designated' in the proposed BOSE as being subject to expectations regarding reasonable steps
Recital 84 outlines that services should "focus on the systems or other elements that may contribute to the risks", and lists a number of examples. Other systems and elements specifically listed across the legislation include: 1. Recommender systems (Recital 70, 84, 88 articles 27, 34 & 35, 38) 2. 'Safety by design' settings for minors (Recital 71) 3. Dark patterns and design of interfaces (articles 25 & 35) 4. Advertising systems (Recital 68, 69, 84, 95 and articles 26, 34 & 35, 39) 5. Content moderation systems (articles 34 & 35, 96) 6. Notice action and complaint mechanisms (recital 89, 96) 7. Trusted flagger systems (articles 22, 35) 8. Terms and conditions (articles 14, 34 & 35)	The duties of care laid out in the Act "apply across all areas of a service, including the way it is designed, operated and used as well as content present on the service," and lists the following areas as requiring measures: 1. Regulatory compliance and risk management arrangements; 2. Design of functionalities, algorithms and other features; 3. Policies on terms of use; 4. Policies on user access to the service or to particular content present on the service, including blocking users from accessing the service or particular content; 5. Content moderation, including taking down content; 6. Functionalities allowing users to control the content they encounter 7. User support measures; 8. Staff policies and practices.	1. Generative AI capabilities 2. Recommender systems 3. User controls 4. 'Safety by design' settings for minors (via best interests proposal in subsection 6(2)(A)) 5. Enforcement of terms of use (14(1A)) 6. Complaints & reporting systems (14(3)) We note that some aspects around staff practices covered by the UK's OSA may be addressed by proposals to amend paragraph 6(3)(f), to add in a suggested example that services assessing whether business decisions will have a significant adverse impact on the ability of end-users to use the service in a safe manner. Further, elements of the DSA's requirements around terms and conditions regarding understandability are being explored in the *Privacy Act Review*.

Figure 3: Systems and elements 'designated' in various safety legislation and proposals, by jurisdiction.

Summary of recommendations from Section 1

For subsection 8A:
- Include 'retraining generative artificial intelligence that has been trained on illegal material' as an example of a reasonable step.
- Include 'ensuring that training materials for generative artificial intelligence capabilities and models comply with the APPs' as an example of a reasonable step.
- Include 'ensuring independent audits of the function of AI systems' as an example of a reasonable step.

For subsection 8B:
- Clarify the breadth of recommender systems covered.
- Include 'ensuring independent audits of the function of recommender systems' as an example of a reasonable step.

For subsection 6(5&6):
- Include an additional requirement that providers take reasonable steps to avoid deploying dark patterns on end-users.
- Include 'allowing users to turn off recommender systems' as an example of a reasonable step.
- Include 'ensuring independent audits of the user control systems' as an example of a reasonable step.

For subsection 14 & 15:
- Include requirements around *adequately* responding to user reports.
- Ensure that data is made available regarding enforcement against end-users and violative content as part of routine transparency reports, and that this is subject to external scrutiny.
- Data made available regarding enforcement of terms of use must be subject to independent audits.

Including additional expectations that:
- Service providers take reasonable steps regarding content moderation systems.
- Service providers take reasonable steps regarding advertising approval systems.
- Service providers take reasonable steps regarding advertising management systems.
- Service providers take reasonable steps regarding all systems and elements involved in the operation of their service. This could be an initial expectation, with specific systems and processes listed beneath it.

2. Improving accountability and transparency requirements

We have so far outlined a range of additional systems, and additional expectations and examples of reasonable steps, that could be required in order to create a comprehensive framework of protections for Australian users. While this could help create a comprehensive framework, this framework is only effective if it is enforced and online services implement changes to their systems and processes. To incentivise this, the proposals for amending the BOSE need to have enhanced requirements around accountability and transparency.

The case for change

The BOSE provides the Office of the eSafety Commissioner with the powers to issue non-periodic reporting notices and periodic reporting notices, that compels online service providers to report on measures taken to implement the BOSE. The 'theory of change' here is that compliance reports create adequate transparency, and that this transparency in turn creates 'reputational' risks that incentivise improvements in services and creates accountability. However, each step of this logic is flawed or simply not working. Specifically:

- The BOSE reporting notices scheme does not create transparency. These powers do not create adequate transparency because they do not sufficiently compel services to disclose information. To date, only one set of non-periodic reporting notices has been issued—regarding CSAM—to seven online services.[61] This resulted in two findings of 'non compliance' with the reporting requirements; or a finding that 29% of services issued with a compliance notice failed to provide adequate details.[62]. That is, in the first and only test of the current scheme, almost a third of companies were found to be not transparent enough.

  As a result, X (formerly Twitter)—who were one of the two services issued with a non-compliance order— was fined $610,500 AUD.[63] However, paying this fine would still be cheaper than the staff required to implement any meaningful CSAM measures to write about. Assuming that a safety team capable of implementing even a single CSAM measure would require a Senior Engineering Manager (average salary band of $530K USD or $800K AUD per year),[64] the lack of incentive becomes obvious. From a business perspective, it is roughly 25% cheaper to pay this fine every year than to employ even one staff member capable of implementing any improvements in their systems. This imbalance is not unique to X and most other large online services pay staff more.[65]

  For these businesses, returning a compliance report can be considered 'a goodwill gesture' that they can afford not to offer if they so choose. Where services can afford to opt-out of compliance reports, they are a broken transparency measure. The hallmarks of effective transparency do not include 'marking your own homework' nor 'being able to choose whether to respond or not'.

[61]Office of the eSafety Commissioner 2022 *Basic Online Safety Expectations: Summary of industry responses to the first mandatory transparency notices*
https://www.esafety.gov.au/sites/default/files/2022-12/BOSE%20transparency%20report%20Dec%202022.pdf
[62]It is worth reiterating that these notices of non compliance were not findings that these services had failed to deliver inadequate safety standards, rather they had even failed the more prosaic test of providing enough information about their safety standards.
[63]Georgie Hewson 2023 'Australia's eSafety commission fines Elon Musk's X $610,500 for failing to meet anti-child-abuse standards' *ABC*
https://www.abc.net.au/news/2023-10-16/social-media-x-fined-over-gaps-in-child-abuse-prevention/102980590
[64]Salary estimates from Levels.fy 2023 *Twitter Salaries* https://www.levels.fyi/companies/twitter/salaries
[65]Business Insider 2023 *Big Tech salaries revealed*
https://www.businessinsider.com/big-tech-salaries-what-you-make-google-apple-amazon-meta-ibm

- The BOSE reporting notice scheme does not create accountability. The Office of the eSafety Commissioner does not have powers to compel online services to implement any changes or improvements to their online services. The issues here are obvious; as long as an online service meets reporting requests and describes their wholly inadequate safety measures in detail, they will still be in compliance with the BOSE.

  The belief that transparency alone leads to accountability is outdated and disproven. It rests on a theory of change that assumes transparency creates significant reputational risk that incentives companies to drive up safety standards. Aside from lack of effective transparency described above, this logic is:

  - *Flawed*: The vast majority of contemporary online services are already regarded as so unsafe and violative that they are held in no regard by the public. For example, globally social media companies sit at the bottom of IPSO Mori's Trustworthiness measure, and any rises in trust are associated only with government regulation.[66] This is also true in Australia, the Edelman Trust Barometer found that social media companies were the least trusted industry, with trust actively declining 7 points annually. [67] Working with YouGov, we polled 1,005 Australian adults regarding the reputation of social media companies when it came to user safety. Only 11% of the population felt these companies had a good or very good reputation, with 45% saying their reputation was mixed and 25% saying their reputation was poor or very poor (see figure 5). Simply put, this is an industry that effectively has no reputation left to risk, so 'reputational risks' will not be an effective driver of change.

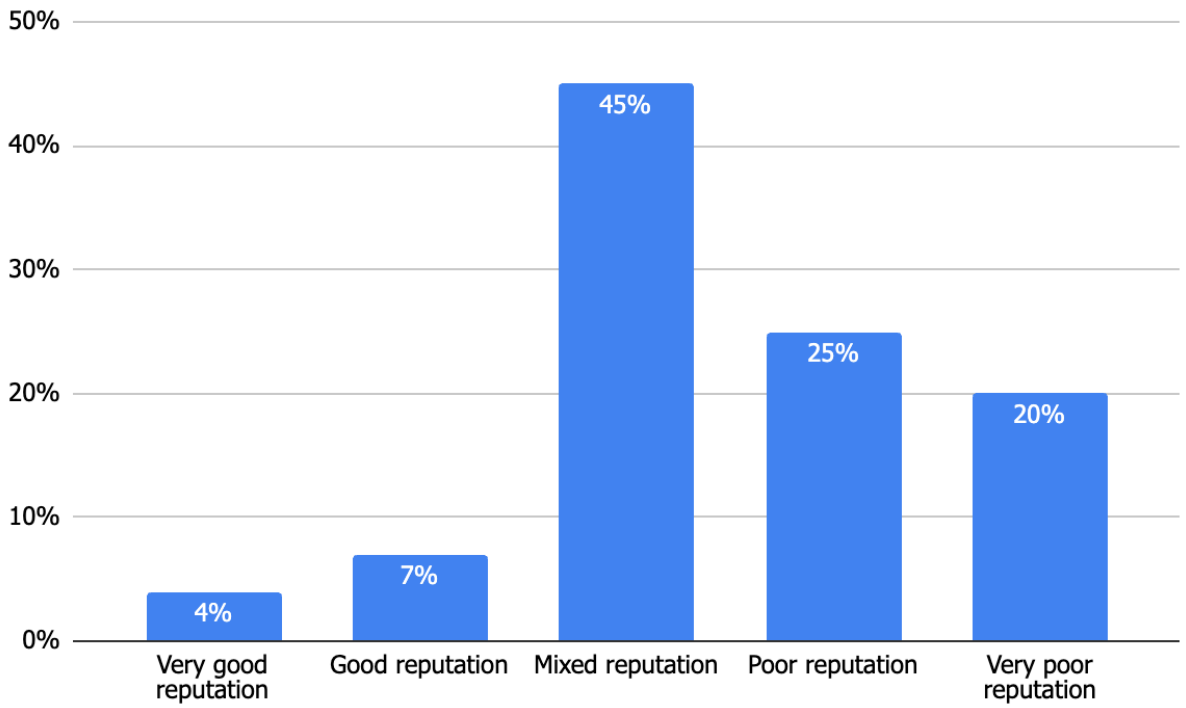

Figure 5: Responses to the question 'Do you think social media companies have a good or poor reputation when it comes to users' safety?', n=1,005.

[66]IPSO Mori 2023 *Trust in Social Media* https://www.ipsos.com/en/trust/trust-social-media.
[67]Edelman 2023 *Edelman Ttrust Barometer 2023: Australia Report* https://www.edelman.com.au/sites/g/files/aatuss381/files/2023-02/2023%20Edelman%20Trust%20Barometer%20Report%20-%20AUS%2002-2023.pdf

- *Demonstrably untrue:* For example, the Facebook Files and revelations leaked by whistleblower Frances Haugen have failed to incentivise a 'safety first' culture at Facebook. They may have temporarily removed $50bn off Meta's valuation[68], but despite this massive financial impact, they have not led to systemic improvements. Likewise, the reputational harms associated with removing trust and safety features at X (formerly Twitter) have been extensive, including, in Australia, admonishment from Digi's independent complaint sub committee[69] and removal from the voluntary Code and findings of non-compliance with the *Online Safety Act* and fines from the Office of the eSafety Commissioner,[70] but this has not led to investments in safety teams or features.

Enhancing transparency (subsection 18A)

Enhancing the transparency requirements in the BOSE—specifically the proposals to amend section 18A to require public periodic transparency reports—is a welcome and necessary step to reduce risks for Australian users. However, the transparency report produced under Digi's *Australian Code of Practice on Disinformation and Misinformation* provides a cautionary tale around the need for clarity and oversight of transparency reports.[71] To avoid the pitfalls evidenced under the Digi Code and adequately ensure transparency around safety standards, a number of amendments to the proposals are necessary.

- **Include additional requirements to report information on a broader range of metrics in transparency reports.** This will allow greater oversight, and comparison between providers. The transparency reports required under the DSA provide a good template for this, and require additional information to the BOSE proposals such as metrics around:

  - Volume and response to regulator orders and other legal requirements:
    - Number of 'take down' orders issued by regulators, median, average and max time to respond to these, and final response;
    - Number of notices received regarding IP, defamation, Privacy and Illegal content notifications received from Australian end-users; median, average and max time to respond to these, and final response;
    - Notices processed using automated means;
    - Data about number of out of court settlement made;

  - Content moderation metrics, including impact on Australian businesses and pages:
    - Number of organic content measures (i.e. how much content they proactively detected) that violated their community guidelines; by violation type; amount detected by automated means; amount detected by human moderators; median, average and max time to detect these, and final response;
      - ➢ Number of organic business entity measures (i.e. how many Australian business accounts were removed and restricted as a result of organic content moderation)

[68]Billy Perrigo 2021 'How Facebook Forced a Reckoning by Shutting Down the Team That Put People Ahead of Profits' *Time Magazine*
https://time.com/6104899/facebook-reckoning-frances-haugen/
[69]Independent Complaints Sub-Committee 2023 *Statements Attributable To The Independent Complaints Sub-Committee*
https://digi.org.au/complaint-by-reset-australia-against-x-f-k-a-twitter-upheld-by-australian-code-of-practice-on-disinformation-and-misinformation-independent-complaints-sub-committee/
[70]Jordan Baker 2023 ''Heinous crimes': Twitter fined $600,000 over child safety failures' *SMH*
https://www.smh.com.au/national/heinous-crimes-twitter-fined-600-000-over-child-safety-failures-20231015-p5ecda.html
[71]Reset.Tech Australia 2024 *Functioning or Failing?* (forthcoming)

        - ➢ Number of organic entity measures (i.e. how many Australian pages or products were removed and restricted as a result of organic content moderation)
    - Number of user-reported content measures (i.e. how much content was reported to the platform by Australian end-users) that violated their community guidelines; by violation type; median, average and max time to detect these; response; number of challenges against response; final outcome
        - ➢ Number of business entity measures following user-reporting (i.e. how many Australian business accounts were removed and restricted after user-reporting)
        - ➢ Number of entity measures following user reporting (i.e. how many Australian pages or products were removed and restricted after user-reporting)
    - Number of 'trusted-flagger' content measures (i.e. how much content was acted on by a platform as a result of Australian fact-checkers or trusted flaggers); amount reported to platform; by violation type; amount subsequently detected by automated means; median, average and max time to detect these; response; number of challenges against response; final outcome;
    - Indicators of accuracy and error rates for automated review processes; both for organic detection and following user reporting;
    - Human resources dedicated to content moderation, including information about; number located within Australia; number dedicated to Australian content or addressing reports from Australian end-users; qualifications and training; support; volume of work (i.e. how much content per hour are they required to review); language addressed;

- Measures against misuse such as number of Australian end-users:
    - Number of accounts suspended or deleted and why; number of challenges and final outcome

- **All data provided to meet requirements regarding systems and enforcement of terms is subject to independent oversight and analysis.** Voluntary transparency is not the same thing as meaningful transparency. Where online service providers can 'pick and choose' what to measure and how to report it, they will remain able to 'mark their own homework'. To create the conditions for meaningful transparency, all data supplied by online services in transparency reports needs to be subject to evidential evaluation. There is precedent for this. For example, under delegated regulation under the EU's DSA, Very Large Online Platforms are required to undertake an independent audit of compliance with regulation.[72]

- **Include specific requirements for transparency around listed systems and processes.** As described in section 1 above, this includes amending:

    - Subsection 8A: To include 'ensuring independent audits of the function of AI systems' as an example of a reasonable step.
    - Subsection 8B: To include 'ensuring independent audits of the function of recommender systems' as an example of a reasonable step.
    - Subsection 6(6): To include 'ensuring independent audits of the user control systems' as an example of a reasonable step.
    - Include 'ensuring independent audits of' any other systems, e.g. content moderation systems, advertising approval systems, advertising management systems and AI systems.
    - Subsections 14 & 15: Data needs to be made available regarding enforcement against end-users and violative content as part of routine transparency reports, and that this is subject to external scrutiny.

- **Include a requirement for researcher access to public interest data.** Beyond independent oversight regarding requirements 18(A)1a-d, enabling independent analysis of emerging harms will ensure the safety of Australian users as the digital threat landscape evolves. To enable this, we appreciate that changes to the *Online Safety Act* would be needed to ensure researcher access to

[72]European Commission 2023 *Commission adopts rules on independent audits under the Digital Services Act* http://digital-strategy.ec.europa.eu/en/news/commission-adopts-rules-independent-audits-under-digital-services-act

public interest data regarding safety. This would enable ongoing assessment of safety threats by academia and civil society. Again, the DSA provides one model for what these provisions and protections could look like.[73] The DSA model places an obligation on services to provide regulators 'within a reasonable period specified in that request, access to data that are necessary to monitor and assess compliance with this Regulation'.[74] Notably, large online platforms operating in the EU already have in-house systems and processes developed to enable this kind of researcher access.

Enhancing accountability

To increase basic online safety standards, the BOSE needs to be able to hold services to account where they breach basic expectations. We appreciate that enhancing accountability may require changes to the *Online Safety Act* itself, but in anticipation of the independent statutory review of the Act, we wanted to briefly describe the nature of possible improvements here to demonstrate that enforcement is possible and desirable. The Terms of Reference for the upcoming *Online Safety Act* should consider a number of steps that could improve accountability, including but not limited to:

- **Introduce an overarching, enforceable duty of care.** The BOSE must be enforceable, and the Office of the eSafety Commissioner must have powers to hold services to account where they fail to implement reasonable steps. Introducing an overarching Duty of Care into Australia's online safety regulation can help to ensure this. The Carnegie Foundation described what an enforceable duty of care might look like in Australian online safety regulations, describing it as having four key aspects:

  1. *the overarching obligation to exercise care in relation to user harm;*
  2. *risk assessment process;*
  3. *establishment of mitigating measures; and*
  4. *ongoing assessment of the effectiveness of the measures.*[75]

  This approach is consistent with the UK OSA and the EU DSA, which would reduce regulatory burden on online services, while ensuring that Australian end-users enjoy the same levels of protection as those in Europe.

- **Create a public facing complaints system for BOSE violations.** Australia's *Online Safety Act* is globally unique in that it creates a much feted public complaints system that end-users who have been harmed by content addressed in the Act—such child bullying or image-based abuse—can seek redress in a way that is enforced by regulators. Australia could build on this and introduce a groundbreaking public complaints facility for end-users who are affected by violations of the BOSE to seek remedy. This would require additional resourcing to effectively operate.

- **Create a presumption that all examples of reasonable steps outlined in the BOSE will be adopted,** except where they are not relevant to a service (for example, the service does not have AI capabilities or any user-controls). This may require changes to the *Online Safety Act*.

- **Increased civil penalties for non-compliance.** The Office of the eSafety Commissioner needs to have sufficient powers to compel online service providers to meet the BOSE. Comparable European

[73]See Article 40, Digital Services Act *Data access and scrutiny* https://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=CELEX:32022R2065
[74]See Article 40, *Digital Services Act* https://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=CELEX:32022R2065 For a comprehensive analysis, see Mathias Vermeulen 2022 'Researcher Access to Platform Data: European Developments' *Journal of Online Trust and Safety* https://tsjournal.org/index.php/jots/article/view/84/31
[75]Carnegie UK 2022 *Submission to the House Select Committee on Social Media and Online Safety* available at https://www.aph.gov.au/Parliamentary_Business/Committees/House/Former_Committees/Social_Media_and_Online_Safety/SocialMediaandSafety/Submissions

and British legislation sets penalties for non-compliance around online safety at 10% of global turnover,[76] and other Australian regulations have these powers.[77]

There is strong public support for increasing accountability and transparency when it comes to user safety (see figure 6). We polled 1,005 Australians about online regulations, and found strong support for accountability (phrased as enforcement) and transparency (phrased as oversight).

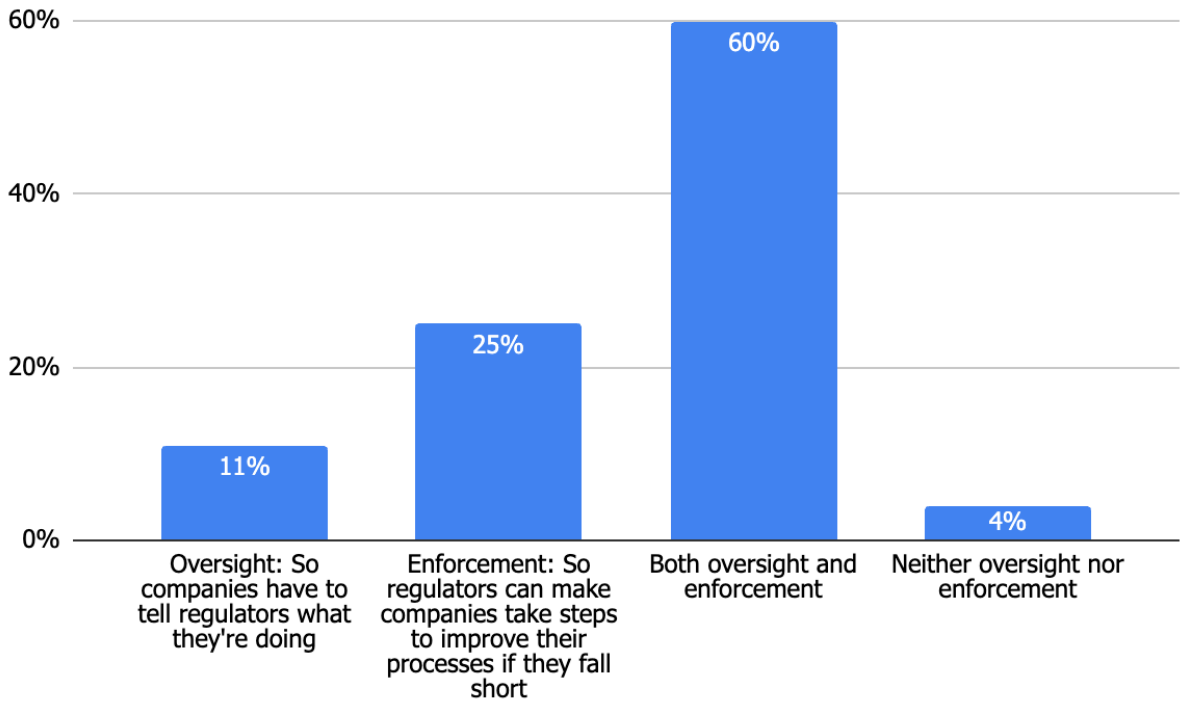

Figure 6: responses to the question 'thinking about online safety regulations for companies, which of these do you think should be required?' n=1,005.

[76]Such as the EU's DSA and the *UK's Digital Markets Bill*
[77]Such as the ACCC for franchising violations (see ACCC nd Fines and penalties https://www.accc.gov.au/business/compliance-and-enforcement/fines-and-penalties) and ASIC for violations of ASIC administered legislation, albeit capped at $782.5million (see ASIC 2023 Fines and Penalties https://asic.gov.au/about-asic/asic-investigations-and-enforcement/fines-and-penalties/)

Summary of recommendations from Section 2

For the *Basic Online Safety Determinations*:

- Include additional requirements to report information on a broader range of metrics in transparency reports.
- All data provided to meet requirements regarding systems and enforcement of terms (Subsections 18(A)1a-d etc) is subject to independent oversight and analysis.
- Include a requirement for researcher access to public interest data.
- Include specific requirements for transparency around listed systems and processes, as described in section 1.

For consideration for the terms of reference for the *Online Safety Act* review*:*

- Introduce an overarching, enforceable duty of care.
- Create a public facing complaints system for BOSE violations.
- Create a presumption that all examples of reasonable steps outlined in the BOSE will be adopted, where they are relevant to a service.
- Increased civil penalties for non-compliance.

3. Ensuring the best interests of the child becomes a primary consideration

Reset.Tech Australia, alongside the Australian Child Rights Taskforce and its members, have been advocating for the introduction of the 'children's best interests' principle into Australian digital regulation for a number of years now. We are extremely pleased to see proposals that requirements that service providers take reasonable steps to ensure that children's best interests are a primary consideration be introduced into the BOSE. The use of the overarching 'best interests' principle will help to harmonise regulations emerging in the privacy space, as well as help to meet Australia's commitments under the Convention on the Rights of the Child.[78]

The case for change

Decisions made by services are frequently made in ways that prioritise profits or other business considerations over young people's safety and wellbeing. For example:

- Internal research leaked from Meta as part of the Facebook Files consistently demonstrates instances where company profits and KPIs were prioritised over children's best interests. For example, Meta knew that Instagram was toxic for teen girls in particular[79] but rather than addressing the issue was instead making plans to launch an Instagram for Kids product, to ensure a conveyor belt of users that are young girls.[80] Further, it showed that Meta chose not to proactively default teens accounts to private accounts (which is safer) because they prioritised interactions on the platforms (which is more profitable).[81] More recent whistleblowers suggest that the problematic de-prioritisation of children's best interests continues within the company. For example, late last year a former engineering director at Meta noted that 13% of young Instagram users aged 13-15 years old have received unwanted sexual advances on the platform.[82] Simple steps were not implemented to curb this, and instead, the platform deployed dark-patterns and made it more difficult for younger users to report this abuse to reduce the cost of managing their user-reporting system.[83] Similarly, recent data made available through the Office of the eSafety Commissioner's transparency requirements highlights how X cut safety staff now operates without Australian trust and safety staff[84] *and* still allows Australian teens to join the platform. The safety of younger users is often not even a consideration, let alone a primary consideration.

[78]Article 3 of the Convention states *"in all actions concerning children, whether undertaken by public or private social welfare institutions, courts of law, administrative authorities or legislative bodies, the best interests of the child shall be a primary consideration.* UN General Assembly 1989 *Convention on the Rights of the Child,* https://www.ohchr.org/en/instruments-mechanisms/instruments/convention-rights-child
[79]*Georgia Wells, Jeff Horwitz* and *Deepa Seetharaman* 2021 'Facebook knew it was toxic for teen girls *WSJ* https://www.wsj.com/articles/facebook-knows-instagram-is-toxic-for-teen-girls-company-documents-show-11631620739
[80]Adam Mosseri 2021 "*Pausing" Instagram for kids* https://about.instagram.com/blog/announcements/pausing-instagram-kids
[81]Meta nd *Should we default teens to privacy settings?* https://www.documentcloud.org/documents/23322914-copy-of-should-we-default-teens-into-privacy-settings__sanitized_opt
[82]Associated Press 2023 'Ex-Meta employee says his warnings of Instagram's harm to teens were ignore' *The Guardian* https://www.theguardian.com/technology/2023/nov/07/meta-facebook-employee-congress-testimony-instagram-child-harm-social-media
[83]Tech Oversight Project 2023 *Statement on Meta's Cover-Up, Implicated by Whistleblower Testimony* https://techoversight.org/2023/11/07/meta-coverup/
[84]Evelyn Manfield 2020 'Online safety regulator lashes X, formerly Twitter, over failure to police hate' *ABC* https://www.abc.net.au/news/2024-01-11/online-safety-x-twitter-failure-online-hate/103307246

- In drafting Australia's online safety codes for class 1A & 1B material, the authors (industry representative groups) chose to set lower levels of protection for Australia's children than those enjoyed by children in other jurisdictions.[85] Put another way, many companies which already have the technical capacity and processes in place to offer European children high levels of safety chose to offer Australian children lower levels of safety. This clearly violates the intent of the Online Safety Act and—paradoxically—the BOSE determinations that instigated these codes. The BOSE currently states that "if a service or a component of a service (such as an online app or game) is targeted at, or being used by, children (services must ensure) that the default privacy and safety settings of the children's service are robust and set to the most restrictive level,"[86] and presumably rests on the *Online Safety Act's* definition of a child which is "an individual who has not reached 18 years". However, this clear requirement *and* the requirements of international regulators was ignored by industry, who chose to interpret this in a way that disregarded children's best interests when they set privacy settings requirements to only protect those up until age 16. This demonstrates that where companies are involved in decision making processes regarding safety, they do not necessarily prioritise children's best interests.

Further, the UN Committee on the Rights of the Child, in their *General comment number 25 on children's rights in relation to the digital environment*, states: *"states parties should ensure that, in all actions regarding the provision, regulation, design, management and use of the digital environment, the best interests of every child is a primary consideration."*[87] This proposal is a welcome step towards advancing children's rights, and should see online service providers share the same responsibility for prioritising children as the Government.

Improving requirements regarding children's best interests being a primary consideration (subsection 6(2A))

- **Including an additional requirement that 'best interests assessments' are undertaken and published,** potentially as part of ongoing transparency measures. Where companies are making decisions regarding the design and operation of their service, requiring children's best interests to be a primary consideration is a powerful way to move towards child-centred digital design. However, online services have always been able to hold children's best interests as a primary concern, but have failed to consistently do so. Requiring online service providers to publish documentation and assessments demonstrating their decision-making processes, and showing that children's best interests are routinely held as a primary consideration, is key to ensuring transparency and ultimately creating the conditions for accountability. Children's 'best interests' assessments are already standard practice in the UK, under the *Age Appropriate Design Code*,[88] and may be required under proposals put forward in the *Privacy Act review*.[89] Harmonising

[85]Reset.Tech Australia 2022 *How outdated approaches to regulation harm children and young people* https://au.reset.tech/news/how-outdated-approaches-to-regulation-harm-children-and-young-people-and-why-australia-urgently-needs-to-pivot/
[86]*Online Safety (Basic Online Safety Expectations) Determination 2022* Subsection 6(C)(3)
[87]UN Committee on the Rights of the Child 2021 *General comment No. 25 (2021) on children's rights in relation to the digital environment*. https://www.ohchr.org/en/documents/general-comments-and-recommendations/general-comment-no-25-2021-childrens-rights-relation, para 12
[88] ICO 2023 *Children's Code Best Interest Framework* https://ico.org.uk/for-organisations/uk-gdpr-guidance-and-resources/childrens-information/childrens-code-guidance-and-resources/how-to-use-our-guidance-for-standard-one-best-interests-of-the-child/best-interests-framework/
[89]Attorney General's Department 2023 *Government Response: Privacy Act Review Report* https://www.ag.gov.au/sites/default/files/2023-09/government-response-privacy-act-review-report.PDF

requirements in the BOSE to ensure 'best interests' assessments could reduce regulatory burden across the industry. They are also strongly supported by civil society.[90]

- As described above, to lead to positive improvements on services for children, these assessments must be subject to independent oversight and regulators must be able to take action where assessments demonstrate that online services have failed to adequately regard children's best interests as a primary consideration.

- **Consult with children and young people around the development of elements of the BOSE and 'best interests' requirements that affect them.** Involving young people in these policy processes produces two distinct advantages. Firstly, it helps to advance their right to participate. Young people have the right to participate in decision making processes that affect them, including decisions made about the governance of the digital world. As the *General Comment No. 25 (2021) on Children's Rights in Relation to the Digital Environment* makes clear, "when developing legislation, policies, programmes, services and training on children's rights in relation to the digital environment, States parties should involve all children, listen to their needs and give due weight to their views."[91] That is, the development and implementation of the BOSE in themselves provide an opportunity to advance children and young people's right to participate. Secondly, meaningfully involving children and young people can help to create better policies and practices and is one way of advancing justice in design.[92] Children and young people have intelligent and articulate insights to share, and their knowledge should position them as key stakeholders in the development and implementation of the BOSE. In appendix 1, we have included some key insights from young people regarding 'best interests' and targeting, and the idea of best interests impact assessments. Although developed to interrogate the application of the best interests principle in a different policy domain, we hope it demonstrates the nature of the insights that engaging with young people can generate. We would be delighted to facilitate workshops or consultations directly with young people to facilitate this.

[90]Reset.Tech Australia, Australian Child Rights Taskforce, ChildFund Australia 2024 *Best interests & targeting* https://au.reset.tech/news/best-interests-and-targeting-implementing-the-privacy-act-review-to-advance-children-s-rights/
[91]UN Committee on the Rights of the Child 2021 *General comment No. 25 (2021) on children's rights in relation to the digital environment*. https://www.ohchr.org/en/documents/general-comments-and-recommendations/general-comment-no-25-2021-childrens-rights-relation.
[92]See for example Sasha Costanza-Chock 2021 *Design Justice Community-Led Practices to Build the Worlds We Need* MIT Press New York

Ensuring requirements regarding children's best interests are prioritised in decision making regarding systems and elements

- **All systems and elements listed—including Generative AI, Recommender Systems User-Controls, Terms of Use, and Content Moderation Systems, Ad Approval Systems, Ad Management Systems, and all systems and elements—need to be required to be designed, function and reviewed to ensure they work in children's best interests.** This is to ensure that alongside overall decision making, all systems and elements must function in children's best interests (and decisions regarding systems be made with children's best interests as a primary consideration). This could be achieved by including a reasonable step within each subsection requiring 'decision making regarding (this system) must consider children's best interests as a primary consideration'.

Summary of recommendations from Section 3

For subsection 6(2A):

- Including an additional requirement that 'best interests assessments' are to be undertaken and published.
- Consult with children and young people around the development of elements of the BOSE and 'best interests' requirements that affect them.

For subsection 6, 8A, 8B, 14 & 15 & 18 (and any additional systems and elements included);

- All systems and elements listed—including Generative AI, Recommender Systems User-Controls, Terms of Use, and Content Moderation Systems, Ad Approval Systems, Ad Management Systems, and all systems and elements—need to be required to be designed, function and reviewed to ensure they work in children's best interests.

4. Hate speech

Australia's digital regulatory framework has consistently overlooked community and societal risks,[93] and the proposals to address collective hate speech within the BOSE are a welcome step towards remedying one important element of this. We note, however, that this is a modest, content-focussed proposal that leaves other societal risks caused by systems and processes unaddressed. For example, the following risks remain unaddressed by the BOSE:

- *Discrimination caused by algorithms.* The algorithmic risk of discrimination is real and the harms are frequent and often serious. For example:
    - The Facebook Files released a trove of evidence documenting harms to minorities as a result of their systems and processes. For example, internal research showed that vulnerable users who were Black "had a much higher concentration of violent content … and sexual content *that they did not want to see in their Feeds"*, than vulnerable White users.[94] Meta also ran a controlled experiment to see if users saw equally harmful content, controlling for variables associated with race. Without controls, African Americans were more likely to encounter harmful content, but once controlled they were still more likely to encounter this content, albeit less so.[95] A comprehensive review undertaken by Meta appears to confirm racial bias was a systemic problem within the platform, stating that 'it's virtually guaranteed that (Meta's) major systems do show systemic biases based on the race of the affected user.'[96]
    - Research into the statistical functions underpinning friend recommender algorithms —such as Twitter's 'Who To Follow'— has shown that they consistently afford minorities less visibility, which can affect users' online networks.[97]
    - Algorithmic promotion on TikTok has been accused of racial bias, and systemically shadow-banning content made by Black creators even when consumers follow those creators.[98]

- *The failure of content moderation systems to adequately protect minorities.* These are frequent and can have catastrophic consequences. For example, failures in algorithmic recommendations and content moderation on Facebook saw the platform fueling intolerance and violence against Rohingya Muslims in the 2016 genocide.[99] Content moderation systems consistently underperform for non-English speakers, as the first round of transparency reports required under the DSA

[93]Reset.Tech Australia 2022 *The future of digital regulation* https://au.reset.tech/uploads/the-future-of-digital-regulations-in-australia.pdf
[94]As made public in Alexis, Kathleen, And Jeffrey Spence Vs Meta Platforms, Inc 2022 Case 3:22-cv-03294 United States District Court Northern District Of California San Francisco Division (FBP 10/22, "Longitude Integrity Harm Project, Facebook paper)
[95] As made public in Alexis, Kathleen, And Jeffrey Spence Vs Meta Platforms, Inc 2022 Case 3:22-cv-03294 United States District Court Northern District Of California San Francisco Division ('The Synthetic Parity Method of ML Equivalency' FBP 39/16 Facebook paper)
[96]As made public in Alexis, Kathleen, And Jeffrey Spence Vs Meta Platforms, Inc 2022 Case 3:22-cv-03294 United States District Court Northern District Of California San Francisco Division ('Comprehensive Study on disparate product impacts by race' FBP 27/17, Facebook paper)
[97]Lisette Espín-Noboa, Claudia Wagner, Marcus Strohmaier and Fariba Karimi 2022 'Inequality and inequity in network-based ranking and recommendation algorithm' *Scientific Rep*orts doi:10.1038/s41598-022-05434-1
[98]Megan McCluskey 2020 'These TikTok Creators Say They're Still Being Suppressed for Posting Black Lives Matter Content' *Time Magazine* https://time.com/5863350/tiktok-black-creators
[99]United Nations Human Rights Council 2018 *Report of the detailed findings of the Independent International Fact-Finding Mission on Myanmar A/HRC/39/CRP.2*

highlight, online service providers often do not recruit human moderators to be 'in the loop' and evaluate content that is reported in non dominant languages.[100]

- *Advertising management systems that discriminate*. Many ad management systems have been found to produce discriminatory outcomes for minorities. For example:
    - Facebook's job advertisement system has been shown to promote job ads differentially between male and female users, even when controlling for qualifications and background.[101]
    - Facebook settled a complaint with the Department of Justice over the use of algorithms that discriminated on protected characteristics, including race and national origin, in selectively delivering advertisements for housing.[102]
    - Younger people are more often targeted for gambling advertising than other age groups,[103] disparately heightening the risks of economic harms (lost money) and psychological harms (addiction) from gambling for the young *vis a vis* the old.

Addressing these risks adequately requires revisions to the *Online Safety Act* itself to ensure an expansion of the harms considered under the core expectations of the BOSE. This has been tried and tested elsewhere. For example, the EU's DSA requires platforms to address four broad categories of risk:

1. Risks from the distribution of illegal content such as CSAM, illegal hate speech and illegal services.[104]
2. Actual or foreseeable impact of the service on the exercise of fundamental rights,[105] which more broadly addresses societal risks such as algorithmic discrimination or harms from content moderation systems.
3. Actual or foreseeable negative effects on democratic processes, civic discourse and electoral processes and security,[106] which are still not addressed within Australia's digital regulatory framework.
4. Actual or foreseeable negative effects on the protection of public health, minors and serious negative consequences to a person's physical and mental well-being, or gender-based violence.[107]

All systems and elements of platforms are required to be assessed against these risks, in order to provide comprehensive protection to communities. A similar, more systemic approach to addressing harms via *Online Services Act* is necessary to effectively protect minorities, such as women and CALD communities. We look forward to engaging in these discussions in the broader review of the *Online Safety Act*. Regardless, the more modest step of embracing a content-based approach—in this case tackling hate speech—is a welcome first step.

We defer to the advice of subject-matter specialists, including the Human Rights Law Centre, AMAN and the Carnegie Trust with respect to proposals for subsection 6.4 regarding a non-exhaustive definition of hate speech.[108]

---

[100]Global Witness 2023 *How Big Tech platforms are neglecting their non-English language users* https://www.globalwitness.org/en/campaigns/digital-threats/how-big-tech-platforms-are-neglecting-their-non-english-language-users/
[101]Basileal Imana, Aleksandra Korolova, and John Heidemann 2021 'Auditing for Discrimination in Algorithms Delivering Job Ads' *Proceedings of The Web Conference 2021*doi.org/10.48550/arXiv.2104.04502
[102]Lauren Feiner 2022 'DOJ settles lawsuit with Facebook over allegedly discriminatory housing advertising' *CNBC* https://www.cnbc.com/2022/06/21/doj-settles-with-facebook-over-allegedly-discriminatory-housing-ads.html
[103]Morgane Guillou-Landreat, Karine Gallopel-Morvan, Delphine Lever, Delphine Le Goff and Jean-Yves Le Reste 2021 'Gambling Marketing Strategies and the Internet: What Do We Know? A Systematic Review' *Frontiers in Psychiatry* doi.org/10.3389/fpsyt.2021.583817
[104]EU 2022 *Digital Services Act* https://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=CELEX:32022R2065, rec 80
[105]EU 2022 *Digital Services Act* https://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=CELEX:32022R2065, rec 81
[106]EU 2022 *Digital Services Act* https://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=CELEX:32022R2065, rec 82
[107]EU 2022 *Digital Services Act* https://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=CELEX:32022R2065, rec 83
[108]See for example, Mohamad Abdalla, Mustafa Ally & Rita Jabri-Markwell 2021 'Dehumanisation of 'Outgroups' on Facebook and Twitter: towards a framework for assessing online hate organisations and actors' *SN Soc Sc*i 1, 238. https://doi.org/10.1007/s43545-021-00240-4, and Carnegie Trust 2022 *Ad hoc advice from Carnegie UK to United Nations Special Rapporteur on Minority Issues*

We note that proposals 6(3)(i) places obligations on online service providers to detect and address hate speech that breaches a service's terms of use. However, some services available in Australia have no policies against hate speech in their terms of use (such as Gab which has policies only regarding illegal speech),[109] while others have 'thin' policies that may not cover the substance intended by subsection 6.4 (such as Gettr that prohibits the use of racial slurs and endorsement of violence or segregation).[110] If the intent of the BOSE is to place obligations on all online service providers to address hate speech, the proposals might require amendments to ensure that all services contain policies regarding hate speech.

https://carnegieuktrust.org.uk/publications/ad-hoc-advice-from-carnegie-uk-to-united-nations-special-rapporteur-on-minority-issues-concerning-guidelines-on-combatting-hate-speech-targeting-minorities-in-social-media/

[109]Gab 2019 *Gab's Policies, Positions, and Procedures for Unlawful Content And Activity On Our Social Network* https://news.gab.com/2019/08/gabs-policies-positions-and-procedures-for-unlawful-content-and-activity-on-our-social-network/

[110]Getter nd *Community Guidelines* https://gettr.com/community-guidelines#hateful-behavior

Conclusion

Reset.Tech Australia warmly welcomes the expansion of the Basic Online Safety Expectations (BOSE) as proposed, and in particular:

- The increased focus on covering more systems;
- Improving transparency and accountability, and;
- The introduction of the children's best interests principle.

The ambition and direction of travel of these proposals is both necessary and forward thinking, and should help to reposition Australia as—once again—world leaders in the ambition to create a safe and secure digital world.

A number of proposals could be specifically strengthened, and some enhanced responsibilities to address additional systems could be included to help realise this ambition.

Appendix 1: Young people's perspectives about the best interests principle

Given the importance of the digital world for young people, we set out to ask young people what they thought about the 'best interests' principle and what it might mean in the digital world to them. We surveyed 1,008 young people aged 15-17 in December 2023 (working with YouGov), and supplemented this with a focussed discussion with three young people aged 15 - 17 years old.

We found widespread support for stronger protections for young people in the digital world, and—aligning with existing policy suggestions—support for centering these around young people's best interests. Importantly, a best interests impact assessment was described as a helpful idea. Although this was undertaken to interrogate the application of the best interests principle in the privacy space, we hope it demonstrates the nature of the insights that engaging with young people can generate.

Young people want stronger protections in the digital world

We asked young people in the survey if they felt 'safe and protected' in the digital world when it came to a range of online issues, including online abuse, encountering distressing content, scams and privacy. The vast majority of young people described feeling unsafe and unprotected in the digital world, with young people feeling least safe when it comes to misinformation, scams, distressing content, privacy and online abuse (see figure 7).

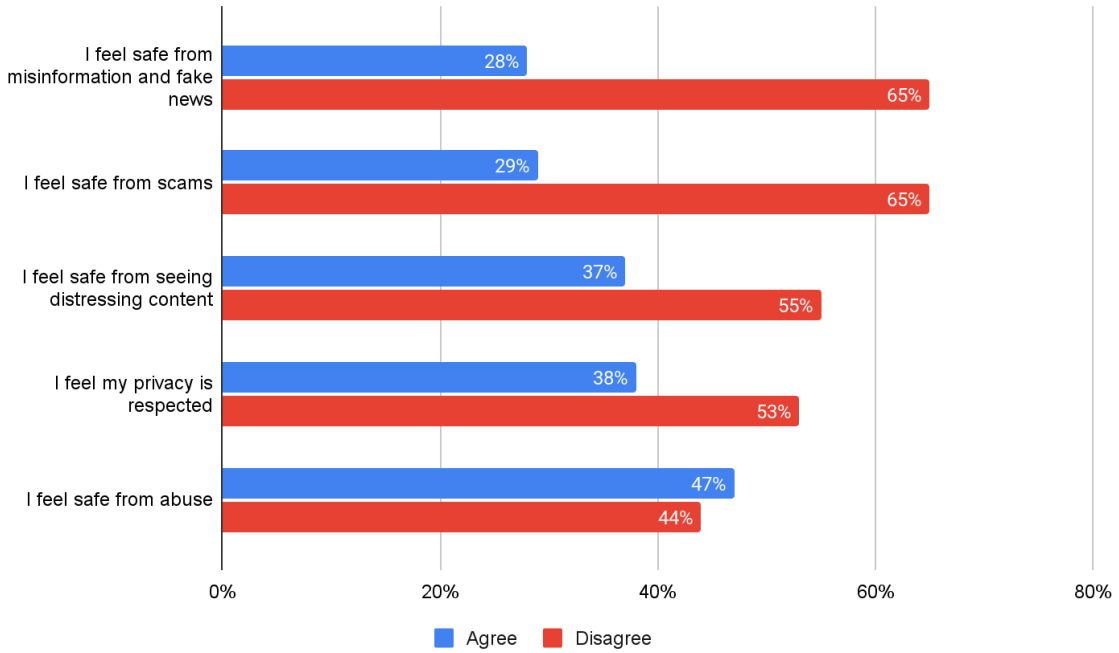

Figure 7: The percentage of young people who agreed or disagreed with various statements about how they felt in the digital world (n=1,008. 'Don't knows' not plotted).

These concerns were echoed by the young people we spoke to who talked about receiving fight content on TikTok, feeling creeped out by face scanning and just routinely facing risks in the online

environment. They were keen to stress that the digital world is still a hugely beneficial part of their lives, but that these sorts of risks exist.

We asked if young people wanted action from the Government around these and there was overwhelming support (see figure 8).

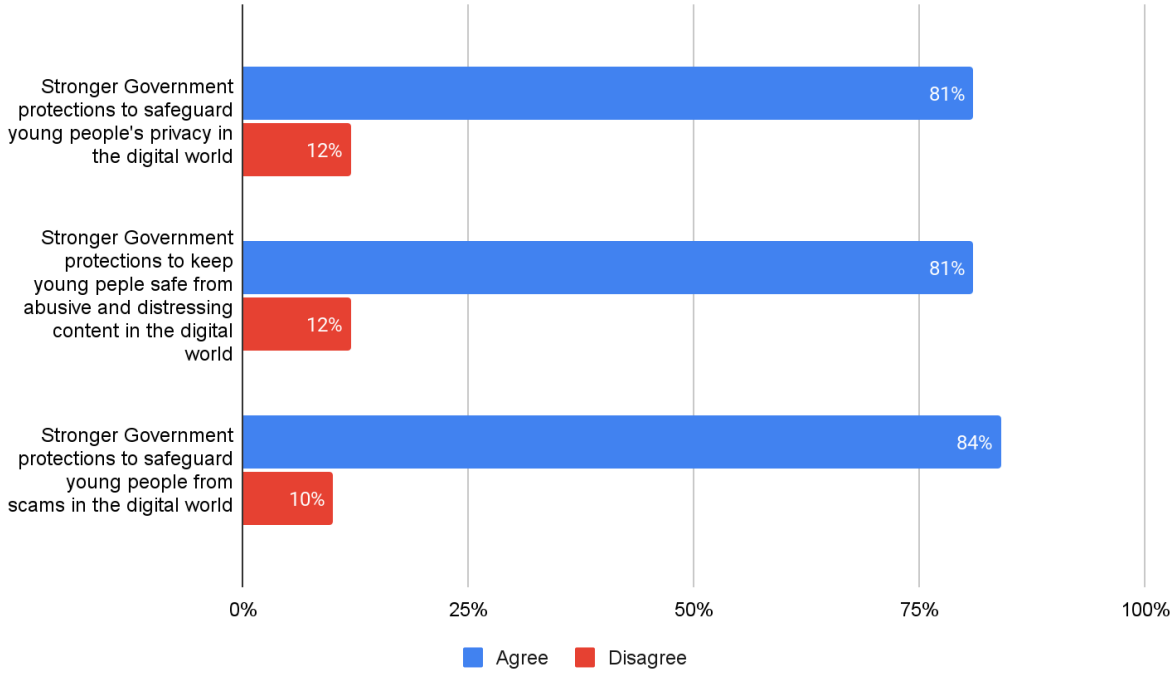

Figure 8: The percentage of young people who agreed or disagreed with various statements about believing we should have particular protections in place (n=1,008. 'Don't knows' not plotted).

Protections need to protect their rights

We asked young people, in the privacy domain, about three of the measures being proposed that involved the use of 'children's best interests as a primary concern', which was framed in the poll as 'respect for young people's rights'. We found strong support for these measures that may also flow over into supports in the online safety space. For example:

- Requirements that all targeting happens in children's best interests (described as a rule to require apps and websites to personalise products for under 18 years old in ways that respect young people's rights): 72% support
- Requirements that data collection happens in children's best interests (described as a rule to require apps and websites to collect data from under 18 years old in ways that respect young people's rights): 70% of respondents agreed that this was desirable
- A Children's Privacy Code (described as a clear set of rules about how to protect young people's privacy): 90% of respondents agreed that this was desirable (see figure 9).

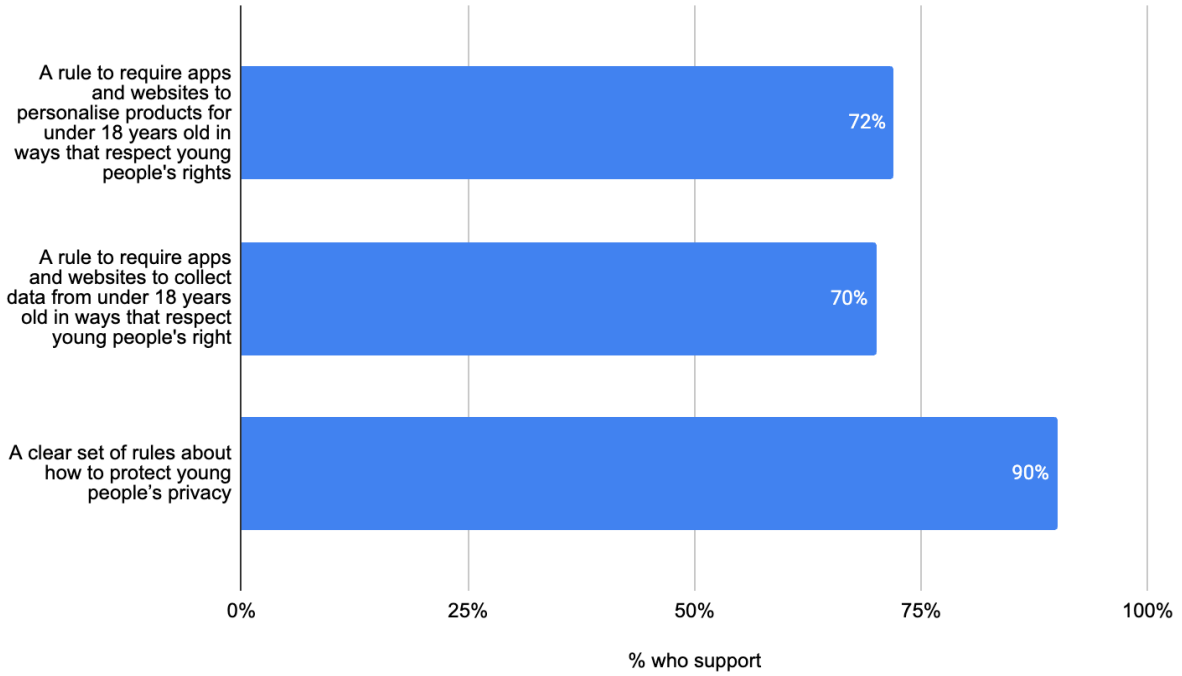

Figure 9: The percentage of young people who supported particular protections (n=1,008).

The young people we spoke to also supported these measures. They noted that what they wanted was to make the digital world act in ways that are better for young people, '*taking it away from young people isn't the answer, it's filtering out the bad'*. The idea that there could be rules to '*dial down the bad*' and *'turn up the good'* felt like the solution to the digital world. They were keen to stress that young people use the digital world daily, and it really affects them; making the digital world work in their best interests could significantly improve their lives.

A best interests impact assessment is a welcome idea

'*Context is important*' said one young person at a focus group, when we asked if they felt—overall—if targeting or data collection worked in young people's best interests or not. When we asked if they felt some sort of impact assessment, or requirement that online services think about how their product might affect young people's rights, would be helpful they described this as '*10/10*' a good idea.

The young people surveyed also supported this proposal. We asked if online services should have to think about and assess how they respect young people's rights in general, which 88% of respondents supported. (See figure 10).

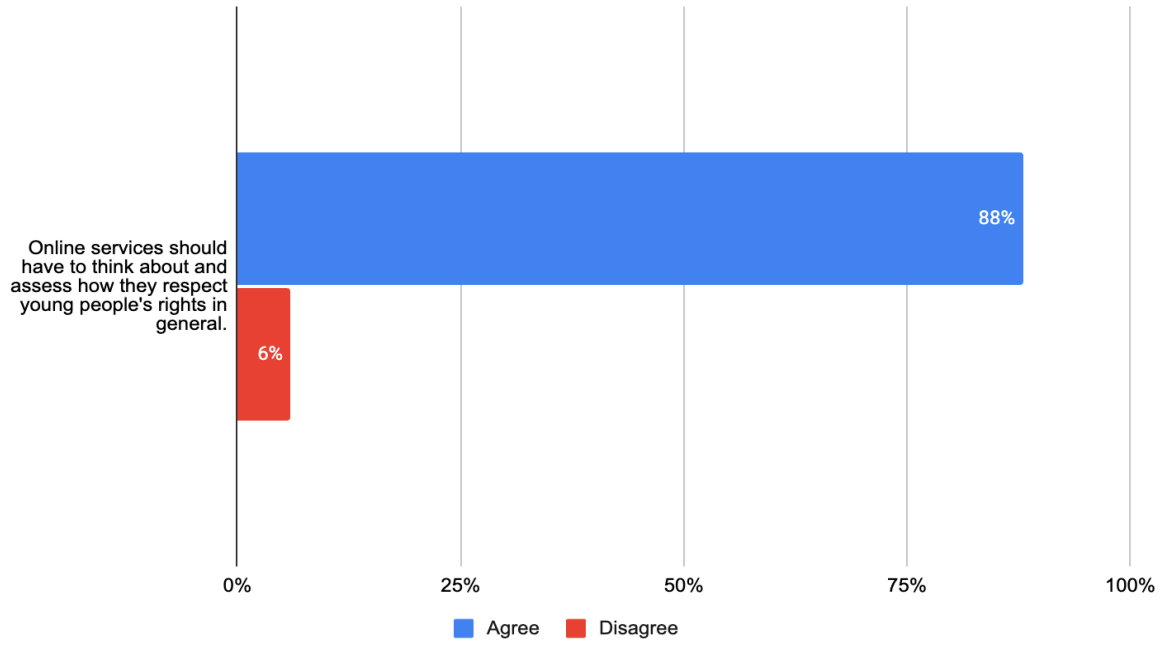

Figure 10: The percentage of young people who agree with the statement that online services that young people use should have to think about and assess how they respect young people's rights in general (n=1,008. 'Don't knows' not plotted).

In our focus group, participants were able to quickly articulate some insightful requirements for best interests impact assessment.

- The assessments are enforced and reviewed somehow. *'The government should implement it, the social media platforms always find a loophole, everyone knows they're just looking for profit at the end of the day. Government should review it and look at it and make sure they're not looking for loopholes or profit to find a way out of it.'*
- The assessments are transparent. We asked who should be able to see these assessments and they say that they '*should be available to their users. All the policies, obviously you don't read them, but someone should look over them to make sure they're ok.'*
- The assessments are 'holistic' and look at all of young people's rights. Specifically *'they should include general overall safety and really think about what features they put in that young people could misuse, like an opportunity for it to go wrong.'*
- The assessments need to be developed in consultation with young people. *'I think for sure, 'cause like we're the ones using it so they should hear from us. If there is anything to add or remove, they're making money, it's just a paycheck. it doesn't affect them, but for us it's a daily part of our lives so maybe they should hear from us. It would help them a lot, if they make it a really good ad and make the best features they will profit.'*