Introducing mandatory guardrails for AI in high-risk settings: proposals paper
eSafety Submission

The eSafety Commissioner (eSafety) welcomes the opportunity to provide a submission to the *Introducing mandatory guardrails for AI in high-risk settings: proposals paper* (Mandatory Guardrails Paper).

In addition to this submission, eSafety is pleased to be a contributing member to the cross-government response for Safe and Responsible Artificial Intelligence (AI). This enables us to contribute our regulatory experience and expertise as Australia's national independent regulator for online safety to this important process.

We also co-authored a joint submission to the Mandatory Guardrails Paper as part of the Digital Platform Regulators Forum (DP-REG). The other DP-REG members are the Australian Competition and Consumer Commission, the Australian Communications and Media Authority and the Office of the Australian Information Commissioner. Building upon the joint submission DP-REG made to the Safe and Responsible use of AI in Australia discussion paper, our joint submission to this process highlighted that DP-REG:

- Shares information about, and collaborates on, cross-cutting issues and activities involving AI and the regulation of digital platforms.
- Supports a coordinated approach to AI across government, while allowing domain-specific regulators to undertake activities within their different frameworks.
- Supports in principle a framework approach (option 2) for implementing mandatory guardrails.
- Emphasises that the success of the framework approach (option 2) is dependent on important principles guiding its implementation.

One of the key principles is that the framework approach allows expert regulators to retain the ability to bring their own distinct lens to proportionately enforcing mandatory guardrails, as informed by each regulator's understanding of the domains and harms they regulate.

As such, the key purpose of this submission is to outline eSafety's role and existing regulatory remit and expertise in relation to AI. This is intended to provide the supporting context for why eSafety supports an approach that progresses a coordinated, whole of government approach to AI regulation, while maintaining the subject matter specificity and coherence of interconnected policy, such as online safety, and related legislation, such as the *Online Safety Act 2021* (Cth) (Online Safety Act).

From an online safety perspective, AI can have a variety of purposes and impacts, some beneficial and some harmful. In the context of online harms, AI tools are especially relevant in relation to the generation of harmful material, including deepfake material, such as image-based abuse (non-consensual intimate images) and child exploitation and abuse material.

It is important that any response to regulating AI adopts an approach that mitigates harms and risks, while harnessing and promoting the possible benefits of AI.

eSafety's approach

As Australia's leader in online safety, eSafety's purpose is to help safeguard Australians from online harms and to promote safer, more positive online experiences.

We undertake a range of work to understand, anticipate and respond to online harms and promote online safety. We also share our expertise and experience as a regulator to ensure our in depth understanding of online safety can inform all online safety measures and initiatives across Australia.

Our regulatory approach is based on the three pillars of prevention, protection, and proactive and systemic change. We also have a regulatory approach that is both risk-based and takes into consideration a range of harms, as our core objective is harms minimisation.

We have a multidimensional regulatory remit that gives us a range of powers relating to AI material. For example, our regulatory levers under the Online Safety Act relating to AI include:

- Our four complaints schemes, which apply to both real and synthetic material which constitutes adult cyber abuse, child cyberbullying, intimate image abuse, or class 1 or 2 material.
- Our systemic regulation, such as industry codes and standards with mandatory obligations for eight key sections of the online industry, as well as the ability to compel transparency from certain services under the Basic Online Safety Expectations (BOSE). These include specific expectations and mandatory obligations around AI, including generative AI. There are also measures that are likely to require the proactive use of AI in order to prevent and detect harmful and unlawful material, such as to detect new child sexual abuse material.
- Our Safety by Design initiative. This is where we work with tech companies to shift their design ethos from 'moving fast and breaking things' to moving thoughtfully and anticipating, detecting and eliminating online threats by embedding protections at the front end, and throughout the development and deployment process. This includes the assessment of harms and threats posed by AI, the active engineering out of misuse and constant monitoring for unanticipated harms following deployment.
- Our Tech Trends workstream. This is where we conduct horizon scanning and work with subject matter experts to identify the online safety risks and benefits of emerging technologies, as well as the regulatory challenges and benefits they may present. This enables eSafety to become much more anticipatory, both with respect to harnessing the benefits of emerging technology but also in mitigating the potential risks.
- Our position statement on generative AI provides an overview of the generative AI lifecycle, examples of its use and misuse, consideration of online safety risks and opportunities, and regulatory challenges and approaches. Our position statement on recommender systems and algorithms explores the benefits and risks of recommender systems and their underlying algorithms, as well as impacts on an individual and societal level.
- Our broad ranging and expansive research, education, awareness raising and engagement efforts relating to online safety, including relating to AI. For example,

we are already delivering webinars for teachers around AI in the classroom. AI has also been discussed in some of the key advisory and consultation mechanisms we oversee, including the Youth Advisory Council, the Trusted eSafety Providers and the National Online Safety Education Council.

- Our broad ranging policy analysis and advice work. This includes our submission to the inquiry into the use of generative AI in the Australian education system to highlight online safety considerations for such use. It also includes our submission to the inquiry into the *Criminal Code Amendment (Deepfake Sexual Material) Bill 2024*, which criminalised deepfake image-based abuse.
- Our international engagement and collaboration with online safety regulators around the world, through the Global Online Safety Regulators Network (GOSRN) and various other international forums and networks. We recognise that online safety is a global challenge that requires a collaborative response. We engage in regular cross-jurisdictional and multi-disciplinary dialogue with our international counterparts and colleagues to better understand regulatory best practice, improve regulatory coherence, and protect Australians in the global online environment.

It is also relevant to note that the Online Safety Act is currently undergoing an independent statutory review. The report is due to government by 31 October 2024.

The Terms of Reference detail consideration of generative AI as part of the Online Safety Act review. However, as one of many Terms of Reference, this also shows that the regulation of AI needs to be understood within the context of a regulator's overall remit and regulatory toolkit. It is important that the enabling legislation of regulators remains the primary mechanism for outlining their powers and functions. It is also important that specific areas, such as AI, are not demarcated in a way that could undermine the cohesiveness and comprehensiveness of a regulator's remit.

The need for a technology, platform, service, distribution and device neutral approach to online safety regulation was highlighted as part of the first statutory review of eSafety's governing legislation back in 2018.

Response to Mandatory Guardrails Paper

Overall, eSafety recognises the importance of progressing a coordinated approach to AI regulation across government. We see the benefit in a harmonised approach across sectors to minimise the potential for fragmentation and inconsistency.

However, in progressing a whole of government approach, it is vital that the subject matter specificity and expertise that has emerged in the context of existing regulatory remits is maintained. In relation to online safety, eSafety has already taken significant work to better safeguard Australians from a range on online risks and harms relating to AI, as outlined above.

Proposed principles for defining what is high-risk

Broadly speaking, eSafety supports the proposed principles for defining what is high-risk AI.

We note that how these principles will be applied in practice will be crucial. There is a risk that when principles are framed at a high level they do not capture or clearly align to the adverse impacts and effects that occur within specific domains.

For example, within eSafety's remit, we see a range of adverse impacts associated with AI, including:

- AI generating and/or amplifying harmful and unlawful content. This includes deepfake child sexual abuse material, image-based abuse, online hate and other forms of abuse. This harms the victim-survivor of these materials, those who inadvertently come across such material and contributes to a less safe and rights-respecting online sphere.
- We are also seeing organised criminals and predators 'innovating' in their techniques to ensnare victims through the use of face swapping technologies on video conferencing platforms. It is only a matter of time before automated chatbots can be scaled to engage in a range of social engineering scams, including grooming of children and financial sexual extortion.
- AI is contributing to the generation and amplification of content that promotes bias and discrimination. This includes promoting misogyny, homophobia, racism and other forms of prejudice. Such content can normalise hate or intolerance. It may also lead to an erosion of trust in online content or institutions.
- AI being used to facilitate, or in ways that can contribute to, radicalisation and extremism. This includes radicalisation towards terrorism and violent extremism. For example, AI can be used to create terrorist and violent extremist content. Some models can also analyse social media posts, online interactions, and other data sources in order to tailor propaganda.

Of the proposed principles, the one that most closely relates to eSafety's remit is 'the risk of adverse impacts to an individual's physical or mental health or safety'.

We suggest limiting health and safety to only 'physical and mental' does not capture the full range of safety issues, including psychological and emotional safety, particularly in the context of leveraging AI as a form of coercive control. AI can also be weaponised to perpetrate simulated abuse and assault in hyper-realistic and high sensory immersive environments. When online abuse or assault is happening in real-time in private immersive spaces, it can have a real, traumatic and visceral impact on the victim survivor.

Another example is the proliferation of AI 'nudifying' services. While based on a technology that may have beneficial or neutral applications, the application in this context is negative and nefarious: the primary purpose of these services is to create sexual or intimate imagery of people without their consent. They are unlikely to employ sufficient protections to prevent their misuse for the creation child sexual abuse material. Services like these have proliferated online and are easy to find and use. While these services make it simple and low-cost for the perpetrator, the cost to the victim-survivor can be devastating. Accordingly, while the proposed principles seem to categorise this type of risk as one that may have 'adverse impacts to the broader Australian economy, society, environment and rule of law' (principle e), eSafety suggests it should also be assessed as presenting significant risks to individuals' rights and safety (under principles a, b and c).

eSafety welcomes further discussions on how these proposed principles will be applied in practice and tailored to capture and align with the adverse impacts and effects that occur within specific domains.

Proposed mandatory guardrails for high-risk AI

eSafety supports having mandatory guardrails for AI. We note that it is important to consider how new mandatory guardrails would be mapped against existing legislative and regulatory requirements, to ensure these can work in step with one another.

To illustrate, we can provide some examples of existing mandatory guardrails within eSafety's systemic regulatory powers.

Industry codes and standards under the Online Safety Act address specific online safety issues in eight sections of the online industry across the digital stack. While AI-generated material is treated under the legislation in the same way as 'real' class 1 and class 2 material, the risks associated with AI generated material have necessitated specific requirements in relation to AI-related features in some codes and standards. For example:

- the Search Engine Services (SES) code, which was registered on 12 September 2023 and came into effect on 12 March 2024, requires search engine providers to take steps to ensure AI functionality integrated into search engines does not return search results that contain class 1 material like child sexual abuse material (CSAM). The SES code also requires action to reduce the accessibility of synthetic materials generated by AI via the search engine service, and make clear when a user is interacting with AI for example to generate search results. These existing requirements generally align with proposed guardrails 1 (accountability processes), 2 (risk management processes), 4 (testing and monitoring), and 6 (informing end-users). eSafety also has the power to require providers to submit a report on the steps taken to comply and an explanation as to why these measures are appropriate, in alignment with proposed guardrail 9 (maintain records for third parties to assess compliance).
- the Designated Internet Services (DIS) standard, prepared after the Commissioner's rejection of an industry-drafted code, includes specific obligations for certain 'high impact generative AI' services to prevent AI features from being used to generate outputs of child sexual abuse material and pro-terror material, where the risk of generating X18+ or RC (refused classification) content is considered not to be immaterial. This includes by regularly reviewing and testing models and promptly making adjustments. These requirements specifically align with proposed guardrail 4 (testing and evaluation), although the thresholds and the corresponding obligations that attach are carefully calibrated to target high risk uses in the online safety context. eSafety registered the DIS standard in June 2024, which is expected to come into effect in December 2024. eSafety will have the power to require reporting from services about their compliance.

The Basic Online Safety Expectations (BOSE) outline the Australian Government's expectations that social media, messaging and gaming service providers and other apps and websites will take reasonable steps to keep Australians safe.

- To date, eSafety has issued 27 transparency notices requiring providers to report on the steps they are taking to keep Australians safe online and address unlawful and harmful material and activity. The notices have included specific questions on how AI is used to improve safety on the service. For example, detecting and removing child sexual abuse material and grooming. It also includes how services are addressing

safety risks of AI such as the potential amplification of harmful content via recommender systems. Findings to date have been published on the eSafety website. This work serves to bolster proposed guardrail 8 (transparency).

- In March 2024, notices were issued covering, among other things, generative AI in relation to terrorism and extremism, and child sexual abuse, with appropriate information to be published in due course. eSafety expects to publish a summary of findings later this year.
- On 30 May 2024, the Minister for Communications amended the Online Safety (Basic Online Safety Expectations) Determination 2022 to include an explicit expectation that providers of relevant services with generative AI capabilities will take reasonable steps to:
    - consider end-user safety and incorporate safety measures in the design, implementation and maintenance on generative AI capabilities on the service (consistent with proposed guardrails 1, 2 and 4, among others)
    - proactively minimise the extent to which generative AI capabilities may be used to produce material or facilitate activity that is unlawful or harmful (consistent with guardrails 2 and 3, among others).
- In July 2024 the first 'periodic' notices were issues to eight providers, that require six monthly reports to eSafety for a period of two years, including on generative AI. The first reports are due to eSafety in February 2025.

Regulatory options to mandate and implement guardrails

In considering the regulatory options to mandate guardrails, eSafety reiterates its position that it supports a coordinated approach across government, while allowing domain-specific regulators to undertake activities within their different frameworks. This would ensure the subject matter specificity and policy nuance of specific domains is not lost in efforts to drive cross-government consistency.

On the option to introduce a new AI-specific Act, eSafety considers this presents practical challenges.

The digital technology landscape is constantly and rapidly evolving and this includes the growth and evolution of AI. With new and emerging forms of technology often being created, it raises the question of whether a new technology-specific regime would be required for each emerging technology development. This creates considerable risks that a technology-specific regulatory framework will not keep pace with technological change.

Another question to consider is how the establishment of a monitoring and enforcement regime overseen by an independent AI regulator would interact with existing regimes that capture AI. As noted above, eSafety already has a range of regulatory measures that apply to some AI services, including mandatory guardrails and enforceable transparency measures. An AI-specific Act enforced by a standalone regulator creates a risk that guardrails are enforced in a way that duplicates existing guardrails in other regulatory frameworks focused on reducing specific harms.

An AI specific act also presents the risks of consumer and industry confusion, as well as regulatory burden.

Of the options presented in the proposals paper for implementing mandatory guardrails, we consider that a framework approach (option 2) is the most suitable. We therefore support option 2 in principle. However, as outlined in our DP-REG joint submission, for a framework approach to be successful it will need to:

- allow expert regulators to retain the ability to each bring their own distinct lens to proportionately enforcing mandatory guardrails, as informed by each regulator's understanding of the domains and harms they regulate.
- clearly set out a common set of broad definitions.
- establish the scope of its application and provide guidance on common regulatory options available to regulators.
- reduce regulatory burden and complexity for regulated entities, who have established and productive relationships with domain-specific regulators.
- reduce duplication of effort by industry participants by reducing barriers to information-sharing between existing regulators.
- continue to build capability of regulators to identify, investigate and respond when AI intersects with their regulatory frameworks, including an uplift in technical skills for existing regulators to better understand the technology.
- promote coordination between regulators, including to facilitate the appropriate allocation of complaints from individuals affected by AI.

For example, eSafety considers that under option 2, definitions used in industry codes and standards, the Basic Online Safety Expectations, and the Online Safety Act more broadly could usefully be further harmonised by framework legislation on AI. However, as outlined above, this would only be successful if it was done in way that maintained a recognition of the unique risks, harms and context of online safety, as well as enabled the positive uses of AI online to prevent harms to Australians.

Lastly, as with any policy area, it is vital that this process is undertaken through a process of co-design that is centred in lived experience. This is to ensure the voices and perspectives of those with lived experience of AI can inform the policies and programs that affect them.

We are happy to provide any further information that would assist. We also look forward to continuing to contribute to cross-government activities that support a whole of government approach to AI, while also ensuring particular policy areas, such as online safety, retain their cohesiveness and subject matter specificity.