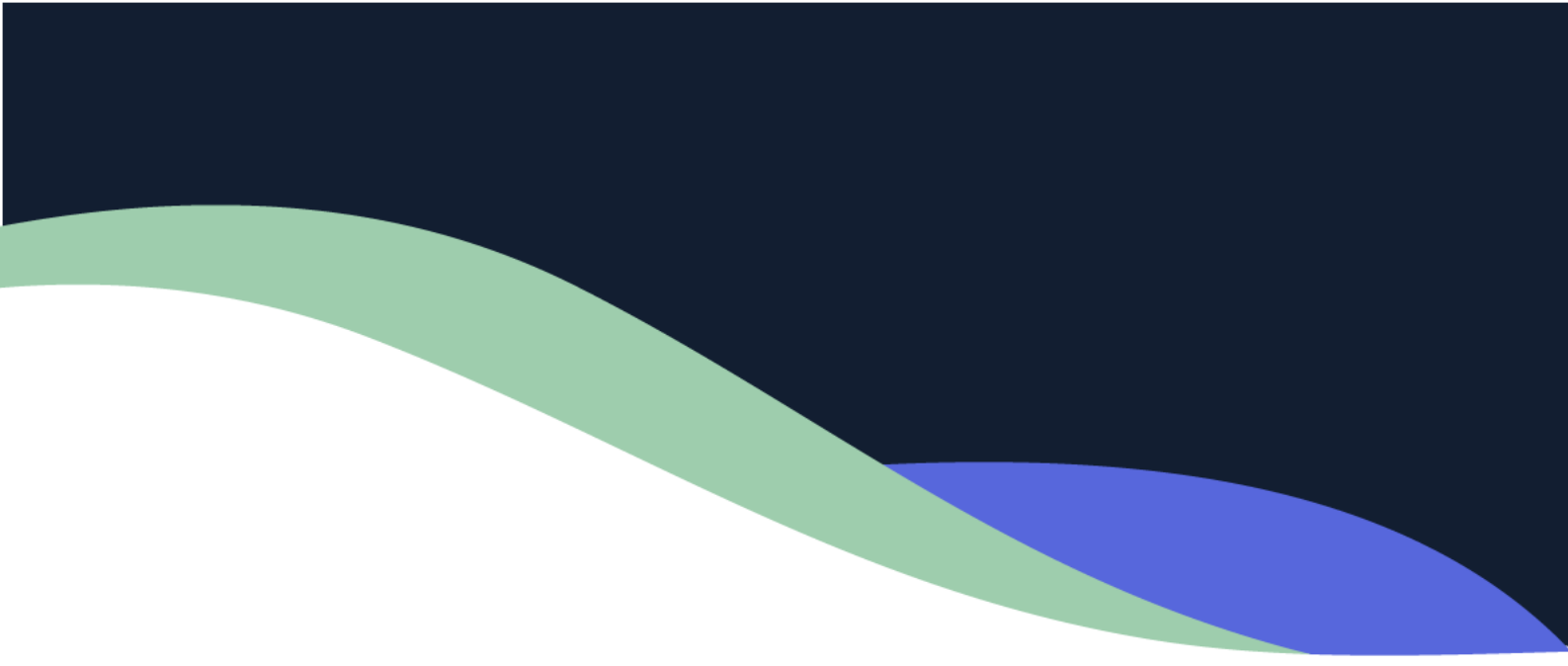




Joint Select Committee on Artificial Intelligence

eSafety submission

September 2026

Contents

Introduction..... 2

eSafety’s regulatory approach to AI 3

AI and online safety reform landscape 15

International engagement on AI safety 16

Conclusion 16

References 17

Introduction

The eSafety Commissioner (**eSafety**) welcomes the opportunity to provide a submission to the Joint Select Committee on Artificial Intelligence (**AI**).

While acknowledging the broad ranging Terms of Reference for this inquiry, this submission focuses primarily on the risks and harms arising from AI, and the online safety regulatory framework as it applies to AI. This is in line with eSafety's remit which is focused predominantly on consumer-facing AI products and services.

In making this submission, we highlight the following key positions:

- **The safety of Australians, including children and young people, must remain at the forefront of AI advancement, development and deployment.** While AI poses significant benefits, these opportunities must not come at the expense of safety. As AI continues to advance and evolve, so too do the ways in which bad actors can exploit AI, exacerbating existing harms and creating new risks where safeguards are absent or inadequate. We are already seeing the weaponisation of AI to create child sexual exploitation and abuse (**CSEA**) material, deepfake image-based-abuse, as well as how AI can amplify technology-facilitated abuse, and how AI companions can expose children to age-inappropriate and harmful material.¹
- **Effective AI safety regulation is necessary to maximise AI benefits, economic productivity and growth.** This includes establishing the right regulatory settings, which requires effective, balanced and proportionate regulation, supported by robust policies, procedures and standards. Effective regulation creates a culture of continuous improvement and innovation, with safety enabling new markets for competition and providing opportunities for organisations to differentiate themselves by innovating on safety features.
- **Strong safeguards help to build trust and create an environment that promotes innovation.** Online harms reduce productivity and undermine consumer trust, whereas safety improves engagement and adoption. It can take years to build trust, and a moment to lose it. Embedding safety throughout the AI lifecycle builds confidence and sends a signal to users that encourages and supports their engagement.
- **eSafety supports a coordinated, whole of government approach to AI regulation that strengthens existing laws and regulatory structures to minimise harms.** Alignment with frameworks and forthcoming reform will build on existing opportunities while minimising fragmentation and inconsistency. eSafety is undertaking compliance and

¹ eSafety's regulatory schemes address AI-facilitated harms. For example, under the Age-Restricted Material Codes, AI companion chatbots capable of generating sexually explicit, high-impact violence or self-harm material need to confirm someone is 18 or older before they access that material. The work we do to address these harms is detailed below in the submission.

enforcement activity which captures certain AI services under our existing regulatory framework. The Government's planned Digital Duty of Care is intended to apply to all existing services under the *Online Safety Act 2021 (Cth)* (the **Online Safety Act**), including AI services. Regulatory coherence and alignment are especially important as AI regulation intersects with a range of remits, including safety, privacy, and security.

This submission provides an overview of eSafety's regulatory approach to AI, outlining the existing online safety framework and our regulatory insights into AI-related risks and harms. It also highlights forthcoming online safety reforms, which provide a critical foundation for a coordinated approach to AI regulation.

eSafety's regulatory approach to AI

eSafety is Australia's national, independent regulator and educator for online safety. eSafety's remit is set out through its powers and functions under the Online Safety Act.

The Online Safety Act captures harms that are facilitated through AI, including where they are AI-generated, AI-modified, or AI-enhanced, where relevant legislative thresholds under regulatory schemes are met.

AI is becoming increasingly integrated across a breadth of online products and services, including standalone apps or services, within large language model (**LLM**) backed AI services and recommender systems, as well as messaging and social media platforms. AI is also integrated within search functionality.

As noted in the introduction, eSafety's regulatory focus is predominantly on consumer-facing AI products and services that have a user-interaction component, such as generative AI tools and AI chatbots (including AI assistants and AI companions). Baseline safeguards and risk management measures are equally relevant considerations to inform the development of legal frameworks for advanced, frontier and open-weight AI models².

eSafety's regulatory approach comprises a holistic regulatory toolkit underpinned by the three pillars of **prevention, protection, and proactive and systemic change**. The next section provides an overview of eSafety's work to address AI risks and harms under these pillars.

² Advanced, frontier AI models are highly capable generative models that process input data (including text, images, video and other information) to generate new content, predictions or other outputs. Open-weight models are a subset of generative models for which the trained parameters (or 'weights') are publicly released, enabling developers to download, run, fine-tune and adapt the model for specific purposes, subject to applicable licence terms.

Prevention

Education and awareness

Underpinning all our education and awareness raising work is the importance of all forms of digital literacy, including AI literacy. Digital literacy is the foundation for people to be able to interact online safely, thoughtfully, competently and critically.

We continue to **provide education and training programs to equip Australians with digital and algorithmic literacy**, including understanding the safety impacts of AI. We are delivering webinars for educators, youth-serving professionals, and parents and carers about aspects of AI, including on AI-generated images and videos, understanding AI companions and how AI is influencing new online risks for children and young people. Throughout the 25/26 financial year and the first 3 months of the 26/27 financial year, over 3000 adults participated in one of these webinars. A new webinar will launch during the *16 Days of Activism* in November 2026, exploring the growing threat of deepfakes targeting girls and women.

eSafety has published online safety advisories, which discuss the risks, harms and impact of AI for children and young people, including in educational settings.³

AI is also an area of focus in **key advisory and consultation mechanisms** which eSafety oversees, including the [eSafety Youth Council](#), the [Parent Advisory Group \(PAG\)](#), the [Trusted eSafety Providers \(TEP\)](#) and the [National Online Safety Education Council \(NOSEC\)](#).

Research on children's engagement with AI assistant and companion apps

In August 2026, eSafety released a **study examining the prevalence and nature of AI assistant and companion app usage** among children aged 10-17, [Talking to machines: Children's experiences with AI assistants and companions](#) (*Talking to Machines*). This study includes exploration of children's usage patterns, motivations, evolution of engagement, emotional response, exposure to potentially inappropriate or harmful advice or interactions, information sharing, and awareness of platform/app reporting options, as well as parental awareness of their children's use of these AI tools.

Key findings include:

- Most children had used an AI assistant (78%), and a notable minority had used an AI companion (8%).

³ Examples of advisories: [AI chatbots and companions – risks to children and young people](#) published February 2025; [Deepfake damage in schools: How AI-generated abuse is disrupting students, families and school communities](#) published June 2025; [Sharing school imagery in the age of AI: risks, deepfakes, and how to protect students and staff](#) published July 2026.

- Most children who had used an AI assistant or companion had used them for functional reasons, such as information, learning or translation (91%), or writing or editing (70%), while many also used them for entertainment (68%).
- More than half of children who had used an AI assistant or companion (54%) reported using them for personal or social reasons, including advice about what to do in a situation (33%), advice about their physical health (33%), to chat about their feelings or challenges (22%), advice about their mental health or wellbeing (20%), or to ‘hang out together’ or chat about their interests (17%).
- Among children who had ever used an AI companion or AI assistant, 20% indicated they used them daily or more frequently. This equates to 16% of all children (aged 10–17) using these tools at least daily.
- 1 in 5 children who had used an AI assistant or companion (20%) had experienced potentially inappropriate or harmful interactions with these tools. This included AI assistants or companions:
 - attempting to control them in some way (13%)
 - sharing or saying hateful things about one or more groups of people – for example, based on that person’s gender, culture, disability, sexuality or religion (9%)
 - giving harmful advice (6%)
 - being offensive, mean or bullying towards them personally (6%)
 - sharing real or AI-generated nude images or videos with them (of the AI/its avatar or of someone else) (4%)
 - showing them real or AI-generated sexual images (for example, real or AI-generated pornography) or engaging in sexual chat (5% of teens – not asked of children under 13).
- Around 1 in 3 children who had used an AI assistant or companion (32%) had shared personal data or potentially sensitive information with it.

In eSafety’s study, AI chatbot tools were categorised as either assistants or companions based on their stated purpose and how they were marketed. Tools classified as AI assistants were those primarily marketed as functional tools to assist users with specific tasks, organisation, education or learning, while AI companions were primarily marketed as sources of friendship, emotional support, romantic companionship or an antidote to loneliness and tended to be engineered to seem caring or intimate.

However, the boundaries between these tool categories are becoming increasingly blurred, with both offering features such as memory, personalisation and human-like interaction. As a result, users, particularly teens, may engage with them in social or relational ways regardless of how they are positioned (Neugnot-Cerioli, 2026).

The prevalence of children in Australia using AI assistants and/or companions and the risks these services can pose highlights the need for robust guardrails. Without strong safeguards, these chatbots have the potential to:

- expose children to potentially harmful age-inappropriate interactions, including sexually explicit chats and other forms of pornographic material, as well as violent material and self-harm and suicide material
- spread misinformation or disinformation and reinforce harmful stereotypes
- pose privacy risks
- create emotional over-reliance
- have negative psychological and social impacts.

(Abid & Farooqi, 2021; Zhang et al., 2025; Internet Matters, 2025; Neugnot-Cerioli, 2026; Robb & Mann, 2025; Common Sense Media, 2025).

The potential impacts on children cannot be dismissed, especially where AI chatbots can generate or share unlawful or age-inappropriate material. The risks of harm can be higher because AI chatbots are designed to feel human-like and are highly interactive. Children and teens can be particularly susceptible to these risks as their brains are still developing. For example, they may be more trusting of the information or advice they receive from AI chatbots and therefore more susceptible to their influence or find it harder to distinguish between simulated AI empathy and genuine understanding (Common Sense Media, 2025; American Psychological Association, 2025).

In order for children in Australia to be able to use AI for the benefits it may have, it is vital that the systems are safe to use. Leveraging and strengthening existing legislation such as mechanisms under the Online Safety Act and adopting a Safety by Design approach can work to mitigate the risks that children face when using these services.

Protection

Under the Online Safety Act, eSafety administers [complaints schemes](#) to address:

- Cyberbullying material targeted at Australian children
- Non-consensual sharing of intimate images (image-based abuse)

- Cyber-abuse material targeted at Australian adults
- Illegal and restricted online content, such as child sexual abuse material and pro terror material.

eSafety can investigate and facilitate removal of material regardless of whether AI was used in its production. For example, this means we can seek removal of child sexual exploitation and abuse material or image-based abuse material, even when it has been digitally altered or generated by AI.

While complaints to eSafety involving synthetic material remain a small proportion of overall complaints, it can be difficult to detect and is expected to increase with the availability of generative AI tools.

Terrorist and violent extremist content

The spread of terrorist and violent extremist content and its role in online radicalisation, particularly of young people, is a concern both in Australia and overseas.

eSafety is concerned by reports that terrorists and violent extremists are moving to capitalise on the emergence of generative AI and are experimenting with ways this evolving technology can be misused to cause harm. eSafety is particularly concerned about the role of livestreaming, recommender systems and generative AI in producing, promoting and spreading terrorist and violent extremist content and activity.

The Christchurch Mosque terrorist attack demonstrated how livestreaming can transform an act of violence into ongoing harmful content, enabling real-time dissemination and inspiring copycat behaviour. We continue to receive complaints about this content circulating online and have identified terrorism manifesto material that has been inspired by the Christchurch event. Recommender systems can further amplify such extremist content by increasing visibility and directing users towards increasingly radical material, as evidenced by the continued proliferation of the incident. We have also identified artificially recreated versions of Christchurch, including gamified 'first person style walk-through' adaptations. This demonstrates how generative AI can now magnify these risks by enabling the rapid, low-cost creation and dissemination of sophisticated extremist material, including synthetic images, audio and video, with the potential to contribute to radicalisation.

Digitally altered or generated intimate images

The prominence of deepfake image-based sexual abuse garners media attention from time to time. The definition of an intimate image in the Online Safety Act makes it clear that synthetic or digitally altered images, including AI deepfake material, are included.

eSafety receives reports of digitally altered images, including sexualised deepfakes that appear to use publicly available photos as source material.

A non-sexual deepfake, such as a face swap, fake dance video or harmful meme, may still cause deep distress. But this content may not meet the legislative threshold for child cyberbullying, image-based abuse or adult cyber abuse under the Online Safety Act, which was not designed to capture material which is purely slanderous or defamatory. While such material may cause harm, eSafety may be unable to seek its removal unless it meets the relevant threshold under its complaint schemes. eSafety has written about AI-generated deepfakes and other forms of image misuse in our most recent advisory on [‘Sharing school imagery in the age of AI: risks, deepfakes, and how to protect students and staff’](#).

Complaints to eSafety of digitally altered intimate images, including deepfakes, jumped by almost 75% last year compared to 2024. eSafety received almost 90 reports in 2025 and approximately half of these involved children. However, reports to eSafety do not represent the true scale of harm.

AI and sexual extortion complaints

eSafety has observed the use of synthetic images in complaints received about the sexual extortion of children and young people, particularly in cases where victims declined to send an image. In the absence of actual images, offenders will create digitally altered images to attempt extortion. Some complaints also reference threats to use AI or create digitally altered images to increase pressure and compliance.

In June 2026 eSafety delivered a sexual extortion education and awareness campaign, ‘If Sextortionists Were Honest’, aimed at men aged 18-24. The campaign utilised AI-generated content ethically and intentionally, which allowed the audience to clearly see how manipulation using AI-generated content works in practice, making it easier for them to recognise and understand the behaviour. This creative innovation sought to raise awareness amongst young Australian men on what to look for, to expose some of the tactics for this kind of extortion and help people recognise the warning signs. Over a 34-day period, the campaign reached over 650,000 men aged 18-24, with over 1 million video views of over 15 seconds or longer. Close to 50,000 users visited the [campaign web page](#). More than 300 people clicked on ‘Report abuse’, and more than 30 reports to eSafety were completed as a result of the campaign.

Case study: AI Deepfakes in Australian Schools

The increasing accessibility of generative AI tools is driving a rise in deepfake abuse within Australian schools, with incidents commonly involving boys aged over 16, creating and sharing sexualised deepfake images or videos of girls using publicly available photographs sourced from social media.

A recent case involved a 16-year-old girl whose Instagram and TikTok images were used to generate sexually explicit deepfake videos using an AI video-generation platform. The manipulated content was distributed online without her consent, resulting in significant emotional distress, reputational harm and police involvement.

eSafety worked with the child, school and police to have the images removed, yet victims experience impacts that are comparable to other forms of image-based abuse, including shame, anxiety, social isolation, concerns for personal safety and ongoing fear that material will re-emerge online.

In response to incidents of this nature, eSafety has established a [Deepfake Incident Response Plan for schools](#), which has been provided to education sector stakeholders through the **National Online Safety Education Council (NOSEC) network**. The plan provides practical guidance for schools to respond quickly and effectively by preserving evidence, accessing appropriate reporting pathways, supporting affected individuals, and managing communication and wellbeing across the school community.

Proactive and systemic change

Codes and standards

Online safety codes and standards under the Online Safety Act place enforceable requirements on a wide range of services, including generative AI services, to protect all Australians from unlawful material and Australian children from age-restricted material. These obligations cover ‘class 1’ material such as child sexual exploitation material, pro-terror material and extreme crime and violence, as well as children’s access to ‘class 2’ material, such as pornography, self-harm and violence.⁴

⁴ Class 1 and class 2 material is defined under the Online Safety Act by reference to Australia’s [National Classification Scheme \(NCS\)](#). Class 1 material is material that is, or would likely be, refused classification. Class 2 material is material that is, or would likely be, classified as either X18+ or R18+.

eSafety has a range of options to enforce non-compliance including seeking civil penalties up to \$54.6 million.

Case study: Progress under the Age-Restricted Material Codes

Since the Age-Restricted Material Codes came into force on 9 March 2026, seven service providers have made safety improvements to their services following engagement with eSafety. This includes services which implemented age assurance and other safety uplifts such as real-time detection of self-harm and suicide-related inputs that trigger in-service referral of the user to appropriate crisis support services.

eSafety's priorities for investigation and enforcement of codes and standards, insofar as they relate to AI, include:

- Generative AI models and services (in particular, 'nudify' services and AI companion chatbots) that can create child sexual exploitation material, online pornography, self-harm material and violence instruction material, and the prevention of AI companion chatbots from serving harmful material to children;
- 'Gatekeeper' services, such as app stores and search engines, which enable access to a wide range of services/apps such as 'nudify' services, including those that may be unresponsive to eSafety outreach.

Case study: Enforcement against 'nudify' services

Nudify services allow users to upload images of real people and generate on-demand, sexually explicit deepfake content, posing serious risks, including potential non-consensual exploitation and harms to children. Abuses enabled by these types of technologies may also include sexual exploitation of children, sexual extortion, image-based sexual abuse, high impact cyberbullying, and misogynistic harassment.

Under the Age-Restricted Material Codes, which cover all sections of the online ecosystem in Australia, AI services which allow users to access or generate age-restricted material, including sexually explicit material, must put in place appropriate age-assurance measures to prevent children under 18 from accessing their service or generating this material. Similarly, the Unlawful Material Codes and Standards require these services to prevent the generation of child sexual exploitation material.

eSafety has taken enforcement action against several of the world's largest nudify services for allegedly failing to comply with their obligations under the Online Safety Act, including failing to implement appropriate safeguards to prevent the creation of synthetic child sexual exploitation material involving Australian school children.

In total, eSafety's enforcement actions have led to eight of the most widely used nudify services in Australia, either withdrawing access from Australia, or putting in place measures to comply with their obligations under the Codes and Standards, such as age-assurance to prevent children's access. These nudify services were receiving over 200,000 visits a month from Australian users. This shows the meaningful and measurable outcomes that eSafety's regulatory actions have brought about to protect Australians, including protecting children from accessing age-inappropriate content. These types of services continue to be a focus for eSafety's compliance and enforcement teams.

Following our standards compliance engagement, AI model distribution service *Hugging Face* also updated its rules to require users to take measures to prevent their models from being used to generate child sexual exploitation material and pro-terror material. As far as we are aware, the Online Safety Act is the only online safety legislation globally that applies to model distribution services, a vital part of the ecosystem.

We also have ongoing investigations into other major services, including xAI's '*Grok*' AI Assistant following reports in January 2026 that the service was generating child sexual exploitation material.⁵

eSafety is also engaging **the most used AI and companion services in Australia** in relation to their obligations. Key services have already taken action to prevent children's use for sexualised chats or encouragement in self-harm as a result of our direct regulatory engagement, with companion services *Nomi* and *Soulkyn* implementing age-checks to prevent children's access.

Basic Online Safety Expectations: Transparency Requirements

The Basic Online Safety Expectations (**BOSE**) outline the Australian Government's expectations that social media, messaging, apps and other websites will take reasonable steps to keep Australian users safe. Under the Online Safety Act, eSafety can require providers of these services to report on how they are meeting any or all of the BOSE through

⁵ These reports also covered the use of Grok to generate deepfake image-based abuse of adults, although this is not something expressly captured by the Online Safety Act's scope for codes and standards.

transparency reporting notices. The obligation to respond to reporting requirements is enforceable and backed by civil penalties and other mechanisms.

In March 2026, eSafety published a [transparency report](#) summarising responses from four AI companion services – *Character.AI*, *Nomi*, *Chai*, and *Chub AI* – to notices given about steps each service was taking to comply with the BOSE. The findings revealed gaps in basic safeguards for children, which applied variously across the services.

For example, while service providers relied on app store ratings and/or self-declaration at signup, none of the service providers had robust age assurance in place, meaning children could still reach adult spaces and features. Under Australia's Online Safety Codes and Standards, these types of age-inappropriate materials should not be accessible by people under 18 – even where it is lawful for adults to access them.

eSafety also observed safety gaps in some of the companion services not checking inputs or outputs across all relevant text, image and video models used to provide their service, to keep their models from being used for child sexual exploitation and abuse, self-harm or pornography (including sexually explicit chats with child users).

These findings highlighted the safety deficiencies on some AI services which can cause great harm to Australian users, particularly children. They also serve as a baseline understanding of some of the issues and gaps prior to commencement of the Age-Restricted Material Codes which focus on class 1C and class 2 material and came into effect in March 2026 (the reporting period for these notices was 1 July 2025 to 30 September 2025). eSafety is now monitoring and enforcing compliance with these codes.

Safety by Design

As part of our proactive and systemic change work, eSafety's **Safety by Design initiative shifts the onus of safety from the user to service providers**. This is about addressing harms before they occur and embedding safety at all stages of the technology product lifecycle. There are **three core principles: service provider responsibility, user empowerment and autonomy**, and **transparency and accountability**.

We have developed a [Safety by Design toolkit](#) focused on addressing child sexual exploitation and abuse. The toolkit aims to ensure that child safety is prioritised at every stage in the development of online products and services.

The toolkit explores, in particular, the use of generative AI to harm children online. It details proactive measures industry can adopt to address risk and encourages the use of AI-enabled tools to detect and report such abuse. Contributors to the toolkit include Google, Meta, OpenAI and Stability AI.

We have been engaging with AI companion and chatbot services to understand risks, harms and best practice online safety interventions in the sector.

We are also developing an additional chapter to the technology-facilitated [gendered violence Safety by Design resource](#) we published in 2024. The new chapter will focus on Safety by Design approaches to technology-facilitated coercive control. The chapter will also include an exploration of how AI is being weaponised to perpetrate that harm alongside how to address and mitigate it.

Case study: Insights from eSafety's Technology-Facilitated Abuse (TFA) Support Service

eSafety's TFA support service provides specialist guidance and advice to frontline workers when they are supporting clients experiencing technology-facilitated abuse as part of family, domestic and sexual violence.

Through the TFA support service's engagement with frontline workers, victim-survivors and the broader domestic, family and sexual violence sector, including through enquiries and sector consultations, eSafety has observed concerns regarding a range of AI-facilitated harms in the context of domestic, family and sexual violence. eSafety has been made aware of emerging concerns about the potential misuse of AI in technology-facilitated abuse, including its capacity to lower the technical barriers to engaging in certain harmful behaviours. This may reduce the time, effort and expertise required to undertake abusive conduct and enable this conduct to be carried out more readily and at greater scale.

Broader concerns identified through these engagements include the potential for AI to amplify existing patterns of technology-facilitated abuse by increasing the speed, scale and accessibility of coercive and controlling behaviours such as monitoring, surveillance, information gathering, impersonation and harassment. This includes the potential use of AI tools to record, transcribe, analyse and translate private conversations, rapidly gather and consolidate personal information from online sources, and to facilitate impersonation through voice cloning, deepfakes and other types of AI-generated content.

There are also concerns that AI may increase the efficiency and scale of coercive and controlling behaviours by enabling individuals to monitor, analyse and act on large volumes of information more quickly and with less effort. AI tools may also support manipulation and abusive communications through the generation of persuasive, targeted, deceptive or high-volume content, and may be misused to facilitate financial abuse, impersonation, harassment and other forms of technology-facilitated harm.

Concerns have also been raised that victim-survivors may use AI chatbots to seek advice, validation and support in relation to experiences of violence and abuse, creating potential risks where responses are inaccurate, incomplete or not trauma-informed.

Regulation which promotes Safety by Design can help prevent harms across various aspects of AI uses. A whole of government approach to ensuring that AI services available on the Australian market have inbuilt safety features from the outset will help to minimise risk of AI-facilitated harms. It is also critical that companies are assessing safety risks, and implementing measures to effectively minimise those risks, before making changes and updates to their AI services.

Tech trends and challenges

As a proactive and anticipatory regulator, we also monitor emerging technologies and evolving harms through our [Tech Trends and Challenges](#) program. Our first tech trends and challenges position statement was written on [deepfakes](#) and steps the purveyors and profiteers of these technologies could do to mitigate the significant risks we foresaw on the horizon.

In 2022, we published our first tech trends and challenges position statement on recommender systems and laid out a blueprint for incorporating Safety by Design, increasing transparency for users and giving them more choice. This [position statement](#) was subsequently updated in May 2026.

In 2023, we published a [position statement on generative AI](#). This paper explored opportunities, risk, emerging good practice and advice for users. In 2025, we also [published a snapshot as part of the Tech Trends converge series that discusses the convergence of generative AI and child sexual exploitation and abuse](#).

We continue to consider the impact of AI on a range of evolving technologies, such as recommender systems and algorithms, including their associated risks, harms, benefits and opportunities.⁶

⁶ In May 2026, eSafety published a [position statement on recommender systems and algorithms](#) which identifies new and specific risks, such as the impacts and increasing influence of AI in terms of both platform and service approaches and content.

AI and online safety reform landscape

Online Safety Act reform and the digital duty of care

eSafety contributes strategic perspectives to online safety law reform, informed by our regulatory and enforcement experience, policy and tech trends analysis, including in relation to AI.

In 2024, the Online Safety Act was subject to an independent statutory review. The [final report](#) was published in October 2024, and in April 2026, the Government published its [response to the Online Safety Act Review](#) with a commitment to implement, or further consider, 64 of the 67 recommendations.

Legislating a Digital Duty of Care (**DDOC**) under the Online Safety Act is a stated high priority for government. While details of the DDOC legislation are still being finalised, it is expected to place the onus on online service providers to proactively keep Australians safe and prevent online harms at a systemic level. This builds on eSafety's existing work to date to keep Australians safe online.

The DDOC is expected to apply to all industry sections currently covered under the Online Safety Act. This includes AI services, including services that provide standalone AI platforms and products, and services which integrate AI into their existing platforms and products.

The DDOC will also apply to design features, including those which may intersect with AI. This includes recommender systems and algorithms. The Government has announced its '[My Feed, My Way](#)' proposal under the DDOC, which would allow the Minister for Communications to make a legislative instrument requiring platforms to provide users with the option to opt-in or opt-out of a personalised recommender feed. Effective regulation of harms arising from recommender algorithms will require access to technical specifications and platform data on recommender algorithms, testing and auditing capabilities, and strong information-gathering and enforcement powers. eSafety would also need to be resourced to support testing and auditing, either through external expertise or by developing in-house capability.

The Government has also committed to work to restrict access to 'nudify' apps and websites.

As reform to the Online Safety Act is underway, we highlight that there is a critical opportunity to coordinate across government, to ensure AI reforms and interventions across various sectors are consistent, future-proofed, robust, and underpinned by safety to maximise AI benefits. It is critical that broader AI legislative reform, standards, and guardrails are consistent with the development of the DDOC and Online Safety Act reform to minimise regulatory fragmentation.

International engagement on AI safety

We recognise that online safety, including AI safety, is a global challenge that requires a collaborative response to achieve a greater degree of regulatory coherence. Through our international engagement and collaboration, we work with online safety regulators around the world. This includes through the [Global Online Safety Regulators Network \(GOSRN\)](#) and various other [international forums and networks](#). 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.

Conclusion

The risks, harms, and safety impacts of AI must be effectively mitigated to realise its benefits. Strong safeguards – in a whole of government approach to AI regulation that leverages existing laws and regulatory structures to minimise harms – are vital to minimise fragmentation and inconsistency.

While eSafety currently has some regulatory power related to AI companions, chatbots and AI generated content with the potential to expand this remit more broadly under Digital Duty of Care, ultimately it is through a nationally coordinated, whole of government approach that we can maximise AI benefits while ensuring AI risks and harms are minimised to keep Australians safe. A strong regulatory outlook, along with a Safety by Design approach, build the foundation for trust, innovation and productivity.

eSafety is pleased to have the opportunity to contribute to the Committee's inquiry. We are happy to provide any further information that would be of assistance.

References

- Abid, M., & Farooqi, J. Z. (2021). Large language models associate Muslims with violence. *Nature Machine Intelligence*, 3 June.
https://www.researchgate.net/publication/352493244_Large_language_models_associate_Muslims_with_violence
- American Psychological Association. (2025a). Artificial intelligence and adolescent well-being: An APA health advisory. <https://www.apa.org/topics/artificial-intelligence-machine-learning/health-advisory-ai-adolescent-well-being>
- Common Sense Media. (2025). *AI risk assessment: Social AI companions*.
<https://www.commonsensemedia.org/articles/social-ai-companions-0?gate=riskassessment>
- Internet Matters. (2025). Me, myself & AI: Understanding and safeguarding children's use of AI chatbots. <https://www.internetmatters.org/hub/research/me-myself-and-ai-chatbot-research/>
- Neugnot-Ceroli, M. (2026). *Adolescents & anthropomorphic AI: Rethinking design for wellbeing*. iRAISE. <https://parispeaceforum.org/publications/adolescents-anthropomorphic-ai-rethinking-design-for-wellbeing/>
- Robb, M. B., & Mann, S. (2025). *Talk, trust, and trade-offs: How and why teens use AI companions*. Common Sense Media. <https://www.commonsensemedia.org/research/talk-trust-and-trade-offs-how-and-why-teens-use-ai-companions>
- Zhang, Renwen, Han Li, Han Meng, Jinyuan Zhan, Hongyuan Gan, & Yi-Chieh Lee. (2025). The dark side of AI companionship: A taxonomy of harmful algorithmic behaviors in human-AI relationships. *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*. <https://dl.acm.org/doi/10.1145/3706598.3713429>