eSafety Commissioner Submission

Inquiry into the use of generative artificial intelligence in the Australian education system

July 2023

Introduction

The eSafety Commissioner (eSafety) welcomes the opportunity to provide a submission to the House of Representatives inquiry into the use of generative artificial intelligence (AI) in the Australian education system.

As Australia's independent regulator and educator for online safety, eSafety aims to safeguard Australians from the risk of serious online harms that can damage their social, emotional, psychological, financial or even physical safety. Our objective is to promote safer, more positive online experiences through a focus on our strategic priorities: prevention, protection, proactive and systematic change, and by developing partnerships to amplify our reach within Australia and around the world.[1]

eSafety provides national leadership in the field of online safety education through our rigorous research, extensive evidence-based programs and resources, and engagement with key stakeholders including government and non-government education bodies, external education providers, students, teachers, and school leaders.

eSafety closely monitors new and emerging technology trends and challenges, and advocates for greater transparency and accountability in efforts to address online safety issues. We also work with local and international stakeholders to examine new research, policy and legislative developments, and provide resources and tools for educators, industry and the Australian community.

The *Online Safety Act 2021* (OSA), eSafety's enabling legislation, establishes our regulatory functions, including administering complaints and investigations schemes for four types of online harms: cyberbullying of children, cyber abuse of adults, the non-consensual sharing of intimate images, and illegal or restricted online content. It also provides eSafety with powers to regulate digital platforms, including through the establishment of industry codes and standards. Further detail on eSafety's role, objectives and measures of success can be found in our 2023-24 corporate plan and our 2022-25 strategy.

Consistent with eSafety's remit, this submission is focused on online safety considerations for the use of generative AI technologies in Australia's education system. We note that other organisations' submissions are focused more specifically on wider issues such as the implications of generative AI for curriculum, pedagogy, and assessment, consistent with their respective remits and expertise.

[1] eSafety Commissioner, *eSafety Strategy 2022-25*, https://www.esafety.gov.au/about-us/who-we-are/strategy (accessed 14/07/2023)

Importantly, however, eSafety does assert that a first principle for supporting students to become safe and responsible users of technology, is that they must have opportunities to learn about, develop skills, and practise using emerging technologies, including generative AI, in appropriate contexts with the guidance of knowledgeable, well-trained teachers.

For the purpose of this submission, eSafety has adopted the definition of generative AI used in the Department of Industry, Science and Resources' June 2023 discussion paper *Safe and responsible AI in Australia*: 'the use of machine learning approaches to generate new code, text, images, audio, video, and simulations'[2].

We trust this submission will be of value to the Committee and would be happy to share any additional information. eSafety is currently developing a detailed position statement on generative AI, as part of our 'tech trends and challenges' series, scheduled for publication in August 2023.

Consideration 1: the strengths and benefits of generative AI tools for children, students, educators and systems

eSafety takes a technology-positive approach, recognising its immense potential to promote productivity, wellbeing and quality of life. We balance this with an evidence and strengths-based approach to identifying, avoiding, and managing risks.

eSafety believes that children and young people must learn about and use generative AI technology in age and developmentally appropriate ways, both in formal school settings and informally with their peers and families. Strength-based education helps to build preventative strategies and resilience so that young people are equipped to safely use current and emerging technologies.

Conversely, simplistic bans or excessive restrictions on the use of emerging technologies, including generative AI, in the classroom have the potential to put students at greater risk of harm through a lack of awareness and understanding of how these technologies work. This approach can also leave students ill-prepared to contribute to the digital economy or apply these technologies effectively in their personal lives.

It is also important to note that students' overall safety and wellbeing is a key factor for effective learning. On this basis, ensuring the safe use of generative AI technologies in the classroom can be expected to have flow-on benefits for student learning.

[2] Department of Industry, Science and Resources, *Safe and Responsible AI in Australia*, https://consult.industry.gov.au/supporting-responsible-ai (accessed 20/07/2023)

eSafety welcomes the announcement of a National Artificial Intelligence (AI) Taskforce 'to develop an evidence-informed framework for Australian education systems and schools to guide the use of generative AI tools'.[3]

Consideration 2: the future impact of generative AI tools on the education workforce

Like all education technologies, AI-enabled tools have the potential to impact educators both positively and negatively, depending on how they are designed and used. One area of concern with respect to online safety, and where eSafety has already received reports from the public, is online abuse involving the education workforce. Educators perform their work in the public eye and can be at greater risk of online abuse from students, parents and the wider community. Examples of abuse experienced by school staff include being filmed without consent and having that footage distributed online for the purpose of humiliating them or damaging their reputation. Female staff are at particular risk of sexualised abuse.

eSafety has received a number of reports of abuse, including reports involving digitally manipulated image-based abuse. These reports do not yet include synthetic images created with generative AI tools. However, there is a risk that generative AI technology may be used to create more sophisticated and realistic images, or do so at larger scale, as the technology becomes more widespread.

A deepfake is a digital photo, video or audio file of a real person that has been manipulated to create an extremely realistic but false depiction of them doing or saying something that they did not actually do or say. Deepfake images created with artificial intelligence software have required multiple still and video images of the target to produce a convincing likeness, which to date has largely limited its use to high profile individuals and celebrities. Generative AI tools may allow bad actors with malicious intent to create authentic looking, high quality deepfakes of individuals without requiring this repository of images to build on. The ability to produce deepfakes more easily and at scale could result in serious and widespread harm to educators.

Educators can report seriously harmful content, including deepfakes, to eSafety. To date, eSafety has had a high success rate in removing this content in partnership with technology platforms.

[3] Education Minister's meeting communique July 2023, https://www.education.gov.au/education-ministers-meeting/resources/education-ministers-meeting-communique-july-2023 (accessed 11/07/2023).

Consideration 3: risks and challenges in the safe and ethical use of generative AI tools

While generative AI tools have a range of educational applications, eSafety is concerned with the potential safety risks for children and young people. We have outlined below several use cases and related online safety risks that have been raised by our stakeholders. Prevention and mitigation of these risks requires a multi-faceted response, as outlined in the final section of this submission.

Online safety risks of generative AI chatbots

Generative AI chatbots present to the user as having 'human-like' qualities and mimicking common conversational traits. For example, chatbots commonly have names that suggest a level familiarity with the user, such as buddy or friend, or that imply a personal or trusted relationship, for example Snapchat's 'My AI' chatbot. Chatbots also seek to maintain user engagement by issuing prompts that seek a response and mimic the behaviour of a 'good listener'.

Crucially, chatbots also present to the user with a high level of authority, expertise and competency. The user of the chatbot often has no way of knowing the limits of knowledge of the application and whether the sources it was trained on are reputable and reliable. This presents acute issues particularly when dealing with children and young people, or with people who may have limited linguistic or intellectual capabilities.

eSafety has concerns about the use of chatbots in education settings, including serious risks to the safety of children and young people. These include:

- age-inappropriate conversations or content, including sexual or violent content, when children and young people engage with a chatbot. This includes where no age assurance is present, where a user can enter an incorrect age, or where the display of content is inadvertent (e.g., a failure of an age estimation algorithm).
- chatbots present with a level of expertise and authority that may not reflect reality. Whereas professionals who work with children and young people are required to complete professional training and accreditation and observe child safeguarding regulations, these requirements do not necessarily apply to chatbots engaging with children and young people. Tech companies have a responsibility to ensure that chatbots operating with children or young people have appropriate safeguards in place to ensure there is no intentional or accidental risk of harm to children and young people.
- a related issue concerns the capacity of chatbots to appropriately identify, respond to and report concerns for the safety and welfare of children and young people. This may include seeking help or making disclosures about experiences, events of circumstances

impacting their safety, health, mental health or wellbeing. If a chatbot fails to assess a circumstance accurately and refer the child or young person to an appropriate support, there may a significantly increased risk of harm (this can include identifying when not to notify a nominated contact person where they may be the perpetrator).

A further concern is the potential for deliberate malicious use of generative AI chatbots by bad actors. eSafety is concerned about the potential for chatbots to be used as a tool for grooming by starting conversations through social media or gaming platforms to manipulate children and young people. The ability to do this at scale – rather than a perpetrator having to directly participate in a conversation – would constitute an alarming development. This demonstrates the need for action in the design of platforms, to anticipate, detect and eliminate the risk upfront, and to build the understanding of children and young people to identify and respond appropriately to grooming behaviours, whether generated by a human or AI.

While acknowledging the risks that generative AI chatbots present, it is important to also focus on the potential benefits they can offer for personalising and targeting education initiatives and scaling up support services. A chatbot can provide relevant and timely information, such as physical and mental health and wellbeing advice, offer referral services or assist with reporting harm and abuse at any time of the day and regardless of staffing or resource limitations. However, it is essential that these functions are designed with similar protections that are currently required of practitioners in relevant professions.

We note there are several generative AI chatbot pilots currently underway in Australia, including a University of Tasmania project (funded by eSafety's Online Safety Grants Program) to support adolescents in dealing with the risks posed by cyberbullying, grooming and image-based abuse, and a trial underway in South Australia[4]. eSafety has also consulted with Kids Helpline who have engaged with a range of international organisations that are already harnessing AI technology to scale up their support services. These include Kids Help Phone in Canada, AinoAid in Tanzania, and Switzerland based Privately.

Cyberbullying, abuse and deep fakes

Generative AI's capability to produce 'human-like' interaction combined with novel high quality personalised content could lead to an amplification of existing cyberbullying and cyber abuse harms. Generative AI tools are being integrated into existing search engines and productivity software that are used in the education system, such as Microsoft 365, Bing and Google Workspace and Bard. The minimum age requirements of 13+ years outlined in the terms of use of companies like OpenAI are unlikely to provide adequate protection

[4] InnovationAus.com, *SA public schools trial 'safe' AI-powered chatbot in nation first*, https://www.innovationaus.com/sa-public-schools-trial-safe-ai-powered-chatbot-in-nation-first/ (accessed

for young children. eSafety continues to see reports of serious cyberbullying from children as young as eight on social media platforms despite 13+ minimum age requirements.

eSafety has not yet received any reports of cyberbullying that involve generative AI. However, it is important to acknowledge there have been upward trends in the number of reports received through the children's cyberbullying reporting scheme.

Generative AI, without safeguards and used with malicious intent, may enable the creation and dissemination at scale of seriously harmful, authentic looking and targeted abuse in the form of text, imagery, audio and video. A particular area of concern is the creation of synthetic media or deepfakes. As noted above, the creation of this material currently requires manual effort and a body of source images to create a convincing likeness. Generative AI may provide a user access to a more powerful, automated and responsive set of tools. For example, the US-based non-profit Thorn has noted the potential for generative AI tools to generate realistic computer-generated child sexual abuse material.[5]

Australia's AI Ethics Framework outlines principles that are designed to ensure AI is safe, secure and reliable. The first principle – human, societal and environmental wellbeing – states that AI systems should benefit society and the environment.[6] This aligns with eSafety's Safety by Design approach which aims to foster more positive, civil and rewarding online experiences for everyone.

There continues to be an important role for age-appropriate preventative education incorporating respect and personal responsibility when using online tools, including generative AI technologies. However, online platforms must also bear a responsibility to protect users and prevent generative AI abuse by putting user safety and rights at the centre of the design and development of products and services. eSafety's forthcoming generative AI position statement will explore emerging good practices and Safety by Design measures for generative AI technologies in more detail.

Bias and vulnerable groups

While all users of generative AI should be protected from harm, eSafety's research shows that certain groups are especially vulnerable to the impact of online harms and experience higher rates of online targeted abuse.[7] Recent research has shown for example that Aboriginal and Torres Strait Islander children and young people are using technology in

---

[5] THORN, *Generative AI: Now is the time for Safety by Design*, https://www.thorn.org/blog/now-is-the-time-for-safety-by-design (accessed 20/07/2023)
[6] Department of Industry, Science and Resources, *Australia's AI Ethics Principles*, https://www.industry.gov.au/publications/australias-artificial-intelligence-ethics-framework/australias-ai-ethics-principles (accessed 14/07/2023)
[7] eSafety Commissioner, *Protecting voices at risk online*, https://www.esafety.gov.au/communities/protecting-voices-risk-online (accessed 19/07/2023)

positive ways for cultural expression and engagement. However, they are almost three times more likely than the national average to have had offensive things said to them because of their race, ethnicity, gender, nationality, sexual orientation, religion, age or disability.[8]

Generative AI tools may replicate any existing stereotypes, biases and discriminatory viewpoints contained in their training datasets. This has led to concerns that generative AI could reinforce discrimination and negatively impact already marginalised sections of the community. As more AI generated content is created and becomes part of the training dataset, this may create a feedback loop to further exacerbate discrimination and amplify targeted abuse. The CSIRO's Data 61 has compiled a repository of biases drawn from real life instances and a toolkit for organisations to identify potential biases and implement mitigation strategies throughout the AI lifecycle.[9] These can include curating training materials to reflect a diversity of communities, values and perspectives and developing content detection and moderation tools.

While mitigation is possible using a Safety by Design approach that involves ongoing testing of AI systems for safety, there is a risk that discriminatory or biased output will not be eliminated altogether. AI and algorithmic literacy and other core competencies (discussed further below) will be important skills for children and young people to identify and understand how biases can be replicated when using generative AI tools and how to challenge discriminatory or exclusionary online content.

Consideration 4: How cohorts of children, students and families experiencing disadvantage can access the benefits of AI

The level of digital literacy and equitable access to technology has a significant impact on the ability of Australians to keep themselves safe online. Children, students and families who experience disadvantage are more likely to miss out on the benefits of new technology, and the opportunity to develop the skills to protect themselves when online.[10]

As an example, research with recent migrants and refugees, including ongoing research by eSafety, has found children typically have greater digital literacy than their parents, limiting

---

[8] eSafety Commissioner, *Online experiences of Aboriginal and Torres Strait Islander children and their parents and caregivers*, https://www.esafety.gov.au/research/online-experiences-aboriginal-torres-strait-islander-children-parents-caregivers (accessed 05/07/2023).

[9] CSIRO Data 61, *Diversity and Inclusion in Artificial Intelligence*, https://research.csiro.au/ss/team/diai/ (accessed 05/07/2023).

[10] Thomas, J, Barraket, J, Wilson, CK, Holcombe-James, I, Kennedy, J, Rennie, E, Ewing, S, MacDonald, T, 2020, *Measuring Australia's Digital Divide: The Australian Digital Inclusion Index 2020*, https://doi.org/10.25916/5f6eb9949c832

parents' ability to support their children to go online safely.[11] [12] The negative consequences of the digital divide may be further exacerbated by generative AI's documented tendency to reinforce biases, for example racial or gender-based discrimination.

While it is necessary to address these risks as part of a suite of measures to improve equity (including digital equity) across education more broadly, it should also be noted that AI has the potential to be a facilitator of access rather than a barrier to it. For example, UNESCO has noted that AI systems such as Google Voice Assistant at the Global Digital Library, Dytective and StorySign provide a range of accessibility benefits for people with disability (literacy difficulties, dyslexia and deafness, respectively). As UNESCO's AI and education guidance points out, core competencies will not just be about learning with AI but learning about AI techniques and learning for a world of human-AI collaboration.[13]

With a continued focus on digital literacy, AI technologies may enhance and improve access to online safety education for those experiencing disadvantage. eSafety promotes digital literacy, critical thinking and resilience as part of its education and family programs. The eSafety Guide aims to help families navigate the safety features of new and emerging technologies. The eSafety Guide has recently been updated to include information on Chat GPT and GPT-4, Google Bard, Bing and Snapchat's My AI chatbot. eSafety will continue to explore innovative ways of delivering online safety education and awareness programs including through emerging technologies.

eSafety's research team is currently developing questions on algorithmic literacy to include in our 2024 youth survey. This research will inform our online safety programs and contribute to the international evidence base on children and young people's digital literacy.

Consideration 5: International and domestic practices and policies in response to the increased use of generative AI tools in education

eSafety is aware that a range of policies and practices are being developed and adopted internationally in response to the rapid escalation in availability and use of generative AI tools. We recommend the Committee refer to the Department of Industry, Science and Resources' *Safe and responsible AI in Australia* discussion paper, which includes a comprehensive summary of international developments in regulating AI technologies.[14]

---

[11] Kenny, Edmee, 2016, *Settlement in the digital age: Digital inclusion and newly arrived young people from refugee and migrant backgrounds*, Centre for Multicultural Youth (CMY)

[12] Worrell, S., 2021, *From Language Brokering to Digital Brokering: Refugee Settlement in a Smartphone Age*. Social Media + Society, 7(2). https://doi.org/10.1177/20563051211012365

[13] UNESCO, AI and education, guidance for policy-makers, https://doi.org/10.54675/PCSP7350, 2021, p13 (accessed 06/07/23)

[14] Department of Industry, Science and Resources, *Safe and Responsible AI in Australia*, https://consult.industry.gov.au/supporting-responsible-ai (accessed 20/07/2023)

Examples of frameworks and guidance on AI technology being developed for educators include publications by UNESCO[15] and Common Sense Media, a US-based non-profit organisation, which is developing a ratings and review framework for AI products to assess their safety, privacy, and suitability for children and young people, with a key focus on education applications.[16]

eSafety also notes work is underway in Australia, led by the National Artificial Intelligence (AI) Taskforce established by Education Ministers, to develop an evidence-informed framework for the use of generative AI in Australian education systems.[17]

Consideration 6: Recommendations to manage the risks, seize the opportunities, and guide the potential development of generative AI tools

eSafety recommends a range of prevention and mitigation strategies to address the online safety risks of generative AI, and amplify the benefits, outlined above. These are organised under our three strategic pillars described in the introduction to this submission.

Prevention

Reducing the likelihood of online and tech-related harms from occurring is critical. As noted, the most effective and enduring way to mitigate risks is via a strengths-based approach – building on existing knowledge, capabilities, and student agency – to identify and address risks before they can cause harm. In the context of the Australian Curriculum, this includes a focus on both technical skills and social and emotional learning (SEL), including AI and algorithmic literacy, media literacy, and Digital Literacy and other General Capabilities.

eSafety recommends a multifaceted approach to addressing the challenges of generative AI technology that includes:

- Ongoing review and adaption of curriculum materials, including the Australian Curriculum and supporting materials, to explicitly teach the technical knowledge and social and emotional skills to safely use new and emerging technologies. This could include reviewing the Online Safety Curriculum Connection to address specific AI-related scenarios.

---

[15] UNESCO, *AI and education: guidance for policy-makers*, https://doi.org/10.54675/PCSP7350 (accessed 14/07/2023)
[16] Common Sense Media, *Common Sense Media Announces New Ratings and Reviews System for AI Products*, https://www.commonsensemedia.org/press-releases/common-sense-media-announces-new-ratings-and-reviews-system-for-ai-products (accessed 21/07/2023)
[17] Education Minister's meeting communique July 2023, https://www.education.gov.au/education-ministers-meeting/resources/education-ministers-meeting-communique-july-2023 (accessed 11/07/2023).

- Continuing to raise awareness of eSafety's Toolkit for Schools and Best Practice Framework for Online Safety Education so that schools and education sectors are better prepared to support students, families and communities, and periodic review of these resources to ensure emerging technologies like generative AI are accounted for.
- Delivering relevant and engaging education programs and resources for, and in consultation with, priority audiences – including children, young people and families, First Nations communities, culturally and linguistically diverse communities, and LGBTIQ+ Australians – and professional learning for educators.
- Partnering with our education stakeholders to raise awareness and build community understanding of generative AI and associated online safety risks. This includes through our National Online Safety Education Council and eSafety Youth Council, and relationships with government agencies, education bodies and providers of online safety education.

Protection

eSafety has a range of regulatory powers under the OSA that we can use to help protect Australians who have experienced online harms. We can remove abusive and harmful content, limit the ability of perpetrators to continue their abuse, and take enforcement action against those who fail to comply with regulatory notices.

eSafety's four complaints-based investigations schemes capture AI-generated images, text, audio, and other content that meets the respective legislative definition of:

- class 1 material (such as child sexual exploitation material and terrorist and violent extremism content) and class 2 material (such as pornography)
- intimate images produced or shared without consent (sometimes referred to as 'revenge porn')
- cyberbullying material targeted at a child
- cyber abuse material targeted at an adult.

Under these schemes, eSafety can provide support to complainants, including assisting in the removal of certain content and providing guidance to minimise the risk of further harm.

The Australian Government has announced that the OSA will be independently reviewed within this term of Government.

Proactive and systemic change

Under this third pillar, eSafety works to bring about a fundamental change in the way that technology products are designed, placing safety of users at the core – not as an

afterthought. We collaborate with national and international partners to develop consistent approaches to shifting the responsibility back onto the tech sector to assess risks and incorporate safety into development processes – for example, through eSafety's Safety by Design initiative.

Safety by Design is a voluntary initiative that encourages industry to anticipate potential harms and implement risk-mitigating and transparency measures throughout the design, development and deployment of a product or service. This includes providing free risk assessment tools and good practice guidance to help companies build in safety features and provide positive online experiences.

Safety by Design should be applied to all AI products and services from the earliest stages of design and throughout their lifecycle. Based on eSafety's recent expert consultations, this could include ensuring:

- generative systems are sourcing high-quality data and information which has been cleaned of illegal, exploitative, and otherwise harmful material
- policies and processes prevent users from generating harmful content
- watermarks and detection tools are used to identify AI-generated materials
- features are evaluated to identify and mitigate risks for diverse user groups
- companies design clear reporting mechanisms and well-defined triage and escalation processes
- system and model cards are used to promote the improvement of models and the enhancement of their understanding by regulators, researchers, and the public.

The OSA provides for the development of industry codes to cover eight sections of the online industry (including search engine providers). Under this co-regulatory model, the online industry is to develop measures to deal with class 1 and class 2 content, and for eSafety to register such codes. If industry codes do not meet the registration requirements, eSafety may determine an industry standard (a regulatory instrument). The risk associated with generative AI technology and its application in search engines was a key reason for the eSafety Commissioner reserving her decision on the draft Internet Search Engines Services code.[18] In response to eSafety's request to re-draft the code, the relevant industry associations submitted a revised code to eSafety for assessment on 6 July 2023.

The OSA also empowers eSafety to require a range of online services to report on the reasonable steps they are taking to comply with the Government's Basic Online Safety

[18] eSafety Commissioner, *eSafety Commissioner makes final decision on world-first industry codes*, https://www.esafety.gov.au/newsroom/media-releases/esafety-commissioner-makes-final-decision-on-world-first-industry-codes (accessed 18/07/2023)

Expectations (BOSE). This is intended to enhance transparency and accountability, and to ensure people can use their services in a safe manner. eSafety has issued 13 reporting notices since August 2022, requiring companies to report on the steps they are taking to implement the BOSE. A report summarising the responses from the first seven notices was published in December 2022.[19] In the future, eSafety could require other service providers to report on the reasonable steps they are taking to ensure the safety of their generative AI functionalities.

eSafety can provide further information to the Committee about our prevention, protection, and proactive and systemic change initiatives upon request.

[19] eSafety Commissioner, *Basic Online Safety Expectations: Summary of industry responses to the first mandatory transparency notices*, https://www.esafety.gov.au/industry/basic-online-safety-expectations/responses-to-transparency-notices (accessed 18/07/2023)