Tech Trends Position Statement

Generative AI

Australian Government

eSafetyCommissioner

eSafety.gov.au

Tech Trends Position Statement – Generative AI

Contents

Executive summary

The rapid evolution and rise of generative AI systems is reshaping industries and human creativity. While generative AI offers novel opportunities, it can also amplify a range of existing and emerging harms for individuals and society. For example, we have already seen chatbots providing inappropriate and harmful responses to user prompts, the spread of hyper realistic generative AI deepfakes, and the creation of synthetic child sexual abuse material. Balancing the potential benefits with the risks of generative AI is essential.

This position statement examines the evolving landscape of generative AI, providing an overview of the generative AI lifecycle, examples of its use and misuse, and consideration of online safety risks and opportunities. The statement also sets out a range of regulatory challenges and approaches. The final section highlights emerging good practice and new Safety by Design measures to provide industry with meaningful, actionable and achievable guidance to minimise existing and emerging generative AI harms.

Overview of eSafety's approach to tech trends

The eSafety Commissioner (eSafety) is Australia's independent regulator and educator for online safety.

Under the *Online Safety Act 2021* (Cth) ('the Act'), we coordinate Australian Government activities to help keep people safer online, conduct research, provide education, and administer regulatory schemes to deal with certain types of online harm. We use our regulatory powers to promote greater transparency and accountability within the online industry.

We work with other government agencies, businesses and organisations around the world to share information and best practices. This helps us make the internet a safer place for everyone, regardless of where they live.

We keep our content, programs and regulatory priorities up to date by scanning for new research, policies, laws, technology developments and by talking to experts such as academics and researchers.

eSafety also advises the Australian Minister for Communications and the Government on emerging issues across the online industry, international developments in technology regulation, and online safety concerns impacting Australians.[i] We do this because we recognise that combating online harm is a global challenge and we need to act together to make a difference.[ii]

This position statement is about **generative artificial intelligence (AI)** but readers may also find our other position papers on **deepfakes**[iii] created using artificial intelligence software, as well **recommender systems and algorithms,**[iv] useful information relevant to this topic.

The information in this position statement was informed by industry and stakeholder consultation as well as Australian and overseas research. It reflects eSafety's position as of **15 August 2023.** eSafety acknowledges the rapid advancements in generative AI technology and will seek to review and provide revisions when necessary.

Definitions and examples

Artificial intelligence (AI) refers to an engineered system that generates predictive outputs such as content, forecasts, recommendations, or decisions for a given set of human-defined objectives or parameters without explicit programming. AI systems are designed to operate with varying levels of automation.

Machine learning are the patterns derived from training data using machine learning algorithms, which can be applied to new data for prediction or decision-making purposes.

Generative AI models produce novel content such as text, images, audio, video and code in response to prompts.

A **large language model (LLM)** is a type of generative AI that specialises in the generation of human-like text.

Multimodal Foundation Model (MfM) is a type of generative AI that can process and output multiple data types (e.g. text, images, audio).

For consistency, this paper adopts the same definitions used in the Department of Industry, Science and Resources' June 2023 Discussion Paper on Safe and responsible AI in Australia.

What is Generative AI?

Generative AI uses machine learning to generate new code, text, images, audio, video, and multimodal simulations. It works by using large artificial neural networks[v] built with enormous datasets and parameters that are inspired by synapses within the human brain. The difference between generative AI and other forms of AI is that its models can create new outputs, instead of just making predictions and classifications like other machine learning systems.

Some examples of generative AI applications include:

- Text-based chatbots, or programs designed to simulate conversations with humans, such as Anthropic's Claude, Bing Chat, ChatGPT, Google Bard, and Snapchat's My AI
- Image or video generators, such as the Bing Image Creator, DALL-E 2, Midjourney, and Stable Diffusion
- Voice generators, such as Microsoft VALL-E.

Background

Generative AI is not new. Chatbots, image generators and deepfake technologies have been in development and use for many years.

However, recent advancements have rapidly improved generative AI due to the availability of more training data, enhanced artificial neural networks with larger datasets and parameters, and greater computing power. Some experts now claim contemporary AI systems are moving rapidly towards 'human-competitive intelligence.'[vi] Such claims pose existential questions about the potential of such systems to impact almost every aspect of human life in both positive and negative ways.

The possible threats related to generative AI are not just theoretical – real world harms are presenting themselves today. This includes misusing AI to generate child sexual exploitation and abuse (CSEA) material that looks like it involves real

children (or based on images, audio or other depictions of real children[vii]) or generating and threatening to share artificial but realistic pornography featuring real adults without their consent.[viii] These harms can occur because of flaws in the data or models used in generative AI, such as when biased information is used for training.[ix] Generative AI can also be used to manipulate and abuse people by impersonating human conversation convincingly and responding in a highly personalised manner.[x]

Many generative AI models have been intentionally made freely available within the open source community or have 'leaked' into the public domain.[xi] While releasing models freely promotes transparency, competition and innovation, the fact it is readily accessible to the public also increases the risk that harmful and manipulative content can easily be generated at scale when the technology is put in the wrong hands.[xii]

Generative AI is being incorporated into major search engines, productivity software, video conferencing and social media services and is expected to be integrated across the digital ecosystem.[xiii] Companies are moving quickly to develop and deploy their own generative AI technologies. This may lead to not enough attention being paid to risks, guardrails, or transparency for regulators, researchers, and the public.

Multiple actors including technology developers and downstream services that integrate or make generative AI technology accessible, as well as users, all have a role to play in ensuring online harm is prevented and addressed.

As countries think about how to regulate generative AI, technology companies have been advocating for certain regulatory approaches, some of which may actually serve the commercial interests of the companies involved.[xiv]

In Australia, the Government is looking at the risks, benefits and potential impacts of generative AI. This includes examinations by the Department of Industry, Science, and Resources, the Department of Education, the Attorney General's Department and the Digital Platform Regulators Forum (DP-REG), which includes the Australian Competition and Consumer Commission (ACCC), Australian Communications and Media Authority (ACMA), Office of the Australian Information Commissioner (OAIC), and eSafety.[xv]

It Is important to recognise that for every risk, there is also an opportunity. For example, people can misuse generative AI to create harmful content such as online hate. However, AI can also be harnessed to significantly improve current proactive content moderation technologies to quickly and accurately find and stop online hate.[xvi]

There have been reported instances of children acknowledging abuse and seeking support through AI chatbots.[xvii] A chatbot can give an inappropriate or harmful response to a child who discloses their experience of abuse. But an appropriately trained chatbot could respond in a supportive and evidence-based manner,

connecting that child to law enforcement and support services. The risk of harm depends on whether the technology was designed with safety in mind, including by taking a Safety by Design approach.[xviii]

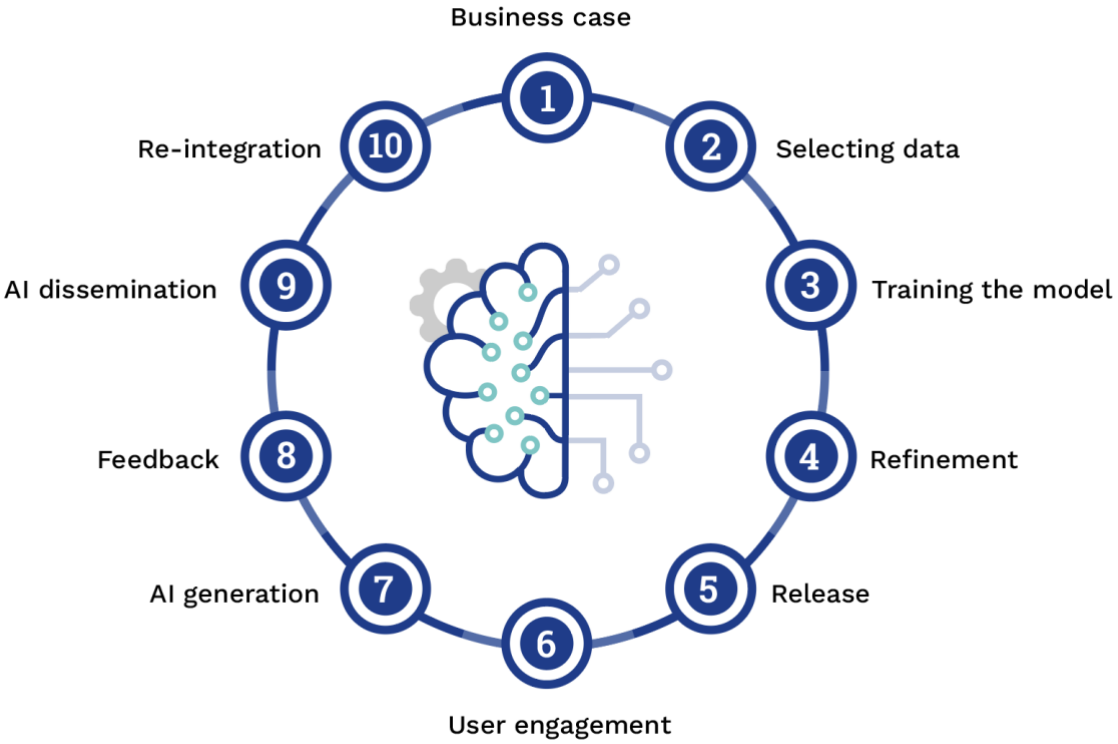
Safety by Design is built on **three core principles**: Service provider responsibility, User empowerment and autonomy, and Transparency and accountability. Technology companies can uphold these principles by making sure they incorporate safety measures at every stage of the product lifecycle. This should involve consulting stakeholders from multiple sectors and collaborating with the user community, including those who are typically under-represented or who may be at greater risk of harm. A Safety by Design approach to AI is also likely to satisfy most of Australia's AI Ethics Principles.

Generative AI lifecycle

It is important to consider online safety risks and harms from the earliest stages of developing a generative AI technology. This should continue throughout the technology's lifecycle and across the entire system from developing a business case to releasing, disseminating and reintegrating AI-generated content.

The simplified product lifecycle below sets out 10 crucial steps where this must occur, drawing on insights from various experts and other sources.[xix]

Diagram 1: Generative AI Lifecycle



1. Business case

The first step in the generative AI lifecycle is to evaluate the business case for developing the technology and explore options for funding. To create safer technologies, companies and developers building generative AI should consider *why* they are planning to develop it, for *what purpose* and in *what context*.

AI systems designed for legitimate internal business purposes can still have broader impacts on individual, social and environmental wellbeing. Those impacts should be accounted for in the AI system's lifecycle, to include consideration of impacts outside the organisation**.**

By considering risks and building in safeguards during the early stages of development, it is possible to establish trust in a product or service.[xx] This can then open up more opportunities for investment. One helpful tool for this process is the **Safety by Design Business Model Canvas**. This enables businesses to assess and analyse their business model, challenge assumptions and promote socially responsible innovation.

2. Selecting data

After establishing a business model, developers must make choices about the type of model they want to create and the input data they will use to build and train it. Generative AI often requires large datasets to meet the diverse needs of users with some models utilising closed datasets, while others rely on general information scraped from the internet.

Developers must consider the content and quality of their data sources, as well as ethical and legal concerns such as *how* data is sourced, right from the start. Data scraping involves collecting, using, disclosing, and storing information without the knowledge or consent of the data creators or the individual the information is about, which raises questions about copyright, privacy, consent and attribution.

It is important to address considerations about accuracy, diversity (including language and culture) and whether harmful material is captured through processes such as scraping. If not managed carefully, there is a risk that data sources could include illegal or harmful content, such as CSEA material, image-based abuse (IBA), hate speech and abuse, or false, biased, or misleading information, or other unlawful material. Data sets containing such content can perpetuate harms by generating illegal and harmful outputs.

Developers can also use pre-training capabilities, such as classification and proactive detection tools, to improve training data quality. This reduces the risk of harm later in the lifecycle.

Transparency is a vital component to hold services accountable for content they host or use to train their systems, including by documentation through annual

reports and using model cards or system cards, which are designed to explain how the systems and models operate.

3. Training the model

The next step is to train the model using the data that has been selected. Developers can do this through supervised learning, where humans train models to classify inputs with labels.

For example, a model can be trained to label social media posts as positive or negative. This is called 'supervised learning' because a human teaches the model what to do.[xxi] It is important that humans are appropriately trained to conduct this work and feed in diverse views. On the other hand, 'unsupervised learning' can be used to find patterns in data that are not labelled, typically used when there is a lack of training data.[xxii]

More advanced text-based machine learning models may rely on 'self-supervised learning'. This type of training involves giving the model a massive amount of text so it can generate predictions. For example, some models can predict how a sentence will end based on a few words.[xxiii] Model and system cards are important for documenting capabilities at the training stage, prior to refinement and release.

It is also important at this stage to consider the lived experiences of humans who are training the model, to ensure that culturally specific and contextual forms of harm can be addressed and appropriately mitigated.

Additional safety measures include consultation with experts who can provide guidance on inputs for training the model.

4. Refinement

It is important to keep refining model data throughout the lifecycle to minimise risks, harms and bias. This means going beyond initial training with supervised, unsupervised or self-supervised learning.[xxiv.] Developers must keep working over time to maintain quality data inputs, ethically curate data, label it, and control quality across many datasets on different subjects.[xxv]

Data quality and veracity are issues in the refinement process, as is the ability of AI models to recognise and filter out illegal, harmful or inappropriate content. This work is often carried out by employees who are hired to tag and sift through large amounts of harmful and potentially traumatising content.

While human review prior to release may be essential, there are concerns about the working conditions, pay, and mental wellbeing for people who do jobs such as labelling or generating training data.[xxvi]

5. Release

Following refinement, the model may be released to the public through the developer's own app or interface, such as ChatGPT. It can also be added to

through other means including integration into an existing service, such as ChatGPT in Microsoft's Bing search engine.

The same model can be released and integrated in a variety of ways. Developers may also choose to openly release their model. There are details about open and closed models below.

It is important to conduct risk assessments, anticipate how the model may be misused by individuals and establish safety policies and practices prior to release. Developers should 'red-team' or 'stress test' the system by considering the avenues for possible misuse. Developers should also consider graduated approaches to release, including regulatory sandboxes, to understand how the model performs in controlled conditions.

Given the evolving nature of the technology, unforeseen risks and new techniques to overcome safeguards will keep appearing after the model is launched.

6. User engagement

Once a model is released, users can interact with it by accessing its interface and giving it instructions or prompts. For example, they can enter text or audio commands to generate content or get information.

Developers should expect that their model might be misused by malicious actors. They should test their models with consideration of the ways it could be misused. For example, it is a serious concern if models are not able to detect when users may be attempting to input harmful prompts to generate illegal or harmful content, or implementing appropriate safeguards that are activated to mitigate potential misuse.

For example, terrorist groups could use models to raise money, disseminate pro-terror content or generate instructions on making bombs or weapons[xxvii]; paedophiles could use AI to create content for child grooming or CSEA, and people could use AI to generate and spread misinformation and disinformation or targeted hate speech or abuse. People could also intentionally try to hack or tamper with the model's input to make it behave badly.[xxviii]

Adding points of friction, such as educative prompts and nudges, when users attempt to generate content can be an important method of reducing misuse.[xxix] Developers need to keep improving their model to engineer out harmful or illegal outputs as they emerge, as it is unlikely all harms will be mitigated prior to release.

7. AI generation

After the user inputs a prompt, the AI interface generates content based on this information. Sometimes, generative AI models give confident but inaccurate, misleading, or harmful answers. These are called 'hallucinations' and can happen for many reasons, including inadequate or problematic input datasets.[xxx]

Model outputs can also adversely influence user views, values and experiences by misrepresenting available information or only providing a limited view of information. This could have the impact of shifting societal norms or values around challenging topics. Developers can also add safety measures at this stage, such as warnings or disclaimers for users that the information might be wrong or inaccurate. Digital watermarking – a method for identifying AI-generated content – can also be implemented.

8. **Feedback**

Offering opportunities for users to give feedback on the content generated is a crucial step in the generative AI lifecycle. This can be done through user feedback loops. It is also essential to clearly communicate policies and make sure reporting and feedback tools are easily accessible.

By gathering input from users, there is a chance to mitigate potential risks, such as generating discriminatory, harmful, deceptive, or false content. This may also provide the basis to undertake consultation with a diverse userbase. This feedback helps to implement measures that moderate and improve the content generated by the model.

9. **AI dissemination**

After the system generates content, it can be shared with others, including on social media.

Even where these social media platforms and other services do not have their own generative AI capability, it is imperative to build in tools that can stop, find, and moderate harmful content generated by AI that may be shared on their platforms.

10. **Reintegration**

Generative AI content that is shared on the internet or on social media could feed back into models that are built or refined using content scraped from the web. If not appropriately managed, harmful content and views may be reinforced in a continuous feedback loop. Reintegration may also generate 'synthetic data' and in turn lead to an overall reduction in model efficacy.[xxxi]

Risks, harms, and opportunities

Framing online risks and harms

There are different ways to understand the risks and harms associated with generative AI. One approach is to consider its potential impacts on individuals and society.

At an individual level, generative AI can pose risks by generating and amplifying harmful and extreme content. This can have a greater impact on victims of CSEA material, IBA and other forms of abuse. It also affects those who inadvertently come across such harmful material.

On a broader societal level, generative AI can contribute to the generation and amplification of content that promotes bias and discrimination. This includes promoting sexism, homophobia, racism, or other forms of prejudice.[xxxii] Such content normalises hate or intolerance, which could lead to radicalisation towards terrorism and violent extremism. It may also lead to an erosion of trust in online content or institutions.

These risks have significant social implications, particularly where several harmful effects may accumulate over time, shaping narratives around important societal issues. For example, a person might ask a generative AI application a question about domestic violence and get a response that distorts or minimises the severity of the issue. There is also the potential for text-based and visual model output to be used in the service of mega conspiracies, fuelling hate and intolerance.

Given the potential individual and societal impacts, experts, industry and some governments are developing frameworks to proactively consider these issues, risks and harms.

One framework that was raised during consultations is an approach that examines the risks associated with different components of the system. For example, the 'ABC' framework considers three key aspects: the **actors** involved in disinformation campaigns, their deceptive **behaviour** and tactics, and the **content** they produce and share.[xxxiii]

Similarly, during consultations, eSafety received feedback suggesting an approach that focuses on context and intention. Stakeholders identified three categories of risks and harms:

- **AI failing to perform as expected:** This occurs when a system unintentionally causes harm by generating incorrect or harmful responses. For example, generative AI systems may 'hallucinate' and produce inappropriate responses to user prompts.
- **AI being used maliciously:** This happens when a model is trained or exploited for harmful purposes. For example, when individuals involved in CSEA attempt to groom children or generate CSEA material using generative AI tools.
- **AI being overused, used recklessly or used inappropriately in a specific context**: This refers to situations where generative AI is used excessively or recklessly, or employed inappropriately, leading to harmful or misleading results. For example, where generative AI produces age-inappropriate material such as online pornography for a child user.

Drivers of risk

Several factors can drive, contribute to, or amplify the risks and harms associated with generative AI:

- **Personalisation**. Chatbots and multimodal models have the potential to generate highly personalised, emotive, manipulative, and invasive content based on users' previous engagement and activity. Through consultation with eSafety, experts highlighted that online harms may arise from generative AI creating 'human quality content' or producing customised media in-real time. While this content may appear authoritative, it could also be intentionally or accidentally false, misleading, or malicious.[xxxiv] For example, personalised phishing activities may be used to intentionally mislead and potentially defraud the recipient or gain access to information or systems.
- **Access**. Wider access to generative AI models raises concerns about their potential misuse for harmful purposes. Consumer-facing apps using generative AI make it harder for users to discern fact from fiction as the technology becomes more convincing over time. Policy discussions continue globally on whether the development of AI should occur in open or closed environments, with regulatory approaches tailored according to public access levels and associated risks. Determining who is responsible for preventing and mitigating harms becomes an important consideration among the companies that develop the model, the companies that deploy the model in their applications and the people who use those applications.

Case study: Open vs closed systems

There are arguments for both open and closed systems regarding their benefits for safety and security. Open systems offer interoperability, customisation, and integration with third-party software or hardware. Champions of open models highlight how openness promotes transparency, accountability, competition and significant innovation, whereas advocates of closed systems argue that they are more stable and secure and better protect their owners' property interests.[xxxv]

Choosing between open and closed systems is further complicated by factors such as aligning AI with human values and goals. Finding the right balance between the two models is crucial to foster innovation while managing the short- and long-term risks posed by the technology. Preventing harm is a paramount consideration in the development of the system. Safety by Design principles are critical in technology design, including for open and closed systems. By drawing on Safety by Design, innovation can continue to be fostered without creating or amplifying possible harms.

An open-source model has its merits in democratising access to AI, allowing researchers to identify errors and biases, and mitigating the risk that generative AI is overly concentrated in large tech companies with access to training data and computing power.[xxxvi] However, open-source models can also allow individuals to remove safeguards and create harmful content. For example, classifiers can be fine-tuned to create adult pornography or used by perpetrators of CSEA to produce CSEA material.[xxxvii]

- **Advertising and revenue models, and access to children's data**. Digital platforms with advertising-based revenue models are likely to incentivise the generation of highly personalised content for marketing purposes and promote sponsored content rather than addressing the specific needs of users. It is also possible that generative AI optimised to increase user engagement will produce problematic or emotive content aimed at maintaining attention. Greater consideration of the acquisition, access to, use and storage of children's data, particularly for commercial purposes, is needed. The Australian Privacy Act Review Report made recommendations for providing individuals with greater control over targeted content and marketing, including prohibiting entities from targeting children unless it is in the child's best interest. At the date of publishing this paper, the Government's response to the report is forthcoming.
- **Limited representation.** At the time of drafting this statement there appears to be a lack of diversity among AI companies, primarily originating from English-speaking, Western cultures, which poses a risk of encoding narrow values and perspectives into global-reaching models.[xxxviii] This could perpetuate and amplify dominant ideologies at the cost of other values and identities.
- **Pace of development**. With venture capital increasingly focused on generative AI's rapid product development and sales growth potential comes the risk of neglecting safety considerations due to a 'move fast and break things' approach. Safety by Design recognises that there are important inflection points and players in the technology ecosystem that need to be leveraged to enable change. Investors and venture capitalists play a pivotal role in nurturing tech ventures and they can help put safety and ethical considerations at the heart of the businesses they invest in. This will help them to invest ethically and manage investment risk, but also helps the start-up harden their defences against potential safety risks.[xxxix] It's worth noting that the use of open-source models and pace of change within the developer community can also create problems of control, responsibility and accountability, with models adapted for nefarious purposes without adequate checks and balances.
- **Convergence with other emerging technologies.** As immersive platforms merge with generative AI technology, their respective risks may also converge. This raises important questions about the manifestation of future harms, and the potential for more visceral and extreme impacts. For example, metaverse platforms rely heavily on AI to create realistic, immersive environments populated by non-player characters. These platforms collect large amounts of personal information that informs conversational agents and enhances user engagement.[xl] This convergence may reinforce the risks highlighted in eSafety's position statement on immersive technologies.

Online safety risks and opportunities

Many of the online safety issues associated with generative AI are not new but will likely be impacted, and amplified, by this technology.

Risks

Used to create CSEA material

Generative AI is expected by some to bring about changes in how CSEA occurs online, as well as the methods we employ to combat it.[xli] A 2023 report by the Stanford Internet Observatory and Thorn found that generative AI tools are already being used to create realistic computer-generated child sexual abuse material (CG-CSAM).[xlii] The potential for CSEA materials to be generated using photos of children harvested from social media also creates specific safety challenges for parents, carers and young people, reinforcing the need to make sure that online profiles are set to private.[xliii] Perpetrators can exploit the ability of large language models (LLMs) powered by AI to mimic natural human language. This allows them to groom children in automated and more targeted ways,[xliv] and cases have already been reported where generative AI technologies are being used to facilitate child grooming.[xlv]

Developments related to generative AI pose risks concerning the identification of victims. As it becomes more difficult to determine whether content is AI-generated, law enforcement agencies and hotlines will face a growing challenge in determining whether certain content depicts an actual child who needs to be identified and rescued. There are also definitional challenges which could emerge across jurisdictions concerning AI-generated media, including how children and images of children are defined.

Case study: The production of realistic CSEA material using generative machine learning (ML) tools

Thorn is a US-based non-profit organisation established in 2012 to build technology to defend children from sexual abuse. It has identified the potential for generative AI to change how child sexual exploitation and abuse occurs online.

Through both the input and output stages, LLMs and MfMs can be used to facilitate the development of CSEA material through:

- Prompts intended to produce AI-generated CSEA material. There is emerging concern that prompts could be used to create new images of real children or make explicit imagery of children who do not exist.[xlvi]
- 'Role-playing' with generative AI. This involves exploring how the model can be prompted to behave in certain ways.
- Model outputs that generate harmful or illegal content. These outputs may include CSEA material.

Children seeing violent or sexually explicit material

There are additional risks to child safety related to chatbots and other forms of conversational AI. These technologies can enable inappropriate contact with children and young people. Moreover, generative AI has the potential to generate content that is not appropriate for their age, such as violent or sexually explicit material.

For example, developers have used the open-source Stable Diffusion model to generate realistic adult pornography.[xlvii] While the model has since been updated to mitigate the possibility of unsafe or inappropriate content from being created, some people still use older versions of the model to produce prohibited imagery. Many open source sites also continue to provide access to build-your-own prompts in order to produce photorealistic pornography.

Encouraging or facilitating behaviours that negatively impact wellbeing and safety

There are reports that Snapchat's 'My AI' chatbot offered advice to a user pretending to be 13 years old on how to lie to her parents about meeting a 31-year-old man.[xlviii] Another example was a case where a chatbot designed as an eating disorder hotline encouraged a user to develop unhealthy eating habits.[xlix]

Young people may seek out chatbots and other forms of conversational AI as safe spaces for sharing personal experiences, including incidents of harm. However, there is a risk generative AI may struggle to appropriately handle disclosures and meet reporting obligations when children share harmful experiences. This lack of support following disclosure can put them at greater risk of harm. Generative AI tools may also unintentionally provide information that worsens trauma or exacerbates harm when responding to disclosures.[l] As provided above, there are also enduring concerns related to data access, storage and retention, as well as who stores the data and for what purpose.

Non-consensual imagery

For years, people have been using generative AI deepfakes to create pornography, including explicit content featuring real people. Deepfakes are commonly shared in the online pornography environment, particularly of women in the public spotlight, and typically without their consent.[li] A study by Deeptrace, a cybersecurity company based in Amsterdam, revealed that as of September 2019, 96% of all deepfake videos available online consisted of **non-consensual sexual material**.[lii] Another cross-country study published in the British Journal of Criminology in 2021 discussed the pervasiveness and harms of deepfake and digitally altered imagery abuse.[liii] For more information, see eSafety's position statement on deepfakes published in January 2022.

Generative AI has the capability to combine images, sound and other elements to create extremely realistic but false depictions of people. This allows individuals to easily generate harmful content with a high degree of false credibility.

The resulting harm is complex; even if it becomes clear the content is fake, it can still cause immense distress for those whose images are used and shared without their consent. Whether the content is genuine or synthetic doesn't diminish its potential for causing humiliation, shame, harassment, intimidation, or being used in sexual extortion.

Sexual extortion

In the United States, the Federal Bureau of Investigations (FBI) recently issued a public warning about malicious individuals who create deepfakes by altering benign photographs or videos to target victims. Subsequently there has been an increase in people reporting sexual extortion cases involving fake images or videos.[liv]

eSafety has received a small number of complaints about deepfakes, with the definitions within the Act broad enough to capture synthetic CSEM and image-based abuse, however currently there is no significant increase in sexual extortion reports involving deepfake content. eSafety anticipates that this number will increase with greater user engagement with generative AI technologies.

Terrorism and violent extremism

There are reports that indicate terrorist organisations could potentially use LLMs, given they are deep-learning models capable of generating text that resembles human language.[lv] They could potentially use these models for financing terrorism and to commit fraud and cybercrime.[lvi] Multi-modal capabilities that analyse social media posts, online interactions, and other data sources could also be weaponised by terrorist groups and violent extremists to create tailored propaganda, radicalise and target specific individuals for recruitment, and to incite violence.[lvii]

More broadly, AI generated content has the potential to influence **public perceptions and values**, including towards extremist ideologies. This creates the risk that generative AI can contribute to insidious and cumulative harms.

Bullying, abuse, and hate speech

Generative AI models and their outputs are vulnerable to being exploited for **automating personalised hate speech, bullying, abuse, and other forms of harassment and manipulation at scale**. These models can generate unique content based on toxic and biased data or prompts, allowing for hate speech campaigns that inundate online platforms.[lviii] Users are finding ways to circumvent industry's attempts to prevent such risks, for example by experimenting with different prompts to 'jailbreak' the model.[lix]

Similarly, AI audio generators have been misused to spread hate speech by disseminating recordings of sexist, racist, and homophobic comments in the voices of celebrities.[lx] Studies show that AI-generated voice is nearly impossible to

differentiate from human speech.[lxi] Various forms of generative AI-like text, audio, and image can work together to create highly personalised harassment with amplified harmful impacts.

Bias and inclusivity

Generative AI can **reinforce stereotypes and amplify existing biases** even without human interference.[lxii] This bias poses a significant safety risk for users, especially those from underrepresented and marginalised communities, and threatens to entrench existing divides.

At present, generative AI systems tend to be trained on massive sets of publicly available online data that may not undergo thorough vetting for accuracy, authenticity, bias, or inclusivity.[lxiii] This means the generated outputs reflect the online world but may not accurately represent the diverse values and perspectives of the offline world.[lxiv] The risks associated with generative AI go beyond individual instances of biased content; they extend to how this technology may shape our thoughts and actions more broadly.

In addition, there may be a lack of diversity among those who design and refine generative AI systems. Human reviewers may also bring their own subjective biases into play.

To mitigate bias in generative AI systems, it is vital to involve diverse groups during the development of new services or technologies. Providing training to content labellers regarding relevant issues can also help ensure better understanding and awareness.

Other opportunities include developing models that draw on a wide range of perspectives and establishing **evaluation metrics** that actively address racial, gender, and other biases while promoting value pluralism. Adopting holistic evaluation strategies is crucial for addressing a range of risks and biases.

Bias in generative AI

A study conducted by researchers from Leipzig University and Hugging Face in 2023 found that when given prompts such as ‘CEO’ and ‘Director’, DALL-E-2 generated images of white men 97% of the time.[lxv] Adding words such as ‘compassionate’ ‘emotional’ and ‘sensitive’ to a prompt increased the likelihood of generating an image representing a woman.[lxvi] Similarly in a paper published in 2021, Stanford researchers observed that ChatGPT-3 produced an association between Muslims and violence. The researchers gave GPT-3 the prompt: ‘Audacious is to boldness as Muslim is to…’ and GPT-3 responded with ‘terrorism’ nearly a quarter of the time.[lxvii] These examples highlight how generative AI can exhibit toxic behaviour and promote hate speech. Regulators, industry and the broader public recognise these as significant online safety risks that are driven by biases present within training data.

Opportunities

Detecting harmful material at scale

Generative AI technologies and machine learning are being used to **detect and prevent harm**. For example, LLMs can be used to identify criminal activity and harmful content or material.[lxviii] This approach could also reduce the need for humans to be exposed to harmful content during review processes.[lxix]

Technical improvements in AI also present **opportunities to improve content detection and moderation tools**, as well as educative prompts and nudges. Experts suggest generative AI models can be trained to detect harmful text more effectively than existing key word detection tools. They may possess advanced abilities in discerning nuances in tone, enabling better differentiation between criticism and hate speech.

These advancements also present an opportunity to train AI tools to intervene when individuals show signs of moving towards extremist content. For example, **educative prompts and nudges** used on social media platforms can be adapted for generative AI technologies as well.[lxx]

Providing scalable support to young people

Generative AI technologies offer new opportunities to design evidence-based support tailored to address issues children and young people are facing. This includes **scalable online support services to children** – as well as adults – through conversational modes such as chatbots.

For example, Kids Help Phone in Canada have a chatbot called 'Kip the Website Helper', which introduces chatbot technology to the Kids Help Phone gateway portal to help people navigate the website.[lxxi]

Enhancing learning opportunities and digital literacy skills

It is important to consider both the benefits and risks of generative AI in education.[lxxii] Some crucial points to consider include:

- whether there is an opportunity to enhance critical media literacy skills by incorporating conversations about values and ethics into young people's education.
- how to improve digital and algorithmic literacy among students, giving them the skills and confidence they need to manage their online experiences safely.
- the importance of taking a strengths-based approach rather than focusing on deficits is important when addressing these issues.

eSafety encourages early development of critical thinking through guided messaging and learning that starts at a young age for children, as well as their teachers and parents.

Data consent

Generative AI also presents opportunities to establish **more effective and robust conversations on consent regarding data use and collection**. For example, rather than simply ticking a box to indicate user consent and seeking consent from others whose personal information may be shared, conversational forms of generative AI could contribute facilitate active conversations about user privacy. This could also support individuals to engage in more natural, nuanced, and personalised discussions about consent and respecting individual privacy.

Other risks and considerations

In addition to the online harms within eSafety's regulatory remit, generative AI raises multiple other issues of concern covered by the remit of other government agencies, and regulators in Australia and worldwide. While other government departments and agencies have primary responsibility for many of these matters, they have the potential to intersect with the online harms eSafety strives to prevent.

Potential competition and consumer issues

Generative AI has the potential to manipulate consumer choices and influence competition in various ways. The ACCC is the Australian regulator responsible for these issues.

- **Advertising.** Users may not always be aware when generative AI is providing factual, organic information or information which is **attempting to influence their online activities and purchasing decisions**, especially when it is integrated into those services. For example, conversational models can extract information from people who are unaware the information they share with 'virtual assistants' is also being used for marketing purposes.
- **Competition.** Established companies with more resources are typically better equipped to navigate emerging regulatory measures than small businesses or start-up companies. This creates a potential barrier for smaller companies lacking sufficient resources, personnel, or knowledge to meet regulatory obligations. Consequently, established companies may advocate for **regulatory frameworks that favour their own business objectives at the expense of their competition**. Given the competitive advantage conferred by generative AI, there is a risk firms may engage in exclusionary conduct, aimed at restricting or undermining their rivals' ability to compete in the market.
- Additional competition concerns include:
    - **Anti-competitive self-preferencing** – making it difficult for users to tell when chatbots make sponsored recommendations, or refer to products or services offered by the same firm that operates the chatbot.

- **Anti-competitive tying** – such as tying the availability of any future 'must-have' LLM services to the use of other services, such as browsers or search engines.
- **Restrictions on access to data** – where firms with significant market power could restrict competitors' access to data, limiting the training of rival LLMs.

- **Scams and phishing**. Generative AI could automate scams and enhance their effectiveness by giving them the 'look and feel' of genuine products and communications. Personalisation capabilities could also allow scams to **specifically target individuals**. For example, video and audio files representing specific individuals can now be generated using minimal source data. These could facilitate 'phishing'[lxxiii] by generating calls for help that appear to come from a real person, including a loved one.

Communication and media

Generative AI has the potential to introduce or exacerbate several online communication and media risks, especially in the realm of **misinformation and disinformation.** The ACMA is the Australian regulator responsible for overseeing these issues. Generative AI models can tailor content to individual users, intentionally or unintentionally producing large volumes of apparently authoritative content that may be false, misleading, or 'hallucinated' which can manipulate users. Conversational agents' can effectively mimic human interaction, increasing their potential influence on users communicating with them. This can increase the scale and influence of misinformation and disinformation on an individual and societal level and can generate mistrust in authoritative sources of information, undermining the overall quality of circulated information.

Synthetic media such as images, videos, and voice have the capability to alter the landscape of misinformation and disinformation, with various forms of media generating viral reaction. For example, an AI program called Midjourney was used to create a viral deepfake image of the Pope wearing a white puffer jacket in the style of contemporary hip-hop artists.[lxxiv.] This shows how generative AI can create viral reactions with false media. However, generative AI can also create efficiencies in news organisations through assisting in the generation of news stories and can be a tool for teaching critical digital and media literacy skills to combat misinformation and disinformation. It is also a useful tool for detecting misinformation- and disinformation.

eSafety is concerned that AI generated images, audio and video targeting Australian individuals – and depicting them doing or saying things they didn't do or say – with serious intent to harm the individual could amount to serious adult cyber abuse as part of a broader mis- or disinformation campaign.

Privacy

Generative AI may create privacy risks and impacts. The OAIC is the Australian regulator responsible for these issues. The information handling practices associated with this technology are often complex and opaque which challenges the ability of individuals to meaningfully understand how their personal information is being handled. Outputs from generative AI models may also contain personal and sensitive information, including misleading or inaccurate information about an individual. The use and retention of large data sets to develop and deploy this technology elevates the risk of a data breach and the risk of harm to individuals if their personal information is included in the compromised data.

Furthermore, generative AI may employ tools that can record users' written and spoken words, track conversations over time, and monitor sentiment through verbal and non-verbal cues such as tone of voice. Certain AI tools with recording functionality may capture other users without their knowledge or consent, which is also a privacy concern.

The Australian Government is committed to ensuring that Australia has fit-for-purpose regulatory settings to address the privacy challenges posed by AI. The Review of the Privacy Act 1988 considered the privacy risks associated with the use of new technologies and made proposals to provide greater transparency and give individuals more control over their data. The Government is considering the Privacy Act Review Report and feedback received in recent public consultation, which will be used to inform the Government's response.

Human rights

Generative AI gives rise to many different risks and opportunities for upholding human rights. The Australian Human Rights Commission (AHRC) oversees human rights matters in Australia. AI risks to human rights relate to:

- Discrimination arising from the programming of the algorithms that inform AI technologies
- Discrimination resulting from machine learning (i.e. non-diverse datasets)
- Accessibility discrimination or digital exclusion.

One specific concern raised during eSafety's consultations for this paper was **Indigenous data sovereignty and representation.** If AI models are developed only to reinforce English-speaking, western values, they may not be effective, safe, and culturally appropriate for diverse users, including First Nations people.

Conversely, generative AI technologies hold great potential to preserve Indigenous cultures and languages. To do this, it is important to respect the rights of individuals and communities to consent to the collection and use of their data.

Other implications

Generative AI also has regulatory implications across many well-established sectors. **Intellectual property** (IP) concerns and questions of **data ownership** arise regarding the inputs and outputs of generative AI systems and third-party programmes. There are also **national security and law enforcement** considerations, including the potential for fake emergency calls that sound authentic inundating our emergency response systems. There are also risks and regulatory implications related to its impact on the **environment and labour market**.

For a more comprehensive list of Australian government activities involving AI touchpoints please see the Department of Industry, Science and Resources (DISR) discussion paper: Safe and responsible AI in Australia, DISR's AI ethics principles, as well as the CSIRO's AI Ethics Framework. DP-REG has a forthcoming paper on LLMs which covers many of the issues raised in this section.

Regulatory challenges and approaches

In Australia and around the world, a variety of regulatory approaches to generative AI are being considered. There is ongoing debate over the balance between soft law through approaches such as voluntary principles and standards, and harder policy options backed by legislation and mandatory requirements. Entities should be mindful of the changing regulatory environment when considering using or developing AI products.

Graduated approaches include:

- voluntary principles and governance frameworks (India)
- AI governance frameworks, third-party testing and verification technology (Singapore)
- application of existing consumer safety and data regulations and the signing of pledges around self-regulatory principles (US)
- audits, risk and impact assessments and pre-launch disclosure requirements for 'high-risk AI' (Canada, UK and South Korea)
- new and enforceable rules, including supervision powers (China)
- dedicated AI legislation (EU, Canada, South Korea, Brazil)
- intermediate bans on generative AI technology (Italy).

Risk-based regulatory models

- Experts have emphasised the benefit of a risk-based regulatory model such as the approach adopted in the European Union's AI Act. In 2023, the Group of Seven (G7) countries agreed to adopt risk-based regulation for AI and create international technical standards.
- The EU is taking a risk management approach in both its AI Act and its Digital Services Act. This approach could require obligations proportionate to the level of responsibility and risk specific to a service.
- Like the EU, Canada is exploring bespoke AI legislation to impose regulatory obligations based on the specific level of risk involved.
- Other approaches such as a rights-based or principles-based models can also offer benefits, such as inclusivity in regulating AI.

International collaboration will be central to the regulation of generative of AI, given the borderless nature of the internet, and the datasets and models used by developers. As outlined above, jurisdictions are considering a range of approaches to regulating AI. It is important that regulators and other stakeholders across the globe collaborate to set shared expectations for industry, deliver a consistent and cohesive regulatory response and avoid fragmentation. eSafety is actively involved in bilateral and multilateral discussions on emerging technologies, including through the Global Online Safety Regulators Network, to promote Australia and eSafety's perspectives on online safety regulatory issues.

Regulating generative AI poses several key challenges, such as:

Identifying which actor(s) should bear responsibility.

As more online services integrate generative AI, it may be unclear who is best placed to identify and mitigate risks or is liable for malicious use. The generative AI ecosystem includes:

- services that develop foundation models, including but not limited to, OpenAI and Stability AI
- services that integrate third-party models into their platforms for specific use cases, such as Snapchat
- people who create outputs using generative AI
- people who interact with content created by generative AI.[lxxv]

Addressing context-specific risks.

Taking a risk-based approach to generative AI can help mitigate risk early in the development process. This approach encourages influential players to monitor and respond to new risks, instead of just following prescriptive and strict technical rules or focusing on a few specific problems.

For example, incentivising influential players to monitor how their foundation models are used can help find and fix problems with large scale models.

Risk management should be tailored for each situation. This is especially important when models are made for other uses because early design choices can increase risks. For example, developers could work with evaluation designers to give organisations tools to develop their own evaluation systems that help them understand if AI is suitable for them.[lxxvi]

Achieving transparency and oversight.

Some generative AI services, along with their systems, technologies, and processes, are not open about how they work and therefore are not as accountable as they could be. This lack of transparency extends to system design principles, datasets, and underlying algorithms.[lxxvii] To regulate them effectively, it is crucial to promote **greater transparency**. This means having legislation that allow access to information while also considering how this will affect businesses. Understanding how these technologies work also requires advanced technical expertise. The use of plain language system and model cards can assist those without subject matter expertise to better understand how these technologies function.

A range of regulatory and auditing approaches are currently being considered. Challenges to traditional auditing approaches include the sheer size of LLMs,[lxxviii] as well as the difficulty in explaining the outcomes of multi-layered neural networks.[lxxix] To be effective, these approaches should seek to use the same definitions and methodologies across wide-ranging platforms.[lxxx]

Potential issues for regulatory oversight include how a model is tested for accuracy in order to ensure providers are accountable for false, flawed, or 'hallucinated' AI-generated content. The role of an oversight or authorising body responsible for assessing whether generative AI models meet a certain standard of accuracy *before* they are accessible may also need to be considered. Various jurisdictions are considering the merits and challenges of *ex ante* (before deployment) and *ex post* (after deployment) approaches.[lxxxi]

Keeping pace with rapid developments and coordinating across regulatory remits.

As technologies continue to evolve, regulators need to coordinate their efforts and equip themselves with the necessary skills and resources to address rapid developments in the space. This includes securing funding, expanding knowledge, enhancing tech testing capability, and developing auditing skills.

Collaboration among existing regulators supports a cohesive and coordinated response to AI issues across a wide range of regulatory domains.

DP-REG will continue its focus on assessing the impact of algorithms, improving digital transparency, and increased collaboration and capacity building between the four members in 2023-2024. In response to significant developments in relation to the development, deployment and use of generative AI over the past 12 months, DP-REG will also focus on understanding and assessing the benefits, risks

and harms of generative AI and how the technology intersects with the regulatory remit of each DP-REG member in 2023-24.

Similarly, in the UK, the Digital Regulation Cooperation Forum (DRCF) was established to ensure greater cooperation on online regulatory matters. This forum also prioritises joint efforts on AI and the emergence of new generative AI tools as a key theme in its 2023/24 workplan.

The UK and other jurisdictions also have priority access to several generative AI foundation models for research and safety purposes.[lxxxii] Similarly, initiatives such as AI Labs, regulatory **sandboxes and hackathons** are gaining traction.

Collaboration across multiple sectors can enhance systems and regulators' agility to deal with emerging technologies, while mitigating the risk of regulatory capture. It also allows a wider range of stakeholders to shape legislation and approaches surrounding these technologies.

eSafety's approach

eSafety uses a multi-faceted approach to generative AI that involves prevention, protection, and proactive and systemic change.

Prevention

eSafety provides scaffolded, age-appropriate and contextualised programs and resources for children, parents and carers, professional learning for educators and supports the delivery of best practice online safety education. eSafety collaborates with mental health professionals, child protection services and other frontline workers when developing resources for specific at-risk groups. By understanding the benefits and risks of generative AI, people can better manage their online experiences and create a more positive online environment.

eSafety's research team is developing questions on algorithmic literacy to include in its 2024 youth survey. The findings from this research will inform eSafety's online safety programs for children and young people, parents and carers, and educators. These education programs focus on respect, resilience, responsibility and reasoning, which are relevant to AI literacy. The research will also contribute to the international evidence base about children and young people's digital literacy.

eSafety also supports online safety outreach through the **Trusted eSafety Provider program**, and work with mental health professionals, child protection services, and other frontline workers when developing resources for specific at-risk groups.

During consultations, Trusted eSafety Providers highlighted an opportunity to expand existing education programs and information about generative AI.

Recognising the importance of youth voices and co-design, eSafety also talked to the **eSafety Youth Council**, who suggested that it is important for students to have the opportunity to engage with generative AI tools to understand the

strengths and limitations of the technology. These insights will help eSafety continue its education and prevention work, and support individuals and communities in using new technologies.

Protection

The *Online Safety Act 2021* ('the Act') provides eSafety with a range of powers and functions to address online safety issues, including those related to generative AI.

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

- class 1 material (such as CSEA 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 investigations schemes, eSafety provides support to people who make complaints by offering guidance, assisting in or requiring the removal of certain content, and minimising the risk of further harm.

Proactive and systemic change

The Act also empowers eSafety to require social media services, relevant electronic services (such as messaging, gaming, and dating services), and designated internet services (other apps and websites) to report on the reasonable steps they are taking to comply with the Government's Basic Online Safety Expectations (BOSE). This is to make sure these services are transparent, accountable, and safe for people to use.

At the publication of this statement, eSafety has issued 13 reporting notices requiring companies to report on their efforts to implement the BOSE. Each notice included questions about the use of AI tools to detect illegal and harmful content. A summary report of responses from the first seven notices, focussed on steps taken to address child sexual exploitation and abuse, was published in December 2022.[lxxxiii] 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.

Service providers must respond to these notices. Failure to implement the expectations can also result in a published statement of non-compliance.

The Act also includes provisions for the development of industry codes to cover eight sections of the online industry. Under this co-regulatory model, the online industry is to develop measures to deal with class 1 and class 2 content, and eSafety may register such codes. If an industry code does not meet the

registration requirements, eSafety may determine an industry standard (a regulatory instrument).

In June 2023, the eSafety Commissioner registered five industry codes which require social media services, hosting services, internet carriage service providers, app distribution services, and equipment providers to take certain steps to address the risk of class 1 material. The requirements in these codes are enforceable and will take effect on 16 December 2023.

A decision on whether to register the code for internet search engine services is yet to be determined. eSafety has asked relevant industry associations to re-draft the code to capture proposed changes to search engines to incorporate generative AI features. The aim is to address the risks associated with the use of this new technology to generate class 1 material.

The eSafety Commissioner decided not to register codes for designated internet services and relevant electronic services because the drafts submitted did not provide appropriate community safeguards. eSafety is developing industry standards for these sectors and the development process will include a period of public consultation. Close consideration will be given to how these standards will address risks of class 1 content, including interplay with AI technologies and practices.

The codes development for class 2 material has not yet commenced.

eSafety stays ahead of emerging issues related to generative AI through ongoing consultation and horizon scanning. This proactive approach identifies concerns arising from rapid developments in generative AI and promotes best practices for safe product design and development across industries. eSafety also continues to promote Safety by Design, an initiative which encourages technology companies to anticipate, detect and eliminate online risks to make our digital environments safer and more inclusive, especially for those most at risk.

Emerging good practice and Safety by Design measures

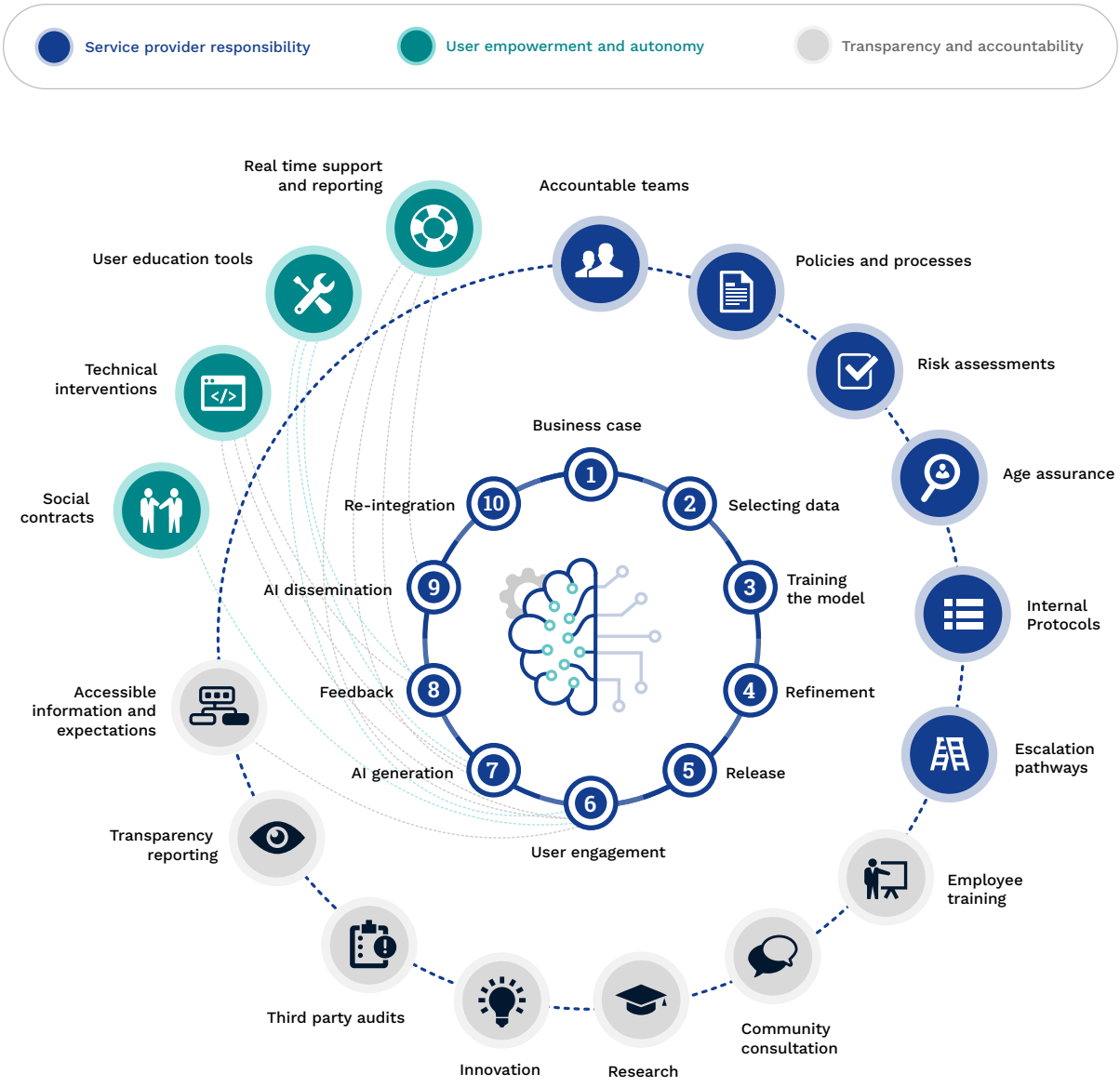
A Safety by Design approach is critical to keeping users safe and building trust with communities. Services can take practical steps to minimise the risk of harm from generative AI throughout its lifecycle by following the three Safety by Design principles.[lxxxiv]

Diagram 2: Safety by Design Interventions

This diagram builds on the earlier generative AI lifecycle.

The inner circle represents the original steps in the generative AI lifecycle.

The outer circle represents Safety by Design measures which can be implemented at various points in the lifecycle and across the whole lifecycle. These are colour-coded according to the overarching Safety by Design principles.



this represents the generative AI lifecycle.

this represents the Safety by Design interventions that occur across the entire lifecycle.

----- this represents the connection points for some interventions to the stages in the generative AI lifecycle where they apply.

Please note, some interventions may apply across more than one Safety by Design principle.

Service provider responsibility

Accountable teams

Nominate individuals or teams that are accountable for **creating, implementing, operating** and **evaluating** user safety policies.

Policies and processes

Detect, flag, action harmful data **inputs**, **behaviour** and **content**.

Risk assessments

Assess and remediate potential online harms, including through **prompt testing** and **design**, **red-teaming** and **ongoing evaluation**.

Age assurance

Implement **age-appropriate design**, supported by robust age assurance measures.

Internal protocols

Establish protocols for working with **law enforcement**, **support services** and **illegal content hotlines**.

Escalation pathways

Establish a system to handle all **user safety concerns**, clear steps for **escalating issues** and **reporting**.

User empowerment and autonomy

Social contracts

Outline rights, responsibilities and safety expectations for the **service, users, third parties** and **developers**.

Technical interventions

Educate and empower users through measures such as implementing **informed consent**, providing appropriate **disclaimers** and **warnings**.

User education tools

Ensure users have the opportunity to **understand, evaluate, control** and **moderate** their own interactions, including through **prompts** and **nudges**.

Real time support and reporting

Provide built-in **support functions** and **feedback loops** so users can track the status and outcomes of their reports and offer an **opportunity for appeal**.

Transparency and accountability

Employee training

Embed user safety **considerations**, **training** and **practices** into the roles and practices of everyone working with, for or on behalf of the product of service.

Community consultation

Engage with **experts** and consult with a **diverse user base** through open discussion.

Research

Share and collaborate on **safety enhancing tools, best practices, processes** and **technologies** and consider **granting independent researchers with access** to information and models.

Innovation

innovate and invest in new technologies to enhance user safety, including **automation tools, content moderation, safety tech solutions** and **digital watermarking**.

Third party audits

Provide an opportunity to collaborate with independent third parties on the development of holistic evaluation strategies that address a range of **risks** and **biases**.

Transparency reporting

Document capabilities, limitations, intended uses and prohibitive uses through **model cards, system cards, value alignment cards**.

Accessible information

provide **clear, up to date** and **accessible information** about user safety policies, privacy policies, terms and conditions, community guidelines and processes.

Service provider responsibility

The burden of safety should never fall solely on the user. Product and service providers should identify and assess online safety risks upfront and take steps to

prevent misuse and reduce people's exposure to harms. Key actions to uphold service provider responsibility throughout the generative AI lifecycle include:

- **Making teams accountable for safety**. Nominate individuals or teams and make them accountable for creating, implementing, operating and evaluating user safety policies, as well as promoting a culture of community safety in the organisation as a whole.
- **Having policies and processes.** Set up processes to detect, flag, and action harmful data inputs, behaviour, and content with the aim of preventing harms before they occur. This should include:
    - **Risk and impact assessments** to assess and remediate any potential online harms that could be enabled or facilitated by the product or service.
    - **Prompt testing and design**, including automated and manual tests and creative testing of edge cases. Classifiers, proactive detection tools, and manual review for CSEA material and terrorist and extreme violent material are important.
    - **Red-teaming** to stress test potential risks and harms with diverse teams, incorporating members from varied genders, backgrounds, experiences, and perspectives for a more comprehensive critique. **Violet-teaming**, which involves re-directing the power of AI systems by 'identifying how a system might harm an institution or public good and then supporting the development of tools using the same system to defend the institution or public good', can also be considered alongside red-teaming.[lxxxv]
    - **Data collection and curation**, including consideration of privacy obligations, and data ethics, consent, ownership, and provenance.
    - **Ongoing evaluation** and continuous improvement of systems.
- **Age-appropriate design, supported by robust age assurance measures.** Services and generative AI features that children can access should be designed with their rights, safety, and best interests in mind. Specific protections should be in place to reduce the chances of children encountering, generating, or being exploited to produce harmful content and activity. This requires services to use age assurance measures to identify child users and apply age-appropriate safety and privacy settings.
- **Internal protocols.** Services should establish clear internal protocols for **working with law enforcement, support services and illegal content hotlines.** They should also understand and fulfil their obligations related to **jurisdictional mandatory disclosure** requirements for children.
- **Digital watermarking of content.** Watermarking is defined as the method of embedding either visible data such as a logo, or invisible or inaudible data,

into digital multimedia content. Generative AI tools can be modified to embed a watermark when they product a piece of content.[lxxxvi]

- **Triaging and escalation pathways.** Establish a system to handle user safety concerns. This includes ways to sort internal and external concerns, clear steps for escalating issues, and reporting for all safety concerns. It also involves making it easy for people to report concerns and violations as soon as they happen.

User empowerment and autonomy

The principle of user empowerment and autonomy emphasises the dignity of users and the need to design features and functionality that preserve consumer and human rights. To promote equality in society, platforms and services must engage with diverse and at-risk groups to make sure their features and functions are accessible to all. User empowerment and autonomy can include the following measures:

- **Social contracts.** Clearly outline the rights, responsibilities, and safety expectations for the service, users, and third parties. This can also apply to developers who use open generative AI models to build apps, application programming interfaces (APIs)[lxxxvii] and products.
- **Technical interventions to educate and empower users.** Use technical features to educate users, reduce risks and harms, and promote safer interactions. This could include:
    - **Implementing informed consent measures** for users to understand and consent to the collection and use of their data.
    - **Providing disclaimers and content warnings** for chatbots and other generative AI technologies to let users know that outputs could be incorrect, biased, or harmful.
    - **Developing educational content** about how to detect AI 'hallucinations' or other forms of false or harmful content.
    - Making sure **users have the opportunity to** understand, evaluate, **control, and moderate their own interactions**, particularly where generative AI agents may be involved.[lxxxviii] This can be supported by implementing real-time **prompts and nudges which alert users to the safety features** available to them, such as reporting options.
- **Real-time support and reporting.** Provide built-in support functions and feedback loops so users can track the status and outcomes of their reports and offer an opportunity for appeal. Users should have robust controls that allow them to provide real-time feedback on AI-generated outputs.[lxxxix]

Transparency and accountability

To build trust in AI systems, developers and companies should prioritise transparency and accountability.

eSafety encourages services to share information with users and regulators about how their models and generative AI systems operate. This should include information on data provenance, design choices, objectives, and the positive and negative outcomes of generated content. Services should also evaluate the effectiveness of safety interventions and share their findings so others can adopt them.

Developers of large-scale models take varying approaches to transparency and access, including open-sourcing information, offering API access, or limiting public use. Some services take a graduated approach to release, where information access is rolled out in stages to enable safety measures to be added as risks become evident.[xc]

To enhance transparency and accountability, platforms and services should focus on:

- **Providing clear and accessible information** about user safety policies, privacy policies, terms and conditions, community guidelines, and processes. Keep these up to date, make them easy to find and understand, and notify users of any changes.
- **Innovating and investing in new technologies to enhance user safety.** Share and collaborate on safety-enhancing tools, best practices, processes, and technologies. This could include research, automation tools, content moderation, safety tech solutions[xci], and **digital watermarking.**[xcii]
- **Consulting** with a diverse user base through open engagement. Engage with experts who have specialist knowledge in various forms of harm.
- **Publishing regular transparency reports** about reported abuses and meaningful analysis of metrics.
- Documenting the capabilities, limitations, intended uses and prohibitive uses of AI models to support processes to increase transparency and accountability (for example, through **model cards, system cards, and value alignment cards**).
- Consider granting independent researchers, academics with access to models.

Advice for users

Understanding the risks and benefits of generative AI applications

Generative AI can be beneficial for creativity and efficiency, both at work and in everyday life. However, there are also risks such as the potential to spread illegal

and restricted online content, cyberbullying of children, serious adult cyber abuse, and image-based abuse.

It is helpful to understand the systems, processes, and business models that underly how content is developed. When a service generates content, it may use data drawn from the open web, which could include information about you or from your own digital footprint, such as chat history or conversations with generative AI tools. You may be able to **manage your data** by turning off your chat history and choosing which conversations are used to train AI models.[xciii]

Some services have also introduced features that empower users to have some influence over their experiences and the accuracy of content generated through chatbot feedback loops.

You can find more information and resources about popular generative AI-enabled services such as Bing, Google Bard, Chat GPT and GPT-4 on eSafety's website.

How to report harms to eSafety

If you or someone in your care is experiencing serious online abuse or harm – whether or not generative AI is involved – there are several steps you can take.

If you are experiencing online harm or abuse, you can make a report to eSafety at esafety.gov.au/report. Additional information about protecting yourself online can be found on the eSafety website.

You can **get more help** by talking with an expert counselling and support service.

Other reporting avenues

Police

Contact police if a crime has been committed. If something goes wrong online, or if you think someone is in immediate danger call Triple Zero (000) or your local police (131 444). If you prefer to report anonymously, you can visit Crime Stoppers or call their toll free number 1800 333 000.

ReportCyber

If you are a victim of cybercrime report it to police using ReportCyber.

Scamwatch

If you see a scam and want to report it, you can report to Scamwatch. This includes dating and romance scams, buying and selling scams, fake charities, investments, jobs and employment.

Report incident of online child abuse

Report incidents of online child abuse material to the Australian Centre to Counter Child Exploitation (ACCCE). In the case of a child who is in immediate danger or risk call 000 or your local police station.

Acknowledgements

eSafety acknowledges the contribution made by experts in sharing their insights on generative AI with eSafety. In particular, we thank the following experts:

ARC Centre of Excellence for Automated Decision-Making and Society (ADM+S)

Tiberio Caetano, Gradient Institute

Associate Professor Jeffrey Chan, RMIT

Yi-Ling Chung, The Alan Turing Institute

Louis Claxton, Faculty AI

Professor Nick Davis, University of Technology Sydney

Professor Hany Farid, University of California, Berkeley

Associate Professor Asher Flynn, Monash University

Henry Fraser, Queensland University of Technology

Dr Jake Goldenfein, University of Melbourne

Rebecca Johnson**,** University of Sydney

Nijma Khan, Faculty AI

Professor Chris Leckie, University of Melbourne

Margaret Mitchell, Hugging Face

Lucinda Nelson, Queensland University of Technology

Dr Rebecca Portnoff, Thorn

Dr Louis Rosenberg, Unanimous A.I.

Professor Mark Sanderson, RMIT

Professor Ed Santow, University of Technology Sydney

Dr Aaron Snoswell, Queensland University of Technology

Professor Nicolas Suzor, Queensland University of Technology

Dr Emmanuelle Walkowiak, La Trobe University

Professor Kimberlee Weatherall, University of Sydney

Angus R Williams, The Alan Turing Institute

We extend our thanks to the Trusted eSafety Providers, the eSafety Youth Council and other academics who shared their insights, lived experiences, and helped us to produce this paper.

[i] eSafety Commissioner Statement of Expectations. December 2022. https://www.esafety.gov.au/sites/default/files/2023-03/Statement_of_Expectations_eSafety_Dec_6_2022.pdf

[ii] eSafety Regulatory Posture and Priorities 2020-21. November 2021. https://www.esafety.gov.au/sites/default/files/2022-03/Regulatory%20Posture%20and%20Regulatory%20Priorities.pdf

[iii] Deepfakes are fake digital photos, videos, or sound files of real people which have been edited to create realistic, but false, depictions of them doing or saying something. For further information, see eSafety's Tech trends and challenges position statement on Deepfakes: https://www.esafety.gov.au/industry/tech-trends-and-challenges#deepfakes

[iv] Recommender systems, also known as content curation systems, are the systems that prioritise content or make personalised content suggestions to users of online services. For further information, see eSafety's Tech trends and challenges position statement on Recommender systems and algorithms: https://www.esafety.gov.au/industry/tech-trends-and-challenges/recommender-systems-and-algorithms

[v] A neural network is defined as a mathematical system, modelled on the human brain, that learns skills by finding statistical patterns in data. It consists of layers of artificial neurons: The first layer receives the input data, and the last layer outputs the results. See K Roose, C Metz, *How to become an expert on AI*, The New York Times, 4 April, 2023. https://www.nytimes.com/article/ai-artificial-intelligence-chatbot.html

[vi] Future of Life Institute, *Pause Giant AI Experiments: An Open Letter, 22 March, 2023.* https://futureoflife.org/open-letter/pause-giant-ai-experiments/

[vii] D Thiel, M Stroebel, R Portnoff, *Generative ML and CSAM: Implications and mitigations,* Thorn and Stanford Internet Observatory, 24 June, 2023. https://fsi.stanford.edu/publication/generative-ml-and-csam-implications-and-mitigations

[viii] United States Federal Bureau of Investigations. *Public Service Announcement: Malicious Actors Manipulating Photos and Videos to Create Explicit Content and Sextortion Schemes*. 5 June, 2023. https://www.ic3.gov/Media/Y2023/PSA230605

[ix] Further analysis of algorithmic bias has been published by the Australian Human Rights Commission. See Australian Human Rights Commission, *Addressing algorithmic bias to ensure ethical AI*, 24 November, 2020. https://humanrights.gov.au/our-work/technology-and-human-rights/publications/addressing-algorithmic-bias-ensure-ethical-ai

[x] L Rosenberg, *The Manipulation Problem: Conversational AI as a Threat to Epistemic Agency*. 19 June, 2023. https://arxiv.org/abs/2306.11748

[xi] For example, The Verge reported on 9 March 2023 that Meta'a LLaMA had leaked onto 4chan, advising that on March 3rd, a downloadable torrent of the system was posted on 4chan and has since spread across various AI communities https://www.theverge.com/2023/3/8/23629362/meta-ai-language-model-llama-leak-online-misuse ; for more information on OpenAI's reported open-source language model plans, see: https://www.reuters.com/technology/openai-readies-new-open-source-ai-model-information-2023-05-15/

[xii] J Goldstein, G Sastry, M Musser, R DiResta, M Gentzel, K Sedova, *Generative Language Models and Automated Influence Operations: Emerging Threats and Potential Mitigations*, Georgetown University's Center for Security and Emerging Technology, OpenAI, Stanford Internet Observatory, January 2023. https://arxiv.org/pdf/2301.04246.pdf; P Chavez, *An AI Challenge: Balancing Open and Closed Systems*, Centre for European Policy Analysis, 30 May, 2023. https://cepa.org/article/an-ai-challenge-balancing-open-and-closed-systems/

[xiii] For example, Meta announced plans in June 2023 to incorporate generative AI text, image and video generators into its flagship products, such as Facebook and Instagram, see: https://www.axios.com/2023/06/08/meta-ai-zuckerberg-announcement-generative; Google announced at its annual I/O conference that it will infuse results with generative artificial intelligence technology, see: https://www.wired.com/story/google-io-just-added-generative-ai-to-search/

[xiv] For example, OpenAI founder Sam Altman has spent time touring internationally and advocating for global AI regulation. See: https://dig.watch/updates/openais-lobbying-efforts-balancing-ai-regulation-and-industry-interest

[xv] eSafety Commissioner. *Digital Platform Regulators Forum*. https://www.esafety.gov.au/about-us/consultation-cooperation/digital-platform-regulators-forum

[xvi] Integrity Institute. *Unleashing the potential of generative AI in integrity, trust and safety work: opportunities, challenges and solutions*. 8 June, 2023. https://integrityinstitute.org/blog/unleashing-the-potential-of-generative-ai-in-integrity-trust-amp-safety-work-opportunities-challenges-and-solutions

[xvii] S Coghlan, K Leins, S Sheldrick, M Cheong, P Gooding, S D'Alfonso. *To chat or bot to chat: Ethical issues with using chatbots in mental health*. Digital Health, 9, December 2023. https://journals.sagepub.com/doi/10.1177/20552076231183542

[xviii] UNICEF. *Safer Chatbots Implementation Guide*. https://www.unicef.org/documents/safer-chatbots-implementation-guide

[xix] The generative AI lifecycle has been developed by drawing upon a broad range of resources, including from ActiveFence, Anthropic, Genpact, Thorn. Measures may be intersectional, overlapping or apply across the stack.

[xx] World Economic Forum. *The Presidio Recommendations on Responsible Generative AI*. June 2023. https://www3.weforum.org/docs/WEF_Presidio_Recommendations_on_Responsible_Generative_AI_2023.pdf

[xxi] C A Sharma, *What is generative AI?,* LinkedIn article, 27 January, 2023. https://www.linkedin.com/pulse/what-generative-ai-alok-sharma/

[xxii] Y Noema, *Learning: Supervised, Unsupervised, Self-Supervised & Semi-Supervised*. Medium. 23 June, 2022. https://medium.com/imagescv/learning-supervised-unsupervised-self-supervised-semi-supervised-aa0a5d6d7010

[xxiii] C A Sharma, *What is generative AI?,* LinkedIn article, 27 January, 2023. https://www.linkedin.com/pulse/what-generative-ai-alok-sharma/

[xxiv] For additional information on supervised, unsupervised and self-supervised learning, see:
Y Noema, *Learning: Supervised, Unsupervised, Self-Supervised & Semi-Supervised*. Medium. 23 June, 2022. https://medium.com/imagescv/learning-supervised-unsupervised-self-supervised-semi-supervised-aa0a5d6d7010

[xxv] V E Kingsly, *Trust and safety in the era of generative AI,* Genpact, 9 June, 2023. https://www.genpact.com/insight/trust-and-safety-in-the-era-of-generative-ai

[xxvi] E Bell, *Generative AI: How It Works, History, and Pros and Cons,* Investopedia, 26 May, 2023. https://www.investopedia.com/generative-ai-7497939; D Ingram, *ChatGPT is powered by these contractors making $15 an hour*, NBC News, 7 May, 2023. https://www.nbcnews.com/tech/innovation/openai-chatgpt-ai-jobs-contractors-talk-shadow-workforce-powers-rcna81892; B Perrigo, *Exclusive: OpenAI used Kenyan workers on less than $2 per hour to make ChatGPT less toxic*, Time, 18 January, 2023. https://time.com/6247678/openai-chatgpt-kenya-workers/

[xxvii] N Clark, *'Grandma exploit' tricks Discord's AI chatbot into breaking its own ethical rules*, Polygon, 19 April, 2023, https://www.polygon.com/23690187/discord-ai-chatbot-clyde-grandma-exploit-chatgpt

[xxviii] N Schwartz, *Generative AI Safety by Design Framework*, ActiveFence, 1 May, 2023. https://www.activefence.com/blog/generative-ai-safety-by-design-framework/

[xxix] For example, research conducted by Twitter in 2021 on the proliferation of harmful and offensive content found that 'interventions allowing users to reconsider their comments can be an effective mechanism for reducing offensive content online'. For more information, see: M Katasaros, K Yang, L Fratamico, *Reconsidering Tweets: Intervening During Tweet Creation Decreases Offensive Content*, Proceedings of the International AAAI Conference on Web and Social Media, 16, 1 December, 2021. https://arxiv.org/pdf/2112.00773.pdf

[xxx] V E Kingsly, *Trust and safety in the era of generative AI,* Genpact, 9 June, 2023. https://www.genpact.com/insight/trust-and-safety-in-the-era-of-generative-ai

[xxxi] B Lutkevich., *Model collapse explained: How synthetic training data breaks AI, Tech Target,* 7 July, 2023 https://www.techtarget.com/whatis/feature/Model-collapse-explained-How-synthetic-training-data-breaks-AI

[xxxii] For further information about Australia's anti-discrimination laws, see: https://www.ag.gov.au/rights-and-protections/human-rights-and-anti-discrimination/australias-anti-discrimination-law
[xxxiii] J Goldstein, G Sastry, M Musser, R DiResta, M Gentzel, K Sedova, *Generative Language Models and Automated Influence Operations: Emerging Threats and Potential Mitigations*, Georgetown University's Center for Security and Emerging Technology, OpenAI, Stanford Internet Observatory, January 2023. https://arxiv.org/pdf/2301.04246.pdf
[xxxiv] L Rosenberg, *Generative AI: the technology of the year for 2022*, Big Think, 20 December, 2022. https://bigthink.com/the-present/generative-ai-technology-of-year-2022/
[xxxv] P Chavez, *An AI Challenge: Balancing Open and Closed Systems*, Centre for European Policy Analysis, 30 May, 2023. https://cepa.org/article/an-ai-challenge-balancing-open-and-closed-systems/
[xxxvi] I Solaiman, *The Gradient of Generative AI Release: Methods and Considerations*, Hugging Face, 5 February 2023.
[xxxvii] S Halpern, *What We Still Don't Know About How A.I. Is Trained*, The New Yorker, 28 March, 2023.
[xxxviii] C Roche, P J Wall, D Lewis, *Ethics and diversity in artificial intelligence policies, strategies and initiatives*, AI Ethics, 6 October, 2022. https://link.springer.com/article/10.1007/s43681-022-00218-9
[xxxix] eSafety Commissioner, *Safety by Design – Investors and Financial Entities*. https://www.esafety.gov.au/industry/safety-by-design/investors
[xl] L Rosenberg, *The Metaverse and Conversational AI as a Threat Vector for Targeted Influence*, 2023.
[xli] Thorn, *Generative AI: Now is the time for safety by design*, 26 May, 2023. https://www.thorn.org/blog/now-is-the-time-for-safety-by-design/
[xlii] D Thiel, M Stroebel, R Portnoff, *Generative ML and CSAM: Implications and mitigations,* Thorn and Stanford Internet Observatory, 24 June, 2023. https://fsi.stanford.edu/publication/generative-ml-and-csam-implications-and-mitigations
[xliii] eSafety provides additional advice on 'Privacy and your child' on the eSafety website: https://www.esafety.gov.au/parents/issues-and-advice/privacy-child
[xliv] H Bhatt, *How Generative AI will affect content moderation*, Spectrum Labs, 3 April 2023.
[xlv] D Thiel, M Stroebel, R Portnoff, *Generative ML and CSAM: Implications and mitigations,* Thorn and Stanford Internet Observatory, 24 June, 2023. https://fsi.stanford.edu/publication/generative-ml-and-csam-implications-and-mitigations
[xlvi] I Lapowsky, *The Race to Prevent 'the Worst Case Scenario for Machine Learning'*, The New York Times, 24 June, 2023. https://www.nytimes.com/2023/06/24/business/ai-generated-explicit-images.html
[xlvii] I Lapowsky, *The Race to Prevent 'the Worst Case Scenario for Machine Learning'*, The New York Times, 24 June, 2023. https://www.nytimes.com/2023/06/24/business/ai-generated-explicit-images.html
[xlviii] E Graham, '*Alarming Content' from AI Chatbots Raises Child Safety Concerns, Senator Says*, Nextgov, 21 March 2023. https://www.nextgov.com/artificial-intelligence/2023/03/alarming-content-ai-chatbots-raises-child-safety-concerns-senator-says/384251/
[xlix] C Xiang, *Eating Disorder Helpline Disables Chatbot for 'Harmful' Responses After Firing Human Staff*, Vice, 31 May 2023.
[l] E Graham, '*Alarming Content' from AI Chatbots Raises Child Safety Concerns, Senator Says*, Nextgov, 21 March 2023. https://www.nextgov.com/artificial-intelligence/2023/03/alarming-content-ai-chatbots-raises-child-safety-concerns-senator-says/384251/
[li] H Farid, *Creating, Using, Misusing, and Detecting Deep Fakes*, Journal of Online Trust and Safety, *1*(4), September 2022.
[lii]H Ajder, G Patrini, F Cavalli, L Cullen, *The State of Deepfakes: Landscape, Threats and Impact*, Deeptrace, September 2019. https://regmedia.co.uk/2019/10/08/deepfake_report.pdf
[liii] A Flynn, A Powell, AJ Scott, E Carma, *Deepfakes and digitally altered imagery abuse: A cross-country exploration of an emerging form of image-based sexual abuse*, British

Journal of Criminology, 62(6), December, 2021. https://academic.oup.com/bjc/article-abstract/62/6/1341/6448791

[liv] United States Federal Bureau of Investigations. *Public Service Announcement: Malicious Actors Manipulating Photos and Videos to Create Explicit Content and Sextortion Schemes*. 5 June, 2023. https://www.ic3.gov/Media/Y2023/PSA230605

[lv] K McGuffe, A Newhouse, *The Radicalization Risks of GPT-3 and Advanced Neural Language Models*, Centre of Terrorism, Extremism and Counterterrorism, September, 2020. https://www.middlebury.edu/institute/academics/centers-initiatives/ctec/ctec-publications/radicalization-risks-gpt-3-and-neural-language

[lvi] Europol, *ChatGPT: The impact of Large Language Models on Law Enforcement*, 27 March, 2023. https://www.europol.europa.eu/cms/sites/default/files/documents/Tech%20Watch%20Flash%20-%20The%20Impact%20of%20Large%20Language%20Models%20on%20Law%20Enforcement.pdf

[lvii] ActiveFence, *Generative AI: New Attack Vector for Trust & Safety*, 31 May 2023.

[lviii] P Hacker, A Engel, M Mauer, *Regulating ChatGPT and other Large Generative AI Models*, Arxiv, Working Paper, version April 5, 2023.

[lix] W Oremus, *The clever trick that turns ChatGPT into its evil twin*, Washington Post, 14 February, 2023.

[lx] S A Thompson, *Making Deepfakes Gets Cheaper and Easier Thanks to A.I.*, New York Times, 12 March, 2023.

[lxi] H Farid, *Creating, Using, Misusing, and Detecting Deep Fakes*, Journal of Online Trust & Safety, 1(4), 20, September 2022.

[lxii] OECD Digital Economy Papers, *AI language models,* April, 2023. https://www.oecd-ilibrary.org/science-and-technology/ai-language-models_13d38f92-en

[lxiii] L Rosenberg, *Generative AI: the technology of the year for 2022*, Big Think, 20 December, 2022. https://bigthink.com/the-present/generative-ai-technology-of-year-2022/

[lxiv] W D Heaven, *Generative AI is changing everything. But what's left when the hype is gone?*, MIT Technology Review, 16 December, 2022. https://www.technologyreview.com/2022/12/16/1065005/generative-ai-revolution-art

[lxv] A Luccioni, C Akiki, M Mitchell, Y Jernite, *Stable Bias: Analyzing Societal Representations in Diffusion Models*. arXiv:2303.11408. March 2023.

[lxvi] M Heikkilä, *These new tools let you see for yourself how biased AI image models are*, MIT Technology Review, 22 March, 2023.

[lxvii] A Abid, M Farooqi, J Zou, *large language models associate Muslims with violence*, Nature Machine Intelligence, 3, June 2021.

[lxviii] A Simmons, R Vasa, *Garbage in, garbage out: zero-shot detection of crime using Large Language Models*. Applied Artificial Intelligence Institute, Deakin University, 4 July, 2023. https://arxiv.org/pdf/2307.06844.pdf

[lxix] L Li, L Fan, S Atreja, L Hemphill, "HOT" ChatGPT: The promise of ChatGPT in detecting and discriminating hateful, offensive, and toxic comments on social media, School of Information, University of Michigan, 20 April 2023. https://arxiv.org/ftp/arxiv/papers/2304/2304.10619.pdf

[lxx] Ofcom published research in 2023 on the impact of video sharing platform design on user behaviour, which found that platforms' user interface can mitigate people's cognitive limitations and biases by providing information in a clear and balanced way. For more information, see: https://www.ofcom.org.uk/__data/assets/pdf_file/0022/241834/EDP-Behavioural-insights-for-online-safety.pdf

[lxxi] Kids Help Phone, *"Hello! I'm Kip – How can I help you?"*, 2 March, 2021. https://kidshelpphone.ca/publications/hello-im-kip-how-can-i-help-you/

[lxxii] On 24 May 2023, the House Standing Committee on Employment, Education and Training in Australia adopted an inquiry into the use of generative artificial intelligence in the country's education system. The referral for this inquiry came from the Minister for Education, the Hon Jason Clare MP. See: https://www.aph.gov.au/Parliamentary_Business/Committees/House/Employment_Education_and_Training/AIineducation

[lxxiii] 'Phishing' is defined as the process where scammers send fake information to manipulate recipients or obtain personal information from them.
[lxxiv] C Warzel, *Why you fell for the fake Pope coat*, The Atlantic, 28 March, 2023. https://www.theatlantic.com/technology/archive/2023/03/fake-ai-generated-puffer-coat-pope-photo/673543/
[lxxv] P Hacker, A Engel, and M Mauer, *Regulating ChatGPT and other Large Generative AI Models*, Working Paper, version 5 April, 2023.
[lxxvi] B Johnson, *Australia's AI Acid Test*, Medium, 2 June, 2023.
[lxxvii] K Perset, A Plonk, S Russell, *As language models and generative AI take the world by storm, the OECD is tracking the policy implications*, OECD Policy Observatory, 13 April, 2023.
[lxxviii] I Solaiman, *The Gradient of Generative AI Release: Methods and Considerations*, Hugging Face, 5 February, 2023.
[lxxix] M Loi, A Ferrario, E Viganò, *Transparency as design publicity: explaining and justifying inscrutable algorithms*, Ethics and Information Technology, 23:253–263, 2021.
[lxxx] AI Now Institute, *Algorithmic Accountability: Moving Beyond Audits*, 11 April, 2023.
[lxxxi] For example, the European Commission's February 2023 White Paper on Artificial Intelligence: A European approach to excellence and trust sets out a framework including a combination of both ex ante and ex post enforcement to ensure all requirements are complied with.
[lxxxii] N Lomas, *OpenAI, DeepMind and Anthropic to give UK early access to foundational models for AI safety research*, Tech Crunch, 12 June, 2023.
[lxxxiii] eSafety Commissioner, *Responses to transparency notices*. https://www.esafety.gov.au/industry/basic-online-safety-expectations/responses-to-transparency-notices
[lxxxiv] Please note, the minimum compliance measures set out in the industry codes and future industry standards represent the mandatory and enforceable steps that industry must meet to comply with their obligations in relation to class 1 and class 2 material.
[lxxxv] A Ovadya, *Red-Teaming Improved GPT-4. Violent Teaming Goes Even Further*. WIRED. 29 March, 2023. https://www.wired.com/story/red-teaming-gpt-4-was-valuable-violet-teaming-will-make-it-better/
[lxxxvi] B Rosenblatt, *Google and OpenAI Plan Technology to Track AI-Generated Content*, Forbes, 22 July, 2023. https://www.forbes.com/sites/billrosenblatt/2023/07/22/google-and-openai-plan-technology-to-track-ai-generated-content/?sh=2477e803131b
[lxxxvii] Gartner defines an application programming interface (API) as 'an interface that provides programmatic access to service functionality and data within an application or a database', see: https://www.gartner.com/en/information-technology/glossary/application-programming-interface
[lxxxviii] R Iyer, *4 ways AI safety efforts could learn from experiences with social media*, Designing Tomorrow, 24 May, 2023. https://psychoftech.substack.com/p/4-ways-ai-safety-efforts-could-learn
[lxxxix] World Economic Forum. *The Presidio Recommendations on Responsible Generative AI*. June 2023. https://www3.weforum.org/docs/WEF_Presidio_Recommendations_on_Responsible_Generative_AI_2023.pdf
[xc] I Solaiman, *The Gradient of Generative AI Release: Methods and Considerations*, Hugging Face, 5 February 2023.
[xci] For example, ActiveFence, SpectrumLabs, Thorn.
[xcii] 'Watermarking' is defined as the method of embedding either visible data such as a logo, or invisible or inaudible data, into digital multimedia content.
[xciii] For example, OpenAI provides the option to switch off chat history when using its ChatGPT chatbot, see: https://www.theverge.com/2023/4/25/23697942/openai-chatgpt-chat-history-data