Australian Government | eSafetyCommissioner

eSafety Submission

Inquiry into the influence of international digital platforms

February 2023

1. Introduction

The Office of the eSafety Commissioner (eSafety) welcomes the opportunity to provide a submission to the Senate Standing Committee on Economics' inquiry into the influence of international digital platforms.

As Australia's independent regulator, educator and coordinator for online safety, eSafety aims to safeguard Australians from harms that can stem from a broad range of digital platforms and online environments that impact their safety online. Our objective is to promote safer, more positive online experiences.

eSafety closely monitors new and emerging tech trends and challenges, including those occurring on major international digital platforms, and advocates for greater transparency and accountability in their efforts to address online safety issues. We also work with local and international stakeholders to examine new research, policy and legislative developments, and provide resources and tools for industry's use.

The *Online Safety Act 2021* (OSA), eSafety's enabling legislation, provides our regulatory functions for online safety, including administering complaints and investigations schemes for 4 types of online harms: cyberbullying of children, cyber abuse of adults, the non-consensual sharing of intimate images, and illegal or restricted online content. It also provides eSafety with powers to regulate digital platforms' broader systems and processes.

Our work forms part of a broader, coordinated response from Australian Government agencies. While we lead efforts in relation to online safety, there are a number of other areas that largely sit with other regulators but intersect with eSafety's remit.

For example, market concentration issues are primarily a matter for the Australian Competition and Consumer Commission (ACCC), but the role of app stores as gatekeepers has the potential to impact user safety if the app store prevents online safety services from reaching Australian end-users. Similarly, issues of data and privacy typically rest with Office of the Australian Information Commissioner (OAIC), though there are many synergies across efforts to protect users' safety and privacy, such as applying the highest safety and privacy settings for users as default settings. eSafety has also seen cases of adult cyber abuse and child cyberbullying which involved the targeting of a person with disinformation and misinformation, for which the industry code is overseen by the Australian Communications and Media Authority (ACMA).

Further detail on eSafety's role, objectives and measures of success can be found in our inaugural 2022-23 corporate plan and our 2022-25 strategy. Our Regulatory Posture and Regulatory Priorities 2021-22 also outlines our current areas of focus and commitments.

In drafting this submission, we have reviewed the Committee's issues paper and provided input against relevant areas of focus and other related matters arising from the paper. We hope this submission will be of value for the Committee and we would be happy to share any additional evidence or clarify any aspects for the Committee's consideration.

Key points from our submission include:

- **Algorithms** are dynamic tools often used by digital platforms to deliver services. While the use of algorithms offer social and economic benefits, their design and purpose can be opaque, and they can be also exploited by users (both businesses and individual end-users) as well as the digital platform, resulting in harm to some individuals. eSafety is undertaking a number of regulatory steps with the aim of increasing platforms' transparency and accountability in relation to how algorithms can impact user safety, including exercising new powers under the OSA.

- As major tech platforms continue to grow and expand into new areas, it is critical that future efforts to **standardise online safety** respond to the dynamic nature of the technology ecosystem and include broad adoption across industry.

- **Children** are at an increased risk of harm online. Responses to issues of children's safety, privacy and other rights, require layered and proportionate approaches that are bolstered through robust safety principles. eSafety takes a role in promoting online environments that meet the needs and best interests of users, including children, and encouraging platforms to implement a range of risk-mitigating measures.

- The emergence of the **metaverse** – enabled by immersive technologies such as virtual reality, augmented reality and mixed reality – has great potential to impact the way Australians learn, communicate, create and experience using new and emerging technologies. eSafety is already working to address potential safety issues that may arise from its use in the ever-evolving technology landscape.

- eSafety observes, participates in and leads a number of **local and international efforts** to enhance our capacity to promote online safety, including the Digital Platform Regulators Forum (DP-REG) and the Global Online Safety Regulators Network.

2. Algorithms and transparency

An algorithm is a coded sequence of instructions that is often used by online service providers to prioritise content a user will see.

These instructions are determined by platforms based on many factors, such as user attributes and patterns, and can involve personalised suggestions to achieve a particular goal, such as discovering new artists, friends, products, activities, and ideas, as well as helping business and creators efficiently reach a target audience.

For these reasons, algorithms are used by almost all digital platforms to amplify, prioritise and recommend content and accounts to their users. Their use and sophistication continues to grow,

with multiple algorithms typically being active within a platform at any given time, all completing different tasks with different outcomes.

In December 2022, eSafety published a position statement on recommender systems and algorithms, which sets out an overview of the use of these technologies by consumer-facing online service providers, new and emerging developments, their safety implications, and global trends relating to their use.[1] This position statement was based on extensive consultation with academics and other subject matter experts.

Manipulation of users and user responses

Algorithms can help drive society and the digital economy by creating speed and efficiency through their ability to collate and disseminate large volumes of data and provide tailored user experiences. They can also reduce online harms, such as helping to identify and filter out abusive and harmful material and bad actors.

However, while algorithms are used across many platforms to great benefit, they can also present an array of safety and other risks to users.

One of these risks is the potential to amplify harmful and extreme content. Many platforms that have the objective of optimising user engagement with their content feeds can, as a result of their algorithms, display increasingly extreme content to a user. Sometimes this content may be borderline to, or even in breach of, their Terms or Service or community standards.

When algorithms identify 'engaging' content, such as that which may be shocking or extreme, they can amplify it to an extensive network of other users, thereby increasing the content's reach and potential impact. Sometimes these algorithms operate in a way that results in the wide dissemination of content before human and editorial oversight is triggered. They can also artificially promote content that has been deemed 'engaging' without balancing other types of content and viewpoints.

On an individual level, these processes can increase the impact experienced by those exposed to harmful material. On a broader societal level, the amplification of content that promotes discriminatory views, such as sexism, misogyny, homophobia, or racism, may have adverse effects, such as normalising prejudice or hate. This may also contribute to radicalisation towards terrorism, violent extremism, and provide users with avenues to find associated groups.

By amplifying some content, algorithms can also end up deprioritising or excluding different viewpoints or valuable ideas contrary to a person's existing beliefs, contributing to what are commonly known as echo chambers or filter bubbles. Echo chambers can impact a person's

[1] We understand algorithms to be the specific computing instructions that form part of wider recommender systems, also known as a content curation systems. While our work program has been primarily focused on recommender systems, for simplicity and consistency with the inquiry, we have referred to these systems as algorithms throughout our submission.

freedom of thought, access to information and autonomy, and can contribute to polarisation. While there is some evidence linking algorithms to filter bubbles, particularly overseas, experts have also highlighted that these issues may be overstated.

Other safety risks stemming from algorithms can directly affect children and young people include:

- how friend/follower suggestions can pressure them to interact with strangers and recommend accounts that may be dangerous
- drawing them into video content loops, which can limit their exposure to diverse content, deliver increasingly problematic content, or result in extended periods of time online
- encouraging dangerous viral challenges
- promoting beauty stereotypes which may be unrealistic or harmful
- normalising the sexualisation of young people
- recommending content that may be appropriate for adults but harmful to children who are not developmentally ready for it, such as violent or sexually explicit material.

These risks and harms can be exacerbated by intersectional factors. A piece of content can be harmful for some people and communities and not others, for example, dieting ads and information that may be suitable for most adults might be harmful to those who have experienced or are at risk of disordered eating, particularly young people. This can make it challenging to prevent specific instances of harm or measure the severity of harms caused by algorithms.

Lack of transparency

Both the inputs used to guide the algorithms and the impact of algorithms employed by digital platforms are typically opaque. By inherent design and the operation of artificial intelligence, they are likely to be in constant flux, changing in response to user needs, economic opportunities, regulation and public sentiment.

Initial limited efforts by industry to increase algorithmic transparency are a step in the right direction, however they generally fall short of offering substantive explanations of the ways in which algorithms may or may not contribute to online harms.

There can be particular difficulties in explaining the functionality of algorithms, especially when there is limited information on their inputs and intended outcomes. Platforms can also be reluctant to share details of their algorithms, given the risk of manipulation by both users and bad actors and the significance of algorithms to their competitive position.

Greater transparency of algorithmic design and the impact is critical to improving public understanding and regulatory responses. While eSafety appreciates the significance of minimising the opportunity for key algorithms to be ‘gamed’ by businesses or bad actors, it is important to ensure that digital platforms are accountable for the impact of their design choices and that users are empowered to make informed decisions.

Regulators are in a unique position to work with platforms to increase transparency measures as well as establish long-term, robust infrastructure to support research and investigation efforts in this space, such as through secure and safe information-sharing regimes.

Regulating algorithms for better outcomes

eSafety has a number of tools at its disposal to promote platforms' transparency and accountability in relation to how algorithms may mitigate or contribute to risks to users' online safety.

Given the complex, evolving, and dynamic nature of algorithms and their use in the online environment, there is no single, fixed regulatory approach to address their potential benefits and harms. Therefore, we consider our work to form part of a broader, multi-faceted response by government, industry and the public, that will require ongoing attention and review.

In March 2022, the House of Representatives Select Committee on Social Media and Online Safety recommended that eSafety, in conjunction with the Department of Infrastructure, Transport, Regional Development, Communications and the Arts (DITRDCA), the Department of Home Affairs (DHA), and other technical experts as necessary, conduct a review of the use of algorithms in digital platforms, examining:

- how algorithms operate on a variety of digital platforms and services
- the types of harm and scale of harm that can be caused as a result of algorithm use
- the transparency levels of platforms' content algorithms
- the form in which regulation should take (if any)
- a roadmap for Australian Government entities to build skills, expertise and methods for the next generation of technological regulation in order to develop a blueprint for the regulation of artificial intelligence and algorithms in relation to user and online safety, including an assessment of current capacities and resources.

We understand the Australian Government is currently considering the Committee's report.

Any regulation or oversight of algorithms should safeguard the rights of users, preserve the benefits of these systems and foster healthy innovation. Important considerations for regulatory efforts targeted at algorithms include:

- harmonising efforts across global government agencies to avoid a fragmented regulatory environment and unnecessary duplication
- understanding the underlying ad-based revenue models which many large digital platforms employ and aligning incentives so that safety considerations are considered in tandem with business incentives
- enhancing education and algorithmic literacy in recognition of the fast-paced nature of technology and that regulation alone is not able to remove all risks.

Some experts, such as the Ada Lovelace Institute, have also proposed several methods to audit algorithms, such as having direct access to a platform's codebase, surveying users of algorithms to gather descriptive data on their experience, and collecting datapoints directly from a platform.

Manual review of code, as discussed in the Ada Lovelace paper, can be highly onerous. Code bases can be extremely large and would require the involvement of broad technical skillsets (including engineers, data scientists and technical auditors) to be able to review a wide array of algorithms, which are continuously being updated. Algorithms are also often written in multiple coding languages which would be difficult, time consuming and expensive for regulators to oversee and evaluate.

The Ada Lovelace paper recommends requesting pseudo-code (plain English descriptions of the code) from developers instead. While this could help mitigate the complexities of human codebase review, securing the technical and experiential skillsets to conduct technical audits would clearly remain a key challenge, particularly for public authorities. The time taken to undertake an audit may also not keep up with the constant changes to codes being used in real time.

Analysis of the source code would also need to be accompanied by collection and review of potentially sensitive user data and training sets. Methods such as site scraping to conduct reviews of the efficacy and impact of algorithms would therefore need to be supported by a clear legislative mandate and information-gathering powers.

Development of industry codes

The OSA provides for the development of mandatory industry codes by sections of the online industry to deal with illegal and restricted content across the digital ecosystem.

In September 2021, we issued a position paper to help guide the digital industry during the industry code development process. We recommended that industry include measures that seek to reduce the promotion and reach of illegal and restricted content within algorithmic systems and user decision-making.

After consulting publicly and with industry, the industry groups tasked with developing the industry codes, submitted eight draft codes to the eSafety Commissioner in November 2022. eSafety provided the industry groups with its initial views in February 2023 raising preliminary concerns with each of the eight draft codes. eSafety has asked the industry associations to resubmit draft codes in early March 2023. eSafety will then decide soon after as to whether the draft industry codes meet the statutory requirement for registration.

If registered, compliance with these industry codes will be mandatory and enforceable, and eSafety will have the power to investigate potential breaches and direct platforms to comply. The eSafety Commissioner has the power to determine industry standards if industry codes are not registered and it is considered necessary or convenient to do so to provide appropriate community safeguards.

Basic Online Safety Expectations

The OSA also provides eSafety with powers to require digital platforms[2] to report on the reasonable steps they are taking to comply with any or all of the Basic Online Safety Expectations (BOSE). This includes proactively minimising material or activity that is unlawful or harmful, protecting children from content that is not age appropriate, and detecting and addressing harmful uses of anonymous services and encrypted services.

The BOSE is intended to provide transparency and accountability for the steps digital platforms are taking to ensure that their end-users are able to use their services in a safe manner.

The obligation for platforms to respond to a reporting requirement is enforceable and backed by civil penalties and other enforcement mechanisms.

Using these powers, eSafety has required platforms to detail how they are using technology to minimise unlawful or harmful material or activity and maintain safe environments. It has also required reporting on the steps taken to ensure that algorithms do not amplify unlawful or harmful material or activity.

Since the OSA commenced, eSafety has issued 12 reporting notices focussed on the steps being taken by platforms to prevent, detect and remove child sexual exploitation and abuse on their services (see case study below). eSafety will issue further notices to platforms on a broader range of harms and issues in the coming months.

Safety by Design

We also aim to produce positive outcomes for Australians and global users by guiding and supporting industry to enhance online safety measures through our Safety by Design initiative.

Safety by Design encourages industry to anticipate potential harms and implement risk-mitigating and transparency measures throughout the design, development and deployment of a product or service. This approach seeks to minimise any existing and emerging harms that may occur, rather than retrospectively addressing harms after they occur. The initiative promotes online safety through 3 guiding principles:

- *Service provider responsibility* – that platforms are responsible for the safety of users
- *User empowerment and autonomy* – that users should be empowered with safety tools and provided with autonomy
- *Transparency and accountability* – that platforms should be transparent and held accountable for their actions.

[2] The Expectations can apply to online services including social media services, relevant electronic services, and designated internet services. For consistency with this inquiry, we have referred to these as digital platforms in this context.

A Safety by Design approach to minimising the risk of harm from algorithms could include:

- **Adjusting algorithms** to focus on more quality-focused metrics instead of, or in addition to, engagement that favours shocking and extreme content, informed by appropriate consultation, public scrutiny and testing.
- **Offering users alternative curation models for their news feeds**, such as a reverse chronological news feed for greater visibility and context of news developments, that are accessible and simple to use.
- **Providing greater choice, control and clear feedback loops for users.** Platforms can empower users to explicitly shape their algorithms, for example by flagging types of content that they do not want to see in suggested posts.
- **Establishing and enforcing content policies, and actively moderating harmful content** to ensure that a platform's pool of available content meets a baseline threshold.
- **Introducing human reviews as a circuit breaker** for content on the path of being amplified, before potentially harmful material can go viral.
- **Introducing additional friction** through design features, such as prompts around whether content has been read and restricting how often content can be shared.
- **Labelling content as potentially harmful or hazardous.**
- **Introducing behavioural cues and prompts that can help users to establish positive patterns of behaviour**, such as those that help users reconsider posting harmful content, or to manage their time spent online.
- **Engaging with at-risk community groups** to develop labels and keywords that users may be using to exploit and cause harm with.

We consider that to better understand the harms that can stem from algorithms, and the factors that contribute to them, it is essential to engage and consult with the platforms and service engineers that develop the algorithms. This provides greater transparency to how they have been designed, their outputs and the rationale behind their use.

Educational response

Recognising the importance of enhancing digital literacy and giving people the skills and confidence to manage their online experiences, eSafety is developing education and training programs to help raise awareness of the potential risks of algorithms and the tools to manage them.

Critical digital literacy is important in knowing algorithms exist and how they work, being able to identify instances of when they are being targeted, as well as understanding how an algorithm may be shaping an individual's experience, beliefs, feelings, or reasoning, and equips them with skills to have a role in their choices around media.

eSafety's current approach to critical digital literacy includes incorporating relevant guidance into our universal prevention programs. This includes being embedded in pre-service and teacher

professional learning, webinars for parents and carers and outreach through our Trusted eSafety Provider program.

We engage with young people through an eSafety Youth Council, where key issues like algorithms are being explored during the 24-month term of the Council. The Council will report to the Australian Government on the challenges they face online and their aspirations and solutions for a safer and more positive online world (further detail provided below).

We are also promoting algorithmic literacy and understanding about how algorithms function and content is personalised, to help give people greater influence over their recommendations. This includes special resources for at-risk groups, especially children, for whom critical thinking skills and media literacy are crucial.

International approaches

eSafety recognises that combatting online harm is a global challenge. That is why we support a harmonised and coordinated response across international policymakers and regulators to reflect the global nature of the online environment.

In addition to the US activity outlined in the issues paper, there are several other jurisdictions that are undertaking similar efforts to address algorithmic transparency in relation to online safety concerns that are relevant to the Australian context. While each of these are focused on their local jurisdictions, there are several common approaches, including increasing transparency and accountability, risk-based regulatory regimes, and systematic reporting requirements.

- In February 2022, the UK Joint Committee recommended a stronger focus on the risks posed by algorithms in its Online Safety Bill. This includes recommending that the largest and highest-risk providers should be placed under a statutory responsibility to commission annual, independent audits of the effects of their algorithms. The Bill also requires services to take a comprehensive, proactive approach to managing risk of online harm, and to ensure that risks to the safety of users are considered as part of product and service design. For example, services will need to consider the risk that algorithms might promote certain harmful content, especially to vulnerable users, or expose children to dangerous engagement with adults.
- The UK Central Data and Digital Office (CCDO) has developed the Algorithmic Transparency Standard, a recording standard that helps public sector bodies provide clear information about the algorithmic tools they use and why they are using them. The Standard is one of the world's first policies for transparency on the use of algorithmic tools in government decision making and is internationally renowned as best practice.
- In September 2022, the UK Digital Regulation Cooperation Forum (DCRF) (the UK equivalent to DP-REG) released a discussion paper: *The benefits and harms of algorithms: a shared perspective from the four digital regulators*, with the intention of fostering debate and discussion among stakeholders.
- The EU Digital Services Act (DSA) is a new set of regulations that require digital platforms to do more to tackle the spread of illegal content and other societal risks on their services in the EU. The DSA includes data access obligation and transparency measures

for major digital platforms, which extends to the algorithms used for recommending content or products to users. It also provides harmonised rules on AI, outlining its risk-based approach to the regulation of AI. It aims to end an era of self-regulation, by requiring platforms to be more transparent about how their algorithmic systems work, and holding them to account for the societal harms associated with the use of their services. The DSA sits alongside the Digital Markets Act which prohibits unfair and anti-competitive practices by large digital gatekeepers.

- The EU Artificial Intelligence Act is a proposed European law on artificial intelligence (AI), the first law on AI by a major regulator. The law assigns applications of AI to three risk categories: first, applications and systems that create an unacceptable risk, such as government-run social scoring are banned; second, high-risk applications, such as a CV-scanning tool that ranks job applicants, are subject to specific legal requirements; lastly, applications not explicitly banned or listed as high-risk are largely left unregulated.
- In mid-2022, the Canadian Department of Heritage convened an Expert Advisory Group on Online Safety. This group was mandated to provide the Minister of Canadian Heritage with advice on how best to design the legislative and regulatory framework to address harmful content online. In this process, many experts stressed that transparency obligations must be structured in a way to allows experts in this field to effectively study online safety. They emphasised that it is crucial for any legislation to ensure that researchers and academics have access to information from regulated services on content moderation, so that they can study the spread and impact of harmful content in Canada. Experts explained that the research could then feed back into the risk-based model, thereby allowing academics and researchers to contribute to the risk management process. They suggested that part of the duty to act responsibly would be to act upon any new research on online safety. Separately, Canada has developed an Algorithmic Impact Assessment tool to support regulators, advocates, public-interest technologists, technology companies, and critical scholars who are identifying, assessing, and acting upon algorithmic harms.
- In May 2019, the Organisation for Economic Cooperation and Development (OECD), which includes Australia as a member, published its guiding Principles on AI, which includes human-centred values and fairness, transparency and explainability, robustness, security and safety, inclusive growth and sustainable development, and accountability.
- The Chinese government has also been developing regulatory tools to govern AI, including a mandatory registration system created by China's internet regulator, the Cyberspace Administration of China (CSC), for recommendation algorithms. This requires platforms to ensure they don't 'endanger national security or the social public interest' and to 'give an explanation' when they harm the legitimate interest of users.
- Singapore's Infocomm Media Development Authority (IMDA) has developed A.I. Verify, an AI governance testing framework and toolkit, to enable industry to demonstrate their deployment of responsible AI. This allows system developers and owners to be more transparent about the performance of their AI systems through a combination of technical tests and process checks. IMDA is currently testing this initiative with industry.

Standardising safety practices

As the breadth of new technologies and digital platforms evolve, so too do new types and methods of online harms. This presents the need for industry to implement consistent, robust and adaptable safety principles across their services and operations.

Siloed implementation of online safety measures, coupled with a lack of transparency around their uptake and efficacy, presents risks of fragmentation, inconsistency, and a duplication of efforts across industry, international regulatory bodies, and other interested stakeholders.

Our Safety by Design initiative provides a framework which enables industry to uplift online safety holistically, irrespective of a platform's size, structure, stage of maturity, jurisdiction of operation, or market reach.

Applying a consistent set of principles, such as those underpinning Safety by Design, can help to avoid a 'splinternet' of expectations and minimise confusion for companies seeking to adopt a recognised system to incorporate, assess and enhance user safety. Additionally, the common vocabularies, shared frameworks for innovation, and interoperability derived from a standardised approach have been found to promote innovation amongst industry sectors.

As major tech platforms continue to grow and exert their power, it is critical that future efforts to standardise online safety capture the diverse needs of the technology ecosystem, with enough flexibility to facilitate widespread adoption regardless of a company's size and industry's market conditions.

Children's data, privacy and safety

Children and young people increasingly rely on digital platforms in their everyday lives to express themselves, learn, play, build their identity, and connect with others. eSafety's research indicates that 94% of children in Australia are already online by the age of 4 years.

eSafety's role in addressing children's safety

Children and young people are especially at risk of online harm, stemming from their lack of digital experiences and broader developmental vulnerabilities.

When eSafety was formed in July 2015 (as the Children's eSafety Commissioner), one of eSafety's main functions was administering a new regulatory scheme in relation to serious child cyberbullying.

Since then, eSafety's remit has broadened to include dedicated abuse schemes to address a range of harms that impact children, including image-based abuse (the non-consensual sharing of intimate images, sometimes referred to a 'revenge porn'), cyberbullying and abuse, as well as exposure to harmful and illegal content such as child sexual abuse material through the Online Content Scheme.

These schemes enable eSafety to receive reports about potentially harmful content and activity, conduct regulatory investigations, and, where appropriate, compel the removal of serious online abuse and illegal and restricted content. This includes powers to issue remedial notices to

relevant online services to ensure harmful content is removed or placed behind an age-gate system, known as a Restricted Access System.

eSafety can also request or require an internet service provider to block material that promotes, incites, instructs in or depicts abhorrent violent material.

We also coordinate our efforts with law enforcement agencies to respond to reports to eSafety from or about children and young people under 18 years. This includes active relationships and dedicated memoranda of understanding (MOUs) with state, territory and Federal police agencies, as well as the Australian Centre to Counter Child Exploitation (ACCCE).

If a person under the age of 18 reports to eSafety that they are the victim of sexual extortion or attempted sexual extortion, we typically:

- refer to the ACCCE for assessment and appropriate action
- provide the child or young person with advice about available supports, prevention, and online safety
- assist with the removal action and/or report social media accounts pending ACCCE clearance to ensure we do not prejudice law enforcement operations.

Additionally, eSafety has a legislated role to improve and promote online safety for Australians, which includes supporting and encouraging online safety education in Australia. This requires a comprehensive approach to producing guidance that addresses a range of online risks for children and young people.

Our statutory functions include:

- supporting and encouraging programs that are relevant to online safety for Australians
- supporting, encouraging, conducting, accrediting, and evaluating educational, promotional and community awareness programs relevant to online safety for Australians
- coordinating the activities of Commonwealth Departments, authorities and agencies relating to online safety for Australians, including children.

Our resources are evidence-based and provide extensive advice to children, young people, parents and carers, and educators about a wide variety of online safety issues. Example resources include guidance to online safety for parents and carers, Online Safety for Every Family videos and factsheets for culturally and linguistically diverse parents and carers, virtual classrooms for education settings, a set of Early Years materials, and materials made specifically for 5–8-year-olds.

eSafety's Youth Council

eSafety's Youth Council, established in April 2022, provides young people a voice about online safety policy. The 24 Council members are aged 13 to 24 years and provide advice to Government about ways to support young Australians to have positive online experiences.

Ahead of Safer Internet Day on 7 February 2023, the Council published an open letter to Big Tech calling for consequences for users who abuse and harass others, in breach of platforms' terms of service. The Council also stated that urgent action is needed to prevent popular online platforms from becoming a haven for trolls, haters and predators.

The letter was published in tandem with new research from eSafety, which has found that almost half of children were treated in a hurtful or nasty way online in the past year, 1 in 10 children have been the target of hate speech online, and 30% of teens have been contacted by a stranger online.

Use of children's data

There is concern that children are being increasingly 'datafied' and that some platforms may be collecting thousands of data points from children, such as information about their activities, location, gender, interests, hobbies, mental health and moods. This can extend to the harvesting of neurological data from immersive technologies, such as the metaverse.

This data can be exploited by third parties to target children with harmful content and behaviour.

Children – like many adults – may not have a strong understanding of the extent to which their activity is being monitored and recorded through data harvesting processes. While this is primarily a matter for OAIC, it also intersects with eSafety's role in promoting online environments that meet the needs and interests of children, and enable them to have safer and more positive experiences.

Importantly, there are some instances where the collection and processing of children's basic data, such as their age, can be in their best interests to ensure their online experience is positive and age appropriate. For example, by identifying users who are children and young adults through supplied age data, platforms can block any incoming messages from unknown adults who may be looking to engage in harmful or illegal activity, such as grooming.

Existing regulatory instruments, such as the UK Age Appropriate Design Code, recognise the importance of services being able to identify which of their users are children and young people so they can create safe, private and appropriate online experiences for them. As part of its review of the *Privacy Act 1988*, the Attorney-General's Department has proposed the creation of an Australian Children's Online Privacy Code, to be modelled of the UK Age Appropriate Design Code, with eSafety being consulted during its development. eSafety is also progressing a roadmap considering how age assurance and other measures can be harnessed to prevent and address harms associated with children and young people's access to online pornography.

The Basic Online Safety Expectations provide examples of reasonable steps that platforms can take to proactively minimise material that is illegal and harmful and protects children. This

includes carrying out child safety risk assessments, and considering age assurance measures to prevent children's access to age-inappropriate content, which would require some access to children's data to implement. Platforms targeted at, or used by, children are also encouraged to consider setting the default privacy and safety settings at the most restrictive level as a reasonable step to ensuring the safe use of a service.

Platforms can also tailor their recommended content to accommodate the best interest of children, such as limiting gambling and alcohol advertisements and sexually explicit material.

eSafety recommends a layered and proportionate response to address these harms that is bolstered through our Safety by Design initiative. For example, platforms can implement a range of risk-mitigating measures such as:

- setting rules for users' minimum age and/or age limits, where appropriate
- detecting and preventing or removing users outside of that age range
- making services safe, private and appropriate for users within that age range, including:
    - higher default safety and privacy settings for younger users
    - provision of tools and controls to help users manage who, what and how they interact with others
    - limiting or providing controls to manage the type of content that is promoted to or accessible by children.

Case study: BOSE reporting

eSafety can use its powers under the OSA to require online service providers to report on their implementation of the Basic Online Safety Expectations. This can include how online service providers are minimising material or activity that is unlawful or harmful for children, such as online pornography, and ensuring that children can use a service in a safe manner.

In August 2022, eSafety issued its first notices to Apple, Meta (and WhatsApp), Microsoft (and Skype), Omegle, and Snap, requiring them to outline the steps they are taking to address child sexual exploitation and abuse (CSEA) on their platforms. As part of the notices, eSafety asked platforms to specify the extent to which platforms are deploying technical tools to identify such content and activity. We received responses from all the providers.

In December 2022, eSafety published a summary of industry responses to these notices and found significant variation in the use of tools and policies to address CSEA. The information received represents a first step towards greater transparency and accountability to address CSEA online.

In February 2023, a second round of notices were issued to Google, Twitter, TikTok, Discord, and Twitch, focussed on child sexual exploitation and abuse, as well as sexual extortion and the safety of algorithmic recommendation systems. This regulatory process is ongoing, and eSafety will publish appropriate information once it concludes.

The metaverse

Immersive technologies and emerging online environments, such as the metaverse, provide a range of opportunities – in entertainment, education, defence, health sciences and other fields. Being able to practise a skill virtually or to understand an experience from an unfamiliar point of view are valuable applications. Immersive experiences can also improve the quality of life and independence of people who are unable to access actual experiences for a variety of reasons, including disability, age, caring responsibilities, transport access or remoteness, and can help people build empathy by experiencing a virtual world from different perspectives.

There are, however, emerging issues with immersive technologies and new online environments, including through the metaverse. In addition to the harms identified in the issues paper, eSafety is concerned that these technologies can be used for cyberbullying, grooming children for online sexual abuse, and image-based abuse. Further, forms of assault might be experienced virtually including through a haptic suit. Augmented realities could also be used to fake a sexually explicit three-dimensional image or video of a real person and interact with it without their consent. While a virtual experience may be considered private due to being physically isolated, there is a risk that an intimate image or video created in that environment could then be livestreamed, stored, or shared without consent.

Alongside these safety risks, there are evolving data and privacy concerns. Immersive technology devices can record vast amounts of data, including biometric information such as fingerprints and location. Such large stores of data could increase identity theft, fraud, and scams.

The use cases for the metaverse and immersive technologies are expected to increase throughout the next decade, with industry investing heavily in its development. Through our Safety by Design initiative, eSafety is proactively working with industry and users to embed risk-mitigating measures into the design, development and deployment of immersive technologies to ensure advancements can be fully and safely enjoyed by all people. We also provide educational and safety resources for parents and carers to support them to increase their digital literacy, such as through our ‘gift guide’, which includes information about immersive technology products to consider before purchasing.

The prospect of a widely adopted metaverse has the potential to also come with a blurring between content and activity, as our online interactions shift from exchanging distinct pieces of content to live and synchronous interactions. This shift is important from a regulatory perspective because current enforcement powers centre on being able to compel the removal of content, rather than harmful activity which is much harder to target. It will be important to consider whether these powers and other online safety measures remain fit for purpose in immersive environments.

eSafety is continuing to consider how its existing regulatory tools (principally our individual reporting schemes, development of industry codes, and the reporting powers under the Basic Online Safety Expectations) can be used to address emerging behaviours and content in the metaverse, such as AI-generated child sexual abuse material.

In some cases, eSafety’s existing tools look to be capable of application to the metaverse. For example, intimate material created in virtual or augmented reality will still be covered by

eSafety's image-based abuse scheme. However, as the use cases for the metaverse develop, further consideration will need to be given to determine whether eSafety's tools remain fit for purpose. Our work forms part of a broader response by the Australian Government and global bodies to promote a safe, private, secure and inclusive metaverse. Digital platforms expanding into the metaverse are already being regulated in relation to safety, privacy, competition, and consumer requirements.

These issues are also being considered at a State level, with the NSW Government recently publishing its report 'The metaverse and the NSW Government' to consider how the NSW Government can play a role as a user, provider and regulator of metaverse applications and platforms.

eSafety also participates in a number of national and international working groups considering metaverse developments, including the Responsible Metaverse Alliance, X Reality Safety Initiative (XRSI) Privacy and Safety Framework 2, World Economic Forum Metaverse Governance Steering Committee, and the World Economic Forum Global Coalition for Digital Safety.

Collaborative efforts to regulate digital platform activities

In addition to the local and international efforts highlighted in this submission, there are several intergovernmental groups, many that eSafety participates in, to enhance online safety on digital platforms. A non-exhaustive list of efforts is provided below.

Digital Economy Regulation Working Group

The Department of Industry, Science and Resources is examining how Australia's regulatory settings and systems can maximise the opportunities artificial intelligence (AI) and automated decision-making (ADM) can offer.

DP-REG

The Digital Platform Regulators Forum (DP-REG) is an initiative of Australian independent regulators, including eSafety, the ACCC, the ACMA, and OAIC, to share information about, and collaborate on, cross-cutting issues and activities on the regulation of digital platforms. This includes consideration of how competition, consumer protection, privacy, online safety and data issues intersect.

In June 2022, the heads of the four DP-REG members met and agreed on a collective set of priorities for 2022–23, which includes a focus on the impact of algorithms, enhancing transparency of digital platform activities and how they are protecting users from harms, and collaboration and capacity building between members.

Further detail on DP-REG's activities can be found in DP-REG's submission to this inquiry.

Global Online Safety Regulators Network

eSafety has worked with regulators in Fiji, Ireland and the UK to form a Global Online Safety Regulators Network (the Network). Through this, we encourage wider international membership and cooperation, with the aim of making sure the approach to online safety between countries is as consistent and coherent as possible.

The Network will share information, best practice, expertise and experience, to support harmonised or coordinated approaches to online safety issues. Members share a commitment to human rights, democracy and the rule of law, and to acting independently of commercial and political influence.

Digital Regulation Cooperation Forum

Similar to DP-REG, the UK's Digital Regulation Cooperation Forum (DRCF) is an intergovernmental group consisting of the Competition and Markets Authority (CMA), the Information Commissioner's Office (ICO), the Financial Conduct Authority (FCA), and the Office of Communications (Ofcom). Formed in mid-2020, DRCF aims to support cooperation and coordination between member regulators on digital regulatory matters. eSafety does not participate in DRCF.

DRCF's workplan for 2022-23 includes greater protections for children online, promoting competition and privacy in online advertising, supporting improvements in algorithmic transparency, and enabling innovation within industry.

DRCF recently published the outputs of its first two research projects looking at the harms and benefits posed by algorithmic processing (including the use of AI) and at the merits of algorithmic auditing.

Social Media (Anti-Trolling) Bill

Finally, we note the Committee's interest in whether the former government's Social Media (Anti-Trolling) Bill could help to improve online safety.

In January 2022, eSafety provided a submission to the consultation for the Bill. This included comments on specific issues the Bill raises, such as:

- highlighting the risk of public confusion in relation to defamation, trolling and adult cyber abuse
- identifying some of the limitations of the proposals, including challenges relating to the collection, verification and utility of information about users
- noting some of the potential unintended consequences, such as the possible exclusion of some Australians from participation in social media if they are unable or unwilling to provide verified contact details to services.

We would welcome discussions with our colleagues at the Attorney-General's Department and DITRDCA to promote a coordinated approach to online harms across Government.

Our submission to the Inquiry into Social Media and Online Safety, as well as our position statements on anonymity and identity shielding, may also be helpful when considering these issues.