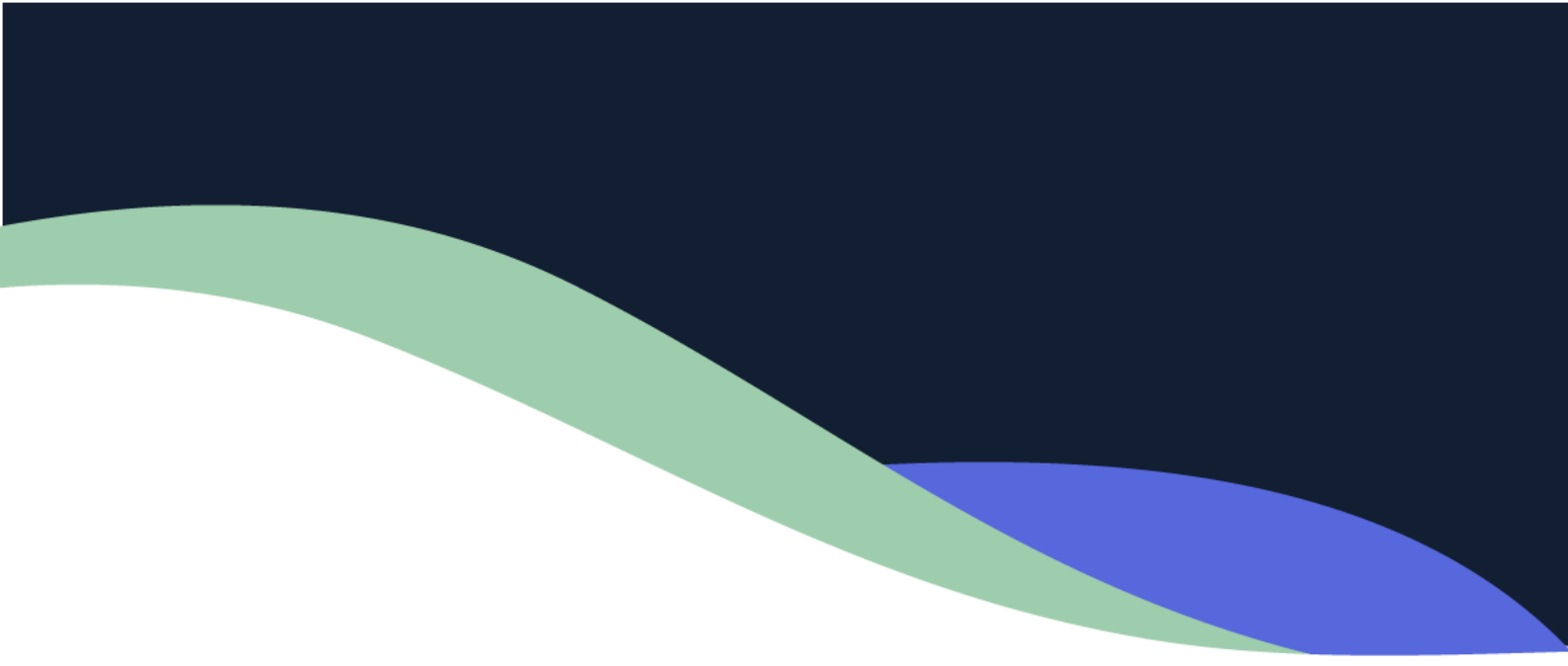




Inquiry into the nature and drivers of radicalisation in Australia

**Parliamentary Joint Committee on Intelligence and
Security (PJCIS)**

Contents

Introduction..... 2

 eSafety’s approach to online radicalisation.....2

 Prevention 3

 Protection.....5

 Proactive and systemic change 11

 Partnerships: Collaborative efforts towards addressing radicalisation ...19

 Online Safety law reform..... 20

Conclusion 21

Introduction

The eSafety Commissioner (eSafety) is Australia's independent regulator, educator, and coordinator for online safety. We aim to safeguard Australians from online harms and to promote safer, more positive online experiences. The *Online Safety Act 2021* (Online Safety Act) sets out our legislative functions. While radicalisation is not defined under the Online Safety Act, online radicalisation can intersect with eSafety's activities where it involves Class 1 or Class 2 material, abhorrent violent conduct material, or online hate that meets the legislative threshold for relevant schemes under the Online Safety Act.

eSafety welcomes the opportunity to make a submission to this inquiry. In line with our remit, our submission focuses on online radicalisation. This submission discusses eSafety's approach to online radicalisation, acknowledging relevant Terms of Reference.

eSafety's approach to online radicalisation

Individuals exposed to, imitating and internalising extremist beliefs and attitudes via the internet can be understood as undergoing online radicalisation. Such radicalised individuals are seen as at an increased risk of committing offences, such as violent acts of terrorism.¹

Exposure to pro-terror and extreme violence material has the potential to cause individuals harm as well as potentially impacting all Australians through the radicalisation of at-risk individuals leading to an increase in real-world violence. Young people are particularly vulnerable to harms from pro-terror and extremist material.²

Through our work across online safety regulation, research and education, eSafety is uniquely placed to inform the Committee about the practices of online service providers, how online services are used, and the steps online services are and are not taking to mitigate radicalisation risks, with the aim to reduce exposure to harmful online content that may contribute to radicalisation pathways.

eSafety takes a risk-based approach to regulating online harms and approaches its remit through four strategic pillars: **prevention, protection, proactive and systemic change and partnerships.**

eSafety's role is one part of a broader ecosystem working collaboratively with law enforcement, government agencies, regulators, industry, academics and civil society

¹ Binder, J. F., & Kenyon, J. (2022). Terrorism and the internet: How dangerous is online radicalization? *Front Psychol.* doi:10.3389/fpsyg.2022.997390

² Commission for Countering Extremism. (2020). COVID-19: How hateful extremists are exploiting the pandemic. London: Commission for Countering Extremism

organisations. We recognise that addressing online radicalisation requires a coordinated response and no single organisation can effectively prevent or respond to these harms in isolation.

Prevention

Research

While eSafety has not conducted research specifically on radicalisation, our research examines associated harms, including exposure to online hate, violent and extremist content.

- In early 2026, eSafety conducted a survey to investigate the use of AI assistants and companions among 1,950 children aged 10 to 17. The [‘Talking to machines’](#) report showed that **almost 1 in 10 children who had used an AI assistant or companion (9%) reported that they had been exposed to online hate** by these tools; specifically, an AI had shared or said ‘mean, offensive or unfair things’ about one or more groups of other people – for example, based on their gender, culture, disability, sexuality or religion. The research also found that most children aged 10 to 17 (78%) had used an AI assistant and 8% had used an AI companion.
- In 2025, eSafety surveyed 3,454 Australian children aged 10 to 17 years to determine the prevalence and nature of online experiences and harms. Concerningly, almost 3 in 4 children (74%) had been exposed online to [content associated with harm](#) at some stage, including **misogynistic ideas** in the form of ‘offensive, sexist or hurtful things about girls or women’ (47%), **online hate** (42%), and seeing **extreme real-life violence** online (22%). Across the various harm types, 69% of children who had been exposed to potentially harmful content online said they **most recently saw it accidentally**, for example in a pop-up or in their social media feed.
- Our 2024 research explored children’s experiences while [gaming online](#), including encounters with online hate and other content associated with harm. The survey sample included 1,799 Australian children aged 8 to 17 who had played video games online in the past year. Of the teen gamers (aged 13 to 17), **1 in 3 (31%) had been exposed to potentially harmful content or ideas via other players in online games** including: **hate speech** (20%), **misogyny** (‘boys or men being better than girls or women’) (11%) and the concept of **people from one race, culture, religion, or nationality being superior** to others (8%).
- eSafety’s [young men online research series](#) (2023-24) explored the online experiences of young men (aged 16 to 21) via focus groups and interviews with young men as well as a workshop with expert practitioners who work with this cohort. Key findings include that, in practitioners’ experience, **algorithms and recommender systems appear to push**

harmful content onto young men and that young men themselves see online communities as places where **abuse, prejudice and discrimination, such as racism, sexism and transphobia, may flourish and be normalised**. The researchers concluded that harmful online content, creators and communities may appeal to young men by meeting their needs for validation, guidance, edginess and belonging.

- eSafety's [Australian adults online survey](#) (2022) found that 1 in 3 adults (34%) had seen online hate in the previous year, increasing to **49% among young adults** aged 18 to 34. The survey included 5,304 adults.

Education and awareness

To support effective prevention of online harms and exposure to content that may lead to self-radicalisation online, eSafety has developed a range of resources for parents and carers, educators, and young people on [dealing with distressing and violent content](#) and [echo chambers](#). In October 2025, eSafety published an online safety advisory on [how violent content is reaching children](#), highlighting that the scale, speed and reach of graphic violent content online has increased significantly, making inadvertent exposure more likely for young people. Recommender systems, livestreams and reposting practices can amplify the spread of gore content, while altered copies can make detection and removal more difficult for platforms.

eSafety also delivers free education to a range of audiences, including webinars for parents and carers and those who support children, young people and their families in the health and welfare sectors. Since July 2024, **over 3,500 people have attended webinars with content relevant to radicalisation** which are designed to prevent online harm, build digital literacy and improve user understanding of potentially harmful behaviours and platform features. This includes sessions on:

- *Influencers, ideology and impact* – exploring how recommender systems and algorithms can influence and reinforce harmful beliefs, and how they can be shaped to build more positive online engagement
- *Exploring the online experiences of boys and young men* – examining how boys and young men engage online and how they may be vulnerable to manipulation or exploitation, sharing strategies to empower boys and young men to navigate the online world safely and respectfully
- *How connection and belonging shapes children and young people's vulnerability online* – providing insights into factors that influence young people's engagement with harmful online content and communities and how families, professionals, and communities can help prevent and respond to online harm.

eSafety also works closely with the national initiative [Step Together](#), which focuses on early intervention and awareness to counter violent extremism, by sharing resources and insights on a regular basis.

Protection

eSafety plays a unique role in protecting Australians from harmful and illegal online content, including content that may contribute to online radicalisation.

Under the Online Safety Act, eSafety administers [complaints schemes](#) to address cyber-abuse material targeted at Australian adults, cyberbullying material targeted at Australian children, the non-consensual sharing of intimate images, and illegal and restricted online content. While these complaint schemes are not directly related to online radicalisation,³ the most relevant is the illegal and restricted online content scheme which includes pro-terror material. These schemes are ‘technology-neutral’, meaning they cover online content regardless of how it is created, including AI-generated material.

Illegal and restricted online content

Illegal and restricted content is divided into the following categories, which are defined under the Online Safety Act by reference to Australia’s National Classification Scheme:

- **Class 1** content includes material that has been or would likely be classified RC (Refused Classification), such as content that promotes, praises, advocates, recruits for, or instructs terrorist acts or violent extremism.
- **Class 2** content includes material that is restricted to adults under the National Classification Scheme but does not meet the threshold for RC. In the radicalisation context, this can include some violent or extremist-themed material that is offensive, disturbing or harmful, but does not actively promote or encourage terrorist activity. Class 2 content can only be actioned when the material is provided from or hosted in Australia.

eSafety's investigations focus on the highest-risk content, including, but not limited to, terrorist propaganda, extremist recruitment activity, manifestos, violent extremist material and content advocating or glorifying terrorist acts.

In 2025-26 financial year, eSafety received 2,102 valid complaints **relating to Terrorism and Violent Extremist Content (TVEC)**, which was an increase of 1,547 complaints (+279%)

³ The adult cyber abuse and child cyberbullying schemes may also be relevant where radicalisation, such as politically motivated violence, terrorism and violent extremism, takes the form of material targeting a particular Australian adult or child. Relevantly, complaints under these schemes may involve abuse targeting a person on the basis of their membership of a group targeted by radicalised or extremist actors.

compared to 2024-25 (555 complaints received). While reports to eSafety about this type of content have increased, these figures may reflect community concern and reporting behaviour rather than a direct measure of prevalence. The majority of reported content relates to promoting or inciting crime or violence, racism, hate speech or political opinion that do not meet the threshold for TVEC or Class 1 material.

Case study: eSafety's response to the Wakeley stabbing

Following the Wakeley stabbing in April 2024, eSafety acted swiftly to prevent the virality of graphic footage depicting the attack. eSafety used its powers under the Online Safety Act to seek the removal of material that was considered likely to be classified as Refused Classification and posed a serious risk of harm to the Australian community, particularly children.

Immediately following the Wakeley attack, most major online services, including Google, Microsoft, Snap and TikTok, worked cooperatively with eSafety to ensure the material could not be accessed for their platforms. Some of these companies took additional, proactive steps to reduce further spread of the material.

In the case of Meta and X Corp, eSafety issued notices on April 16, 2024, requiring them to remove the material under the Online Safety Act. eSafety was satisfied with Meta's response, however, X Corp chose not to remove the video from its platform. Instead, the company geo-blocked to Australia, leaving the material accessible on its platform, including to many users in this country with a VPN.

eSafety was not satisfied with X Corp's response and sought to enforce and uphold its removal notice via actions in the Federal Court and, subsequently, the Administrative Appeals Tribunal.

While eSafety discontinued its Federal Court proceedings, and resolved the Administrative Appeals Tribunal process by withdrawing the notice to X Corp., eSafety's immediate actions in the aftermath of the Wakeley stabbing stemmed the virality of this seriously harmful material. The harm and radicalising potential was nonetheless emphasised by evidence introduced in the sentencing of the UK Southport attacker, who the Court heard was likely to have watched the Wakeley stabbing on the X service six minutes before leaving to kill three young girls at a Taylor Swift themed party, and attempting to murder 10 others.⁴

⁴ [Southport Public Inquiry](#)

The UK Government subsequently established the Southport Public Inquiry into this event. Phase 1 of the Inquiry established the events leading up to the Southport attack and the attack itself, which included the attacker's ability to access the stabbing video on X. In discussing evidence to the Inquiry from Ms Deanna Khananisho of X Corp, the Inquiry report stated:⁵

- *44. Nevertheless, I am conscious of the context. X had fought litigation in order to continue hosting the footage of the Bishop Emmanuel stabbing. Ms Khananisho agreed that X had, over recent years, erred on the side of freedom of speech and the sharing of more material. She also commented that removing the footage “under the guise of safety, that isn’t justice; that is tyrannical overreach”.*
- *45. Ms Khananisho stated that “X is committed to ensuring that its young users are protected from age inappropriate and sensitive content online and that violent and terrorist content is removed or otherwise treated appropriately”. If that was the case, then, given the context, it was incumbent on X to ensure that it protected children from violent content with appropriate age verification. In my view, the age verification measures that X had in place in 2024, which were simply a requirement for the user to state their age, were inadequate and failed to meet X’s purported commitment to young users.*

Phase 2 of the Inquiry continues and is examining a range of issues, including the role of the internet and social media in influencing and enabling violent fixated individuals to prepare and carry out violent attacks. The Inquiry will report on Phase 2 in May 2027.

The Southport attack demonstrates how the widespread availability of violent content online can contribute to radicalisation risks and reinforces the importance of content removal and other measures to prevent the amplification of seriously harmful material.

Case study: Availability of terrorist and violent extremist material in 2025

In 2025, eSafety investigated the availability of manifestos of known terrorists, such as those by the perpetrators of the Christchurch and Buffalo attacks, as part of its Operation Catalyst, each of which was classified as Refused Classification by the National Classification Board. The operation identified 180 URLs hosting copies of 10 terrorist manifestos across 52 clearnet and deep web services. Action taken by eSafety through notifications and formal notices resulted in 133 confirmed removals, six instances of geo-blocking in Australia, and the loss of access to a further two URLs due to domain deactivation. While the operation was a success, it demonstrates the easy access to well-

⁵ [Southport Public Inquiry](#) (April 2026) Phase 1 Report, volume 1, page 254.

known manifestos that clearly pose a risk of radicalisation and instruction in terrorism to Australians, including on mainstream services.

eSafety also investigated the online circulation of graphic violent material depicting three high-profile incidents in the United States: the assassination of Charlie Kirk, the murder of Ukrainian student Iryna Zarutka, and the murder of hotel manager Chandra Mouli Nagamallai in Dallas.

Footage depicting these incidents was widely disseminated across online platforms over a short period and was accessible to Australian users, including children. The rapid proliferation of this material highlighted the challenges associated with the timely identification, assessment, and removal of graphic violent content in an online environment.

In response, eSafety sought urgent classification assessments from the Classification Board. The Board classified instances of material depicting all three attacks as Refused Classification (RC). Following applications by X Corp, some instances of the material were subsequently assigned lower classifications by the Classification Review Board, such that – despite depictions of graphic murder – eSafety was not able to use its removal powers to prevent Australian end-users from encountering this material.

eSafety engaged directly with major technology companies to communicate their broader obligations under the applicable online safety framework, including to limit children's access, and seek the voluntary removal of material.

Class 2 material that is R18+ content and provided from Australia is required to be placed behind a restricted access system (RAS) to prevent children under the age of 18 from accessing it. If a RAS is not in place, eSafety can give the online service or its hosting service provider a remedial notice to remove or restrict access to the material.

Under the *Criminal Code 1995* (Cth), eSafety may also give notices to online services relating to abhorrent violent material (AVM), which is perpetrator-produced material depicting abhorrent violent conduct. These are not removal notices, but are intended to make certain online service providers aware of AVM on or hosted by their services. If a service is later prosecuted for failing to remove or cease hosting AVM, the notice can be used in legal proceedings to show the failure was reckless.

Online Content Incident Arrangement and ISP blocking

eSafety has a key role supporting ongoing inter-agency coordination in response to terrorist and extreme violence incidents by the National Situation Room (NSR), which is run by the National Emergency Management Authority (NEMA).

NEMA and eSafety are developing new governance arrangements to replace the original Online Content Incident Arrangement put in place after the Christchurch attack in 2018, aimed at preventing the viral, rapid, and widespread distribution of online abhorrent violent conduct material.⁶ However, existing operational practices remain in effect while those new arrangements are developed and the NSR continues to identify and refer terrorist and violent extremist material to eSafety for assessment and action.

Under eSafety's protocol with Australian internet service providers, an Online Crisis Event (OCE) can be declared by the eSafety Commissioner when material that depicts abhorrent violent conduct is shared or spread online in a manner likely to cause significant harm to the Australian community, in circumstances warranting a rapid, coordinated, and decisive response by industry and government, even if the material is on a platform or hosting service outside Australia.

eSafety has powers under Part 8 of the Online Safety Act to request or require ISPs to prevent access to the relevant material for a limited time by taking steps like blocking domain names, URLs, and IP addresses. These powers are intended to prevent the rapid distribution online of material closely connected to the OCE, as occurred.

eSafety has not declared an OCE nor implemented the ISP blocking powers since they came into effect. There has not been an incident involving abhorrent violent conduct since these powers have been introduced. Additionally, as the use of notice powers has provided a sufficient – and more targeted – way to reduce the risk of harmful material. Declaring an OCE is a voluntary, non-legislative step and is not a condition precedent to the exercise of relevant powers by eSafety. In practice, eSafety's online tracking and rapid engagement with platforms has enabled relevant material to be effectively managed across the internet and major social media platforms without the need for declaring an OCE.

Relationships with Law Enforcement and National Intelligence Community

Law enforcement agencies

⁶ Abhorrent violent conduct occurs when a person engages in a violent terrorist act, murders or attempts to murder another person, tortures another person, rapes another person, or kidnaps another person using violence or the threat of violence. Abhorrent violent conduct material is material which promotes, incites, instructs in, or depicts abhorrent violent conduct.

eSafety has robust referral and information-sharing arrangements with law enforcement, including a Letter of Exchange with the AFP's Counter Terrorism and Special Investigations Command focused on information sharing on terrorism, violent extremism, radicalisation and hate crime related issues.

eSafety has specific powers and obligations regarding the sharing of information with law enforcement agencies. eSafety may disclose information to law enforcement agencies and other regulatory bodies if it is satisfied that the information will assist them to perform their functions or powers. eSafety must notify a member of an Australian police force if material is of a sufficiently serious nature.

National Intelligence Community

eSafety's work to prevent and minimise online harms is distinct from, but complementary to, the efforts of the work of the national intelligence community. Accordingly, eSafety engages as required with members of the national intelligence community on matters of national security related to its remit. While eSafety can engage with intelligence agencies in relation to a possible OCE, existing information-sharing provisions under the Online Safety Act are more limited than those applying to information sharing with law enforcement agencies.

Case study: Opportunities for information sharing and collaboration

As many online threats are global, the response to threats must be increasingly coordinated across roles and jurisdictions.

In the event of a major incident involving online material of the incident, or where material of the incident is likely to go online, eSafety engages with law enforcement and industry to ascertain the need to exercise any functions or powers under the Online Safety Act. This occurred on the evening of the Bondi attack in December 2025.

Announced by AFP Commissioner in May 2026, the [Counter Terrorism Online Centre](#) establishes new capability to provide an early warning system for the nation's Joint Counter Terrorism Teams about emerging hate groups and individuals who use social media, gaming platforms, online forums and the dark web to incite or commit violence. The focus on online radicalisation and extremist content presents potential opportunities for collaboration between the Department of Home Affairs, the Counter Terrorism Online Centre and eSafety, particularly in relation to terrorist and violent extremist material and platform engagement.

Different legal frameworks across countries affect how harmful content is classified and hashed. There have been calls for international standards for harmful content classification, similar to those used to classify child sexual abuse material. eSafety recognises that international classification standards would support improved consistency.

Hash-sharing databases – an important form of information sharing – are vital for effective content takedown. Examples include Tech Against Terrorism’s (TAT) [Terrorist Content Analytics Platform](#) and the Global Internet Forum to Counter Terrorism’s (GIFCT) [Hash-Sharing Database](#).

To support information-sharing on classifying harmful content, eSafety recommends the committee explore access to sources such as the TAT Terrorist Content Analytics Platform and the GIFCT Hash-Sharing Database, as well as greater sharing of hashes between Government, law enforcement and regulators to detect and remove known terrorism content.

Proactive and systemic change

eSafety drives proactive and systemic change across the online ecosystem. Services providers across the technology stack have **a range of obligations under the Online Safety Act designed to enhance the online safety of Australians.**

Codes and Standards

The [online safety codes and standards](#) place enforceable requirements on a wide range of services to protect all Australians from illegal and restricted online content. Failure to comply with the codes and standards may result in civil penalties of up to \$54.6 million.

The Unlawful Material Codes and Standards are most relevant to online radicalisation. They contain mandatory measures to address the most seriously harmful online content, such as pro-terror material and extreme crime and violence. Pro-terror material may include ideologically motivated extremism towards Australians. Mandatory requirements in relation to pro-terror material include requiring services to detect and prevent the creation and distribution of such material; proactively remove the material as soon as practicable; provide effective user-reporting mechanisms, and appropriate and timely responses to user reports; and notify appropriate authorities of the material.

Case studies: Enforcement of the Unlawful Material Codes and Standards

[Civil penalty proceedings against Telegram](#)

In July 2026, eSafety commenced civil penalty proceedings in the Federal Court against the provider of the Telegram messaging service, for its alleged failure to detect and remove pro-terror material from the platform, including videos of terrorist executions and mass shootings, under the [Relevant Electronic Services \(RES\) Standard](#).

The continued proliferation of pro-terror material and extreme violence on services such as Telegram risks normalising extremist content, causing real harm to Australians who are exposed to it, and facilitating recruitment and radicalisation of at-risk individuals. These proceedings are important in creating specific and general deterrence, should the Court decide there have been breaches of the RES Standard.

eSafety secures court-enforceable undertaking from Roblox

In August 2026, eSafety accepted a court-enforceable undertaking from Roblox to put in place robust measures designed to protect children from grooming (which may include recruitment into violent extremism) and sexual exploitation and abuse. The undertaking following an investigation and testing by eSafety that found, amongst other things, that adult accounts on Roblox were able to find and add children as ‘connections’, view information such as their interests and friends lists, and engage with them on the service.

While Roblox disagreed that it had breached its obligations, it entered into the enforceable undertaking committing it to put in place additional measures to ensure that adults could not contact children without parental consent, and that children’s accounts were private by default. Importantly, the undertaking includes independent auditing requirements and eSafety will evaluate compliance having regard to the independent audit report.

Such audits will become increasingly important across industry to ensure that compliance is independently tested and evaluated.

Age-Restricted Material Codes

The Age-Restricted Material Codes promote children’s right to be protected from accidental and unsolicited exposure to age-inappropriate material, including high-impact violence material. They impose a range of obligations on services, including services with AI chatbot functions, such as implementing age assurance to restrict children’s access to high-impact violence.

Basic Online Safety Expectations: Transparency requirements

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

themselves are not enforceable, eSafety has powers to compel information from providers on their compliance with the BOSE. These powers are backed by civil penalties and form the basis for the transparency scheme under the Online Safety Act. Transparency is critical to ensuring the accountability of online service providers, and a powerful lever for incentivising safety improvements.

Case study: [Transparency report to online gaming and gaming-adjacent services](#)

On 1 April 2026, eSafety gave non-periodic reporting notices to the providers of Roblox, Fortnite, Minecraft and Steam. The notices asked the service providers to explain how they were meeting the BOSE during the period from 1 October 2025 and 31 March 2026. They asked for detailed information about the systems, processes, technologies and resources each service provider used to help protect children from harms including violence material and activity such as pro-terror, violent extremism, and online hate.

In October 2026, eSafety published its transparency report summarising providers' responses to these notices, *Playing it safe: how online game services are (or are not) keeping kids safe online*.

Key findings from the report relevant to radicalisation include:

- All four services allowed users to create or play different games as part of their services. Their level of control over the games and their creation varied. For example, Roblox stated it used AI classifiers to monitor for harmful signals and behaviours across all games and at all times, including signals relating to terrorism and violent extremism. By contrast, Steam stated it required game developers to complete a survey on the content in their game and it undertook a review of games prior to publication, checking for certain information such as game titles against violence material keyword lists. Steam only retested games based on factors such as the number and content of user reports and the known history of the game and developer, not based on proactive tools or classifiers.
- Where Minecraft, Roblox and Fortnite said they used multiple indicators to detect users who had re-registered (or attempted to) with new details after being banned or suspended for CSEA or violence material related offences, Steam reported that it used minimal indicators. The fewer indicators used, the more likely it is that perpetrators can return and continue to carry out harms such as child sexual exploitation, sexual extortion and terrorism or extreme violence.
- For both Minecraft and Steam, their trust and safety staff did not operate fluently in any of the five most commonly spoken languages in Australia other than English (i.e. Mandarin, Arabic, Vietnamese, Cantonese and Punjabi). Instead, all services reported

that their Trust and Safety staff used language translation software when needed. Services that rely on language translation tools risk losing important context and cultural or linguistic nuance.

- Services differed in the extent to which they participated in cross-industry information sharing initiatives such as Tech Against Terrorism, GIFCT, and the Christchurch Call Foundation. While Roblox, Minecraft and Fortnite all participated in these kinds of initiatives to at least some degree, Steam did not participate in any information-sharing initiatives at all. Services that do not participate in signal-sharing initiatives may be at significant risk of failing to detect and remove unlawful and harmful material and activity, including activity which promotes, incites or instructs in matters of extreme violence.

Background to the report: Violence material and activity in online games

Online games services present unique challenges due to the combination of features such as real-time player chat, pseudonymous play, immersive environments, and easy interaction across gaming communities. There is evidence online games have been used to promote, incite or instruct players in crime and violence. This has included violent extremists exploiting gaming spaces to spread ideas, recruit users or build communities.⁷

While online hate does not always lead to violence or radicalisation, extremist groups often spread online hate as part of their rhetoric, including threatening violence against a particular group. A 2024 study of Australia and six other countries found that 38% of respondents saw endorsements of violence against a particular group on online game and gaming-adjacent platforms.⁸ The report also notes that communities associated with gaming “can serve as sites of misogynistic and gendered harassment of women and members of the LGBTQI+ community – the normalisation of which can lower resilience to radicalisation and have offline consequences”.⁹ Perpetrators are known to move children from games that promote extremist ideas to more private spaces that allow indoctrination and mobilisation.¹⁰

⁷ See eg NYU Stern Center for Business and Human Rights (2023), [Gaming the System: How Extremists Exploit Gaming Sites and What Can Be Done to Counter Them](#), NYU website; UNICEF (2025). [Protecting Children in Online Gaming: Mitigating Risks from Organized Violence](#), UNICEF website; Radicalisation Awareness Network (2021). [Digital Grooming Tactics on Video Gaming & Video Gaming Adjacent Platforms: Threats and Opportunities](#), EU Migration and Home Affairs website.

⁸ White, J., Wallner, C., Lamphere-Englund, G., Frankie, L., Kowert, R., Schlegel, L., Kingdon, A., Phelan, A., Newhouse, A., Saiz Erasquin, G., & Regeni, P. (2024). [Radicalisation through Gaming: The Role of Gendered Social Identity](#) (p.22), RUSI website.

⁹ White, J., Wallner, C., Lamphere-Englund, G., Frankie, L., Kowert, R., Schlegel, L., Kingdon, A., Phelan, A., Newhouse, A., Saiz Erasquin, G., & Regeni, P. (2024). [Radicalisation through Gaming: The Role of Gendered Social Identity](#) (p.22), RUSI website.

¹⁰ NYU Stern Center for Business and Human Rights (2023), [Gaming the System: How Extremists Exploit Gaming Sites and What Can Be Done to Counter Them](#), NYU website.

Sadistic online exploitation, sometimes known as ‘nihilistic violent extremism’ is characterised by the deliberate use of the internet to harm, manipulate or exert control over others for personal pleasure, gratification, a sense of power, ideological beliefs or to corrupt victims and inflict ongoing harm.¹¹

Perpetrators have been known to groom their victims, as well as coercing them through sexual extortion and other kinds of online blackmail. Certain groups associated with sadistic online exploitation have been known to use Neo-Nazi, white supremacist, or other extremist language.¹²

Prior to the recent report on gaming, eSafety also gave non-periodic reporting notices to a selection of other online service providers, including Google, Meta, WhatsApp, Reddit, Telegram and X, to report on steps they were taking to [tackle the risk of terrorist and violent extremist material and activity \(TVE\)](#) on their services for the period of 1 April 2023 to 29 February 2024. The notices required answers to questions about the tools, processes and resources they use to ensure safety. Responses from Google, Reddit, Meta and WhatsApp, as well as information provided by Telegram after the deadline, revealed that although these service providers did have measures in place to detect and address TVE on their services, they were not applied consistently or comprehensively. [Summaries of the responses](#) were published on eSafety’s website in March 2025, and a further addendum in July 2025.

Social Media Minimum Age obligation

The [Social Media Minimum Age](#) (SMMA) obligation requires providers of age-restricted social media platforms to take reasonable steps to prevent Australian children under the age of 16 from having accounts. The SMMA obligation aims to protect young Australians from pressures and risks that users can be exposed to while logged into social media accounts. These come from design features that encourage them to spend more time on screens, while also serving up content that can harm their health and wellbeing. eSafety has [raised concerns](#) with the extent to which platforms are complying with the obligation, and investigations remain underway.

Safety by Design

eSafety’s [Safety by Design](#) initiative shifts the onus of safety from the user to service providers. This approach seeks to embed safety in the process, rather than retrospectively addressing harms after they occur. There are three core principles: **service provider**

¹¹ Australian Centre to Counter Child Exploitation (2025). [Sadistic Online Exploitation](#), ACCCE website.

¹² Resolver (2026). [Weaponized loneliness: Understanding a hybrid threat to young people globally](#), Resolver website.

responsibility, user empowerment and autonomy, and transparency and accountability. Key Safety by Design measures to prevent and address radicalisation include:

- Implementing processes in place to detect, surface, flag and remove illegal and harmful behaviour, contact and content with the aim of preventing harms before they occur.
- Providing technical measures and tools that adequately empower users to manage what content they see, and that are set to the most secure privacy and safety levels by default.
- Establishing clear protocols and consequences for service violations that serve as meaningful deterrents and reflect the values and expectations of the users.
- Publishing information about safety tools, policies, and processes, and their impact and effectiveness.

Practical resources are provided via the [Safety by Design industry guides and assessment tools](#), including educative content on intersectional risk factors for online harms, insights into perpetrator motives, and exploration of human rights in the digital context. The recent [Preventing child sexual exploitation and abuse online guide](#) covers sadistic online exploitation and organised networks with nihilistic violent ideologies.

Impact of emerging and evolving technologies on online radicalisation

As a proactive and anticipatory regulator, eSafety monitors emerging and evolving technologies and harms through our [Tech Trends and Challenges](#) program.

Recommender systems and algorithms

In May 2026, eSafety published a [position statement on recommender systems and algorithms](#), which explores the benefits and risks of recommender systems and their underlying algorithms, as well as impacts on an individual and societal level.

Recommender systems can reduce online harms, for example by helping to identify and filter out abusive and harmful material and bad actors. However, they can also present risks, including the potential to amplify harmful and extreme content, especially where they are optimised for user engagement. Systems that prioritise user engagement with content feeds can display increasingly extreme content to a user. This can amplify content that promotes division, false narratives, and undermines democratic values.

Generative AI

In August 2023, eSafety published a [position statement on generative AI](#). Generative AI can both be trained on, and contribute to the generation and amplification of, content that promotes bias and discrimination. Such content has the potential to normalise hate,

intolerance, or extremist views. It may also lead to an erosion of trust in online content or institutions. Multi-modal capabilities that analyse social media posts, online interactions, and other data sources could also be weaponised by extremists to create tailored propaganda, radicalise and target specific individuals for recruitment, and to incite violence.

However, generative AI technologies and machine learning also present opportunities to improve the detection and prevention of harm. For example, generative AI models can be trained to detect harmful text more effectively than existing key word detection tools, potentially improving content detection and moderation. They may possess advanced abilities in discerning nuances in tone, enabling, for example, better differentiation between criticism which is permitted under relevant terms of use and hate which is not. There may also be 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.

Case study: Terrorist and violent extremist material and generative AI

eSafety is concerned by reports that perpetrators are misusing generative AI to cause harm, including to gamify graphic violent content, such as terrorism, violent extremism, and gore. eSafety is particularly concerned about the role of livestreaming, recommender systems and generative AI in producing, promoting and spreading terrorist and violent extremist content and activity.

The Christchurch Mosque terrorist attack in 2019 demonstrated how livestreaming can transform an act of violence into ongoing harmful content, enabling real-time dissemination and inspiring copycat behaviour. We continue to receive complaints about this content circulating online and have identified terrorism manifesto material that has been inspired by the Christchurch event.

We have also identified artificially recreated versions of Christchurch, including gamified 'first person style walk-through' adaptations. This demonstrates how generative AI can magnify these risks by enabling the rapid, low-cost creation and dissemination of sophisticated extremist material, including synthetic images, audio and video, with the potential to contribute to radicalisation.

Without adequate detection and moderation, high-risk user-generated environments can thrive in online gaming environments, such as games which recreate mass shootings or incorporate terrorist propaganda.

Case study: Preventing the generation of pro-terror material in AI

Following our compliance engagement in relation to the Unlawful Material Standards, AI model distribution service *Hugging Face* updated its rules to require users to take measures to prevent their models from being used to generate pro-terror material and child sexual exploitation material. As far as we are aware, the Online Safety Act is the only online safety legislation globally that applies to model distribution services, a vital part of the ecosystem (just as app stores are for ‘traditional’ services).

These obligations are essential given the rise of AI models that have been demonstrated to be able to generate terrorist propaganda and instructional material. For example, the organisation Tech against Terrorism found that ‘open-weight models’ developed by a range of companies easily had their safety guardrails removed via a process of ‘abliteration’. Tech against Terrorism found that this process of removing guardrails on average occurred within three days of a model being published for the first time. Abliterated models were more capable of generating usable instructions on how to cause a mass casualty attack, even when the prompter specifically claimed to be a terrorist.

Amongst other things including independent safety benchmarks, Tech against Terrorism calls for companies that distribute models such as Hugging Face, Civitai and Github to raise the barrier to abliterated copies, reflecting eSafety existing work and powers under the Online Safety Act’s Codes and Standards.

End-to-end encryption (E2EE)

In October 2023, eSafety published a [position statement on end-to-end encryption \(E2EE\)](#). E2EE generally prevents information or data being read or modified by others, including the messaging service itself, while in transit. In contrast, when using standard transit encryption, data can be accessed by the service provider. This access is often material in detecting illegal content like TVEC.

eSafety encourages E2EE services to adopt a Safety by Design approach and measures to reduce misuse. Measures can include use of behavioural signals, effective user reporting, restricting invite links and working with law enforcement to disrupt illegal activity. Importantly, the Codes and Standards under the Online Safety Act do not exempt E2EE services from taking measures to ensure safety from unlawful material. Instead, there are key protections for cyber-security and privacy, but a recognition that even where E2EE services cannot take certain measures because of these considerations, or technical infeasibility, they must still take ‘alternative appropriate action’ in relation to harms such as child sexual exploitation material and pro-terror material.

Partnerships: Collaborative efforts towards addressing radicalisation

We recognise that online safety issues, including the scale and evolving nature of radicalisation, are global challenges that requires a coordinated and collaborative response. Through our [international engagements and collaboration](#), we work closely with online safety regulators, governments, researchers, and civil society around the world to promote greater regulatory coherence and share expertise. A key mechanism for collaboration is the [Global Online Safety Regulators Network](#) (GOSRN), which provides a forum for sharing information, emerging trends and regulatory approaches.

Through GOSRN, members can learn from practical regulatory responses implemented in other jurisdictions, including regulators implementing the EU's Terrorist Content Online Regulation and Digital Services Act, and UK Online Safety Act, among others. GOSRN also facilitates engagement with experts in radicalisation through its observer stream, including with the Institute for Strategic Dialogue, INTERPOL and the Centre for the Study of Organized Hate.

eSafety also participates in a range of other multilateral forums that address elements of radicalisation. This includes engagement with the Christchurch Call Foundation on its [Crisis Response Protocol](#), the [Project Catalyst](#) initiative looking at violent misogyny violent extremism, and [exploration on radicalisation in gaming](#). eSafety also represents Australia in the [Global Partnership for Action on Gender-based Online Harassment and Abuse](#), which has explored the links between gender-based violence, misogyny and radicalisation. Through these [international engagements](#), eSafety contributes to international efforts to better understand emerging risks, strengthen regulatory and policy responses to radicalisation, and support a safer online environment globally.

In addition to the work highlighted above, eSafety also collaborates with local and international regulators to enhance online safety on digital platforms. Notably, the Digital Platform Regulators Forum (DP-REG) is an initiative of Australian independent regulators, including eSafety, the Australian Competition and Consumer Commission, the Australian Communications and Media Authority, and the Office of the Australian Information Commissioner, to share information about, and collaborate on, cross-cutting issues and activities on the regulation of digital platforms.

In 2023, DP-REG released a [working paper](#) on the harms and risks of algorithms, which discusses their potential role in mitigating or amplifying extreme content and radicalisation.

Online Safety law reform

Online Safety Act reform and the Digital Duty of Care

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

The *Online Safety Amendment (Digital Duty of Care) Bill* is currently before Parliament. If passed in its current form, the Digital Duty of Care would require services to take appropriate action, including effectively mitigating risks identified in risk assessments, to ensure a safe online environment. A safe online environment is one where all Australians are protected from seriously harmful material and conduct, and children are protected from material and conduct harmful to them. Harms relevant to radicalisation are captured, including material or conduct relating to support for a terrorist organisation or terrorist act, and violent extremist material.

eSafety supports a Digital Duty of Care which builds on and strengthens eSafety's critical work to date to hold industry to account.

Reforming the National Classification Scheme

Decoupling the Online Safety Act from the National Classification Scheme was a key recommendation in the Online Safety Act review, to which the Government noted. The Online Safety Act's current reliance on the National Classification Scheme creates operational challenges and limitations. The challenges are particularly significant for immersive and interactive environments such as online games. The National Classification Scheme was not created with the online environment in mind, which impacts how the classification criteria applies to online content, such as livestreamed activity, immersive content and conduct and user-generated content.

The reliance on the National Classification Scheme impacts eSafety's ability to address content, as eSafety may need to refer online content to the Australian Classification Board for a formal classification decision. The Classification Board's decision informs eSafety's regulatory actions under the Online Safety Act.

Accordingly, eSafety's recommendation is to remove the connection between the Online Safety Act's Online Content Scheme and the National Classification Scheme. This will promote clarity and alleviate operational challenges due to Part 9's tie to the National Classification Scheme, which currently limits eSafety's ability to address online harms to the definitions of Class 1 and Class 2 material. If this is not feasible, eSafety supports reforming

and modernising the National Classification Scheme, including a specific online-content categorisation to ensure it is fit for purpose in addressing online content.

Conclusion

eSafety is pleased to have the opportunity to provide context and evidence on online radicalisation, from its vantage point as Australia's online safety regulator. We will continue to draw on the levers available to us to support Australian's communities to have safer, more positive experiences online. We welcome the opportunity to provide additional evidence or clarify any aspects of the submission should it assist the Committee's inquiry.