Australian Government | eSafetyCommissioner

eSafety submission

Inquiry into right wing extremist movements in Australia

April 2024

About eSafety

The eSafety Commissioner (eSafety) is Australia's national independent regulator for online safety. Our purpose is to help safeguard Australians from online harms and to promote safer, more positive online experiences.

The *Online Safety Act 2021* (OSA) sets out our legislative functions, which include coordinating online safety activities across the Australian Government, supporting and conducting educational and community awareness programs, and administering four complaints schemes. The issue of online extremism can intersect with eSafety's activities where it involves Class 1 or Class 2 material, abhorrent violent conduct material, cyber abuse or bullying, or online hate.

eSafety welcomes the opportunity to submit to the inquiry. This submission focuses on the following terms of reference, as they relate to eSafety's remit:

- (a)(v) the role of the online environment in promoting extremism and
- (c) measures to counter violent extremism in Australia.

Online extremism

Online extremism can take different forms, and may impact communities in different ways. For example, it may take the form of online hate or abuse, with members of at-risk groups being more likely to encounter online hate or be targeted with online abuse from those who hold extremist views. In other circumstances, some internet users, particularly young people, may be at risk of being radicalised towards extremist views by various types of content and activity they encounter online (as well as offline). Counter-terror experts and former extremists have described hateful and extremist online content and activity as having a propagandising effect that desensitises individuals to hateful attitudes towards perceived outgroups, normalises and glorifies acts of violent extremism, and exposes susceptible individuals to potential recruitment and further radicalisation by established extremist individuals and groups.[1] Similarly, the Royal Commission of Inquiry into the Terrorist Attack on Christchurch Mosques on 15 March 2019 drew a connection between the internet use of the individual who carried out the attack and his extremist ideology.[2]

---

[1] See eg Heather J Williams et al., *Mapping White Identity Terrorism and Racially or Ethnically Motivated Violent Extremism*, Rand Corporation, 2022; Lydia Khalil and Sam Roggeveen, *Rise of the Extreme Right*, Lowy Institute, 2022.

[2] See New Zealand Royal Commission of Inquiry into the Terrorist Attack on Christchurch Mosques on 15 March 2019: The Report: Chapter 3 World travel 15 April 2014 – 17 August 2017, paragraph 2.

eSafety recognises that combating online harm is a global challenge, which cuts across many different disciplines and portfolios. We work as part of a cross-agency, cross-sector, and multi-jurisdictional online safety ecosystem, across our three foundational pillars of prevention, protection, and proactive and systemic change.

Prevention

eSafety works closely with Australian communities to understand their online experiences, to inform our initiatives, and to co-design meaningful resources and support so that they can engage safely and confidently online. While eSafety does not have a specific focus on extremism, some of our research activities and resources are relevant to understanding and preventing online harms which may be related to extremism.

Research

eSafety has not conducted research into ideologically motivated extremism. However, we have conducted some research into online hate. This includes our most recent research release on online gaming, which explores children and young people's experiences online, including their experience of online hate. Findings include:

- 20% of teen gamers had seen or heard other players share or use hate speech.
- 11% had seen or heard other players expressing or sharing misogynistic ideas relating to the belief that men are superior to women.
- 8% had seen or heard other players expressing or sharing ideas that people from one race, culture, religion, or nationality being better than other people.

Our *Online safety in Australia* series looks at adults' experience of online harms. Topline data was released in 2023. A series of reports with more in-depth analysis will be released later in 2024. Themes will include the overall experience and perpetration of online harms, as well as experience of and exposure to online hate with a focus on at-risk communities.

Resources

In relation to extremism which takes the form of online abuse, our resources provide targeted advice on identifying online abuse, ways to prevent further contact from those perpetrating abuse, skills and strategies to deal with online abuse, and how to report different types of abuse to eSafety or other relevant services.

eSafety also has a range of resources to help people develop the critical skills they need to navigate the online world more safely. For example, Young and eSafe is a resource for educators designed to help young people cultivate respect, responsibility, empathy, resilience, and critical thinking in relation to online content and conduct. eSafety has also developed resources for parents and carers, educators, and young people on dealing with distressing and violent content.

Protection

eSafety administers complaints-based regulatory schemes to address:

- cyber abuse material targeted at Australian adults (**adult cyber abuse**)
- cyberbullying material targeted at Australian children (**cyberbullying**)
- non-consensual sharing, or threatened sharing, of intimate images (**image-based abuse**)
- Class 1 and Class 2 material (**illegal and restricted content**).

These schemes cover both natural and synthetic (including AI-generated) online content.

Where the online material reported to us meets the relevant legislated thresholds, eSafety can issue removal notices to the online service on which the content is available and the hosting service provider that hosts the content. In some instances, other enforcement options are also available, as set out in our Compliance and Enforcement Policy.

We work with online service providers and others to achieve positive outcomes for victims and survivors of online harm. This may involve facilitating rapid removal of abusive content posted online; referrals to law enforcement, mental health providers, or legal services; and providing tips and strategies for how to mitigate further harm.

Illegal and Restricted Content

Illegal and restricted content covered by eSafety's online content scheme is divided into the following categories:

- **Class 1 material** includes material that would be refused classification under the National Classification Scheme. This includes online content that advocates terrorist acts, or that promotes, incites, or instructs in matters of crime or violence.
- **Class 2 material** includes material that would be classified R 18+ or X 18+ under the National Classification Scheme. This includes high-impact material that may be inappropriate for general public access and/or for children and young people under 18 years old, such as material featuring high-impact violence, crime and suicide, death, and racism.
- **Material that depicts abhorrent violent conduct** is a subset of Class 1 material, and includes material that depicts, promotes, incites, or instructs in 'abhorrent violent conduct' such as terrorist acts, murder, torture, rape, or violent kidnapping.

eSafety can give removal notices in relation to Class 1 material to the online service where the content is available and the hosting service provider that hosts the content for the service. The vast majority of reports to eSafety relate to child sexual exploitation and abuse material (87%). Less than 1% of reports relate to content that advocates terrorist acts or material that incites or instructs in matters of crime or violence.

For Class 2 material that is X 18+ content and provided from Australia, eSafety can give a removal notice to the online service or the hosting service provider of the service, requiring the recipient to remove the material. Class 2 material that is R 18+ 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.

Complaints relating to this type of content are very rare and to date eSafety has not issued a Class 2 removal or remedial notice.

Under the *Criminal Code 1995* (Cth), eSafety may also give notices 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.

eSafety usually approaches online service providers informally in the first instance, asking them to remove Class 1 or Class 2 material. Informal requests often lead to faster removal of the material compared to formal action, resulting in fewer Australians being exposed to harmful online content. eSafety's powers under the OSA provide further enforcement options if needed.

eSafety actions in response to Class 1 material depicting the Wakeley stabbing

Following the Wakeley stabbing on 15 April 2024, eSafety has been working to respond to the circulation of associated Class 1 material depicting gratuitous or offensive violence with a high degree of impact or detail. eSafety continues to work closely with others across government, as well as with online services, in an effort to remove this material quickly. This has included issuing Class 1 removal notices in some circumstances. In addition, eSafety has provided the community and stakeholders with advice about what they can do if they encounter distressing content.

Online Content Incident Arrangement (OCIA) and ISP blocking

The eSafety Commissioner plays a key role in the whole-of-government Online Content Incident Arrangements (OCIA) framework administered by the Department of Home Affairs. The OCIA is Australia's domestic crisis response protocol aimed at preventing the viral, rapid, and widespread distribution of online abhorrent violent conduct material. It is designed to enhance communication flows and information sharing between government and industry stakeholders during a potential Online Crisis Event (OCE).

Under the OCIA as well as eSafety's protocol with Australian internet service providers, an 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.

In the event of an OCE, eSafety has powers under Part 8 of the OSA 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, for example, after the 2019 Christchurch mosque attacks.

The eSafety Commissioner has not declared an OCE nor implemented the ISP blocking powers since they can into effect, as the use of notice powers has provided a sufficient – and more targeted – way to reduce the risk of this type of online material.

Relationships with the National Intelligence Community

eSafety's work to prevent and minimise online harms is distinct from, but complementary to, the efforts of those working in law enforcement and national security. Accordingly, eSafety engages as required with members of the National Intelligence Community on matters of national security interest.

eSafety has Memoranda of Understanding that support the exchange of information with all state and territory police agencies and the Australian Federal Police. These relate to serious material such as child sexual abuse material, terrorist material, and crime and violence material.

Adult Cyber Abuse and Child Cyberbullying schemes

eSafety has powers to investigate and take action to remove adult cyber abuse and child cyberbullying.

Adult cyber abuse is when someone sends seriously harmful material to a person who is 18 or older, or posts or shares seriously harmful content about them, using an online or electronic service or platform. That material must, as judged by an ordinary reasonable person, be both:

1. likely to have been intended to cause serious harm to a particular Australian adult; and
2. menacing, harassing, or offensive in all the circumstances.

The Child Cyberbullying scheme covers material targeting a child under 18. The material must, as judged by an ordinary reasonable person, be likely to have the effect of seriously threatening, intimidating, harassing, or humiliating a particular Australian child.

These regulatory schemes may be relevant where online extremism takes the form of material targeting a particular Australian, although complaints of this nature are relatively uncommon. Information about the nature of complaints and their outcomes can be found in our annual report.

Proactive and systemic change

eSafety works with the online industry to promote proactive and preventative change to better address the risks of existing and emerging online harms through their systems and processes.

Industry codes and standards for illegal and restricted online content

In addition to the complaints schemes outlined above, eSafety has the power to register industry codes and develop standards for eight sections of the online industry. These industry codes and standards are designed to protect Australians from illegal and restricted online content, by setting obligations for relevant industry sectors to proactively deal with this material at a systemic level. Like the four investigations schemes, the industry codes and standards apply to online content whether or not it is generated by AI.

The industry bodies tasked with developing the industry codes adopted a two-phase approach. The first phase targets the most seriously harmful Class 1 material, including child sexual exploitation material, pro-terror material, and extreme crime and violence material.

To date, eSafety has registered six industry-developed codes for the first phase, which focuses on Class 1 material such as pro-terror material, which advocates terrorist acts as defined in the Criminal Code, outside of public debate, entertainment, or satire.[3] The codes apply to social media services, app distribution services, hosting services, internet carriage services, equipment providers, and search engine services.

Examples of measures that social media services must now take to prevent and address pro-terror material include:

- Using systems, processes and/or technologies to detect and remove certain types of pro-terror material.
- Implementing systems, processes, and technologies that enable the provider to take appropriate enforcement action against end-users who breach policies prohibiting pro-terror material.
- Providing tools which enable Australian end-users to report, flag, and/or make a complaint about pro-terror material accessible on the service.

In addition, search engines which incorporate generative AI functionality are now required to:

- improve systems, processes, and/or technologies that aim to reduce the safety risks to end-users concerning synthetic materials generated by artificial intelligence.
- research detection technologies that assist end-users in identifying deep fake images that are accessible from the service.
- make clear when a user is interacting with any features using artificial intelligence, for example, to generate search results in the form of longer form answers, summaries, or materials.

eSafety declined to register two codes developed by industry bodies and has drafted industry standards for Relevant Electronic Services (such as messaging and gaming services) and Designated Internet Services (such as websites offering generative AI functionality which meet the proposed threshold, and online file and photo storage services). eSafety is now closely considering submissions received during consultation, and what amendments should be made before finalising the standards.

The second phase of industry codes development, focusing on children's access to Class 2 material, is expected to formally begin once the first phase concludes. In addition to content like

[3] Pro-terror material is defined in Annexure A to the head terms of the consolidated industry codes of practice.

online pornography, Class 2 material also includes high-impact material such as high-impact violence and themes like crime and suicide, death, and racism.

Basic Online Safety Expectations

The Basic Online Safety Expectations are designed to improve online service providers' safety standards, transparency, and accountability. Services are required to have terms of use, policies, and procedures to ensure the safety of users. Services are also required to take steps to ensure that penalties for breaches of their terms of use are enforced against all accounts created by the end-user.

With regard to extremism, the *Online Safety (Basic Online Safety Expectations) Determination 2022* (the Determination) establishes expectations that service providers will:

- take reasonable steps to ensure that end-users are able to use the service in a safe manner (s6(1)) and to proactively minimise the extent to which material or activity on the service is unlawful or harmful (s6(2)).
- take reasonable steps to minimise certain material, including Class 1 material and material that depicts abhorrent violent conduct.
- ensure they have terms of use, policies, and procedures in relation to the safety of end-users, as well as policies and procedures for dealing with reports and complaints (s14).
- take reasonable steps to ensure that penalties for breaches of their terms of use are enforced against all accounts held by those end-users (s14(2)).

Additionally, the Explanatory Statement to the Determination states that services should use their terms, policies, and procedures to address harmful material that is not necessarily unlawful or explicitly referenced in the OSA, such as online hate against a person or group of people on the basis of race, ethnicity, disability, religious affiliation, caste, sexual orientation, sex, gender identity, serious disease, disability, asylum seeker/refugee status, or age.

The OSA provides eSafety with powers to require online service providers to report on the steps they are taking to comply with any or all of the Basic Online Safety Expectations. When deciding which providers to give a notice to, the OSA requires eSafety to have regard to specified criteria. These criteria and other considerations are summarised in the BOSE Regulatory Guidance. The obligation for services to respond to a reporting requirement is enforceable and backed by civil penalties and other enforcement mechanisms. Information obtained from the reporting notices are published in transparency summaries where appropriate.

On 21 June 2023, eSafety issued a reporting notice to Twitter (subsequently X), requiring it to explain what it is doing to minimise online hate, including how it is enforcing its terms of use and hateful conduct policy. A summary of the response was published on 11 January 2024. Notably, the response showed that as of May 2023, no tests were conducted on Twitter recommender systems to reduce risk of amplification of hateful conduct, and URLs linking to websites dedicated to harmful content were not blocked on Twitter. Additionally, relevant trust and safety and moderation staffing levels had decreased following the company's acquisition in October 2022.

On 18 March 2024, eSafety issued reporting notices to Google, Meta, WhatsApp, Reddit, Telegram, and X (formerly Twitter) requiring them to report on the steps they are taking to tackle the risk of terrorist and violent extremist material and activity on their services. The notices require answers to questions about the tools, processes, and resources they use to ensure safety. eSafety will publish appropriate information on the findings to improve transparency and accountability.

The Department of Infrastructure, Transport, Regional Development, Communications and the Arts is currently considering feedback received through its consultation on an amended Determination, with new expectations regarding the safety of generative AI and recommender systems. The draft amended determination also proposes that detecting and addressing hate speech, which breaches a service's terms of use, is a reasonable step to ensure safe use of a service.

Safety by Design

eSafety also aims to produce positive outcomes for Australians by guiding and supporting the online industry to enhance 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 three guiding principles:

1. **Service provider responsibility**: The burden of safety should never fall solely upon the user. Every attempt must be made to ensure that online harms are understood, assessed, and addressed in the design and provision of online platforms and services.
2. **User empowerment and autonomy**: The dignity of users is of central importance. Products and services should align with the best interests of users.
3. **Transparency and accountability**: Transparency and accountability are hallmarks of a robust approach to safety. They not only provide assurances that platforms and services are operating according to their published safety objectives, but also assist in educating and empowering users about steps they can take to address safety concerns.

A Safety by Design approach can seek to address myriad online safety issues, including extremism, by promoting a proactive approach to user safety. For example, it could include measures to respond to extremism and hate, such as:

- having individuals or teams accountable for the creation, evaluation, and implementation of relevant policies.
- putting tools and processes in place for detecting and actioning content that violates these policies.
- ensuring that community guidelines and reporting processes are accessible and easy to understand.

- carrying out open engagement with a wide user base including independent experts and key stakeholders.
- committing to consistently innovate and invest in safety-enhancing technologies.
- publishing information about safety tools, policies, and processes, and their impact and effectiveness.
- ensuring that design features and functionality preserve fundamental user and human rights.

Practical resources are provided via the Safety by Design 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.

We have focused our Safety by Design work on diverse, marginalised, and at-risk groups to make sure their needs are effectively considered, incorporated, and actioned in the design of online products and services. We also consider that education and empowering people will always form the basis for addressing the social and behavioural issues that manifest online.

Tech Trends: recommender systems and algorithms, and generative AI

To make sure our content and programs reflect current information, technological developments and global trends, eSafety monitors new and emerging trends through our Tech Trends workstream.

In December 2022, we released a paper on recommender systems and algorithms. This was developed based on extensive consultation with academics and other subject matter experts and summarises eSafety's approach to these systems, including proactive safety measures for industry and advice for users.

Recommender systems, and their underlying algorithms, are integral to many online services, which use them to amplify, prioritise, and recommend content and accounts to their users, and to deliver relevant search results. They 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.

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.

Collaborative efforts to regulate digital platforms

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.

DP-REG's current strategic priorities include assessing the impact of algorithms, improving digital transparency, as well as understanding and assessing the benefits, risks and harms of generative AI and how the technology intersects with the regulatory remit of each DP-REG member. In 2023, 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.

Review of the Online Safety Act 2021

The Australian Government has announced an independent review of the OSA to be conducted during 2024. The terms of reference indicate the review will be broad ranging, and will include consideration of eSafety's existing statutory schemes, including those outlined above. It will also consider whether additional arrangements are warranted to address online harms not explicitly captured under the existing statutory schemes (such as online hate), or potential online safety harms raised by emerging technologies like generative AI and recommender systems.

The review will include a period of public consultation, commencing with the release of an Issues Paper in the first half of 2024. eSafety will be closely working with the Australian Government to ensure the OSA remains fit for purpose and adequately reflects Australians' needs and expectations. The Final Report of the Review is expected to be provided to the Minister for Communications by 31 October 2024.