Updated Position Statement

End-to-end encryption

October 2023

Australian Government | eSafetyCommissioner

eSafety.gov.au

Updated Position Statement – End-to-end encryption

Contents

Definitions

What is encryption?

Encryption is a way to prevent unauthorised access to information and make it readable only by a person who has the 'key'. Encryption is not new and, in its modern form, has been used for more than 40 years as a tool to promote privacy and security. It is primarily employed for the secure transmission and storage of information and can help to prevent data breaches and hacking. It facilitates legitimate, positive, and safe communication.

Virtually all mainstream communications are encrypted 'in transit', which involves protecting the information as it travels over the internet, for example from your mobile device to a company's servers. There is also encryption to protect data 'at-rest', for example on a hard drive, mobile device, or in cloud storage.

What is end-to-end encryption?

End-to-end encryption (E2EE) describes a means of securing communications from one device, 'sender', or 'end point', to another intended recipient. E2EE transforms standard text, imagery, and audio into an unreadable format while it is still on the sender's system or device. Unlike standard encryption in transit, data transmitted via E2EE can only be decrypted and read once it reaches its final destination, rather than by an intermediary. For these reasons, it is seen as more secure.

Popular examples of services that use E2EE include iMessage, WhatsApp, Signal, and parts of Skype and Telegram. Meta is in the process of implementing E2EE across its Facebook and Instagram messaging services, and has also begun testing its use on the "Quest" virtual reality platform.

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 could be used for multiple purposes and often plays an important role in preventing spam and malware, as well as detecting illegal content like child sexual exploitation material (CSEM), child sexual exploitation and abuse (CSEA), and terrorist and violent extremist content (TVEC), as detailed below.

Online safety risks

E2EE presents online safety challenges as it can create digital hiding places that can enable significant individual and community harm, both physical and psychological, as well as limiting the ability of services, regulators, and law enforcement to investigate wrongdoing.

For example, E2EE can prevent or limit the detection of online CSEA and the grooming of children by blocking technologies that can identify illegal material and activity. As a result, E2EE can conceal the production, storage, exchange, and proliferation of CSEA, enabling offenders' abuse to go undetected and allowing them to continue connecting with and abusing new victims. It also exposes survivors of past abuse to ongoing trauma as images of their abuse continue to be circulated.

eSafety has received more than 130,000 reports about illegal and restricted content since 2015, with 3 in 4 about CSEM. According to the US-based National Centre for Missing and Exploited Children (NCMEC) in 2022, of approximately 32 million reports, more than 31.8 million CSEA reports were made from electronic service providers, most of which check for this content using well-established photo matching technologies.[i] These technologies involve checking if content on a service matches the unique 'digital fingerprint' of previously confirmed CSEM. The error rate of these technologies is 1 in 50 billion to 1 in 100 billion.[ii] Services then report this content to designated organisations such as NCMEC – enabling material to be tagged, traced, and removed across the internet.

NCMEC estimates that 80% of reports – i.e. 25 million reports – would be lost if the electronic service providers that report large numbers of CSEA were to implement default E2EE.[iii]

E2EE can also reduce or restrict the ability for services to ensure user safety or to act on cyberbullying, cyber abuse, and non-consensual sharing of intimate images. If messages were E2EE, platforms would need to rely on proactive detection tools on unencrypted parts of a service, or wait for users to report after the harm has been done.

Relationship to privacy and security

E2EE can strengthen the privacy and security of online social interactions and information exchanges. This can mean greater security of data while a communication is in transit, and stronger defence against efforts to tamper with a communication, as any messages altered in transit cannot be decrypted and viewed.[iv] E2EE is increasingly being adopted and promoted by services for their message functions to demonstrate a commitment to user privacy and to build user trust.

Benefits of E2EE: protecting activists online

Many users can benefit from the added security of E2EE, but it is particularly important for people who might be persecuted if their communications were revealed. E2EE has the potential to help human rights activists to keep sensitive information confidential and secure from authoritarian governments. For example, activists who organise or communicate with one another online, or use the internet to raise awareness of human rights abuses, can be at risk from attempts to compromise their accounts.

The public key encryption process used in E2EE is a valuable way of providing resistance against hostile uses of surveillance and spyware. E2EE, when used in combination with other account authentication and cyber security measures, can help defend against security breaches that could have dangerous consequences for vulnerable users.

Limitations of E2EE

E2EE does not provide perfect privacy and security. Most E2EE services continue to collect identifying information or metadata, which can be used by platforms to track users in ways they may not anticipate, for example, by tracking the businesses that users engage with and monetising this information.

In addition, services that claim to be E2EE may only be referring to encryption of data while it is in transit, with data potentially unencrypted at the start and end points. Lack of awareness of this may lead many users into a false sense of security when using E2EE. Furthermore, E2EE services and communications can also be vulnerable where the security of an endpoint becomes compromised. There is also always a possibility that a recipient of an encrypted message may report or share the communication received.

While E2EE can be presented as a critical standard to protect privacy and build users' trust in online services, in practice, it can be applied differently across platforms and with varying effectiveness. Services are sometimes opaque about where and what data is encrypted. In the absence of meaningful transparency or auditing of E2EE services against an agreed standard, services can use E2EE as a marketing tool to attract users rather than an objective standard of safety, security, or privacy.

eSafety's approach

E2EE provides offenders with an increased level of protection from being detected and held accountable, while decreasing protections for victim-survivors. While E2EE can be an important measure for protecting sensitive information, offenders can also exploit it to hide their illegal actions, allowing ongoing abuse and further harm to victims, as well as exposing survivors to re-traumatisation if their images continue to circulate online.

eSafety's position is that safety, privacy, and security are not mutually exclusive, and each can be maintained through thoughtful and intentional design. **eSafety does not expect companies to design systematic vulnerabilities or weaknesses into E2EE services. Our focus is on working with industry to find proactive and systemic solutions, including the prevention and detection of illegal content. Deployment of E2EE services does not absolve platforms and services of responsibility for hosting illegal content and facilitating the sharing of CSEA and other harmful material.**

Where complaints through our reporting schemes relate to services with E2EE, eSafety would expect the service provider to take reasonable steps to prevent further harm, for example, by warning, suspending, or closing the responsible account.

Through the Basic Online Safety Expectations, eSafety can seek information on how encrypted and E2EE services are taking reasonable steps to develop and implement processes to detect and address unlawful or harmful material or activity. Greater transparency through the Basic Online Safety Expectations will help to improve understanding of the harms occurring and the steps available to industry, including those available to providers which offer E2EE services.

For example, eSafety asked Apple and WhatsApp about their implementation of E2EE in notices issued in August 2022. Information provided in response to the notices confirmed that:

- Apple takes no steps to detect known CSEA in iMessage, or to detect livestreamed CSEA on FaceTime. There is no in-service user reporting option, although Apple said users could report via an email address.
- WhatsApp checks for known and new CSEA on parts of its service that are not E2EE (such as profile pictures, group names and descriptions, and user reports). WhatsApp also confirmed to eSafety that, although it checks links sent in E2EE messages for "suspicious content", this process is a simple "rule-based function" that 'happens in the context of an E2EE chat, where WhatsApp has no ability to access the message content.' Other than this 'suspicious link detection', WhatsApp confirmed that it does not perform any checks on any other content for suspicious material or malware, for example in a gif, file, or photo.[v]

In Australia, six industry codes to regulate class 1 content (including CSEM and TVEC), have been registered by the eSafety Commissioner. More details on the registered industry codes are available on the eSafety register for industry codes and industry standards. Industry Standards will be developed by eSafety for other sections of the online industry, including 'relevant electronic services' which include open and closed communication such as text messaging, email, and private online messaging, including E2EE messaging.[vi] The commitments or requirements in industry codes and industry standards are enforceable and apply to all industry participants within the relevant online section and in the scope of code or standard requirements.

Other jurisdictions across the globe are introducing legislation with safety requirements on E2EE services.

eSafety will continue to examine developments in E2EE, monitor global trends, collaborate with experts and stakeholders, and facilitate proactive and systemic change by advocating for Safety by Design and the implementation of harm mitigations at the earliest stages of development.

Guidance for industry

When designing new features and products, user safety should be considered alongside privacy and security, all of which are critical to user trust and retention. One poll undertaken by YouGov for the National Society for the Prevention of Cruelty to Children in 2021 found that users' support for E2EE actually increases if it is rolled out with children's safety protected.[vii] In particular, the vast majority of respondents supported services having the technical ability to detect child abuse images (92%) and to detect adults sending sexual images to children (91%).

Safety by Design

eSafety suggests all services, including those that are E2EE, should adopt a Safety by Design approach. Safety by Design encourages online services of all sizes to anticipate, detect and eliminate online safety risks – making our digital environments safer and more inclusive, especially for those most at risk, such as children. All services, whether encrypted or not, should offer critical features like user reporting. Other steps that are available to most E2EE services include:

- Requiring multi-factor authentication and user sign up information, where users register for the service using a phone number, email address, or other identifier. While multi-factor authentication provides individuals with account security, the process of account authentication (for example, through an authentication link), can add friction to the service and help prevent recidivism for accounts that have been identified as breaching the law or terms of service. By requiring information and identifiers to create an account, multi-factor authentication may help reduce the risk of a banned user creating a new account with the service, including anonymous and pseudonymous accounts.
- Making design choices such as ensuring that E2EE is opt-in by default, rather than opt-out. This would require people to choose E2EE should they wish to use it, therefore allowing certain detection technologies to work for communication between users that have not opted in to E2EE.
- Using nudges and prompts that educate users, including those most at-risk of harms, on risks such as accepting friend requests from people they do not know.[viii]
- Providing nudges, prompts, and warnings to users to warn when sending or receiving images that contain nudity. These checks can happen on users' devices and do not undermine E2EE. Services could incorporate safety

features such as interstitial warnings that blur or block content, and flag safety and reporting information.

- Restricting or limiting a user's ability to share material with large numbers of people instantaneously.
- Providing accessible user reporting so that people can seek help and alert services to harmful content.[ix]
- Using classifiers, metadata, and behavioural signals to flag accounts for review.
- Continuing to detect illegal content on the parts of their services that are not E2EE such as profile pictures and group names, by using photo matching, machine learning, artificial intelligence, and other detection technologies.
- Monitoring and restricting the use of joining and invite links for encrypted services, which are often shared on unencrypted online spaces to facilitate the exchange of CSEA and TVEC.[x]
- Reducing the ability for anyone to discover groups consisting of predominately children.
- Ensuring appropriate escalation pathways are in place for reports of CSEA and other harmful material – including to law enforcement and relevant experts.
- Explicitly and publicly stating how they employ measures to deter the use of their services for sharing harmful content like CSEA and TVEC as part of enhancing overall transparency.[xi]
- Working with law enforcement and relevant experts (for example, experts in relation to CSEA and TVEC) to identify and block access to E2EE channels associated with illegal activity.

Any integration of E2EE should only occur after a thorough risk assessment has been completed to minimise the potential for harm across communication channels, and to ensure the full range of safety, privacy, and security factors have been considered and potential harms mitigated. If a service chooses to introduce E2EE, it should be transparent about how it will manage risks to user safety. As with any update to a service's features or functionalities, approaches to E2EE require careful design to prevent the introduction of new vulnerabilities with consequences for safety, privacy, and security.

Tackling CSEA

There are many standard trust and safety measures that can be introduced to prevent and mitigate a wide range of online harms. However, it is important to consider the unique nature of CSEA in determining whether such measures are likely to be sufficient to prevent and address this particular form of violence.

For example, providing user reporting mechanisms and monitoring behavioural signals are key trust and safety measures, which many services can, do, and should employ to combat harms. But there are limitations of relying exclusively on these measures to prevent CSEA:

- **Services should make user reporting easier and more accessible, but this should not shift sole responsibility to victims for their own safety.** While easy-to-access user reporting is an essential feature, children and other vulnerable users cannot be expected to carry the burden of reporting their abuse. Often they aren't aware that images of their abuse are being shared, or they may feel ashamed or scared to report.
- **Child sexual exploitation and abuse offenders will not report each other.** E2EE services risk creating safe environments for offenders to share imagery or content between themselves. Neither reporting tools nor metadata tools are effective at detecting offender-to-offender communication. Photo matching is a critical tool for uncovering and reporting content and banning users. A single known image is often the tip that uncovers broader offending, including contact abuse.
- **A combination of tools is needed to prevent harm at scale.** While it is important to enable reports to be made about specific accounts and items of content, photo matching and other content-based tools can detect, remove, block, and report content at scale, and with high rates of accuracy. Service providers should be employing technological solutions that support detecting CSEA material proactively.
- **Reporting options should be backed up by measures which enable services to investigate.** Currently, where services have user reporting options, this typically results in only a couple of recent messages being accessible to moderators. This is unlikely to provide enough context and information to confirm abuse such as grooming is taking place.
- **Behavioural signals, such as an adult contacting multiple children, are crucial complementary measures but are insufficient replacements** for photo matching techniques, or other tools that assess the content of communications. As behavioural tools only give a probabilistic assessment of abuse, they should be used alongside other tools and techniques. Services with limited internal expertise should consider deploying third party safety tech and content moderation tools.

Accordingly, these measures should be applied as part of a broader set of risk mitigations to protect the rights and best interests of children.

Case study: existing tools

It is argued by some that the detection of illegal or harmful content cannot occur without "breaking E2EE", or "building a backdoor". This is overly simplistic and not representative of the technical options available to industry that do not require breaking or weakening E2EE. Indeed, several services already detect harmful content within E2EE services to varying degrees. For example, some services scan for malware on users' devices alongside E2EE and promote the privacy-preserving benefits of these approaches.

Apple has already deployed solutions to detect some harmful content within an E2EE service, namely its Communication Safety feature for iMessage. Apple uses machine learning to warn children when receiving or sending photos that contain nudity. These checks happen on users' devices and do not undermine E2EE. Although Apple's intervention is limited in that it does not prevent the sharing of illegal material or activity, or enable accounts to be banned by the service, it demonstrates – at scale – that device side tools can be used alongside E2EE, without weakening encryption and while protecting privacy.

Apple previously announced proposals to roll out device-based scanning of iCloud Photo uploads for known CSEM using its NeuralHash algorithm, which reportedly had a negligible false positive rate of 1 in a trillion when combined with a threshold of CSEM.[xii] Apple has since withdrawn its plans.

Emerging good practice

There are existing measures, as well as emerging technologies, which can also assist in detecting harmful and illegal activity online without weakening encryption. How best to apply these measures will depend on the harms being considered, but we strongly encourage all services to take a Safety by Design approach. This means protections should be considered and built in from the design stage, not retrofitted once harm has been done.

Services should invest in innovative approaches to prevent the dissemination of illegal and harmful content and to detect illegal and harmful content – without compromising privacy and security – to ensure user safety on E2EE services.

There are several existing and emerging methods for proactively detecting CSEA that can be utilised by online services alongside E2EE:

- **Proactive, device-based detection tools operating at the point of transmission,** rather than during transmission. These "client side" tools can involve photo-matching for known imagery and video content, and be implemented by a specific application, or offered by device operating systems as an opt-in tool. Communication services already commonly use client-side scanning to detect malware. There are some concerns as to whether this technology can be securely deployed at scale. Nevertheless, the technical directors of the UK's National Cyber Security Centre and Government Communications Headquarters have suggested that "client-side scanning techniques" can be "implemented safely in many situations."[xiii]
- **Proactive detection tools that operate within dedicated 'secure enclaves'.[xiv]** This proposal would involve an E2EE communication being sent from a user's device, and then checked for known CSEM using photo matching tools by a dedicated server, before the communication is sent on to the recipient. Such a system can be audited to ensure it fulfills only one function, and that no access is provided to the service provider or other third parties. This avoids the need to store hashes on a users' device, or for the computation to take place on a user's device.

- **AI and proactive technical solutions can also operate at device-level** – such as nudity detection tools combined with age estimation tools or grooming classifiers. These tools, which may be enabled through on-device machine learning, have a higher false positive rate than photo-matching tools, in part due to a lack of training data.[xv] Hence these tools are predominantly used to help prioritise content for review or to issue safety prompts and warnings to users, particularly children. Warnings can also deter people who are sharing CSEA for reasons other than having an explicit interest in this material, for example out of anger, reminding them that possession is illegal.[xvi]
- **Proactive scanning tools based on homomorphic encryption.** Homomorphic technology performs photo hashing on encrypted data without the need to decrypt the information. This technology is also being researched as a way to enable targeted advertising in encrypted environments.[xvii] While promising, it is currently difficult to effectively scale forms of homomorphic encryption to operate at required internet scale and speeds, and across the diversity of user devices.

Robust safeguards and transparency and audit mechanisms are important to ensure tools are only used for their stated purposes.

Each of these potential solutions carry trade-offs, which can be managed through taking into account the level of intervention needed in proportion to the risk of harm. Just as E2EE can take various forms, so too can the implementation of proactive tools. For example, services can combine behavioural signals and device-based photo matching by relying on signals to flag suspicious accounts, and then only including those devices in the photo-matching process.

These tools and processes are developing and improving over time. It is important that as well as taking practical and effective steps now to tackle CSEA and other harmful content on encrypted services, online service providers should monitor and invest in technology and systems that enhance safety protections and detect such content.

While further technical development and greater transparency may be needed for some solutions, these examples demonstrate that content-based detection is possible alongside E2EE. As with any technical intervention, our collective focus should be on mitigating potential negative consequences, while maximising the benefits for the child, user, and public safety.[xviii]

Guidance for users

Users are advised to take extra care when communicating on E2EE services, particularly when they do not know the person they are communicating with. These services can also heighten and conceal the risk of children and adults co-mingling.

Anyone who encounters child sexual exploitation images, videos, livestreaming, or other content that incites the production or sharing of this type of material on encrypted services should report it to the relevant platform and to eSafety.

- You can make a report to eSafety at **esafety.gov.au/report**.

People who encounter other abuse on E2EE services where there is an Australian connection, including serious cyberbullying, adult cyber abuse, and threatened or actual sharing of intimate images, should report it to the relevant platform and to eSafety, and block or mute the individual or account to prevent further contact.

- You can collect evidence* and make a report to eSafety at **esafety.gov.au/report**.

If you suspect a child is a victim of online child exploitation or grooming you can report to the Australian Federal Police via the Australian Centre to Counter Child Exploitation (ACCCE).

- In Australia the police are a first point of call to ensure an operational response can be prioritised. This includes calling 000 if there is immediate danger to a child, reporting via the Australian Centre to Counter Child Exploitation's online reporting portal (https://www.accce.gov.au/report), or reporting in confidence to Crime Stoppers (https://crimestoppers.com.au/).

*Please note that possessing, creating, or sharing intimate images or videos of people who are under 18 may be unlawful, even if they are intended as evidence of cyberbullying, adult cyber abuse, or image-based abuse. For more information about relevant laws in Australia, visit Youth Law Australia.

Date published: 13 October 2023

[i] National Center for Missing & Exploited Children, *2022 Cyber Tipline Reports by Electronic Service Providers (ESP)*, National Center for Missing & Exploited Children, 2023, accessed 25 August 2023.

[ii] Hany Farid, *'Statement to the House Committee on Energy and Commerce Hearing: Fostering a Healthier Internet to Protect Consumers'*, 2019, accessed 15 September 2023.

[iii] Michelle DeLaune, *United States Senate Committee on the Judiciary, "Protecting Our Children Online"*, Testimony of President and CEO National Center for Missing & Exploited Children, 14 February 2023.

[iv] Ben Lutkevich, Madelyn Bacon, *DEFINITION end-to-end encryption (E2EE)*, Tech Target, 19 January 2023.

[v] See eSafety Industry Basic Online Safety Expectations website material https://www.esafety.gov.au/industry/basic-online-safety-expectations/responses-to-transparency-notices

[vi] Please note, the minimum compliance measures set out in the industry codes and future industry standards represent the mandatory and enforceable steps that industry must meet to comply with their obligations in relation to class 1 and class 2 material.

[vii] NSPCC, *Private messaging and the rollout of end-to-end encryption*, discussion paper, 2021.

[viii] Laura Draper, *Protecting Children in the Age of End-to-End Encryption*, Joint PIJIP/TLS Research Paper Series, 80, 2022.

[ix] Note C3P developed CSEM recommendations to clarify and streamline the process for users to report CSEA. These include (1) creating reporting categories specific to CSAM and online CSEA, (2) including CSAM-specific options in easy-to-locate reporting menus, (3) ensuring reporting functions are consistent across the platform, (4) allowing users to report content that is visible without creating or logging in to an account, and (5) eliminating mandatory personal information fields in reporting forms. (see Canadian Centre for Child Protection, Reviewing Child Sexual Abuse Material Reporting Functions on Popular Platforms, 2020, https://www.protectchildren.ca/pdfs/C3P_ReviewingCSAMMaterialReporting_en.pdf)

[x] *Terrorist Use of End-to-End Encryption: Insights from a Year of Multi-Stakeholder Discussion*, Tech Against Terrorism, accessed 23 January 2023; *End (-to-End Encrypted) Child Sexual Abuse Material*, Cyber Peace Foundation, 2 October 2020.

[xi] *Terrorist Use of End-to-End Encryption: Insights from a Year of Multi-Stakeholder Discussion*, Tech Against Terrorism, accessed 23 January 2023.

[xii] Apple, *CSAM Detection: Technical Summary*.

[xiii] Levy, Ian, and Crispin Robinson, Thoughts on child safety on commodity platforms, *arXiv*, 2022, doi = 10.48550/ARXIV.2207.09506.

[xiv] Hany Farid, *An Overview of Perceptual Hashing*, Journal of Online Trust and Safety, 1(1), 2021, https://doi.org/10.54501/jots.v1i1.24.

[xv] Laura Draper, *Protecting Children in the Age of End-to-End Encryption*, Joint PIJIP/TLS Research Paper Series, 80, 2022.

[xvi] Laura Draper, *Protecting Children in the Age of End-to-End Encryption*, Joint PIJIP/TLS Research Paper Series, 80, 2022.

[xvii] Kris Holt, *Facebook is reportedly trying to analyze encrypted data without deciphering it*, Engadget, 3 August 2021; Toubiana, Vincent, Arvind Narayanan, Dan Boneh, Helen Nissenbaum, and Solon Barocas, *Adnostic: Privacy preserving targeted advertising,* In Proceedings Network and Distributed System Symposium. 2010.

[xviii] In Australia this is articulated as an expectation within the Basic Online Safety Expectations (see s8). Section 8 applies to services that are encrypted in any form, including those using 'in transit' encryption such as Transit Layer Security, encryption at rest, and those using end-to-end encryption. The reasonable steps that a provider should take to develop and implement processes to detect and address material or activity that is unlawful or harmful may depend on the nature of the encryption implemented on the service, and whether encryption is utilised on some, or all parts of a service. Services which utilise encryption in transit and/or at rest, and which have a high risk of **unlawful** material and activity occurring on their service, should take reasonable steps to detect this material through both algorithmic moderation processes such as hash matching, AI classifiers and human review. Further details are set out in the Basic Online Safety Expectations guidance on the section 6 and 11 Expectations, 2023, https://www.esafety.gov.au/about-us/who-we-are/regulatory-schemes. Find out more about the Online Safety (Basic Online Safety Expectations) Determination 2022 – referred to as 'the Determination' – and read the explanatory statement on the Federal Register of Legislation at legislation.gov.au.

eSafety.gov.au