AFP Submission to the eSafety Industry Standards public consultation – January 2024

Content warning: This submission includes statistics and information relating to child sexual exploitation and abuse.

AFP

afp.gov.au

1) Introduction

The AFP is committed to our vision of policing for a safer Australia. The AFP's key priorities are to protect lives, protect Australia's way of life and maximise our impact on the criminal environment. As a policing agency with a global footprint, the AFP leverages its relationships with international law enforcement partners and private industry to gather critical insights, information and intelligence into threats. The AFP is also committed to the innovative use of technology to disrupt criminal activity and will continue to build on this success by applying inventive and practical approaches. The AFP will continue to work across government and with industry to develop specialist technical capabilities to detect, deter and disrupt crime at the first possible instance. The AFP also acknowledges that any use of technology must fall within ethical guidelines and be consistent with legal requirements and community expectations.

The AFP acknowledges our close working relationship with the Office of the eSafety Commissioner (eSafety) and thanks eSafety for their efforts in drafting *the Industry Standards for Relevant Electronic Services and Designated Internet Services* (Industry Standards). It is crucial that registered Industry Standards are regularly reviewed and tested to ensure they are fit for purpose, noting technological advances and the emergence of new capabilities for digital industry. This draft of the Industry Standards will be useful to increase transparency and accountability for digital industry and reduce accessibility and exposure to harmful material online. The AFP is supportive of any methodology which asks digital industry to take proactive steps to create, monitor and maintain a safe online environment.

The AFP's submission contains comments relating to Class 1A and Class 1B material defined by eSafety as child sexual exploitation material, pro-terror material, extreme crime and violence material. The following material was considered in completing the AFP submission:

- The two draft Industry Standards (Relevant Electronic Services and Designated Internet Services)
- The two eSafety Fact Sheets
- The eSafety Discussion Paper
- A discussion between members of the AFP and eSafety about the draft Industry Standards conducted by teleconference on 14 December 2023.

The AFP acknowledges online child abuse material (CAM) is becoming more prevalent, commodified, organised and extreme. In the 2022/23 financial year, the AFP-led Australian Centre for Countering Child Exploitation (ACCCE) received 40,232 reports of online child exploitation. The number of young people being investigated by the Joint Counter Terrorism Teams is increasing across several Australian state and territory jurisdictions. Individuals as young as 12 years of age are adopting violent extremist ideologies.

Advancements in technology, including artificial intelligence (AI) and end-to-end encryption (E2EE) will further impact on law enforcement ability to identify offenders online. Since 2013, the AFP has seen a continued increase in the volume of unintelligible E2EE communications due to the increased adoption of E2EE. In the 2022/23 financial year, 96.1% of the 61.4 million 'sessions' intercepted by the AFP used E2EE or other forms of encryption. This represented an increase from 90% in the 2018/19 financial year.

The AFP notes eSafety's recognition of the limitations within certain technology, such as those which offer and use E2EE, when detecting and removing CAM and pro-terror material. Further, the

AFP welcomes the ongoing evolution of these types of technology, noting that "as detection technologies are developed and tested, relevant electronic services currently unable to meet the requirements in sections 21 and 22 (including because the service is end-to-end-encrypted) may find that detection becomes feasible."[1]

The AFP acknowledges the declaration by eSafety that privacy and safety are not mutually exclusive, and can both be maintained through good design. For example, the Industry Standards reflect the regard held for online privacy for Australians with file storage and do not require services to continuously monitor private communications.

2) Definitions and terminology

The AFP notes definitions used in the Industry Standards refer to 'child sexual exploitation material' and 'child sexual abuse material'. In identifying child abuse material, the AFP relies upon the definition of CAM as it appears in the *Criminal Code Act 1995* (Cth) (Criminal Code). It may be useful to ensure consistency in the definition of CAM through amendments to the Industry Standards.

Additionally, the Industry Standards refer to "threat to life or physical safety". From an Australian law enforcement perspective, the Criminal Code also refers to "threat to life or serious harm". It may be useful to align the terminology to reduce any confusion with digital industry on what material is covered by the definitions.

The AFP recommends a review and further clarify on the definition of "enforcement authority", which currently includes a police force or other Law Enforcement Agency (LEA), or other organisations including NGOs, who receive and forward reports to LEA. The AFP is of the view that providers should be required to report only to LEAs who are able to identify and take action for a child in imminent danger of sexual exploitation. If services are permitted to report to non-LEA entities, who in turn report to an LEA, an unnecessary step may be created in the reporting chain, resulting in delays to investigation, identification and rescue of child victims.

The Industry Standards refer to situations where Relevant Electronic Services and Designated Internet Services may not be able to 'detect and remove' CAM or pro-terror material. In those instances, the services are required to 'take appropriate alternative action' to deter and disrupt the distribution of the material. Examples of alternative action include using hash matching; machine learning; AI; and other detection technologies on parts of the service that are not encrypted. The AFP would recommend including reporting to LEA as an alternative action noting that LEA may have supplementary information and ability to detect and remove the material. Even if the service is unable to ascertain whether material is CAM or pro-terror material, the service should still be encouraged to report the material to LEA to allow an appropriate assessment to be conducted and intelligence information to be gathered.

The Industry Standards and explanatory material contain many references to hash matching as a technology to deter and disrupt CAM and pro-terror material. The AFP utilises both hash matching and actual viewing of material to ascertain whether the material is CAM. To ensure correct identification at least two AFP members must agree that the material is CAM before it is designated and uploaded into the Australian Victim Identification Database. It may be useful for

[1] Relevant Electronic Services Fact Sheet, page 8.

the Industry Standards to provide further explanatory material on the requirements of the hash matching technology. For instance, scheduling of hash database updates, how the hash sets are governed and checked for authenticity and accuracy and the specifications and processes for adding hashes to the database. Further, some services maintain commercially available hash sets however, this may raise issues for smaller services that could not afford to maintain a hash database.

Additionally, consideration should be given whether the distinction between "known" and "not known" CAM and pro-terror material is needed throughout the Industry Standards given the main objective is to prevent exposure to any material of CAM and pro-terror nature.

3) Scope of industry standards

eSafety's Discussion Paper on the Industry Standards discusses the inclusion of online gaming platforms into the Industry Standards. It is noted, eSafety draws a distinction between those gaming services that have the ability to allow users to simply communicate with each other and those which go further and have multi-functional aspects to communicate between users. From a legal standpoint, it would be possible for a user to commit an offence – for example, Use of a Carriage Service for Child Abuse Material contrary to s.474.22(1) of the Criminal Code – utilising a gaming service communication function. The AFP confirms it has seen examples of online grooming via in-game chat and agrees it is appropriate to include gaming services with communication functionality in the Industry Standards.

The AFP is very supportive of the Industry Standards including a requirement or even a mandate, for reporting tools or mechanisms to be available to end-users to report CAM or pro-terror material. The AFP is also supportive of the services then taking appropriate and timely action to respond to reports of breaches of terms of use and report to LEA. End-user and public reporting of online CAM or pro-terror material is crucial in enabling the AFP to achieve its core functions. The AFP will continue to work closely with eSafety to publicly promote mechanisms to report CAM.

Throughout the Discussion Paper and in the draft Industry Standards, eSafety refers to services "acquiring and using off-platform information that can help identify and block the registration of potential end-users who have distributed child sexual material and/or pro-terror material in other environments". "Off-platform information" should be further clarified to determine the types of information that should be utilised or relied upon by services noting legislative and privacy implications could also apply to the use of "off-platform information". The AFP suggests that unless services were relying on sufficiently verified information (for example, notifications to the service from law enforcement) that this suggestion be narrowed in scope or further defined.

eSafety is seeking to place requirements on service providers that are best-placed to prevent the use of generative AI features to create and disseminate class 1A and class 1B material. The use of generative AI to create and/or disseminate CAM or pro-terror material is acknowledged by the AFP and international law enforcement but is difficult to quantify. There is general agreement that it is an emerging issue and that the rapid pace of development for this technology poses significant challenges to law enforcement.

It is not possible to say definitively whether eSafety's Industry Standard requirements and risk ratings will prevent of the use of generative AI technology to create and/or disseminate CAM or pro-terror material, but the AFP views the requirements as a positive initiative. The AFP agrees with the proposal to include "detect and remove" and "deter and disrupt" obligations for high impact generative AI, machine learning model platform services and enterprise providers. These

obligations are appropriate and review of the Industry Standards will indicate whether these obligations are effective in eliminating the creation and dissemination of CAM and pro-terror material. There may be scope to include further obligations following future review of the Industry Standards as technology advances for these services.

4) Compliance measures

The AFP acknowledges eSafety's efforts to draft compliance measures that strike "a balance between flexibility and enforceability".

For CAM and pro-terror material the AFP recommends explicitly including provisions within the "Compliance Measures" divisions to require Relevant Electronic Services and Designated Internet Services to immediately report instances of CAM or pro-terror material to LEA. Currently the Industry Standards require reporting as a compliance measure only where there is evidence of a serious or immediate threat to the life or physical safety of a person in Australia. It may be difficult for service providers to determine whether the person is in Australia so technical solutions may need to be used to identify content that has an Australian nexus. The main requirement would be that LEA is notified as soon as practicable.

Alternatively, the Industry Standards could require the service provider to notify LEA and remove the pro-terror material, but also provide LEA a copy of the material so that it can be assessed and used in a prosecution. An update to the Industry Standards, section 15(2)(b) could note:

> "*..believes in good faith that the material affords evidence of a serious offence under Part 5.3 of the Criminal Code Act 1995 or section 80.2C of the Criminal Code Act; "the provider must, as soon as practicable, report the matter and a copy of the material to an enforcement authority, or otherwise as required by law*".

These amendments would allow LEA to investigate and disrupt the commission of a serious part offence before it got to the stage of being an immediate threat to life.

Additionally, consideration should be given to including an obligation in the Industry Standards for content identified as pro-terror to be taken down from public view and preserved by the service provider for a minimum period of time, for example 90 days, to facilitate appropriate assessment and lawful acquisition by enforcement bodies, for instance, through law enforcement preservation requests. This would greatly assist in the investigation of any criminal offences arising from dissemination of the material, as opposed to just terminating the account and removing the content, which deletes any evidence of the offence. Retention of the material by the service provider is not an offence.

Some of the obligations and compliance measures in the Industry Standards are based on monthly active user thresholds. The AFP accepts eSafety not wanting to disproportionally burden smaller service providers, however it is unclear how the 'monthly active user' thresholds were decided and how they will be assessed or checked by eSafety. It would be useful to include further explanatory material on the thresholds, i.e. are services meant to self-report changes in their monthly active user and if so how often. The thresholds lead to determining which obligations apply to which services so should be subject to review noting public interest in a service provider can change drastically in a short timeframe.

The AFP welcomes the requirement for all service providers to have trust and safety personnel to oversee safety within their operations. The personnel should also be responsible for considering

the implications of future technical updates on their services and their ability to continue to comply with the Industry Standards.

5) Conclusion

The AFP welcomes continued engagement with eSafety and other relevant government departments in the drafting of the Industry Standards.