From: International
Sent: Friday, 30 September 2022 11:56 AM
To: s 22 [redacted]
Subject: Action in the United States – ISF Roundup September 2022 [SEC=OFFICIAL]

eSafetyinternational

International, Strategy and Futures Roundup

15 - 28 September 2022

Catch up on the latest global online safety developments in the International, Strategy and Futures fortnightly newsletter. Please note, article selections are not endorsement.

Action in the United States

California passes social media transparency bill

On 13 September 2022, the Governor of California, Gavin Newsom, approved Bill B-587 Social media companies: terms of service. In February 2021, Assemblymember Jesse Gabriel introduced the bill with the aim of 'pulling back the curtain and require tech companies to provide meaningful transparency.' Social media companies that earn more than US$100 million per year will have to publish content moderation policies and report information such as data on violations of terms of service to the Attorney-General. They will also have to publish this information online.

Texas and Florida look to the Supreme Court

A legal battle is unfolding over a Texas law that stops large social media platforms from removing political posts. The law makes it possible for individuals or the Texas Attorney General to sue social media platforms with more than 50 million monthly US users for taking down political viewpoints. A coalition of tech companies is contesting the law, arguing that it undermines their discretion to moderate dangerous or illegal content on their platforms. Despite the Appeals Court ruling to uphold the law, it is yet take effect and analysts expect a Supreme Court appeal from tech groups is likely.

In Florida, the Attorney General asked the Supreme Court to review Senate Bill 7072, which aims to stop social media companies from moderating users' political speech. While an Appeals Court upheld the Texas law, it determined it is unconstitutional for Florida to prevent social media companies from issuing bans to political figures under that state's law. The Washington Post explores the implications of these Appeals Court rulings and a potential Supreme Court review.

Social media companies testify before US Senate Committee on trust and safety

Executives from Meta, TikTok, YouTube and Twitter appeared before the Senate Homeland Security Committee to respond to questions on safety, privacy and moderation. Senators asked questions about the number of employees working on trust and safety, moderation efforts on non-English language content, and data privacy breaches. Following its appearance before the Committee, YouTube announced an updated content moderation policy that aims to curb violent extremist content, including content not affiliated with designated terrorist organisations.

Senators call for a new US tech regulator

Republican Senator Lindsey Graham and Democrat Senator Elizabeth Warren are developing a proposal to regulate the US tech sector, with an enforcement regime akin to that of the European Union. Graham claims existing regulatory bodies have not kept pace with technology.

International bulletin

New software to understand algorithms under Christchurch Call

Prime Minister of New Zealand, Jacinda Ardern, announced a joint initiative with the US, Twitter and Microsoft to develop new software tools to research how algorithms drive individuals towards terrorist and violent content. The announcement coincided with the Christchurch Call Leaders' Summit, which brought together governments, online service providers and civil society to discuss their continued efforts to combat terrorist and violent extremist content online.

Expert Group backs complaints mechanism for online harms in Ireland

Irish Minister for Tourism, Culture, Arts, Gaeltacht, Sport and Media, Catherine Martin, published the Report of the Expert Group convened to examine the practicability of including an individual complaints mechanism in the Online Safety and Media Regulation Bill 2022. The Group recommended the Irish Government introduce an individual complaints mechanism on a phased basis, prioritising those complaints where the online content in question relates to children.

Industry updates

Apple News hack

Fast Company, which provides content for Apple News' subscription service was 'hacked' resulting in offensive two-sentence push notifications being sent to some iPhone users including some who were not subscribers. Traditionally known for its 'walled garden', the hack was remedied by Apple, which disabled Fast Company's channel to prevent further offensive alerts going out to its customers.

Roblox rolls out voice chat and age recommendations

Popular online game platform Roblox, which relies on user-generated content, is rolling out voice chat for users aged 13 and over. Announced at a recent developer conference, the new voice chat allows users who verify their age to have reduced chat filters to communicate more freely with other players. For players under 13, Roblox will maintain its text only chat functionality.

Parler's 'uncancellable cloud'

Parler announced it is restructuring with a new venture called Parlement Technologies. Marketing itself as 'a free speech social media platform' and an alternative to 'mainstream platforms like Facebook and Twitter', the company will now provide cloud infrastructure for businesses 'at risk of being forced off the internet'.

TikTok's approach to content moderation

Forbes reports that TikTok used a 'two-tiered system' and 'more lenient policy enforcement system' to give preferential treatment to influencers and celebrities. Forbes says TikTok used dedicated queues to 'prioritize and protect the posts of people with more than 5 million followers when they break TikTok's content rules'. According to Engadget, TikTok was using this moderation system as recently as last year.

Twitch's design allegedly enabling abusers to track children

Bloomberg reports that Amazon's livestreaming service Twitch has moderation tools that have proved insufficient at preventing young children from livestreaming themselves and being groomed by adults. There are few barriers to prevent children livestreaming and adult viewers can communicate through text anonymously with children. Twitch relies on user reports and automated solutions to identify abuse, but many grooming tactics are designed to escape automated detection.

Instagram’s suicide and self-harm policies

At a UK inquest into the suicide of 14-year-old Molly Russell, Elizabeth Lagone (Meta Head of Health and Well-Being Policy) said Meta’s guidelines on suicide and self-harm were subsequently updated in consultation with experts. Meta changed its guidelines to ban ‘all graphic suicide and self-harm content’ while allowing users to talk about their feelings related to those issues providing those comments are not ‘graphic, promotional, or show methods or materials’.

YouTube

Bloomberg details how YouTube algorithms went from relying on machines alone to more actively involving humans, particularly on YouTube Kids. Mark Bergen’s *Like, Comment, Subscribe: Inside YouTube’s Chaotic Rise to World Domination* suggests many women creators on the YouTube platform are dealing with ‘vicious harassment, bullying, and stalking’ with ‘toxicity on the platform’ escalating, despite harassment policies being updated in 2019.

Emerging issues

Calls to regulate facial technology

The Human Rights Law Centre (HRLC) welcomed a University of Technology Sydney report on facial recognition technology (FRT), which proposes a legal framework for regulating the use of FRT. The HRLC called on the Attorney-General to urgently regulate the technology to prevent human rights harms. The proposed legal framework would require developers and deployers of FRT – including, in some cases, age assurance tools that rely on facial analysis – to undertake impact assessments, provide safeguards and oversight mechanisms and prohibit use in high-risk contexts. Kieran Pender, Senior Lawyer for the HRLC says 'Our current laws were not drafted to address the challenges posed by facial recognition technology' and it 'has the potential to disproportionately impact women and people of colour', as well as marginalised communities.

Social media users evade content moderation using coded 'algospeak'

Social media users are increasingly using 'algospeak' (codewords, emojis, and deliberate typos) to circumvent automated content moderation tools. The machine learning tools employed by platforms can detect overtly violative material, like hate speech, but struggle with euphemisms like 'camping (abortion)' or 'cheese pizza (CSEM)'.

Report shows increase in calls for violence on 'Incel' forum

The Centre for Countering Digital Hate (CCDH) released a report on the 'Incelosphere', a reference to forums that host discussions of the 'involuntary celibate' movement, which promotes hatred and violence against women and other groups. Studying posts on one 'incel forum', CCDH found a 59 percent increase between January 2021 and January 2022 in the use of terms and codewords related to acts of mass violence.

Research

Distinguishing deepfakes from reality

In the *Journal of Online Trust & Safety*, the article Creating, Using, Misusing, and Detecting Deep Fakes forecasts that deepfakes will become indistinguishable for humans, and outlines technologies available to detect deepfakes.

Algorithm audits

The University of Zurich's empirical investigation into the TikTok recommender algorithm found that the active choice of 'following' another user had the largest influence on the algorithm. Researchers were concerned that video view rate had a similar level of influence as 'liking' a video. As videos cannot be 'unwatched', this limits the control a user has over the feed.

Mozilla released research into YouTube's algorithm, finding that user controls like clicking 'dislike' or 'not interested' had little effect on which videos the platform would recommend. Similar to TikTok, this research raises concerns about users' ability to protect themselves from harmful or problematic content.

Report advocates for Safety by Design to protect children in virtual reality

In a report exploring the risks to children emerging on immersive social and gaming platforms, the Bracket Foundation encourages tech companies and government to prioritise ‘Safety by Design’ when designing and regulating virtual reality. The report highlights a range of safety solutions including age assurance and AI-supported moderation. It concludes that comprehensive regulatory frameworks are needed to standardise safety measures.

Australian Government | eSafetyCommissioner esafety.gov.au

Unsubscribe

From: s 22
Sent: Friday, 30 September 2022 2:30 PM
To: s 22; s 22; s 22; s 22
Subject: Online risks material [SEC=UNOFFICIAL]

This is interesting from the International team

Social media users evade content moderation using coded 'algospeak'

Social media users are increasingly using 'algospeak' (codewords, emojis, and deliberate typos) to circumvent automated content moderation tools. The machine learning tools employed by platforms can detect overtly violative material, like hate speech, but struggle with euphemisms like 'camping (abortion)' or 'cheese pizza (CSEM)'.

s 22
Manager – Education and Training

Australian Government | eSafetyCommissioner

☎ s 22

s 22

esafety.gov.au

eSafety acknowledges the Traditional Custodians of country throughout Australia and their continuing connection to land, waters and community. We pay our respects to Aboriginal and Torres Strait Islander cultures, and to Elders past, present and emerging.

Investigation

INV-2022-26816

Created	Investigation Status	Investigator	Outcome
8/11/2022 2:36 AM	Completed	s 22	Class 1 material

Summary

Locator
s 7(2) cheesepizza s 7(2)

Content Service Type
Hosting Service

Presentation Type
Web Page

Complainant Reason
Child sexual abuse / child abuse / Paedophile activity

Type
OSA Section 38

Priority
Critical

Technical Element
No

Investigator
s 22

Investigation Status
Completed

Created
8/11/2022 2:36 AM

Commenced
8/11/2022

Finalised
8/11/2022

Derived
No

Source Type
CYR

Referrer Url

Title: s 7(2) **. Checked 8/11/2022 at 9:17 am.**
s 37(2)(b)

s 22 **8/11/2022 10:59 AM**

Content

Locator Documents

Investigation Documents

Content Traces

Source IP	Primary Record	g Country	Created On
s 7(2)	Yes	United States	8/11/2022 11:01 AM
s 7(2)	No	United States	8/11/2022 2:36 AM
s 7(2)	No	United States	8/11/2022 9:21 AM

1 3 of 3 (0 selected) Page 1

Regulatory Notices

Name	Notice Type	Legislative Reference	D
		No Regulatory Notice records are available in this v	

0 - 0 of 0 (0 selected) Page 1

INHOPE

Reference Number	**Referral Sent**	**Report Status**
3067563	**8/11/2022 12:39 PM**	**Open**

Actions

Report to LEA	**Report to ISP**	**Content Removed**	**Content Unavailable**
Moved	**Not Illegal**	**Not Legally Accessible**	

Status Update

Last status update	**Next status update**
7/02/2023	

Related Emails

Date Sent/Received	From	To	Su
s 37(2)(b)			

1 - 2 of 2 (0 selected) Page 1

Status **Completed**

From: s 22@eSafety.gov.au>
Sent: Thursday, 5 January 2023 11:49 AM
To: s 22; s 22; s 22; s 22; s 22;
s 22; s 22; s 22

🍕= cheese pizza = child porn

Investigation

INV-2023-02104

Created	Investigation Status	Investigator	Outcome
24/01/2023 10:24 AM	Terminated	s 22	

Summary

Locator
s 7(2) cheesepizza s 7(2)

Content Service Type
Hosting Service

Presentation Type
Web Page

Complainant Reason
Child sexual abuse / child abuse / Paedophile activity

Type
OSA Section 38

Priority
Low

Technical Element
No

Investigator
s 22

Investigation Status
Terminated

Created	Commenced	Finalised
24/01/2023 10:24 AM	25/01/2023	25/01/2023

Derived
No

Source Type
CYR

Referrer Url

Title:
Locator updated to directed URL.
s 22 25/01/2023 10:10 AM

Content

Locator Documents

Investigation Documents

Content Traces

Source IP	Primary Record	g Country	Created On
s 7(2)	Yes	United States	25/01/2023 10:10 AM
s 7(2)	No	United States	24/01/2023 10:25 AM

1 - 2 of 2 (0 selected) Page 1

Regulatory Notices

Name	Notice Type	Legislative Reference	D
		No Regulatory Notice records are available in this v	

0 - 0 of 0 (0 selected) Page 1

INHOPE

Reference Number	**Referral Sent**	**Report Status**

Actions

Report to LEA No	**Report to ISP** No	**Content Removed** No	**Content Unavailable** No
Moved No	**Not Illegal** No	**Not Legally Accessible** No	

Status Update

Last status update	**Next status update**

Related Emails

Date Sent/Received	From	To	Su
		No E-mail records are available in this view.	

0 - 0 of 0 (0 selected) Page 1

Status	Terminated

Investigation

INV-2023-02378

Created	Investigation Status	Investigator	Outcome
28/01/2023 4:20 AM	Completed	s 22	Class 1 material

Summary

Locator
s 7(2) cheesepizza s 7(

Title:
s 37(2)(b)

s 22 30/01/2023 2:56 PM

Content Service Type
Hosting Service

Presentation Type
Web Page

Complainant Reason
Child sexual abuse / child abuse / Paedophile activity

Type
OSA Section 38

Priority
Critical

Technical Element
No

Investigator
s 22

Investigation Status
Completed

Created
28/01/2023 4:20 AM

Commenced
30/01/2023

Finalised
30/01/2023

Derived
No

Source Type
CYR

Referrer Url

Content

Locator Documents

Investigation Documents

Content Traces

Source IP	Primary Record	g Country	Created On
s 7(2)	Yes	United States	30/01/2023 2:47 PM
s 7(2)	No	United States	24/01/2023 10:25 AM

1 - 2 of 2 (0 selected) Page 1

Regulatory Notices

Name	Notice Type	Legislative Reference	D
		No Regulatory Notice records are available in this v	

0 - 0 of 0 (0 selected) Page 1

INHOPE

Reference Number 3153039	**Referral Sent** 30/01/2023 3:29 PM	**Report Status** Open

Actions

Report to LEA	**Report to ISP**	**Content Removed**	**Content Unavailable**
Moved	**Not Illegal**	**Not Legally Accessible**	

Status Update

Last status update 1/05/2023	**Next status update**

Related Emails

Date Sent/Received	From	To	Sı
s 37(2)(b)			

1 - 1 of 1 (0 selected) Page 1

Status Completed

From: s 47F
Sent: Sunday, 23 April 2023 2:03 AM
To: eSafety Hotline
Subject: CyberCrime: Instagram - Child Exploitation
Attachments: Screenshot_2023-03-24-09-13-22-40_1c337646f29875672b5a61192b9010f9.jpg;
Screenshot_2023-03-24-10-22-41-96_1c337646f29875672b5a61192b9010f9.jpg;
Screenshot_2023-03-24-10-27-30-26_1c337646f29875672b5a61192b9010f9.jpg;
Screenshot_2023-03-24-10-25-44-50_1c337646f29875672b5a61192b9010f9.jpg;
Screenshot_2023-03-24-10-30-01-94_1c337646f29875672b5a61192b9010f9.jpg;
Screenshot_2023-03-24-10-27-26-22_1c337646f29875672b5a61192b9010f9.jpg;
Screenshot_2023-03-24-10-26-40-06_1c337646f29875672b5a61192b9010f9.jpg;
Screenshot_2023-03-24-10-30-26-26_1c337646f29875672b5a61192b9010f9.jpg;
Screenshot_2023-03-24-10-31-26-10_1c337646f29875672b5a61192b9010f9.jpg;
Screenshot_2023-03-24-10-31-40-74_1c337646f29875672b5a61192b9010f9.jpg;
Screenshot_2023-03-24-10-32-03-37_1c337646f29875672b5a61192b9010f9.jpg;
Screenshot_2023-03-24-10-47-02-22_1c337646f29875672b5a61192b9010f9.jpg;
Screenshot_2023-03-24-10-47-38-63_1c337646f29875672b5a61192b9010f9.jpg;
Screenshot_2023-03-24-10-47-48-94_1c337646f29875672b5a61192b9010f9.jpg

You don't often get email from s 47F. Learn why this is important

Hello, although I am not anonymous I would like to remain anonymous.

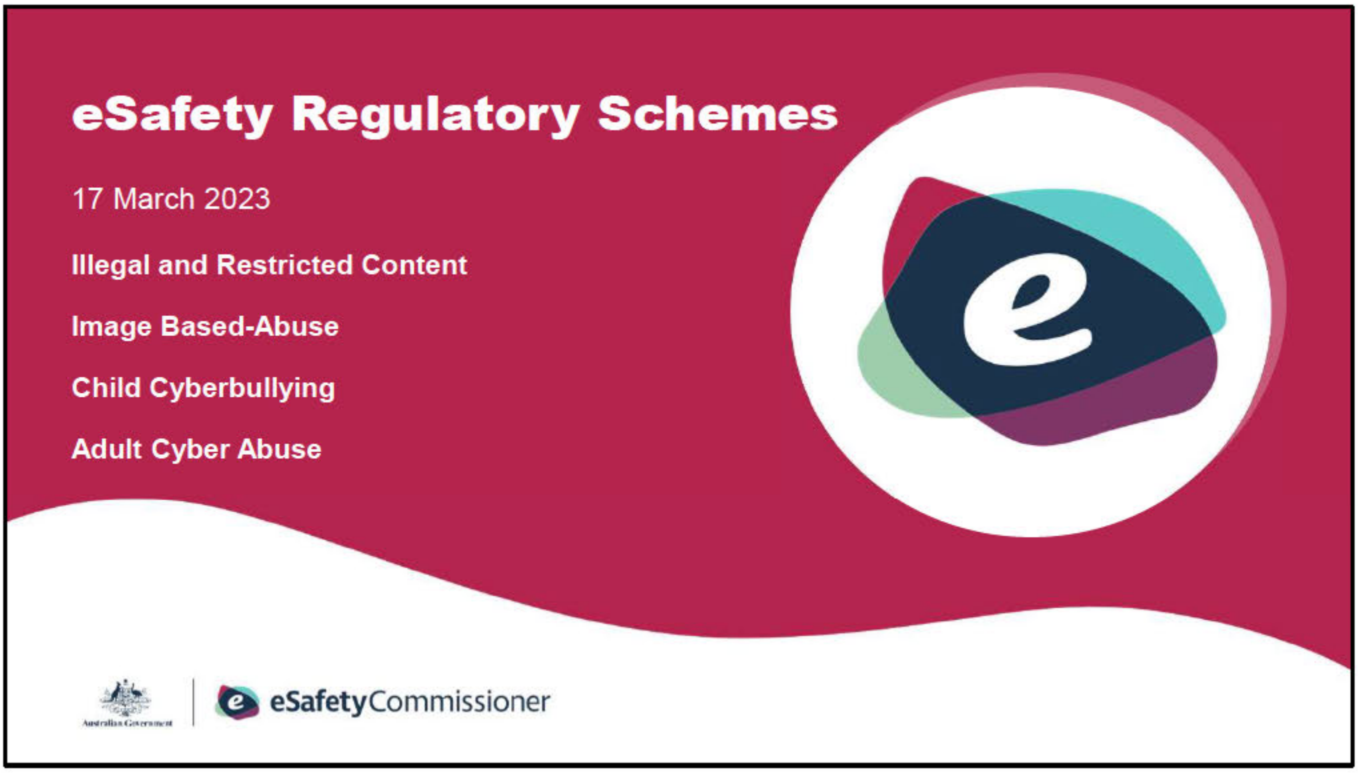
I have discovered that Instagram is being used to market child exploitation materials, child abuse, bestiality, rape and even murder. Many of the participating accounts marketing "s 7(2)" and using language such as s 7(2), Cheese Pizza, Pizza, s 7(2), s 7(2), s 7(2) etc...

I have made diligent efforts to contact Instagram, s 7(2) and even Telegram to report such activity in effort to have these accounts and content removed with minute luck.

This activity is widespread throughout each of these platforms and not only have I discovered the sale of such content but also groups and communities even as well as people offering children and slaves for sale, on Instagram I was offered live shows of children being abused and from Instagram I was coached to telegram where I was offered children for purchase and even an adult slave, which I reported to each of the respective platforms with little acknowledgement, that being said I do think I was the person solely responsible for delaying "Instagram for kids" which I feel is a bad idea.

I will leave things here yet I have attached some screenshots for your convenience, unfortunately I don't have all the screenshots I have submitted yet I will affirm children and people are being trafficked which is marketed on Instagram and moved to other platforms.

regards,



eSafety Regulatory Schemes

17 March 2023

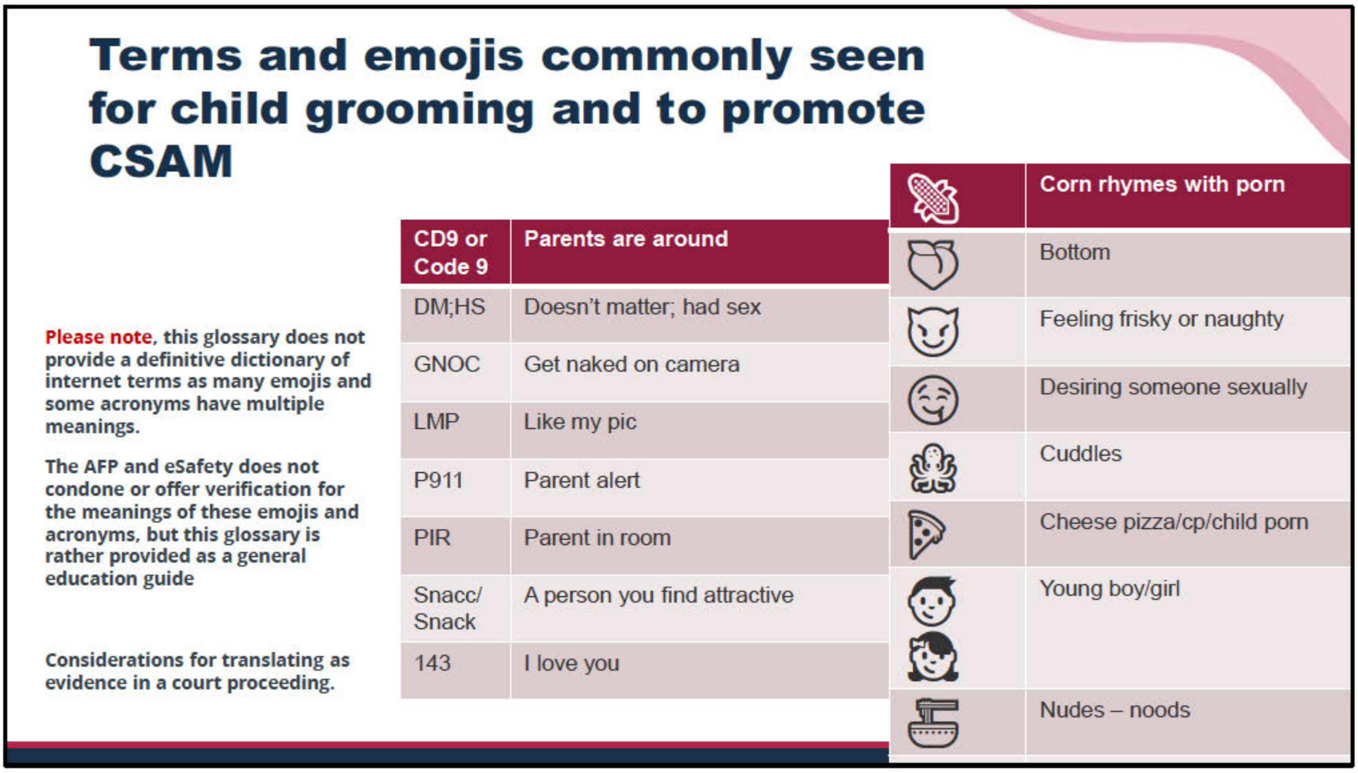
Illegal and Restricted Content

Image Based-Abuse

Child Cyberbullying

Adult Cyber Abuse

Australian Government | eSafetyCommissioner



Terms and emojis commonly seen for child grooming and to promote CSAM

Please note, this glossary does not provide a definitive dictionary of internet terms as many emojis and some acronyms have multiple meanings.

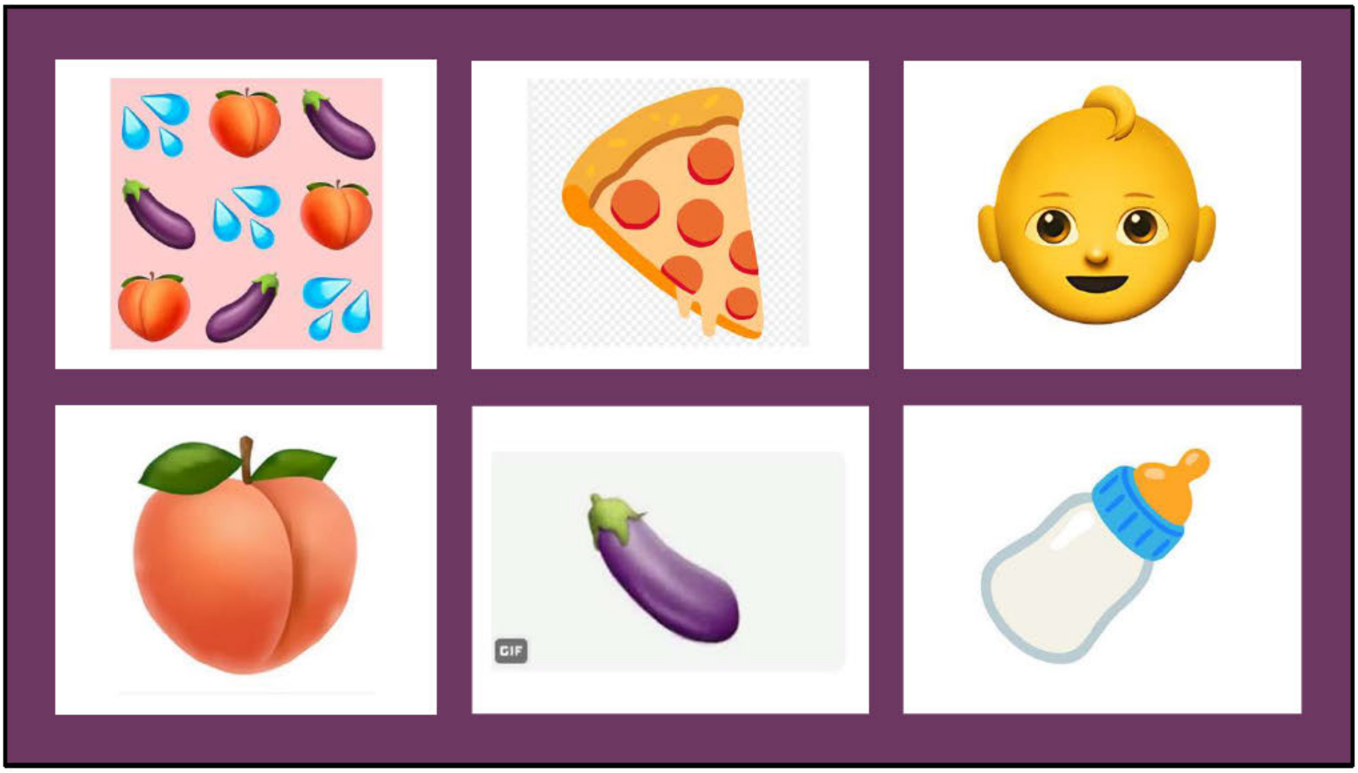
The AFP and eSafety does not condone or offer verification for the meanings of these emojis and acronyms, but this glossary is rather provided as a general education guide

Considerations for translating as evidence in a court proceeding.

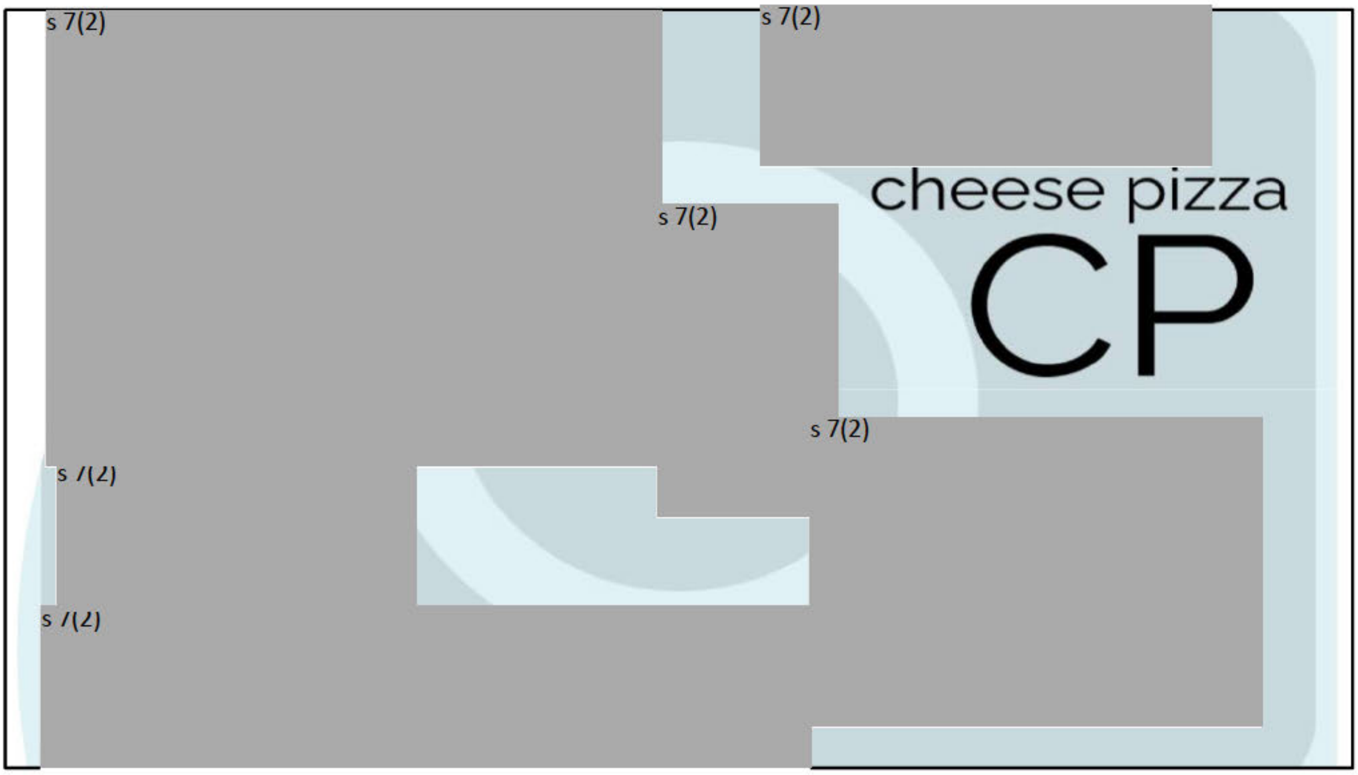
CD9 or Code 9	Parents are around
DM;HS	Doesn't matter; had sex
GNOC	Get naked on camera
LMP	Like my pic
P911	Parent alert
PIR	Parent in room
Snacc/ Snack	A person you find attractive
143	I love you

🌽	Corn rhymes with porn
🍑	Bottom
😈	Feeling frisky or naughty
🤤	Desiring someone sexually
🐙	Cuddles
🍕	Cheese pizza/cp/child porn
👦 👧	Young boy/girl
🍜	Nudes – noods

Due to its phallic use, the hashtag was once banned (2015) on Instagram's search function



A few more emojis which are indicative of how the totally innocent images can be misused and reinterpreted for communication and or the potential content one may expect to find.



s 7(2)

s 7(2)

s 7(2)

s 7(2)

s 7(2)

s 7(2)

s 7(2)

CP = Child Porn
Cheese Pizza abbreviates into CP

s 7(2)

s 7(2)

We typically observe CSAM on the clear or free web, without any pay wall or viewing restriction. Whilst the dark web provides a greater level of anonymity and protection that help persons evade law enforcement, it is

How would you present this as evidence – interpreted in the courts.

From: s 22 @eSafety.gov.au>
Sent: Thursday, 24 August 2023 2:50 PM
To: s 22

> s 22
>
> thanks s 22 just to clarify, detecting signals/patterns (and metadata?) are measures that "encrypted" services can take on their unencrypted surfaces?

to my knowledge yes - the main example is WhatsApp (and apparently fb messenger in secret chat) - profile pics are not encrypted, so if someone's profile pic is of cheese pizza that could be a signal that they share CSAM from that account

Tech giants including Apple, Google and Meta will be forced to do more to tackle online child sexual abuse material and pro-terror content, including "deepfake" child pornography created using generative AI, in world-first industry standards laid out by Australia's eSafety Commissioner.

Following more than two years of work, and after rejecting draft codes created by the tech industry, eSafety Commissioner Julie Inman Grant will release draft standards on Monday covering cloud-based storage services like Apple iCloud, Google Drive and Microsoft OneDrive, as well as messaging services like WhatsApp, requiring them to do more to rid their services of unlawful content.

Inman Grant, a former Twitter executive, said that she hopes Australia's industry standards would be the "first domino" of similar regulations globally to help tackle harmful content.

She said the requirements would not force tech companies to break their own end-to-end encryption, which is turned on by default on some services, including WhatsApp.

All major tech platforms have policies that ban child sex abuse material from their public services, but Inman Grant said they have not done enough to police their own platforms.

"We understand issues around technical feasibility, and we're not asking them to do anything that is technically infeasible."

"But we're also saying that you're not absolved of the moral and legal responsibility to just turn off the lights or shut the door and pretend this horrific abuse isn't happening on your platforms.

"What we've found working with WhatsApp, it's an end-to-end encrypted service, but they pick up on a range of behavioural signals that they've developed over time, and they can scan non-encrypted parts of the services, including profile anpd group chat names, and things like cheese pizza emojis, which is known to stand for child pornography."

"These and other interventions enable WhatsApp to make 1.3 million reports of child sexual exploitation and abuse each year," she added.

The standards will also cover child sexual abuse material and terrorist propaganda created using open-source software and generative AI. A growing number of Australian students for example are creating so-called "deepfake porn" of their classmates and sharing it in classrooms.

“We’re seeing synthetic child sexual abuse material being reported through our hotlines, and that’s particularly concerning to our colleagues in law enforcement, because they spend a lot of time doing victim identification so that they can actually save children who are being abused,” she said.

“I think the regulatory scrutiny has to be at the design phase. If we’re not building in and testing the efficacy and robustness of these guardrails at the design phase, once they’re out in the wild, and they’re replicating, then we’re just playing probably an endless and somewhat hopeless game of whack-a-mole.”

Inman Grant’s office has commenced public consultation on the draft standards, a process that will run for 31 days. She said the final versions of the standards will be tabled in federal parliament and come into effect six months after they’re registered.

“The standards also require these companies to have sufficient trust and safety, resourcing and personnel. You can’t do content moderation if you’re not investing in those personnel, policies, processes and technologies,” she said.

“And you can’t have your cake and eat it too. And what I mean by that is, if you’re not scanning for child sexual abuse, but then you provide no way for the public to report to you when they come across it on your services, then you are effectively turning a blind eye to live crime scenes happening on your platform.”

The introduction of the standards comes after social media giant X – formerly known as Twitter – refused to pay a $610,500 fine from the eSafety Commissioner for allegedly failing to adequately tackle child exploitation material on its platform.

X has filed an application for a judicial review in the Federal Court.

“eSafety continues to consider its options in relation to X Corp’s non-compliance with the reporting notice but cannot comment on legal proceedings,” a spokesman for the commissioner said.