eSafety Submission

Senate Standing Committee on Legal and Constitutional Affairs Legislation Committee Criminal Code Amendment (Deepfake Sexual Material Bill) 2024

23 July 2024

Introduction

The eSafety Commissioner (eSafety) welcomes the opportunity to make a submission to the Senate Standing Committee on Legal and Constitutional Affairs Legislation Committee on the *Criminal Code Amendment (Deepfake Sexual Material) Bill 2024I* (the Bill).

In summary, eSafety broadly supports the Bill. In this submission, we highlight aspects of the Bill we support and areas we suggest be given further consideration.

We also outline the importance of eSafety's whole of community and multidimensional regulatory remit complementing a criminal justice response to image-based abuse. Relevantly, we operate a complaints scheme for the non-consensual sharing of intimate images that applies to both real and synthetic material, including AI and deepfakes.

We go into further detail below about how our complaints scheme for image-based abuse is complemented by a range other powers and functions that allows us to address image-based abuse at the individual, community, industry, structural and systemic level.

It is by interlocking the approaches of eSafety and the criminal justice system that we can mutually reinforce each other's schemes. This will allow image-based abuse to be addressed more holistically and comprehensively. This will ultimately lead to better safety outcomes and response options for the community, especially victim-survivors of image-based abuse.

Overview of eSafety

As you know, eSafety is Australia's independent regulator, educator and coordinator for online safety. We aim to safeguard Australians from online harms and to promote safer, more positive online experiences.

The *Online Safety Act 2021* (Online Safety Act) sets out our legislative functions. Our regulatory approach is comprised of the three pillars of prevention, protection, and proactive and systemic change.

We outline these pillars below.

- **Prevention:** While eSafety acts as an important safety net for Australians online, our primary goal is to prevent online harms from happening in the first place. This work falls under our prevention pillar. Through research, education and training programs, we aim to build the capacity of Australians to interact safely online. We seek to provide Australians with the practical skills and confidence to be safe, resilient and positive users of the online world, and to know where to seek help if issues arise.
- **Protection:** Where online harm does occur, eSafety offers tangible, rapid assistance. This work falls under our protection pillar. Our individual complaints mechanisms allow us to investigate and take action to remove certain types of content relating to four types of harm: cyberbullying of children, cyber abuse of adults, the non-consensual sharing of intimate images, and illegal or restricted online content. All our schemes apply to both real and synthetic material, including AI and deepfakes.
- **Proactive and systemic change:** With the rapid evolution of technology, eSafety knows we need to be at the forefront of anticipating, mitigating and responding to online harms. This work falls under our proactive and systemic change pillar. This includes our powers to regulate digital platforms' broader systems and processes, including through the Basic Online Safety Expectations (BOSE) and industry codes and standards. It also includes our Safety by Design initiative, as well as our work anticipating and responding to emerging tech trends, opportunities and challenges.

These pillars reflect our broad and holistic legislative remit. The way the pillars work together reflects how eSafety's various functions work together to create a multidimensional regulatory toolkit.

We take a risk and harms-based approach to our work. This approach is underpinned by our core mission of safeguarding Australians at risk of online harm.

This complements the role other agencies play in investigating and prosecuting crimes perpetrated online.

We recognise that combating online harm is a global challenge. We therefore work as part of a cross-sector and multi-jurisdictional online safety ecosystem.

Understanding image-based abuse

To begin, it is first helpful to understand that image-based abuse is a form of sexual violence. Within research, policy and practice, it has three key forms:[1]

1. The creation of intimate images without consent.
2. The sharing or distribution of intimate images without consent.
3. Threats to share or distribute intimate images without consent.

Image-based abuse can also include the non-consensual creation, sharing or threatened sharing of intimate images where:[2]

- The image is natural or digitally altered imagery (e.g. deepfakes).
- The victim has taken the intimate image, such as a 'selfie', and/or it has been produced within the context of an intimate relationship.
- The image has been created coercively (e.g. during a sexual assault) or covertly (e.g. 'upskirting' or surreptitious filming) or has been obtained through hacking or by being stolen.
- The image-based abuse is part of a broader blackmail attempt. This is known as sexual extortion and is outlined further below.

Understanding deepfake technology

Creating a deepfake typically involves using a combination of machine learning and face-tracking algorithms. This allows a fake face to be seamlessly stitched onto a real one. This gives anyone, including those with malicious intent, the capability to control what an image appears to say and do.

While this technology is not new, the ability to create a realistic impersonation of a person in video is a relatively new frontier.

[1] Henry, N, McGlynn, C Flynn, A Johnson, K, Powell, A and Scott. AJ 2020, Image-based sexual abuse: A study on the causes and consequences of non-consensual nude or sexual imagery, London and New York, NY: Routledge; Powell A and Henry N, 2017 Sexual Violence in a Digital Age, London: Palgrave Macmillan.; Powell A, Henry, N and. Flynn, A, 2018 mage-based sexual abuse, in *Routledge handbook of critical criminology*, 2nd ed., New York, Routledge, pp. 305-315. Powell, A Henry, N Flynn, A and Scott AJ 2019, Image-based sexual abuse: The extent, nature, and predictors of perpetration in a community sample of Australian residents, *Computers in Human Behavior,* vol. 92, pp. 393-402.; Henry. N and Flynn A 2019 Image-based sexual abuse: Online distribution channels and illicit communities of support, *Violence Against Women,* vol. 25, no. 6, pp. 1932-1955,

[2] Flynn, A Powell, A. Scott AJ and Cama, E. 2021 Deepfakes and digitally altered imagery abuse: A cross-country exploration of an emerging form of image-based sexual abuse. ., *The British Journal of Criminology,* vol. XX, pp. 1-18,; Henry, N Flynn, A and Powell, A, 2019. Image-based sexual abuse: Victims and perpetrators, Australian Institute of Criminology, Canberra: ACT.; Flynn A and Henry, N 2021. Image-based sexual abuse: An Australian reflection, *Women & Criminal Justice,* vol. 31, no. 4, pp. 313-326

Thousands of open-source AI apps have proliferated online and are often free and easy to use by anyone with a smartphone. These apps make it simple and cost-free for a perpetrator to use, while incurring a significant, lingering and incalculable cost on the victim-survivor.

We are concerned that these apps are using sophisticated monetisation tactics and are increasingly being found on mainstream social media platforms. This boosts their visibility and availability, particularly to younger audiences.

We are also concerned about multi-modal forms of generative AI. These can create even more hyper-realistic deepfake imagery, including deepfake image-based abuse.

We believe that one of the biggest problems is that deepfake detection tools are significantly falling behind the rapid proliferation of powerful AI models for imagery, audio and video. These tools are also applied at the back-end of this abuse, which means harm has already occurred.

We unfortunately cannot just deploy benevolent AI to tackle malevolent AI. What we need is effective detection tools. This will require multiple technical interventions, including manual forensics analysis to effectively flag media manipulation that may be so photo-realistic that it isn't discernible to the naked eye.

AI is a new and powerful force in what is an already interconnected world. It has the capacity for good and can also have immense social benefits.

But without greater regulation of AI and greater transparency about how these technologies are being applied, we cannot achieve the levels of safety and accountability we need to prevent harm, including in the context of deepfake image-based abuse.

We explore the harms of this type of abuse, and possible interventions and initiatives, further below.

Research

Before outlining our feedback on the Bill and eSafety's regulatory remit, we want to highlight the impacts image-based abuse has on victim survivors. This not only underscores the seriousness of the issue, but also ensures lived experience can inform our work in this space.

We note the impacts of image-based abuse are likely to only increase with the proliferation of new forms of this abuse. This includes the significant increase in the creation of AI deepfakes, with one study finding a 550% increase in deepfake videos between 2019 and 2023.[3]

A 2019 study found that deepfake image-based abuse (what the report termed non-consensual deepfake pornography) accounts for 96% of the total deepfake videos online.[4] Of the top 5 deepfake pornography websites examined, 100% of the videos were of women.

A global systematic review and meta-analysis (*n* = 19 papers) found that:

- 8.8% of participants across 20 independent samples had their intimate images shared without consent.
- 7.2% of participants had been threatened with distribution of intimate images.
- 17.6% of participants had experienced non-consensual creation of sexually explicit material.[5]

[3] Home Security Heroes, (2023). *2023 State of Deepfakes: Realities, Threats and Impact*. 2023 State Of Deepfakes: Realities, Threats, And Impact (homesecurityheroes.com)

[4] Patrini, G. (2019). *Mapping the Deepfake Landscape*. Deepfake Labs. deepfake_report.pdf (regmedia.co.uk)

[5] Patel, U. & Roesch, R. (2022). The prevalence of technology-facilitated sexual violence: A meta-analysis and systematic review, *Trauma, Violence, & Abuse, 23*(2), 428-443.

In a 2019 survey involving participants aged 16 to 64 years in Australia, New Zealand and the United Kingdom (*n* = 6,109), just over one in three (37.7%) respondents had experienced at least one form of image-based abuse since the age of 16. The findings were comparable across the three countries:

- 35.2% in Australia.
- 39.0% in New Zealand.
- 39.0% in the United Kingdom.[6]

Of respondents surveyed across Australia, New Zealand and the United Kingdom:

1. 14% reported that someone had shared or created digitally altered nude or sexual images of them without their context.
- 34% those whose images had been taken without their consent had their images digitally altered.[7]

A recent large-scale survey of 16,693 adults from 10 countries found that:

- 14.5% of respondents had experienced threatened sharing of intimate images since the age of 18.
- 15.9% of Australian respondents had experienced threatened sharing since the age of 18.[8]

The same survey also had findings specifically relating to deepfakes:

- 2.2% of respondents (3.7% of Australian respondents) had experienced 'deepfake pornography' video victimisation since the age of 18.
- 1.2% (2.6% Australian) had experienced someone creating a 'deepfake pornography' content of them.
- 1.3% (1.9% Australian) had experienced someone posting or sending 'deepfake pornography' content of them.
- 1.2% (2.3% Australian) had experienced a threat to share 'deepfake pornography' content of them.[9]

The harms caused by image-based abuse have been consistently reported. This includes negative impacts on mental health and career prospects, as well as social withdrawal and interpersonal difficulties.[10] Victim-survivors have also described how their experiences of image-based abuse victimisation radically disrupted their lives, altering their sense of self, identity and their relationships with their bodies and with others.[11]

A survey of 4,274 Australians aged 16–49 years found that victims of image-based abuse were almost twice as likely as non-victims to report high levels of psychological distress.[12] More specifically, moderate to severe depression or anxiety was reported in:

- 80% of those who had experienced threats to share an intimate image without consent.
- 75% of those who had experienced intimate images shared without their consent.

---

[6] Powell, A., Scott, A.J., Flynn, A., & McCook, S. (2024) A multi-country study of image-based sexual abuse: Extent, relational nature and correlates of victimisation experiences, *Journal of Sexual Aggression, 30*(1), 25-4.

[7] Henry, N., McGlynn, C., Flynn, A., Johnson, K., Powell, A., & Scott., A.J. (2020). *Image-based sexual abuse: A study on the causes and consequences of non-consensual nude or sexual imagery.* Routledge, New York, NY

[8] Henry, N., & Umbach, R. (2024). Sextortion: Prevalence and correlates in 10 countries. *Computers in Human Behavior*, 158, 108298.

[9] Umbach, R., Henry, N., Beard, G. F., & Berryessa, C. M. (2024). Non-Consensual Synthetic Intimate Imagery: Prevalence, Attitudes, and Knowledge in 10 Countries. In Proceedings of the CHI Conference on Human Factors in Computing Systems (1-20).

[10] Henry, N., McGlynn, C., Flynn, A., Johnson, K., Powell, A., & Scott., A.J. (2020). *Image-based sexual abuse: A study on the causes and consequences of non-consensual nude or sexual imagery.* Routledge, New York, NY; Citron, D.K., & Franks, M.A. (2014). Criminalizing revenge porn. *Wake Forest Law Review 49*, 345–391; Paradiso, M.N., Rollè, L., & Trombetta, T. (2024). Image-based sexual abuse associated factors: a systematic review. *Journal of Family Violence*, 39, 931-954.

[11] McGlynn, C., Rackley, E., Johnson, K., et al. (2021). 'It's torture for the soul': The harms of image-based sexual abuse. *Social & Legal Studies 30*(4), 541–562.

[12] Henry, N., Flynn, A., & Powell, A. (2019). Responding to revenge pornography: prevalence, nature and impacts, report to the Criminology Research Advisory Council, Australian Institute of Criminology. CRG_08_15-16-FinalReport.pdf (aic.gov.au)

- 67% of those who had experienced intimate images taken without their consent.

In addition, many victims of image-based abuse reported feeling highly fearful for their safety. This included:

- 46% of those who had experienced threats to share an intimate image without consent.
- 39% of those who had experienced intimate images shared without consent.
- 28% of those who had experienced intimate images taken without their consent.

Further, image-based abuse can have significant social impacts on victim-survivors. This includes reduced contact with others, reputational concerns and social isolation.[13] For example, in the survey of respondents of image-based abuse across Australia, New Zealand and the UK:

- 78.8% reported reputational concerns.
- 55.7% reported impacts on their relationships with others.[14]

Feedback on the Bill

As noted above, eSafety broadly supports the Bill. Our feedback falls into two categories:

1. The interaction between the complaints scheme administered by eSafety under the Online Safety Act to address the non-consensual sharing of intimate images, also known as image-based abuse, and the proposed offence under the Bill.
2. The aspects of the Bill that provide welcome clarity to the operation and coverage of the existing 474.17A offence under the Criminal Code.

Interaction with the Online Safety Act

We note there are some differences in coverage and application between the image-based abuse complaints schemes under the Online Safety Act and the Bill. It is important to understand these differences, in order to understand the different recourse and redress options for victim-survivors of image-based abuse.

We have outlined above that within research, policy and practice image-based abuse applies to the non-consensual creation, sharing or threatened sharing of intimate images.

It is then helpful to understand how the Online Safety Act addresses image-based abuse under our complaints scheme. The Online Safety Act:

- Does not apply to the creation of intimate images without consent.
- Applies to the posting of intimate images without consent, which is a subset of sharing or distribution.
- Applies to the threatened posting of intimate images without consent.
- Applies in circumstances where the posting or threatened posting is by an end-user of a social media service, relevant electronic service or designated internet service.[15]
- Applies to both natural and synthetic material, including AI and deepfakes.

---

[13] Paradiso, M.N., Rollè, L., & Trombetta, T. (2024). Image-based sexual abuse associated factors: a systematic review. *Journal of Family Violence*, 39, 931-954; McGlynn, C., Rackley, E., Johnson, K., et al. (2021). 'It's torture for the soul': The harms of image-based sexual abuse. *Social & Legal Studies 30*(4), 541–562.

[14] Powell, A., Scott, A.J., Flynn, A. & Henry, N. (2020). *Image-based sexual abuse: an international study of victims and perpetrators – a summary report*, Royal Melbourne Institute of Technology, Melbourne. Image-based sexual abuse: An international study of victims and perpetrators – A Summary Report. — Monash University

[15] These services are defined at sections 13, 13A and 14 of the Online Safety Act 2021 (Cth). They are also outlined further below.

An intimate image is defined at section 15 of the Online Safety Act. In summary, it includes the depiction of private parts, the depiction of private activity and the depiction of a person without attire of religious or cultural significance. It includes still and moving visual images (the material) and it is immaterial whether the material has been altered. However, the circumstances must be those in which an ordinary reasonable person would reasonably expect to be afforded privacy. It also requires the end-user or depicted person to be ordinarily resident in Australia.

As with all our investigations schemes, the core objective of the scheme is to alleviate harm. We do this by removing harmful material and restraining the ability of individuals to perpetrate harm, where the relevant statutory threshold requirements are met.

It is then helpful to understand how the Bill proposes to address image-based abuse. The Bill:

- Applies to the transmitting of intimate images without consent.
- Applies to the creation of intimate images without consent, but only where the images are also transmitted, whereby it is an aggravated offence.
- Applies where the transmission or creation combined with transmission occurs by a carriage service.
- Applies to both natural and synthetic material, including AI and deepfakes.

The key distinctions between the Online Safety Act and Bill are therefore:

- The Online Safety Act does not apply to the creation of intimate images without consent.
- The Bill does not apply to threats to transmit intimate images without consent.

This means there will be a distinction between the Bill and Online Safety Act in coverage. For example, the Bill will not apply to sexual extortion, unless the sexual extortion leads to an intimate image being shared.

Sexual extortion is a subset of image-based abuse. It is a form of blackmail where someone threatens to share a nude or sexual image or video of someone unless that person gives in to their demands.

Sexual extortion is a serious issue. It remains the most-reported form of image-based abuse to eSafety. In the 2023-2024 financial year, as of 31 March 2024, sexual extortion reports where material was posted online comprised about 3% of the total number of reports, whereas reports about threatened sharing comprised about 55% of reports.

Approximately 75% of reports received from adults involve sexual extortion. Unlike other forms of image-based abuse, sexual extortion is predominantly experienced by boys and young men. The high proportion of sexual extortion reports we receive means that we receive significantly more reports from boys and men (68%) than girls and women (30%) under this scheme.

It may be useful to consider whether other criminal measures at the federal, state or territory level are needed to address sexual extortion more effectively.

We also note the Bill contains an aggravated offence that is linked to the Online Safety Act. This is where the offender has already been subject to 3 or more civil penalties orders relevant to the Online Safety Act. We support this provision. We note, however, that given eSafety is generally able to take action informally, this is likely to only be enlivened in rare and serious circumstances.

Clarity on the existing 474.17A offence

We welcome that the Bill clarifies the operation and coverage of the existing 474.17A offence under the Criminal Code in several important ways.

As noted above, we welcome that the new offence will apply regardless of whether the material is unaltered or has been created or altered in any way using technology. This aligns with all eSafety's investigative schemes, which apply to natural and synthetic versions of online material.

We also welcome that the new offence removes the requirement that a lack of consent be regarded as offensive by a reasonable person. Consistent with the approach taken by the Online Safety Act, this recognises that image-based abuse is inherently harmful and that this does not need to be demonstrated through an offensive criterion.

Similarly, the new offence in the Bill, unlike the Online Safety Act, sensibly does not rely on the depicted person having a 'reasonable expectation of privacy'.

The background information in the Bills Digest (No.81, Bills Digests Alphabetical Index 2023-24) references our Image-Based Abuse Regulatory Guidance to note that the new Criminal Code definitions should similarly cover material that 'appears to depict' certain persons and that it is 'immaterial whether material has been altered'. It notes that this definition is likely to extend to sexualised deepfake images.

We welcome that the new offence will clarify the position. However, we note that paragraph 64 of the Explanatory Memorandum to the Bill states intimate images need to 'reasonably or closely resemble an individual'.

In our experience, close-up images of naked torsos or genitalia are often shared in a context where they are made to seem like a person they are not. For example, the intimate image may be posted with a person's name or a non-intimate image of a person's face. This leads people to believe it is the named or pictured person. For example, our Image-Based Abuse Regulatory Guidance takes the following approach:

> 'Intimate images can include photos and videos that have been digitally altered (for example, photoshopped images or deepfakes). They also include images or videos which have been shared in a way that will make people think they show a specific person (for example, a nude photo tagged with a person's name even though it is not of them).'

The Bill will strengthen the protections for image-based abuse and lead to an enhanced criminal justice response, which will work in conjunction with eSafety's image-based abuse scheme under the Online Safety Act. This will provide victim-survivors with more choice. As outlined below, we must then ensure that regardless of the choice victim-survivors make, they are treated with a trauma-informed approach.

The Bill's implementation would benefit from a broad community education and awareness raising campaign to reduce stigma and promote help-seeking among victim-survivors, while highlighting the harms of image-based abuse and that perpetration is never okay. Specific training and upskilling on trauma-informed approaches for law enforcement and the judicial system should be considered, especially in respect of interactions with victim-survivors, as we know image-based abuse can be incredibly traumatising for victim-survivors.

Statutory Review of the Online Safety Act

As the Committee may know, an independent statutory review of the Online Safety Act is currently underway. The Terms of Reference for the review are broad ranging and include consideration of eSafety's existing statutory schemes, including those outlined above. The Terms of Reference also considers 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 technologies, such as generative AI and recommender systems.

Public consultation commenced with the release of an Issues Paper on 29 April 2024. Submissions closed on 21 June 2024. eSafety made a public submission touching on our goals for the review. These align with the goals and broader contextual points we are raising in this submission.

The Final Report of the review is expected to be provided to the Minister for Communications by 31 October 2024.

eSafety's remit

eSafety has a range of regulatory levers that will work in conjunction with the Bill to provide better support to victim-survivors of image-based abuse – and just as importantly, drive systemic change to this issue.

We want to outline the history of eSafety's work addressing image-based abuse, our existing initiatives and interventions, and where we think reform would allow us to more effectively and systemically address the issue.

When the eSafety Commissioner began in January 2017, one of her first priorities as Commissioner was to change the lexicon around image-based abuse. The Commissioner wanted to shift from the term 'revenge porn', an inherently victim blaming term, to 'image-based abuse', which reinforces the nature of the act: abuse.

In January 2020, we released our Deepfake trends and challenges position statement. At the time, it was a forward leaning and anticipatory analysis of an emerging issue. In the position statement, we provided an overview of:

- The concept of deepfakes
- How deepfakes are created
- The risks and harms of deepfakes
- How to spot a deepfake
- Advice for dealing with deepfakes
- eSafety's approach to deepfakes
- How eSafety can help with deepfakes

The fact that in just over four and half years, this is now a commonplace form of abuse speaks to the complex, evolving and dynamic nature of the online safety ecosystem. It also speaks to how challenging, but necessary, it is to regulate.

More broadly, eSafety undertakes a range of education and awareness raising programs relating to image-based abuse. We are firmly of the view that education and capacity building is a lifelong journey that should begin as early as possible. This should be based on the 'Four Rs of Online Safety' — respect, resilience, responsibility and reasoning.

But we also believe that educators and parents are not armed with all of the information they need – and may not be prepared for the onslaught of AI-generated harms that may continue to rapidly proliferate. As such, we are also stepping up our efforts to raise awareness among the community about AI-related harms specifically. We have a number of mechanisms to do this:

- Our engagement with children, young people, parents and carers, both directly and through representative groups.

- Our key consultation mechanisms and stakeholder groups we oversee. This includes the Trusted eSafety Providers, the eSafety Youth Council and the National Online Safety Education Council.
- Our training with educators, student wellbeing officers, pre-service teachers, university support staff, parents and carers, sporting organisations, corporates and youth serving professionals. For example, we are already delivering webinars for teachers around AI in the classroom.

Image-Based Abuse Scheme

From beginning a national image-based abuse portal in October 2017, eSafety now administers a formal scheme relating to image-based abuse, which we have outlined above.

In the 2023-24 financial year, eSafety received 7,267 reports of image-based abuse.

The scheme includes formal powers, including removal notices and civil penalty powers. However, we have only needed to issue a small number of removal notices, as our informal action has been successful in most instances (over 80%), which makes giving removal notices unnecessary.

In addition, over half of our reports concern threatened sharing, which means fewer removal notices. In response to threatened sharing, we typically alert platforms to the account user's behaviour. This may result in account deletion.

While we have a high success rate with our scheme, we can also speak to our unsuccessful removal requests. This generally occurs when intimate images are posted to pornography sites, particularly hack, leak or expose sites. The challenge is that many of these sites do not have contact points and use services to hide their true hosting information. We also have challenges with automated scraping and sites operated by "bots", especially when these sites appear to have no human moderation. The remainder of our unsuccessful removal requests are comprised of uncooperative message boards, archive sites, file sharing, hosting providers, image or video sharing sites, as well as many sites being operated and hosted overseas.

Where eSafety is unable to effect removal of intimate images, we take steps to limit their discoverability. While we don't have specific regulatory powers to issue link deletion powers, we have good success with informal requests to search engine providers to deindex the images from search engine results. This makes it harder for people to find. We also share instances of unsuccessful removal requests with other areas of eSafety to leverage industry change.

We also refer eligible complainants to online prevention tools, mainly StopNCII.org (for adults) and Take It Down (for under 18s).

We also help complainants understand their options under criminal offences at the federal, state and territory level. This helps complainants make an informed decision about the most appropriate avenue for them in their circumstances.

When material or activity appears to meet a criminal threshold or there is evidence that a person is under a real threat of physical harm, we work directly with law enforcement. This includes referring specific matters for criminal investigation. We have memorandums of understanding with policing organisations across the country to facilitate our engagement. This helps achieve a degree of coordinated, cross-agency and multi-jurisdictional effort. This is especially true in cases where under 18s experience image-based abuse, which we refer for assessment and triage to the Australian Centre to Counter Child Exploitation.

We will take enforcement action against perpetrators in appropriate circumstances. For example, we currently have one matter before the Federal Court being pursued under our civil penalty scheme.

Lastly, though very important, we will help connect complainants to additional support, such as counselling and legal services, when needed.

More broadly, we use the insights and intelligence we gain from our reporting schemes to inform and shape our other work programs across eSafety. This includes our prevention work with the community and, as outlined below, our systemic work with industry.

Systemic powers

We have a range of measures to address online harms systemically. We outline some of the key ones below. We draw particular attention to how these systemic powers relate to artificial intelligence.

Codes and standards

Under the Online Safety Act, the Commissioner can request industry bodies to develop codes to regulate illegal and restricted online material across eight industry sections. These are outlined at section 135 of the Online Safety Act. They are summarised below:

- **App distribution services** – services distributing apps that can be downloaded and accessed by end-users in Australia. For example, app stores/marketplaces.[16]
- **Designated internet services** – services allowing end-users in Australia to access material using an internet carriage service, where the service is not a Social Media Service or Relevant Electronic Service. For example, file storage services managed by end-users in Australia, websites and apps.[17]
- **Equipment services** – manufacturers, suppliers, maintainers and installers of equipment that is used to access online services,[18] such as mobile phones; laptops; tablets; internet-enabled devices (such as smart TVs and gaming consoles); immersive technologies (such as virtual reality headsets); wi-fi routers. This section of the online industry includes manufacturers of these devices, operating system providers, as well as businesses and retail outlets that install, sell and/or repair/maintain such devices.
- **Hosting services** – services which host stored material in Australia (for example, a service with data centres located in Australia).[19]
- **Internet carriage services** – a listed carriage service that enables end-users to access the internet.[20]
- **Internet search engine services** – electronic services designed to collect, organise (index) and/or rank information on the world wide web (WWW) in response to end-user queries and return search results to end-user queries.[21]
- **Relevant electronic services** – services that can be accessed by end-users in Australia, including but not limited to instant messaging services; Short Message Services and Multimedia Message

[16] Online Safety Act 2021 (Cth), s 5 (definition of 'App Distribution Services'). Excludes links to an app and download of apps from third party websites.
[17] Online Safety Act 2021 (Cth), s 14.
[18] Online services' are social media services, relevant electronic services, designated internet services and internet carriage services: Online Safety Act 2021 (Cth), s 135(2)(h).
[19] 6 Online Safety Act 2021 (Cth), s 17. eSafety notes that 'internet service provider' is defined in s 19 of the OSA as 'a person who supplies, or proposes to supply, an internet carriage service to the public.'
[20] 7 Online Safety Act 2021 (Cth), s 5.
[21] Excludes search functionality within platforms where content or information can only be surfaced from that which has been generated/uploaded/created within the platform itself and not from the WWW more broadly.

Services; chat services; online multiplayer gaming services; email services; online dating services; enterprise messaging services.[22]

- **Social media services** – services enabling online social interaction between end-users. For example, social networks; public media sharing networks; discussion forums; consumer review networks.[23]

Illegal and restricted content is defined by reference to the National Classification Scheme. The Classification Scheme does not distinguish between 'real' material and synthetic material. This means that guardrails set out in the codes and standards can apply equally to AI-generated content, including deepfakes.

If these codes meet appropriate community safeguards, the Commissioner can register them. This makes them binding on all industry participants. If a code fails to meet these requirements, eSafety can develop an enforceable industry standard for that section of the online industry instead. This ensures appropriate protections are in place for the community.

The development of these codes and standards has occurred in two phases. Phase 1 applies to 'class 1' material, such as child sexual exploitation and pro-terror content. There are currently six phase 1 industry codes in operation. They apply to social media services, app distribution services, hosting services, internet carriage services, equipment providers, and search engine services.

In addition to these codes, there are two phase 1 industry standards which have been registered and will take effect on 22 December 2024. They apply to relevant electronic services and designated internet services.

The second phase of industry codes development focussed on class 2 material formally commenced on 1 July 2024. 'Class 2' material is material that is inappropriate for children. It includes online pornography and other high-impact material, as defined by reference to the National Classification Scheme.

To summarise:

- Deepfake intimate images depicting children will likely constitute class 1 child sexual exploitation material, which is covered by the phase 1 codes and standards.
- Deepfake intimate images depicting adults may constitute class 2 material, which is likely to be covered by the phase 2 codes. However, those codes are concerned with mitigating harm to children associated with accessing such material, as opposed to mitigating harm to adults who may be depicted in content.

I will now outline the application of phase 1 codes and standards to deepfakes in more detail. They contain important protections for deepfake class 1 material. For example:

- The **Social Media Services Code** requires certain social media service providers to disrupt and deter users from creating and sharing child sexual abuse material and pro-terror material, regardless of whether the material is genuine or synthetic.[24]
- The **Search Engine Services Code** requires certain providers to make improvements to reduce the safety risks to end-users concerning synthetic materials generated by AI on the search engine

[22] Online Safety Act 2021 (Cth), s 13A.
[23] Online Safety Act 2021 (Cth), s 13. Online social interaction does not include online business interaction.
[24] Schedule 1 – Social Media Services Online Safety Code (Class 1A and Class 1B Material, minimum compliance measures 10.

service, and to research detection technologies that assist end-users in identifying deepfake images that are accessible from the service.[25]

- The **App Distribution Services Code** requires app stores to take appropriate action against apps that breach agreements to comply with Australian content laws and regulations.
- The **Designated Internet Services Standard** contains specific categories with distinct risk profiles and proportionate obligations. eSafety's fact sheets available at fact sheets detail how services are treated within the DIS Standard.

There are also two specific categories within the Designated Internet Services Standard (the Standard) that set requirements on key services in the generative AI ecosystem.

1. The Standard regulates high risk consumer facing generative AI services under the 'high impact generative AI DIS' category.

   A service will fall within the category of 'high impact generative AI DIS' if it has not incorporated adequate safety controls that mean the risk of the service generating high impact material (X18+ or RC) is immaterial.[26] Services within this category will be expected to:

   - have a suite of requirements including providing reporting tools
   - enforce terms of use covering child sexual exploitation and pro-terror material
   - deploy safeguards specific to the generative AI context.

   As outlined above, where generative AI services fail to incorporate sufficient controls to reduce the risk of producing high impact material, they will meet the 'high impact generative AI DIS' category under the Designated Internet Services Standard. This may include some apps that generate pornography or 'nudify' images without effective controls to prevent their application to children.

2. The Standard also contains world leading rules for model distribution platforms, introducing friction to address the unique risks of open-source models.

   Emerging evidence indicates most computer-generated child sexual exploitation material, including deepfake child sexual exploitation material, is based on an openly released generative AI model.[27] If it is openly available, it enables anyone to download and adjust the model, including by removing its safeguards and even finetuning it to produce harmful material.

   Putting requirements on model distribution platforms recognises the important role that these services play in the digital ecosystem and AI supply chain. The use of proportionate obligations aims to incentivise these services to consider the risks of the models they host. This includes the risk that they may be used to generate synthetic child sexual exploitation material and pro-terror material.

   eSafety will publish regulatory guidance on the Designated Internet Services Standard prior to its commencement.

---

[25] Schedule 6 – Internet Search Engine Services Online Safety Code (Class 1A and Class 1B Material), minimum compliance measures 1(e) and (f).
[26] See definition of a 'high impact generative AI DIS' in section 6 of the *Online Safety (Designated Internet Services— Class 1A and Class 1B Material) Industry Standard 2024*.
[27] Kapoor, Sayash, et al. "Position Paper: On the Societal Impact of Open Foundation Models." Forty-first International Conference on Machine Learning.

Basic Online Safety Expectations

The Online Safety Act allows eSafety to require reports on services' implementation of the Basic Online Safety Expectations (Expectations). These reporting powers aim to improve the transparency and accountability of online services. They have the goal of incentivising improvements in the safety of services on a broad range of harmful and unlawful material and activity. The obligation for services to respond to a reporting requirement is enforceable and backed by civil penalties and other enforcement mechanisms.

The Expectations apply to a narrower section of the online industry than codes and standards. They align with the services that eSafety's complaints schemes cover social media services, relevant electronic services and designated internet services.

However, the Expectations cover a wider range of unlawful and harmful material or activity than industry codes and standards, which focus on Class 1 and Class 2 material only. This includes material or activity that falls within the remit of the Online Safety Act or impedes the online safety of Australians more broadly.

The Expectations apply specifically to non-consensual intimate images, regardless of whether an image is 'real' or artificially generated or altered. They also apply to related activity, such as sexual extortion and generation of deepfakes. Providers are expected to take a range of foundation steps to ensure online safety, including:

- Reasonable steps to ensure that end-users are able to use the service in a safe manner (section 6(1)).
- Reasonable steps to proactively minimise the extent to which material or activity on the service is unlawful or harmful (section 6(2)).
- Reasonable steps to make available controls that give end-users choice and autonomy to support safe online interactions (section 6(5)).
- Reasonable steps to consider end-user safety and incorporate safety measures in the design, implementation and maintenance of generative AI capabilities on the service (section 8A(1)) and proactively minimise the extent to which generative AI capabilities may be used to produce material or facilitate activity that is unlawful or harmful (section 8A(2)).
- Reasonable steps to minimise the extent to which non-consensual intimate images are provided on the service (section 11).
- Have terms of use, policies and procedures and standards of conduct in relation to online safety (including but not limited to, material specified in the Act such as non-consensual intimate images) and take reasonable steps to detect breaches of terms of use, policies and procedures or standards of conduct and enforce any specified penalties for these breaches (section 14(1), (1A) and (2)).
- Provide clear and readily identifiable report and complaint mechanisms in relation to this material (section 13(1)(c)) and review and respond to these reports and complaints within a reasonable period of time (section 14(3)).

eSafety has published Regulatory Guidance in relation to the Expectations. This includes additional examples of steps providers can take to ensure they are compliant with each applicable expectation.[28]

[28] Regulatory schemes | eSafety Commissioner

As noted above, while the Expectations are not enforceable, our reporting powers allow us to improve the transparency and accountability of providers and incentivise improvements in safety.

To date, we have focused our use of these powers on understanding what steps providers are – and are not – taking to ensure safety on issues like child sexual exploitation and abuse, image- based abuse, and generative AI. Reports providing findings in response to the transparency notices are available on the eSafety website.

eSafety can also publish statements regarding providers' compliance and non-compliance with the Expectations.

Safety by Design

We work with tech companies to shift their design ethos from 'moving fast and breaking things' to moving thoughtfully and anticipating, detecting and eliminating online threats before they occur. This is 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 image-based abuse, including that perpetrated through deepfakes. Steps that can be taken by service providers include:

- Developing community guidelines, terms of service and moderation procedures that address image-based abuse, and consistently and fairly implementing them.
- Establishing clear protocols and consequences for service violations related to image-based abuse that serve as meaningful deterrents.
- Putting processes in place to detect, surface, flag and remove image-based abuse, with the aim of preventing harm before it occurs.
- Putting in place infrastructure that supports internal and external triaging, clear escalation pathways and reporting on image-based abuse, alongside readily accessible mechanisms for users to flag and report concerns and violations at the point they occur.
- Carrying out open and meaningful engagement with a wide user base including diverse and at-risk groups, independent experts and other key stakeholders.

- Committing to consistently innovate and invest in safety-enhancing technologies and collaborate and share tools, best practices, processes and technologies with others.

To help industry, we have developed practical resources via the Safety by Design assessment tools. This includes educative content on intersectional risk factors for online harms, insights into perpetrator motives and exploration of human rights in the digital context.

Conclusion

In summary, eSafety welcomes and supports the Bill. Combined with eSafety's functions and powers, it will lead to a more holistic and comprehensive range of support, redress options, and choice for victim-survivors of image-based abuse.