DIGITAL RIGHTS WATCH©

All or nothing?

The relationship between privacy and safety in addressing online harms

Polling, policies and strategies for harmful material, including Child Sexual Abuse or Exploitation Material, detection and removal, and possible privacy implications

NOVEMBER 2024

Report commissioned by Digital Rights Watch
Polling by Essential Media
Rapid desktop review prepared by Scheherazade Bloul, PhD
Acknowledgements: Dr. Caitlin McCrane and Xan Cooper for assistance.
Editing: Hella Ibrahim

Contents

Executive summary 3
Framing the problem 6
Safety and privacy as counterposed 9
2.1 Cyber detection and identification strategies 9
Peer-to-Peer (P2P) Network Monitoring 11
Automated Multi-Modal CSAEM Detection Tools 11
Web Crawlers for CSAEM Site Identification 11
Facial Recognition Technology 12
2.2 Age verification and age assurance 12
2.3 Client-side scanning 13
2.4 Limits of Machine Learning (ML) technologies 14
2.5 Considerations for content removal and banning 15
2.6 Detection methods 16
Improving safety without while impacting privacy in limited ways 18
3.1 Case study on CSAEM prevention and intervention 18
Types of Initiatives 19
Treatment Approaches 19
Current research limitations on CSAEM prevention strategies effectiveness 20
3.2 Encryption 21
3.3 Pop-Up Warning Messages 21
3.4 Platform design 22
3.5 Privacy literacy (education) 23
3.6 User reporting 23
Future research opportunities 25
Conclusion 27
Appendices 29
Appendix A: Acronyms and abbreviations 30
Appendix B: Scope and methodology 31
Appendix C: Limitations 32
Appendix D: Policy landscape 33
Online Safety Act 2021 (The Act) 34
Informing policy 35
Appendix E: The Essential Report – Digital Rights Watch 38
References 45

Executive summary

This report was commissioned on the occasion of an event (with the same name) which took place at the SXSW festival 2024 in Sydney featuring the Privacy Commissioner, Carly Kind, the eSafety Commissioner, Julie Inman Grant, Professor Michael Salter from UNSW and Chair of Digital Rights Watch, Lizzie O'Shea.

The report consists of two parts:

1. Polling conducted by Essential Media, asking a statistically representative sample of Australians about issues at the intersection of privacy and safety

2. A rapid desktop review of current academic literature, policy frameworks and technological interventions that concern the relationship between online safety and privacy, with a focus on Child Sexual Abuse or Exploitation Material (CSAEM) and other online harms.

Key Insights from polling

The starting point for most people is that the web is a safe space, but we do hold concern for both older people and children. We are uncomfortable with the amount of personal information being collected by private companies, while holding mixed views of government holding our data.

Australians understand the challenges and recognise the importance of privacy and safety, and the dynamic between the two:

- Most participants think that the online web is not that safe or not safe at all for children (69%) and older people (51%).
- Almost half (46%) think that the collection and use of online information by governments and private companies makes the online web less safe.
- 69% of participants think that online safety and privacy are equally important.

Critically, we do not believe that we should have to trade off our privacy for a safer internet. This is born out with strong support for privacy reforms, which is even more supported than banning children from social media, although this too has high levels of support. We are more sceptical about the idea of using AI to police web safety, unconvinced that these techno-centric approaches will create safer online environments.

This polling research should reinforce the need to carefully think through approaches to improving online safety and ensuring that any form of age or identity verification includes appropriate safeguards and protections.

The complete dataset obtained via polling is at Appendix E on page 38.

Key insights from the rapid desktop review

The rapid desktop review explores the complex relationship between online safety and privacy, with a focus on Child Sexual Abuse or Exploitation Material (CSAEM) and other online harms. The review synthesises current academic literature, policy frameworks and technological interventions, highlighting gaps in research and presenting evidence-based recommendations. It adopts a critical media studies perspective, challenging dominant techno-solutionist approaches that tend to frame safety as incompatible with privacy. It examines some holistic approaches that integrate both elements while addressing the socio-cultural factors that contribute to online harm.

The debate surrounding online safety often positions privacy and safety in opposition, leading to an overreliance on surveillance technologies that prioritise detection methods at the expense of user privacy. This binary framing oversimplifies the complexities of digital life, ignoring broader socio-political and economic contexts (see Burgess et al. 2022).

Platforms, driven by profit and engagement algorithms, unintentionally amplify harmful content such as extremism, misinformation, and CSAEM, complicating efforts to ensure safety (Zubhoff 2019; Marwick et al. 2024). These challenges are compounded by the fragmentation of global laws, which vary in definitions of harm, hate speech, and extremist content, leading to uneven enforcement and governance issues (Phillips et al. 2022; Kansok-Dusche et al. 2023).

Current responses prioritise techno-carceral solutions, which ignore the broader socio-cultural and economic dimensions of online abuse (Stardust et al. 2023; Marwick et al. 2024).

The findings suggest that by promoting user-centred safety designs, education, and structural interventions, governments can better safeguard individuals in digital spaces, particularly children, young people, and marginalised communities.

The challenges of online safety and privacy cannot be addressed through simplistic techno-solutionism. Instead, a holistic approach is needed, one that recognises the interplay between privacy, safety, and the socio-cultural conditions enabling harm. Policy interventions and evidence-based policy responses must prioritise privacy as a core component of safety, rather than treating the two as mutually exclusive. This approach involves fostering digital literacy, supporting harm-reduction strategies, and developing privacy-preserving technologies.

The rapid desktop review makes up the remainder of this report.

This report was made possible by an unconditional grant made by Apple Inc, and other sources of funding provided to Digital Rights Watch as outlined in the organisation's annual reports.

1

SECTION 1:

FRAMING THE PROBLEM

Framing the problem

Public discourse and policy debate around online safety often sets privacy against safety, fuelling a securitisation logic that prioritises surveillance in detection methods.

This binary framing oversimplifies the complexities of digital life, ignoring broader socio-political and economic contexts. The complexity of the often-counterposed **online safety versus online harm at the expense of privacy** paradigm extends far beyond the binary of privacy versus security, encompassing deeper issues tied to the **economic structures** and **governance of digital platforms themselves** (Zubhoff 2019; Gilbert 2024; Marwick et al. 2024; Stardust et al. 2024).

Platforms, driven by engagement and profit, often use algorithms that inadvertently amplify harmful content such as extremism and misinformation, further complicating efforts to ensure user safety (e.g. Ashe et al. 2020; Ehsan and Stott, 2020; Miller-Idriss, 2020; Skocylis & Andrews 2022).

There is a need to understand online harms critically, particularly in understanding political violence. A pertinent example of this in the southern hemisphere is the Christchurch attacker, who while live-streaming the terror attack on a Muslim community in Aotearoa/New Zealand, did not so in a vacuum. The War on Terror post 9/11 and subsequent mainstreaming of far-right political discourse in Australia and across the world (Bloul et al. 2021; Miller-Idriss, 2020; Skoczylis & Andrews, 2022; Winter & Mondon, 2020) contributed to and intensified the abhorrent violence.

This issue is compounded by the global fragmentation of laws and regulations, where differing cultural and legal definitions of harm – whether around extremist content, sexual content, including of children, or hate speech – create uneven enforcement and governance issues (e.g. Phillips et al. 2022; Kansok-Dusche et al. 2023) in ever-changing landscapes of online abuse (Thomas et al 2021).

Marginalised communities (such as LGBTIQA+, sex workers, content moderators in the Global Souths; overpoliced communities etc) bear the disproportionate burden of surveillance (Stardust et al. 2023, 2024), while the structural conditions that enable online harm remain unaddressed. Researchers suggest that approaches that simply prioritise law enforcement and tech industry partnerships **promote techno-carceral solutionism** (e.g. Stardust et al. 2023), while **relying on technology alone as a silver bullet leads to techno-solutionism to essentially socio-cultural, political and economic problems** (e.g. Marwick et al 2024; Angel & boyd 2024).

Within this landscape, the Online Safety Act 2021 (the Act) in Australia sets out **a regulatory framework for addressing online harm**, including CSAEM and other harmful content (see Appendix D). The Act ostensibly provides the eSafety Commissioner (eSafety) with significant powers, including the ability to issue takedown notices and enforce new industry codes. These codes have been in force since May 2024 and more will come into effect in December 2024 (eSafety 2024).

However, concerns persist regarding the effectiveness of such measures, particularly on platform self-regulation (e.g. Human Rights Law Centre [HRLC] 2024). Platforms use an array of technological tools to detect such content, while inaccuracies, false positives/negatives and duplicate data have the potential to infringe on privacy and free expression, as well as the jurisdictional reach of content removal (e.g. eSafety Commissioner v X Corp). These legal frameworks often treat privacy as secondary to safety, overlooking **privacy's role as a protective mechanism** and as **necessary for freedom of political thought foundational to liberal democracies** (Froomkin & Colangelo 2020; DRW submissions 2023; 2024).

DRW's polling research reveals that **Australians are highly aware of the challenges surrounding online privacy and safety**, with a significant majority recognising the balance between the two. A large portion of respondents believe that the internet is not safe for children (69%) or older people (51%). Nearly half of participants (46%) think that the collection and use of personal information by governments and private companies negatively impacts online safety. 69% of Australians consider both privacy and safety

to be equally important, **suggesting a widespread demand for a more secure and private digital environment.**

Australian news media, particularly print newspapers, often present parents as primarily responsible for regulating children's digital technology use while portraying the government's role as less prominent unless the issue relates to CSAEM (Duffy et al. 2024). Media reporting on the issue has significantly increased since 2020, reflecting a rise in reported cases (e.g. Duffy et al 2024). Sharing CSAEM predates the internet (Landau 2023), while emerging technologies, such as self-generated content, livestreaming, and automated technologies and Artificial intelligence (**AI**) content, have exacerbated global proliferation of reported CSAEM (Thorn 2023).

However, **the true scale of this growth is unclear.** Some of the increase may be attributed to better detection tools which are nevertheless ineffective in end-to-end encryption (E2EE) environments (Draper 2022; Landau 2023). Some E2EE platforms have implemented user-reporting features to flag harmful content, but whether content is reported to law enforcement is another matter (see Mayer 2019).

Platforms and law enforcement employ detection and removal strategies that are mired in concerns around efficacy, privacy and bias. **Reliance on automated tools, including AI-powered content moderation, risks undermining users' privacy while potentially perpetuating discriminatory outcomes or false positives**.

For example, Facebook reported that 90 percent of the CSAEM content it flagged in late 2021 was identical or visually similar to previously reported material, with **just six videos accounting for half of its reports to National Center for Missing and Exploited Children (NCMEC) in the US** (Meta 2021; see also Cronin 2022).

In Australia, a recent non-randomised survey study of 13,300 participants found that 0.8% reported intentionally viewing CSAEM content, which differed from the previous estimates of 2.2-4.6% (Brown 2023). **Differing numbers suggesting various challenges in accurately assessing the scale of this issue, as self-reporting on sensitive topics like CSAEM is prone to variability depending on the study's methodology and context**.

Police and government use some studies for statistics and profiling to detect and identify users of producers of CSAEM. If the statistics identify certain marginalised demographics without identifying others, it provides a biased basis from which risk assessments can be made (particularly if automated). It can lead to potentially overpoliced marginalised communities and can put people in early contact with the criminal justice system (Farrell et al 2024; Sentas 2022; Ulbrick 2021).

For example, following an eight-year campaign by the Youth Justice Coalition, NSW police ended their **unlawful use** in 2023 of the **Suspect Target Management Plan (STMP)**. This plan applied **risk assessments based on aggregated data** to allocate police resources in an attempt to prevent crime (LECC 2023). Aboriginal children were **10 times more likely** to be flagged under this system compared to the general population (Sentas & Pandolfini 2017).

There is a lack of knowledge (or accessibility) on what technological tools are being used by law enforcement in Australia to detect, prevent, and identify crimes, including viewing CSAEM. However, we know from some research from the Australian Institute of Criminology that law enforcement uses a variety of methods to detect content and identify users. These tools include but are not limited to: **peer-to-peer network monitoring**, **automated multi-modal CSAEM detection tools**, image hashing **web crawlers to identify CSAEM sites**, **pop-up warning messages**[1], and **facial recognition** (Edwards et al 2021; InHope 2021; Westlake et al. 2022; Brown 2023).

1 When users enter search terms that could be identified as associated to CSAEM, pop-up messages could appear, warning them about the legal consequences of accessing illegal content (Prichard et al. 2011; Prichard et al. 2021). Research reviewed in this area **but outside of CSAEM-related content** (e.g Silic & Black, 2017; Anderson et al., 2016; Clayton et al., 2020; Caillon et al., 2021) consistently shows that online warnings can influence user behaviour and suggest that internet users are generally responsive to warnings related to hazardous online activities

2

SECTION 2:

SAFETY AND PRIVACY AS COUNTERPOSED

Safety and privacy as counterposed

Policies that regulate online spaces can treat safety and privacy as counterposed, particularly when addressing issues like online hate speech, violent content, and/or CSAEM material (see for e.g. Smith et al. 2024). These policies require platforms to identify and remove harmful content as defined under the Act, which can involve using user data and potentially infringing on user privacy.

For example, efforts to combat CSAEM often involve surveillance and monitoring of online spaces, which can raise privacy concerns (ECPAT 2019; Marwick et al 2024; Angel & boyd 2024).

Within the disciplines of traditional criminology and crime prevention, situational crime prevention (SCP) theory (see Clarke 1980) and crime prevention through environmental design (CPTED) underpin common strategies used by police to prevent and control cybercrimes (see Krone et al. 2020; Ho et al. 2022). SCP is a primary prevention measure aimed at reducing crime opportunities rather than focussing on individual characteristics of potential offenders (Clarke 1995).

SCP is rooted in a utilitarian, risk-oriented approach, focusing on minimising opportunities for crime through strategies such as increasing the effort required to commit crimes, amplifying risks for offenders, reducing potential rewards, and eliminating provocations and excuses (Cornish & Clarke, 2003; Cozens & Love 2017). SCP as a strategy has been widely adopted by many Western governments (see UK Home Office, 2016; Ministry of Justice New Zealand 2017; Canadian Council on Social Development, 2020) including Australia informs the criminological framework of policy as part of a broader environmental approach to crime prevention (see Australian Institute of Criminology 2012; Clancey et al 2016).

However, this preoccupation with managing crime through technocratic control mechanisms overlooks the broader sociopolitical and economic conditions that give rise to criminal behaviours in the first place. Rather than interrogating crime's root causes, such as systemic inequality or state violence, SCP operates within neoliberalist logic, where individuals are responsible for navigating the risks imposed by their environment and crime is treated as a disruption to the smooth functioning of markets and social order (Kleinig 2000; Freilich & Greene-Colozzi 2024; Parnaby 2006).

In this context, SCP serves not only as a crime control strategy but as an expression of the neoliberal state's broader shift toward governance through risk management (Ho et al., 2022; Parnaby 2006). In other words, SCP manages the symptoms rather than the roots causes (Kleinig 2000; Raymen 2016).

2.1 Cyber detection and identification strategies

Some primary technological strategies platforms use to detect CSAEM or other harmful content includes image hash databases, web-crawlers, visual detection algorithms, and deep/machine learning methods (e.g. Lee et al. 2020).

The primary technology to assess and detect CSAEM is image hashing (Westlake et al., 2012; da Silva Eleuterio et al., 2012) which is a technique used to generate a unique, fixed-size representation (hash) of an image, allowing for efficient comparison, retrieval, and detection of duplicate or similar images in large datasets (InHope 2024; Fonseca-Bustos et al. 2022, 2024).

Hashing and matching technologies enable platforms to detect previously reported CSAEM by converting digital files into unique hash values – essentially digital fingerprints – using either perceptual or cryptographic hashing methods. These hash values are then compared against known CSAEM databases to identify and remove illegal content. This method's effectiveness relies heavily on the size and comprehensiveness of the hash databases available to companies.

Platforms employ a variety of methods to detect CSAEM, with **PhotoDNA**, AI, and **Content Safety API** being some of the most common technologies (Teunissen & Napier 2022).

Perceptual Hashing remains a well-established and highly efficient technique, using digital fingerprints to match known CSAEM against databases. **PhotoDNA**, developed by Microsoft in 2009, compares the digital fingerprints (hashes) of images and videos against known CSAEM databases and has been widely adopted by platforms such as Microsoft, Meta, Snapchat, and Twitter to detect abusive content (Thorn 2016; Allen 2011; Davis 2020; Snap Inc 2024).

Despite its effectiveness, PhotoDNA can only identify previously known and catalogued CSAEM, leaving new or altered material undetected, and investigations have revealed vulnerabilities in its application, particularly on Bing (Keller & Dance 2019). **Moreover, it cannot operate in end-to-end encrypted settings, where content access is restricted.**

AI and Machine Learning (ML) play a growing role in CSAEM detection, particularly for platforms like Meta and Google, where AI is used to proactively detect unknown or newly generated CSAEM (Davis 2018; Google n.d., 2024; see also the Safer n.d.).

AI Image Content Analysis offers the potential to identify both known and newly created CSAEM, but its large-scale application is hindered by issues of accuracy and reliability. There is also little transparency about how these AI systems function, and empirical evaluations on their effectiveness are scarce. The availability of training data is severely restricted by laws prohibiting CSAEM possession, and AI's performance is particularly unpredictable in encrypted environments where it may encounter unfamiliar data and images, complicating verification processes.

Meta and Google have both integrated **Google's Content Safety API**, which uses AI to triage and prioritise CSAEM for human review, making it available to other companies and NGOs including the Internet Watch Foundation (Canegallo 2021). Platforms like Google and YouTube also use the **CSAI Match** tool, which specifically targets CSAEM video content (Google 2024).

These tools, combined with deterrent measures such as pop-up messages and warnings, are used by several platforms, but gaps remain in detecting new or substantially altered CSAEM and in ensuring transparency around the algorithms used (Prichard et al. 2022; Gladstone 2019).

Research on the cyber strategies employed by police to 'combat' CSAEM narrows the fields to a few strategies including **peer-to-peer network monitoring**, **automated multi-modal** CSAEM detection tools, **web crawlers** to identify CSAEM sites, **pop-up warning messages**, and **facial recognition** (Edwards et al 2021). The authors argue that these strategies have the potential to significantly reduce CSAEM availability and distribution by increasing the detection risk for offenders and making it more difficult to access and produce CSAEM.

Further research is needed to fully assess the effectiveness of these strategies and to explore the potential ethical and legal challenges associated with their use (Edwards and Christensen, 2021).

Peer-to-Peer (P2P) Network Monitoring

Mechanism: Law enforcement employs investigative tools (e.g. metadata analysis) to monitor data exchanged over peer-to-peer networks. By tracking IP addresses, hash values, and file names, investigators identify the dissemination of CSAEM and locate perpetrators. Additionally, it supports the takedown of websites hosting CSAEM.

Limitations: P2P monitoring is labour- and resource-intensive. It is also vulnerable to obfuscation techniques such as steganography, which offenders use to conceal illicit content. Analysing activities on P2P networks is challenging due to their complex structures (Lee et al. 2020). Wolak et al. (2014) found that less than 1% of computers on the Gnutella network contributed significantly to CSAEM trafficking, suggesting that focusing on high-contribution users could reduce CSAEM circulation.

However, these estimates are conservative, as only known CSAEM was tracked and the study spanned over 100 countries. Research by Bissias et al. (2016) used unique identifiers to measure P2P activities, avoiding the overestimation caused by shared IP addresses, and found that while CSAEM trafficking via P2P networks is declining overall it remains active, particularly on the BitTorrent network. Other studies, such as those by Al Mutawa et al. (2015) and Shavitt et al. (2013), argue for considering evolving technologies like social networks, mobile apps and the darknet to stay ahead of CSAEM distribution trends.

In Australia, there is limited research on P2P monitoring. The focus is on finding and targeting possible CSAEM viewers, while a Brown et al. (2023) non-randomised sample study discusses that most CSAEM viewers in their sample had unintentionally viewed the content.

Automated Multi-Modal CSAEM Detection Tools

Mechanism: Tools like PhotoDNA (used by Meta, Microsoft, TikTok, Twitter/X) leverage machine learning algorithms to analyse multimedia files, examining features such as skin tone, shapes, text, and colour patterns to flag potential CSAEM. Law enforcement perspectives (Edwards and Christensen) argue the method significantly enhances the efficiency and precision of detection efforts, thereby alleviating the workload of digital forensic teams and increasing the likelihood of identifying offenders.

Limitations: Such tools can be bypassed through techniques like steganography. Furthermore, effective use requires specialised training for law enforcement personnel to properly interpret flagged content.

Web Crawlers for CSAEM Site Identification

Mechanism: Automated web crawlers are deployed to scan websites and follow internal links to map and identify networks involved in CSAEM distribution. These scripts expedite the identification of CSAEM distribution hubs, enabling law enforcement to carry out targeted takedowns of key nodes in these networks.

Limitations: The success of this strategy depends heavily on the accuracy of keywords and hash values used. Additionally, law enforcement must first pinpoint relevant sites and terminology for the crawlers to track.

HASHING: The current database is limited to searches for CSA images, Edwards & Christensen (2021) suggest it necessary to expand its scope to include other formats, such as videos, for comprehensive detection of CSAEM. Additionally, the database is ineffective in identifying newly uploaded content, as it relies on pre-existing hash values, rendering it incapable of detecting CSAEM that has not previously been classified (Westlake et al., 2012).

Text-Based Analysis evaluates conversations for grooming or enticement patterns and can pre-emptively identify risky interactions. However, keeping AI models updated with evolving language, particularly slang used by children, presents a challenge. This method is also constrained in encrypted environments where privacy and free speech concerns restrict companies from acting on flagged conversations.

Facial Recognition Technology

Mechanism: AI-powered facial recognition tools, technology such as Clearview AI which Australian police forces trialled (Office of the Australia Information Commissioner [OAIC] 2021), are used to match faces in CSAEM with images from law enforcement databases, to identify victims and offenders.

Limitations: Facial recognition technology raises significant ethical and legal concerns, particularly regarding privacy (see OAIC 2021; Matulionyte 2024), data security, and the risk of biased or inaccurate matches as detailed by Stardust et al (2024). Racialised misidentification in facial recognition software can increase the likelihood of discriminatory policing practices (Buolamwini & Gebru, 2018). Furthermore, there is a notable lack of robust evaluation research on its efficacy. Because facial recognition technology is rapidly developing, it is unclear what further limitations may arise (Edwards & Christensen 2021).

These strategies' effectiveness is not fully understood because there has been limited evaluation research. The tech literature surveyed that highlights a reduction in online CSAEM is "a joint endeavour by both law enforcement and industry" (Edwards & Christensen, 2021, p. 10; see also Bleakley et al. 2024). This suggests some research trends acknowledge the issue is multifaceted, but one that requires collaboration between police and industry, rather than other approaches that consider communities, academics, civil society organisations, government, industry. Authors note that there is a need for further research in the field.

2.2 Age verification and age assurance

Age assurance methods are designed to verify the age of users to comply with legal requirements and protect minors from inappropriate content. However, these methods often face significant compromise user privacy. One primary limitation of age assurance methods is their reliance on collecting and processing sensitive personal data. For instance, many systems require users to provide biometric data, such as voice or facial recognition, which can be exploited if not properly secured (Jarvie & Renaud 2021).

Research indicates that biometric characteristics, including age-related features, can be used to identify individuals, raising substantial privacy concerns (Meden et al 2021; Maouche et al., 2022; Pathak 2012). The collection of such data not only increases data breaches risk but also poses ethical dilemmas regarding consent and the potential for misuse (Gupta et al., 2019; Meden et al 2021).

Moreover, existing age verification techniques often lack transparency and user control, which are essential for maintaining trust. Users may be unaware how their data is being used or shared, leading to heightened privacy concerns. Studies show trust is a significant predictor of privacy concerns, and a lack of transparency can exacerbate these worries (Wang et al., 2019; Zheng et al., 2015). This is particularly relevant in environments where users are required to disclose sensitive information to access services, as the perceived risk can deter individuals from engaging with platforms that employ stringent age verification measures (Wang et al., 2019).

Another aspect is the effectiveness of these methods in accurately determining age without infringing on privacy. Many traditional age verification techniques, such as document verification or third-party databases, can be intrusive and may not guarantee accuracy. For example, government-issued IDs can expose users to identity theft if data is mishandled (Saleh et al., 2023; Bhattacharjee 2024). The challenge of balancing privacy with the need for accurate age verification remains unresolved, as many systems struggle to implement robust privacy-preserving techniques while ensuring compliance with legal standards (Meden et al., 2021; Bhattacharjee 2024).

There are significant gaps and concerns over everyday use of age assurance tools. A rapid evidence review of 61 studies on age assurance and parental controls found that existing measures are often ineffective and sometimes even counterproductive, particularly in respecting children's rights to privacy and autonomy

(Smirnova et al 2021). The current age assurance measures, such as self-declaration tools, are easily circumvented and fail to protect children in high-risk situations (Nikitin et al 2016; Williams et al. 2017; Williams et al. 2020).

More robust methods like third-party age verification raise concerns about privacy (Nikitin et al., 2016; Williams et al., 2015), data minimisation and online fraud, particularly for children (Williams et al. 2017). Additionally, there is limited research on how age assurance is used within families, making it difficult to assess the full impact of these measures on children's well-being, resilience, and rights. Many parents seek flexibility in age restrictions, preferring solutions that adapt to their family values rather than rigid, universal approaches.

However, current tools often prioritise certain conceptualisations of safety over other rights, such as participation, learning and privacy, leading to overly restrictive measures that can hinder development (Smirnova et al 2021). These tools are often designed without considering children's perspectives, which is critical to ensuring effectiveness and respecting children's evolving capacities (see Digital Child 2024; Livingstone 2013).

Parental control tools tend to focus on restrictive mediation, but excessive restriction can damage parent-child relationships and prevent children from learning essential online coping skills. The diverse needs of families, including cultural norms, parenting styles and the specific challenges faced by vulnerable or marginalised groups are often overlooked in these tools' design.

eSafety announced that the age assurance market was immature with significant gaps, citing feasibility and technical concerns and recommending media literacy and education (2023a). Stardust et al (2024) raise several ethical and practical limitations of using age verification software to restrict access to online pornography. Privacy risks arise from the collection and storage of sensitive data (such as the metric used in face scanning through a user's camera.

Stardust et al (2024) propose that governments and policymakers redirect resources towards comprehensive, age-appropriate sex education that includes discussions about pornography, consent, and online safety89. Additionally, they recommend promoting media literacy to help young people critically evaluate online content and make informed decisions.

2.3 Client-side scanning

Client-side scanning (CSS) is a technology that allows for data analysis on a user's device as opposed to server-side scanning, where analysis is conducted once server data is located. Abelson et al. (2024) argue that CSS is not a reliable solution for preventing crime or protecting public safety, as it creates significant individual privacy and security risks.CSS can also be easily evaded where people can use a variety of methods to avoid detection or trigger false alarms (Abelson et al 2024).

The Irish Council for Civil Liberties (ICCL) published a report indicating that CSS tools had high rates of false alarms and low accuracy (Cronin 2022). The report found 20% constituted actual CSAEM, with fewer than 10% deemed actionable (Cronin 2022). Many false positives involved innocent individuals sharing non-offensive content, raising concerns about overreach and unlawful surveillance practices. This case challenges scanning technologies accuracy, suggesting that the high rate of false positives could impede law enforcement's ability to effectively address CSAEM.

Dutch law enforcement has expressed similar concerns, noting their limited capacity to manage the anticipated surge in reports under the EU's forthcoming regulation (European Digital Rights [EDR], 2022). The ICCL contends that these inefficiencies raise critical questions about whether the European Commission's approach is truly serving the children's best interests. EDR argues that more holistic approaches are necessary. Rather than "naïve faith in technology as a silver bullet", policy approaches should focus on:

> education, awareness-raising and empowerment of survivors, social and structural change, reforming the police and other institutions, investing in child protection services, and bringing together child rights and digital rights groups, as well as other experts, to work together on solutions. *(European Digital Rights, 2022)*

The potential for "scope creep" raises concerns about CSS being expanded beyond its intended use to target content such political sensitive material (e.g. The Australian state targeting journalists and whistleblowers), with policy decisions and software updates being the only barriers to this expansion. CSS introduces new security vulnerabilities by widening the attack surface, making it susceptible to hackers, corrupt officials or abusive individuals (The Internet Society, 2022). Adversaries can employ techniques to evade detection or trigger false positives, further undermining CSS's reliability.

Moreover, the technical and ethical challenges of deploying CSS at scale, such as ensuring fairness, addressing software bugs, and navigating jurisdictional complexities, erode trust between users and technology providers. CSS poses significant privacy risks by allowing the examination of content on personal devices, effectively blurring the boundary between private and public spaces.

In an open letter address to EU and UK lawmakers, researchers called on government to reconsider introducing policies that would necessitate the use of scanning technologies arguing that "the scanning technologies that currently exist and that are on the horizon are deeply flawed" (Martin 2023). The flaws lie in false negatives and positives as discussed in (Abelson et al. 2024).

The case studies and research indicate that CSS cannot be deployed safely, effectively, or accurately. CSS is both technically ineffective and impractical as a strategy for addressing violent online political extremism and curbing the distribution of CSAEM (Anderson & Gilbert 2022).

The limitations of CSS lie in its inability to offer comprehensive protection against these threats while simultaneously raising serious privacy concerns. Its technical flaws make it unreliable, as the scanning process can be circumvented or compromised, allowing classified harmful content to evade detection. Implementing CSS at scale would require invasive access to private communications, which poses significant questions around surveillance and can potentially further undermine user trust and violate privacy rights. In practice, CSS does little to address the structural issues that facilitate the spread of CSAEM. It is a blunt tool that fails to offer meaningful solutions to these complex problems.

2.4 Limits of Machine Learning (ML) technologies

Machine learning (ML) technology, while capable of processing vast datasets, is inherently limited by its inability to fully grasp context and nuance, often resulting in misclassifications due to biased or incomplete training data, as has been found in studies on extremist content detection online (Costello et al 2018; Khan et al., 2022). For example, counterterrorism applications face criticism due to the complexity of terrorist acts and issues like class imbalance and high-dimensional data (Brown et al 2023; Shortland et al., 2017; L'Heureux et al., 2017). Consequently, even accurate models may misclassify many individuals, offering limited security benefits relative to privacy costs (Verhelst et al., 2020; Begoli et al., 2019; Dunson, 2018; Soghoian, 2008).

Machine learning techniques applied to metadata in mass surveillance are often presented as less invasive than full data collection, but this argument is flawed (Begoli et al., 2019; Dunson, 2018). Metadata, such as the time and length of communications, can still reveal highly sensitive information when analysed by machine learning algorithms.

Studies have shown that algorithms can de-anonymise data, identifying individuals from seemingly innocuous information such as credit card transactions (de Montjoye et al., 2015) or mobile phone activity (Mayer et al., 2016). Additionally, the distinction between metadata and data is often ambiguous – for example, an IP address might be considered metadata in one system but data in another (Feigenbaum & Koenig, 2014). This lack of clarity increases the risk of overreach and privacy breaches.

As machine learning algorithms become more sophisticated, the potential to extract personal information from metadata will only grow, making the collection of metadata a significant infringement on privacy. Verhelst et al. (2020) recommend that policymakers critically reassess the assumption that metadata analysis constitutes only a minor invasion of privacy. Given machine learning algorithms' increasing sophistication, metadata can reveal highly sensitive personal information.

2.5 Considerations for content removal and banning

The Online Safety Act 2021 (the Act) concentrates on content control over design principles (Reset Australia 2024). While the primary goal of content removal is to protect people online, research indicates that these measures can inadvertently lead to alternative means to share harmful content. For example, although outside of CSAEM, countermeasures such as content removal and deplatforming can lead groups to migrate to less regulated spaces, such as anonymous forums and encrypted messaging apps (Casilli et al 2013). This shift complicates detection efforts, as traditional monitoring strategies may not extend to these alternative channels.

One significant concern is that removing harmful content can push individuals toward less regulated spaces on the internet, where they may encounter even more extreme or harmful materials (e.g. Brennan et al. 2022; Casilli et al 2013).

For instance, Casilli et al. (2013) found that censorship of pro-anorexia/bulemia online communities – websites and social media groups promoting anorexia and bulimia – altered the network's structure, leading to more insular and less interconnected communities, thereby complicating healthcare professionals and support organisations' ability to effectively intervene and aid.

A decade later, Brennan et al. (2022) argue that while certain online content related to self-harm is widely recognised as harmful, removal can lead individuals to seek out alternative platforms that may not have robust moderation policies in place. This can create echo chambers where harmful behaviours are normalised and reinforced, ultimately exacerbating the very issues that content removal seeks to mitigate.

While automated systems can quickly identify and remove harmful content, they may lack the contextual understanding necessary to differentiate between harmful and benign expressions of self-harm. Lewis and Seko argue that using machine learning algorithms to filter content can help identify high-risk individuals, but these systems must be designed carefully to avoid misclassifying content that may serve a supportive purpose for some users (Lewis & Seko, 2015).

Often content removal measures, through automated detection, discriminate against LGBTQIA+ groups, sex education content, and activist content (e.g. Are 2023, 2024; Stardust et al 2024).

The challenge lies in developing algorithms that can accurately assess user-generated content's intent and context, which requires ongoing refinement and validation. As well as algorithmic challenges, the ethical implications of content removal must be addressed.

2.6 Detection methods

The detection of hateful and abusive content is seen as paramount to mitigate the proliferation of online harmful behaviour (Schneider and Rizoiu, 2023). While automatic detection methods have garnered significant attention, most existing models rely on train-test splits within the same dataset, leading to limited generalisation across different hate speech datasets (Arango et al. 2019; Swamy et al. 2019; Fortuna et al. 2021).

A major impediment to building generalisable models is the absence of a universally agreed-upon, operational definition of hate speech, compounded by the broad nature of hateful content and resulting in inconsistent labelling criteria and narrow dataset scopes. Recent studies (e.g. Yuan & Rizoiu 2025) suggest improving models so they draw from multiple datasets to create a broader representation of hate speech. Results indicate such a model performs better than existing models on unfamiliar datasets, and is comparable to state-of-the-art models on familiar datasets.[2]

However, such studies focus of improving the surveillance and technological capacities of speech detection based on binary classification (problematic vs harmless), lack of contextual awareness of a particular moment speech is occurring, database initial biases in terms of how models define and classify hate speech (often based on governmental working definitions).

2 By applying the MTL classifier to more than 300,000 tweets, Yuan and Rizoiu (2025) observed patterns of hateful and abusive speech, particularly noting that right-leaning figures produce significantly more problematic content, largely concentrated around six specific topics (including Islam, women, ethnicity, and immigrants). They note that MTL training helps to more clearly distinguish hateful and abusive speech from neutral speech.

3

SECTION 3:

IMPROVING SAFETY WITHOUT WHILE IMPACTING PRIVACY IN LIMITED WAYS

Improving safety without while impacting privacy in limited ways

While the existing body of literature often treats safety and privacy as dichotomous concepts, neglecting the potential for integrative approaches that recognise their interdependence.

For instance, while some studies have focused on the privacy paradox – where users express concern for privacy yet engage in behaviours that compromise it (Hirschprung et al. 2022) – there is a lack of research that explores how policies can be designed to simultaneously enhance both safety and privacy. Yet, in relation to responding to potential harm caused by spreading and engaging with CSAEM content[3], research suggests a holistic approach to addressing the problem (Draper 2022; Gannoni et al 2023).

3.1 Case study on CSAEM prevention and intervention

In an international review, the Australian Institute of Criminology (AIC) surveyed 74 articles on initiatives aimed at preventing child sexual abuse material (CSAEM) offending. The primary themes emerging from the review found that there are a diversity of approaches targeting varying audiences using varied methods. Some focus on convicted offenders (Babchishin et al 2015) while others focus on self-identified individuals at risk (Beier 2021), and some aim to educate parents, carers and professionals (e.g. Hudson 2018).

Gannoni et al (2023) suggest early intervention, reaching individuals before they engage in harmful behaviours or escalating existing behaviours is effective. Such approaches often adopt a harm reduction approach, acknowledging that individuals may struggle with problematic sexual thoughts but can learn to manage them and avoid offending, which necessarily requires a different policy approach. These initiatives provide a combination of education, therapy, self-help resources, and community support to address the complex factors contributing to CSAEM offending.

3 This is one type of identified online harm, there are others including misogynistic content, harassment, bullying, extreme violence and torture content among others.

Types of Initiatives

Gannoni et al. (2023) detail initiatives falling into several broad categories:

- **Helplines & Online Resources:** These offer anonymous, confidential support and information to individuals concerned about their own behaviour or that of others. Examples include Stop It Now! operating in various countries, "Otanvastuun" (I Take Responsibility) in Finland (see Save the Children Finland 2022), Troubled desire in Germany (Schuler et al 2021) and "Prevent It!" in Sweden
- **Therapeutic Programs:** These provide individual and group therapy, often using cognitive behaviour therapy (CBT) techniques to address problematic sexual thoughts and develop coping mechanisms. Examples include "Kein Täter Werden" (Don't Offend) in Germany (Beier et al. 2021), "ReDirection Self-Help Program" in Finland, and The "Aurora Project" in the UK.
- **Community Support Programs:** These aim to reintegrate convicted offenders into the community and reduce recidivism through support groups and practical guidance.
    - Examples include Circles of Support and Accountability (CoSA) operating in multiple countries and "WellStop" in New Zealand.
    - Peer support models like PartnerSPEAK who offer "effective support for this underserved client group including the reduction of shame and isolation." (Jones et al., 2023, p. 715)
- **Educational Campaigns:** These aim to raise public awareness, educate about risks, and promote preventative measures against CSAEM offending. Examples include "Keep it Real Online" in New Zealand and "The Eggplant" TV series, also in New Zealand.

Treatment Approaches

- **Cognitive Behaviour Therapy (CBT):** This approach is widely used to help individuals identify and challenge distorted thoughts and develop healthier coping mechanisms for managing their sexual urges. As the source states, the Otanvastuun program in Finland uses CBT to "challenge misbeliefs and sexual thoughts towards children".
- **Relapse Prevention:** Many programs incorporate relapse prevention strategies, helping individuals identify triggers for offending and develop strategies to avoid or manage these situations effectively.
- **Good Lives Model:** This model focuses on identifying an individual's needs and goals and developing prosocial ways to fulfill those needs, reducing the likelihood of resorting to harmful behaviours.
- **Restorative Justice:** This approach, used by programs like CoSA, focuses on repairing harm and fostering accountability through facilitated dialogues between offenders, victims, and community members.

Current research limitations on CSAEM prevention strategies effectiveness

Gannoni et al (2023) indicate further research is needed to assess the long-term impact of these programs and identify best practices for prevention. Due to the nascent research area, there are also methodological limitations found in their study:

- **Small sample sizes:** can limit the ability to generalise findings and detect statistically significant effects.
- **Short follow-up periods:** Longer follow-up periods (e.g., years) are needed to determine whether any positive treatment effects are maintained over time.
- **Lack of matched (non-treatment) comparison groups:** Many studies did not include a matched comparison group, which makes it difficult to determine if observed changes are due to the intervention or other factors. It is not always clear if reductions in dynamic risk factors are attributable to treatment or other time-related factors, such as the resolution of legal proceedings.
- **Reliance on retrospective assessments of impact:** Many studies rely on retrospective self-report data (e.g. asking participants to reflect on their behaviours before and after treatment), which may be biased by problems with recall and social desirability (e.g., participants may over-report positive behaviour changes).
- **Difficulty conducting randomised controlled trials:** The sources highlight challenges in conducting randomised controlled trials (RCTs) in sex offender treatment research. For example, it is often deemed unethical to deny eligible individuals treatment for the purpose of creating a control group.

In addition to these methodological limitations, Gannoni et al (2023) mention a number of broader limitations impacting research in this area:

- **Lack of research on CSAEM use among specific populations:** No evaluations were identified that examined the impact of prevention programs on CSAEM offending among specific populations.
- **Lack of research on undetected offenders and those at risk of offending:** Much of the research focuses on individuals who have been detected by authorities for CSA or CSAEM offenses. However, there is a lack of research on individuals who have offended but have not been detected, and those who are at risk of offending, but have not offended.
- **Lack of consensus on terminology:** There is inconsistency CSAEM terminology across studies and programs (e.g., child abuse material, child exploitation material, child pornography). This may lead to difficulty identifying all relevant initiatives in literature searches and comparing research findings across studies. This has been reported elsewhere (Teunissen & Napier 2022)
- **Lack of focus on CSAEM-specific risk factors:** Most of the studies included in the review focused on general risk and protective factors for sexual offending rather than examining factors specific to CSAEM offending. This is likely because less is known about the characteristics of individuals who engage in CSAEM offenses compared to contact sexual offenders.

3.2 Encryption

The Office of the eSafety Commissioner's transparency report details various efforts by tech companies to balance safety and user privacy, particularly in the detection and prevention of CSAEM on their platforms. While the focus is on the general techniques being used, there is limited detail on specific privacy-enhancing technologies. End-to-end encryption (E2EE) is emphasised as a key factor in safeguarding privacy – only the sender and recipient can access communications.

This environment creates challenges for detecting harmful content based on known technologies. Companies like Apple had explored additional approaches such as ML powered CSS but have abandoned these plans due to the impact on privacy. Meta platforms such as WhatsApp have adopted E2EE self/user-reporting mechanisms and limits on post sharing capabilities to detect and limit harmful content.

Fang and Qian (2021) propose a privacy-preserving machine learning framework that uses homomorphic encryption and federated learning to ensure that sensitive data remains confidential during processing (Fang & Qian, 2021). Their work illustrates how advanced encryption techniques can facilitate secure data analysis without compromising user privacy.

Fully homomorphic encryption (FHE) has opened new avenues for privacy preservation in cloud computing. Brännvall et al. discuss FHE's potential in enabling federated learning applications while complying with stringent regulatory frameworks (Brännvall et al., 2023). This aligns with Avgerinos et al's (2023) findings, who detail a scalable privacy-preserving framework that leverages various encryption techniques to facilitate secure data processing across borders. Such frameworks are crucial in ensuring compliance with regulations like the GDPR, which mandates strict data privacy measures.

3.3 Pop-Up Warning Messages

Mechanism: When users enter search terms commonly associated with CSAEM, pop-up warning messages appear about the legal consequences of accessing illegal content (Prichard et al. 2011; Prichard et al. 2021). This strategy seeks to deter potential offenders, particularly first-time users, by 'raising awareness' of CSAEM criminality and challenging 'cognitive distortions' that may normalise such behaviour.

Pritchard et al (2022) found that various methods have been employed to evaluate the effectiveness of online messages, including self-report surveys (Ullman, 2017; Zaikina-Montgomery, 2011), uncontrolled field experiments (Silic & Back, 2017), controlled laboratory experiments (Anderson et al., 2016), controlled field experiments (Junger et al., 2017), meta-analyses (Hancock et al., 2020), crowdsourced recruitment (Clayton et al., 2020), and randomized controlled experiments utilizing honeypots and naïve participants (Maimon et al., 2014; Prichard et al., 2021). Hunn et al. (2023) found that the technology can be used across sectors and leads to more collaboration across parties to develop messages. This approach is grounded in SCP theory (and psychological approaches that suggest disrupting thought patterns (justifications, rationalisation etc) that lead to certain harmful behaviours (Prichard et al. 2011, 2022).

Research reviewed in this area but outside of CSAEM-related content consistently shows that online warnings can influence user behaviour and suggest that internet users are generally responsive to warnings related to hazardous online activities such as cyber-attacks (Testa et al., 2017), malicious software use (Silic & Black, 2017; Anderson et al., 2016), malware exposure (Haddad et al., 2020), online piracy (Ullman & Silver, 2018), fake news (Clayton et al., 2020), phishing (Neupane et al., 2016), online gambling (Caillon et al., 2021; Gainsbury et al., 2015) and disclosing personal information (Carpenter et al., 2014).

Limitations: Research on pop-up warnings long-term effectiveness is limited. Further research is needed to assess long-term impacts and potential site displacement of this strategy (Prichard et al., 2021, 2022, 2024). Research into CSAEM-specific pop-up messages is nascent but used by some platforms such as Google. Though warnings may deter offenders, they lack ongoing support; incorporating referrals to relevant services could address this limitation (Baines, 2018; Prichard et al. 2022, 2024).

3.4 Platform design

The academic literature highlights the need for a holistic approach to platform design, particularly in ensuring user safety through responsible content moderation. There seems to be some alignment within different fields of research on having safety by design principles embedded from the outset of product development by the platforms children use. However, how this is designed varies. For example, Draper (2022) argues that pre-verified uploads on adult content platforms requiring content creators to obtain consent and verify the age of individuals before uploading material is necessary to include while Stardust et al (2024) suggest this comes with its own limitations around body types.

Scholars like Sanderson et al. (2022) argue that AI-driven safety mechanisms must be designed with clear ethical standards in mind. These include themes such as privacy, transparency, fairness, human oversight, and the promotion of human values (Fjeld et al., 2020). However, critics point out that current frameworks tend to focus too much on post-hoc explanations of AI outputs, often neglecting proactive design principles and the broader societal impacts of AI systems (Morley et al., 2020; Shneiderman, 2020).

The European Union's regulatory framework, particularly the Digital Services Act (DSA), sets a valuable precedent for mandating safety-by-design mechanisms in content moderation (Schneider & Rizoiu, 2023). In contrast, Australia's Online Safety Act lacks direct focus on platform design principles, though eSafety advocates for integrating safety into platforms from the outset (see Appendix D).

This proactive approach can prevent harmful content amplification through algorithmic safeguards. Murthy (2021) suggests platforms like YouTube can implement algorithmic changes to improve accountability while maintaining user privacy, aligning with Bleakley et al.'s (2024) argument for balancing regulation with privacy concerns in moderating CSAEM content.

Some platforms such as WhatsApp, an E2EE messaging service, implements several safety design features aimed to limit unwanted contact and potential abuse, including prohibiting users from searching for individuals they don't already know and requiring a user to possess another user's phone number to initiate contact.

In terms of children's privacy, platforms can implement specific design features that disable data-sharing capabilities and enhance privacy protections. Witzleb et al. (2020) note that platforms like Apple and Google have restricted app tracking and location sharing, while YouTube disables interactive features for content aimed at children. Platforms such as Facebook and TikTok have introduced safeguards like inbox filtering and parental reminders to limit interactions for minors. However, platforms like Google and Microsoft have faced criticism for privacy-intrusive default settings that require user intervention to activate protective measures. For instance, YouTube's content filter must be enabled by parents, in contrast to Facebook's default privacy settings that limit teens' sharing options and visibility to advertisers while Snapchat turns off location sharing by default for all users.

A user-centred approach to privacy, as detailed by Barth et al. (2021), involves making data handling practices more transparent through visual privacy ratings. Providing clear, accessible information about how data is used empowers users to make informed decisions, enhancing both trust and safety without infringing on privacy (Ganesh & Bright, 2020). This approach places importance of transparency in design, not just to meet regulatory demands, but to foster a digital environment that prioritises individual user well-being at its core.

However, policies like transparency and awareness can create the impression that just knowing about surveillance, like profiling or targeted ads, can help individuals avoid its impact. But this emphasis on awareness often limits stronger solutions, such as outright bans on certain practices (Padden & Öjehag-Pettersson 2024). For instance, the EU's regulation on political advertising doesn't ban micro-targeting but requires citizens to be aware when they are targeted, aiming to reduce manipulation of democratic discussions without restricting advertisers' freedom of expression (Padden & Öjehag-Pettersson 2024). Such solutions like industry codes of practice shift responsibility from individuals, yet often favour business interests, and awareness-based policies may give a false sense of control.

As such, privacy safeguard (or privacy-by-design) principles such as data minimisation with limits the collection, use and storage of personal data to what would be adequate, *necessary* or relevant to a specific purpose is detailed as a standalone principle in the EU GDPR, UK GDPR, and the *California Privacy Rights Act* (CPRA). In Australia, data minimisation is captured through a few principles but does not have its own standalone principle. The APP 3 refers to collection of personal information when it is "reasonably necessary" for an entity's function or activity and APP11 requires entities to take reasonable steps to de-identify or destroy personal information they no longer need. Privacy Act Review discussed data minimisation but did not introduce it as a principle citing sufficient protection under the current APPs. Such a (lack of) focus on minimising data collection and retention often repositions privacy as a technical issue rather than a political one. In doing so, it shifts the responsibility to individuals and organizations to protect privacy within economic structures that continue to rely heavily on data extraction.

3.5 Privacy literacy (education)

One significant gap is the lack of comprehensive studies that explore the interplay between online privacy literacy and users' perceptions of safety. While some studies have indicated that higher privacy literacy correlates with a greater sense of safety on social networking sites (SNSs) (Bartsch & Dienlin, 2016), there is insufficient empirical evidence examining how these dynamics play out in various demographic groups or across different online platforms.

Existing literature does not adequately address the potential for privacy literacy to mitigate safety concerns, particularly in contexts involving vulnerable populations, such as children or individuals experiencing domestic violence (Sabri et al., 2023). This gap suggests a need for research that investigates how enhancing privacy literacy can lead to improved safety outcomes in online environments.

3.6 User reporting

User Reports represent a critical method, enabling individuals to report abusive content without breaching encryption (Draper 2022). Platforms typically provide users with the ability to report content they believe violates community guidelines. According to Anderson and Gilbert (2022), user reporting remains the most effective method for detecting grooming and CSAEM, particularly in E2EE environments. This participatory approach not only empowers users but also helps platforms identify harmful content that may have evaded automated detection systems (Schoenebeck et al., 2023). However, its success is dependent on users being both aware of and willing to utilise reporting mechanisms, as some may be reluctant to report content due to fears of retaliation or disbelief that their reports will lead to action (Sumner et al., 2021). Additionally, many countries lack the necessary infrastructure, such as tipline access, to effectively process and respond to user reports (Draper 2022).

Further, large tech companies often hinder this process by making it difficult for users to contact moderators who often sift through large amounts of reports. Addressing user complaints incurs significant operational costs. Some have argued mandating platforms provide accessible and efficient mechanisms for users to report illegal or harmful material, ensuring timely removal and response.

Tech companies already implement similar processes for copyright violations, and the law should compel them to extend the same level of protection to vulnerable users, such as women and children. When encountering harassment or abuse, these users must have the ability to swiftly contact moderators who can secure evidence, remove harmful material, and block offenders seeking to exploit them. This would shift the balance of responsibility towards greater accountability and user safety within digital environments.

eSafety's reporting mechanism reveals several potential limitations in its approach. While the system is intended to empower users to report harmful content, its reliance on user-driven reporting shifts the burden of content moderation onto victims rather than addressing the systemic issues inherent in platform design. This approach places the responsibility on individuals – often vulnerable users such as children, women, and those targeted by abuse – to navigate what they may perceive to be complex reporting processes, which can be time-consuming and emotionally taxing.

4

SECTION 4:

FUTURE RESEARCH OPPORTUNITIES

Future research opportunities

As this review is limited in its scope (see Appendix B), a more comprehensive systematic review into the privacy implications for methods and strategies used to detect and minimise online harms would need to be undertaken.

A review that also discusses methodologies, ideological frameworks, research paradigms, and critically appraises relevant studies. However, there are several avenues for further research outlined in the existing literature reviewed in this study. These include but are not limited to:

Privacy as Safety

Instead of framing privacy as a hindrance to security, research should investigate how privacy functions as a safety mechanism in online environments and how protecting personal data contributes to safeguarding users, especially vulnerable groups like children. Some studies have explored differing algorithms and access to varied perspectives to respond to mis- and disinformation (e.g. Aïmeur et al 2019; Sharma et al 2019; Gupta et al 2022; Alatawi et al. 2021).

Long-term Impact of Prevention Initiatives

Further research is needed to assess the long-term efficacy of harm reduction strategies aimed at CSAEM prevention and violent content and behaviours online (Edwards et al, 2021). While early evidence suggests holistic approaches are necessary, studies are needed to evaluate these interventions' sustainability and impact across diverse populations (Gannoni et al. 2023).

Bias in AI Detection Tools

If industry is going to continue down the path of automation, which it does with the volumes of data content and interactions online has amassed, the potential for bias in AI-powered detection tools is a critical area for further investigation. These technologies risk replicating societal biases, disproportionately affecting marginalised communities while overlooking context in content moderation processes.

Understanding Socio-Cultural Factors

More research is needed on the socio-cultural and structural factors that contribute to the creation and distribution of CSAEM and other online harms. Further, there is no agreement on definitions of what constitute harm online as there is no shared moral definitions across the world. This complicates the terrain and object of research. More research is needed into the content that gets flagged as harmful in politically tense environments such as the expanding wars in Palestine, Lebanon, the Sahel region and in Ukraine. Some studies suggest a more holistic approach is needed to promote healthy sexual development in the digital age (e.g. Stardust et al. 2024). Proposals include that governments and policymakers redirect resources towards comprehensive, age-appropriate sex education that includes discussions about pornography, consent, and online safety (Gillett et al. 2022; Stardust et al., 2024; Marwick et al. 2024).

5

SECTION 5:

CONCLUSION

Conclusion

The research reviewed for this limited review across a varied field suggest more holistic approach are needed to combat harms. Many call for more transparency and accountability, a radical change in platform and algorithmic design, a consideration of socio-political contexts.

Discussion on safety, particularly for children, and risk (the probability and consequences of harm) to harms (the actual outcome, either objective or subjective) is not new (Livingston 2013). While cyberspace introduces new risks, harm has always been experienced offline (e.g. Livingstone 2013; boyd 2024) and responses should not cut off the the potential empowering capacity of online spaces for people, in particular young people and children (e.g. eSafety 2024b). The technocratic approach employed by policy reflects a broader societal obsession with risk mitigation, where 'prevention' becomes synonymous with control, and crime is depoliticised into a series of calculable risks to be managed rather than a phenomenon rooted in social structures. In situating SCP within the context of neoliberal risk management, we can critically interrogate how such crime prevention measures serve to reinforce existing power structures while diverting attention from the deep social inequalities that underlie criminalised behaviours.

Technological solutions for mitigating online harms expose significant shortcomings. These interventions often operate within a reductive binary framework of harm, which fails to capture the complexities of digital life. The reliance on technologies such as client-side scanning (CSS), machine learning (ML), and automated content removal reflects broader logics of securitisation and governmentality, where the governance of online spaces increasingly revolves around risk management and pre-emptive control. This shifts the focus from addressing the structural causes of harm to simply containing its most visible manifestations.

These technologies embody a carceral logic that extends digital surveillance and control into both private and public domains, reinforcing existing power structures (Shkabatur 2011; Raher 2024). The governance of online harms, intertwined with state and corporate surveillance mechanisms, seeks not only to detect and prevent deviant behaviour but also to reproduce and strengthen these structures. CSS, for instance, illustrates this invasive extension of control, where personal devices are monitored, collapsing the boundaries between public and private life. This mechanism reflects a panoptic model of governance, where individuals are constantly subjected to potential surveillance, reinforcing compliance through fear of observation (Raher 2024).

The deployment of machine learning (ML) to detect extremist content similarly reduces complex social phenomena to quantifiable data, reinforcing a technocratic logic that obscures the political, social, and economic contexts that give rise to radicalisation (Hernández et al 2022). The limitations of such approaches reveal an underlying tendency to individualise and depoliticise harm, neglecting the socio-economic and political drivers that foster harm to others. This aligns with neoliberal governance strategies that frame deviance as an issue of risk management rather than addressing the root causes of inequality and disenfranchisement.

Content removal and deplatforming strategies further illustrate this neoliberal risk logic, where the focus is on managing harmful actors by removing them from mainstream platforms, rather than engaging with the underlying ideologies and grievances that drive them. These strategies often displace harmful content to more obscure, less regulated spaces, reinforcing echo chambers where extremist views can proliferate. The individualisation of harm through such measures neglects the intersectional nature of online behaviours, where users may simultaneously be victims and perpetrators, particularly in cases of 'radicalisation' or self-generated harmful content.

The securitisation logics underpinning these technological solutions are emblematic of the broader expansion of state and corporate power into digital spaces (Manokha 2023; Whitehead & Collier 2023). By framing online harms as issues of national security, surveillance technologies are justified as necessary protections for public safety. However, these measures often normalise the erosion of privacy and the expansion of control into intimate spheres of life (e.g. Chambers 2024). The processes of governmentality embedded in these technologies not only manage populations but also perpetuate power imbalances, with the regulation of digital spaces increasingly shaped by market-driven imperatives that align corporate interests with state surveillance.

These technological approaches also operate within a binary framework of harm that simplifies the complexities of digital interactions. Harm is often framed as a clear dichotomy between "good" and "bad" actors, neglecting the fluidity and intersectionality of online experiences. This is particularly evident in cases of online radicalisation or self-generated child sexual abuse material, where individuals may be both victims and perpetrators. A more nuanced understanding of harm would account for the ways individuals navigate intersecting power relations online, shaped by socio-economic, racial, and gendered identities.

A holistic response to online harms requires moving beyond these technological fixes and addressing the socio-political conditions that give rise to harm. Approaches centred on education, community support, and social justice offer more sustainable, long-term solutions. Rather than relying solely on surveillance and control, policies must engage with the structural inequalities that contribute to online exploitation and radicalisation. This includes addressing economic disenfranchisement, political alienation, and social isolation, which are often the underlying drivers of harmful online behaviour.

Furthermore, privacy must be re-centred in these discussions, particularly as current technological solutions frequently sacrifice privacy in favour of enhanced surveillance. Privacy-enhancing technologies, such as end-to-end encryption, should be integrated into digital governance as a safeguard against the overreach of state and corporate power. The balance between safety and privacy should not be viewed as a zero-sum game but rather as a dynamic relationship requiring careful consideration of rights, agency, and structural change (Mann et al 2018).

Ultimately, technological solutions for online harms are constrained by the same power dynamics that shape the digital economy. The commodification of user data, driven by profit motives, often undermines the very protections these technologies claim to offer. To develop more equitable and effective approaches, it is necessary to move beyond the binary logics of harm and embrace strategies that focus on structural change, social justice, and the rights of marginalised communities. This requires rethinking the political economy of digital platforms, the role of state regulation in online spaces, and the social structures that sustain inequality both on and offline. By shifting towards a holistic approach that integrates prevention, education, and privacy protections, a more just and inclusive digital environment can be achieved.

APPENDICES

Appendix A:

Acronyms and abbreviations

AI	Artificial intelligence
AIC	Australian Institute of Criminology
APPs	Australian Privacy Principles
BOSE	Basic Online Safety Expectations
CBT	cognitive behaviour therapy
CoSA	Circles of Support and Accountability
CSAEM	Child Sexual Abuse or Exploitation Material
CSS	client-side scanning
DRW	Digital Rights Watch
DSA	EU Digital Services Act
E2EE	end-to-end encryption
EDR	European Digital Rights
eSafety	Office of the eSafety Commissioner
FHE	Fully homomorphic encryption
HRLC	Human Rights Law Centre
ICCL	Irish Council for Civil Liberties
ML	Machine Learning
NCMEC	National Center for Missing and Exploited Children
The Act	Online Safety Act 2021 [Australia]
P2P	peer-to-peer
RAS	Restricted Access System
RCTs	randomised controlled trials
SCP	situational crime prevention
STMP	Suspect Target Management Plan
UK OSA	UK Online Safety Act 2023

Appendix B:

Scope and methodology

This desktop review is limited in scope due to constraints on available time and access to certain data sources. While it draws on existing literature, policy documents, and relevant case studies, it does not encompass a comprehensive exploration of all possible perspectives or datasets related to safety and privacy policy as well as academic and grey literature. As such, this review should be viewed as an initial analysis, highlighting key themes in the current research and avenues for further research. Further in-depth investigation will be required to address areas that remain underexplored or for which data is not readily available.

In this limited desktop review, the focus is placed on examining sources relevant to the issue of Child Sexual Abuse Material (CSAEM) and to a lesser extent, online spaces and technocultures that impact people's mental health, particularly for young people. The review does not contend with exploring other harmful online content such as classified extremist material as this required a critical approach beyond the main literature that focuses on radicalisation – a research area that is ripe with uncertainty and political ambiguity (see Marwick et al 2022). The review draws from diverse but scattered evidence, including research from academic literature, law enforcement agencies, government agencies, studies with offenders or those at risk of offending, and accounts from young people and survivors. Key sources were selected based on their relevance to current policy discussions, particularly under the framework of the Online Safety Act 2021, which governs online safety and online privacy, in relation to other legislation, in Australia. The review employs thematic analysis to assess which areas of policymaking, particularly those that balance safety and privacy, have been well-researched and where gaps remain. Consideration was also given to the effectiveness of various policy interventions, such as encryption and age assurance, as well as those that enhance safety without compromising privacy, like design choices and limits on data collection. The focus was on synthesising evidence that could inform avenues for further research into more effective and balanced policymaking.

Appendix C:

Limitations

This desktop review has several limitations that impact the scope and depth of its findings. First, due to the rapid nature of the review, time constraints and limited resources restrict its ability to provide an exhaustive analysis of all potential sources of evidence on harmful online content and policy considerations. The selection of sources was primarily focused on readily accessible academic research, policy documents, and reports, which may introduce selection bias – the inadvertent prioritisation of well-known or easily available sources – potentially overlooking key information or minority perspectives (see Clark et al. 2021). The review also faced challenges related to its multidisciplinary nature. The researcher is versed in the field of critical media studies but less so in other disciplines such as psychology, computer science, and law. Coupled with inconsistent terminology across different disciplines may have led to difficulties in synthesising findings and the potential misinterpretation of key concepts, complicating the overall analysis and limiting the potential to observe bias in literature may further influence findings, as original studies themselves may have presented incomplete or skewed views.

The review's geographical and English language focus also poses limitations. While it draws heavily on Australian legislation, it offers less in-depth analysis of international frameworks, limiting the review's ability to provide a comprehensive global comparison. Moreover, the review does not examine the technological nuances in detail which may leave out important technical considerations. Additionally, the rapid review method may have resulted in missing context or practical insights, and there was no time for meta-analysis or systematic comparisons across studies, limiting the depth of the findings. As the policy and technological landscape is rapidly evolving, some of the reviewed policies and technologies may quickly become outdated, reducing the long-term relevance of this analysis.

The review's gaps in interdisciplinary insights further constrain the scope of the findings, particularly as the topic spans multiple fields. The limited capacity for critical appraisal also meant that the researcher's ability to thoroughly assess the methodological rigour of included studies was restricted. Lastly, complex topics may not be easily distilled in such a rapid review, leading to oversimplified conclusions or missing nuances that are crucial to a deeper understanding (Clark et al. 2024). These limitations stress the need for more comprehensive, long-term research to address the gaps identified in this rapid limited desktop review.

Appendix D:

Policy landscape

Policies referred to in the report:

- Copyright Act 1968
- Criminal Code Amendment (Sharing of Abhorrent Violent Material) Act 2019
- Cyber Security Strategy
- Digital Services Act [EU]
- Enhancing Online Safety for Children's Safety Act 2015
- Misinformation and Disinformation Bill
- Online Safety Act 2021 [Australia]
- Online Safety Act 2023 [UK]
- Privacy Act 1988

The regulatory landscape of the internet in Western countries has increasingly shifted towards a more interventionist approach, characterised by increased scrutiny of tech giants, the implementation of frameworks aimed at safeguarding user rights and territorialising or governing online space (e.g. Belli & Venturini 2016; Davies 2014). This shift is largely a response to growing public concern over privacy, data security, and the monopolistic practices of major digital platforms, which have prompted calls for more stringent regulations to protect people online and ensure accountability (e.g. Namara et al. 2020). In Australia, the legislative landscape concerning online privacy and safety draws upon various regulatory frameworks, including the Privacy Act 1988[4] and the Online Safety Act[5] (discussed in Mann et al 2018). The Privacy Act establishes guidelines for the handling of personal information, while the Act empowers eSafety to combat harmful online content, such as cyberbullying, classified extremist content, violent content and CSAEM (Pothong, 2019; Budak & Rajh, 2018). These frameworks reflect a broader trend towards increased government attention to platform governance, which some claim is moving towards more structured oversight of digital platforms (for limitations see also Smith et al. 2024).

Currently, the Australian policy landscape is marked by a tension between privacy and safety, often framed as a binary opposition (see Mann et al 2018; 2020). The framing of privacy as a safety concern has been utilised to justify extensive surveillance measures, often at the expense of individual rights. This phenomenon is evident in the rhetoric surrounding new laws and policies that prioritise security over privacy (e.g. metadata retention), creating a narrative that positions these two concepts as a zero-sum game (e.g. Mann et al. 2018). The implications of this framing normalises surveillance practices. The Act exemplifies this tension, as it grants eSafety extensive powers to remove harmful content while potentially infringing on individual privacy rights due to technological detection methods arising out of these policies, some of which are detailed in this report.

4 This Act has undergone several amendments to enhance privacy protections, particularly in response to the growing concerns about data breaches and misuse of personal information in the digital age (see McDuie-Ra 2023). The introduction of the ***Australian Privacy Principles (APPs)*** under this Act emphasises the need for transparency, accountability, and the protection of individuals' personal data (Wozniak et al., 2020). However, the effectiveness of these regulations is often debated, especially in light of rapid technological advancements and the increasing prevalence of surveillance practices justified under the guise of safety.

5 Note: The ***Cyber Security Strategy 2020*** outlines a roadmap for responding to cyber threats, particularly in critical infrastructure sectors, while the ***Copyright Act 1968*** and the ***Criminal Code Amendment (Sharing of Abhorrent Violent Material) Act 2019*** address intellectual property rights and harmful content moderation, respectively (Moulds et al., 2022).

This Act establishes 'Basic Online Safety Expectations' (BOSE) and a series of industry codes for service providers, which include obligations to protect users from various forms of online harm (Chiner et al., 2019). However, critics argue that this approach oversimplifies the relationship between privacy and safety, often prioritising the removal of harmful content at the expense of user privacy (Tuikong, 2022) that can lead to various other ways in which that content proliferates (e.g. Casilli et al 2013; Alderton 2022). The critiques of this approach highlight the dangers of framing privacy and safety as mutually exclusive. For instance, the reliance on surveillance technologies to enforce safety measures can lead to invasive practices that undermine user privacy and uphold surveillance logics. When users do not believe that existing regulations effectively protect their privacy, they often resort to self-designed strategies for data protection, indicating a lack of trust in regulatory frameworks (Li et al., 2020) as well as in the platforms themselves. Further, the emphasis on rapid content removal often places disproportionate responsibility on individuals to report harms, such as cyberbullying, while allowing platforms to maintain self-regulatory practices that fail to address systemic issues within their business models (You & Wang, 2019) or the offline practices that constitute these behaviours. The framing of safety as a justification for surveillance practices often leads to a simplification of the relationship between individual rights and the structural power of global platforms (Radovnikovic et al., 2020; Fisk, 2014).

Online Safety Act 2021 (The Act)

The Act in Australia provides the main comprehensive legal framework for addressing online harm, including CSAEM and other classified illegal or violent or abusive material. The Act builds upon previous frameworks established by the Enhancing Online Safety for Children's Safety Act 2015, which laid the groundwork for coordinated online safety efforts across various sectors, including government, industry, and civil society (Jang, 2023). Yet it marks a shift in Australia's regulatory landscape, extending protections against online harm beyond children to include adults through its adult cyber abuse scheme. Through the Image-Based Abuse Scheme, the Act empowers eSafety to act in cases of non-consensual sharing of intimate content, addressing violations of personal privacy in digital contexts. The regulation of abhorrent violent content mandates ISPs to block access to classified dangerous material, while the Basic Online Safety Expectations (BOSE) establish obligations for online platforms to actively mitigate harmful behaviour, reinforced by civil penalties for non-compliance, thereby strengthening the regulatory oversight of digital service providers. Such an approach is hailed co-regulatory between government and industry, yet relies on industry self-regulation "in all but name" (Farthing 2022, p.1) as industry leads in developing the code of practice, resulting in too much power given to platforms and inadequate safety standards (Reset Australia 2024). The reliance on industry-developed codes for online safety can be interpreted as a market-driven approach, wherein the responsibility for user protection is outsourced to private corporations that may prioritise profit over substantive safety measures (Schmoelz, 2023). This dynamic reducing complex social issues related to abuse and exploitation to mere algorithmic solutions, which often fail to address the root causes of hateful digital cultures (Pfister & Yang, 2018) and present technological solutions to societal issues. While the Act imposes substantial responsibilities on online service providers to ensure user safety, it simultaneously reinforces the commodification of user data by mandating transparency and accountability without addressing the underlying systemic issues inherent in platform capitalism (Gorwa, 2019). This paradox highlights the tension between regulation and corporate self-regulation, potentially entrenching the power of major tech firms under the guise of user safety (Maguire & Murphy, 2013)

Under the Act, eSafety has expanded powers to investigate and enforce compliance with new industry codes requiring digital platforms to take stronger measures against harmful content, including CSAEM, ensuring rapid removal of such material from online platforms. Key provisions of the Act related to CSAEM include powers to issue takedown notices for CSAEM, with platforms required to comply within 24 hours or face penalties. The Act mandates online platforms to report CSAEM, facilitating criminal investigations, and criminalises the non-consensual sharing of intimate images, including self-generated exploitation material. The Act also promotes international cooperation for the rapid removal of such content hosted on foreign servers. The onus on reporting and taking down content is placed upon platforms who use a range of technological detection tools.

Content restrictions under the Act

The Act focuses on content, illegal and restricted online content includes material showing or promoting child sexual abuse, terrorism, or extreme violence. Under the Act, eSafety can issue removal notices to online platforms to eliminate this content and restrict access to material unsuitable for those under 18. The Act classifies harmful material as either "Class 1" or "Class 2" content[6], with stricter measures for Class 1, including child sexual abuse and criminal activity. For internationally based content (hosted outside of Australia), eSafety works with global networks like INHOPE for removal of CSAEM and collaborates with Australian law enforcement for domestic cases. Additionally, the Act includes powers to block material related to "abhorrent violent conduct", such as terrorism. The Restricted Access System (RAS) dictates that age-inappropriate content is restricted to adults.

Informing policy

Public discourse

Moral panics globally, are a way to gage hegemonic public discourse in disruptions to the status quo (Cohen 1972), often around children, childhood, and youth (Cohen 1972; Drotner 1999; Krinsky 2008; Nicholas & O'Malley 2013). In relation to new technologies, moral panics espoused and sometimes articulated through media coverage and policy decisions produce commonsense about childhood and the anxieties of children's access to the internet and various platforms (Facer 2012). Early legislation in the US were fuelled by moral panics (Marwick et al. 2024). In the 1990s, panics over online content and conduct involved cyberporn and piracy by the mid-2000s an added layer of concerns over cyberbullying and online predators "involved overblown threats fuelled by media hysteria, misinterpreted research, and inaccurate statistics." (Marwick et al., 2024, p. 6).

In Australia, media coverage follows the same trends – the industry frames children's digital technology use negatively (Duffy et al. 2024; Jeffery 2018; Facer 2012). Such media framings are generally focus on discourses of CSAEM, screen time, concerns over parenting. In a study of 604 newspaper articles from 2002-2022, Duffy et al. (2024) found close to half (41.6%) of articles concerning children's use of digital technologies pertained to CSAEM, where media focus on this topic has increased overtime, particularly since 2019 where articles discussed the increase in CSAEM. Such panics, position parental and governmental as having lost 'control' and can oriented "political and public concerns away from 'real world' sexual abuse of children" (Jewkes & Wykes 2012, p 9355). Such societal attitudes that focus on the content or material available on the internet expound common myths that sexual attraction to kids is a problem of late-modern societies, carried out by strangers who are individually pathologised (Jewkes & Wykes 2012).

Such framings of internet safety can inform rushed legislation (e.g. the recent social media ban for children in Australia), policy decisions (e.g. the Misinformation and Disinformation Bill excluding mainstream media who are significant sources of misinformation), increased surveillance of young people (Fisk 2016) or through predictive analytics to detect potential offenders. As such, these problems can be framed as a technological one, requiring technological solutions rather than the holistic approaches research and civil society suggest (e.g. Dezuanni et al 2024; ARC centre for the Digital Child 2024. The Act's framing as a progressive initiative must be critically evaluated against the backdrop of neoliberal governance and the commodification of safety in the digital economy (Pfister & Yang, 2018; Viale et al., 2017).

6 Class 1 material, which has been or is likely to be refused classification under the National Classification Code. This includes child sexual exploitation material, pro-terrorist material, and material that promotes or incites crime; Class 2 material, which has been or is likely to be classified X18+ or R18+ under the National Classification Code. This includes non-violent sexual activity, or anything that is 'unsuitable for a minor to see.'

Field of research informing policy

Policy-making in this field on draws on a range of evidence sources, including a variety of policy submissions when drafting bills, a range of government research (e.g. eSafety), academic research, reports from local and international organisations. **Law enforcement data** provides empirical information on patterns of offending, offender characteristics, and the operational outcomes of regulatory frameworks (Gurriell, 2021). This data is instrumental in assessing the effectiveness of current enforcement strategies. Additionally, **research with offenders** contributes to understanding the behaviours and motivations behind the production and consumption of CSAEM, which can inform the design of targeted prevention and intervention strategies (Salter, 2023).

Engaging with **survivor testimonies**, particularly from young people affected by online exploitation, offers critical perspectives on the psychological and social impacts of harmful content. This information can guide the development of more responsive and informed policy measures (Bleakley et al., 2023). Similarly, **mental health research** provides evidence on the broader psychological effects of exposure to CSAEM and extremist content, underscoring the need for targeted mental health interventions (Gewirtz-Meydan, 2023). Moreover, **public health data** helps to contextualise the societal-level implications of online harm, contributing to a more comprehensive understanding of the public health burden posed by CSAEM and related content (Koebe, 2024).

In terms of well-researched areas, the effectiveness of existing laws and regulations regarding CSAEM moderation has been studied yet the field is still nascent considering continuing changes in technological tools, interfaces, and migrating communities such as AI-generated content. Research has shown that while some measures, such as mandatory reporting and content moderation, have improved detection rates, they often fall short in addressing the root causes of online exploitation (Bleakley et al., 2023). Additionally, the psychological toll on law enforcement personnel involved in CSAEM investigations has been documented (Gewirtz-Meydan, 2023).

Conversely, several areas remain underexplored. For instance, the long-term mental health impacts of exposure to CSAEM and extremist content on young people require further investigation. Additionally, the effectiveness of educational programs aimed at preventing online exploitation and promoting digital literacy among young users has been underexamined (Gannoni et al 2023). Furthermore, the intersection of privacy rights and safety measures in the context of CSAEM regulation remains a complex and underexamined issue, necessitating a more comprehensive approach that considers both individual rights and collective safety (Salter, 2023).

Gaps in policy

Australia's Online Safety Act suffers from several gaps, particularly in its regulatory framework (see HRLC 2024; Reset Australia 2024). A recent comparative analysis of the enforcement mechanisms within the UK Online Safety Act 2023 (UK OSA), the EU's DSA and Australia's proposed Basic Online Safety Expectations (BOSE) (Reset Australia 2024).

One of the main issues is the **industry-led approach to drafting codes of practice**. In the current model, platform companies play a central role in shaping the codes that govern their own operations (HRLC 2024; Reset Australia 2024). This industry-driven process often prioritises business interests over public safety, resulting in weaker safety standards. In contrast, in the UK, the regulatory body Ofcom takes charge of drafting codes, ensuring a more balanced approach that better reflects public safety concerns rather than corporate priorities. Public polling shows that 73% of Australians support a shift towards regulator-led code development, further highlighting dissatisfaction with the current industry-centric process.

In terms of **transparency and accountability**, Australia's Online Safety Act lags behind international frameworks. The current Act relies heavily on platforms' self-reporting of their safety practices, which allows companies to control the narrative and limit external scrutiny. By comparison, the EU's Digital Services Act (DSA) mandates platforms to conduct independent audits, provide prescriptive transparency reports, and share data with researchers to assess systemic risks. These measures are designed to promote

accountability and ensure that platforms are complying with safety standards, something the Act fails to address adequately.

Another significant gap is in **enforcement powers**. eSafety has limited authority to enforce compliance, with penalties capped at $610,500 AUD, even for major platforms like X (formerly Twitter). This fine is a fraction of what global platforms can easily absorb, making it an ineffective deterrent. In contrast, the UK OSA allows regulators to impose fines of up to 10% of a company's global turnover, creating a much stronger incentive for compliance. Moreover, the Australian Act lacks provisions allowing the regulator to compel platforms to make necessary safety changes or, in extreme cases, restrict access to services – powers that are essential for ensuring platforms meet their obligations.

The Australian Online Safety Act predominantly focuses on content moderation, without sufficiently addressing the systemic risks posed by platform design and operation (Martinez-Martin & Kreitmair, 2018; Reset Australia 2024). This content-centric approach overlooks the role that platform architectures play in facilitating online harm (HRLC 2024). In contrast, the UK OSA takes a more "**hybrid approach**", addressing both **system design** and **content moderation**. UK OSA introduces a **duty of care** for platforms, requiring them to proactively identify and mitigate risks associated with their design choices, algorithms, and features. The focus on "**safety by design**" mandates that platforms incorporate safety considerations from the outset, particularly in services targeting children. The UK OSA also requires regular **risk assessments** and empowers the regulator, **Ofcom**, to enforce corrective actions where necessary (Reset Australia 2024). This systemic approach recognises that online harms are not solely the result of content but are embedded in platform infrastructures and their economic models. However, such an approach can lead introduction of **generalised surveillance and monitoring posing privacy risks and limits to speech** where service providers could be incentivised to over-censor (see Anderson & Gilbert 2022).

Appendix E:

The Essential Report – Digital Rights Watch

Date:	09/10/24
Prepared By:	Essential Research
Data Supplied by:	qualtrics

The Research Society

Our researchers are members of The Research Society.

About this poll

Key Insights

The research conducted for Digital Rights Watch by Essential identifies some important nuance around the current approaches to improving online safety.

- The starting point for most people is that the web is a safe space, but we do hold concern for both older people and children.
- We are uncomfortable with the amount of personal information being collected by private companies, while holding mix views of government holding our data.
- Critically, we don't believe that we should have to trade off our privacy for a safer internet.
- This is born out with strong support for privacy reforms, which is even more supported than banning children from social media, although this too has high levels of support.
- We are more sceptical about the idea of using AI to police web safety, unconvinced that these techno-centric approaches will create safer online environments.

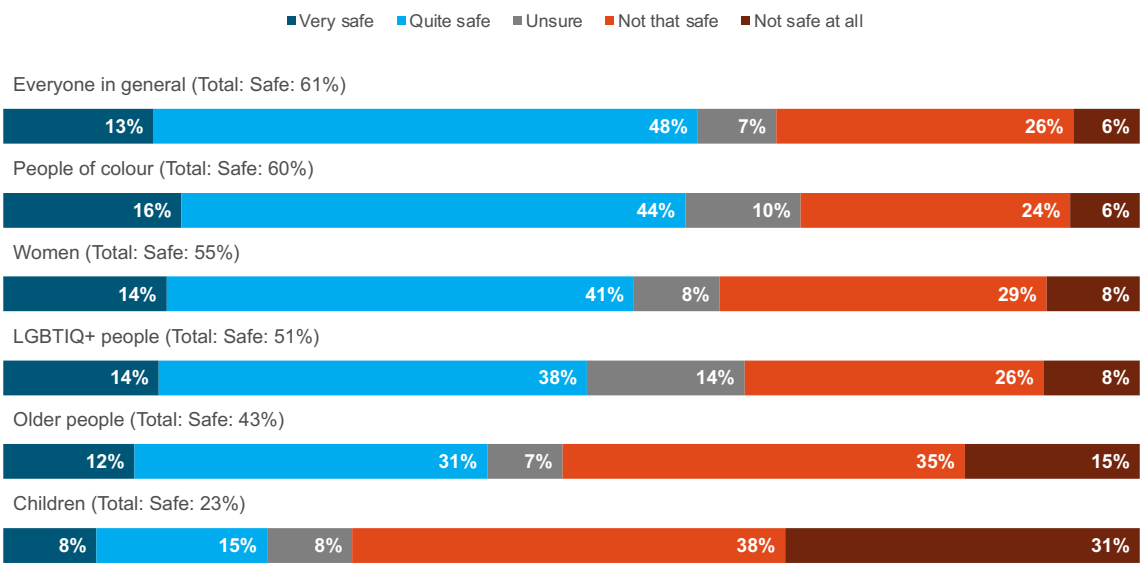
The research shows important nuances between different sections of the public:

- Younger people are more likely to be comfortable with the collection of online information
- Mid-age respondents re ore likely to support a tarde-off of privacy for safety
- Older respondents are more likely to support government regulation, but less likely to embrace automated solutions

This research should reinforce the need to carefully think through approaches to improving online safety and ensuring that any form of age or identify verification includes appropriate safeguards and protections.

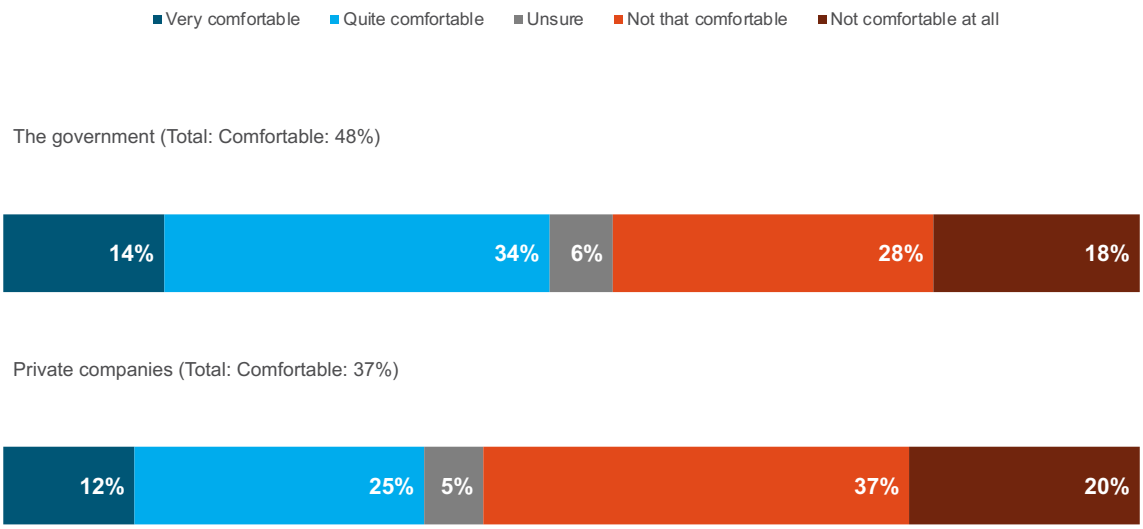
The web is safe for most people – but we are concerned for the young and the old

Q *How safe do you think the online web is for the following groups?*



We are not comfortable with private companies collecting our personal information, while we have mixed feelings about the government

Q *How comfortable are you with the collection and use of online information (e.g. web usage, personal information) by the following?*



Government

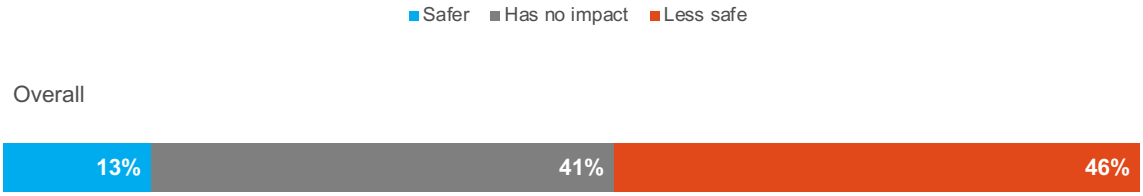
		Gender		Age		
Column %	**TOTAL**	**Male**	**Female**	**18-34**	**35-54**	**55+**
Very comfortable	**14%**	18%	11%	19%	12%	12%
Quite comfortable	**34%**	36%	32%	36%	34%	32%
Not that comfortable	**28%**	23%	33%	25%	29%	30%
Not comfortable at all	**18%**	18%	18%	13%	17%	23%
Unsure	**6%**	5%	6%	7%	7%	3%
Very comfortable + Quite comfortable	**48%**	54%	43%	55%	47%	43%
Not that comfortable + Not comfortable at all	**46%**	41%	51%	38%	46%	53%

Private Companies

		Gender		Age		
Column %	**TOTAL**	**Male**	**Female**	**18-34**	**35-54**	**55+**
Very comfortable	**12%**	14%	9%	20%	13%	3%
Quite comfortable	**25%**	26%	25%	34%	26%	18%
Not that comfortable	**37%**	35%	40%	30%	36%	44%
Not comfortable at all	**20%**	20%	21%	10%	18%	30%
Unsure	**5%**	5%	5%	5%	6%	4%
Very comfortable + Quite comfortable	**37%**	40%	34%	54%	39%	21%
Not that comfortable + Not comfortable at all	**58%**	55%	61%	40%	54%	75%

Data collection is seen as a safety issue rather than a safety solution

Q *Do you think that the collection and use of online information by the government and private companies makes the online web safer, less safe or has no impact?*

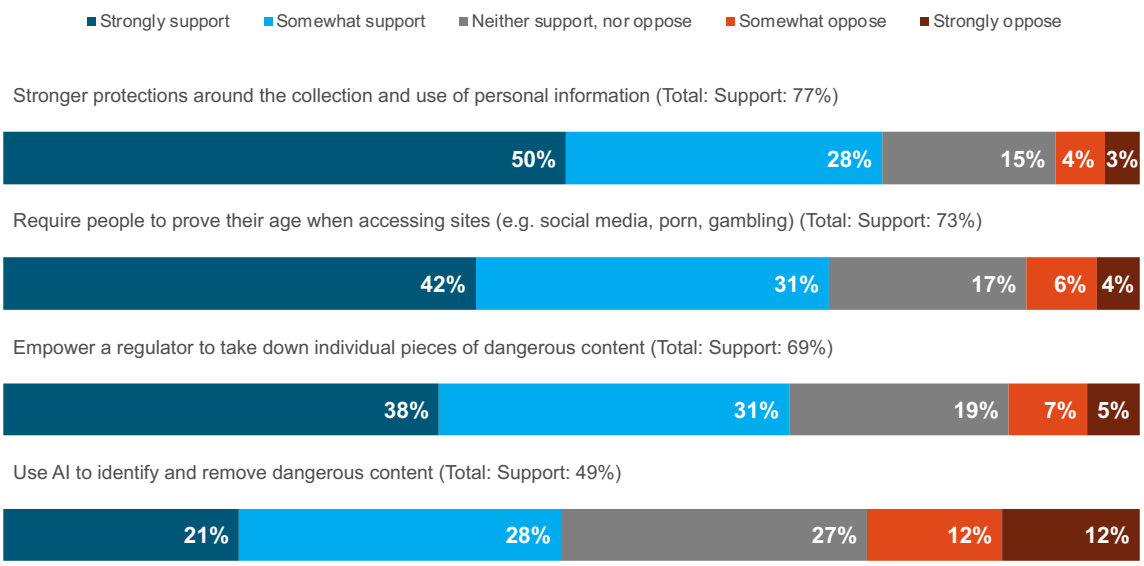


Q *Do you think that the collection and use of online information by the government and private companies makes the online web safer, less safe or has no impact? by National Banner*

		Gender		Age		
Column %	**TOTAL**	**Male**	**Female**	**18-34**	**35-54**	**55+**
Safer	**13%**	15%	11%	22%	14%	5%
Has no impact	**41%**	41%	40%	40%	46%	36%
Less safe	**46%**	43%	49%	39%	40%	58%

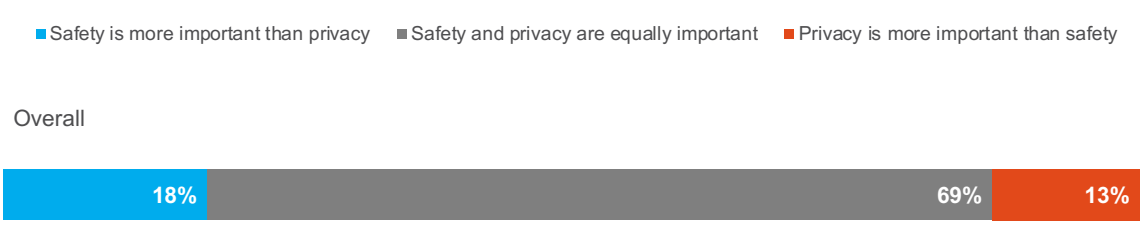
Privacy reform is key to broader regulation, with less enthusiasm for automated approaches

Q *To what extent would you support or oppose the following policies to make the online web safer?*



Most people don't want to trade off their privacy for safety

Q *Thinking about privacy and safety online, which of the following is closest to your view?*



		Gender		Age		
Column %	**TOTAL**	**Male**	**Female**	**18-34**	**35-54**	**55+**
Safety is more important than privacy	**18%**	20%	16%	24%	16%	15%
Safety and privacy are equally important	**69%**	62%	75%	56%	70%	78%
Privacy is more important than safety	**13%**	17%	9%	20%	13%	7%

Appendix:
Methodology, margin of error and professional standards

This report summarises the results of a fortnightly omnibus conducted by Essential Research with data provided by Qualtrics. The survey was conducted online from 2nd to 6th October 2024 and is based on 1,139 participants.

The weighting efficiency applied to the results at a national level is 80%, which gives an effective sample size of 909. The maximal margin of error at this effective sample size is ±3.3% (95% confidence level).

Each fortnight, the team at Essential Media Communications discusses issues that are topical and a series of questions are devised to put to the Australian public. Some questions are repeated regularly (such as political preference and leadership approval), while others are unique to each week and reflect current media and social issues.

Full text for standard voting and regular political preferences can be found in the link above. No questions were asked prior to these questions which have material influence on results.

All Essential Research staff hold Research Society membership and are bound by professional codes of behaviour. This research is compliant with the Australian Polling Council Quality Mark standards.

REFERENCES

References

Abbas, M. S. (2021). (In) Securitisation of Everyday Spaces: State of Exception, Spaces of Terror. *Terror and the Dynamism of Islamophobia in 21st Century Britain: The Concentrationary Gothic*, 301-377.

Abelson, H., Anderson, R. J., Bellovin, S. M., Benaloh, J., Blaze, M., Callas, J., Diffie, W., Landau, S., Neumann, P. G., Rivest, R. L., Schiller, J. I., Schneier, B., Teague, V., & Troncoso, C. (2024). Bugs in our pockets: the risks of client-side scanning. *Journal of Cybersecurity*, *10*(1). https://doi.org/10.1093/cybsec/tyad020

Abelson, H., Anderson, R., Bellovin, S.M., Benaloh, J., Blaze, M., Callas, J., Diffie, W., Landau, S., Neumann, P.G., Rivest, R.L. and Schiller, J.I. (2024). Bugs in our pockets: The risks of client-side scanning. *Journal of Cybersecurity*, *10*(1),

Academia, C. (2023, July 4). *CSA Academia Open Letter*. Google Docs. https://docs.google.com/document/d/13Aeex72MtFBjKhExRTooVMWN9TC-pbH-5LEaAbMF91Y/edit?tab=t.0

Aïmeur, E., Amri, S., & Brassard, G. (2023). Fake news, Disinformation and Misinformation in Social media: a Review. *Social Network Analysis and Mining*, *13*(1). https://doi.org/10.1007/s13278-023-01028-5

Alatawi, F., Cheng, L., Tahir, A., Karami, M., Jiang, B., Black, T., & Liu, H. (2021). A survey on echo chambers on social media: Description, detection and mitigation. arXiv preprint arXiv:2112.05084.

Albrecht, M., & Haddadi, H. (2023). *Open Letter from Security and Privacy Researchers in relation to the Online Safety Bill*. https://haddadi.github.io/UKOSBOpenletter.pdf

Alderton, Z. (2022). *Preventing Harmful Behaviour in Online Communities: Censorship and Interventions*. Routledge.

Ananny, M., & Gillespie, T. (2016). Public platforms: Beyond the cycle of shocks and exceptions. IPP2016 The Platform Society.

Anderson, B. B., Vance, A., Kirwan, C. B., Jenkins, J. L., & Eargle, D. (2016). From warning to wallpaper: Why the brain habituates to security warnings and what can be done about it. Journal of Management Information Systems, 33(3), 713-743.

Anderson, R., & Gilbert, S. (2022). *The Online Safety Bill | Policy Brief*. The Bennet Insititue for Public Policy Cambridge, University of Cambridge. https://www.bennettinstitute.cam.ac.uk/wp-content/uploads/2022/09/Policy-Brief-Online-Safety-Bill.pdf#page=5.57

Arango, A., Pérez, J., & Poblete, B. (2019, July). Hate speech detection is not as easy as you may think: A closer look at model validation. In Proceedings of the 42nd international acm sigir conference on research and development in information retrieval (pp. 45-54).

Are, C. (2023). Flagging as a silencing tool: Exploring the relationship between de-platforming of sex and online abuse on Instagram and TikTok. *new media & society*, 14614448241228544.

Are, C. (2024). 'Dysfunctional'appeals and failures of algorithmic justice in Instagram and TikTok content moderation. *Information, Communication & Society*, 1-18.

Ashe, S. D., Busher, J., Macklin, G., & Winter, A. (Eds.). (2020). Researching the far right: Theory, method and practice. Routledge. (n.d.).

Australian Centre To Counter Child Exploitation [ACCCE]. (2020, February). *ACCCE Research Report | ACCCE*. Www.accce.gov.au. https://www.accce.gov.au/resources/research-and-statistics/understanding-community-research

Australian Institute of Criminology. (2012). *National Crime Prevention Framework*. Special reports. Canberra: Australian Institute of Criminology. https://doi.org/10.52922/sr100614

Avgerinos, N., d'Antonio, S., Kamara, I., Kotselidis, C., Lazarou, I., Mannarino, T., Meditskos, G., Papachristopoulou, K., Papoutsis, A., Roccetti, P. and Zuber, M., (2023, July) A Practical and Scalable Privacy-preserving Framework. In *2023 IEEE International Conference on Cyber Security and Resilience (CSR)* (pp. 598-603). IEEE.

Babchishin, K. M., Hanson, R. K., & VanZuylen, H. (2015). Online child pornography offenders are different: A meta-analysis of the characteristics of online and offline sex offenders against children. Archives of sexual behavior, 44, 45-66.

Barth, S., Ionita, D., & Hartel, P. (2022). Understanding online privacy—a systematic review of privacy visualizations and privacy by design guidelines. *ACM Computing Surveys (CSUR)*, *55*(3), 1-37.

Bartsch, M. and Dienlin, T. (2016). Control your facebook: an analysis of online privacy literacy. Computers in Human Behavior, 56, 147-154. https://doi.org/10.1016/j.chb.2015.11.022

Belli, L. and Venturini, J. (2016). Private ordering and the rise of terms of service as cyber-regulation. *Internet Policy Review*, 5(4). https://doi.org/10.14763/2016.4.441

Begoli, E., Bhattacharya, T., & Kusnezov, D. (2019). The need for uncertainty quantification in machine-assisted medical decision making. *Nature Machine Intelligence*, *1*(1), 20-23.

Beier, K. M. (2021). Pedophilia, hebephilia and sexual offending against children. Pedophilia, hebephilia, and sexual offending against children. The Berlin Dissexuality Therapy (BEDIT). Springer.

Benbow, D. I. (2024). 'Don't panic, don't panic': an analysis of a purported pro-eating disorder website/online content moral panic and legal and policy responses. *Information & Communications Technology Law*, 1–23. https://doi.org/10.1080/13600834.2024.2404283

Benjamin, R. (2019). *Race after technology: Abolitionist tools for the new Jim code*. John Wiley & Sons.

Bhattacharjee, B. (2024). Facial Recognition Technology Balancing Ethical Considerations and Privacy Rights. *Available at SSRN 4885585*.

Bleakley, P., Martellozzo, E., Spence, R., & DeMarco, J. (2024). Moderating online child sexual abuse material (CSAM): Does self-regulation work, or is greater state regulation needed?. *European Journal of Criminology* , *21* (2), 231-250.

Bloul, S., Sammak, T., & Hussein, S. (2021, November 3-5). *After Christchurch: Rupture or continuity?* Paper presented at the AVERT International Research Symposium (online conference), Deakin University.

Bossio, D., & Barnet, B. (2023). The News Media Bargaining Code: Impacts on Australian journalism one year on. *Policy & Internet*, *15*(4). https://doi.org/10.1002/poi3.361

Bossio, D., Carson, A., & Meese, J. (2024). A different playbook for the same outcome? Examining Google's and Meta's strategic responses to Australia's News Media Bargaining Code. *New Media & Society*, 14614448241232296.

boyd, d. (2024, October 8). *Risks vs. Harms: Youth & Social Media*. Substack.com; Data: Made Not Found (by danah). https://zephoria.substack.com/p/risks-vs-harms-youth-and-social-media

Brännvall, R., Linge, H., & Östman, J. (2023). Can the use of privacy enhancing technologies enable federated learning for health data applications in a Swedish regulatory context?. *Swedish Artificial Intelligence Society*, 58-67.

Brown, O., Smith, L. G. E., Davidson, B. I., Racek, D., & Joinson, A. (2023). Online risk signals of offline terrorist offending.. https://doi.org/10.31234/osf.io/hej3r

Brown, R. (2023). Prevalence of viewing online child sexual abuse material among Australian adults. *Trends and Issues in Crime and Criminal Justice*, (682), 1-17.

Buolamwini, J., & Gebru, T. (2018, January). Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on fairness, accountability and transparency* (pp. 77-91). PMLR.

Burgess, J., Albury, K., McCosker, A., & Wilken, R. (2022). *Everyday data cultures*. John Wiley & Sons.

Bursztein, E., Clarke, E., DeLaune, M., Elifff, D.M., Hsu, N., Olson, L., Shehan, J., Thakur, M., Thomas, K. & Bright, T. (2019). Rethinking the detection of child sexual abuse imagery on the internet. In *The world wide web conference* (pp. 2601-2607).

Caillon, J., Grall-Bronnec, M., Saillard, A., Leboucher, J., Péré, M., & Challet-Bouju, G. (2021). Impact of warning pop-up messages on the gambling behavior, craving, and Cognitions of online gamblers: A randomized controlled trial. *Frontiers in Psychiatry*, *12*, 711431.

Canadian Council on Social Development, 2020. Situational Prevention. https://www.ccsd.ca/resources/CrimePrevention/sit.htm

Caillon, J., Grall-Bronnec, M., Saillard, A., Leboucher, J., Péré, M., & Challet-Bouju, G. (2021). Impact of Warning Pop-Up Messages on the Gambling Behaviour, Craving, and Cognitions of Online Gamblers: A Randomized Controlled Trial. Frontiers in Psychiatry, 12, 711431–711431.

Canegallo K 2021. Our efforts to fight child sexual abuse online. https://blog.google/technology/safety-security/our-efforts-fight-child-sexual-abuse-online/

Carpenter, S., Zhu, F., & Kolimi, S. (2014). Reducing online identity disclosure using warnings. Applied Ergonomics, 45(5), 1337-1342. https://doi.org/10.1016/j.apergo.2013.10.005

Casilli, Antonio A., Pailler, Fred and Tubaro, Paola (2013) Online networks of eating-disorder websites: why censoring pro-ana might be a bad idea. Perspectives in Public Health, 133 (2). pp. 94-95. ISSN 1757-9139 (Print), 1757-9147 (Online) (doi:10.1177/1757913913475756)

Chambers, M. (2024). Institute of Public Affairs Submission to the Statutory Review of the Online Safety Act 2021

Clancey, G., Fisher, D., & Yeung, N. (2016). A recent history of Australian crime prevention. *Crime Prevention and Community Safety*, *18*, 309-328.

Clark, T., Foster, L., Bryman, A., & Sloan, L. (2021). Bryman's social research methods. Oxford university press.

Clarke, R.V., 1980. Situational crime prevention: Theory and practice. Brit. J. Criminology, 20, p.136.

Clayton, K., Finley, C., Flynn, D. J., Graves, M., & Nyhan, B. (2021). Evaluating the effects of vaccine messaging on immunization intentions and behavior: Evidence from two randomized controlled trials in Vermont. *Vaccine*, *39*(40), 5909-5917.

Cornish, D. B., & Clarke, R. V. (2003). Opportunities, precipitators and criminal decisions: A reply to Wortley's critique of situational crime prevention. Crime Prevention Studies, 16, 41–96.

Costello, M., Barrett-Fox, R., Bernatzky, C., Hawdon, J., & Mendes, K. (2018). Predictors of viewing online extremism among America's youth. Youth & Society, 52(5), 710-727. https://doi.org/10.1177/0044118x18768115

Cozens, P., & Love, T. (2017). The dark side of crime prevention through environmental design (CPTED). In *Oxford Research Encyclopedia of Criminology and Criminal Justice*.

Cronin, O. (2022, October 19). *An Garda Síochána unlawfully retains files on innocent people who it has already cleared of producing or sharing of child sex abuse material.* Irish Council for Civil Liberties. https://www.iccl.ie/news/an-garda-siochana-unlawfully-retains-files-on-innocent-people-who-it-has-already-cleared-of-producing-or-sharing-of-child-sex-abuse-material/

da Silva Eleuterio, P. M., de Castro Polastro, M., & Police, B. F. (2012, September). An adaptive sampling strategy for automatic detection of child pornographic videos. In *Proceedings of the seventh international conference on forensic computer science, Brasilia, DF, Brazil* (pp. 12-19).

Davies, T. (2014). Digital rights and freedoms: a framework for surveying users and analyzing policies., 428-443. https://doi.org/10.1007/978-3-319-13734-6_31

Davis A 2018. New technology to fight child exploitation. https://about.fb.com/news/2018/10/fighting-child-exploitation/.

Davis A 2020. Facebook joins industry effort to fight child exploitation online. https://about.fb.com/news/2020/06/fighting-child-exploitation-online/

Deldari, E. (2024). Users' perceptions of online child abuse detection mechanisms. Proceedings of the Acm on Human-Computer Interaction, 8(CSCW1), 1-26. https://doi.org/10.1145/3637424

De Montjoye, Y. A., Radaelli, L., Singh, V. K., & Pentland, A. S. (2015). Unique in the shopping mall: On the reidentifiability of credit card Metadata. *Science*. https://doi.org/10.1126/science.1256297.

Desmarais, A. (2024, October 7). *Little progress on suppression of pro-Palestine content, NGOs claim*. Euronews; Euronews.com. https://www.euronews.com/next/2024/10/07/human-rights-ngos-say-social-media-platforms-continue-to-censor-pro-palestine-content

Dezuanni, M., Hourigan, A., & Rodriguez, A. (2024). Principles for a better Children's Internet. *Australian Research Council Centre of Excellence for the Digital Child*, Queensland University of Technology.

Digital Child (2024, July 2). *Principles for a Better Children's Internet.* Digital Child. https://digitalchild.org.au/research/publications/reports/principles-for-a-better-childrens-internet/

Digital Rights Watch [DRW]. (2022). *Submission to the Joint Committee on Law Enforcement regarding the inquiry into*. https://digitalrightswatch.org.au/wp-content/uploads/2023/02/DRW-Submission-Law-enforcement-capabilities-in-relation-to-child-exploitation-Jan-2023.pdf#page=10.08

Draper, L. (2022). *Protecting Children in the Age of End-to-End Encryption Protecting Children in the Age of End-to-End Encryption*. https://digitalcommons.wcl.american.edu/cgi/viewcontent.cgi?article=1082&context=research

Duffy, G., Fordyce, R., & Mannell, K. 2024 Digital Child Working Paper 2024-03, Analysing Australian news media reporting about the role of digital technologies in children's lives. ARC Centre of Excellence for the Digital Child, Brisbane, Australia.

Dunson, D. B. (2018). Statistics in the big data era: Failures of the machine. *Statistics & Probability Letters, 136,* 4–9. https://doi.org/10.1016/J.SPL.2018.02.028.

ECPAT International (2019). Ethical considerations in research on sexual exploitation involving children. Bangkok: ECPAT International.

Edwards, G., Christensen, L., Rayment-Mchugh, S., & Jones, C. (2021). *Trends & issues in crime and criminal justice Cyber strategies used to combat child sexual abuse material*. https://www.aic.gov.au/sites/default/files/2021-09/ti636_cyber_strategies_used_to_combat_csam.pdf

Ehsan, R., & Stott, P. (2020). Far-right terrorist manifestos: A critical analysis. Henry Jackson Society. (n.d.).

eSafety Commissioner. (2022). *Register of industry codes and industry standards for online safety | eSafety Commissioner.* ESafety Commissioner. https://www.esafety.gov.au/industry/codes/register-online-industry-codes-standards

eSafety Commissioner. (2024a). *Industry codes and standards | eSafety Commissioner*. ESafety Commissioner. https://www.esafety.gov.au/industry/codes

eSafety Commissioner. (2024b). Tipping the balance: LGBTIQ+ teens' experiences negotiating connection, self-expression and harm online. Australian Government.

European Digital Rights, Edr. (2022, October 15). *News from Ireland question effectiveness and lawfulness of online scanning for tackling child sexual abuse: Lessons for the EU*. European Digital Rights (EDRi). https://edri.org/our-work/breaking-irish-story-shows-that-eus-csam-proposal-can-never-work/

European Digital Rights. (2022, October 15). *News from Ireland question effectiveness and lawfulness of online scanning for tackling child sexual abuse: Lessons for the EU*. European Digital Rights (EDRi). https://edri.org/our-work/breaking-irish-story-shows-that-eus-csam-proposal-can-never-work/

Facer, K. (2012). After the moral panic? Reframing the debate about child safety online. Discourse: Studies in the cultural politics of education, 33(3), 397-413. https://doi.org/10.1080/01596306.2012.681899

Fang, H., & Qian, Q. (2021). Privacy preserving machine learning with homomorphic encryption and federated learning. *Future Internet*, *13*(4), 94.

Farrell, P., Ting, I., Lim, A., Brettell, T., Shatoba, K., & Fisher, J. (2024, March 17). Raymond can't count how often he's been searched by police. Data reveals he's just one of millions. *ABC News*. https://www.abc.net.au/news/2024-03-18/how-proactive-policing-quotas-sent-nsw-police-searches-soaring/103579210

Farthing, R. (2022). *How outdated approaches to regulation harm children and young people and why Australia urgently needs to pivot*. Reset Tech Australia. https://au.reset.tech/uploads/report_-co-regulation-fails-young-people-final-151222.pdf

Feigenbaum, J., & Koenig, J. (2014). On the feasibility of a technological response to the surveillance morass. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*. https://doi.org/10.1007/978-3-319-12400-1_23.

Fjeld, J., Achten, N., Hilligoss, H., Nagy, A., & Srikumar, M. (2020). Principled artificial intelligence: Mapping consensus in ethical and rights-based approaches to principles for AI. Berkman Klein Center Research Publication, (2020-1).

Fonseca-Bustos, J., Ramírez-Gutiérrez, K. A., & Feregrino-Uribe, C. (2022). Robust image hashing for content identification through contrastive self-supervised learning. *Neural Networks*, *156*, 81-94.

Fonseca-Bustos, J., Ramírez-Gutiérrez, K. A., & Feregrino-Uribe, C. (2024). A robust self-supervised image hashing method for content identification with forensic detection of content-preserving manipulations. *Neural Networks*, *177*, 106357.

Fortuna, P., Soler-Company, J., & Wanner, L. (2021). How well do hate speech, toxicity, abusive and offensive language classification models generalize across datasets?. *Information Processing & Management*, *58*(3), 102524.

Freilich, J. D., & Greene-Colozzi, E. A. (2024). The Ideological and Theoretical Positionality of Situational Crime Prevention in Criminology. *Crime & Delinquency*, 00111287241286015.

Freilich, J. D., & Newman, G. R. (2017). Situational crime prevention. In H. Pontell (Ed.), Oxford research Encyclopedia of criminology and criminal justice. Oxford University Press. https://doi.org/10.1093/acrefore/9780190264079.013.3.

Froomkin, M., & Colangelo, Z. (2020). Privacy as safety. *Washington Law Review*, *95*, 141.

Gainsbury, S., Aro, D., Ball, D., Tobar, C., & Russell, A. (2015). Determining optimal placement for pop-up messages: evaluation of a live trial of dynamic warning messages for electronic gaming machines. International Gambling Studies, 15(1), 141-158.

Ganesh, B., & Bright, J. (2020). Countering extremists on social media: Challenges for strategic communication and content moderation. *Policy & Internet*, *12*(1), 6-19.

Gannoni A, Voce A, Napier S, Boxall H & Thomsen D (2023). *Preventing child sexual abuse material offending: An international review of initiatives*. Research Report no. 28. Canberra: Australian Institute of Criminology. https://doi.org/10.52922/rr78764

Gilbert, J. (2024). Techno-feudalism or platform capitalism? Conceptualising the digital society. *European Journal of Social Theory*, 13684310241276474.

Gillett, R., Stardust, Z., & Burgess, J. (2022). Safety for Whom? Investigating How Platforms Frame and Perform Safety and Harm Interventions. *Social Media + Society*, *8*(4). https://doi.org/10.1177/20563051221144315

Gladstone, N. (2019). Child sexual abuse material spreading 'exponentially' on social media. Sydney Morning Herald, 2 July. https://www.smh.com.au/national/child-sexual-abuse-material-spreading-exponentially-onsocial-media-20190701-p520nn.html

Goldsteen, A., Ezov, G., Shmelkin, R., Moffie, M., & Farkash, A. (2022). Data minimization for GDPR compliance in machine learning models. *AI and Ethics*, *2*(3), 477-491.

Google. (n.d.). *Fighting child sexual abuse online*. Protectingchildren.google. https://protectingchildren.google/

Google. (2024). *Google Transparency Report*. Google. https://transparencyreport.google.com/child-sexual-abuse-material/reporting?hl=en

Gupta, A., Kumar, N., Prabhat, P., Gupta, R., Tanwar, S., Sharma, G., ... & Sharma, R. (2022). Combating fake news: Stakeholder interventions and potential solutions. *Ieee Access*, *10*, 78268-78289.

Gupta, B. B., Rahulamathavan, Y., Yamaguchi, S., Brooks, T., & Yan, Z. (2019). IEEE access special section editorial: Recent advances in computational intelligence paradigms for security and privacy for fog and mobile edge computing. *IEEE Access*, *7*, 134063-134070.

Haddad, A., Sauer, J., Prichard, J., Spiranovic, C., & Gelb, K. (2020). Gaming tasks as a method for studying the impact of warning messages on information behavior. *Library Trends*, *68*(4), 576-598.

Hall, A., Wilson, C., Stanmore, E., & Todd, C. (2019). Moving beyond 'safety' versus 'autonomy': a qualitative exploration of the ethics of using monitoring technologies in long-term dementia care. BMC Geriatrics, 19(1). https://doi.org/10.1186/s12877-019-1155-6

Hegemann, H., & Kahl, M. (2017). (Re) Politicizing Security? The Legitimation and Contestation of Mass Surveillance after Snowden. *World Political Science*, *13*(1), 21-56.

Hernández, A. D., Owen, R., Nielsen, D. S., & McConville, R. (2022). Addressing contingency in algorithmic (mis) information classification: Toward a responsible machine learning agenda. arXiv preprint arXiv:2210.09014.

Heylen, K. (2023). Enforcing platform sovereignty: A case study of platform responses to Australia's News Media Bargaining Code. *New Media & Society*, *1*(18), 146144482311660. https://doi.org/10.1177/14614448231166057

Hirschprung, R., Klein, M., & Maimon, O. (2022). Harnessing soft logic to represent the privacy paradox. Informatics, 9(3), 54. https://doi.org/10.3390/informatics9030054

Ho, H., Ko, R., & Mazerolle, L. (2022). Situational Crime Prevention (SCP) techniques to prevent and control cybercrimes: A focused systematic review. *Computers & Security*, *115*, 102611.

Hudson, K. (2018). Preventing child sexual abuse through education: The work of Stop it Now! Wales. Journal of sexual aggression, 24(1), 99-113.

Human Rights Legal Centre. (2024). *Rights-First: Principles for Digital Platform Regulation*. Human Rights Law Centre. https://static1.squarespace.com/static/580025f66b8f5b2dabbe4291/t/6705cc6360afa12be510e645/1728433270650/Report_Rights+First+Principles+for+Digital+Platform+Regulation.pdf

Human Rights Watch. (2023). Meta's Broken Promises. *Human Rights Watch*. https://www.hrw.org/report/2023/12/21/metas-broken-promises/systemic-censorship-palestine-content-instagram-and

Hunn, C., Watters, P., Prichard, J., Wortley, R., Scanlan, J., Spiranovic, C., & Krone, T. (2023). How to implement online warnings to prevent the use of child sexual abuse material. *Trends and issues in crime and criminal justice*, (669), 1-14.

Ingole, C. (2023). Machine learning-based privacy policy analysis and visualization for GDPR compliance.. https://doi.org/10.21203/rs.3.rs-2977464/v1

INHOPE. (2021). *What kinds of technology are used in the fight against Child Sexual Abuse Material?* Inhope.org. https://www.inhope.org/EN/articles/what-kinds-of-technology-used-in-the-fight-against-child-sexual-abuse-materialare?locale=tr

InHope. (2024). *INHOPE | The Facts*. Inhope.org. https://inhope.org/EN/the-facts

Jang, Y. (2023). Online safety for children and youth under the 4cs framework—a focus on digital policies in australia, canada, and the uk. Children, 10(8), 1415. https://doi.org/10.3390/children10081415

Jarvie, C., & Renaud, K. (2021, October). Are you over 18? A snapshot of current age verification mechanisms. In *2021 Dewald Roode Workshop*.

Jewkes, Y., & Wykes, M. (2012). Reconstructing the sexual abuse of children:'Cyber-paeds', panic and power. *Sexualities*, *15*(8), 934-952.

Jones, K. (2019). Online disinformation and political discourse: Applying a human rights framework.

Kansok-Dusche, J., Ballaschk, C., Krause, N., Zeißig, A., Seemann-Herz, L., Wachs, S., & Bilz, L. (2023). A systematic review on hate speech among children and adolescents: Definitions, prevalence, and overlap with related phenomena. *Trauma, violence, & abuse*, *24*(4), 2598-2615.

Keller, D. (2021, September). *Written evidence submitted to the Draft Online Safety Bill Joint Committee*. Parliament of the United Kingdom. https://committees.parliament.uk/writtenevidence/39173/pdf/

Khan, S., Kamal, A., Fazil, M., Alshara, M., Sejwal, V., Alotaibi, R., & Alqahtani, S. (2022). Hcovbi-caps: hate speech detection using convolutional and bi-directional gated recurrent unit with capsule network. Ieee Access, 10, 7881-7894. https://doi.org/10.1109/access.2022.3143799

Kleinig, J. (2000). The burdens of situational crime prevention: An ethical commentary. *Ethical and social perspectives on situational crime prevention*, 37-58.

Krone T, Smith RG, Cartwright J, Hutchings A, Tomison A & Napier S 2017. Online child sexual exploitation offenders: A study of Australian law enforcement data. Report to the Criminology Research Advisory Council. Canberra: Australian Institute of Criminology. https://www.aic.gov.au/crg/reports/crg-5812-13

Krone, T., Spiranovic, C., Prichard, J., Watters, P., Wortley, R., Gelb, K., & Hunn, C. (2020). Child sexual abuse material in child-centred institutions: situational crime prevention approaches. *Journal of sexual aggression*, *26*(1), 91-110.

Landau, S. (2023a, January 24). *Finally Some Clear Thinking on Child Sexual Abuse and Exploitation Investigation and Intervention*. Default. https://www.lawfaremedia.org/article/finally-some-clear-thinking-on-child-sexual-abuse-and-exploitation-investigation-and-intervention

Landau, S. (2023b, October 12). *The Shapeshifting Crypto Wars*. Default. https://www.lawfaremedia.org/article/the-shapeshifting-crypto-wars

Law Enforcement Conduct Commission [LECC]. (2023, October 10). *Media Release - Operation Tepito Final Report*. Lecc.nsw.gov.au. https://www.lecc.nsw.gov.au/news/media-release-operation-tepito-final-report?expand=actions

Lee, H. E., Ermakova, T., Ververis, V., & Fabian, B. (2020). Detecting child sexual abuse material: A comprehensive survey. *Forensic Science International: Digital Investigation*, *34*, 301022

Lee, Y. (2024, September 26). *Thorn's Safety by Design for Generative AI: 3-Month Progress Report on Civitai and Metaphysic*. Thorn. https://www.thorn.org/blog/safety-by-design-for-generative-ai-3-month-progress-report/

Livingstone, S. (2013). Online risk, harm and vulnerability: Reflections on the evidence base for child Internet safety policy. *ZER: Journal of Communication Studies*, *18*(35), 13-28.

López, A. B., Pastor-Galindo, J., & Ruipérez-Valiente, J. A. (2024). Frameworks, Modeling and Simulations of Misinformation and Disinformation: A Systematic Literature Review. *arXiv preprint arXiv:2406.09343*.

Maimon, D., Alper, M., Sobesto, B., & Cukier, M. (2014). Restrictive deterrent effects of a warning banner in an attacked computer system. Criminology, 52(1), 33-59. https://doi:10.1111/1745-9125.12028

Mannell, K., Bloul, S., Sefton-Green, J., & Willcox, M. (2024). Digital media and technology use by families with infants, toddlers, and young children: A scoping review and call for forward momentum. *Journal of Children and Media*, 1-24.

Manokha, I. (2023). gdpr as an Instance of Neoliberal Governmentality: a Critical Analysis of the Current 'Gold Standard'of Data Protection. *Political Anthropological Research on International Social Sciences (PARISS)*, *4*(2), 173-218.

Mäntymäki, M., Minkkinen, M., Birkstedt, T., & Viljanen, M. (2022). Defining organizational AI governance. AI and Ethics, 2(4), 603-609.

Maouche, M., Srivastava, B. M. L., Vauquier, N., Bellet, A., Tommasi, M., & Vincent, E. (2022, September). Enhancing speech privacy with slicing. In *Interspeech 2022-Human and Humanizing Speech Technology*.

Martin, A. (2023). *Researchers call for UK, EU to heed scientific evaluation of client-side scanning proposals*. Therecord.media. https://therecord.media/encrypted-messaging-client-side-scanning-uk-eu-proposals

Martin, A. (2023). *Researchers call for UK, EU to heed scientific evaluation of client-side scanning proposals. The Record*. https://therecord.media/encrypted-messaging-client-side-scanning-uk-eu-proposals

Marwick, A., Clancy, B., & Furl, K. (2022). Far-Right Online Radicalization: A Review of the Literature. *The Bulletin of Technology & Public Life*. https://doi.org/10.21428/bfcb0bff.e9492a11

Marwick, A., Smith, J., Caplan, R., & Wadhawan, M. (2024). Child Online Safety Legislation (COSL) - A Primer. *The Bulletin of Technology & Public Life*. https://doi.org/10.21428/bfcb0bff.de78f444

Matulionyte, R. (2024, August 22). *Australia's privacy regulator just dropped its case against "troubling" facial recognition company Clearview AI. Now what?* The Conversation. https://theconversation.com/australias-privacy-regulator-just-dropped-its-case-against-troubling-facial-recognition-company-clearview-ai-now-what-237231

Mayer, J. (2019). *Content Moderation for End-to-End Encrypted Messaging*. https://www.cs.princeton.edu/~jrmayer/papers/Content_Moderation_for_End-to-End_Encrypted_Messaging.pdf

Mayer, J., Mutchler, P., & Mitchell, J. C. (2016). Evaluating the privacy properties of telephone metadata. *Proceedings of the National Academy of Sciences of the United States of America*. https://doi.org/10.1073/pnas.1508081113.

McDonagh, K., & Heng, Y. K. (2011). Risk, human rights and the bureacratisation of counter-terrorism.

McDuie-Ra, D. (2023). Pandemic surveillance and mobilities across Sydney, New South Wales. *Geographical Research,* 62(1), 45-57. https://doi.org/10.1111/1745-5871.12618

Meden, B., Rot, P., Terhörst, P., Damer, N., Kuijper, A., Scheirer, W.J., Ross, A., Peer, P. and Štruc, V. (2021). Privacy–enhancing face biometrics: A comprehensive survey. *IEEE Transactions on Information Forensics and Security*, *16*, pp.4147-4183.

Meta. (2021, February 23). *Preventing Child Exploitation on Our Apps*. About Facebook. https://about.fb.com/news/2021/02/preventing-child-exploitation-on-our-apps/

Miller-Idriss, C. (2022). Hate in the homeland: The new global far right. New Jersey : Princeton University Press

Mittelstadt, B. (2019). Principles alone cannot guarantee ethical AI. Nature machine intelligence, 1(11), 501-507.

Molitorisz, S., Meese, J., & Hagedorn, J. (2021). From shadow profiles to contact tracing: qualitative research into consent and privacy. Law Technology and Humans, 3(2), 46-60. https://doi.org/10.5204/lthj.1874

Mondon, A., & Winter, A. (2020). *Reactionary democracy: How racism and the populist far right became mainstream*. Verso Books.

Morley, J., Floridi, L., Kinsey, L., & Elhalal, A. (2020). From what to how: an initial review of publicly available AI ethics tools, methods and research to translate principles into practices. Science and engineering ethics, 26(4), 2141-2168.

Mott, G. (2018). *A critical reflection on the construction of the cyberterrorist threat in the United Kingdom of Great Britain and Northern Ireland*. Nottingham Trent University (United Kingdom).

Murthy, D. (2021). Evaluating platform accountability: Terrorist content on YouTube. *American behavioral scientist*, *65*(6), 800-824.

Nikitin, D., Timberlake, D. S., & Williams, R. S. (2016). Is the e-liquid industry regulating itself? A look at e-liquid internet vendors in the United States. Nicotine & Tobacco Research, 18(10), 1967-1972.

Noble, S. U. (2018). Algorithms of oppression: How search engines reinforce racism. In *Algorithms of oppression*. New York university press.

Office of the Australian Information Commissioner [OAIC]. (2021, November 3). *Clearview AI breached Australians' privacy*. OAIC. https://www.oaic.gov.au/news/media-centre/clearview-ai-breached-australians-privacy

Padden, M., & Öjehag-Pettersson, A. (2024). Digitalisation, democracy and the GDPR: The efforts of DPAs to defend democratic principles despite the limitations of the GDPR. *Big Data & Society*, *11*(4), 20539517241291815.

Page Jeffery, C. (2018). Too sexy too soon, or just another moral panic? Sexualization, children, and "technopanics" in the Australian media 2004–2015. Feminist Media Studies, 18(3), 366-380. https://doi.org/10.1080/14680777.2017.1367699

Parnaby, P. (2006). Crime prevention through environmental design: Discourses of risk, social control, and a neo-liberal context. *Canadian Journal of Criminology and Criminal Justice*, *48*(1), 1-30.

Pathak, M. A. (2012). *Privacy-preserving machine learning for speech processing*. Springer Science & Business Media.

Petit, P. (2020). 'Everywhere Surveillance': Global Surveillance Regimes as Techno-Securitization. *Science as Culture*, *29*(1), 30-56.

Phillips, W., & Milner, R. M. (2021). *You are here: A field guide for navigating polarized speech, conspiracy theories, and our polluted media landscape*. MIT Press.

Prichard, J., Scanlan, J., Krone, T., Spiranovic, C., Watters, P., & Wortley, R. (2022). Warning messages to prevent illegal sharing of sexual images: Results of a randomised controlled experiment. *Trends and Issues in Crime and Criminal Justice*, (647), 1-15.

Prichard, J., Scanlan, J., Watters, P., Wortley, R., Hunn, C., & Garrett, E. (2022). Online messages to reduce users' engagement with child sexual abuse material: a review of relevant literature for the reThink chatbot.

Prichard, J., Watters, P. A., & Spiranovic, C. (2011). Internet subcultures and pathways to the use of child pornography. Computer Law & Security Review, 27(6), 585-600.

Prichard, J., Wortley, R., Watters, P. A., Spiranovic, C., Hunn, C., & Krone, T. (2021). Effects of Automated Messages on Internet Users Attempting to Access "Barely Legal" Pornography. SexualAbuse.

Prichard, J., Wortley, R., Watters, P., Spiranovic, C., & Scanlan, J. (2024). The effect of therapeutic and deterrent messages on Internet users attempting to access 'barely legal'pornography. *Child Abuse & Neglect*, *155*, 106955.

Raher, S. (2024). Data Privacy in Carceral Settings: The Digital Panopticon Returns to Its Roots. *Northwestern University Law Review*. 119, pp.73-104

Raymen, T. (2016). Designing-in crime by designing-out the social? Situational crime prevention and the intensification of harmful subjectivities. *British Journal of Criminology*, *56*(3), 497-514.

REPHRAIN. (n.d.). *Open Letter from Security and Privacy Researchers in relation to the Online Safety Bill*. https://haddadi.github.io/UKOSBOpenletter.pdf

Reset Australia (2024). *Accountability, the Online Safety Act and the Basic Online Safety Expectations: Can safety standards be enforceable?*. Reset Tech Australia. Policy briefing paper. https://au.reset.tech/uploads/Accountability-&-Safety-requirements-0424-V2.pdf#page=5.15

Robertson, K., Khoo, C., & Song, Y. (2020). To Surveil and Predict: A Human Rights Analysis of Algorithmic Policing in Canada'. Citizen Lab and International Human Rights Program, p 127.

Sabri, B., Saha, J., Lee, J., & Murray, S. (2023). Conducting digital intervention research among immigrant survivors of intimate partner violence: Methodological, safety and ethnical considerations. *Journal of family violence*, *38*(3), 447-462.

Safer (2020, June 30). About Safer: Defend your platform. Defend children. Safer: Building the Internet We Deserve. https://safer.io/about/

Salter, M., & Whitten, T. (2022). A comparative content analysis of pre-internet and contemporary child sexual abuse material. *Deviant behavior*, *43*(9), 1120-1134.

Sanders, C. B., & Hannem, S. (2012). Policing "the risky": Technology and surveillance in everyday patrol work. *Canadian Review of Sociology/Revue canadienne de sociologie*, *49*(4), 389-410.

Sanderson, C., Lu, Q., Douglas, D., Xu, X., & Whittle, J. (2022). Towards implementing responsible ai.. https://doi.org/10.48550/arxiv.2205.04358

Sanderson, C., Douglas, D., Lu, Q., Schleiger, E., Whittle, J., Lacey, J., & Hansen, D. (2023). AI ethics principles in practice: Perspectives of designers and developers. IEEE Transactions on Technology and Society, 4(2), 171-187.

Save the Children Finland 2022. Are you concerned about your sexual interest in children? https://www.pelastakaalapset.fi/en/our-work-in-finland/child-protection-and-finnish-hotline/otanvastuun/

Schneider, P. and Rizoiu, M. (2023). The effectiveness of moderating harmful online content. Proceedings of the National Academy of Sciences, 120(34). https://doi.org/10.1073/pnas.2307360120

Shkabatur, J. (2011). A Global Panopticon-The Changing Role of International Organizations in the Information Age. *Mich. J. Int'l L.*, *33*, 159.

Schoenebeck, S., Lampe, C., & Triêu, P. (2023). Online harassment: Assessing harms and remedies. *Social Media+ Society*, *9*(1), 20563051231157297.

Schuler, M., Gieseler, H., Schweder, K. W., von Heyden, M., & Beier, K. M. (2021). Characteristics of the users of troubled desire, a web-based self-management app for individuals with sexual interest in children: descriptive analysis of self-assessment data. JMIR mental health, 8(2), e22277.

Sentas, V., & Pandolfini, C. (2017). *Policing Young People in NSW A study of the Suspect Targeting Management Plan*. https://www.piac.asn.au/wp-content/uploads/2017/10/17.10.25-YJC-STMP-Report.pdf

Sharma, K., Qian, F., Jiang, H., Ruchansky, N., Zhang, M., & Liu, Y. (2019). Combating fake news: A survey on identification and mitigation techniques. *ACM Transactions on Intelligent Systems and Technology (TIST)*, *10*(3), 1-42.

Shneiderman, B. (2020). Bridging the gap between ethics and practice: guidelines for reliable, safe, and trustworthy human-centered AI systems. ACM Transactions on Interactive Intelligent Systems (TiiS), 10(4), 1-31.

Shortland, N., Nader, E., Imperillo, N., Ross, K., & Dmello, J. (2017). The interaction of extremist propaganda and anger as predictors of violent responses. Journal of Interpersonal Violence, 36(3-4), NP1391-1411NP. https://doi.org/10.1177/0886260517747599

Silic, M., & Back, A. (2017). Deterrent Effects of Warnings on User's Behavior in Preventing Malicious Software Use. In Proceedings of the 50th *Hawaii International Conference on System Sciences*

Skoczylis, J., & Andrews, S. (2022). Strain theory, resilience, and far-right extremism: the impact of gender, life experiences and the internet. *Critical Studies on Terrorism*, 15(1), 143-168. (n.d.).

Smirnova, S., Livingstone, S., & Stoilova, M. (2021). Understanding of user needs and problems: A rapid evidence review of age assurance and parental controls.

Smith, G. (2022, August). *Reimagining the Online Safety Bill*. Cyberleagle. https://www.cyberleagle.com/2022/08/reimagining-online-safety-bill.html

Smith, M., Nolan, M., & Gaffey, J. (2024). Online safety and social media regulation in Australia: *eSafety Commissioner v X Corp*. *Griffith Law Review*, 1–17. https://doi.org/10.1080/10383441.2024.2405760

Snap Inc. (2024). *Snapchat Transparency Report | Snapchat Transparency*. Values.snap.com. https://values.snap.com/privacy/transparency

Soghoian, C. (2008). *Insecure flight: Broken boarding passes and ineffective terrorist watch lists. Policies and research in identity management* (pp. 5–21). Boston, MA: Springer.

Stardust, Z., Gillett, R., & Albury, K. (2023). Surveillance does not equal safety: Police, data and consent on dating apps. *Crime, Media, Culture*, 19(2), 274-295. https://doi.org/10.1177/17416590221111827

Stardust, Z., Obeid, A., McKee, A., & Angus, D. (2024). Mandatory age verification for pornography access: Why it can't and won't "save the children." *Big Data & Society*, *11*(2). https://doi.org/10.1177/20539517241252129

Sumner, S.A., Ferguson, B., Bason, B., Dink, J., Yard, E., Hertz, M., Hilkert, B., Holland, K., Mercado-Crespo, M., Tang, S. and Jones, C.M. (2021). Association of online risk factors with subsequent youth suicide-related behaviors in the US. *JAMA network open*, *4*(9), pp.e2125860-e2125860.

Swamy, S. D., Jamatia, A., & Gambäck, B. (2019, November). Studying generalisability across abusive language detection datasets. In *Proceedings of the 23rd conference on computational natural language learning (CoNLL)* (pp. 940-950).

Testa, A., Maimon, D., Sobesto, B., & Cukier, M. (2017). Illegal roaming and file manipulation on target computers: Assessing the effect of sanction threats on system trespassers' online behaviors. *Criminology & Public Policy*, *16*(3), 689-726.

Teunissen, C., & Napier, S. (2022). Child sexual abuse material and end-to-end encryption on social media platforms: An overview (No. 653). *Australian Institute of Criminology*. Trends & issues in crime and criminal justice. https://www.aic.gov.au/publications/tandi/tandi653

The Internet Society. (2022). *CC BY-NC-SA 4.0 Client-Side Scanning What It Is and Why It Threatens Trustworthy, Private Communications*. https://www.internetsociety.org/wp-content/uploads/2020/03/2022-Client-Side-Scanning-Factsheet-EN.pdf#page=4.59

Thomas, K., Akhawe, D., Bailey, M., Boneh, D., Bursztein, E., Consolvo, S., ... & Stringhini, G. (2021, May). Sok: Hate, harassment, and the changing landscape of online abuse. In *2021 IEEE Symposium on Security and Privacy (SP)* (pp. 247-267). IEEE.

Thorn. (2024). *2023 Impact Report*. Thorn. https://www.thorn.org/about/our-impact/2023-impact-report/

Ulbrick, A. (2021). *Predictive Policing and Young People Discriminatory impacts of pre-emptive and racialised policing in Victoria*. Flemington and Kensington Community Legal Centre Inc. https://www.policeaccountability.org.au/wp-content/uploads/2020/03/2021-12-17_PredictivePolicingAndYoungPeople.pdf

UK Home Office (2016). Modern Crime Prevention Strategy. UK Home Office

Ullman, J. R. (2017). The development and testing of potential music piracy warnings (79. University of Nevada, Las Vegas.

Ullman, J.R., & Silver, N.C. (2018). Perceived effectiveness of potential music piracy warnings. Proceedings of the Human Factors and Ergonomics Society Annual Meeting 62(1): 1353–1357. https://doi.org/10.1177/1541931218621309

Verhelst, H. M., Stannat, A. W., & Mecacci, G. (2020). Machine learning against terrorism: How big data collection and analysis influences the privacy-security dilemma. *Science and Engineering Ethics*, 26(5), 2975–2984. https://doi.org/10.1007/s11948-020-00254-w

Wang, L., Sun, Z., Dai, X., Zhang, Y., & Hu, H. (2019). Retaining users after privacy invasions. Information Technology and People, 32(6), 1679-1703. https://doi.org/10.1108/itp-01-2018-0020

Westlake, B., Bouchard, M., & Frank, R. (2012, August). Comparing methods for detecting child exploitation content online. In *2012 European intelligence and security informatics conference* (pp. 156-163). IEEE.

Westlake, B., Brewer, R., Swearingen, T., Ross, A., Patterson, S., Michalski, D., Hole, M., Logos, K., Frank, R., Bright, D., & Afana, E. (2022). *Trends & issues in crime and criminal justice Child Sexual Abuse Material Reduction Research Program Developing automated methods to detect and match face and voice biometrics in child sexual abuse videos*. https://www.aic.gov.au/sites/default/files/2022-03/ti648_developing_automated_methods_face_and_voice_biometrics_v2.pdf

Whitehead, M., & Collier, W. G. (2023). Incidental governmentality: Big tech and the hidden rationalities of government. *Digital Geography and Society*, *5*, 100071.

Williams, P., & Kind, E. (2019). Data Driven Policing: The Hardwiring of Discriminatory Policing Practices Across Europe. *European Network Against Racism*. <statewatch.org/media/documents/news/2019/nov/data-driven-profiling-web-final.pdf>

Williams, R. S., & Ribisl, K. M. (2012). Internet alcohol sales to minors. Archives of pediatrics & adolescent medicine, 166(9), 808-813.

Williams, R. S., Derrick, J., & Phillips, K. J. (2017). Cigarette sales to minors via the internet: how the story has changed in the wake of federal regulation. Tobacco control, 26(4), 415-420.

Williams, R. S., Derrick, J., & Ribisl, K. M. (2015). Electronic cigarette sales to minors via the internet. *JAMA pediatrics*, 169(3), e1563-e1563.

Williams, R. S., Phillips-Weiner, K. J., & Vincus, A. A. (2020). Age verification and online sales of little cigars and cigarillos to minors. Tobacco regulatory science, 6(2), 152.

Witzleb, A. N., Paterson, P. M., Wilson-Otto, J., Tolkin-Rosen, G., & Marks, M. (2020). Privacy risks and harms for children and other vulnerable groups in the online environment.

Wortley, R. (2013). Situational prevention of child abuse in the new technologies. In Understanding and preventing online sexual exploitation of children (pp. 188-203). Routledge.

Wortley, R., & Smallbone, S. (2012). Internet child pornography: Causes, investigation, and prevention. Bloomsbury Publishing USA.

Wortley, R., & Tilley, N. (2017). Does situational crime prevention require a rational offender?. In The future of rational choice for crime prevention (pp. 8-29). Routledge.

Yuan, L., & Rizoiu, M. A. (2025). Generalizing Hate Speech Detection Using Multi-Task Learning: A Case Study of Political Public Figures. *Computer Speech & Language*, *89*, 101690.

Zheng, H., Yuan, Q., & Chen, J. (2015). A framework for protecting personal information and privacy. Security and Communication Networks, 8(16), 2867-2874. https://doi.org/10.1002/sec.1212

Zielinska, O. A., Mayhorn, C. B., & Wogalter, M. S. (2017). Connoted hazard and perceived importance of fluorescent, neon, and standard safety colors. *Applied ergonomics*, 65, 326-334.

Zuboff, S. (2019), The Age of Surveillance Capitalism: The Fight for a Human Future at the New Frontier of Power, New York: Public Affairs.


DIGITAL
RIGHTS
WATCH©