A Conversation on Privacy, Safety, and Security in Australia: Themes and Takeaways

DECEMBER 2023

FUTURE OF PRIVACY FORUM

AUTHORED BY

Amie Stepanovich
Vice President of U.S. Policy, Future of Privacy Forum

Felicity Slater
Policy Fellow, Future of Privacy Forum

ACKNOWLEDGEMENTS

The authors would like to thank the many experts whom we consulted for their contributions to this report.

The Future of Privacy Forum (FPF) is a non-profit organization that serves as a catalyst for privacy leadership and scholarship, advancing principled data practices in support of emerging technologies. Learn more about FPF by visiting fpf.org.

TABLE OF CONTENTS

I. INTRODUCTION

On 27 October 2023, the Future of Privacy Forum, in partnership with the UNSW Allens Hub for Technology, Law and Innovation, convened a multidisciplinary meeting of experts on technology, privacy, safety, and security to discuss benefits, challenges, and unanswered questions associated with the forthcoming industry standards for the regulation of certain online content.

For purposes of the meeting, participants were asked to assume the existence of industry standards that satisfy the Online Safety Act's statutory requirements. As such, the goal was not to solicit arguments about any specific approach but rather to provide an opportunity for experts to discuss underlying opportunities and challenges in regard to the creation of industry standards, particularly in regard to partially or entirely end-to-end encrypted services.

While meeting participants were not in full agreement in regard to any specific point, there were many themes that came up multiple times within the conversation as well as areas of broad consensus on certain points. The items below represent key themes and takeaways from that meeting.

The meeting was held in Sydney, NSW, under the Chatham House Rule. A participant list is included at the end of this document for reference only; no participant should be interpreted as having endorsed this report in whole or in part.

II. BACKGROUND

The Online Safety Act of 2021[1] ("Online Safety Act") mandates the development of industry codes or standards[2] to provide appropriate community safeguards with respect to certain online content, including child sexual exploitation material ("CSEM"), pro-terror material, crime and violence material, and drug-related material. Through September 2023, the eSafety Commissioner ("eSafety") has registered six industry codes or standards that cover: Social Media Services, App Distribution Services, Hosting Services, Internet Carriage Services, Equipment, and Internet Search Engine Services.

In May 2023, however, the Commissioner rejected proposed codes or standards for relevant electronic services ("RES") and designated internet services ("DIS") on account that they "do[] not provide appropriate community safeguards." Specifically in regard to the RES code, the Commissioner identified the following four items as reason for the rejection:

1. there is no requirement on closed communication and encrypted RES Providers with capability to deploy systems, processes or technologies to detect and remove known (i.e. pre-identified) child sexual abuse material and known pro-terror material to take such steps
2. requirements on certain RES Providers to take action and invest in disruption and deterrence of child sexual abuse material and pro-terror material fail to address the omission identified above, due to enforceability concerns
3. there is no requirement on closed communication RES Providers (such as email providers) to have trust and safety personnel, and
4. there is no requirement on certain RES Providers (those which consider themselves to be not capable of reviewing and assessing materials on their services) to enforce their own policies relating to class 1A and 1B material.[3]

Under the Online Safety Act, the rejection of the RES and DIS codes by the Office of the eSafety Commissioner ("eSafety") initiated a process in which the Commissioner drafted industry standards for these sectors. A draft of the industry standards was published on 20 November 2023 and is open for public comment until 21 December 2023.[4]

III. THEMES & TAKEAWAYS

In a full-day meeting on 27 October 2023, experts on technology, privacy, safety, and security were asked to consider the implications of an industry standard to create community safeguards for certain online content and to discuss secondary benefits of the development of such a standard, potential challenges, and unanswered questions about the standard's application. The conversation was far-ranging, and at no point was there a unanimous consensus on any specific point or recommendation. However, several themes emerged throughout the meeting with varying levels of agreement. These themes are discussed individually below.

1. Participants agreed broadly on the goals of the Online Safety Act and the mission of the eSafety Commissioner.

At the outset of the meeting, participants emphasized the importance of eSafety's efforts to combat the escalating problem of CSEM, pro-terror content material, and extreme crime and violence material (Class 1A Materials) as well as content depicting crime and violence and drug use (Class 1B Materials). Participants were largely united in emphasizing that materials falling into these categories are contemptible, and that their continued circulation must be addressed. Furthermore, several participants noted that, particularly in the case of CSEM content, continued production and distribution of this material appears to be escalating.[5]

Several participants commended eSafety for taking a role as a leader in this space, noting that eSafety has played an important leadership role in drawing attention to the proliferation of abusive, violent, and illegal content online and in conceiving of possible solutions. In difficult policy arenas such as this one, where there is no clear solution, participants emphasized that clear regulatory language and open processes can help drive consensus. Some participants praised specifically eSafety's commitment to a graduated regulation, creating different levels of regulation for different types of content. Some participants noted that, in general, progress is often imperfect and that good policy solutions, even when imperfections persist, can be far more valuable than big, imprecise statements.

2. Several participants found deficits in the length and scope of the public consultation available throughout the process.

Several participants expressed regrets about the limitations in the public consultation process on the development of the industry codes, as well as the current process for the finalization of industry standards.[6] In particular, participants noted that, because of limited consultation windows, it could be challenging for academics, civil society groups, and others with constrained resources to participate effectively.

Participants emphasized that a democratic process should involve increased give-and-take between eSafety and interested parties, an element that

several participants felt was absent from the process. One participant did note that one aspect on which the previously rejected proposed code for RES had done "ok" was on not conceding ground on privacy, but still noted that the process was largely influenced by the private sector and with relatively little input from other experts.

Finally, many participants reiterated that consultation processes such as this one should give specific voice to people with relevant experience; a diverse array of members of the online safety community, as well as survivors, must be a key part of any conversation about online safety in a way that elevates, where possible, their many, and sometimes competing, priorities.

3. Participants identified several potential benefits of an industry code beyond its intended scope.

Some participants underscored that tools developed to address Class 1A and 1B content could likely be important instruments to help combat the spread of other types of objectionable content online, including non-consensual intimate imagery ("NCII") (also sometimes referred to as "revenge porn"). Such images are circulated using some of the same methods — including encrypted messaging services and/or through small online communities — as CSEM and other pornographic material. Discussants thus emphasized that developing ways to combat the spread of Class 1A and 1B material, beginning at the moments when those who create and share it connect with one another, will create processes and tools that may be able to help address the proliferation of other objectionable material.

In addition, many participants discussed that efforts to address content within the Online Safety Act could create critical opportunities for greater education for all people in regard to online threats, including content beyond the Act's central scope, including mis-, dis-, and mal-information.

4. Participants broadly opposed any approach that would require otherwise encrypted messaging services to utilize content hashing and/or client-side scanning.

Participants indicated that any requirements for organizations to utilize content hashing and/or client-side scanning requirements would be both over- and under- inclusive and would likely result in new harms to individuals. Participants raised a number of specific and distinct arguments against these practices, noting that not only could such tools be easily circumvented, but could also perpetuate abuse and even facilitate the identification of individuals, including whistleblowers, political dissidents, and others who have not violated the provisions of the Online Safety Act.

Participants with significant technical expertise drew attention to research that has demonstrated that any image can be deliberately modified to alter its "hash" or "fingerprint". These alterations, invisible to the human eye, would mean that individuals sharing known CSEM or other relevant images would be able to bypass detection.[7] In addition, it is possible that a generated image may accidentally match to known CSEM material or could be deliberately modified to do so, falsely indicating that individuals have shared offending content when they have not. Such individuals could be pushed out of their online communities in ways that result in isolation, harassment, abuse, or other harm.[8]

Participants raised particular alarm around cases where a hash of known CSEM was hidden behind content that is targeted to a specific community of individuals, such as directions to a government protest or instructions on how to securely transmit whistleblower material. In these cases, it was discussed that such content could be weaponized by oppressive regimes or other bad actors to target and identify people within those communities. In the scenarios outlined above, client-side scanning could represent not only a security risk but also a significant risk to individual privacy.

Despite these risks, participants also explained that mandates to engage in these practices would not even reach all actors. For instance, it would only be able to match to known content and not any material that is newly generated. This was discussed to be an increased limitation in the context of new technologies, including generative AI, which can create new imagery at a previously impracticable rate, or immersive technologies like augmented or virtual reality, where new content is created constantly as individuals interact with one another. Further, if individuals did want to transmit known imagery, they could simply choose to do so outside the scope of the industry code, either using a service that is not definitionally covered or that falls outside of Australia's jurisdictional reach.

5. Many participants discussed the need for unique treatment for different types of content based on distinctions in context.

Several participants pointed out important distinctions between the categories of content covered by the Online Safety Act. For example, content that constitutes CSEM is unlikely to fall outside of legal restrictions regardless of the context in which it is shared (though with regard to the possibility of false positives, discussed above). By comparison, content determined to constitute "pro-terror content" may not only be used for recruitment into terrorist organizations or other restricted purposes but also by researchers, academics, or others who study or respond to this material, including for reasons of determining effective counter-messaging.

In light of this, participants emphasized that any response must create adequately distinct requirements for the identification of both known CSEM material and other material, including known pro-terror or crime and violence content, particularly in cases of services that utilize end-to-end encryption. Encrypted services where the provider is not technically able to see the full context of the conversation would be unable to distinguish between bad actors and anyone else when they are flagged for sharing material in contravention of the Online Safety Act. Participants discussed that not only could a false positive result in long-term reputational, professional, and/or personal harm (even if later corrected for), but could also have a broad chilling effect on important research into counter-messaging as well as a drain on already-limited enforcement and investigative resources.

6. Participants flagged previous cases of mission drift in regard to certain legal authorities and warned of similar evolution.

Several participants raised for discussion previous occasions where programs were authorized and implemented for a specific purpose and then subsequently expanded in practice beyond the previously-indicated limits. Specifically, some participants pointed to the expansion of Australia's Metadata Retention Scheme as an example of mission drift having occurred in the recent past and as a concrete question of how requirements in the industry standards, once established, could be repurposed toward other ends.[9] Some participants even raised that such an expansion had already occurred in part, explaining that the underlying authorities were chiefly justified by the need for more authority to address CSEM but were also scoped to include a significant amount of other content, including content related to terrorism and crime.

On the topic of mission drift, several participants voiced worry about what other material the Online Safety Act could be applied to in the future, particularly related to imagery specifically tied to certain historically-marginalized communities, such as queer communities. Participants explained that this sort of expanded scope could lead to the internet becoming a hostile and/or dangerous place for people with queer identities, political activists, or others implicated by any expansion.

7. Participants flagged an important role for greater education, both for individuals as well as enforcers.

Some participants emphasized the importance of education and personal empowerment as tools to respond to the spread of certain online content, although other participants questioned the extent to which the issues should be made a matter of personal responsibility, with some pointing to the documented difficulty for individuals to engage in privacy self-management online. One participant in particular indicated that users are often powerless to act on issues around privacy even when aware of data practices. However, another participant pointed to research that found that those in the age group from 18-24 are "privacy actives," who care about data privacy and, when they were able to access and understand information on data practices, were willing to spend money or switch service providers to ensure that their data was protected adequately.

Participants generally discussed avenues where increased education could make a difference: education on internet context, threats, and resources for both survivors and targets of CSEM and education for law enforcement on the life cycle of CSEM and related materials, as well as how abusive communities recruit and connect with new offenders. Regarding the former, some participants questioned if schools should be required to teach more on internet literacy and online self-defense. However, in terms of the latter, participants specifically flagged studies demonstrating that many individuals are recruited into communities of abusers through public channels, utilizing specific types of "gateway" content, and that an understanding of this process could create more tangible pathways for investigation without creating some of the risks associated with other approaches. Participants broadly believed that more research on education could provide new avenues for enforcement and preventative interventions.

8. Participants supported a broad public dialogue on effective responses and solutions.

Many participants broadly supported efforts that could be deployed to launch a more comprehensive public conversation and dialogue on the means and methods, and associated risks possible to best address the increasing circulation of CSEM. Other participants emphasized that the current approach constituted an encroachment upon the privacy rights of individuals and that any subsequent approach would need to be better formulated to protect both the rights to privacy and anonymity.

In regard to possible alternatives, one participant raised a program run by international hotel chain Marriott International, which blocked access to CSEM materials on hotel Wi-Fi networks, after studies found that this material was often accessed and circulated from hotels.[10] Other participants discussed the possibility of developing regulatory interventions aimed at preventing individuals from joining abusive communities at the recruitment and early outreach stages (also discussed above). Such interventions, some participants explained, may be best able to prevent new CSEM material from being created and circulated instead of only responding to the spread of existing content.

9. Participants identified a large number of unanswered questions in regard to the creation, implementation, and enforcement of industry codes that left much uncertainty.

Participants broadly agreed that the process so far, and the expected future processes, have failed to answer many important questions. For instance, beyond recognizing the broad scope of the problems to be addressed, participants said greater clarification was needed on specific contours on what challenges currently existed that required further action. Defining the specific

problem or problems, participants indicated, could help reach consensus on solutions that would be suitable while preserving privacy and other rights.

The multitude of other questions raised by participants during a long stretch of conversation included the following: what resources will be available to enforce the new codes and the role of law enforcement; will the codes require new enforcement authority; what new resources will be needed for investigation and enforcement, given the number of currently un-addressed cases involving CSEM; if research has been done into the efficacy of other responses to CSEM and how eSafety will collect data on current and future programs (this was raised several times and once in particular in regard to the centrality of this research to questions of necessity and proportionality of responses); the extent to which end-to-end encrypted services are utilized by government and/or law enforcement in Australia and what impact the industry codes will have on that use; how many investigations have been specifically halted because of the use of end-to-end encrypted services and what were the circumstances of those investigations; and the impact that industry codes may have on military intelligence partnerships, either positive or negative.

10. Participants recognized that Australia has played a leadership role globally on issues related to Online Safety and is likely to continue to do so.

Many participants encouraged authorities to undertake a public commitment to achieve the best possible outcomes, motivated by the important role of Australia in setting standards that other countries across the world are following. Participants broadly supported eSafety's work as a matter of global importance and significance, noting Australia's global influence and associated responsibility.

IV. CONCLUSION AND NEXT STEPS

FPF is grateful for the opportunity to host such an important conversation at a momentous time in Australia's campaign for online safety. We are thankful to each of our participants, as well as our moderator, for taking time out of their lives to engage in this discussion. The above represents a summary of the day's many discussions as compiled in a series of notes recorded by the FPF team, with all identifying information removed as required under the Chatham House Rule, save for instances where we have received specific consent to attribute a specific idea or quote to the speaker.

A copy of this report will be submitted to the Office of the e-Safety Commissioner for Australia. If anyone has questions about the content in this report, the Future of Privacy Forum, or our work, please contact info@fpf.org for more information, and your remarks will be routed to the appropriate person.

V. APPENDIX

A. About FPF

FPF is a non-profit organization that works globally, serving as a catalyst for privacy leadership and scholarship, advancing principled data practices in support of emerging technologies. FPF regularly serves as a forum to bring together industry, academics, consumer advocates, and other thought leaders to explore the challenges posed by technological innovation and develop privacy protections, ethical norms, and workable business practices. Through research, publications, educational meetings, expert testimony, and other related activities, FPF works with organizations and governments to shape best practices and policies, in the United States and globally.

B. Participant List

The below list is for reference only, and no participant should be seen as having endorsed this report in whole or in part. Some participants have been omitted on request.

Note: eSafety have redacted the names of the 17 participants listed.

C. Statement from eSafety Commissioner

The Office of the e-Safety Commissioner did not participate in this meeting. However, the office submitted a statement in advance that was provided to all participants. Here is that statement:

1. eSafety is working closely with drafters to prepare industry standards for both the relevant electronic services (RES) and designated internet services (DIS) industry sections.
2. eSafety will hold public consultation on draft versions of the standards before they are determined. eSafety welcomes written submissions from all stakeholders and industry participants.
3. We are hoping to carry out this consultation before the end of the year.
4. The industry standards will provide a proportionate and practical approach for RES and DIS to address, minimise and prevent illegal content including child sexual abuse material and pro-terror content.
5. eSafety recognises the importance of private communication and online file/photo storage, and the expectations of end-users. We do not see privacy and online safety as mutually exclusive.
6. eSafety had extensive dialogue with industry representatives throughout the codes development, our positions remain the same in relation to the standards:
    - We do not expect companies to design vulnerabilities into E2EE services
    - Proactive steps can be taken to detect known, pre-identified CSAM and pro-terror materials
    - Privacy-preserving tools such as hash matching are widely used and operate without reviewing the specific content of messages, they can be deployed on unencrypted surfaces of a service such as user profile and group images and description.
    - We recognise that not all services will be able to deploy detection tools, but they can still implement other reasonable tangible measures/interventions in relation to harmful materials

ENDNOTES

1 *Online Safety Act 2021,* Australian Government Federal Register of Legislation (Feb. 3, 2023), https://www.legislation.gov.au/Details/C2022C00052.

2 *Online Safety Act 2021 – 133 Industry Standards,* Australian Government Federal Register of Legislation (Feb. 3, 2023), https://www.legislation.gov.au/Details/C2022C00052/Html/Text#_Toc94782355

3 *Summary of Reasons – Relevant Electronic Services Code,* eSafety Commissioner (May 31, 2023), https://www.esafety.gov.au/sites/default/files/2023-05/eSafety_summary_Relevant_electronic_service_providers.pdf.

4 *Online Safety (Relevant Electronic Services—Class 1A and Class 1B Material) Industry Standard 2024,* eSafety Commissioner (Nov. 20, 2023), https://www.esafety.gov.au/sites/default/files/2023-11/Draft%20Online%20Safety%20%28Relevant%20Electronic%20Services%20-%20Class%201A%20and%20Class%201B%20Material%29%20Industry%20Standard%202024%20_0.pdf.

5 *See, e.g., Subject Matter Expert Working Group Reports*, U.S. Department of Justice, https://www.justice.gov/d9/2023-06/sme_wg_reports_combined_2.pdf.

6 *eSafety Commissioner makes final decision on world-first industry codes*, eSafety Commissioner (Jan. 6, 2023), https://www.esafety.gov.au/newsroom/media-releases/esafety-commissioner-makes-final-decision-on-world-first-industry-codes.

7 Qingying Hao, Licheng Luo, Steve T.K. Jan, & Gang Wang, *It's Not What It Looks Like: Manipulating Perceptual Hashing based Applications*, University of Illinois at Urbana-Champaign, https://dl.acm.org/doi/pdf/10.1145/3460120.3484559.

8 *See, e.g., ningyu1991/ArtificialGANFingerprints*, GitHub, https://github.com/ningyu1991/ArtificialGANFingerprints; *see also*, Brad Dwyer, *ImageNet contains naturally occurring NeuralHash collisions*, RoboFlow (Aug. 19, 2021), https://blog.roboflow.com/neuralhash-collision/ (demonstrating the potential for accidental generation of images with fingerprints that match other images). For an analysis of the harm that can result from a false positive, see, e.g., Kashmir Hill, *A Dad Took Photos of His Naked Toddler for the Doctor. Google Flagged Him as a Criminal*, The New York Times (Aug. 21, 2023), https://www.nytimes.com/2022/08/21/technology/google-surveillance-toddler-photo.html.

9 *Government commits to significant metadata reform*, Attorney-General's Department (Feb. 21, 2023), https://ministers.ag.gov.au/media-centre/government-commits-significant-metadata-reform-21-02-2023.

10 *See Marriott International Joins the Internet Watch Foundation; Deploys Cisco Technology to Combat Online Child Sexual Abuse Material*, Marriot International (Feb. 22, 2023), https://news.marriott.com/news/2023/02/22/marriott-international-joins-the-internet-watch-foundation-deploys-cisco-technology-to-combat-online-child-sexual-abuse-material; *See also* Sam Sabin, Inside Marriott's fight against online child sex abuse, Axios (Aug. 30, 2023), https://www.axios.com/2023/08/30/marriotts-fight-online-child-sex-abuse#.

1350 EYE STREET NW | SUITE 350 | WASHINGTON, DC 20005 FPF.ORG | info@fpf.org