GNI Submission on the Draft Online Safety Industry Standards in Australia

I. Introduction

The Global Network Initiative (GNI) is a multistakeholder organization that brings together more than 90 prominent academics, civil society organizations, tech companies, and investors from around the world. Members' collaboration is rooted in a shared commitment to the advancement of the GNI Principles on Freedom of Expression and Privacy, which are grounded in international human rights law and the UN Guiding Principles on Business and Human Rights (UNGPs). For over a decade, the GNI Principles and corresponding Implementation Guidelines have guided tech companies to assess and mitigate risks to freedom of expression and privacy in the face of laws, restrictions, and demands, including in politically sensitive contexts.

GNI thanks the eSafety Commissioner for the opportunity to provide feedback on the draft industry standards recently published for consultation pursuant to the Online Safety Act. As we have previously noted, our global and multistakeholder membership is closely following developments in Australia, in no small part because of the important role that Australia plays as a model for democratic governance in Asia, throughout the Commonwealth, and globally. We appreciate the extensive consultation that eSafey has allowed in multiple stages of the law-making process, several of which GNI has been a part of. As a general matter, we appreciate the Commissioner's efforts to craft a balanced and practical approach to its online safety objectives. However, we continue to believe that the standards can be further improved, specifically by clearly articulating the criteria and processes for warranted exceptions and ensuring appropriate and non-discriminatory use of proactive detection technologies.

II. Background

In 2020, GNI conducted an analysis using human rights principles of existing and proposed governmental efforts to address various forms of online harm related to user-generated content — a practice we refer to broadly as "content regulation." This included the evaluation of Australia's 2019 Online Safety Discussion Paper. After extensive consultations with GNI members and outside stakeholders, including governments, in a wide range of jurisdictions, GNI published a policy brief titled "Content Regulation and Human Rights: Analysis and Recommendations," ("Policy Brief") which set out a range of observations and suggestions on

how to regulate content in a manner that upholds and strengthens human rights.

That analysis informed our May 2021 submission to the Australian government on the then-proposed Online Safety Bill, as well as our October 2022 analysis of the *Consolidated Industry Codes of Practice for the Online Industry Phase 1 (class 1A and class 1B material)*. In the two submissions, we noted serious concerns including: the overly broad and undifferentiated application of the Bill to companies across the spectrum of services; extraterritorial reach; limited exemptions for content in the public interest; an inflexible emphasis on a 24-hour takedown window; and lack of definitional clarity around thresholds for certain categories of content. GNI is pleased to see many of these concerns now being addressed, and they continue to inform our analysis of the resulting draft industry standards that have been put forward for feedback.

GNI hopes that this feedback will be useful and remains eager to engage with relevant authorities, industry associations, and civil society to ensure that Australia's approach to online safety is consistent with the country's long-standing commitments to international human rights principles, including through its engagement in the Freedom Online Coalition and the recent Declaration for the Future of the Internet. As we have stated in our last submission, if carefully balanced and subject to appropriate safeguards, including regarding transparency in implementation, independent scrutiny and oversight, and opportunities for adjustment going forward, GNI is hopeful that Australia's approach can help demonstrate effective and rights-protecting content regulation.

III. Scope of Application and Technical Feasibility Exception

GNI has concerns that the draft industry standards allow eSafety to govern a broad range of services without sufficient distinctions between the different types and sizes of covered companies and organizations based on their position in the technology ecosystem. We appreciate the consideration of the principle of proportionality by the eSafety Commissioner in framing content moderation obligations and note that the technical feasibility provision under section 7 of the respective standards makes an exception for services based on the financial cost of observing the proposed risk assessment and content regulation measures.

However, the breadth of obligations under the standards and the lack of clarity as to how this exception will operate in practice raise serious concerns about the ability of smaller and/or not-for-profit organizations, media outlets, and individual web hosts and bloggers to comply.

These types of services are an important part of what makes today's Internet a rich and resilient resource for such a wide variety of users, and their ability to comply with the standards should therefore be of paramount concern. Going forward, the standards may also create barriers to the creation of new sites, platforms, and services, many of which have traditionally been developed through organic, academic, and/or non-commercial means by actors who might have difficulty satisfying the standards as drafted. For instance, there is an active and ongoing effort globally to federate the storage and management of Internet content in order to empower users and address perceived challenges around content ownership, data protection, and competition. The draft standards in their present form could create significant obstacles to these approaches.

In order to provide greater flexibility, clarity, predictability, and accountability, eSafety should publish clear criteria that it would use to evaluate and articulate responses to Section 7 requests and commit to transparency around resulting determinations. It should also commit to using objective means to occasionally assess the practical consequences of the standards to assess whether they are working as intended and identify any unintended consequences, especially for smaller and/or non-profit entities.

IV. Lack of Clarity on Exceptions to Class 1A and Class 1B Materials for Research, Documentation, and Media Purposes

While we acknowledge the important efforts by eSafety to regulate violent and harmful materials online, the draft standards appear to neglect the existence of projects and services that require hosting materials that may fall under the categories of class 1A and class 1B materials for research, documentation, counter-messaging, or facilitating accountability against terrorist content or CSAM. There is currently no clear provision that allows safeguards for academic and media outlets, leaving an important blind spot that risks the viability of important websites and repositories such as the Syrian Archive or the Terrorist Content Analytics Platform.

While we appreciate the Commission indicating orally during the consultation on the draft online safety standards that it may consider applications for making exceptions to class 1A and class 1B materials required for research and documentation purposes, we feel that it is important to confirm and clarify how this will work in practice in the standards themselves.

V. Proactive Detection of Content

Finally, we continue to have concerns about the emphasis on the role of proactive detection technologies in the form of hash matching, keyword detection, and artificial intelligence in the Discussion Paper and draft standards. While it can often be reasonable and efficacious to apply these forms of technology, GNI has consistently cautioned against overreliance on automated tools to proactively detect and remove content due to their lack of accuracy and the tendency of resulting, erroneous content detection and action to fall disproportionately on certain user communities, including, LGBTQ, Muslims, Arabic speakers, and victims of terrorist and extremist ideologies. Moreover, there are no safeguards in the draft standards to ensure that services that offer end-to-end encrypted communications do not run afoul the requirement to detect objectionable content. Providers that offer this service cannot access their users' communications content in order to meet a detection mandate. In addition, over-reliance on the detection and removal of problematic content can lead both service providers and regulators to underappreciate and fail to address deeper challenges and alternative, potentially more rights-aligned solutions. Finally, the costs associated with these technologies can create substantial barriers to entry for smaller and/or not-for-profit services, further exacerbating the concerns articulated in sections III and IV of this submission.

If these technologies continue to be required under the standards, we strongly suggest that eSafety also: (i) facilitate and participate in the development of separate standards for their development and use; (2) require periodic audits to test their reliability and effectiveness at both the individual service and aggregated levels, with particular attention paid to possible discriminatory impacts; and (3) create an exception for end-to-end encrypted services from obligations to scan content to detect terrorist content, CSAM or other objectionable content so these services can continue to make private and secure communications possible.

VI. Conclusion

As Australia seeks to ensure the safety of its citizens both on and offline through the online safety codes, GNI urges care in ensuring that the resulting industry standards avoid unintended consequences and preserve freedom of expression and privacy rights for all users. Given the vast scope of the standards and the wide range of actors in the technology ecosystem, regulators should:

- clarify the scope and application of technical feasibility exceptions to apps and services across the technology ecosystem;
- articulate additional, clear exceptions for research and media purposes;

- limit and clarify circumstances where proactive detection of harmful content is required; and
- facilitate the development of clear standards for proactive detection technologies, as well as their periodic auditing.

We encourage eSafety to consider these recommendations as they revise and update the codes for registration. We look forward to further engagement on future industry codes and technology policy development.