DRAFT ONLINE SAFETY PLAN (OSP): KEY CONCERNS AND RECOMMENDATIONS 1 APRIL 2026

We, ARTICLE 19, the Centre for Independent Journalism (CIJ), and the Sinar Project are responding to the Malaysian Communications and Multimedia Commission (MCMC)'s public consultation on the proposed regulatory framework for the Online Safety Plan (OSP) and the Draft Risk Mitigation Code and Draft Child Protection Code which was published on 12 February 2026 and which will be part of the subsidiary instruments on the implementation of the Online Safety Act 2025 (ONSA).

Our recommendations to strengthen the proposed regulatory framework are rooted in international human rights law, in particular standards on freedom of expression, privacy, non-discrimination and equality, as well as in the United Nations Convention on the Rights of the Child (UNCRC) and its General Comment No. 25 (2021) on children's rights in relation to the digital environment.

The Draft Online Safety Plan (Draft OSP) contravenes Section 13(3) of the Online Safety Act (ONSA) with respect to the 'duty to implement measures to mitigate risk of exposure to harmful content.' Indeed,

Section 13(3) of the ONSA provides that *'The measures implemented under this section shall not unreasonably or disproportionately limit a user's expression.'*

MAIN CONCERNS:

1. **The focus on 'harmful content' will do little to protect users:** The measures outlined in the Draft OSP do little to address concerns about the systems, business models, and services of platforms, particularly social media, when they rely heavily on the moderation of 'harmful content.' This approach results in disproportionate and unnecessary interference with legitimate internet use and comes at the cost of significant infringements on rights, particularly freedom of expression and information.

2. **Privatised enforcement:** The Draft OSP reinforces the requirement that platforms must police users' speech across vast categories of content, acting as private judges over whether such content can remain online. At the same time, the government is granted vast discretionary powers to determine what constitutes 'harmful content.'

3. **Democratic deficit:** Since the government can penalise service providers for failing to fulfil their responsibilities and can dictate what information can be shared and accessed online, the Draft OSP has significant implications for the rights of internet users and service providers. This is especially concerning when the criteria for 'harmful content' set out in the Draft OSP appear difficult to understand or implement. We also note with concern that the Draft OSP appears to broaden the category of 'harmful content' even more than the ONSA does, despite not undergoing the full legislative process and parliamentary scrutiny.

4. **General monitoring obligations and risks to freedom of expression:** We would like to stress that the combination of vague and undefined categories of 'harmful content' with expectations of detection and removal effectively incentivises proactive monitoring of user content. This creates a high risk of over-removal of lawful expression, as platforms are likely to err on the side of caution to

avoid liability. Such a system amounts to generalised content surveillance and leads to disproportionate restrictions on freedom of expression, failing to meet the requirements of legality, necessity, and proportionality under international human rights law.

5. The government is looking to introduce a **social media ban for children under 16 through the Children Protection Code**. ONSA does not provide a basis to ban users under the age of 16 from social media. In fact, the duty to protect the online safety of child users in section 18, which includes measures **'to prevent access of a user identified to be a child to a content suspected to be a harmful content**' or **'to control personalized recommendation systems suitable for child users'** seems to be predicated on the assumption that children will have access to social media. The Child Protection Code, introducing a social media ban, therefore seems at odds with the ONSA itself.

We also note that the OSP references other legislation, for example, the Digital Services Act (DSA). However, we believe that it is incorrect to state that the DSA requires companies to 'take reasonable steps to minimise the availability and spread of harmful or illegal content through mechanisms such as content moderation processes, accessible reporting and redress mechanisms, and transparency and accountability measures.' In fact, the DSA is widely described as 'content agnostic,' meaning it regulates risks related to the **systems and processes** underlying content moderation, advertising or recommender systems - very much unlike the ONSA, it does not define categories of illegal or harmful user speech or focus on when and how such speech ought to be restricted.

GENERAL RECOMMENDATIONS

1. The Draft OSP presents an opportunity to move away from business models that incentivise harm towards a framework grounded in human rights and safety by design. However, regulatory

responses must not rely solely on content restriction. Instead, the government should require service providers to urgently and effectively conduct **Human Rights Due Diligence and Child Rights Impact Assessments (CRIAs),** and to identify and mitigate any risks associated with their services, including those specifically related to children's rights. Such an approach would lead to better human rights protection for minors and adults alike. These assessments should be transparent and subject to public and regulatory scrutiny.

2. Any measures adopted under the OSP must fully **comply with international freedom-of-expression standards, in particular Article 19(3) of the International Covenant on Civil and Political Rights (ICCPR)**. Restrictions on speech must be clearly defined in law, pursue a legitimate aim, and be necessary and proportionate. Vague categories such as 'harmful' content risk leading to overbroad and arbitrary enforcement, and should be clearly defined and narrowly interpreted to avoid undue restrictions on lawful expression.

3. The Draft OSP and the Child Protection Code should also be aligned with the **UN Convention on the Rights of the Child (UNCRC) and General Comment No. 25 on Children's rights in relation to the digital environment,** ensuring that children's rights are protected holistically. This includes:

    - **Withdraw the proposal to ban social media** for children under 16 and instead prioritise measures in line with international best practices and well-established systemic risk management approaches, requiring tech companies to assess how their products may impact children and mitigate risks upstream while upholding their rights to freedom of expression and privacy.
    - **Moving beyond a narrow focus on content to address the full range of systemic risks,** including those related to content, contact, conduct, and commercial practices, while ensuring that protective measures do not unduly restrict children's access to information.

4. A fundamental requirement of the online safety regulatory framework is that services must comply with their published terms of service. It is advisable for **MCMC to conduct a comprehensive analysis of how these services have modified their public terms since 2025**. This analysis will help ensure that companies are not merely lowering their standards to meet compliance requirements but are instead enhancing their systems and processes to fulfil the intent of the Act.

5. In their 90-day and annual reports, services should be **required to report specifically on harms to children and users in general**, and on their prevalence, based on established regulatory standards rather than company standards, and to make these reports public.

6. **Adopt a whole-society approach** that not only examines social media in isolation but also proactively engages directly with children, civil society, women's rights groups, children's rights groups, parents, health services, educational settings, and other relevant stakeholders. By collaborating with these groups, the government can gain valuable insights, develop comprehensive strategies, and implement impactful measures to safeguard children and individuals from online harm.

This is outlined in detail below:

Key Components	Points of Considerations	Concerns	Recommendations
3.2 Content monitoring and content management	1. Whether the proposed scope and level of detail for describing content management and moderation approaches in the OSP are clear, proportionate, and practical; and 2. Any suggestions on additional aspects of content management and moderation that should be explained in the OSP to support transparency and effective regulatory oversight?	**Main Concerns:** ● In our view, because the Online Safety Act (ONSA) is problematic and does not lay a foundation to protect freedom of expression and other rights, in particular through focusing on seeking to regulate user generated content rather than the business model and systems and processes of the social media platforms themselves, any additional or subsidiary legislation enacted is likely to also fail to respect human rights. ● For example, the proposed Draft OSP reinforces the idea that platforms must police users' speech across vast categories of content, acting as private judges over whether such content can remain online. At the same time, the government is granted vast discretionary powers to determine what constitutes 'harmful content.' ● Since the government can penalise Service Providers for failing to fulfil their responsibilities and can dictate what information can be shared and accessed online, the Draft OSP has significant implications for the rights of internet users and service providers. Furthermore, the lack of vetting by Parliament or public scrutiny for these interventions raises additional concerns.	1. One of our recommendations is that the government require service providers to urgently and effectively conduct **Human Rights Due Diligence and Child Rights Impact Assessments (CRIAs)**, and to identify and mitigate any risks associated with their services, including those specifically related to children's rights. Such an approach would lead to better human rights protection for minors and adults alike. 2. It is recommended that the Draft OSP be revised to encompass the comprehensive spectrum of risks that children face, specifically in the areas of content, contact, conduct, and consumer/contract issues. 3. Any legislation seeking to regulate social media companies should ensure that human rights and freedom of expression lie at the heart of its right, as protected by Article 19 of the International Covenant on Civil and Political Rights (ICCPR). We are aware that Malaysia is not a party to the ICCPR. Yet, in 2021, the UN Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression noted that

		● The measures outlined in the Draft OSP do little to address concerns about platforms' systems, business models, and services, particularly social media, that rely heavily on the moderation of 'harmful' content. This approach results in disproportionate and unnecessary interference with the rights to freedom of expression and privacy, among others. This is especially concerning when the criteria for 'harmful content' set out in the Draft OSP are vague, thereby increasing the risk of incentivising the removal of broad categories of protected speech. ● The Draft OSP predominantly addresses 'harmful content'; however, it does not include sufficient measures to prevent the wider range of risks that children may face online. For example, **c**ontact risks (such as grooming, online bullying, or strangers attempting to exploit children), conduct risks (such as sharing personal information or participating in risky challenges), or children being exposed to unfair commercial practices, hidden terms, or exploitation through in-app purchases or misleading advertising. ● Therefore, the Draft OSP needs to clearly explain how companies should address the impact of their design choices on children's and the public's	the content of Article 19 of the ICCPR is based on Article 19 of the Universal Declaration of Human Rights (UDHR) and thus should inform Malaysia's obligations under international law. Under Article 19(3) of the ICCPR, restrictions to the freedom of expression are permissible only when 'provided by law,' and necessary for 'the rights or reputations of others' or 'for the protection of national security or of public order or of public health and morals.' The principles of legality, legitimacy, necessity and proportionality under Article 19(3) of the ICCPR should be applied throughout in line with freedom of expression standards. 4. As pointed out in General Comment No. 25 to the UNCRC, 'age assurance systems should be consistent with data protection and safeguarding requirements.'

		lives. **Specific Concerns:** **3.2.1 - Content moderation obligations** - In terms of an important omission, we note that the section on content moderation does not consider at all that platforms should also take steps to mitigate any risks that content moderation systems pose to freedom of expression, by potentially removing vast amounts of protected speech, in particular when automated systems are used that cannot properly account for nuance and context. - It also does not require transparency towards users for content moderation decisions. The user should be able to access a statement of reasons and appeal such a content moderation decision. - We note that almost all the provisions are problematic and do not align with freedom of expression standards. **We only provide commentary on a number of distinct issues; the information below summarises but does not comprehensively list our concerns.** ● **Para (a)&(b)**	

		- The lack of a threshold and clarity on what constitutes 'harmful content' is already a major issue under the ONSA; as a result, the proposed measures lack clarity. We believe implementation will be difficult. - The Draft OSP requires proactive monitoring by the licensed service providers. This is not only a huge problem from a privacy standpoint. It is also likely to result in more lawful content or forms of protected speech being taken down from the internet and to extensive censorship. - The inconsistency and non-clarity of using 'priority harmful content' and 'harmful content' under the proposed draft raised concerns that the platforms once again need to proactively monitor every 'harmful content', not just 'priority harmful content' as per schedule one of the Act. ● **Para (c)** - For age assurance, we recommend incorporating language that requires privacy-preserving age assurance systems, which are proportional to the risk and purpose. As pointed out in General Comment No. 25 to the UNCRC, 'age assurance systems should be consistent with data protection and safeguarding requirements.'	

		• **Para (e)** - There is no clear distinction between the types of content being uploaded or live-streamed. When it comes to 'priority harmful content,' such as child sexual abuse material (CSAM), it is essential to prevent these images from being uploaded in the first place, thereby blocking access. However, because the term 'harmful content' is not clearly defined in the law, proactive monitoring of live streams that involve legitimate expression may be adversely affected. • **Para(f)** - Same as above, because the term 'harmful content' is not clearly defined in the law, disabling or suspending accounts that involve legitimate expression may have negative consequences for freedom of expression. • **Para (h)** This appears to broaden the categories of 'harmful content' even more than is the case under the ONSA, which does not make reference to this exceptionally broad category of content 'that may undermine national security, public order and social cohesion' - it is also completely unclear what compliance with this provision would look like as 'putting place in safeguards	

		to address [such content]' could mean basically anything. ● **Para (k),** we argue that this overly imposes broad duties on platforms to be ever ready to remove content at the government's order, which can be deleterious by bypassing the court process and increasing government intervention. ● **Para 3.2.2 Detection and Removal Systems** - We are concerned that requirements for 'proactive monitoring,' particularly through automated tools, risk leading to the over-removal of protected speech. Such systems are prone to error and lack contextual understanding. We are concerned that the Draft OSP may incentivise platforms to err on the side of removal to avoid liability, thereby undermining freedom of expression. - While the framework mentions transparency and handling of false positives, it does not clearly ensure robust user safeguards, such as notice, the right to appeal, and independent oversight. Without strong procedural protections, users' content may be removed arbitrarily or without adequate remedy. - The obligation on providers to detect and remove 'harmful' content,	

		particularly when this concept is vague, effectively shifts responsibility for adjudicating the legality of speech to private companies. - Although it may be argued that monitoring merely enables companies to detect potentially illegal or other problematic content, in practice, mere detection is almost always coupled with removal or other actions that reduce the availability of such content. This is deeply problematic, given that content-monitoring technology is not nearly as accurate as users sometimes perceive it to be. In particular, hash-matching algorithms, e.g., for CSAM detection and natural language processing tools, may not be accurate or contextualised, especially in a multilingual country such as Malaysia, in distinguishing legal or illegal content, because legality may vary by context, such as news reporting, satire, or parody. Vast amounts of legitimate content may therefore be removed. Moreover, these technologies infringe on users' privacy rights because they require the analysis of individuals' communications.	
3.3 User empowerment and control	The Commission invites views on whether the proposed approach to describe user-controlled content filters in the OSP is clear, proportionate, and	● We welcome measures that give users, including children, more control over their online experiences. This includes the ability to block, filter or mute content and other users. However, it is important to note that promoting user controls and	● It is essential that filtering tools are fully transparent, optional, and user-controlled, with clear information on how content is filtered, to ensure users can make informed choices

	practical, including whether the level of explanation appropriately supports user understanding and empowerment.	online tools alone is not an effective way to ensure safety, especially for children. Such measures should be viewed as complementary to other safety-by-design strategies. ● Additionally, we emphasise that the MCMC and Service Providers should avoid shifting responsibility for safety onto children or users; this responsibility rests with the Service Providers. While it is beneficial to offer users means to manage the content they see and the individuals they interact with, these tools should complement the proactive safety measures taken by service providers.	and retain control over their online experience.
3.4 Child Protection		**Main Concerns:** **1. Social media ban for children under 16** ● The government is looking to introduce the social media ban through the Child Protection Code, which is 'to be issued by the Commission under section 80 of ONSA for the purpose of specifying the measures that Licensed Service Providers shall implement to ensure safe use of their services by child users in compliance of the duty under section 18 of ONSA.' ● As we have stated previously, the ONSA is problematic, lacks clarity, and does not lay a foundation for protecting freedom of expression and other rights.	1. Overall, instead of opting for an outright ban, the government must **withdraw the proposal to ban social media** for children under 16 and instead prioritise measures in line with international best practices and well-established systemic risk management approaches, requiring tech companies to assess how their products may impact children and mitigate risks upstream while upholding their rights to freedom of expression and privacy. 2. The design of services is crucial in influencing children's exposure to risks. While these practices may aim to increase business profits, they can have detrimental effects on children's rights, safety, and well-being. Service Providers must ensure that their products and services are designed to uphold children's rights. Additionally, they should identify and mitigate potential risks to

		We believe that any legislation aimed at regulating social media companies must prioritise transparency and accountability. ● Even so, **we believe that the ONSA does not provide a basis to ban users under the age of 16 from social media. In fact, the duty to protect the online safety of child users in section 18, which includes measures 'to prevent access of a user identified to be a child to a content suspected to be a harmful content' or 'to control personalized recommendation systems suitable for child users' seems to be predicated on the assumption that children will have access to social media. The Child Protection Code, introducing a social media ban, therefore seems at odds with the ONSA itself.** **2. The ban undermines children's human rights** The proposed social media ban would most likely violate children's rights and best interests. This is why: ● A blanket social media ban for children would likely violate international human rights standards, including legality, necessity, and proportionality.	children before harm occurs, as emphasised in General Comment 25 and reflected in the Digital Services Act (DSA). More details can be found at: 5Rights' Disrupted Childhood report. 3. The Draft OSP must align with international best practices and be explicitly grounded in children's rights and best interests, with reference to the Convention on the Rights of the Child and General Comment 25 on children's rights in relation to the digital environment. The rationale for this is that children have established rights and protections under the Convention, and their experiences in a technology-driven world must adhere to these standards. General Comment 25 outlines how children's rights apply in the digital environment, emphasising that their best interests must be a primary consideration. ➔ 'Should ensure that, in all actions regarding the provision, regulation, design, management and use of the digital environment, the best interests of every child is a primary consideration. Ensure transparency in the assessment of the best interests of the child and the criteria that have been applied' **(General Comment 25, Paragraphs 12-13, in reference to UNCRC Article 3).** ➔ 'Requiring the business sector to undertake child rights due diligence, in particular to carry out child rights impact assessments and disclose them to the public, with special consideration given to the differentiated and, at times, severe

		● Such a ban would infringe on children's rights to express themselves and access information, as protected by Malaysia's Federal Constitution, Article 19 of the UDHR and Article 12 of the UNCRC. ● Children's rights apply until the age of 18. Measures to protect children's rights should therefore cover all children (not just under-16s) and apply to all digital products and services that are likely to be accessed by or impact children (not just social media). ● Social media bans also fail to address tech companies' harmful business models and practices, nor do they create better or safer spaces for children. Instead, these approaches may disincentivise tech companies – both those within and beyond the restriction's scope – from providing age-appropriate and rights-respecting digital experiences for children ● Limiting access to online platforms undermines children's rights, their media literacy development, and their ability to engage meaningfully in societal issues. ● Children's autonomy and self-development depend on their freedom to access information and communicate online. ● If faced with a ban, children might migrate to a less safeguarded	impacts of the digital environment on children' **(General Comment 25 Para. 38, in reference to UNCRC Art. 4).** ➔ 'Taking all appropriate measures to protect children from risks to their right to life, survival and development. Risks relating to content, contact, conduct and contract encompass, among other things, violent and sexual content, cyberaggression and harassment, gambling, exploitation and abuse, including sexual exploitation and abuse, and the promotion of or incitement to suicide or life-threatening activities, including by criminals or armed groups designated as terrorist or violent extremist' **(General Comment 25, Para. 14, in reference to UNCRC Art. 6).** ➔ 'The government should ensure that digital service providers offer services that are appropriate for children's evolving capacities' **(General Comment 25 Para. 20, in reference to UNCRC Art. 5).** ➔ 'The government should take appropriate steps to prevent, monitor, investigate and punish child rights abuses by businesses' **(General Comment 25 Para. 38, in reference to UNCRC Art. 4).** ➔ 'Requiring all businesses that affect children's rights in relation to the digital environment to implement regulatory frameworks, industry codes and terms of services that adhere to the highest standards of ethics, privacy and safety in relation to the design, engineering, development, operation, distribution and marketing of their products and services' **(General Comment 25 Para 39 in reference to UNCRC Art.4).**

		environment, including 'the dark web'. ● The ban will also lead to isolation when excluding children from shared social spaces, including digital ones, which can increase feelings of exclusion instead of alleviating distress. ● The proposed ban aims to address concerns about the exploitation of children's data by requiring age verification, which would eliminate online anonymity. This could increase the processing of personal data and raise privacy concerns for children. ● The ban would create a dangerous 'cliff-edge' phenomenon. At the age of 16, children would be compelled to suddenly navigate high-risk environments with insufficient preparation. Consequently, removing opportunities for gradual, supported engagement would not cultivate resilience; rather, it would merely defer the onset of risk. **3. Enforcing the social media ban through age verification undermines the rights of all users** ● Age assurance should be used to provide children with age-appropriate digital experiences and must be lawful, rights-respecting, privacy-preserving, risk-based, and proportionate.	➔ 'Prohibiting by law the profiling or targeting of children of any age for commercial purposes on the basis of a digital record of their actual or inferred characteristics, including group or collective data, targeting by association or affinity profiling' **(General Comment 25 Para. 42 in reference to UNCRC Art. 4).** ➔ 'Ensuring that automated systems or information filtering systems are not used to affect or influence children's behaviour or emotions or to limit their opportunities or development' **(General Comment 25 Para. 62 in reference to UNCRC Art. 13-14).** ➔ 'Regulate against known harms and proactively consider emerging research and evidence in the public health sector, to prevent the spread of misinformation and materials and services that may damage children's mental or physical health. Measures may also be needed to prevent unhealthy engagement in digital games or social media, such as regulating against digital design that undermines children's development and rights **(General Comment 25 Para. 96, in reference to UNCRC Art. 24).** ➔ 'Introducing or using data protection, privacy-by-design and safety-by design approaches and other regulatory measures, States parties should ensure that businesses do not target children using those or other techniques designed to prioritize commercial interests over those of the child. Examples of such techniques are opaque or misleading advertising or highly persuasive or gambling-like design features' **(General**

		• The government's proposal that all social media platforms operating in Malaysia will be required to adopt mandatory electronic Know-Your-Customer (e-KYC) verification using government-issued documents, such as MyKad, passports, and MyDigital ID, inherently involves expanding surveillance technology. Once mass surveillance systems are established, they can be easily exploited by governments, private corporations, and malicious actors alike. This raises significant privacy and safety concerns for everyone involved. It also effectively erases the possibility of anonymous online expression, a key enabler of the right to free expression that must be protected. • A document-based system may also lead to exclusion or discrimination, particularly for individuals without recognised identity documents and who are already facing high levels of structural discrimination. Using government-issued records could disproportionately exclude vulnerable communities, such as undocumented individuals, refugees, the LGBTQIA+ community, the elderly, and those in rural areas with limited connectivity. This exclusion undermines their access to essential information and support networks, potentially widening existing inequalities and reinforcing systemic barriers.	**Comment 25 Para. 110, in reference to UNCRC Art. 31).**

3.5 Transparency and reporting	a) the proposed ninety (90) day period for Licensed Service Providers to prepare and submit the initial OSP following the commencement of the relevant regulations; b) the proposed annual submission of the OSP; and c) the proposed requirement for Licensed Service Providers to notify the Commission and submit an updated OSP where there are material changes to its contents.	It is crucial to ensure that the Draft OSP clearly outlines what Service Providers must report to meet their compliance obligations.	In their 90-day and annual reports, services should be required to publicly report on identified human rights risks to children and users in general, allowing for public scrutiny.
3.6 Governance and Accountability	The Commission invites views on whether the proposed approach to describing governance and accountability arrangements in the OSP is clear, proportionate, and practical in supporting effective oversight and compliance with obligations under ONSA.	We welcome proposals to appoint staff to report on compliance. While accountability among senior staff is crucial to ensuring compliance, responsibility for making design decisions that affect the safety of children and general users should not rest solely with top management. It is essential that safety standards are implemented at all levels of the organisation so that they are thoroughly understood and embraced throughout the organisation.	It is essential that safety standards are implemented at all levels of the organisation so that they are thoroughly understood and embraced throughout the organisation.

DRAFT RISK MITIGATION CODE AND DRAFT CHILD PROTECTION CODE:

KEY CONCERNS AND RECOMMENDATIONS

1 APRIL 2026

I. OVERVIEW

The **Risk Mitigation Code (RMC)** and **Child Protection Code (CPC)** to be issued under Part III of the Online Safety Act 2025 (ONSA) aim to operationalise statutory duties imposed on Licensed Service Providers to mitigate 'harmful content' and ensure the safety of child users online.

While the policy objective of enhancing online safety is both legitimate and necessary in light of the documented proliferation of online harms, including scams, child exploitation and abuse, and non-consensual sharing of intimate images, the proposed Codes fall significantly short of international human rights standards and will likely pose further limitations to freedom of expression and privacy. Its overbreadth, proportionality of restrictions and implementation feasibility raise added concerns, heightening existing risks of unwarranted surveillance and creating a chilling effect on online discourse.

We find that online safety regulation must move beyond reactive or enforcement-centric approaches toward a rights-respecting, risk-based, and safety-by-design framework that focuses on the architecture and business models of the platform service providers rather than imposing restrictions that would ultimately undermine the users' human rights. Without such recalibration, the Codes risk undermining fundamental freedoms, exacerbating inequalities, and introducing systemic risks to privacy and data protection.

II. INTERNATIONAL HUMAN RIGHTS STANDARDS AND FRAMEWORK

The submission is grounded in international human rights principles and standards, including the normative frameworks of the International Covenant on Civil and Political Rights (ICCPR), Convention on the Rights of the Child (CRC), Convention on the Elimination of All Forms of Discrimination against Women (CEDAW), General Comment 25 (2021) on children's rights in relation to the digital environment, interpretive guidance from the UN Human Rights Committee and UN Human Rights Council, and emerging global norms on digital governance.

These standards apply fully to digital environments and require that regulatory measures addressing online harms be clearly defined, necessary and proportionate to address the human rights impact, and be accompanied by robust safeguards.

a. Freedom of Expression and the Three-Part Test

The right to freedom of expression is protected under Article 19 of the ICCPR and applies equally online. This right includes the freedom to seek, receive, and impart information and ideas of all kinds, regardless of frontiers.

Under international law, any restriction on expression must satisfy the **three-part test** of legality, necessity and proportionality, and legitimacy, as elaborated in General Comment No. 34 of the UN Human Rights Committee.

The UN Special Rapporteur on freedom of expression has repeatedly warned that regulatory frameworks imposing broad or unclear obligations on platforms may lead to private enforcement, effectively delegating censorship functions without adequate safeguards.

b. Non-Discrimination and Substantive Equality

The principles of non-discrimination and equality are enshrined in Articles 2 and 26 of the ICCPR, Article 2 of the CRC, and Articles 1 and 2 of the CEDAW. These provisions require not only formal equality, but also substantive equality, which addresses structural disadvantage and disproportionate impacts. The UN Human Rights Council (2018) has also highlighted the risks of algorithmic discrimination, noting that digital systems may reproduce or exacerbate existing inequalities if not properly designed and regulated.

There should also not be any form of exclusion or discrimination against children's right to participate, access information and freely express themselves. Article 12 of the CRC further emphasises children's right to participate in decisions affecting them, including in digital environments. This reinforces the need for regulatory approaches that do not treat children solely as passive subjects of protection, but as rights-holders with evolving capacities.

c. Privacy and Data Protection by Design

The right to privacy is protected under Article 17 of the ICCPR, which prohibits arbitrary or unlawful interference with an individual's privacy, family, home, or correspondence. This protection extends to the digital environment, including the processing of personal data by both state and non-state actors.

International standards, including UN General Assembly resolution on the right to privacy in the digital age, establish key data protection principles, including lawfulness and legitimacy, purpose limitation, data minimisation, security and accountability.

These principles underpin the concept of privacy by design and by default, which requires that privacy safeguards be embedded into systems at the outset.

The UN High Commissioner for Human Rights has emphasised that states must ensure that regulatory frameworks do not compel private entities to engage in disproportionate data collection or surveillance practices. This obligation extends to third-party service providers and

government-mandated systems, particularly where large-scale identity verification or data sharing is involved.

Privacy must therefore be embedded by design and by default, including through data minimisation, decentralised assurance models, and independent audits.

d. Due Process, Access to Justice and Accountability

Due process and access to justice are fundamental components of the rule of law and apply to both state action and platform governance. In the digital context, it requires transparency, procedural fairness, and access to effective and timely remedies.

The UN Guiding Principles on Business and Human Rights (UNGPs) establish that companies have a responsibility to respect human rights. In the context of content moderation, this translates into requirements for:

- clear and accessible rules
- human rights due diligence process to identify, prevent, mitigate and account for how impacts on human rights are addressed
- timely user notification when content is removed or restricted
- reasons for enforcement actions
- access to timely appeal or redress mechanisms

Independent oversight is also critical to ensuring accountability. Without mechanisms for external review and audit, enforcement actions risk becoming opaque and arbitrary, particularly where platforms are required to cooperate with government authorities.

e. Intermediary Liability Principles

Intermediary liability frameworks play a central role in shaping how platforms regulate user-generated content. International standards caution against imposing strict or expansive liability, which may incentivise removal of protected speech.

Global good practices (e.g., EU Digital Services Act) emphasise safe harbour protections, risk-based obligations, and procedural accountability rather than strict liability.

The Manila Principles on Intermediary Liability recognise that intermediaries should not be held liable for third-party content unless they have knowledge of illegality and fail to act. Closely linked to this is the prohibition of general monitoring obligations, which would require platforms to proactively monitor all user content. Such obligations are widely regarded as disproportionate and incompatible with both privacy and freedom of expression.

III. PART A: RISK MITIGATION CODE (RMC)

The Risk Mitigation Code (RMC) adopts a risk-based framework, requiring (i) ‘harmful content’ risk assessments and (ii) implementation of mitigation measures. Several concerns arise when assessed against international human rights standards. In particular, its structure and scope risk undermining freedom of expression, privacy, and due process, as outlined above.

1. Scope of Application and Calibration of Risk-Based Duties *(Para. 2.1)*

The RMC applies to Licensed Service Providers, which covers application service providers or content application service providers licensed under the Communications and Multimedia Act 1998. Such providers include social media companies, as under the Malaysian social media licensing framework, they are subject to licensing if they have 8 million users or more in Malaysia. At face value, this threshold reflects proportionate targeting of systemic actors, consistent with comparative models such as the European Union's tiered obligations for Very Large Online Platforms (VLOP) under the Digital Services Act (DSA).

MAIN CONCERNS:

1. However, while the threshold introduces proportionality at the entry point, the RMC does not sufficiently differentiate obligations for online services of different roles, size, and impact in the online ecosystem. All Licensed Service Providers are required to:

    - Conduct 'suitable and sufficient' risk assessments *(para 3.1)*
    - Implement 'reasonable, proportionate and effective' mitigation measures *(para 4.1)*
    - Establish internal assurance functions *(para 4.4)*

2. The broad and undifferentiated application to Licensed Service Providers, without sufficient differentiation based on service type, functionality, systemic risk profile, and the specificity and scale of the harm, creates proportionality concerns. Even among large platforms, risk exposure varies significantly, especially for public messaging services and public content-sharing platforms.

3. The RMC also risks shifting liability onto platforms in ways that incentivise removal of protected speech to avoid liability and regulatory exposure. This is inconsistent with the right to freedom of expression and could in certain instances lead to pre-publication censorship.

Recommendations:

1) Regulatory approaches should adopt a **risk-based framework** that differentiates between platforms based on their service type, functionality, and systemic harm and impact. A tiered and graduated approach is critical to ensure effectiveness and proportionality by aligning regulatory obligations with actual risk exposure and not merely by the number of users.

2) Ensure that the RMC is brought in line with freedom of expression standards, respecting the principles of legality, legitimacy, necessity and proportionality throughout.

2. Scope of 'Harmful Content' and Legal Certainty *(para. 2.6)*

The RMC adopts the definition of 'harmful content' set out in the First Schedule of ONSA, which includes categories such as:

- Obscene or indecent content
- Content causing distress, fear, or alarm
- Content that may promote feelings of ill-will or disturb public tranquillity

MAIN CONCERNS:

1. As we had previously stated in regard to the ONSA, the categories are too broad and open to subjective interpretation, raising concerns under the requirement of the 3-part test of ICCPR. Broad and subjective categories risk inconsistent interpretation across platforms, over-enforcement by private actors, and chilling effects on freedom of expression.

2. In the platform context, such ambiguity is particularly problematic because enforcement is delegated to private actors. Platforms, faced with uncertainty and potential liability, are likely to adopt over-inclusive moderation policies, leading to the removal of protected content. The resulting chilling effect is likely to impact: political expression and dissent; communities at risk and minority voices, including racial, religious, and sexual and gender-based voices; media and journalists and artistic and cultural expression. Such outcomes are incompatible with the requirement that restrictions on expression be based on a legitimate aim and both necessary and proportionate.

Recommendations:

The RMC should incorporate interpretive guidance that narrows the categories of 'harmful content' to bring it closer in line with international standards on freedom of expression, including clear thresholds to reduce subjectivity within overbroad categories.

3. Risk Assessment Obligations Without an Explicit Human Rights Lens (*paras 3.1 - 3.4*)

The RMC requires Licensed Service Providers to conduct ‘suitable and sufficient harmful content risk assessments, taking into account factors such as user demographics, service design, and behavioural patterns' *(para 3.2)*.

MAIN CONCERNS:

1. While this reflects a structured approach, the framework is primarily operational and technical, with no explicit reference to the fundamental human rights impacts.

2. The absence of an explicit human rights dimension in the risk assessment and requirements for a human rights due diligence represents a significant accountability gap and creates a risk that mitigation measures may inadvertently produce systemic harms, including:

    - Disproportionately restrict freedom of expression
    - Produce discriminatory moderation outcomes
    - Exclude vulnerable and marginalised communities

Recommendations:

1) The RMC must incorporate explicit requirements for annual **human rights due diligence** and **independent audits**, which are to be conducted with multi-stakeholder inclusion and participation.

2) The annual human rights due diligence reports and independent audits, incorporating an **intersectional, gender, and children's rights lens**, are to be submitted to the Commission and the Online Safety Committee for **review** and be made **public**.

3) To adopt the principles of the UNGPs, particularly corporate responsibility to respect human rights. Such will also be in line with the recently launched National Action Plan on Business and Human Rights 2025-2030, aimed at 'ensuring that human rights are respected, protected and upheld at every stage of economic activity.'

4. Mandatory Registration Requirements

The RMC requires platforms to adopt safe design measures, including:

- Ensuring content is communicated only by registered users *(para 4.2.4 (a))*
- Verifying advertisers against government-issued records *(para. 4.2.4 (b))*

MAIN CONCERNS:

1. Requiring that content may only be communicated by registered users, particularly where this entails identity verification, constitutes a significant interference with the rights to freedom of expression and privacy, and therefore should not be introduced through an implementing code. Such a measure should instead be subject to full parliamentary scrutiny.

2. The RMC relies heavily on data-intensive interventions, yet Malaysia's data governance framework remains fragmented:

    - The **Personal Data Protection Act (PDPA) 2010 does not apply to government entities**, creating accountability gaps.
    - Third-party vendors may access sensitive data with limited oversight and redress mechanisms.
    - Centralised identity systems risk expanding the attack surface, and as breaches would expose multiple layers of data, it would thus result in systemic risks rather than just individual risks.

3. These measures in the RMC raise significant concerns.

    - **Freedom of Expression – Mandatory registration fails to meet the 3-Part Test as stated above. In particular:**

- ***Legality:*** The RMC lacks clarity regarding the type of data collected, who has access to this data, and the duration for which it will be retained. As this information is not clearly defined in the RMC, it creates legal uncertainty and increases the risk of arbitrary interference.
- ***Necessity:*** Is identity verification necessary to address potential harm? We believe that while traceability can help mitigate certain risks, mandatory user registration, including identity verification, poses significant threats to freedom of expression and privacy, especially for vulnerable users and at-risk communities.
- ***Proportionality:*** The collection of identity information is excessively broad, and the RMC has serious implications for the rights of internet users and service providers. Moreover, the absence of parliamentary oversight or public scrutiny for these measures raises further concerns.

- Mandatory registration **limits anonymity and pseudonymity**, which are recognised as critical for the exercise of free expression, particularly for vulnerable groups, human rights defenders, and communities at risk who may face risks of retaliation or harm.

- **Privacy and Data Protection –** Identity verification entails large-scale collection and processing of personal data, raising risks such as:

  - Data breaches and security vulnerabilities
  - Profiling and surveillance
  - Interoperability and cross-border data risks
  - Function creep beyond the original regulatory purpose

4. The government can penalise service providers for failing to meet their responsibilities. As a result, licensed service providers are **subject to strict compliance obligations** and potential liabilities. This situation may lead to a risk-averse approach, causing them to over-collect user data beyond what is strictly necessary.

5. The RMC does not include explicit human rights safeguards, including clear requirements for judicial authorisation for data access; user notification; an independent oversight body; and remedies for misuse or data breach.

6. By requiring identity-linked participation, the RMC introduces a system of traceability that may deter users from engaging in protected expression. In the absence of robust safeguards, the proposed requirement may exceed what is necessary to achieve its stated objectives.

7. Crucially, the mandatory collection of verified identity data from user creates an avenue for both state and corporate surveillance. Implementing **electronic Know Your Customer (e-KYC)** will most likely entail linking account activity to real-world identification and metadata (IP addresses, devices, timestamps). This infrastructure expands the scope for more intrusive monitoring of online behaviour (e.g., who comments on which posts and interacts with whom, and even the inference of attributes such as emotions or political beliefs), whether by state actors or through

compelled cooperation with platforms. According to the digital rights group Electronic Frontier Foundation, there is no foolproof ID verification method, risking that users' private information will be collected and stored by companies or states that lack transparent privacy standards or robust data protection. Taken together, these capabilities transform what may appear to be a narrow identity verification requirement into a broader surveillance infrastructure.

8. A further concern is the risk of function creep, whereby data collected for legitimate purposes is later repurposed for unrelated uses. While in practice, identity data gathered through mandatory registration frameworks may be accessed or shared with law enforcement. This risk is particularly acute in Malaysia, where gaps in the legal framework and specifically the exclusion of government entities from the PDPA limit safeguards against such secondary uses.

9. Without strict purpose limitation, independent oversight and clear legal thresholds for access, the data collected through the e-KYC system may undermine trust and disproportionately impact vulnerable communities. In addition, individuals such as investigative journalists, activists, human rights defenders, and whistleblowers rely heavily on anonymity to protect their safety in both online and offline realms. The absence of anonymity can expose them to harassment, intimidation and physical harm. Similarly, this method of verification could also very likely exclude the estimation of at least hundreds of thousands of communities at risk, such as refugees and stateless individuals, who are unable to obtain documents from the government.

10. This requirement would also disproportionately exclude persons or communities at risk, including undocumented persons, refugees, LGBTQI+ community, persons of older age, and those living in rural or remote areas with limited connectivity. By limiting access to Licensed Service Providers, mandatory **e-KYC** would curtail individuals' ability to participate in online discussions, access to essential information, and connect with support networks. In practice, such a policy could widen existing inequalities and reinforce systemic barriers that already leave vulnerable communities behind.

The public's right to privacy must be given an added layer of protection. Any such interference with a person's privacy must be subject to principles of legality, necessity, and proportionality.

Recommendations:

1) The requirement for mandatory identity-based user registration should be abandoned. It is not appropriate to introduce such a measure in an implementing instrument. It should instead have been subject to full parliamentary scrutiny. In addition, while traceability may assist in addressing certain harms, blanket user registration requirements, including identity requirements, poses significant risks to freedom of expression and privacy, particularly for vulnerable users and communities at risk. Alternative approaches that preserve **anonymity and pseudonymity**, while enabling targeted mitigation of harm in specific circumstances, must align with international human rights standards.

2) The RMC must uphold the requirements for **data minimisation**, and require the Licensed Service Provider to mitigate risk by shifting from **"Who are you?/Who is the user?"** (user

identification) to **"What are you doing?/What harmful act is carried out?"** (user behaviour leading to 'harmful content'). This would require:

-Identifying, addressing and responding to fake, imposter or spam accounts. Algorithms can be used for network analysis and bot detection to detect patterns of abuse or violations (eg: spam, coordinated inauthentic behaviour) by spotting a high volume of similar, rapid-fire posts from disparate, new accounts.

- Preventing known techniques used by perpetrators to target and abuse others. Platforms can compare a user's shared content against a database of known 'harmful content' (eg: hash matching repositories of child sexual abuse materials, scams, malware identification, etc) without knowing or requiring the user's identity.

3) As an alternative to mandatory government-issued ID verification, Licensed Service Providers can mitigate harm from **paid advertisements or in their online marketplace**, such as scams, fraudulent goods or misleading services by adopting a multi-layered approach. Illustrative example may include: the approach adopted by Google with a Financial Services Verification (FSV) that requires financial advertisers to prove they are licensed and legitimate entities. We note that despite such systems, large platforms still face significant challenges with ad fraud and illicit advertising. If, however, MCMC wanted to consider such practices, this would require further consultation to determine such a framework, as it would need to consider, among others, who will be responsible for determining licensing validity; how to monitor the ads; how cross-border regulatory recognition is handled; and how verification is maintained over time.

4) The RMC should be strengthened through **robust data protection safeguards**, including the application of privacy-by-design principles. This includes limiting data collection to what is strictly necessary, avoiding centralised identity systems, and ensuring that both private entities and government bodies are subject to clear accountability mechanisms. In this regard, addressing existing gaps in Malaysia's data protection framework, by amending the PDPA, including removing the exemption under section 3 of the PDPA, will be critical to ensuring that regulatory measures do not inadvertently facilitate surveillance or misuse of personal data.

5. Algorithmic Systems *(para 4.2.4 (c))*, Transparency and User Redress

The RMC requires platforms to test and adapt their algorithmic systems. While these obligations recognise the role of algorithms, the Code does not impose corresponding requirements for:

- Transparency
- Independent auditing
- User explanation

Algorithmic systems play a central role in shaping the visibility of expression, access to information, and public discourse. Without transparency or auditing requirements, there is a risk that algorithmic systems may:

- Suppress certain forms of expression or access to information
- Amplify 'harmful' or polarising content
- Produce discriminatory moderation outcomes

The UN Human Rights Council (2018) has highlighted the risks of algorithmic decision-making, particularly where systems lack transparency and accountability. Further, the EU DSA requires the platforms to:

- be audited by an independent auditor at least once a year and adopt measures that respond to the auditor's recommendations.
- allow vetted researchers to access platform data when the research contributes to the detection, identification and understanding of systemic risks in the EU.
- provide an option in their recommender systems that is not based on user profiling.
- make their risk assessment and transparency reports public.

In the absence of safeguards, algorithmic governance remains opaque, limiting users' ability to understand or challenge decisions that affect their rights.

Recommendations:

1) The RMC must incorporate **meaningful algorithmic transparency obligations** to protect users from opaque, addictive algorithms, and automated content moderation.

Given the central role of algorithmic systems in shaping access to information and exposure to content, Licensed Service Providers should be required to adapt key global good practices, including the EU DSA, to:

- Ensure transparency of the recommender system and algorithms used for recommending content or products to users. Consistent and detailed disclosure of specific input data and weights is essential for a meaningful understanding and evaluation of recommender system design, risk, and mitigation strategies.

- Provide clear multilingual and plain language user explanations of how such systems operate and why specific content is shown. This would enhance accountability while empowering users to make informed decisions about their online experiences.

- Include user options to opt-out of profiling and have the option to access content feeds not based on personal data profiling and algorithmic suggestions.

- Mandatory transparency in ad targeting, requirements for Ads to be clearly labelled, including who is paying for them and why a user sees them, with a ban on targeting based on sensitive personal data (e.g., race, religion, sexual orientation).

- Ban manipulative "Dark Patterns" that include interfaces designed to mislead users or disrupt their autonomy, including in "nudging" them into specific choices through the use of subliminal techniques, or purposefully manipulative or deceptive techniques that exploit the vulnerabilities of users. (See EU AI Act, Article 5)

2) Mandate that the Licensed Service Providers provide access to data to independent researchers to monitor compliance. Learning from the EU's DSA in conducting compliance of these requests, these data should be reliable and adequate to meet researcher's requirements to conduct any transparency assessment including potential impact on users' physical and mental health. The process and tools to access the data should also be made as user-friendly as possible.

IV. PART B: DRAFT CHILD PROTECTION CODE (CPC)

The Child Protection Code (CPC) seeks to enhance online safety for children, a legitimate and important objective. However, its measures must be assessed against the full spectrum of children's rights, including privacy, access to information, and participation, as protected under the CRC. The government is also looking to introduce the social media ban through the CPC, which is 'to be issued by the Commission under section 80 of ONSA for the purpose of specifying the measures that Licensed Service Providers shall implement to ensure safe use of their services by child users in compliance of the duty under section 18 of ONSA.

1. AGE VERIFICATION, IDENTITY-BASED ACCESS CONTROLS AND SOCIAL MEDIA BAN

The CPC requires platforms to:

i. **Implement a social media ban for children under 16, which undermines children's human rights *(para. 3.1)***
ii. **Implement age verification mechanisms *(para. 3.1)***
iii. **Verify users against government-issued records *(para. 3.2)***

i. Social Media ban for children under 16 undermines children's human rights *(para. 3.1)*

The proposed ban of social media for children under 16 is a threat to the fundamental human rights of children.

Children have rights and protections under the CRC, including to safety, privacy, protection from exploitation, and the freedoms of expression and information, which Malaysia has ratified. CRC General Comment 25 (2021) on children's rights in relation to the digital environment affirmed that children's rights apply fully in the digital world. It also clarified States' obligation to protect children's rights in the digital world and businesses' responsibility to respect these. Therefore, any measures relating to their experiences in a technology-driven world must comply with the rights under the CRC.

In addition, children, like anyone else, enjoy the right to privacy and freedom of expression as guaranteed by the ICCPR. While Malaysia is not a party to the ICCPR, the rights therein are grounded in the Universal Declaration of Human Rights (UDHR). Section 4(4) of the Human Rights Commission of Malaysia Act 1999 recognises the UDHR as a guiding tool for fulfilling its responsibilities to promote and protect human rights in the country.

The proposed social media ban would most likely violate children's rights and best interests, as:

- A blanket social media ban for children would likely violate international human rights standards, including legality, necessity, and proportionality.
- Such a ban would be a form of discrimination against children and infringe on children's rights to express themselves and access information, as protected

by Articles 8 and 10 of Malaysia's Federal Constitution, Articles 2 and 19 of the UDHR and Articles 2 and 12 of the CRC.

- Children's rights apply until the age of 18. Measures to protect children's rights should therefore cover all children (not just under-16s) and apply to all digital products and services that are likely to be accessed by or impact children (not just social media).
- Social media bans also fail to address tech companies' harmful business models and practices, nor do they create better or safer spaces for children. Instead, these approaches may disincentivise tech companies, both those within and beyond the restriction's scope, from providing age-appropriate and rights-respecting digital experiences for children.
- Limiting access to online platforms undermines children's rights, their media literacy development, and their ability to engage meaningfully in societal issues.
- Children's autonomy and self-development depend on their freedom to access information and communicate online.
- If faced with a ban, children might migrate to a less safeguarded environment, including 'the dark web'. If official sign-ups become impossible, underage users can still access the same platforms through accounts belonging to parents or older siblings', use of VPN or even borrowed identities or informal "account rental" practices.
- The ban will also lead to isolation when excluding children from shared social spaces, including digital ones, which can increase feelings of exclusion instead of alleviating distress.
- The proposed ban aims to address concerns about the exploitation of children's data by requiring age verification, which would eliminate online anonymity. This could increase the processing of personal data and raise privacy concerns for children.
- Focusing only on age-gating can miss how online harm actually happens. It happens throughout the age cycles and through direct messages, group chats, algorithmic bubbles and echo chambers, impersonation, scams, doxxing, coercion, and the speed at which abuse can scale. Even if children below 16 are kept off Licensed Services, the harm engine remains intact and will continue to manifest without stronger product safeguards and timely and effective response systems.
- Thus, the ban would create a dangerous 'cliff-edge' phenomenon. At the age of 16, children would be compelled to suddenly navigate high-risk environments with insufficient preparation and resilience. Consequently, removing opportunities for gradual, supported engagement would not cultivate resilience; rather, it would merely defer the onset of risk.

Age assurance should be used to provide children with age-appropriate digital experiences and must be lawful, rights-respecting, privacy-preserving, risk-based, and proportionate.

While globally a number of countries are rolling out age assurance measures, the BBC reported that the banning for under-16 had not been effective when interviewing children and teenagers in Australia. In the UK, there was a surge of VPN use of up to 1800% upon the implementation of age verification checks.

Ultimately, a social media ban would simply change registration streams and reduce visible accounts, while leaving behaviours and harms largely intact.

Thus, we would like to reiterate that the proposed blanket social media ban does not address the root issues of social media companies' business models and services. Children should not be prohibited from accessing the digital world; they should be able to do so safely and in ways that protect their rights. Companies that exploit children online should be prevented from doing so.

ii. Implement age verification mechanisms *(para. 3.1)*

In March 2026, more than 400 security and privacy scientists and researchers from 32 countries signed a joint letter urging a pause in the use of age-assurance technologies until there is substantial evidence of their effectiveness and societal implications.

In the context of Malaysia and the related CPC, there is a lack of transparency and no published **Roadmap for the implementation of Age Verification**. Some concerns are as below:

Age verification is not a simple plug-and-play solution; rather, it should be treated as a layered approach that comprises multiple solutions because it is an intersection between privacy, security, and rights of children and the general public.

- There is currently no clear framework publicly established and disclosed on the standards and methods guiding and governing the duties of the platforms and the regulators as they undertake age verification measures. Without a comprehensive framework, it creates a 'regulatory vacuum' whereby the regulators and government agencies or providers may default to large scale and invasive data collection.
- There is also no clarity on how the age verification measures are proportionate to the risks and harms, and in line with the obligations to promote and protect human rights, including upholding users' right to privacy.
- Further, there is insufficient publicly available information on the procurement process surrounding the selection of a third-party vendor who will implement age verification measures for the proposed social media ban.
- The proposed implementation timeline of July 2026, as signalled by the Minister of Communications, is therefore premature given the profound technical and ethical complexities of age assurance or verification.

While protecting children is crucial, implementing blanket identity checks across the internet is too dangerous and counterproductive. The core challenge in regulating online access is recognising that prohibition may not eliminate demand and may instead drive children towards less regulated platforms, underscoring the need for more effective and nuanced solutions. As

pointed out in General Comment No. 25 to the CRC, age assurance or age verification systems should be consistent with data protection and safeguarding requirements.

> ***Example:***
>
> By contrast, comparative models, such as the Australian eSafety Commissioner's roadmap spanned 5 years of continuous consultations and sandboxing before reaching the enforcement stage. Adopting an accelerated 'fast track' route risks a policy derivative that is functionally flawed, failing to address the critical concerns regarding data security and privacy.

iii. Verify users against government-issued records *(para. 3.2)*

a) The government's proposal that all social media platforms operating in Malaysia will be required to adopt mandatory e-KYC verification using government-issued documents, such as **MyKad, passports, and MyDigital ID**, inherently involves expanding surveillance technology. Once mass surveillance systems are established, they can be easily exploited by governments, private corporations, and malicious actors alike.

b) While intended to protect children, the requirement to verify age against government-issued records raises significant concerns:

- **Privacy Risks –** Large-scale collection and processing of sensitive data will likely result in centralised repositories that creates systemic vulnerabilities, including surveillance, data breaches, and misuse.
- **Proportionality –** Identity-based verification may be more intrusive than necessary, particularly where less intrusive alternatives exist.
- **Chilling Effects –** An identity-based access system may also lead to exclusion or discrimination, particularly for individuals without recognised identity documents and who are already facing high levels of structural discrimination. Using government-issued records could disproportionately exclude at-risk individuals and communities, such as undocumented individuals, refugees, the LGBTQI+ community, the elderly, and those in rural areas with limited connectivity. This exclusion undermines their access to essential information and support networks, potentially widening existing inequalities and reinforcing systemic barriers.

c) **Gaps in data protection and governance in Malaysia** make centralised ID-based verification extremely vulnerable to exploitation and unmitigated data breaches.

- The integration of third-party vendors (eg: via integration of MyDigitalID – a government agency established in 2024 under the supervision of the Prime Minister's Department) into government-mandated age verification systems exposes a critical accountability gap within the PDPA 2010. While we note the

latest amendments to the PDPA (2024) are intended to closely align with the EU GDPR, it still stands that **Section 3(1) of the PDPA** explicitly does not apply to Federal and State Government agencies, thus creating a liability shield extending to private contractors that are acting on behalf of the state. Additionally, Section 17(2) of the Data Sharing Act allows public agencies to engage third parties for "data migration, data integration or data analytics work" with compliance obligations, but lacks specific oversight mechanisms, effectively allowing for private contractors to gain access to vast government datasets. If a data breach occurs, users are likely to find themselves in a legal conundrum, especially after cases such as the Mysejahtera app data breach, the alleged Mytax portal data breach, the alleged JPN data breach, and the MySPR data leak. We believe a strong focus on strengthening foundational data governance and having independent oversight will be required before any implementation of an age verification system.

d) The **technical feasibility** of implementing current systems would also require further rigor.

- It is understood that the government is looking into implementing a zero-knowledge proof (ZKP) concept of age verification. Numerous technical organisations have raised concerns over the use of ZKP, especially on security fragility, insufficient privacy guarantees, client-side scanning, and as raised in the point above, having data centralised risk of creating a 'honeypot' for cyberattacks to occur.

Based on media reports, a specific vendor/technology provider is already engaged in a 'Regulatory Sandbox' but it is not explicitly mentioned in the public draft code. MCMC is yet to disclose its technical partners or the specific technologies that are currently being explored in the regulatory sandbox, raising concerns on transparency and accountability.

Malaysia risks implementing a system that may not be interoperable with global platforms, especially when countries that have an age assurance framework aligned closely with global standards. For example, **ISO/IEC 27566-1:2025** represents the first international standard for age assurance.

Example:

In Europe, the euCONSENT conducted a pilot on the use of 'tokenisation', interoperable, and double-blind approach as part of data minimisation efforts. In addition, a full whitepaper on the technology used and how it will be privacy-preserving was also released to the public.

In Australia, an Age Assurance Technology Trial was conducted to look at the feasibility of age assurance mechanisms. The report of the trial acknowledged that no one-size-fits-all solution exists, especially due to the layers of technologies involved in any digital service, such as the device, the browser, network, operating system and the platform. Implementing an age assurance mechanism at one layer may mean that it is still not

effective on another, however privacy considerations and service delivery would be affected disproportionately if implemented on all layers.

At the international level, technical standards on age verification or assurance systems such as IEEE 2089 and ISO/IEC 27566 have been developed, where some countries adopted these standards into their legal systems to ensure standards are met in terms of privacy, security, accessibility, performance, and functionality.

Recommendations:

1) **Withdraw the proposal to conduct age verification and ban social media** for children under 16 and instead prioritise measures in line with international best practices and well-established systemic risk management approaches, requiring tech companies to assess how their products may impact children and mitigate risks upstream while upholding their rights to freedom of expression, right to information and privacy.

2) Urgently and effectively require Licensed Service Providers to conduct **Human Rights Due Diligence and Child Rights Impact Assessments (CRIAs)** and identify and mitigate any risks associated with their services, including specifically on children's rights.

3) To carefully consider the Committee on the Rights of the Child's General Comment 25 and UNICEF's recommendations on alternatives to age-gating for child online safety, focusing on digital literacy, curated safe environments, and product design over restriction. Solutions may include mandatory "minor modes" with limited screen time/content filters, and improved parental controls.

2. CONTENT MODERATION *(PARA. 4)*

The CPC requires platforms to establish clear and robust systems to detect, remove, and prevent access to 'harmful content' ***(para. 4.2.1)***. The CPC's emphasis on restricting access to 'harmful content', while protective in intent, must be balanced against children's rights to access information and express themselves, and prevent measures that may result in over-blocking, limiting access to educational and support content.

Further, the CPC requires platforms to ensure that search and recommendation systems are appropriate for children ***(para. 7.1)***. It further requires filtering of 'harmful content' and user control mechanisms ***(para. 7.2.1)***. However, the lack of transparency and explicit user control mechanisms raises concerns as it may result in over-filtering of protected content, restricted access to information limited user understanding, and reduced accountability. Children and parents must be able to understand and manage these systems to make informed choices. Without such mechanisms, algorithmic decision-making remains opaque, limiting accountability and potentially undermining children's rights.

Under the CRC, children have the right to seek, receive, and impart information, as well as to participate in matters affecting them, including in digital environments. The UN Committee on the Rights of the Child has further emphasised that digital regulation should balance protection with empowerment, ensuring that children are not unduly restricted in their development and participation. Emphasising children's rights to participation and access to information can foster understanding and support for policies that respect their agency, rather than solely focusing on restrictions.

Content moderation practices without stringent human rights safeguards may significantly limit their exposure to public discourse on issues that concern them, particularly given that social media has become a key source of information for young people. While such measures do not entirely isolate them, over-moderation and removal will restrict their access to diverse perspectives on issues, such as politics, the economy, and global developments. Reduced exposure to information and debate may affect the development of critical thinking skills necessary for informed civic participation. Overly restrictive systems risk limiting access to beneficial content, including other educational resources and support services.

Recommendations:

1) The CPC should more explicitly embed **safety-by-design principles**, ensuring that protections are integrated into platform architecture from the outset. This includes designing systems that reduce exposure to 'harmful content', enhance user control, and prioritise well-being, particularly for children, without resorting to overly restrictive or intrusive measures.

2) The CPC should be revised to encompass the comprehensive spectrum of risks that children face, not only in the areas of content but also in the areas of contact, conduct, and consumer/contract issues.

3) Key features that may be considered:

- Regularly assess whether the services pose risks to children and teens, while noting that risks and harm may differ based on services, functionalities, and maturity of users.

- Introduce multilingual and plain language key safety features (eg: privacy settings, default protections, reporting tools) upon setting up of any user account.

- Reporting and feedback tools that are easy to find, understand and use.

- Accessibility features for all, including children with additional needs or disabilities.

- Clear notification or warnings when children are interacting with AI features.

3) Invest in and train human moderators to spot threats to children's safety and well-being, such as grooming and non-consensual intimate image (NCII).

4) Consult with children of all ages, maturity and (dis)ability levels in designing safety features and protocols.

3. ALGORITHMIC CONTROLS *(PARA. 7)* AND THE BEST INTERESTS OF THE CHILD

The CPC's provisions on algorithmic restrictions aim to reduce exposure to 'harmful content' and to ensure that the search and recommendation systems are suitable and appropriate for child users.

The CPC's requirements on search and recommendation systems include:

- Safe search by default ***(para. 7.2.1)***
- Enabling child users and parents to manage personalised recommendation systems ***(para. 7.2.3)***
- Restrictions on 'harmful' recommendations, through filtering systems ***(para. 7.2.2)***
- Limits on displaying, promoting, or recommending 'harmful content' to child users ***(para. 7.2.4)***

On one hand, these measures seem to align with safety-by-design principles and the best interests of the child standard. However, the absence of transparency and independent auditing requirements limits accountability and makes it difficult to assess effectiveness.

The principle of the **best interests of the child**, as set out in Article 3 of the CRC, requires that all measures affecting children be designed to promote their well-being. This includes ensuring that algorithmic systems do not:

- Prioritise engagement on their platform over well-being
- Reinforce 'harmful' behaviours or content
- Limit access to beneficial information

Without independent oversight and transparency, it is difficult to assess whether these systems meet this standard.

Recommendations:

1) Similar to the recommendations above related to the RMC, the CPC should require:

- Transparency and user explanations on the use of algorithms and recommender systems.

- Human Rights Due Diligence and Child Rights Impact Assessments by companies, which should be assessed for compliance by the Commission and Online Safety Committee, and to be made public.

- Independent audit.

4. PARENTAL CONTROLS *(PARA.5)* AND CHILDREN'S AUTONOMY

The CPC mandates parental control tools, including monitoring and usage restrictions ***(para. 5.1)***. While beneficial, excessive monitoring may interfere with children's privacy and autonomy, particularly for older children.

Parental controls can support child safety, but they must be implemented in a manner consistent with children's evolving capacities and rights. Article 5 of the CRC recognises that children's capacities and rights evolve with age and maturity, requiring flexible and proportionate approaches.

Recommendations:

1) CPC should ensure that parental controls complement and do not take away the Licensed Service Providers' obligations in ensuring their design and architecture have built-in safety features.

2) Ensure that parental controls are not overly intrusive and that children retain meaningful agency and autonomy in digital environments.

5. PRIVACY AND SAFETY SETTINGS *(PARA. 6)*

The CPC requires high-privacy default settings and restrictions on contact with unknown users ***(para. 6.2)***.

These provisions would however require:

- Strong enforcement and accountability.
- Independent oversight to guarantee that privacy protections are not undermined by other practices, such as profiling or behavioural advertising.
- Consistent implementation and stronger safeguards to ensure compliance with data protection principles.

Recommendations:

1) The CPC should set limits on how data is collected and used to make recommendations.

2) Ensure default settings are private, safe and secure from the start, and include:

(i) limiting contact;

(ii) turning of risky features (geolocations, autoplay of videos, microphone and camera, contact syncing, tracking);

(iii) preventing strangers from seeing or downloading content;

(iv) requiring explicit permission before sharing contact information;

(v) managing notification (turn off push notification or alerts during sleep time);

(vi) reducing excessive use; and

(vii) protecting mental health (turn off filters that negatively affect body image, self-esteem or self-harm)

V. CONCLUSION

Safety of the internet and in digital spaces cannot be achieved through isolated technical or regulatory fixes. It requires a holistic socio-technical approach grounded in human rights, and adopts privacy-preserving engineering and participatory governance.

Malaysia's current approach risks becoming a "stop-gap" regulatory model, which is focused on control rather than systemic accountability and building resilience of its users. A shift toward safety-by-design, risk-based regulation, and rights-respecting governance is essential to ensure that online safety measures do not undermine the very rights they seek to protect. A key element to support this would be the requirement to **review and amend the ONSA**, as the parent Act.

In the same vein, the RMC and CPC should adopt **a whole-of-society approach** that not only examines social media in isolation but also proactively engages directly with children, civil society, women's rights groups, children's rights groups, parents, health services, educational institutions, and other relevant stakeholders. By collaborating with these groups, the government can gain valuable insights, develop comprehensive strategies, and implement impactful measures to safeguard children and individuals from online harm.