Joint Memorandum by the Online Safety Advocacy Group on the proposed Penal Code (Amendment) Bill & Online Safety Bill

PART 2: Specific Recommendations

Introduction

The Online Safety Advocacy Group comprises the following civil society organisations (CSOs) and individuals working on the issues of freedom of opinion and expression, child rights, gender equality, and women's rights:

1. Centre for Independent Journalism (CIJ)
2. Justice for Sisters
3. KRYSS Network
4. Childline Foundation
5. Protect and Save the Children (PS The Children)
6. End CSEC Network Malaysia (ECPAT Malaysia)
7. CRIB Foundation (Child Rights Innovation & Betterment)
8. Voice of the Children
9. Kemban Kolektif
10. Women's Centre for Change (WCC)
11. Monsters Among Us (MAU)
12. Sarawak Women for Women Society (SWWS)
13. Association of Women Lawyers (AWL)
14. Johor Women's League (JEWEL)
15. Women's Aid Organisation (WAO)
16. Sisters in Islam (SIS)
17. Sinar Project
18. Maha Balakrishnan

Summary

This Memorandum is Part 2 of a collective two-part response from the Online Safety Advocacy Group regarding the government's proposed online safety and anti-cyberbullying laws ("Government's Proposals"). Our response is split into two parts due to the government's urgent request for our feedback on policy on or before 8 October 2024. As a result:

(a) Part 1 of the Memorandum contains the Online Safety Advocacy Group's position on the policy rationale for the Government's Proposals and was submitted to the government on 8 October 2024.

(b) Part 2 of the Memorandum contains detailed responses and counter-proposals to the government's specific provisions regarding the proposed Online Safety Bill (OSB) and cyberbullying.

BHEUU has shared draft language for the Government's Proposals relating to cyberbullying, but it has not shared a copy of its draft Online Safety Bill. For the purposes of providing our feedback and recommendations, we shall refer to any parent Act on online safety laws as "the parent Act".

Recommendations

1. Guiding Principles

It is our position that the policy rationale for online safety and anti-cyberbullying laws must be to better enable and strengthen responses, approaches, and mechanisms for user empowerment and protection, and be survivor-centric.

It is also our collective, unequivocal position that the parent Act should include a clear and express statement of the overarching purpose of the legislation, which must include the protection and balancing of human rights. In so doing, the Malaysian Government should take note of:

(a) the international human rights principles as enshrined under the Universal Declaration of Human Rights, Convention on the Rights of the Child (CRC), Convention on the Elimination of All Forms of Discrimination against Women (CEDAW), UN Guiding Principles on Business and Human Rights (UNGP), recommendations by UN Special Procedure Mandate Holders, and other international human rights law; and

(b) the European Union Digital Services Act, the Digital Marketing Act, and the General Data Protection Regulation (GDPR), among others.

For examples of online safety laws that incorporate express statements of purpose and compliance with human rights, we refer to the Canadian Online Harms Bill, Australia's Online Safety Act 2021, and the UK's Online Safety Act 2023. In this regard, **Box 1** below contains relevant extracts from the Canadian Online Harms Bill, which is currently pending passage in the Canadian Parliament.

BOX 1: Extracts from the Canadian Online Harms Bill

Section 9: The purposes of this Act are to

(a) promote the online safety of persons in Canada;
(b) protect children's physical and mental health;
(c) considering that exposure to harmful content online impacts the safety and well-being of persons in Canada, mitigate the risk that persons in Canada will be exposed to harmful content online while respecting their freedom of expression;
(d) enable persons in Canada to participate fully in public discourse and exercise their freedom of expression online without being hindered by harmful content;
(e) reduce harms caused to persons in Canada as a result of harmful content online;
(f) make content that sexually victimizes a child or revictimizes a survivor and intimate content communicated without consent inaccessible online;
(g) ensure that operators are transparent and accountable with respect to their duties under this Act; and
(h) contribute to the development of standards with respect to online safety.

Section 27:

When making regulations and issuing guidelines, codes of conduct and other documents, the Commission must take into account

(a) freedom of expression;
(b) equality rights;
(c) privacy rights;
(d) the needs and perspectives of the Indigenous peoples of Canada; and
(e) any other factor that the Commission considers relevant.

1.1 Promoting and Protecting the Right to Freedom of Opinion and Expression

1.1.1 The right to freedom of expression is constitutionally codified under Article 10 of the Malaysia Federal Constitution, Article 19 of the Universal Declaration of Human Rights (UDHR)**,** and Article 19 of the International Covenant on Civil and Political Rights (ICCPR). Although the Malaysian Parliament may pass laws to restrict the right to freedom of expression, it may only do so on limited, narrow grounds, and it must satisfy strict legal principles and constitutional norms. Therefore, any laws to regulate online safety and anti-cyberbullying must clearly be shown to have met these standards.

1.1.2 Under international human rights law and domestic legal principles, restrictions on the right to freedom of expression must meet the "three-part test," which mandates that restrictions must be:

- **Provided for by law**, any law or regulation must be formulated with sufficient precision to enable individuals to regulate their conduct accordingly;
- In pursuit of a **legitimate aim**, listed exhaustively as respect of the rights or reputations of others; or the protection of national security or public order, or public health or morals;
- **Necessary and proportionate** in a democratic society, i.e. if a less intrusive measure can achieve the same purpose as a more restrictive one, the least restrictive measure must be applied.

The three-part test applies to electronic communication or expression disseminated over the Internet.

1.1.3 Additionally, to address the issues related to incitement, "hate speech," Article 20(2) ICCPR provides that any advocacy of national, racial, or religious hatred that constitutes incitement to discrimination, hostility, or violence must be prohibited by law. The UN has developed the Rabat Plan of Action at the international level, which provides the closest definition of what constitutes incitement law under Article 20(2) ICCPR (refer to paragraph 3.2.7 below).

1.2 Duty of care and safety by design

1.2.1 We accept that a legal duty of care must be imposed, but it should not be a broad and undefined "duty of care." Instead, the legal duty of care must be delineated in the parent Act so that there is no opportunity for arbitrary application nor the vesting of unfettered powers to regulatory authorities and that any abuse of power can and will be held accountable[1].

[1] Some jurisdictions have further suggested implementing a series of overlapping duties of care (see link below), which would impose obligations to both adopt specific approaches as well as distinguish different types of harm and related liabilities. While taking note of the primary duty or duties of care, caution must be taken to avoid over-moderation or censorship in the name of protection.

https://www.google.com/url?q=https://thepolicymaker.jmi.org.au/the-dangers-of-pluralisation-a-singular-duty-of-care-in-the-online-safety-act-2/&sa=D&source=docs&ust=1730120036390568&usg=AOvVaw3nl9Jdtewweweww;;A25RPlqEbnWmogh

1.2.2 In ensuring “safety by design,” the parent Act and related regulations should focus on prevention by minimising online threats by anticipating, detecting, and eliminating online harms before they occur. In the UK Online Safety Act, service providers must ensure that their measures to comply with the Act’s requirements, such as content removal and moderation, do not disproportionately impact users' freedom of expression. They need to carefully balance the removal of illegal or harmful content (which should be clearly defined) with ensuring that users can still freely express themselves on the platform. Determination of illegal or harmful content must meet international human rights standards and be subject to the three-part test under 1.1.2 above.

1.3 Transparency Obligations

Transparency should be an essential requirement for any new regulatory framework. In that regard, the parent Act should impose a legal duty on the following actors that is framed in the following terms:

1.3.1 Service providers, including social media and tech companies, must be required to be transparent, open, and honest with users about how their data is being used with adequate specificity so that users are fully informed. This may be achieved by imposing legal requirements to issue privacy notices and policies that clearly outline what data is collected, how it is used, with whom it is shared, and also the individual’s rights to erasure of their personal and private information/data and to be forgotten.

1.3.2 Social media platforms and tech companies should be required to provide essential information and explain to the public how their algorithms are used to present, rank, promote, or demote content.

1.3.3 The parent Act must require platforms and providers to be transparent about the companies’ terms of service community standards, and technological resources used to ensure compliance or for digital advertising.

1.3.4 The parent Act or subsidiary regulations regulating the activities of dominant platforms should require the establishment of clear notice and action rules and internal redress.

1.4 Privacy and Data Protection

1.4.1 Any new laws regulating online harms, including cyberbullying, must include statutory protections for privacy and personal data and must be based on the principles of lawfulness and fairness.

1.4.2 Lawfulness: Personal data must be processed lawfully and for legitimate purposes, including in relation to obtaining consent, the ease of withdrawal of consent, the ease of review and erasure of personal data, and the requirement for encryption and anonymity. As in the UK Online Safety Act, Malaysian laws must also mandate that providers be obligated to ensure that their compliance with the Act respects users' privacy rights, especially concerning content moderation, risk assessments, proactive technology, or content scanning.

1.4.3 Fairness: Data processing must uphold user empowerment and privacy. Data processing should be done in a way that is fair to the user and should not put them at risk of harm in any way.

1.4.4 Children's personal data and privacy: Children's rights to personal data and privacy must be given an added layer of statutory protection. Interference with a child's privacy is only permissible if it is neither arbitrary nor unlawful. Any such interference should, therefore, be subject to principles of legality, necessity, and proportionality**,** be designed to observe the best interests of the child, and must not conflict with the provisions, aims, or objectives of the CRC. Legislation should include strong safeguards, transparency, independent oversight, and access to remedy.

1.5 Agency and Self-determination

1.5.1 We recommend that the key principles articulated in the General Data Protection Regulation (GDPR), including the right to access their data, request corrections, object to processing, data portability, and the right to be forgotten (erasure), be incorporated into the parent Act and related primary or subsidiary legislation. This would include the statutory obligation on social media platforms to communicate these rights to users, empowering them to control their data, thereby allowing them to maintain their agency and self-determination.

1.5.2 Consent is the lawful basis for data processing. It must be freely given, specific, informed, and unambiguous. Legislation must require users to provide a clear affirmative action (opt-in) to indicate their agreement and allow them the right to withdraw consent at any time.

1.5.3 Regarding the child's agency, service providers should be guided by General Recommendation 25 of the CRC. Legislation should require service providers to take into account the evolving capacities of the child, including their gradual acquisition of competencies, understanding, and agency, depending on their age and stage of development.

1.6 Due process

1.6.1 The parent Act must impose obligations on service providers to promptly notify data subjects and relevant authorities (ideally within 72 hours) in the event of a data breach that is likely to result in a risk to individuals' rights and freedoms. This ensures transparency in case personal data is compromised.

1.6.2 In this regard, it is recommended that the United Nations Guiding Principles on Business and Human Rights framework (BHR)[2] be adapted into the legislation under the obligations of social media companies to:

[2] Special Rapporteur on FoE, Report of 6 April 2018, A/HRC/38/35, para 11.

(a) take reasonable measures to avoid causing or contributing to adverse human rights impacts and to prevent or mitigate such impacts directly linked to their operations, products, or services by their business relationships, even if they have not contributed to those impacts (principle 13, BHR);

(b) conduct periodic due diligence that identifies, addresses, and accounts for actual and potential human rights impacts of their activities, including through regular risk and impact assessments, meaningful consultation with potentially affected groups and other stakeholders, and appropriate follow-up action (principles 17−19, BHR);

(c) take reasonable prevention and mitigation measures to give effect to internationally recognised human rights principles to the greatest extent possible (principle 23, BHR);

(d) conduct periodic reviews of their efforts to respect rights that include consultation with stakeholders, and frequent, accessible and effective communication with affected groups and the public (principles 20−21, BHR); and

(e) to put in place appropriate, accessible remediation measures for users, including through operational-level grievance mechanisms (principles 22, 29 and 31, BHR).

1.7 Remedy

The parent Act should provide clear and comprehensive pathways for users and victims to access effective remedies in the event harm does occur. Such statutory remedies should include non-judicial grievance mechanisms that can be implemented by public bodies and businesses, and in that regard:

- User-friendly reporting and appeals processes: Users should have a clear, accessible process for reporting online harms, appealing content moderation decisions, and seeking redress if they feel their rights have been violated.
- Legal remedies: Where appropriate, public bodies and businesses can support victims in seeking legal remedies or cooperating with law enforcement to ensure justice for serious cases of online abuse. This includes survivors from communities at risk, such as refugees, undocumented persons, and migrants.

2. Recommendations Specific to the Proposed Penal Code (Amendment) Bill on Cyberbullying

2.1 As stated in our Joint Memorandum - Part 1, creating new anti-bullying laws in the Penal Code is neither necessary nor proportionate.

2.2 Notwithstanding, any laws relating to cyberbullying should (i) clearly distinguish between adults and children in terms of processes, procedures, safeguards, repercussions, and enforcement measures; (ii) clearly define the related harm; and (iii) ensure that the sanctions be necessary and proportionate to the related harm.

2.3 Please refer to **Annex 1** for our detailed responses and recommendations on the proposed Penal Code (Amendment) Bill sections, including our proposals for alternative legislative language.

3. Further Recommendations on the Proposed Online Safety Bill (OSB)

3.1 *Independent Oversight Mechanism:*

We recommend that an independent Commission be established to promote online safety and ensure the administration and enforcement of the legislation. In that regard, the parent Act should include the provisions stated in subparagraphs 3.1.1 to 3.1.4 below:

3.1.1 The powers and functions of the independent Commission should include:

a) Promoting online safety and reduce online harms against all persons in Malaysia, guided by State's treaty obligations, international standards on freedom of expression and fundamental human rights;

b) Formulating guidelines or directives on emerging standards or good practices with respect to online safety;

c) Monitoring, evaluating, and auditing platforms' obligations on safeguards, user protection, access to due process, and remedies, in adherence to human rights principles;

d) Administering an accessible complaint, appeals, and redress system for users;

e) Investigating cases and enforcing decisions where platforms fail to act appropriately, particularly in fulfilling their obligations under the Act. The decisions of the Commission may be subjected to judicial review;

f) Collecting, analysing, interpreting, and disseminating information relating to online safety;

g) Supporting, encouraging, conducting, accrediting, and evaluating educational, promotional, and community awareness programs that are relevant to online safety for all in Malaysia;

h) Coordinating activities of the Malaysian Communications and Multimedia Commission (MCMC), authorities, and other agencies relating to online safety for all in Malaysia, including cross-border collaboration and capacity building.

In exercising its powers and functions, the Commission should ensure meaningful consultation with stakeholders, including children and young people.

3.1.2 The Commission should report to and be accountable to Parliament for its functions and powers under this legislation.

3.1.3 Composition of the Independent Commission:

a) The Commission should consist of five to seven members appointed by a Dewan Rakyat select committee (and duly endorsed by a resolution of the Dewan Rakyat). The Chair and Deputy Chair of the Commission should be selected from among the appointed Commissioners by the Commissioners themselves.

b) Commissioners should be Malaysian citizens, with representatives from social media platforms, civil society stakeholders, relevant experts, and industry stakeholders, and they should have extensive knowledge or practical experience on online safety, law, child rights, or other human rights matters. At any given time, there should be at least three female members of the Commission.

c) The Commissioners should be appointed from a list of candidates selected in accordance with the following procedure:

i. an open call for applications should be published that lays out clear and specific selection criteria;

ii. candidates should be selected from those applying, and due regard should be given to the promotion of equal opportunity, merit, competence, and diversity.

d) Any person who holds public office, whether appointed or elected, who is actively involved in politics, or who is registered with any political party shall not be appointed as a Commissioner.

3.1.4 Term of Office:

a) A Commissioner's term of office should be five years, but every Commissioner should be eligible for reappointment once for another period of five years.

b) A Commissioner may at any time resign from office by letter addressed to the Speaker of the Dewan Rakyat.

3.1.4 While a Commission based on the criteria set out above is our preferred choice of regulatory and oversight body, we are open to discussing other options that are equally independent and empowered.

3.1.5 The Malaysian Communications and Multimedia Commission (MCMC), as the regulatory body may be acceptable as an alternative to an independent commission if the Malaysian Communications and Multimedia Commission Act 1998 and the Communications and Multimedia Act 1998 are amended to ensure that the MCMC's roles and responsibilities as an oversight body under the proposed Online Safety Bill are separated from its current functions under the other two laws, to ensure independence, accountability, transparency, and fairness.

3.2 ***Harmful Content Covered under the OSB***

3.2.1 *Kandungan 1 (Bahan penganiayaan seksual kanak-kanak)*

(a) In addition to the feedback we provided in Part 1 of our Joint Memorandum, we propose that where the perpetrator is a child, legislation should provide for restorative or transformative justice approaches for the child involved whenever possible, in addition to preventive measures and safeguards.

(b) Children with disabilities or those from communities at risk, such as refugees, undocumented, and migrant children, may be more exposed to the risk of sexual exploitation and abuse in the digital environment. In addition to the legislation, regulations providing for the dissemination of safety information and protective strategies relating to the digital environment should be put into place and made available in accessible formats.

3.2.2 *Kandungan 2 (Kandungan seksual)*

(a) We reiterate the focus should be on non-consensual dissemination of intimate images (NCII) rather than sexual or intimate images per se.

(b) We propose the following definition of "intimate image":

> **"intimate image"**, in relation to a person, means any visual representation (including any accompanying sound or document) made by any means including any photographic, film, video or digital representation, including deepfake, —
> (a) of what is, or purports to be the person's genitals, buttocks or anal region and, in the case of a female, her breasts;
> (b) of the underwear covering the person's genitals, buttocks or anal region and, in the case of a female, her breasts;
> (c) in which the person is nude; or
> (d) in which the person is engaged in sexual activity.

(c) The parent Act should include a definition of NCII that incorporates situations in which[3]:

> (i) no consent is given by the person depicted in the intimate image, or there is recklessness as to whether consent is given by the person depicted in the intimate image
>
> (ii) there is a lack or absence of knowledge of the dissemination by the person in the intimate depiction, and/or an invasion or violation of privacy is perceived by the person

[3] Additional references:
https://www.legislation.govt.nz/act/public/2015/0063/latest/whole.html
https://www.cybercrimejournal.com/pdf/YarDrewVol13Issue2IJCC2019.pdf

(iii) the dissemination is done either by persons known or unknown to the person intimately depicted

(iv) the dissemination is done, whether for financial gain or not, to the disseminator

(d) It is important to note that a person may, with consent, share intimate images of themselves for a limited purpose and period, and with specific recipients. In such situations, expectations of privacy must be preserved. Hence, further dissemination of the images outside such bounds of consent ought to be considered a violation of privacy and a form of NCII.

3.2.3 *Kandungan 3 (Kandungan berkaitan dengan penipuan kewangan dalam talian)*

(a) We reiterate that content related to online financial scams should be dealt with separately. Such content is not a priority for this legislation, as it involves deeper topics related to cybercrime and cybersecurity, which require greater involvement with law enforcement in both physical and digital spaces.

(b) We reserve the right to comment on the draft bill once it is shared if we need to address other implications.

3.2.4 *Kandungan 4 (Kandungan yang digunakan untuk membuli)*

(a) We propose the following definition of cyberbullying:

"cyberbullying" means the commission of one or more of the following acts:

(a) communicating threatening, intimidating, humiliating, or derogatory material;
(b) disseminating false or private information; or
(c) creating and sharing any such material or information

through the use of any electronic or digital means or technology where it is reasonable to believe that the act or acts in question could cause severe physical, mental, emotional, or sexual harm to another person.

3.2.5 *Kandungan 5 (Kandungan yang boleh mengapikan* violence *atau* terrorism*)*

(a) Expanding on Part 1 of our Joint Memorandum, we propose that the scope of this type of content be limited to grave forms of violent extremism and terrorism and be clearly defined to prevent overreach or be used as an attempt to stifle legitimate expression. In this regard, it is proposed that the definition in the Canadian Online Harms Bill be adopted.

BOX 2: Extracts from the Canadian Online Harms Bill

Section 2: Definitions

content that incites violent extremism or terrorism means content that actively encourages a person to commit — or that actively threatens the commission of — for a political, religious or ideological purpose, an act of physical violence against a person or an act that causes property damage, with the intention of intimidating or denouncing the public or any section of the public or of compelling a person, government or domestic or international organization to do or to refrain from doing any act, and that, given the context in which it is communicated, could cause a person to commit an act that could cause

(a) serious bodily harm to a person;

(b) a person's life to be endangered; or

(c) a serious risk to the health or safety of the public or any section of the public.

3.2.6 *Kandungan 6 (Kandungan yang berkemungkinan mendorong kanak-kanak untuk menyebabkan kemudaratan kepada diri sendiri (*harm themselves*))*

(a) Expanding on Part 1 of our Joint Memorandum, we propose that the definition of this type of content include content that purports to encourage, promote, or provide instructions that would cause a child to commit or engage in any act of self-injury, disordered eating or dying by suicide.

3.2.7 *Kandungan 7 (Kandungan yang membangkitkan kebencian (hate speech))*

(a) Expanding on Part 1 of our Joint Memorandum, we propose that, in addressing hate speech, the parent Act must expressly meet or incorporate the following principles and conditions:

- We note that the government has not yet offered its definition of hate speech, which it intends to incorporate into the parent Act. While there is no uniform definition of hate speech under international human rights law, we believe the government must apply Article 20(2) of the ICCPR, which states that "any advocacy of national, racial or religious hatred that constitutes incitement to discrimination, hostility or violence shall be prohibited by law."
- A high legal threshold must be applied when prohibiting or circumscribing any speech as hate speech. The high threshold for defining hate speech and setting the obligation to prohibit speech that leads to incitement should incorporate the 3-part test under Article 19(3) ICCPR, i.e., they must be: (a) provided by law, (b) in pursuit of a legitimate aim, including protecting the rights of others, and (c) necessary and proportionate to that aim.
- Further, any statutory restrictions based on hate speech must be in accordance with the Rabat Plan of Action, adopted in 2012, which prohibits

advocacy of national, racial, or religious hatred that constitutes incitement to discrimination, hostility, or violence and includes conclusions and recommendations to combat speech that violates Article 20(2) of the ICCPR. The Rabat Plan of Action proposes the following 6-part test to determine whether a speech falls under Article 20(2):

- *Whether the social and political context is conducive to violence*
- *influence and reach of the speaker*
- *intent of the speaker to incite violence against a target group*
- *content and form of expression*
- *extent of the expression and its dissemination, i.e., severity*
- *likelihood and imminence of violence, discrimination, or hostility occurring as a direct consequence of the expression.*

- In addition, the parent Act must expressly protect specific categories of lawful speech from restriction or prohibition. Expressions or speeches that may be profoundly offensive but do not meet the above criteria must not be restricted. This includes, but is not limited to, blasphemy, expression against the state and public officials, and defamation. The Canadian Online Harms Bill makes a clear distinction in reflecting the extreme nature of hate speech by excluding speech solely because it expresses dislike or disdain or discredits, humiliates, hurts, or offends.

(b) Considering the above, and in line with international human rights standards, we propose that the scope of hate speech covers the content or communication that expresses vilification, detestation, or advocacy of hatred that constitutes incitement to discrimination, hostility, or violence towards individuals or groups on the basis of specific protected characteristics.[4]

(c) If there is a failure to meet the conditions set out under (a) above, it is proposed that 'hate speech' content be excluded from the Online Safety Bill.

3.3 ***Duties under the OSB***

Duties under the OSB should adopt best practice standards that focus on systems, policies, and processes that enable digital platforms to balance safety responsibly, privacy, freedom of expression, and other fundamental values. This provides an incentive for digital platforms to invest in policies, products, tools, and programs that empower users to make informed decisions and have control over their experiences and interactions online. It also provides greater flexibility for digital platforms to respond and adapt quickly and appropriately to ever-changing risks of online harm.

[4] Under international human rights standards, protected characteristics include (among others) characteristics relating to race, religion, culture, ethnicity, national origin, gender, age, marital status, socioeconomic status, political persuasion, educational background, geographic location, sexual orientation or physical or mental ability.

3.3.1 Duty to Act Responsibly

We wish to put forward the following recommendations:

(a) Online service providers must ensure that adequate measures are adopted to reduce the risk and prevalence of harmful content online; specifically, they must:

i. Provide safeguards to reduce the risk of harm arising from online child sexual exploitation and abuse (CSEA), including implementing, enforcing, and/or maintaining policies, processes, products, and/or programmes that seek to:

- prevent known child sexual abuse material (CSAM) from being made available to users or accessible on their platforms and services
- prevent search results from surfacing child sexual abuse material
- adopt enhanced safety measures to protect children online from peers or adults seeking to engage in harmful sexual activity with children (e.g., online grooming and predatory behaviour).

ii. Implement, maintain and raise awareness of product or service-related policies and tools for users to report cyberbullying, NCII, or other defined harmful content covered under the parent Act.

iii. Strengthen social media platforms' due diligence obligations, with a specific focus on mandatory human rights impact assessments, and ensure that risk assessment and mitigation measures respect necessity and proportionality requirements.

iv. Label harmful content when communicated in multiple instances or artificially amplified through automated communications by computer programmes. This requirement would include harmful content shared widely by bots or bot networks.

v. Submit a digital safety plan to the independent Commission, including:

- An assessment of the risks that the users will be exposed to in relation to harmful content;

- A description of measures that the service providers implement to mitigate the risk;

- A description of the indicators used to measure its effectiveness as well as an assessment of the effectiveness of the measures in mitigating the risk;

- Information related to design features that are integrated into services to mitigate the risks;

- Information related to resources, including human resources, allocated to prevent or mitigate the risk; and

- Ensuring that the plan does not contain personal information.

vi. Publish the digital safety plan and make it available in an accessible and easy-to-read format in the multiple languages used in Malaysia.

vii. Ensure that the measures implemented do not unreasonably or disproportionately limit users' right to expression on the regulated platform.

(b) Online service providers must adopt measures to empower users to have more control and make informed choices through:

i. Ensuring that users are empowered to have control or make informed decisions about the content they see on the platform and/or their experiences and interactions online by:
- Implementing, enforcing, and/or maintaining policies, processes, products, and/or programmes that seek to provide users with appropriate control over the content they see, the character of their feed and/or their community online.
- Launching and maintaining products that provide users with controls over the appropriateness of the ads they see.

ii. Making available to users tools to:
- Block other users from finding or communicating with them on the service. This includes giving them the option of filtering out unverified users, which will help stop anonymous trolls from contacting them;
- Help users reduce the likelihood of encountering certain types of harmful content. These categories of content are to be set out in the parent Act and must be effective and easy to access and
- Enable users to flag harmful content

iii. Notify users who have flagged harmful content of the fact that it was flagged and of any measures undertaken

iv. Establishing a robust remedial mechanism for users of the service, including the following:

- Make available a child-friendly reporting mechanism, including a single unified hotline, taking into account gender, language, disability, etc, so that child victims or children-at-risk know where to report and can report without fear;
- make a 'human' content moderator who is easily identifiable and accessible, available to users of the service;
- hear users' complaints or concerns with respect to harmful content in multiple languages used in Malaysia;

- direct users to internal and external resources to address their complaints or concerns, including the internal complaints mechanism or the independent Commission;
- take a balanced approach to digital due process, including providing notice and an opportunity to object and the right to be heard (counter-notice), except in narrow, exceptional circumstances – such as child sexual abuse material (CSAM) – to users whose content has been flagged; and
- periodic audits or reviews are required to ensure that particular types of users (especially communities at risk or vulnerable groups) or content categories are not unduly impacted and that the content remains relevant and consistent with evolving norms and technologies.

v. Ensure measures are in place to ensure the use of the algorithm and data is fit for purpose by (a) disclosing its use, (b) studying the risk and understanding its limitations, and (c) identifying and managing bias. In this regard, ensure that privacy, ethics, and human rights are safeguarded by regularly peer-reviewing algorithms to assess for unintended consequences and act on this information.

vi. Harm can also arise from the way content is disseminated, such as when an algorithm repeatedly pushes content to a child in large volumes over a short space of time. Providers will then need to mitigate and effectively manage any identified risks. This includes considering their platform's design, functionalities, algorithms, and other features likely to meet the illegal content and child safety duties.

vii. Subject to exceptional circumstances involving child sexual exploitation and abuse, which may require otherwise, online service providers must, in no later than one year, uphold the duty to destroy content, data and documents that were retained as part of its due process related to risk assessment, digital safety plan and other measures.

(c) Online service providers must enhance the transparency of policies, processes, and systems

i. Online service providers must ensure transparency of their safety and harm-related policies, systems, processes, products, tools, and measures that aim to reduce the risk of online harm by publishing and making it accessible to users. This includes disclosures on content curation methods and processes and deployment of Artificial Intelligence (AI) where applicable, such as deciding which content should be presented to users (in terms of frequency, order, priority, discoverability, and so on) based on the platform's design. These should be made available in multiple languages and accessible in an age-, gender- and disability-appropriate manner.

ii. Transparency is to be maintained by clearly explaining how decisions are informed by algorithms, such as promotion, demotion, and other forms of ranking of content and content curation using algorithmic recommendation systems aimed at maximising users' engagement. The

duties include (a) providing plain English and BM documentation of the algorithm; (b) making information about the data and processes available (unless a lawful restriction prevents this disclosure); and (c) publishing information about how data are collected, secured, stored and used.

iii. Online service providers must also specifically consider and publish their assessment on how algorithms could impact users' exposure to harmful content – and children's exposure to content that is harmful to them – as part of their risk assessments. Thus, annual transparency reports should also contain relevant information such as information on the algorithms used and their effect on users' experience, including children.

iv. Online service providers must publish annual transparency reports on:

- Measures undertaken to reduce the spread and prevalence of harmful content;
- Annual compliance with government requests for content take-down and measures for complaints and redress;
- Information related to the volume and type of harmful content that was accessible over its services, including:

    - volume and type of harmful content that was moderated;
    - manner in which the harmful content was moderated;
    - timeframe within which the content was moderated;
    - number of times the contents were flagged by users as being harmful content;
    - manner in which the platform triaged and assessed the flags; and
    - measures undertaken to address flagged content.

3.3.2 Duty to Protect Children

We wish to put forward the following recommendations:

(a) Online service providers must ensure that the digital space is safer for children through child safeguarding policies and child protection procedures. Design features must be implemented in the best interest of children. This would include technical approaches, tools and services for parents and children, such as:
   i. Age- and gender-appropriate, safety by design features
   ii. Parental control tools or family safety settings that place the safety and rights of the child at the centre of the design
   iii. Age-differentiated experiences with password-protected content
   iv. Block/allow lists
   v. Purchase/time controls
   vi. Default settings related to warning labels for children. Online service providers are to apply concise and intelligible content labelling, for example on the age-appropriateness or trustworthiness of content.
   vii. Filtering and safe search settings for the service's internal search function

viii. Design features to limit children's exposure to harmful content, including explicit adult content, cyberbullying content and content that incites self-harm.

(b) Online service providers must issue customer terms and conditions and/or acceptable use policies to explicitly state their position on the misuse of their services to store or share child sexual abuse material and the consequences of any abuse.

(c) Where possible, online service providers must consider the use of age verification to limit access to content or material that, either by law or policy, is intended only for persons above a certain age. At the same time, online service providers should recognize the potential for misuse of such technologies in ways that could restrict children's right to freedom of expression and access to information.

(d) Online service providers must clearly describe available content and corresponding parental controls or family safety settings. This includes making language and terminology accessible, visible, clear and relevant for all users – including children, parents and caregivers – especially in relation to terms and conditions, costs involved in using content or services, privacy policies, safety information and reporting mechanisms.

(e) Online service providers must adopt the highest privacy standards when it comes to collecting, processing, storing, sale and publishing of personal data, including location-related information and browsing habits, gathered from persons under 18. Default privacy settings and information about the importance of privacy should be appropriate to the age of the users and the nature of the service.

(f) For services directed at or likely to attract a main audience of children, online service providers must consider the risks posed to children by access to, or collection and use of, personal information (including location information), and ensure those risks are properly addressed. In order to aid understanding and assist users in managing their privacy, online service providers should ensure that any materials or communications used to promote services, provide access to services, or by which personal information is accessed are age-, gender-, and linguistically appropriate, and that the language and style of such materials or communications are clear and simple.

(g) In line with 3.3.1 (b) (iv), child-friendly reporting mechanisms, including a single unified hotline, must be made available to child users who have concerns about content and behaviour. Furthermore, reporting needs to be followed up appropriately, with timely provision of information about the status of the report. Although online service providers can vary their implementation of follow-up mechanisms on a case-by-case basis, it is essential to set a clear time frame for responses, communicate the decision made regarding the report, and offer a method for following up if the user is not satisfied with the response.

(h) Online service providers must integrate due diligence on child online protection issues into existing human rights or risk assessment frameworks (e.g., at the corporate level, product or technology level) to determine whether the online service may be causing or contributing to adverse impacts through its own activities, or whether adverse impacts may be directly linked to its operations, products or services or business relationships.

(i) Online service providers must educate users on how to manage concerns relating to Internet usage – including spam, data theft and inappropriate contact such as bullying and grooming – and describe what actions users can take and how they can raise concerns on inappropriate use.

(j) In exceptional circumstances, specifically involving child sexual abuse and exploitation, online service providers are to ensure that specific safety-by-design measures (including hash scanning) are extended to private messaging services, without undermining safety features of private messaging services such as anonymity and end-to-end encryptions. At the same time, online service providers should guard against potential misuse of such measures in ways that could restrict children's rights to freedom of expression and access to information.

3.3.3 Duty to Make Priority Harmful Content Inaccessible

(a) While the government should protect children from harmful and untrustworthy content in accordance with their rights and evolving capacities, it should also recognise children's rights to information and freedom of expression. Hence, any restrictions on the operation of any Internet-based, electronic or other information dissemination systems should be in line with Article 13 of the CRC. Laws should not intentionally obstruct or enable other actors to obstruct the supply of electricity, cellular networks or Internet connectivity in any geographical area, whether in part or as a whole, which can have the effect of hindering a child's access to information and communication. it should also recognize children's rights to information and freedom of expression.

(b) Expanding on Part 1 of our Joint Memorandum, we reiterate that this duty would require online service providers to make only two specific categories of harmful content inaccessible to their users: (1) child sexual abuse material (CSAM); and (2) non-consensual intimate images (NCII), including sexualised deepfakes.

(c) In this regard, it is proposed that the model introduced in the Canadian Online Harms Bill be adopted.

3.3.4 Insert Additional Duty: Duty to Promote Freedom of Expression and Opinion

(a) Any efforts to address safety and harmful content online should respect freedom of expression and opinion and other fundamental human rights. The parent Act, while

targeting certain types of harmful content for heavy censorship, must aim to strike a balance between risk mitigation and respecting freedom of expression.

(b) As such, except for the narrow but critical instances of CSAM and NCII, the parent Act should not require mandatory proactive steps on the part of covered social media platforms to identify, manage, and remove harmful content on their services.

4. Conclusion

In conclusion, while the proposed Online Safety Bill and proposed amendments on cyberbullying are commendable in their efforts to protect children and society at large from online harm, it is crucial that these measures are carefully balanced with fundamental human rights, particularly the right to freedom of speech and expression, the right to personal liberty and privacy, and the safeguarding of children within the criminal justice system.

The Online Safety Advocacy Group urges the government to ensure that any regulatory framework is effective in safeguarding children and sensitive to potential overreach that could stifle public discourse, infringe on privacy, or limit access to information. Achieving this balance is essential for a safe and open digital environment where children and society are protected while fundamental freedoms are upheld.

Date: 11 November 2024

ANNEX 1: RESPONSES AND RECOMMENDATIONS RELATED TO THE PROPOSED PENAL CODE (AMENDMENT) BILL

No.	New Section	Government's proposed new offence with CSOs' proposed changes tracked	Government's proposed wording with CSOs' proposed changes tracked	Responses, recommendations and rationale by CSOs
1.	507B	Act done with intent to cause or recklessness as to whether harassment, ~~alarm,~~ fear ~~or~~, distress or invasion of privacy is caused	Whoever with intent to cause harassment, ~~alarm~~, fear, ~~or~~ distress or invasion of privacy to any person or being reckless as to whether or not harassment, fear, distress or invasion of privacy is caused, by any means— (a) uses any threatening, abusive or ~~insulting~~ degrading words or behaviour; or (b) makes any threatening, abusive or ~~insulting~~ degrading communication, and as a result causes the person harassment, ~~alarm,~~ fear ~~or~~ distress or invasion of privacy, shall be punished with imprisonment for a term which may extend to five years or with fine or with both.	1. The phrases "Whoever" and "any person" are too wide and could unintentionally criminalise thousands of children, based on the current prevalence of cyberbullying. As the intent of the proposed Bill is to protect children from cyber aggression and threats, censorship, data breaches and digital surveillance, it is important to ensure that children will not be caught under this section. Children should not be prosecuted for expressing their opinions in the digital environment, unless they violate restrictions provided by criminal legislation which are compatible with Article 13 of the UN Convention on the Rights of the Child (CRC). 2. The word "alarm" should be deleted and/or replaced as its meaning is unclear. 3. We propose to include the phrase "invasion of privacy" to cover violations of privacy, and the word "insulting" to be replaced with "degrading" as it connotes a more severe degree of harm. 4. Instead of a separate Section 507C covering offences done without criminal intent, we propose that this section also apply to cover instances, where despite the lack of criminal intent, the person was reckless. To this end, we propose to include the phrase "or being reckless as to whether or not

			Explanation Repetition, sharing, circulation, forwarding or dissemination of such threatening, abusive or degrading words, behaviour or communication is considered as using words, behaviour or communication as described under this section.	xxx is caused". The language on recklessness is also found in the UK Online Safety Act 2023, New Zealand's Harmful Digital Communications Act 2015 and Ireland's Harassment, Harmful Communications and Related Offences Act 2020. 5. The person who commits this offence need not be the primary and/or first person who uses the offensive words, behaviour or communication. Repetition, sharing, circulation, forwarding or dissemination should also be considered as an offence where there is intent or recklessness. 6. Additionally, we propose that an exemption clause and/or an explanatory statement be included to distinguish children from adults in terms of processes, procedures, safeguards, repercussions, and enforcement measures. 7. There should also be an explanatory statement that sets out the intent of the proposed amendments to the Penal Code to address "risks relating to content, contact, conduct that encompass cyber aggression and harassment, and the promotion of or incitement to suicide or life-threatening activities posed by adult offenders upon victims, including child victims".
2.	~~507C~~	~~Act done without criminal intent to cause harassment, alarm, fear or distress~~	~~Whoever by any means –~~ ~~(a) uses any threatening, abusive or insulting words or behaviour; or~~ ~~(b) makes any threatening, abusive or insulting communication,~~	1. Section 507C is similar to Section 507B except that it covers acts done without criminal intent (although the result to the victim may be indicative of intent, in which case, section 507B would be more appropriate). We propose that a section that provides for a lack of criminal intent should not be included in the Penal Code as the Penal Code is highly punitive in nature.

			~~which is heard, seen or otherwise perceived by any person likely to be caused harassment, alarm, fear or distress shall be punished with imprisonment for a term which may extend to three years or with fine or with both.~~ ~~(2) In any proceedings for an offence under subsection (1), it is a defence for the accused to prove —~~ ~~(a) that the accused had no reason to believe that the words or behaviour used, or the communication made by the accused would be heard, seen or otherwise perceived by any person; or~~ ~~(b) that the accused's conduct was reasonable and would not result in causing harassment, alarm, fear or distress to any person.~~	2. Hence, we recommend that the proposed Section 507C be deleted in entirety.
3.	507D	Harassment, ~~alarm,~~ fear or distress to any person by any means by publishing any identity information	Whoever by any means publishes any identity information of a person with the intent to cause, or being reckless as to whether it causes harassment, ~~alarm,~~ fear or distress to any person shall be punished with imprisonment for a term which may extend to five years or with fine or with both. Explanation Repetition, sharing, circulation, forwarding or dissemination of such	1. We propose deleting the word "alarm," as suggested in our feedback on Section 507B above. 2. We also propose to include the phrase "or being reckless as to whether or not xxx is caused" to cover instances where despite the lack of criminal intent, the person was reckless. The language on recklessness is not new, and we have listed some examples of legislation where this has been incorporated in our feedback on Section 507B above. 3. The person who commits this offence need not be the primary and/or first person who published the identity information. Repetition, sharing, circulation, forwarding, or dissemination should also be considered a publication, and

			identity information is considered as an offence under this section.	such person should be considered to have committed this offence where there is intent or recklessness. 4. Importantly, Sections 507B and 507D should also cover instances where such acts encourage, promote, or provide instructions for an act of deliberate self-injury and suicide.
4.	507E	Publication of any identity information with criminal intent to cause violence against any person	Whoever by any means publishes any ~~identity information~~ personally identifiable information of any person with the intent, or being reckless as— (a) to cause the person to believe that violence will be used against any person; ~~or~~ (b) to facilitate the use of violence against any person~~,~~; or (c) to encourage, promote or provide instructions for an act of self-injury, disordered eating, or dying by suicide; shall be punished with imprisonment for a term which may extend to seven years or with fine or with both. Explanation Repetition, sharing, circulation, forwarding or dissemination of such identity information is considered as an offence under this section..	1. The phrase "identity information" indicates that the information in the plain sense contains a person's identity. We propose that the phrase be amended to "personal identifiable information" and/or to define it to mean "any information relating to an identified or identifiable person including without limitation by reference to a name, an identification number, location data, an online identifier or to one or more factors specific to the physical, physiological, genetic, mental, economic, cultural or social identity of that person"[5]. 2. We have similarly included the phrase "or being reckless as", as suggested for earlier sections above. 3. The person who commits this offence need not be the primary and/or first person who publishes the identity information. Repetition, sharing, circulation, forwarding, or dissemination should also be considered a publication, and such person should be considered to have committed this offence where there is intent or recklessness. 4. As suggested for previous sections, Section 507E should also cover instances where such acts encourage, promote, or provide instructions for an act of self-injury, disordered eating or dying by suicide.

[5] Borrowed from definition of "personal data" under the EU General Data Protection Regulation

5.	~~507F~~	~~Publication of any identity information without criminal intent to cause violence against any person~~	~~Whoever by any means publishes any identity information of any person with knowledge or having reasonable cause to believe that it is likely —~~ ~~(a) to cause the person to believe that violence will be used against any person; or~~ ~~(b) to facilitate the use of violence against any person,~~ ~~shall be punished with imprisonment for a term which may extend to five years or with fine or with both~~.	3. Section 507F is similar to Section 507E, except that it covers acts done without criminal intent (although the result to the victim may be indicative of intent, in which case, Section 507E would be more appropriate). We propose that a section that provides for a lack of criminal intent should not be included in the Penal Code as the Penal Code is highly punitive in nature. 4. Hence, we recommend that the proposed Section 507F be deleted in entirety.
6.	507G	Definition of 'identity information'	For the purposes of sections 507D, 507E and 507F— "~~identity information~~" "personally identifiable information" means any information that, whether on its own or or in conjunction with other information, identifies or purports to identify an individual, including (but not limited to) any of the following: (a) the individual's name, residential address, email address, telephone number, date of birth, national registration identity card number identification number, passport number, signature (whether handwritten or electronic) or password; (b) any photograph or video recording	1. We propose that the phrase "identity information" be replaced with the phrase "personally identifiable information" as it is more consistent with the proposed definition. 2. The definition should also cover other personally identifiable information used online. See also our feedback on Section 507E on the definition.

			of the individual; (c) any information about the individual's family, employment or education; and <u>(d) any social media user identification, username, alias or other digital identifier uniquely associated with the individual on any electronic, digital, or social media platform.</u>	
7.	507H	Fear, provocation or facilitation of violence	*Whoever by any means uses towards any person any threatening, abusive or insulting words or behaviour, or makes any threatening, abusive or insulting communication to any person either—* *(a) with the intent—* *(i) to cause the person believe that violence will be used against any person; or* *(ii) to provoke the use of violence by any person against any person; or* *(b) where—* *(i) the victim is likely to believe that such violence mentioned in paragraph (a)(i) will be used; or* *(ii) it is likely that such violence mentioned in paragraph (a)(ii) will be provoked,*	1. This Section 507H seems to be overlapping with Section 507E. We propose a further discussion in order to understand the intent. 2. Alternatively, we propose that the phrase "provocation or facilitation of violence" be included in all the preceding Sections 507 B-F and that this section be removed in entirety. 3. We also propose that language in the preceding sections include acts and/or content which encourage, promote, or provide instructions for an act of self-injury, disordered eating or dying by suicide.

			shall be punished with imprisonment for a term which may extend to five years or with fine or with both.	