1. INTRODUCTION

The Human Rights Commission of Malaysia (SUHAKAM) welcomes the invitation from the National AI Office (NAIO) to participate in the "*Bengkel Konsultasi Pihak Berkepentingan Sektor Awam bagi Cadangan Rang Undang-Undang Tadbir Urus AI*" and the opportunity to comment on the framework of the AI Governance Bill (the Bill) prepared for public consultation.

SUHAKAM is established under the Human Rights Commission of Malaysia Act 1999 [Act 597] and operates in accordance with the Paris Principles as an A-status National Human Rights Institution (NHRI). Among SUHAKAM's statutory functions are to advise and assist the Government in formulating legislation and policies and to recommend to the Government the ratification of international human rights instruments. Regulation of AI falls squarely within this mandate given the significant and well-documented implications of AI for civil, political, economic, social and cultural rights. As AI systems are increasingly deployed in both the public and private sectors, the Bill presents an important opportunity to ensure that technological innovation develops in a manner that respects, protects and fulfils human rights while fostering public trust and accountability.

SUHAKAM commends NAIO for undertaking early and iterative public consultation. This submission is intended to be constructive; it seeks to identify areas where the proposed framework aligns with internationally recognised human rights principles as well as areas that, in SUHAKAM's view, require further strengthening to ensure that human rights considerations are embedded throughout the Bill rather than addressed as standalone safeguards. The comments below are organised as follows:

i) Definitions and Scope of the Bill
ii) The Central AI Authority
iii) AI Principles
iv) AI Risk and Harm Framework
v) AI Incident Reporting
vi) AI Enabling Powers

vii) AI Sandbox Function

SUHAKAM's comments are grounded in the Federal Constitution and informed by Malaysia's international human rights obligations and internationally recognised standards on AI governance. While not all of the instruments are legally binding on Malaysia, they provide persuasive international guidance and reflect an emerging global consensus on rights-respecting AI governance.

2. DEFINITIONS AND SCOPE OF THE BILL

2.1 Definitions of Artificial Intelligence (AI), AI System and AI Lifecycle

In principle, the definitions of AI and AI Systems are appropriately technology-neutral and intended to be future-proof which SUHAKAM welcomes given the rapid pace of AI development.

AI:

SUHAKAM notes the absence of a universally accepted definition of AI. As presented during the stakeholder consultation, the Bill focuses on what the system does rather than only how it is built. Accordingly, the definition is anchored in cognitive functions typically associated with human intelligence, including learning, reasoning, inference, prediction, perception, language processing, recommendation, content generation, decision support and output generation. On this approach, a system may still fall within the definition of AI whether it performs below, at or beyond human capability.

SUHAKAM supports this functional and technology-neutral approach. However, from a human rights perspective, the definition should also ensure that the Bill remains capable of regulating future AI systems that may not necessarily emulate human cognition, nevertheless generate outputs capable of materially affecting individuals' rights, interests or opportunities. A definition that is overly tied to present-day technological concepts may unintentionally create regulatory gaps as AI technologies continue to evolve.

SUHAKAM recommends clarifying that the key characteristic of AI is not just its ability to simulate human thinking, but also its capacity to produce outputs that can influence decisions, predictions, recommendations, or content in ways that may significantly impact individuals or communities.

AI systems:

SUHAKAM supports the proposed approach that a system should be regarded as an AI system where it incorporates or features AI. This provides an appropriately broad and future-proof basis for determining the scope of the Bill and reduces the risk that regulatory obligations may be avoided through technical distinctions regarding system architecture.

AI lifecycle:

The Bill recognises that risks arise throughout the lifecycle rather than only at the point of deployment. SUHAKAM welcomes this lifecycle approach. However, the current definition should expressly recognise that many human rights risks arise before an AI system is developed or placed into operation, particularly during data sourcing, collection, annotation, labelling, model training, validation and testing. The ASEAN Guide on AI Governance and Ethics similarly recognises that AI governance begins well before model development, including project governance and problem statement definition[1] as well as data collection and processing[2], while emphasising that privacy-by-design should be embedded throughout every stage of the AI lifecycle. This reflects the understanding that many AI-related risks arise long before deployment. Discriminatory datasets, the exclusion of minority groups, inaccessible training data and biased annotation processes may all give rise to systemic discrimination long before an AI system is developed and placed into operation. Accordingly, a genuinely preventive approach to AI governance requires these early stages to fall within the statutory definition of the AI lifecycle.

International human rights standards reinforce this approach. Convention on the Rights of the Child (CRC) General Comment No. 25 (2021) on Children's Rights in relation to the Digital Environment (General Comment No. 25) requires States to mandate child rights impact assessments and businesses to undertake child rights due diligence for digital systems affecting children.[3] Article 9 of the Convention on the Rights of Persons with Disabilities (CRPD) requires that information and communication technologies be accessible[4], highlighting that inaccessible training data and annotation processes may themselves become a source of downstream discrimination against persons with disabilities (PWDs). Similarly, the UN General Assembly Report of the Working Group on Discrimination against Women and Girls on Women's and Girls' Rights and Artificial Intelligence and related Digital Technologies (Women's and Girls' Rights and Artificial Intelligence and related Digital Technologies) highlights that biased datasets, gender-insensitive design and the underrepresentation of women in AI development can reinforce structural discrimination and produce unequal outcomes for women and girls throughout the AI lifecycle. These standards demonstrate that many AI-related human rights risks originate well before deployment, therefore, support a lifecycle definition that extends to the earliest stages of AI deployment.

[1] Pages 27 – 29.
[2] Pages 27, 29 and 30.
[3] Paragraphs 23 and 38.
[4] Article 4 of CRPD.

In response to whether a common definition across the industry is necessary, SUHAKAM acknowledges that defined terms may require variance to adapt to sector-specific contexts. Nonetheless, aligned definitions of AI-related terms across sectors ensure some form of legal certainty as highlighted in the European Union Artificial Intelligence Act (EU AI Act).[5] Conversely, varying or overly vague definitions may create regulatory fragmentation in the markets, undermine legal certainty for developers, deployers, users or other AI actors and hinder consistent implementation and enforcement. Similarly, the UN Report on Governing AI for Humanity recommends maintaining a register of AI-related definitions to promote a common understanding between industry players and users[6] thereby supporting greater regulatory coherence and interoperability.

Recommendations:

- Ensure that the statutory definitions remain sufficiently technology-neutral and impact-based so that AI systems capable of materially affecting individuals' rights, interests or opportunities remain within the scope of the Bill regardless of their underlying technical architecture.
- Align the statutory definitions, where appropriate, with internationally recognised functional definitions while preserving sufficient flexibility to accommodate future technological developments.
- Test the final definitions against concrete rights-relevant use cases, including algorithmic recruitment, biometric identification, predictive policing, immigration decision-making, healthcare diagnostics and automated eligibility determinations for public services.
- Expand the definition of the AI lifecycle to expressly include data sourcing, collection, labelling, annotation, training, validation and testing, recognising that many forms of discrimination and other human rights harms originate during these stages, consistent with CEDAW, CRC and CRPD.
- Require high-risk AI systems to undertake human rights impact assessments and bias and discrimination auditing of training data as lifecycle obligations, ensuring that the Bill's preventive intent of the lifecycle approach is operationalised from the earliest stages of AI development rather than only after development and deployment.

[5] Section 3 of the EU AI Act.
[6] Page 13.

2.2 Developer (Provider) and Deployer (Adopter)

As presented during the consultation, AI has no legal personality. As such, legal responsibility must rest with identifiable natural or legal persons.

SUHAKAM supports the distinction between developers and deployers. However, from a human rights perspective, effective accountability depends not merely on distinguishing between these actors but also on allocating responsibility throughout the AI value chain according to each actor's degree of control over, contribution to or ability to prevent the harm. This approach is consistent with the UN Guiding Principles on Business and Human Rights (UNGPs) which recognise that responsibility for adverse human rights impacts may be shared among multiple actors depending on their respective roles and contributions.

Both the Doha Declaration on Artificial Intelligence and Human Rights (Doha Declaration) and the GANHRI Annual Conference 2026 Outcome Statement: The Role of National Human Rights Institutions in Promoting and Protecting Human Rights in the Digital Space (GANHRI Outcome Statement) go further than the UNGPs' general due diligence standard by calling for mandatory human rights due diligence, including human rights impact assessment, particularly in relation to high-risk AI systems. Similarly, the ASEAN Guide on AI Governance and Ethics recommends that deployers undertake risk-based assessments of AI systems before commencing any data collection, processing or modelling stages.[7] The EU AI Act likewise requires certain high-risk AI systems to undergo a fundamental rights impact assessment.[8] The prominence of human rights and risk assessment obligations across these international instruments demonstrates that effective accountability requires more than assigning legal responsibility after harm occurs; it also requires proactive identification, assessment and mitigation of foreseeable risks throughout the AI lifecycle. The UNICEF Guidance on AI and Children similarly emphasises the need for child rights impact assessments on AI-enabled systems.[9] Accordingly, SUHAKAM recommends that these more robust standards inform the allocation of developer and deployer obligations under the Bill.

SUHAKAM also notes a structural gap not addressed in the current proposal. Where harm results from the interaction between a developer's design choices and a deployer's implementation decisions, the current framework may create uncertainty regarding accountability with each actor potentially

[7] Page 29.
[8] Article 27 of the EU AI Act.
[9] Page 15.

attributing responsibility to the other. This accountability gap is particularly significant where AI systems continue to influence decisions after deployment, making effective oversight essential throughout the operational lifecycle of the system. The EU Independent High-Level Expert Group on Artificial Intelligence acknowledges the need for meaningful human oversight in AI Systems via human-in-the-loop (HITL), human-on-the-loop (HOTL) or human-in-command (HIC) approaches[10], depending on the nature and level of risk posed by the AI system. These governance models help ensure that responsibility remains with identifiable human actors and that AI systems do not operate without meaningful human accountability. Without a mechanism for shared accountability and appropriate human oversight, this risks weakening both regulatory enforcement and affected individuals' access to effective remedy.

Recommendations:

- Clearly define the legal responsibilities across the AI value chain, including developers, deployers and other relevant actors where responsibilities are shared and human oversight is required.
- Require developers and deployers to undertake proportionate human rights due diligence, including human rights impact assessments for high-risk AI systems before development and deployment, consistent with the Doha Declaration, the GANHRI Outcome Statement and international good practice.
- Require developers and deployers to understand the intended purpose, limitations and foreseeable risks of AI systems before development and deployment.
- Require deployers to maintain sufficient documentation regarding the source, functionality and deployment of AI systems to facilitate accountability, regulatory oversight and access to effective remedy.
- Require child rights impact assessments where AI systems are reasonably likely to affect children or process children's data, consistent with the UNICEF Guidance on AI and Children and CRC General Comment No. 25.
- Require AI systems to incorporate appropriate human oversight measures proportionate to the level of risk, ensuring that meaningful human intervention remains possible throughout the AI lifecycle.
- Introduce a joint-responsibility mechanism so that, where harm results from the interaction of a developer's design choices and a deployer's implementation choices, both actors can be held accountable rather than each being able to point to the other.

[10] Page 16.

2.3 Scope of the Bill

SUHAKAM supports the proposed scope covering AI systems placed on the market, designed, developed or used in Malaysia as well as systems used by a deployer established in Malaysia regardless of where the system is hosted. This provides a relatively comprehensive jurisdictional basis and appropriately recognises the cross-border nature of AI development and deployment.

Nevertheless, SUHAKAM considers that three aspects of the proposed scope warrant further clarification from a human rights perspective.

First, the Bill does not expressly provide that it applies to AI systems designed, developed, procured or deployed by public authorities. Given the increasing use of AI across public administration, the Bill should expressly clarify that such systems fall within its scope. Public authorities increasingly rely on AI in Government functions, including policing, immigration, criminal justice, healthcare, education, social protection and other public functions that directly affect the exercise of constitutional rights. In light of the significant powers exercised by the State, AI systems deployed by public authorities should be subject to at least the same, and where appropriate higher standards of legality, transparency, accountability and independent oversight than those applicable to private-sector actors. The GANHRI Outcome Statement gives this concrete content, specifically identifying digital identity systems, integrated public service platforms, civil registration systems and interoperable databases as public-sector AI applications warranting heightened legal frameworks, ex ante safeguards, meaningful human oversight and independent scrutiny.[11]

Secondly, extraterritorial jurisdiction and extraterritorial enforcement were discussed during the consultation. Given that many high-impact AI systems, including foundation models are developed and hosted outside Malaysia, effective extraterritorial jurisdiction is essential to ensure that individuals within Malaysia are not deprived of legal protection or effective remedies merely because the responsible developer or provider is located overseas. Accordingly, the Bill should establish or be supported by practical mechanisms for international cooperation, regulatory coordination and cross-border enforcement to ensure that its extraterritorial provisions can be implemented effectively. This would be consistent with the Doha Declaration's call for international cooperation on AI governance.[12]

[11] Paragraph 16 of the GANHRI Outcome Statement.
[12] Pages 2 and 4 of the Doha Declaration.

Thirdly, SUHAKAM notes that the proposed scope focuses primarily on the human rights impacts arising from the development and deployment of AI systems, however, it does not expressly recognise the broader environmental impacts associated with the AI ecosystem. While AI regulatory frameworks in more than 100 countries have largely focused on issues such as privacy, bias and discrimination, misinformation and disinformation, safety, security and cybersecurity, comparatively less attention has been given to the environmental impacts of AI. In reality, the AI ecosystem can have significant environmental implications throughout its lifecycle, including greenhouse gas emissions, high energy and water consumption, electronic waste, pollution and biodiversity loss. These impacts arise from, among others, the manufacturing of AI chips and graphics processing units (GPUs), the extraction of critical minerals and rare earth elements, the construction and operation of data centres and the training and deployment of increasingly complex AI models. Recognising these broader impacts, the EU AI Act became the world's first comprehensive AI regulatory framework to expressly recognise certain environmental considerations associated with AI systems.[13] While environmental regulation may extend beyond the primary objectives of this Bill, SUHAKAM considers that adopting a broader AI ecosystem perspective would better support responsible and sustainable AI governance in Malaysia.

Recommendations:

- Expressly provide that the Bill applies to AI systems designed, developed, procured or deployed by public authorities.
- Require enhanced safeguards for AI systems used by public authorities where decisions may significantly affect fundamental rights.
- Strengthen the Bill's extraterritorial jurisdiction and enforcement mechanisms to ensure that affected individuals have meaningful access to accountability and remedy irrespective of where the AI system originates.
- Consider adopting a broader AI ecosystem approach that recognises the interrelationship between (i) AI applications across industries, (ii) AI models and platforms, (iii) AI infrastructure, (iv) AI accelerators and specialised hardware and (v) AI computing infrastructure[14] while taking into account their associated human rights and environmental impacts.

[13] Available at <https://theconversation.com/ai-laws-overlook-environmental-damage-heres-what-needs-to-change-279047>

[14] Available at <https://news.skhynix.com/en/exploring-the-ai-ecosystem-ai-value-chain/>

2.4 Exemptions

2.4.1 Personal Use Exemption

SUHAKAM supports the position that the personal use exemption should not apply where an individual's use of AI poses significant risks to the rights or interests of others or where such use becomes integrated into commercial or public activities. However, the concepts of a 'threshold test' and a 'sensitive domain' should be clearly defined within the Bill or accompanying guidance. Without clear statutory definitions or objective criteria for their application, the scope of the exemption may be interpreted inconsistently, thereby creating legal uncertainty for regulators, developers, deployers and affected individuals alike.

2.4.2 National Security Exemption

SUHAKAM notes with concern that the proposed exemption for AI systems ‘used solely for national defence or security’ is drafted in broad terms without accompanying safeguards, oversight or review mechanisms. The Bill also does not define the terms ‘national defence’ or ‘national security’, creating uncertainty regarding the scope of the exemption and the circumstances in which it may properly be invoked.

SUHAKAM's central position is that a blanket exemption of this kind is not in line with human rights standards. This conclusion does not rest primarily on general international derogation doctrine developed for public emergencies alone. It rests, first and most directly, on Malaysia's own constitutional order; even Malaysia's most extraordinary rights-limiting powers, the anti-subversion power under Article 149 of the Federal Constitution and the emergency power under Article 150 are not unconditional, undefined or immune from independent scrutiny. If the Federal Constitution itself does not permit a blanket, indefinite, unreviewable suspension of rights even under these two express constitutional gateways, an ordinary statutory exemption in a sectoral bill which invokes neither Article 149 nor Article 150 of the Federal Constitution cannot properly claim broader or less accountable insulation from scrutiny than the Federal Constitution's own extraordinary powers allow.

The Federal Constitution as the primary framework:

Malaysia already possesses an express constitutional architecture for exceptional security powers, comprising Article 149 of the Federal Constitution (action against subversion, organised violence and acts prejudicial to the public) and Article 150 of the Federal Constitution (action to combat

emergency). As Emeritus Professor Shad Saleem Faruqi observes, these two Articles are frequently described as the ‘dark underbelly’ of the Federal Constitution because they arm the Government with extraordinary powers, yet even these powers are subject to defined limits that are directly instructive for how the Bill's national security exemption should be constrained.

i) Legality and legitimate aim

Article 149 legislation may only be enacted for one or more of the specific purposes enumerated in Article 149(1)(a) to (f) of the Federal Constitution and the enacting statute must contain a recital that action has been taken or threatened by a substantial body of persons to cause the relevant harm. Article 150 of the Federal Constitution likewise requires an actual or threatened emergency affecting the security, economic life or public order of the Federation before the Yang di-Pertuan Agong (YDPA) may act. In both cases, the constitutional gateway to exceptional power is triggered only by a defined factual and legal predicate, not by an open-ended reference to ‘national security’ or ‘national defence’.

By contrast, the Bill's proposed exemption contains no equivalent statutory trigger. It exempts systems ‘used solely for national defence or security’ without requiring any defined factual basis, threat assessment or authorising instrument. This provides a materially weaker legality standard than Malaysia's own constitutional framework governing exceptional power.

ii) Bounded scope of permissible limitation

Even where Article 149 of the Federal Constitution is validly engaged, its scope is narrow; it permits legislation to abridge only the rights in Article 5 (personal liberty), Article 9 (freedom of movement), Article 10 (speech, assembly and association) and Article 13 (right to property) and cannot be used to violate other constitutional guarantees. Article 150 of the Federal Constitution is broader but is itself subject to the six entrenched topics under Article 150(6A) of the Federal Constitution; Islam, religion generally, Malay customs, native customs of Sabah and Sarawak, language and citizenship which cannot be touched even by a simple parliamentary majority during an emergency.

This demonstrates that even Malaysia's constitutional framework for exceptional powers does not authorise a blanket, undifferentiated suspension of rights and safeguards. The Bill's proposed exemption, by contrast, would remove an entire category of AI systems from the whole of the Bill's protective architecture; the AI Principles, the Risk and Harm Framework,

incident reporting, and Central AI Authority oversight without any equivalent boundary on which safeguards may be displaced and which may not.

iii) Necessity, proportionality and bona fides

Malaysian courts have consistently emphasised that exceptional powers exercised under Article 149 of the Federal Constitution remain subject to the rule of law. In Selva Vinayagam Sures v Timbalan Menteri Dalam Negeri [2021] 1 MLRA, the Federal Court held that preventive detention powers must be exercised bona fide for the statutory purpose, that every statutory precondition authorising the exercise of power must be strictly established, and that the material provisions conferring such powers must be strictly construed while safeguards protecting individual rights must be liberally interpreted. The Court further held that failure to satisfy the statutory requirements rendered the exercise of power unlawful. In PP v Khairuddin Abu Hassan & Anor [2017] 4 CLJ 541, the courts held that an alleged security offence not falling within the specific grounds enumerated in the enabling legislation could not be prosecuted under that legislation's special procedures, notwithstanding the Government's characterisation of the conduct as a security matter. These examples demonstrate that Malaysian law consistently treats exceptional security powers as temporary and subject to periodic review. The Bill's exemption should adopt a similar approach.

iv) Time limitation

Malaysia's own legislative practice demonstrates that exceptional security-related powers are not treated as permanently open-ended. The Security Offences (Special Measures) Act 2012 (SOSMA) contains a sunset clause in Section 4(11) of SOSMA and separately requires periodic parliamentary review of its detention-for-investigation provisions every five years. The Dangerous Drugs (Special Preventive Measures) Act 1985 likewise requires periodic parliamentary confirmation. Article 150(7) of the Federal Constitution provides that emergency legislation automatically lapses six months after an emergency proclamation ceases, absent further action. This body of domestic practice provides a ready-made, locally grounded precedent for time-limiting the Bill's national security exemption, rather than requiring reference to foreign derogation frameworks.

v) Independent oversight and judicial review

Article 151 of the Federal Constitution entitles preventive detainees, even under Article 149 legislation, to be informed of the grounds of arrest, to make representations to an independent Advisory Board and to have that Board's recommendations transmitted to the YDPA. More

significantly, the Federal Court's decision in Dhinesh a/l Tanaphll v Lembaga Pencegahan Jenayah & Ors [2022] 3 MLJ 356 marks a decisive shift in Malaysian constitutional jurisprudence; the Court unanimously struck down provisions of the Prevention of Crime Act 1959 (POCA) that ousted judicial scrutiny of detention grounds, holding that such an ouster clause relegates judicial review to 'nothing less than a clerical function' and affirming that the Constitution is supreme and that judicial power cannot be encroached upon by the legislature or executive. This confirms that, under current Malaysian constitutional law, even Malaysia's most sensitive security-related detention powers cannot be placed entirely beyond the reach of the courts.

Therefore, a national security exemption in the Bill that excludes independent oversight altogether would sit in tension with the trajectory of Malaysia's own constitutional jurisprudence which has moved toward, not away from, independent and judicial scrutiny of executive action taken in the name of security.

Defining 'national defence' and 'national security':

Rather than leaving these terms undefined, SUHAKAM recommends the Bill draw on Malaysia's own official policy instruments which already provide operative definitions capable of grounding a precise statutory exemption.

Malaysia's National Defence Policy defines the objective of national defence as the protection of national interests founded on sovereignty, territorial integrity and economic prosperity[15], implemented principally through the Malaysian Armed Forces across its core areas, offshore economic zones and strategic maritime and air lines of communication.[16] This is a comparatively narrow, military-institutional concept.

Malaysia's National Security Policy, by contrast, defines national security expansively as a state of freedom from any threat, internal or external, to nine enumerated National Core Values[17] and identifies a wide range of qualifying threats including fragility of national unity, challenges to the democratic system, irregular migration, extremism and terrorism, cyber security, natural disasters, economic and social crises, transnational crime, pandemics and infectious disease, energy security, and food security.[18] If adopted wholesale for the purposes of this exemption, the National Security

[15] Page 1.
[16] Pages 3 and 4.
[17] Paragraph 1.2.
[18] Paragraph 4.

Policy's broad definition could extend the exemption to AI systems used in public health, disaster management, immigration, economic regulation or food security. Several of these are precisely the sectors that SUHAKAM has elsewhere identified as requiring heightened rather than reduced, regulatory scrutiny.

SUHAKAM further notes that the National Security Policy itself identifies 'People's Security' and the protection of citizens' constitutionally entrenched rights as a National Core Value[19] and includes as an explicit primary strategy the guarantee of human rights entrenched in the Federal Constitution and in international law to which Malaysia is a party.[20] Importantly, the National Security Policy itself identifies 'People's Security' and the protection of constitutionally guaranteed rights as National Core Values, demonstrating that human rights protection is regarded as an integral component of national security rather than a competing interest. A blanket exemption that removes AI systems used for security purposes from human rights-relevant safeguards would sit uneasily with the security policy's own stated commitments.

Applying the test: legitimate aim, necessity, proportionality, time limitation and independent oversight:

Drawing together the constitutional principles discussed above, SUHAKAM recommends that any national security exemption under the Bill satisfy the following cumulative requirements:

i) Legitimate aim

The exemption should apply only to AI systems used for a defined defence or security purpose and should not extend to the broader range of non-traditional threats better addressed through the Bill's ordinary risk-tiered framework.

ii) Necessity

The exemption should apply only where compliance with the Bill's ordinary safeguards would be genuinely incompatible with a legitimate defence or security operation and not as a default exclusion for any system with a security-adjacent label.

iii) Proportionality

Any relaxation of the Bill's requirements should be limited to what is strictly necessary to protect the specific operational sensitivity in question; the exemption should identify which

[19] Paragraph 5.1.8.
[20] Paragraph 6.15.

safeguards may be modified and preserve the remainder rather than disapplying the Bill's protective architecture in its entirety, consistent with the bounded-scope principle drawn from Article 149 of the Federal Constitution above.

iv) Time limitation

The exemption should be subject to periodic parliamentary or ministerial review, for example, the five-year review cycle under the Security Offences (Special Measures) Act 2012 and the three-year review cycle already built into the National Security Policy itself so that reliance on the exemption remains demonstrably current and justified.

v) Independent oversight

The exemption should not exclude independent oversight or judicial review. At minimum, an Advisory Board-style mechanism, modelled on Article 151 of the Federal Constitution or an equivalent independent reviewer with appropriate security clearance should be empowered to examine reliance on the exemption and the exemption should not include an ouster clause preventing judicial scrutiny, consistent with the Federal Court's reasoning in Dhinesh Tanaphll.

Military AI systems and international humanitarian law:

Even where a national security exemption is justified under the constitutional framework discussed above, AI systems used in defence and military contexts do not operate in a legal vacuum. As currently drafted, the exemption may unintentionally remove some of the highest-risk AI applications and systems from regulatory oversight, including systems used for biometric identification, predictive policing, border management, intelligence gathering or other activities that may significantly affect the enjoyment of human rights, including the rights to privacy under Article 12 of the Universal Declaration of Human Rights (UDHR) and liberty and security under Article 3 of the UDHR. While certain operational information relating to national security may legitimately remain confidential, confidentiality should not translate into complete exemption from legal accountability or independent oversight.

This position is reinforced by international standards. Article 3 of the Council of Europe Framework Convention on Artificial Intelligence and Human Rights, Democracy and the Rule of Law (Council of Europe AI Convention) provides that activities relating to national security must nevertheless comply with applicable international law, including international human rights law, and respect democratic institutions and processes. Meanwhile, the EU AI Act specifies that the use of AI Systems for the purpose of law enforcement should therefore be prohibited, except in exhaustively listed and

narrowly defined situations where the use is strictly necessary to achieve a substantial public interest.[21] Where the national security or defence exemption is invoked in relation to AI systems used by the Malaysian Armed Forces in armed conflict or military operations, such use does not occur in a legal vacuum. As a State Party to the Geneva Conventions 1949, Malaysia remains bound by the Fourth Geneva Convention's protections for civilian persons in time of war and by the customary principles of international humanitarian law that govern the conduct of hostilities; distinction between combatants and civilians, proportionality in attack, precaution in the planning and execution of operations and the overarching principle of humanity. These obligations apply independently of and are not displaced by any exemption granted under the Bill.

This point is reinforced by recent international NHRI consensus. The GANHRI 2026 Outcome Statement expresses concern regarding the growing use of AI and neurotechnology in military and security management contexts, including autonomous and predictive technologies, given the associated risks to life, liberty and due process.[22] The Doha Declaration on Artificial Intelligence and Human Rights calls on States to prohibit military AI systems that do not operate consistently with international human rights law and international humanitarian law.[23] SUHAKAM treats this international consensus as reinforcing rather than substituting for Malaysia's existing treaty obligations under the Geneva Conventions and customary international humanitarian law.

Accordingly, SUHAKAM considers that the national security exemption should not operate as a blanket exclusion from the Bill's regulatory framework. Rather, it should constitute a narrowly tailored exception that remains subject to legality, necessity, proportionality, accountability, independent oversight and judicial review. Such safeguards are essential to protect human rights and to maintain public confidence that AI systems deployed for national security purposes continue to operate within the rule of law.

2.4.3 Interaction with Other Laws

SUHAKAM further recommends that the Bill clearly articulate its relationship with existing legislation. AI systems are increasingly regulated indirectly through a range of sector-specific laws governing data protection, cybersecurity, consumer protection, financial services, healthcare, communications and criminal law. In the absence of clear provisions governing this interaction, there

[21] Paragraph 33 of the EU AI Act.
[22] Paragraph 17 of the GANHRI Outcome Statement.
[23] Page 3 of the Doha Declaration.

is a risk of regulatory overlap, inconsistent compliance obligations, uncertainty regarding the competent regulator and potential gaps in enforcement. Accordingly, the Bill should clarify that compliance with this legislation does not diminish obligations arising under other applicable laws while also providing a clear framework for coordination and allocation of responsibilities where multiple legislative regimes apply.

2.4.4 Additional Definition

The ASEAN Guide on AI Governance and Ethics and the OECD Recommendation of the Council on Artificial Intelligence include a definition of an 'AI Actor'[24]; a broad term for any individual or organisation involved with at least one stage of the AI Lifecycle. AI development and usage is a complex, cross-industry endeavour that includes actors who may not directly fall into the definition of developer or deployer. Researchers, monitoring bodies, advisors and other actors should also be required, in a proportionate manner, to adhere to the principles of the proposed Bill.

Recommendations:

- Define the terms 'threshold test' and 'sensitive domain' to enhance legal certainty.
- Define the term 'AI Actor' to ensure miscellaneous roles in AI development also uphold the AI principles relevant to them.
- Define the terms 'national defence' and 'national security for the purposes of this exemption, ensuring that the exemption is confined to clearly identified defence and security functions.
- Require the exemption to identify a defined triggering predicate, analogous to the recital required under Article 149 of the Federal Constitution rather than applying on a bare assertion that a system is used 'solely' for national defence or security.
- Limit the scope of any relaxation of the Bill's requirements to what is specifically necessary for the operational sensitivity in question rather than disapplying the Bill's protective architecture in its entirety.
- Require any reliance on the exemption to demonstrate legality, necessity and proportionality, supported by documented reasons and subject to strict judicial scrutiny, consistent with Selva Vinayagam Sures.
- Subject the exemption to periodic review, modelled on the review mechanisms under the existing laws and policies.

[24] Page 15 of the ASEAN Guide on AI Governance and Ethics and Paragraph I (Page 8) of the OECD Recommendation of the Council on Artificial Intelligence.

- Establish an independent oversight mechanism, modelled on the Advisory Board under Article 151 of the Federal Constitution and expressly preserve access to judicial review, consistent with the Federal Court's reasoning in Dhinesh Tanaphll.
- Expressly preserve Malaysia's obligations under international humanitarian law, including the Geneva Conventions and customary international humanitarian law, for AI systems used in armed conflict or military operations.
- Clarify that the exemption does not displace any express prohibition applicable to Tier 1 (Unacceptable Risk) AI systems under the Bill's Risk and Harm Framework unless Parliament expressly provides otherwise.
- Require AI systems operating under the exemption to be subject to accountability safeguards, including appropriate documentation, audit trails, internal governance, independent oversight, and periodic review while recognising that genuinely operationally sensitive information may remain confidential where objectively justified.
- Clarify the relationship between this Bill and existing legislation to prevent regulatory overlap, conflicting obligations or gaps in enforcement.
- Establish mechanisms for cooperation and coordination among competent authorities where AI systems fall within multiple regulatory regimes.

3. THE CENTRAL AI AUTHORITY

SUHAKAM supports the creation of a Central AI Authority (CAIA) to address fragmentation and set consistent baseline standards. The CAIA model addresses a real institutional problem; fragmented, sector-siloed AI governance.

However, while the proposed framework establishes a strong institutional foundation for AI governance, the functions and powers of the CAIA are not yet expressly framed around the protection and promotion of human rights. As drafted, rights protection appears primarily through the AI Principles rather than through CAIA's own statutory mandate. Given that the CAIA will operationalise the National AI Principles, issue governance instruments, supervise compliance, investigate AI-related incidents and exercise enforcement powers, its institutional design will directly influence how AI systems affect human rights in practice.

Accordingly, SUHAKAM's central recommendation is that the protection and promotion of human rights should be embedded as a structural and non-delegable function of the CAIA rather than treated solely as an outcome of the AI Principles.

3.1 Core Functions

AI safety function:

The proposed AI safety function bundles development and maintenance of a 'risk-and-harm framework' together with a mandate to 'improve AI adoption'. Combining a safety/protective mandate with a promotional/adoption mandate may create a structural conflict of interest; a body institutionally incentivised to demonstrate adoption metrics may face competing institutional incentives when required to identify or regulate AI systems that present significant risks to individuals or communities.

Investigation and enforcement functions:

The description in the Bill is oriented toward system-level correction; producing findings 'to support corrective actions' with no explicit reference to individual remedy. SUHAKAM notes that a technical findings report that fixes the system going forward does not by itself compensate, restore or vindicate the individual who was harmed. This distinction is important because regulatory compliance and effective remedy serve different purposes; the former protects the public interest while the latter protects the rights of individuals adversely affected by AI systems.

AI enablement function:

SUHAKAM notes that capacity-building runs in one direction only; toward public authorities, regulators and regulated organisations. There is no reciprocal capacity-building for rights-holders, such as the public and civil society who will need to understand AI-related harms and incidents. This may result in an imbalance whereby institutional compliance capacity develops more rapidly than public awareness of AI-related rights and available avenues for redress. A system that builds compliance capacity for duty-bearers but not claim-making capacity for rights-holders risks disproportionately strengthening regulatory compliance while leaving affected individuals less equipped to recognise, challenge or seek remedies for AI-related harms.

Recommendations:

- Require the CAIA to establish objective and measurable compliance benchmarks, particularly for AI systems that present significant risks to human rights.
- Institutionally separate the promotion of AI adoption from the AI safety function or otherwise establish governance arrangements (for example, separate reporting lines, decision-making processes or performance indicators) that safeguard the independence of CAIA's protective functions.
- Require investigations involving identifiable individuals to consider system-level corrective measures as well as appropriate avenues for individual remedy, including ensuring that affected persons are informed of the investigation's outcome where appropriate.
- Expand the AI enablement function to include public rights-literacy initiatives, enabling individuals and communities to understand AI-related risks, available safeguards and mechanisms for seeking redress.

3.2 Key Powers

SUHAKAM notes with concern that the discretion CAIA holds over whether to issue a mandatory or non-mandatory legal instrument where necessary is currently unconstrained. Absent clearer statutory criteria, this discretion may result in inconsistent regulatory responses, particularly where AI systems have significant implications for human rights. If left unconstrained, CAIA could default to voluntary codes even where rights are materially at stake, simply because that is politically or commercially easier.

The recurrence of an 'improve AI adoption' objective within the AI safety function's key powers reinforces the conflict-of-interest concern raised above.

The enforcement toolkit (code of practice, directives, notices, interim measures, administrative penalties) is appropriately robust. However, SUHAKAM notes with concern that the Bill gives comparatively little attention to the procedural safeguards governing CAIA's own exercise of these powers. The Bill does not address due process safeguards such as the right to be heard before an interim measure is imposed, the publication of reasons for significant regulatory decisions or a right of appeal or judicial review. Without these, CAIA's considerable powers may not be accompanied by corresponding safeguards that promote transparency, procedural fairness and public confidence.

Administrative penalties, as presently framed, are primarily punitive and regulatory in nature. While they may promote deterrence and compliance, they do not provide an effective remedy to individuals who have suffered harm. Accordingly, the Bill should ensure that CAIA's enforcement powers are complemented by mechanisms that facilitate corrective measures and access to appropriate avenues for compensation or other forms of redress.

Recommendations:

- Specify that any instrument addressing a Tier 1 or Tier 2 risk to a human right must be mandatory and reserve voluntary instruments for Tier 3/baseline matters.
- Build in due process safeguards for CAIA's own enforcement powers; a right to be heard before interim measures, published reasons for directions/penalties and a clear right of appeal or judicial review.
- Allow CAIA's enforcement orders to include victim-directed remedial measures, such as correction, deletion, human review or re-decision of an AI-assisted outcome and other appropriate corrective measures. Where affected individuals or communities may be entitled to compensation or other civil relief, CAIA should be empowered to refer or facilitate the matter to the appropriate court, tribunal or other redress mechanism, in addition to administrative penalties payable to the State.

3.3 Relationship with Sectoral Regulators

The co-regulatory model proposed in the Bill is a pragmatic approach that recognises the technical expertise and existing statutory mandates of sectoral regulators. Nevertheless, from a human rights

perspective, this is also the aspect of the institutional framework where the risk of inconsistent protection is most pronounced.

Fragmentation risk survives despite the stated purpose:

The Bill's stated rationale for CAIA is to prevent fragmentation and inconsistent standards across sectors. Yet the co-regulatory model, by allowing each Sectoral Lead to implement sector-specific guidance based on its own legal authority, technical expertise and governance capacity risks reintroducing exactly the fragmentation CAIA was created to solve.

A person harmed by an AI system in a well-resourced, mature sector (for example, finance) may receive robust protection while the same harm in a less-resourced sector (for example, education, gig-economy platforms, social welfare administration) may not, purely because of differential institutional capacity rather than differential risk to the person.

This raises broader concerns relating to equality before the law and the consistent enjoyment of rights as the level of protection afforded to individuals should not depend on the institutional capacity of the sector in which AI is deployed.

Composition of Sectoral Leads:

The composition of Sectoral Leads should not be limited to regulators and industry representatives. Consideration should also be given to including independent expertise, civil society organisations and representatives of affected communities, particularly where AI systems have significant implications for human rights.

Given Sectoral Leads will advise on sector-specific legal instruments and implement enforcement functions, broader stakeholder participation would strengthen transparency, legitimacy and public confidence in sector-specific AI governance.

Additionally, Ministries dealing with the most rights-vulnerable populations, including welfare, education, health, women/family/community are not necessarily the best-resourced. Whether such Ministries are designated Sectoral Leads despite limited capacity or are not designated leads at all, the community who are marginalised bear the consequences.

Overcompliance / duplicate-reporting concern:

Industry's concern about reporting to 'one more' body is a legitimate compliance-burden issue. However, administrative efficiency should not be achieved at the expense of effective oversight or protection against AI-related harms. The Bill should not resolve an overcompliance concern by narrowing what must be reported or diluting incident-reporting thresholds as this would risk weakening regulatory visibility over AI-related harms affecting individuals and communities.

Federal-State jurisdictional gap:

The unresolved question raised in consultation about the source of power for the Bill to reach State-level bodies is not merely a technical issue. Many areas in which AI may have significant implications for rightsholders, including land administration, Syariah-related processes and native customary rights fall within State jurisdiction or involve shared Federal-State responsibilities. If the Bill's reach stops at the Federal boundary without a State-level coordination mechanism, an entire category of high-vulnerability AI use could fall into a protection vacuum.

Recommendations:

- Retain a non-delegable floor function regardless of a Sectoral Lead's maturity where CAIA must be able to directly monitor and intervene on rights-critical matters in every sector, including sectors with no sufficiently capacitated regulator.
- Require Sectoral Lead structures to include independent/public-interest representation (which could include SUHAKAM) with conflict-of-interest and recusal rules for industry representatives.
- Address duplicative reporting through administrative streamlining rather than through narrower reporting thresholds.
- Require CAIA to pursue formal intergovernmental coordination instruments so that State-administered functions involving AI are not left ungoverned.
- Direct CAIA's enablement function to prioritise capacity-building toward rights-sensitive but low-resource Ministries and not only general public-sector capacity building.

3.4 Where SUHAKAM Can Play a Role

SUHAKAM recommends the following concrete entry points which map onto its existing statutory functions under Act 597 rather than a general consultative role:

- Standard-setting/advisory - consider providing for independent human rights representation which may include SUHAKAM on any advisory or ethics panel established under the CAIA. The UNICEF Guidance on AI and Children also recommends establishing child-focused AI ethics committees to monitor AI usage from a child's perspective and advise on proactive governance approaches grounded in the best interest of the child.[25] SUHAKAM, particularly the Office of the Children's Commissioner (OCC) along with other relevant child experts may assist the CAIA in ensuring future governance initiatives remain child-focused and ethical.
- Institutional representation - provide for independent public-interest or human rights representation which may include SUHAKAM within Sectoral Lead governance arrangements where AI systems significantly affect vulnerable or at-risk groups.
- Complaints and inquiry - a formal referral pathway between SUHAKAM's existing complaints mechanism and CAIA's incident reporting mechanism with SUHAKAM retaining its independent functions for systemic AI harms, especially affecting groups unlikely to use formal statutory channels, for example, indigenous communities, refugees, persons with disabilities.
- Monitoring and public reporting - access to anonymised/aggregate National AI Incident Repository data to report on the human rights dimension of AI incidents and also serving as an informal check on inconsistent sectoral implementation.
- Capacity building - co-delivery of rights-focused training to Sectoral Leads and public agencies, complementing CAIA's technical/compliance training and public rights-literacy programmes so affected communities know how to raise AI grievances.
- Sandbox oversight - consultation on sandbox exit reports for rights compliance before full deployment.
- National security exemption oversight - an independent check, with appropriate clearance arrangements, on systems claimed to be exempt on national security grounds given the current absence of any safeguard in that exemption.
- Federal-State coordination - inclusion in any intergovernmental coordination mechanism addressing the jurisdictional gap identified above given SUHAKAM's mandate covering human rights.
- Governance instruments review - consultation on instruments for human rights compliance before issuance.

[25] Page 46.

4. AI PRINCIPLES

SUHAKAM appreciates NAIO's framework which draws appropriately from the OECD, UNESCO, Council of Europe, European Union, ISO/IEC and ASEAN. SUHAKAM's review, informed by the stakeholder feedback during the consultation and cross-referenced against the CEDAW, CRC, CRPD, ICCPR, the Doha Declaration and the GANHRI Outcome Statement, recommends the following refinements to ensure the five principles fully reflect Malaysia's human rights obligations:

4.1 Principle 1: Human Dignity, Agency and Rights

SUHAKAM appreciates that Principle 1 is appropriately positioned as the foundational principle of the proposed AI governance framework. Its emphasis on human dignity, agency and rights reflects a welcome commitment to a human-centred approach to AI governance and is broadly consistent with international developments.

However, in its current form, the principle functions primarily as a statement of overarching values rather than an operative standard capable of guiding consistent implementation by regulated persons. Requiring AI governance to 'uphold human dignity' and AI systems to 'remain subject to human-centred values' articulates an important normative objective but does not yet provide sufficient guidance on the concrete measures expected of developers, deployers, actors and other regulated persons to give practical effect to these concepts. Stakeholder feedback during the consultation reinforces this concern, noting that implementation requires greater specification. In particular, the slow development of Malaysia's personal data protection and privacy framework demonstrates that one of the most immediate dignity-related harms, namely, the loss of individual control over personal data, identity and autonomy remains insufficiently operationalised within the broader regulatory landscape.

From a human rights perspective, this distinction is significant. Human dignity is the normative foundation from which a range of substantive rights derive, including privacy, equality, autonomy and access to effective remedies. Accordingly, Principle 1 provides a clearer basis for translating those values into concrete regulatory obligations.

Regarding the issue of overlapping with other legislation, SUHAKAM cautions that avoiding overlap with other laws should not inadvertently create regulatory gaps or dilute the protection of human rights. While complementary legislation, for example, the Personal Data Protection Act 2010 [Act

709] may regulate particular aspects of AI governance, those regimes serve different regulatory purposes and should not be regarded as substitutes for the Bill's own responsibility to safeguard human dignity and other human rights. These are distinct but complementary harms requiring different legal responses and remedies.

Recommendations:

- Operationalise Principle 1 by translating the concepts of human dignity, agency and rights into concrete, measurable and auditable obligations capable of consistent implementation and enforcement.
- Clarify that the Bill should be interpreted consistently with the Federal Constitution and Malaysia's international human rights obligations, recognising that existing legislation addressing related matters (such as personal data protection) complements rather than replaces the Bill's own human rights safeguards.

4.2 Principle 2: Transparency and Explainability

SUHAKAM welcomes Principle 2 which appropriately recognises that transparency and explainability are essential safeguards for ensuring that individuals affected by AI-assisted decisions are informed, understand how those decisions were reached and are able to challenge them where appropriate. These procedural safeguards are fundamental to protecting accountability, procedural fairness and access to effective remedies.

However, the principle currently focuses primarily on the individual-facing dimension of transparency. While informing affected individuals is essential, international human rights standards increasingly recognise that transparency also serves a broader public-interest function by enabling independent oversight, democratic accountability and public trust in the governance of AI systems.

This distinction is particularly important where AI is deployed by public authorities or in contexts involving significant consequences for the enjoyment of human rights. AI systems used in public administration, justice, immigration, policing, healthcare, education, social protection or other essential public services may influence decisions affecting large numbers of people and, in some cases, entire communities. In these circumstances, transparency should operate as an ongoing governance safeguard that enables meaningful public scrutiny before, during and after the deployment of AI systems.

The GANHRI Outcome Statement reflects this broader understanding by requiring AI use by public authorities to remain subject to clear legal frameworks, appropriate ex ante safeguards, meaningful human oversight and independent scrutiny, particularly where AI systems affect access to public services, justice, migration procedures, policing, social protection, education, healthcare or other areas with significant consequences for rights.[26] These safeguards are especially important in the context of digital identity systems, integrated public service platforms, civil registration systems, interoperable databases and other large-scale public-sector data infrastructures, where failures may affect individual rights as well as public confidence in Government decision-making.

In this regard, transparency serves several distinct but complementary purposes. First, it enables individuals to understand decisions that affect them, exercise their right to challenge those decisions and obtain timely and appropriate redress.[27] Secondly, it enables regulators, oversight bodies and independent institutions to scrutinise whether AI systems are operating lawfully and consistently with human rights. Thirdly, it promotes broader democratic accountability by allowing the public to understand how AI is being deployed in areas that significantly affect society. This democratic inclusion and public understanding can only be achieved by ensuring AI documentation is easily understandable to the system's target audience which may include children. AI information should be prepared with due regard to children's understanding of the AI systems they interact with, including providing access to child-friendly, age-appropriate explanations of how to use the system and how to obtain legal redress and support if needed.

The Doha Declaration further reinforces this approach by calling for greater transparency in the development and deployment of AI systems, mandatory disclosure of AI use across sectors, including political discourse, public services and military applications, greater algorithmic explainability, independent audits and the publication of human rights impact assessments.[28] These measures recognise that transparency is not an end in itself but a mechanism through which accountability, public oversight and effective remedies can be realised.

Recommendations:

- For AI systems deployed by public authorities, strengthen Principle 2 by requiring mandatory ex ante disclosure, meaningful human oversight and independent scrutiny, particularly where such systems may significantly affect the enjoyment of human rights.

[26] Paragraph 16 of the GANHRI Outcome Statement.
[27] Paragraph 16 of the GANHRI Outcome Statement.
[28] Page 3 of the Doha Declaration.

- Require AI-generated content to be labelled, consistent with the Doha Declaration's recommendation, particularly for content moderation and public-facing generative AI applications.
- Require regulated persons to provide clear, meaningful and accessible explanations of the principal data sources, factors and reasoning that contributed to an AI-generated prediction, recommendation, content or decision while recognising that the level of explanation should be proportionate to target audience, especially if it includes children, as well as the nature and impact of the AI system.
- Ensure that individuals adversely affected by AI-assisted decisions are provided with sufficient information to enable them to effectively challenge those decisions, seek review where appropriate and access available remedies.
- Consider requiring the publication of human rights impact assessments or equivalent summaries for high-risk public-sector AI systems where appropriate, in order to strengthen public trust and facilitate independent oversight.

4.3 Principle 3: Accountability and Redress

SUHAKAM welcomes Principle 3 which appropriately recognises that responsibility for AI systems must remain attributable to identifiable natural or legal persons and that individuals who suffer harm should have access to effective remedies. This reflects an important human rights principle that the use of AI should not diminish legal accountability or create situations in which responsibility for harmful outcomes becomes diffuse or impossible to determine.

However, while the principle articulates the correct normative objective, stakeholder feedback during the consultation indicates that significant uncertainty remains regarding how accountability should operate in practice, particularly where multiple actors are involved throughout the AI lifecycle. Developers, deployers, vendors, integrators, data providers and end-users may each exercise different degrees of control over the design, deployment and operation of an AI system. Without greater clarity regarding their respective responsibilities, there is a risk that accountability becomes fragmented, making it more difficult for affected individuals to identify responsible parties and obtain effective remedies.

From a human rights perspective, this issue extends beyond regulatory certainty. Accountability is the mechanism through which rights become enforceable in practice. Where legal responsibilities are unclear or dispersed across multiple actors, affected individuals may face unnecessary barriers in

identifying the appropriate respondent, challenging unlawful or harmful AI-assisted decisions and obtaining timely and effective remedies.

SUHAKAM also notes that Principle 3 currently focuses primarily on accountability after harm has occurred. Accountability should also not be understood solely as a reactive obligation that arises after harm has occurred. Effective AI governance requires accountability to operate throughout the AI lifecycle by requiring organisations to identify, assess and mitigate reasonably foreseeable human rights risks before AI systems are developed, deployed and used. Embedding preventative accountability within the Bill would better align the proposed framework with internationally recognised approaches to responsible AI governance and human rights due diligence.

The Doha Declaration's operative recommendation is particularly relevant in this regard. It calls for robust accountability and redress mechanisms through the establishment of clear legal responsibilities for all actors involved in AI development and deployment, accessible remedy options for affected individuals, appropriately trained oversight bodies and targeted capacity building for policymakers, public officials and the general public.[29] Similarly, the GANHRI Outcome Statement emphasises that private actors, including developers and deployers bear an independent responsibility to respect human rights and must ensure meaningful and effective access to remedy for individuals whose rights are adversely affected by their technologies, including through accessible and effective complaint mechanisms.[30]

SUHAKAM also notes that accountability and redress are not gender-neutral in their operation. The GANHRI Outcome Statement expressly recognises the growing risks posed by technology-facilitated gender-based violence[31] which should be considered alongside CEDAW's General Recommendation No. 40 (2024) on the Equal and Inclusive Representation of Women in Decision-Making Systems, particularly its observations regarding women's digital literacy and the continued underrepresentation of women in STEM education, careers and leadership positions.[32] These factors may affect women's representation in the development of technological advances, their ability to participate in AI processes and their practical ability to access accountability and effective remedies where AI-related harms occur. Similarly, AI systems increasingly affect children through educational technologies, generative AI applications and digital platforms, as well as persons with disabilities through automated accessibility tools, assistive technologies and AI-supported public services. The CRC (as

[29] Page 4 of the Doha Declaration.
[30] Paragraph 25 of the GANHRI Outcome Statement.
[31] Paragraph 9 of the GANHRI Outcome Statement.
[32] Page 6.

elaborated in the General Comment No. 25), Joint Statement on Artificial Intelligence and the Rights of the Child and the CRPD (in particular, Articles 5, 7 and 9) provide well-developed international standards regarding accessibility, equality, participation and protection from discrimination that could usefully inform both the Bill's AI Principles and its broader risk-management framework. Accordingly, Principle 3 should be designed from the outset to be accessible, inclusive and responsive to the distinct circumstances of women, children, persons with disabilities and other groups in vulnerable situations rather than assuming that a single complaint mechanism will operate equally effectively for all users.

Recommendations:

- Clarify that accountability under the Bill extends across the entire AI lifecycle and applies proportionately to all actors exercising meaningful control over the design, development, deployment, operation or modification of AI systems.
- Allocate legal responsibilities across developers, deployers, vendors, integrators, data providers and other relevant actors according to each actor's degree of control over, contribution to or ability to prevent or mitigate the relevant harm, including where responsibility is shared.
- Require a public, plain-language, publicly available statement of responsibility and the appropriate point of contact to accompany every AI system above a baseline risk tier so that responsibility is clearly identifiable by affected individuals, regulators and the wider public.
- Strengthen Principle 3 by recognising accountability as both a preventative and corrective obligation, requiring organisations to identify, assess and mitigate reasonably foreseeable human rights risks before deployment while ensuring effective remedies remain available where harm nevertheless occurs.
- Require developers and deployers of higher-risk AI systems to conduct and document human rights impact assessments and maintain appropriate governance documentation, decision records and audit trails sufficient to demonstrate compliance, facilitate regulatory oversight and support investigations or judicial proceedings where necessary.
- Ensure that individuals adversely affected by AI-assisted decisions have access to timely, accessible and effective complaint mechanisms, meaningful human review where appropriate, and remedies that are capable of addressing both individual harm and systemic deficiencies identified during investigations.
- Ensure redress mechanisms are explicitly designed to be accessible, inclusive and responsive to the needs of women, children, persons with disabilities and other marginalised groups, including individuals who may be reluctant or unable to access traditional complaint mechanisms.

4.4 Principle 4: Safety, Security and Robustness

SUHAKAM welcomes Principle 4 which recognises that AI systems should be designed, developed, deployed and maintained in a manner that is safe, secure and technically robust throughout their lifecycle. However, from a human rights perspective, the concept of 'AI safety' should not be understood solely in technical or operational terms. An AI system may satisfy requirements relating to cybersecurity, accuracy, resilience and technical robustness while nevertheless producing harms and rights-infringing outcomes. These harms may include algorithmic discrimination, unjustified surveillance, exclusion from essential services, erosion of privacy, disproportionate impacts on communities who are marginalised and the denial of meaningful human oversight in decisions affecting rights. Accordingly, technical safety and human rights protection should be understood as complementary dimensions of responsible AI governance.

The GANHRI Outcome Statement's particular emphasis on the growing use of AI and neurotechnology in military and security contexts, including autonomous and predictive technologies[33] is especially relevant. It recalls States' commitments to assess the impact of emerging technologies on deprivation of liberty, the rule of law, access to justice and the prevention of torture and other cruel, inhuman or degrading treatment, including the use of AI-assisted decision-making and facial recognition technologies by law enforcement authorities. These considerations are directly relevant to the Bill, particularly given the current breadth of the proposed national security exemption.[34]

Recommendations:

- Clarify that Principle 4 encompasses both technical safety and the prevention, mitigation and ongoing management of foreseeable human rights risks arising from the design, development, deployment, use and termination of AI systems.
- Require Principle 4's safety duties to apply with particular rigour to law enforcement, security and military-adjacent AI uses (for example, facial recognition, predictive policing and autonomous systems), thereby addressing the regulatory gap created by the Bill's current national security exemption, as highlighted in SUHAKAM's earlier comments on the scope of the Bill.

[33] Paragraph 17 of the GANHRI Outcome Statement.
[34] Paragraph 24 of the GANHRI Outcome Statement.

- Strengthen the principle by requiring continuous post-deployment monitoring, recognising that AI-related risks may emerge or evolve over time as systems learn, adapt or are deployed in new contexts.
- Clarify the boundary between the CAIA's Principle 4 obligations and those of other relevant authorities, regulators and existing legal frameworks by establishing a coordinated and coherent compliance pathway that minimises duplication while maintaining effective oversight.

4.5 Principle 5: Responsible Data Stewardship

SUHAKAM welcomes Principle 5 which recognises that responsible data governance is fundamental to trustworthy AI systems. Given that AI systems depend heavily on the quality, integrity and governance of data throughout their lifecycle, this principle provides an important foundation for ensuring reliable and accountable AI outcomes.

However, from a human rights perspective, data governance is also intrinsically linked to the protection of privacy, equality, autonomy and other human rights. While the Bill appropriately emphasises responsible management of data, it does not expressly recognise privacy as a distinct human right.[35] Although robust data governance contributes significantly to privacy protection, it does not by itself guarantee protection against unlawful or disproportionate interference with an individual's private life. Other AI frameworks such as the EU AI Act explicitly provides for several privacy protections, such as prohibiting the use of real-time remote biometric identification systems in publicly accessible spaces for the purposes of law enforcement, except for narrow exceptions of substantial public interest such as searching for victims of trafficking or abduction, or identifying suspects of serious criminal offences with these exceptions only enforced along with strict safeguards and judicial authorisation.[36] Therefore, responsible data stewardship should be framed as supporting broader human rights protections with a focus on data minimisation and privacy-preserving measures rather than operating solely as an information governance obligation.

SUHAKAM further notes that poor data stewardship frequently results in systemic rather than individual harms. Inaccurate, incomplete, historically biased or unrepresentative datasets may disproportionately disadvantage particular communities, including women, children, persons with disabilities, indigenous peoples, migrants, refugees and other communities who are marginalised. These harms often arise because the underlying data reflects or reproduces existing structural

[35] Article 17 of the ICCPR.
[36] Article 5 of the EU AI Act.

inequalities. Therefore, responsible data stewardship should encompass ongoing assessment of data quality, representativeness and potential discriminatory impacts throughout the AI lifecycle.

The Doha Declaration's data protection recommendations provide a clear and precise standard; legislation grounded in legality, data minimisation, purpose limitation, integrity, confidentiality, accountability and transparency together with independent data protection authorities with the mandate and resources to effectively oversee compliance and enforce accountability, including requirements for AI developers to retain training-data records for disclosure in the public interest.[37]

Recommendations:

- Either expand Principle 5 or add a further principle that expressly recognises privacy and the protection of private life as fundamental rights while requiring privacy-by-design, data minimisation and purpose limitation throughout the AI lifecycle.
- Require organisations to adopt appropriate measures to identify and mitigate risks arising from inaccurate, incomplete, unrepresentative or historically biased datasets, particularly where AI systems are deployed in high-impact contexts.
- Strengthen guidance on representative datasets and data quality assurance to minimise discriminatory outcomes affecting vulnerable or marginalised communities.
- Do not treat existing Government data-classification practices as satisfying Principle 5. Instead, require the designation of a clearly accountable role (for example, a Data Protection Officer or AI Data Steward) with explicit responsibility for safeguarding rights and ensuring compliance with this Principle, separate from official secrecy or information-classification functions.
- Anchor the proposed data repository in the principles of legality, necessity, proportionality, data minimisation, purpose limitation and accountability outlined above, with independent oversight, preferably undertaken in coordination with the relevant data protection authority and other competent regulators, consistent with the GANHRI Outcome Statement's call for greater coordination between NHRIs, data protection authorities, equality bodies and other oversight institutions.
- For emergency scenario guidance, require the application of an explicit necessity and proportionality test, together with clear sunset clauses or periodic review mechanisms so that temporary relaxation of data safeguards does not become a permanent or unreviewed exception. This would also be consistent with SUHAKAM's earlier concerns regarding the breadth of the Bill's national security exemption.

[37] Page 3 of the Doha Declaration.

- Promote periodic review of datasets used in higher-risk AI systems to ensure that they remain accurate, relevant and appropriate for their intended purpose over time.

4.6 The Phased 'Socialise then Codify' Implementation Approach

SUHAKAM recognises the pragmatic rationale for phased implementation, namely, to build awareness, institutional readiness and capacity before introducing binding legal obligations. However, from a human rights perspective, this sequencing may nevertheless give rise to significant risks, particularly in light of the emphasis placed by both the Doha Declaration and the GANHRI Outcome Statement on the urgency of establishing effective safeguards for AI governance.

The GANHRI Outcome Statement warns that digitalisation is advancing more rapidly than the development of legal safeguards, institutional capacity, public awareness and effective oversight, creating a risk that rights-infringing practices become normalised before they are adequately understood, regulated or remedied.[38] This concern is directly relevant to the proposed phased implementation of the Bill. If the initial socialisation phase is not accompanied by a minimum set of immediately applicable safeguards, AI systems that adversely affect human rights may already be deployed and cause harm before binding legal obligations take effect. For example, a discriminatory automated eligibility system or an opaque public-sector facial recognition system could undermine individuals' rights, erode public confidence and become institutionally embedded before the regulatory framework is fully operational. While SUHAKAM appreciates the need for a gradual implementation process, it considers that safeguards essential to the protection of human rights should apply from the outset rather than being deferred to a later phase.

Recommendations:

- Adopt a two-track approach rather than a single linear phase:
  - (i) a small set of non-derogable, immediately binding obligations, constituting the minimum human rights floor recommended by SUHAKAM, including the prohibition of Tier 1 unacceptable-risk practices, mandatory incident reporting for serious harms and mandatory human rights impact assessments for high-risk public-sector systems, to apply from the commencement of the Act; and
  - (ii) the broader socialisation-then-codification track for the remaining, more technical or sector-specific requirements relating to the five AI principles.

[38] Paragraph 5 of the GANHRI Outcome Statement.

4.7 Due Regard vs. Mandated Human Rights Due Diligence

SUHAKAM considers that this represents one of the key areas where the proposed Bill could be further strengthened to align with the emerging international NHRI consensus. Both the Doha Declaration[39] and the GANHRI Outcome Statement[40] advocate a shift towards binding, legally mandated human rights due diligence and human rights impact assessments rather than sole reliance on a discretionary due-regard standard that is calibrated primarily by the regulated entity's own assessment of risk.

SUHAKAM acknowledges that the Bill adopts a principles-based approach to AI governance. Nevertheless, a 'due regard' obligation, on its own, may not provide sufficient assurance that foreseeable human rights risks have been systematically identified, assessed and mitigated. By contrast, human rights due diligence requires organisations to proactively identify, prevent, mitigate and account for adverse human rights impacts throughout the AI lifecycle, as a result, strengthening accountability before harm occurs.

Recommendation:

- Reframe the 'due regard' obligation as the standard for baseline (Tier 3) and general application of the AI Principles. However, require mandatory, legally binding human rights due diligence and human rights impact assessments, proportionate to the level of risk, with appropriate public disclosure and subject to independent audit for any Tier 2 (high-risk) system, particularly those used by public authorities or deployed in justice, healthcare, social protection, law enforcement, migration or education.

4.8 Missing Principles

Non-discrimination and equality:

None of the five principles expressly recognises non-discrimination or equality as a standalone principle, notwithstanding that algorithmic discrimination and bias are among the most well-documented and consequential AI-related harms globally, with disproportionate impacts on women, children, ethnic and religious minorities, persons with disabilities, older persons and low-income communities.

[39] Pages 3 and 5 of the Doha Declaration.
[40] Paragraphs 24 and 25 of the GANHRI Outcome Statement.

As iterated, both the Doha Declaration and the GANHRI Outcome Statement identify discriminatory profiling, algorithmic bias and gendered digital harms as central human rights risks arising from AI. Similarly, the CEDAW Addendum to General Recommendation No. 30 on Women in Conflict Prevention, Conflict and Post-Conflict Situations in relation to the Women and Peace and Security Agenda (CEDAW Addendum to General Recommendation No. 30) reinforced UN General Assembly Resolution on Seizing the Opportunities of Safe, Secure and Trustworthy Artificial Intelligence Systems for Sustainable Development (UN General Assembly Resolution 78/265) which recognises that algorithmic discrimination and bias embedded in AI systems may exacerbate existing gender inequalities and emphasises the need for governance frameworks that mitigate bias in datasets and AI systems.[41] Likewise, the ASEAN Guide on AI Governance and Ethics recognises Fairness and Equity as a core principle, calling on deployers to implement safeguards to ensure algorithmic decisions do not further exacerbate or amplify existing discriminatory or unjust impacts.[42] These instruments demonstrate an emerging international consensus that non-discrimination and equality are not merely implicit values and provide compelling support for SUHAKAM's recommendation that Non-Discrimination and Equality be recognised as a standalone AI Principle rather than being left implicit within the broader principle of human dignity.

Special attention to the principle of non-discrimination and fairness in the context of children should also be given due consideration. The UNICEF Guidance on AI and Children highlights the need for AI systems to actively support and benefit children in marginalised situations, including girls, children from minority groups, children with disabilities, children who speak non-prioritised languages and children with low digital literacy. It further recommends that data used to develop AI systems be assessed for equity and representativeness to minimise bias and reduce the risk of exclusionary or discriminatory outcomes affecting children.[43]

Furthermore, Article 8 of the Federal Constitution guarantees equality before the law and protection against discrimination. Although Malaysia has not ratified the ICCPR, Article 26 of the ICCPR remains a widely recognised international benchmark on equality before the law and equal protection of the law without discrimination. Together with Malaysia's obligations under CEDAW, CRC and CRPD, these standards reinforce the importance of recognising non-discrimination and equality as an express AI Principle. Doing so would strengthen the Bill's alignment with both Malaysia's

[41] Paragraph 82.
[42] Pages 12 and 13.
[43] Pages 25 and 26.

constitutional framework and evolving international human rights standards on human-centered AI governance.

Environmental sustainability:

SUHAKAM notes that many international AI governance frameworks increasingly recognise environmental sustainability as an emerging component of responsible AI governance. Therefore, consideration could be given to including sustainability as an additional National AI Principle, encouraging developers, deployers and regulators to minimise the environmental impacts of AI systems throughout their lifecycle while supporting responsible innovation.

Best interest of the child:

Children who form the most vulnerable group in our society require specialised attention in all aspects of the AI lifecycle due to the malleability of their neurological development, combined with an unprecedented rise in the prominence of AI usage. The CRC calls for the best interests of the child to be a paramount consideration in all matters concerning children[44], including AI systems that directly or indirectly impact them. The UNICEF Guidance on AI and Children outlines three main principles for a child-centred framework; protection, provision, and participation of children.[45] The principle of prioritising child-centred frameworks, including consulting children and giving due regard to their views[46] should be obligated to be upheld by all parties in the development, deployment, use and termination of AI systems.

Recommendations:

- Add a standalone principle on Non-Discrimination and Equality, requiring developers and deployers to identify, test for and mitigate discriminatory bias across the AI lifecycle, including during training data selection, model design, deployment and ongoing monitoring.
- Consider incorporating environmental sustainability into the National AI Principles, recognising the environmental impacts of AI throughout its lifecycle, including energy and water consumption, greenhouse gas emissions, electronic waste and the extraction of critical minerals. Guidance issued by CAIA could encourage developers and deployers, particularly of higher-impact AI systems to consider these impacts when designing, developing and deploying AI systems.

[44] Article 3 of the CRC.
[45] Page 8.
[46] Article 12 of the CRC.

- Add a standalone principle on the Best Interest of the Child, requiring developers and deployers to carry out child risk impact assessments, implement safeguards that protect children's best interest, and actively consult children to ensure AI system features are child-friendly and beneficial for their development.

4.9 Updating the International Benchmarks Referenced for AI Principles

SUHAKAM recommends that the Bill's international benchmarks for the AI Principles be updated to include the additional instruments referenced by SUHAKAM.

These instruments represent the most recent international consensus on AI governance and human rights. They are more recent, specifically focused on the intersection between AI and human rights and provide practical guidance for Governments, regulators and NHRIs. Their inclusion would strengthen the Bill's human rights foundation and visibly align Malaysia's AI governance framework with the evolving international NHRI consensus.

4.10 Where SUHAKAM Can Play a Role

Consistent with its statutory mandate under Act 597 and its commitments under the Doha Declaration and the GANHRI Outcome Statement, SUHAKAM proposes the following concrete contributions to support the implementation of the AI Principles:

Recommendations:

- Ensure the Bill reflects Malaysia's commitments under CEDAW, CRC and CRPD while taking into account SUHAKAM's commitments under the Doha Declaration, the GANHRI Outcome Statement and the other international standards referenced throughout this feedback.
- Provide SUHAKAM with a formal advisory role in relation to the implementation, monitoring and periodic review of the AI Principles, particularly on matters relating to human rights, equality and access to effective remedies.
- Monitor rightsholders' digital harms as a standing SUHAKAM function consistent with the Doha Declaration's specific call to NHRIs to monitor AI-driven attacks on rightsholders and the GANHRI Outcome Statement's broader commitment to receive and follow up on complaints involving digital harms to rights-holders in situations of vulnerability.

5. AI RISK AND HARM FRAMEWORK

5.1 What is 'Harm' under the Bill

SUHAKAM supports the Bill's approach of anchoring its regulatory framework in actual harm and intent rather than purely technical classification. This outcomes-based orientation is sound and avoids over-regulating benign technical features. Nevertheless, from a human rights perspective, while this definition serves as a useful starting point, it may not fully capture important categories of AI-related harm that international human rights standards and the more recent NHRI instruments treat as fundamental concerns, such as:

- Discrimination and unequal treatment are not covered as a standalone harm category. Instead, such harms would appear to fall within 'unlawful deprivation of fundamental liberty' or 'contravention of any written law', meaning that protection depends largely on the scope of existing domestic equality and anti-discrimination laws. As a result, certain forms of AI-enabled discrimination, particularly indirect or structural discrimination may not clearly fall within any specific statutory provision despite producing significant adverse human rights impacts. The CEDAW Addendum to General Recommendation No. 30 reinforced UN General Assembly Resolution 78/265 which recognises that AI systems may reproduce algorithmic discrimination and emphasises the importance of mitigating bias embedded in datasets and strengthening the governance of safe, secure and trustworthy AI systems. Although framed within conflict settings, CEDAW Committee's observations reflect broader concerns regarding discriminatory AI systems and reinforce the need for discrimination to be recognised as a standalone category of harm under the Bill.[47]
- Collective, societal or democratic harms are inherently diffuse. They may not constitute a 'harm' to any single identifiable individual under the current definition. However, the GANHRI Outcome Statement treats such harms as a central consideration for AI governance.[48]

Recommendations:

- Expand the statutory definition of 'harm' to expressly include discriminatory or unequal treatment (whether direct or indirect and regardless of intent).
- Delink the discrimination limb of 'harm' from any intent requirement, consistent with the settled principle in international equality law reflected in CEDAW and the CRPD that indirect

[47] Paragraph 82 of the CEDAW Addendum to General Recommendation No. 30.
[48] Paragraph 21 of the GANHRI Outcome Statement.

and structural discrimination may amount to unlawful discrimination irrespective of the actor's intent.

- Include a residual category capturing cumulative or societal harm so that harms operating at population scale are not excluded merely because no single victim can be identified.

5.2 What is 'Risk' under the Bill

The three risk factors; likelihood, severity and scale as well as duration and reversibility provide a useful general-purpose risk methodology. However, they remain largely population-neutral. They ask how serious the harm is and how extensive or enduring its effects may be, but not who is most likely to bear those impacts. The UN Report on Governing AI for Humanity suggests utilising a vulnerability-based approach in assessing AI-related risks. This approach frames risks by who will likely be negatively impacted by AI, such as children, marginalised communities, women and other groups.[49]

The GANHRI Outcome Statement recognised that children, persons with disabilities, older persons, minorities, LGBTQI+ and human rights defenders face heightened exposure to digital exclusion, surveillance, abuse or discriminatory impacts. Consequently, an AI system presenting the same level of technical risk may have markedly different human rights consequences depending on the characteristics and vulnerabilities of the affected population. This approach is also reflected in CEDAW, CRC and CRPD which require assessments that take account of the specific circumstances and needs of different groups. The current three-factor test does not expressly require such a differentiated assessment.

Recommendations:

- Add a fourth risk factor; the vulnerability and disproportionate impact on affected individuals or groups, requiring assessors to consider whether the affected population includes groups with heightened exposure to AI-related harms and to weight likelihood and severity accordingly.
- Require meaningful consultations with rights holders and affected communities, including women, children, persons with disabilities and others in vulnerable situations when conducting risk assessments for systems affecting access, particularly to public services.
- Require that consultations involving children comply with Article 12 of the CRC, including methods for child consultations should accommodate diverse groups of children, with child safeguards in place to ensure the adoption of age-appropriate methods, maintain transparency

[49] Paragraph 28.

of representation, counter risks of tokenism or manipulation, avoid collection of unnecessary data and protect child consultants from any form of retaliation. The consultations involving children must comply with the General Comment No. 12 (2009) on the Right of the Child to be Heard by the UN Committee on the Rights of the Child which enumerates steps for the implementation of child participation, which requires:

i) Prior consultation preparation that ensures children are well informed and consenting of their participation;
ii) Allowing the child to exercise his/her right to be heard in an encouraging and confidential manner;
iii) Assessment of the capacity of child participants;
iv) Giving feedback or information on the outcome of consultations to child participants; and
v) Implementation of complaints, remedies, and redress mechanisms for when children's right to be heard is disregarded or violated.[50]

- Require human rights impact assessments, including a distinct child rights impact assessment where relevant, for systems that are reasonably likely to have a material effect on children or other persons in vulnerable situations.

.

5.3 Tier 1: Unacceptable Risk

The listed examples, including serious manipulation, exploitative targeting of vulnerability, unlawful social scoring, serious biometric abuse and developing an AI system with the intent of causing harm provide an important foundation. However, SUHAKAM considers that the framework could be further strengthened in two respects when assessed against the Doha Declaration and the GANHRI Outcome Statement.

First, the intent criterion may not adequately capture discrimination-based harms. A system that produces serious, systemic discriminatory outcomes, for example, a recruitment or credit-scoring system that reliably disadvantages women or persons with disabilities could warrant classification as a Tier 1 system or at least a heightened Tier 2 system based on its actual or reasonably foreseeable effects without requiring proof that it was developed with an intention to cause harm. CEDAW and the CRPD do not condition a finding of discrimination on intent and SUHAKAM considers that the Bill's highest-risk category should similarly recognise discriminatory effects, irrespective of intent.

[50] Paragraphs 40 - 46.

Second, SUHAKAM notes that two categories explicitly identified by the Doha Declaration and the GANHRI Outcome Statement are not presently reflected within Tier 1.

The first concerns military AI systems that do not operate in compliance with international human rights law and international humanitarian law. The Doha Declaration calls on States to prohibit such systems[51] while the GANHRI Outcome Statement separately raises concern regarding the growing use of AI and neurotechnology in military and security management contexts, including autonomous and predictive technologies, given their implications for the rights to life, liberty and due process.[52] This concern is further reinforced by the CEDAW Addendum to General Recommendation No. 30 which recognises that AI-enabled military technologies, including lethal autonomous weapon systems may amplify gender bias embedded in training datasets, resulting in discriminatory or erroneous targeting decisions with serious humanitarian consequences.[53] The CEDAW Committee further notes growing international concern regarding the need to regulate such systems consistently with international humanitarian law and international human rights law.[54] These considerations assume added significance in light of SUHAKAM's earlier observations regarding the Bill's proposed national security exemption. As presently drafted, military and security-related AI may both fall outside the scope of the Bill and remain absent from the list of prohibited practices, potentially creating a significant regulatory gap.

The second concerns fully automated decision-making in sensitive domains without meaningful human review. The Doha Declaration calls on States to ensure meaningful human decision-making in the administration of justice, healthcare, social protection, law enforcement and military contexts while emphasising that AI deployment across sectors should be complemented by appropriate human oversight.[55] This suggests an important distinction that could be more clearly reflected in the Bill, namely, between AI systems that support human decision-making and those that effectively replace human decision-makers in contexts with significant human rights implications.

[51] Page 3 of the Doha Declaration.
[52] Paragraph 17 of the GANHRI Outcome Statement.
[53] Paragraph 88.
[54] Paragraph 89.
[55] Page 3 of the Doha Declaration.

Recommendations:

- Reframe Tier 1 criteria so that discriminatory systems can be classified as unacceptable risk based on demonstrated or reasonably foreseeable discriminatory effect, not only on demonstrated intent.
- Add an explicit Tier 1 category for AI systems designed for or used in military or security contexts in a manner inconsistent with international human rights law or international humanitarian law taking into account the concerns raised in the Doha Declaration, the GANHRI Outcome Statement and the CEDAW Addendum to General Recommendation No. 30 regarding AI-enabled military technologies, including autonomous weapon systems and AI applications that may amplify discrimination or unlawful targeting and require this category to apply notwithstanding the Bill's national security exemption.
- Add an explicit Tier 1 (or specially heightened Tier 2) category for fully automated decision-making without meaningful human review in the administration of justice, healthcare, social protection, law enforcement or migration.
- Explicitly include AI systems designed to create or facilitate non-consensual intimate imagery within the 'serious manipulation'/'exploitative targeting' examples given the GANHRI Outcome Statement's specific identification of technology-facilitated gender-based violence (TFGBV) and non-consensual intimate imagery as a distinct, serious digital harm.

5.4 Tier 2: High Risk

The mandatory measures listed, such as risk assessment, documentation, traceability, human oversight, testing, validation, monitoring and incident notification, provide a comprehensive technical compliance framework. However, SUHAKAM notes that they do not explicitly require bias or discrimination auditing, or accessibility compliance, both of which are central considerations under Article 9 of the CRPD on accessibility and the equality provisions of CEDAW and the CRPD.

In addition, the Bill does not specify which real-world domains should ordinarily be classified as Tier 2. The GANHRI Outcome Statement identifies digital identity systems, integrated public service platforms, civil registration systems, interoperable databases and other large-scale public sector data infrastructures as warranting heightened scrutiny.[56] It likewise highlights decisions affecting access to justice, migration, policing, social protection, education and healthcare[57] and further recommends

[56] Paragraph 16 of the GANHRI Outcome Statement.
[57] Paragraph 12 of the GANHRI Outcome Statement.

that very large online platforms and AI developers and deployers conduct regular assessments of high-risk AI systems.[58]

Providing greater clarity regarding the categories of AI systems that should ordinarily fall within Tier 2 would promote greater regulatory certainty and facilitate more consistent implementation across sectors.

Recommendations:

- Add mandatory bias/discrimination testing and accessibility compliance to the list of Tier 2 obligations alongside the existing technical safeguards.
- Establish a presumptive list of Tier 2 domains drawn directly from the GANHRI list above; justice, healthcare, migration, policing, social protection, education, digital identity and public-service infrastructure so that classification does not depend solely on case-by-case technical assessment. Within the education sector, this should include AI systems used for student admissions, assessment of learning, disciplinary decision-making, behavioural or emotion inference, student profiling, biometric monitoring and decisions affecting access to educational support or opportunities. The Bill should provide that AI systems reasonably likely to have a material effect on a child's education, development or future opportunities are presumed to fall within Tier 2 (High Risk) unless demonstrated otherwise.
- Ensure that the required risk assessment for Tier 2 systems expressly incorporates a human rights impact assessment component, consistent with SUHAKAM's recommendations under the AI Principles Fact Sheet, rather than being confined to technical or operational risk considerations.

5.5 Tier 3: Low Risk

Governing Tier 3 systems through baseline obligations and voluntary instruments is generally appropriate for genuinely low-impact applications. However, from a human rights perspective, it may not fully capture systems that are individually low in severity yet produce serious cumulative or societal harm at scale. For example, content-recommendation or engagement-optimisation algorithms where each individual interaction may appear relatively insignificant can collectively affect information diversity, civic space and democratic participation. These broader societal impacts are expressly recognised in the GANHRI Outcome Statement which identifies online harassment,

[58] Paragraph 25 of the GANHRI Outcome Statement.

information manipulation and threats to democratic participation as significant AI-related human rights concerns.[59]

Recommendation:

- Introduce a scale-based escalator within the risk framework so that systems deployed at mass or population scale are subject to enhanced regulatory scrutiny where their aggregate societal impact on information integrity, civic space or democratic participation is significant, even if the severity of individual interactions would otherwise place them within Tier 3.

5.6 How Regulations May Be Created under the Framework

The three regulation-making pathways; by class of AI system, by domain of deployment and by sector provide a flexible and practical regulatory framework. The domain-based approach is particularly well suited to operationalising the GANHRI-identified high-risk domains, including justice, healthcare, migration, policing, social protection and education through subsidiary legislation.

Nevertheless, SUHAKAM considers that two aspects of the proposed framework could benefit from further clarification.

First, split duty-holding by pathway may create accountability gaps. Since class-based regulation makes the developer the primary duty holder and domain-based regulation makes the deployer the primary duty holder, there may be situations where harm arising from the interaction between a compliant AI model[60] and a deployer's implementation choice[61] results in uncertainty regarding legal responsibility. In practice, this could allow responsibility to become fragmented with each party relying on the other's regulatory pathway. This mirrors the developer/deployer accountability concern SUHAKAM has raised in its feedback on the Bill's Definitions and Scope.

Second, sector-based codes of practice risk reproducing inconsistencies in sectoral regulatory maturity. This echoes the concern SUHAKAM has previously highlighted regarding the CAIA. Where a Sectoral Lead has differing levels of technical expertise, regulatory capacity or human rights experience, there is a risk that sector-specific guidance may inadvertently adopt a lower standard of

[59] Paragraph 21 of the GANHRI Outcome Statement.
[60] An AI built to follow safety rules and work correctly.
[61] How a company or person sets up, connects or uses the AI in their system.

protection than that envisaged by the national AI framework. This could undermine the objective of maintaining consistent national standards for Tier 1 and Tier 2 AI systems.

Recommendations:

- Include a joint-responsibility mechanism so that where harm results from the interaction of class-level design and domain-level deployment choices, both developer and deployer can be held accountable, rather than allowing responsibility to be displaced between the respective regulatory pathways.
- Require that sector-specific codes of practice maintain standards that are consistent with, and not materially lower than, the national Tier 1 and Tier 2 classifications, consistent with SUHAKAM's recommendation regarding the CAIA that substantive human rights standards remain uniform regardless of sector maturity.

5.7 Rightsholders / Communities at Risk and Intersectionality

As reflected throughout SUHAKAM's observations on the definition of harm, the assessment of risk and the proposed tiering framework, the impacts of AI systems are not experienced uniformly across society. Certain groups may face heightened exposure to discrimination, exclusion or other adverse human rights impacts because of existing structural inequalities or particular vulnerabilities.

From a human rights perspective, incorporating an explicit intersectionality requirement would help ensure that AI systems are assessed not only according to their overall technical risk but also according to the differentiated impacts they may have on particular individuals and communities. This would strengthen the Bill's alignment with the principle of substantive equality reflected across international human rights law and would better support proportionate, risk-based regulation.

Recommendation:

- Require the Bill or its subsidiary legislation relating to the AI Risk and Harm Framework to include an express requirement that risk and harm assessments consider the actual and potential impacts of AI systems on:

    i) Women and girls - including technology-facilitated gender-based violence (TFGBV), non-consensual intimate imagery, algorithmic discrimination and other gender-specific digital harms, consistent with CEDAW, the CEDAW Addendum to General Recommendation No. 30 and Paragraph 9 of the GANHRI Outcome Statement.

ii) Children - including through child rights impact assessments for AI systems reasonably likely to affect them, applying the best interests of the child as a primary consideration under Article 3 of the CRC and respecting the child's right to be heard under Article 12 of the CRC, as further elaborated in General Comment No. 25.

iii) Persons with disabilities - through accessibility-by-design under Article 9 of CRPD, non-discrimination, including reasonable accommodation under Article 5 of CRPD and meaningful consultation with representative organisations under Article 4(3) of CRPD.

iv) Other groups identified by the GANHRI Outcome Statement as facing heightened exposure to AI-related harms - older persons, minorities, indigenous peoples, migrants, stateless persons, persons living in poverty, LGBTQI+ persons, human rights defenders, journalists and NHRIs themselves.

5.8 Protection of Children from AI Risks

While children are one of the groups identified above as requiring heightened consideration in AI risk assessments, SUHAKAM considers that the complexity of children's rights and the unique ways in which AI systems may affect them justify dedicated safeguards within the Bill.

Accordingly, SUHAKAM recommends the adoption of a child-centred risk and harm framework supported by mandatory child rights impact assessments for AI systems reasonably likely to affect children. Such assessments should treat the best interests of the child as a primary consideration, assess risks in light of children's evolving capacities, facilitate meaningful participation of children from diverse backgrounds, adopt data-minimising and privacy-preserving approaches, and incorporate age-appropriate transparency and safety-by-design throughout the AI lifecycle.

Recommendations:

- Expand the definition of ‘risks’ under the Bill to include content or system functionalities that may promote harm or unhealthy behaviours in children, encourage anthropomorphic qualities or human-like emotional bonds with child users and otherwise reasonably prejudice the best interests of the child.
- For AI systems likely to be used by or affect a child, require Tier classification and risk assessments to expressly evaluate the likelihood, severity and scale as well as and duration and reversibility of risks in the context of child users.

- Require sector-specific codes of practice governing AI systems likely to affect children to conduct child risk impact assessments as well as implement child-specific safeguards and protection frameworks addressing child-centred risks.
- Require AI systems reasonably likely to be accessed or used by children to adopt child-safe-by-design and child-rights-by-design approaches throughout the AI lifecycle, including during design, development, deployment, monitoring and post-deployment review.

6. AI INCIDENT REPORTING

SUHAKAM's feedback in this section draws upon the General Comments and General Recommendations of the UN treaty bodies on access to justice, recognising that a national incident-reporting and user-complaint system is, in practical terms, an access-to-justice mechanism.

CEDAW General Recommendation No. 33 on Women's Access to Justice sets out six interlinked components, namely, justiciability, availability, accessibility, good quality, accountability and remedies which provide a useful framework for assessing whether individuals affected by AI systems are able to report harm as well as to pursue appropriate forms of redress. The report of the UN High Commissioner for Human Rights on Right to Access to Justice under Article 13 of CRPD highlights the importance of equality, participation and effective remedy. Similarly, CRC General Comment No. 25 emphasises child-friendly processes and appropriate and effective remedial judicial and non-judicial mechanisms.

These international standards demonstrate that an effective incident-reporting framework should also ensure that individuals whose rights have been adversely affected by AI systems are able to report incidents through accessible procedures and, where appropriate, be directed or referred to the relevant court, tribunal, regulator or other competent redress mechanism capable of providing an effective remedy. While the incident-reporting framework should facilitate access to justice, it should not be understood as replacing existing judicial or statutory mechanisms for determining liability or awarding compensation unless expressly provided for by law.

SUHAKAM recommends applying these access-to-justice principles systematically to the design of the incident-reporting framework.

6.1 Why Malaysia Needs AI Incident Reporting

The stated rationale; sectoral silos, the need for systemic learning and the objective of complementing rather than displacing existing mechanisms is sound and SUHAKAM welcomes this approach. However, SUHAKAM considers that the rationale could be further strengthened by recognising a complementary human rights objective.

As presently framed, the rationale is predominantly institutional, focusing on better governance and regulatory coordination. In SUHAKAM's view, the framework should also expressly recognise

facilitating access to effective remedy for affected individuals as a co-equal objective. This would be consistent with the third pillar of the UN Guiding Principles on Business and Human Rights on access to remedy as well as the access-to-justice principles reflected in CEDAW General Recommendation No. 33, CRC General Comment No. 25 and the report of the UN High Commissioner for Human Rights on Article 13 of the CRPD.

The articulation of the framework's objectives is important, as its stated purpose will inevitably influence its design and implementation. A framework conceived primarily as a regulatory learning mechanism may not adequately prioritise ensuring that individuals are able to navigate appropriate complaint pathways and obtain effective remedies through the competent authority.

Recommendation:

- Add facilitating access to effective remedy for affected individuals as an explicit, co-equal purpose of the incident-reporting framework alongside systemic learning and regulatory coordination, including by requiring the framework to provide clear pathways for referral to the appropriate court, tribunal, regulator or other competent redress body where individual remedies fall outside CAIA's statutory powers.

6.2 The Dual-Intake Model

The combination of statutory reporting (developer/deployer-filed) and user complaints (public-filed) is an appropriate structural approach and directly responds to concerns raised during the consultation that users and affected groups are often the first to experience actual impacts while developers and deployers possess the operational information necessary to support technical investigations.

Nevertheless, when assessed against CEDAW General Recommendation No. 33's six access-to-justice components, the level of detail and procedural safeguards afforded to the two reporting pathways appears uneven.

Justiciability and availability:

The statutory report pathway is well-specified with prescribed channels and defined technical content. By comparison, the user complaint pathway is described only as being submitted ‘through Central AI Authority channels’ with no detail regarding the available entry points, whether multiple or decentralised complaint mechanisms will be established to reach marginalised communities, such as rural communities, indigenous peoples or B40 communities or whether reliance on a single national

reporting channel may inadvertently create barriers for persons without reliable internet access, digital literacy or other necessary resources.

Accessibility:

CEDAW General Recommendation No. 33 requires that justice mechanisms be legally and practically accessible, affordable, physically reachable, available in relevant languages, procedurally simple and supported by assistance where needed.

The report of the UN High Commissioner for Human Rights on Right to Access to Justice under Article 13 of CRPD further emphasises that accessibility requires procedural accommodations[62] and personnel trained to communicate effectively with complainants with diverse disabilities.[63] CRC General Comment No. 25 similarly requires complaint mechanisms to be child-friendly.[64] These important accessibility considerations are not presently reflected in the description of the user complaint pathway.

Good quality:

CEDAW General Recommendation No. 33 requires proceedings to be timely and gender-responsive and designed to minimise the risk of secondary victimisation.

This is particularly relevant in cases involving tech-facilitated gender-based violence (TFGBV), such as non-consensual intimate imagery, which is specifically identified in the GANHRI Outcome Statement as a serious digital harm.[65] In such cases, the design of the complaint mechanism itself can significantly influence whether affected individuals are willing and able to seek redress, and poorly designed procedures may inadvertently compound the harm experienced by complainants.[66]

Recommendations:

- Design the user complaint pathway to the same level of specificity as the statutory report pathway, including multiple, decentralised entry points, multilingual and offline/non-digital intake options, simplified, low-literacy-friendly complaint forms and dedicated procedures for complaints involving gender-based violence and children.

[62] Paragraphs 24 to 32.
[63] Paragraphs 41 and 60.
[64] Paragraph 44.
[65] Paragraph 9 of the GANHRI Outcome Statement.
[66] Paragraph 25 of the GANHRI Outcome Statement.

- Require complaint-handling personnel to receive training on engaging with complainants with disabilities, consistent with the report of the UN High Commissioner for Human Rights on the Right to Access to Justice under Article 13 of the CRPD as well as training on child-sensitive procedures, consistent with CRC General Comment No. 25.
- Establish an independent regulatory body that has auditing oversight, including the power to request audits and that utilises child rights experts' expertise to specifically check for child rights infringements.
- Ensure the user complaint pathway is resourced and prioritised to the same extent as the statutory pathway, recognising that affected persons are often the first to know that harm has occurred and noting the concerns raised by stakeholders regarding the potential reluctance of developers and deployers to self-report.

6.3 When Must an Incident be Reported

The trigger of 'any event, outcome or credible allegation where an AI system is suspected of causing harm' is ultimately dependent upon the definition of 'harm' adopted by the Bill. As set out in SUHAKAM's feedback on the AI Risk and Harm Framework, the current definition; death, personal injury, unlawful deprivation of liberty and contravention of written law may not fully capture discrimination, unequal treatment or diffuse societal harms. Accordingly, unless the definition of 'harm' is broadened, the incident-reporting obligation may not extend to several categories of AI-related harm that the GANHRI Outcome Statement and the Doha Declaration identify as among the most significant in practice.

NAIO's presentation offers a broader, technically-framed definition of AI incident covering any event, failure, weakness, misuse, unexpected effect or material circumstance that causes or may cause AI harm. SUHAKAM welcomes this broader formulation at the level of triggering events. Nevertheless, it remains anchored to the concept of 'AI harm', therefore, its effectiveness will depend upon the breadth of the underlying statutory definition of harm.

SUHAKAM further notes that the 'credible allegation' standard for triggering a report from a user complaint should not be interpreted as requiring complainants to establish credibility without assistance before their complaint is accepted. CEDAW General Recommendation No. 33 is explicit that the burden placed upon individuals seeking justice should not be so onerous as to render access

to justice practically inaccessible, particularly for complainants with limited legal or technical literacy.[67]

Recommendations:

- Align the incident-reporting trigger directly with the broadened definition of 'harm' recommended in SUHAKAM's feedback on the AI Risk and Harm Framework so that discrimination, unequal treatment and societal or democratic harms are reportable.
- Clarify that 'credible allegation' is a threshold to be assessed by the CAIA upon receipt of a complaint, rather than a burden that complainants must independently satisfy before their complaints are accepted. A good-faith, low-threshold approach to complaint intake should be adopted.
- Explicitly extend 'near miss' reporting to include rights-related near misses, such as bias identified through internal testing but not corrected before deployment and not only technical or safety-related near misses, consistent with the mandatory bias-auditing obligation SUHAKAM has recommended for Tier 2 systems.

6.4 What Happens after Reporting

SUHAKAM considers this stage to be particularly important because it determines whether the incident-reporting framework functions solely as a regulatory learning mechanism or also as an effective avenue for facilitating access to justice and effective remedies.

The described sequences; confirm the incident, determine response and recovery, prevent recurrence, produce a technical findings report and apply differentiated logging/preservation duties provide a structured process for regulatory oversight and systemic learning. However, the framework does not expressly require consideration of the appropriate pathway to remedy for an identifiable individual who has suffered harm.

CEDAW General Recommendation No. 33's sixth component, remedies, requires that justice mechanisms provide a range of effective remedies, including restitution, compensation, rehabilitation, satisfaction and guarantees of non-repetition, which should be implemented without undue delay. The Doha Declaration's operative recommendation on accountability similarly calls for

[67] Paragraph 25(a)(iii).

accessible remedy options for affected individuals, in addition to corrective action directed at the system.[68]

NAIO's presentation frames the process as investigation, safety assessment, corrective action and regulatory learning. While these are important components, SUHAKAM considers that the framework could be further strengthened by incorporating an explicit remedy-determination/referral stage. Positioning this stage alongside corrective action would better reflect the principle that AI governance should protect both the public interest in improving AI systems and the individual rights of persons adversely affected by those systems. Where the appropriate remedy falls outside CAIA's statutory powers, the framework should facilitate referral to the appropriate court, tribunal or other competent redress mechanism.

Recommendations:

- Add an explicit remedy-determination/referral stage to the post-reporting process. Where an investigation confirms that an identifiable individual has suffered harm, the CAIA's findings should include consideration of the appropriate corrective measures within its statutory powers (for example, correction, deletion, suspension of deployment or re-decision by a human) and where individual remedies (for example, compensation) fall outside CAIA's jurisdiction, provide for referral to the appropriate court, tribunal or other competent redress mechanism.
- Ensure that the complainant or their representative is informed of the outcome of any investigation triggered by their complaint, including any corrective measures taken and any available avenues for review, appeal or referral to an appropriate court, tribunal or other competent redress mechanism where they consider the outcome to be inadequate, consistent with the accountability component of CEDAW General Recommendation No. 33 under which justice mechanisms must themselves be transparent and accountable.
- Apply differentiated logging and record-preservation duties as proposed. However, ensure that retained records may, subject to appropriate safeguards and applicable privacy protections, be accessed by the affected individual or their legal representative for the purpose of pursuing appropriate legal or administrative remedies.
- Build appropriate whistleblower protections for employees and other persons acting on behalf of developers and deployers who report suspected AI-related incidents in good faith, recognising the concerns raised during the stakeholder consultation regarding potential reluctance to report such incidents.

[68] Page 4 of the Doha Declaration.

6.5 National AI Incident Repository

The policymaking and systemic learning function of the repository is valuable and SUHAKAM supports its inclusion within the proposed framework. However, from a human rights perspective, two important considerations could be addressed more explicitly.

First, the privacy of complainants and affected individuals. A national repository aggregating incident data should not inadvertently create a secondary privacy risk. Consistent with Article 12 of the UDHR and the data protection standards referenced in SUHAKAM's feedback on the AI Principles, the Bill should require repository entries to be anonymised or aggregated before central storage while ensuring that any identifying information is managed under a separate and appropriately restricted access regime.

Second, public transparency of aggregate findings. Neither the Doha Declaration[69] nor the GANHRI Outcome Statement[70] envisages an incident repository as serving regulators alone. Both call for greater transparency regarding how AI-related harms are identified, monitored and addressed, thereby enabling the public and civil society to hold both industry and Government accountable. Therefore, a repository that operates solely as an internal regulatory resource may miss an important opportunity to strengthen public confidence, institutional accountability and evidence-based policymaking.

Recommendations:

- Require anonymisation or aggregation of repository entries as the default approach a separate, access-controlled process for any identifying information required for individual remedy or law enforcement purposes.
- Require the CAIA to publish a regular (for example, annual) public report drawing on aggregate repository data, covering patterns of AI-related harm, the sectors and populations most affected and the regulatory measures taken in response.
- Provide SUHAKAM and other relevant independent oversight bodies with access to anonymised/aggregate repository data for the purposes of monitoring systemic human rights impacts and informing evidence-based policy recommendations.

[69] Pages 3 and 5 of the Doha Declaration.
[70] Paragraphs 15, 16 and 25 of the GANHRI Outcome Statement.

6.6 Rightsholders / Communities at Risk

Consistent with the access-to-justice principles reflected in the relevant UN treaty body General Comments, General Recommendations and reports, SUHAKAM recommends that the user complaint pathway be explicitly designed to accommodate the needs of the following groups:

- Women and girls - gender-responsive intake procedures and protection from secondary victimisation for AI-facilitated gender-based violence complaints, consistent with CEDAW General Recommendation No. 33.
- Children - mechanisms that are free, safe, confidential and accessible that enable children to complain directly without parental involvement where this would create a conflict or risk; provide support and protection against retaliation; and ensure prompt safeguarding referrals, consistent with CRC General Comment No. 25.
- Persons with disabilities - through procedural and communication accommodations, accessible complaint mechanisms and appropriately trained personnel, consistent with Article 13 of the CRPD and the report of the UN High Commissioner for Human Rights on the Right to Access to Justice under Article 13 of the CRPD.
- Other marginalised communities, including rural, indigenous people and persons with low digital literacy - through non-digital, decentralised and multilingual complaint channels to ensure that access to the incident-reporting framework does not, in practice, depend upon digital literacy or geographic location.

Incorporating these measures would help ensure that the accessibility of the incident-reporting framework is assessed in terms of whether affected individuals are realistically able to access and benefit from it in practice.

6.7 Where SUHAKAM Can Play a Role

Consistent with its statutory mandate under the Act 597 and its commitments under the Doha Declaration and the GANHRI Outcome Statement, SUHAKAM proposes the following potential contributions in relation to the AI Incident Reporting Framework:

- Provide technical advice on the design of the user complaint pathway, drawing upon CEDAW General Recommendation No. 33, CRC General Comment No. 25 and the report of the UN High Commissioner for Human Rights on the Right to Access to Justice under Article 13 of the

CRPD with the objective of ensuring that the user complaint mechanism receives the same level of design attention and procedural safeguards as the statutory reporting pathway.

- Advocate for a five-stage post-reporting process that expressly incorporates a remedy-determination/referral stage alongside investigation, safety assessment, corrective action and regulatory learning.
- Contribute to the development of the definitions of ‘AI incident’ and ‘AI harm’ so that they are consistent with the broadened definition of harm recommended in SUHAKAM's feedback on the AI Risk and Harm Framework, ensuring that discrimination, unequal treatment and diffuse societal harms are capable of being reported and appropriately addressed.
- Support the establishment of appropriate governance arrangements that enable SUHAKAM and other independent oversight bodies to access anonymised repository data for systemic human rights monitoring, trend analysis and policy development, subject to appropriate safeguards.

7. AI ENABLING POWERS

SUHAKAM recognises the genuine institutional challenge that this fact sheet seeks to address. AI governance is dynamic, technically complex and cross-sectoral and a rigid, one-time statute risks the 'regulate and forget' problem identified by NAIO. Accordingly, delegating detailed operational matters to subsidiary legislation while Parliament retains oversight through the primary Act is a legitimate and well-established regulatory approach that enables the framework to respond to technological developments over time.

At the same time, delegated powers of this breadth, covering classification thresholds, prohibition of AI systems, incident-reporting procedures, sandbox arrangements and technical and organisational safeguards also require appropriate rule-of-law safeguards. From a human rights perspective, such powers should be accompanied by clear legal precision and foreseeability, meaningful parliamentary or independent oversight, opportunities for public participation before rules take effect and effective avenues for review or remedy where delegated instruments adversely affect rights.

Article 8 of the UDHR guarantees an effective remedy for acts violating fundamental rights which necessarily presupposes that the legal framework governing those rights is sufficiently transparent, accessible and capable of independent scrutiny. Therefore, SUHAKAM considers that the current framework could be strengthened by expressly incorporating these safeguards, as none of the three enabling powers nor the discretion to issue mandatory or voluntary governance instruments is presently accompanied by corresponding procedural safeguards.

This concern is reinforced by NAIO's own description of the CAIA's strategic coordination model which identifies a dual mandate whereby CAIA operates both as a regulator and as the national coordinator responsible for AI enablement. As highlighted in SUHAKAM's earlier observations on the CAIA and the AI Principles, combining promotional and regulatory functions within a single institution may create an inherent tension between encouraging AI adoption and ensuring robust protection of human rights. A body whose institutional success is measured partly by adoption and readiness may, in certain circumstances, face competing policy considerations when determining whether binding regulatory intervention is necessary. Therefore, appropriate statutory safeguards would help reinforce public confidence in the independence and consistency of regulatory decision-making.

7.1 The Three Key Enabling Powers

Power to develop governance instruments:

This power is necessary to ensure that the Bill remains responsive to technological developments. However, SUHAKAM notes with concern that the discretion to issue governance instruments as either mandatory or voluntary is currently not accompanied by any express criteria linked to the potential impact on human rights. Without such guidance, there is a possibility that matters with significant human rights implications may be addressed through voluntary instruments where binding regulatory requirements would provide greater legal certainty and protection.

Power to prohibit use:

This is an important and appropriately robust power for addressing Tier 1 unacceptable-risk systems. Nevertheless, SUHAKAM considers that the Bill could be strengthened by incorporating procedural safeguards governing the exercise of this power. At present, there is no requirement to publish reasons, provide notice, offer affected developers or deployers and, where appropriate, affected individuals/communities an opportunity to be heard or expressly recognise a right of appeal or judicial review.

Given that the exercise of this power may have significant implications for livelihoods, continuity of services or access to essential public services, its exercise should be subject to principles of procedural fairness and legality. This approach would be consistent with the non-arbitrariness principles reflected in the UDHR as well as Article 14 of the Council of Europe AI Convention which requires States Parties to ensure that effective remedies are available domestically for violations of the Convention's provisions.

Capacity building:

As presently framed, this function is directed solely at the public sector by promoting safe and responsible AI adoption within Government. While this objective is important, SUHAKAM considers that the proposed capacity-building framework could adopt a broader rights-based approach.

This reflects the concern SUHAKAM has already raised in its feedback on the CAIA's AI Enablement Function, namely that there is currently limited emphasis on capacity building for rightsholders, civil society and other stakeholders who will also need to understand, participate in and engage with Malaysia's evolving AI governance framework.

Recommendations:

- Establish clear, binding criteria linking the severity of rights impact to the choice between mandatory and voluntary instruments so that Tier 1 and Tier 2 matters affecting fundamental rights are addressed through appropriate legally binding measures where necessary, rather than voluntary instruments alone.
- Introduce due process safeguards for the power to prohibit use, including a requirement to publish reasons, provide affected parties with an opportunity to be heard before, or where necessary promptly after, a prohibition is imposed (subject to appropriately limited emergency exceptions) and recognise a clear right of appeal or judicial review.
- Broaden the capacity-building function to include human rights literacy, AI literacy and access-to-justice initiatives for the public, civil society and other relevant stakeholders.

7.2 Why These Powers are Needed

SUHAKAM agrees that Parliament cannot be expected to legislate exhaustively for a rapidly evolving technology through a single legislative exercise, therefore, SUHAKAM welcomes the proposed shift from a 'regulate and forget' model towards one that is capable of adapting and learning over time.

However, SUHAKAM also notes that the GANHRI Outcome Statement highlights a complementary consideration that is directly relevant to this approach. It observes that digitalisation is advancing more rapidly than the development of legal safeguards, institutional capacity, public awareness and effective oversight, thereby increasing the risk that rights-infringing practices become normalised before they are adequately understood, regulated or remedied.[71]

Accordingly, while an adaptive regulatory framework is both necessary and appropriate, it should not result in the postponement of core human rights protections pending further regulatory experience or institutional learning.

Recommendation:

- Adopt the two-track approach recommended by SUHAKAM in this submission, comprising (i) a limited set of immediately applicable, non-derogable human rights protections, including the prohibition of Tier 1 practices, mandatory incident reporting for serious harms and mandatory human rights impact assessments for high-risk public-sector AI systems and (ii) the continued

[71] Paragraph 5 of the GANHRI Outcome Statement.

development of the broader technical and operational framework through the enabling powers provided under the Bill.

7.3 What Future Regulations May Cover

The five listed areas; classification and thresholds, incident-reporting procedures, technical and organisational safeguards, sandbox arrangements and sector/domain/class-specific standards provide an appropriate foundation for future subsidiary legislation. However, from a human rights perspective, several important implementation areas are not expressly identified.

These include procedures relating to remedy and redress, methodologies for bias and discrimination testing, accessibility standards, methodologies for human rights impact assessments and procedures for public participation in future rule-making.

While these matters could potentially be addressed through subsidiary legislation, the current drafting presents them as issues that regulations may address rather than matters that should ordinarily form part of the regulatory framework. SUHAKAM considers that greater clarity in this regard would help ensure that future regulations give appropriate attention to both technical governance requirements and human rights implementation.

Recommendations:

- Add to the list of matters that subsidiary regulations should address: (i) remedy and redress procedures for affected individuals, (ii) bias and discrimination testing methodology, (iii) accessibility standards consistent with CRC and CRPD, (iv) human rights impact assessment methodologies for high-risk systems and (v) public participation and consultation procedures for the development of future governance instruments.
- Require that any regulation governing sandbox arrangements incorporate the human rights safeguards SUHAKAM recommended in this submission, including informed consent, heightened protections for children and other persons in vulnerable situations and the principle that non-derogable human rights protections should not be suspended during testing.

7.4 Governance Instruments: Mandatory vs Voluntary

The proposed toolkit of governance instruments, comprising mandatory instruments (such as rules, directives, enforcement orders and subsidiary regulations) and voluntary instruments (including

reports, standards, guidelines and voluntary codes) provides the CAIA with a flexible range of regulatory tools. However, SUHAKAM considers that three specific design features would benefit from further clarification and safeguards.

Reports used to 'calibrate risk tiers and justify binding measures':

If CAIA's own reports form the evidentiary basis for future binding rules that affect rights, the underlying data and reasoning should, to the greatest extent possible, be transparent and subject to public scrutiny, consistent with the transparency expectations reflected in both the Doha Declaration and the GANHRI Outcome Statement as well as the right to seek and receive information under Article 19 of the UDHR. Without an appropriate degree of transparency, there is a risk that binding regulatory measures may be justified by evidence or reasoning that affected stakeholders and the public are unable to examine or meaningfully challenge.

Standards described as 'consensus instruments':

Consensus-building processes for technical standards typically draw on industry, regulators and technical experts. While such expertise is essential, absent an explicit requirement, affected communities, civil society and independent human rights institutions may not be adequately represented in these consensus processes. This may inadvertently reproduce the same representation gap that SUHAKAM has identified in relation to the proposed Sectoral Lead arrangements under the CAIA.

In particular, for standards relating to accessibility, equality and non-discrimination, international human rights standards require the meaningful participation of rights holders. The obligations under CEDAW, CRPD and CRC to closely consult with and actively involve persons affected by decisions concerning them provide useful guidance that could be more explicitly reflected in the framework.

Voluntary codes that 'operate immediately':

Providing a mechanism for a rapid, non-binding response to emerging risks is both practical and appropriate, particularly in relation to genuinely novel or lower-risk AI applications. However, the current framework does not appear to prevent voluntary codes from becoming the principal or long-term regulatory response to Tier 1 or Tier 2 risks where the development of binding instruments is delayed.

The UN Guiding Principles on Business and Human Rights make clear that the State's duty to protect human rights is fulfilled through appropriate legislative and regulatory measures, rather than reliance

on voluntary business self-regulation. Accordingly, voluntary instruments are best understood as complementing legally binding regulation, rather than replacing or indefinitely postponing it.

Recommendations:

- Require CAIA reports used to justify binding measures to be published or, where necessary, released in an appropriately redacted or aggregated form that protects legitimate confidentiality while preserving public accountability for the underlying evidence and reasoning.
- Require consultation and standard-setting consensus processes to include meaningful participation by civil society, affected communities and independent human rights bodies, consistent with the participatory obligations reflected in CEDAW, CRPD and CRC.
- Require that any voluntary code addressing a Tier 1 or Tier 2 risk be accompanied by a clear timeframe for the development and issuance of a corresponding binding instrument, thereby ensuring that voluntary measures function as an interim regulatory mechanism rather than a long-term substitute for enforceable legal protections.

7.5 Cross-Cutting Recommendations on Delegated Rule-Making

Having regard to the observations set out above, SUHAKAM recommends that the Bill establish a structured accountability and oversight framework governing CAIA's exercise of its enabling powers which should include the following safeguards:

- A requirement for mandatory public consultation on draft governance instruments before finalisation, including meaningful engagement with affected communities identified through the vulnerability-weighted risk assessment.
- A requirement for CAIA to publish reasons for significant regulatory decisions, including the exercise of the power to prohibit use and the issuance of binding rules, directives or enforcement orders.
- An express right of appeal or judicial review in relation to governance instruments and enforcement actions that adversely affect the rights or legitimate interests of identifiable persons.
- A requirement for periodic statutory review of the enabling powers framework with SUHAKAM identified as one of the key consultees to assess whether the proposed ‘adapt and learn’ model continues to provide an effective and rights-respecting response to emerging AI risks and harms.

8. AI SANDBOX FUNCTION

SUHAKAM welcomes the Bill's clarification that the sandbox is intended to function as a governance mechanism rather than a deregulated exception and that it is not designed as a licence for unmanaged experimentation, instead, it provides a structured supervisory pathway. SUHAKAM agrees that this is the appropriate conceptual foundation for a regulatory sandbox. However, its effective implementation will depend on ensuring that appropriate human rights safeguards are embedded throughout the sandbox framework. SUHAKAM's principal observations are set out below.

8.1 What is a Sandbox

SUHAKAM agrees with the Bill's characterisation of a sandbox as a controlled environment for understanding the opportunities and risks of specific AI innovations, operating by reference to the AI Principles, the Risk and Harm Framework and the supervisory functions of the CAIA. This provides an appropriate conceptual anchor for balancing innovation with regulatory oversight. However, the effectiveness of this approach will depend on ensuring that the operation of the sandbox remains consistent with the human rights principles underpinning the broader legislative framework.

The recommendations below are intended to ensure that the practical operation of the sandbox remains aligned with the objectives described in the Bill.

Recommendations:

- State expressly, on the face of the Bill, that participation in the sandbox does not derogate from the non-derogable rights floor recommended in SUHAKAM's submission, including the prohibition of Tier 1 unacceptable-risk practices and that any statutory power to relax regulatory requirements does not infringe constitutional rights and human rights.
- Require sandbox exit reports to include a human rights impact assessment or, at minimum, a human-rights-informed risk assessment before an AI system transitions to full deployment.

8.2 Why Malaysia is Building a Sandbox

The three stated rationales; encouraging innovation, enabling flexible testing and generating evidence for policy appropriately recognise the benefits that regulatory sandboxes can offer to developers and regulators. However, SUHAKAM notes that none expressly identifies the protection of test participants and the wider public as a parallel objective of the sandbox framework.

The way in which the purpose of a regulatory mechanism is articulated often shapes how it is implemented in practice. A framework that is framed primarily around innovation and regulatory learning may inadvertently place comparatively less emphasis on the protection of individuals participating in, or affected by, sandbox testing.

Recommendation:

- Emphasise a fourth explicit rationale for the sandbox; to facilitate the safe testing of AI innovations while safeguarding the rights, safety and dignity of participants and the wider public, consistent with the Doha Declaration's call for human rights due diligence and human rights impact assessments throughout the AI lifecycle, including during pre-deployment testing.

8.3 What the Sandbox Could Do: A Function-by-Function Review

The nine proposed sandbox functions span a wide spectrum of activities, ranging from low-risk internal technical testing to testing that directly involves identifiable members of the public. SUHAKAM's assessment indicates that the safeguards proposed for functions involving purely technical testing are generally appropriate. However, the framework becomes progressively less developed as the degree of human participation and potential rights impact increases.

Functions 1 and 2 (isolated code execution and non-production development environment):
These functions are appropriately characterised as low-risk technical activities undertaken within controlled environments. SUHAKAM does not have specific human rights concerns regarding these functions as presently described.

Function 3 (integration platform to simulate or test data exchange between systems):
Where this function involves the use of real or identifiable data to test system-to-system data exchange, it may give rise to privacy and data protection considerations that are not expressly addressed elsewhere in the fact sheet. Accordingly, SUHAKAM considers that the safeguards described for Function 4, including anonymisation, synthetic data generation, logging and monitoring should apply equally where Function 3 relies on real or real-derived data.

Function 4 (controlled access to data through anonymisation/synthesis, logging, output checking):

SUHAKAM welcomes this function as the most comprehensively developed privacy safeguard within the proposed sandbox framework. Nevertheless, it could be strengthened by expressly grounding it in the data protection principles of legality, necessity, proportionality and purpose limitation reflected in the Doha Declaration.

In addition, SUHAKAM recommends clarifying that the concept of 'safety' within this function extends beyond the technical integrity of the testing process to include the protection of the rights and interests of data subjects.

Function 5 (evaluating models for accuracy, fairness, robustness and safety, producing assurance evidence for certification):

SUHAKAM welcomes the express reference to fairness, recognising that fairness assessments are fundamental to human rights-based AI governance. However, the Bill does not specify the standards against which fairness is to be evaluated or the protected characteristics that should be considered.

SUHAKAM recommends that fairness assessments be informed by the equality and non-discrimination standards reflected in CEDAW, CRC and CRPD. Furthermore, where fairness evaluations are undertaken for certification purposes, consideration should be given to independent verification of the assessment methodology to enhance public confidence and reduce potential conflicts of interest. This recommendation is further supported by the CEDAW Addendum to General Recommendation No. 30 which highlights the increasing prevalence of AI-enabled and technology-facilitated violence against women and girls[72] and calls for regulatory frameworks capable of addressing gender-specific harms.[73] Therefore, fairness testing should also include gender-based impacts alongside other protected characteristics.

Function 6 (testing AI-enabled products with limited real users, time-bound pilots or restricted markets):

From a human rights perspective, this function warrants the highest level of scrutiny as it involves real, identifiable individuals interacting with AI systems that have not yet been fully deployed.

[72] Paragraphs 83 and 84.
[73] Paragraph 86.

The Bill currently describes no requirement for informed consent, no express right for participants to withdraw, no enhanced protections for children or other persons in vulnerable situations and no dedicated complaints or grievance mechanism for sandbox participants.

These safeguards are particularly important given that the EU AI Act's provisions on real-world testing outside regulatory sandboxes for high-risk AI systems expressly require informed consent, participant protection and the ability to withdraw from participation at any time without detriment.[74]

Function 7 (assessing new workflows, organisational processes or policies in a controlled setting):

Where this function is used to trial a new public-service process, for example, a new eligibility assessment process for social assistance, individuals may be affected without necessarily being aware that they are participating in an experimental process.

Accordingly, many of the participant-protection concerns identified under Function 6 arise equally in this context. Additional safeguards are warranted because access to public services should not depend on whether an individual happens to be assigned to an experimental workflow rather than the standard administrative process.

Function 8 (simulated environments for training and competence-building, such as cyber ranges or flight simulators):

This function is generally low risk for third parties. However, where simulated environments are developed using real or real-derived case data, for example, law enforcement records used to construct training scenarios, the data protection safeguards recommended under Function 4 should apply equally to this function.

Function 9 (permitting regulated activities or innovations to be tested under a relaxed regulatory environment):

The description of this function appears to require further clarification given the Bill's earlier statement that the sandbox is not intended to operate as a deregulated exception.

Accordingly, SUHAKAM recommends that the concept of a ‘relaxed regulatory environment’ be clearly defined. It should expressly exclude any relaxation of the non-derogable human rights protections identified throughout this submission.

[74] Article 61 of the EU AI Act.

Recommendations:

- Extend the anonymisation/synthetic-data safeguards described for Function 4 to any data used under Functions 3 and 8 involving real or real-derived data.
- Require independent verification of fairness/bias evaluation methodology under Function 5, drawing upon the equality standards reflected in CEDAW, CRC and CRPD.
- For Functions 6 and 7, introduce a mandatory informed consent framework, including clear notification that participation forms part of an experimental process, the right to withdraw at any time without detriment, enhanced protections for children and other persons in vulnerable situations, appropriate independent ethics review where relevant and an accessible participant grievance mechanism.
- For Function 7 specifically, require that any public-service trial operate in parallel with the standard administrative process so that an individual's access to public services is not adversely affected by participation in an experimental workflow, and ensure that participants retain the same review and appeal rights available under the standard process.

8.4 Human Rights Concerns of Key Consultation Findings

Statutory power to suspend laws:

SUHAKAM notes with concern the proposal discussed during the consultation regarding a statutory power to suspend existing legal frameworks within sandbox operations. Unlike ordinary delegated rule-making, such a power would temporarily disapply legal protections that have already been established by Parliament for a defined testing population, potentially without those individuals having sought or consented to a reduced level of legal protection.

Given the breadth of this proposed power, SUHAKAM considers that any authority to suspend legal requirements should be accompanied by clear statutory limits and robust procedural safeguards to ensure consistency with the rule of law and the protection of fundamental rights.

Recommendations:

- Enumerate, on the face of the Bill, a clear and exhaustive list of legal protections that can never be suspended within a sandbox.
- Limit any permissible suspension and require that any decision to suspend a legal requirement be published together with reasons, consistent with SUHAKAM's related recommendation

under the AI Enabling Powers Fact Sheet concerning procedural safeguards for the exercise of regulatory powers.

- Clarify which authority is empowered to approve any suspension and require that any decision affecting a rights-sensitive sandbox proposal be made only after consultation with an independent human rights-focused body, a role that SUHAKAM would be well placed to support pursuant to its statutory mandate.

Post-sandbox continuation and graduation:

The stakeholder consultation appropriately raises the question of what occurs after a sandbox trial concludes and an AI system progresses to full deployment, including whether mechanisms exist to support post-sandbox scaling.

From a human rights perspective, this transition gives rise to at least two issues that are not expressly addressed in the current framework. First, whether participants whose data or interactions contributed to the development of the AI system are informed when that system subsequently enters commercial or public deployment. Second, whether data and other information generated during the sandbox are retained, anonymised or deleted pursuant to a clear legal basis once the sandbox concludes.

Recommendations:

- Require a documented human rights impact assessment, consistent with SUHAKAM's recommendations in this submission as a prerequisite for graduation from the sandbox, alongside any technical, commercial or operational readiness assessment.
- Require that sandbox participants be informed where their data or interactions contributed to an AI system that subsequently progresses to full deployment and establish clear rules governing the lawful retention, anonymisation, deletion or continued use of sandbox-generated data following graduation.

Extension of existing sandboxes and sectoral capacity:

The proposal to extend existing regulatory sandboxes, including those operating within the financial sector or broader innovation initiatives such as Meranti to accommodate AI governance represents a pragmatic use of existing regulatory infrastructure.

Nevertheless, SUHAKAM considers that existing sector-specific sandbox models may require additional safeguards before being applied to AI systems operating in other sectors. A framework originally designed to address prudential or consumer protection risks may not adequately capture

the broader human rights considerations arising in sectors such as healthcare, education, social protection or public administration, including issues relating to discrimination, privacy and the rights of children and persons with disabilities.

This reflects SUHAKAM's broader observation throughout this submission that sectoral expertise, while valuable, should not be assumed to encompass the specialised human rights considerations associated with AI governance.

Recommendations:

- Where an existing sandbox is extended to cover AI systems beyond its original sectoral mandate, require that its governance framework be supplemented by the human rights safeguards recommended throughout this submission, including informed consent, vulnerability assessments, independent fairness verification and appropriate human rights oversight, rather than assuming that existing regulatory arrangements can be applied without modification.

8.5 Rightsholders and Participant Representatives

Beyond the specific safeguards recommended above, SUHAKAM highlights several additional considerations that should inform any sandbox function involving the participation of real individuals.

Children:

Consistent with the best interests principle under Article 3 of the CRC, the child's right to be heard under Article 12 of the CRC and CRC General Comment No. 25, SUHAKAM recommends that children should not be included in sandbox testing under Functions 6 or 7 unless their participation is necessary and demonstrably in their best interests; an independent child rights and ethics review finds the risks proportionate and adequately mitigated; guardian consent and the child's informed assent, according to age and maturity, are obtained; and the child may withdraw without detriment.

Persons with disabilities:

Consistent with Articles 4(3) and 12 of CRPD, any informed consent process for sandbox participation should respect the legal capacity of persons with disabilities to make decisions on their own behalf, supported where necessary through supported decision-making arrangements rather than substituted decision-making. Where a sandbox trial relates specifically to disability services or

assistive technologies, representative organisations of persons with disabilities should be closely consulted in accordance with Article 4(3) of the CRPD.

Participant representativeness:

SUHAKAM also notes an important intersection between technical assurance and human rights considerations. If the limited user population recruited for a pilot under Function 6 is not broadly representative of the population likely to use the AI system once deployed, for example, if participation is concentrated among digitally literate urban users, then the fairness evidence generated under Function 5 may not accurately predict the system's performance across the wider population.

Accordingly, SUHAKAM recommends that sandbox governance incorporate an assessment of participant representativeness before fairness assurance evidence is relied upon for certification or broader deployment decisions.

9. CONCLUSION

SUHAKAM appreciates NAIO's efforts to develop a coherent, forward-looking and principle-based national AI governance framework as well as the openness and constructive engagement demonstrated throughout this consultation process.

Overall, SUHAKAM is encouraged by the Bill's adoption of a risk-based, principles-driven approach to AI governance. The recommendations contained in this submission are intended to strengthen the framework by ensuring that the promotion of innovation is accompanied by clear, effective and enforceable safeguards for human rights. In particular, SUHAKAM considers that embedding human rights considerations from the outset, including through stronger accountability mechanisms, meaningful human rights due diligence, effective access to remedy, protection for persons in vulnerable situations and transparent governance arrangements will enhance both the legitimacy and long-term effectiveness of Malaysia's AI governance framework.

As artificial intelligence continues to evolve rapidly, an adaptive regulatory framework will be essential. Equally important is ensuring that such adaptability is underpinned by the principles of legality, transparency, accountability, non-discrimination and meaningful public participation. Maintaining this balance will help foster public trust while ensuring that technological advancement remains consistent with Malaysia's constitutional values and international human rights commitments.

Therefore, the recommendations above are offered in a constructive and collaborative spirit with the shared objective of ensuring that Malaysia's AI Governance Bill enables responsible innovation while providing a clear, proportionate and enforceable baseline of human rights protection for everyone in Malaysia, particularly those who may be disproportionately affected by AI-related harms.

SUHAKAM would welcome the opportunity to continue providing technical assistance as the Bill progresses through the legislative process and looks forward to continued collaboration with NAIO in further refining the proposed framework ahead of its tabling in Parliament.

Finally, SUHAKAM respectfully recommends that NAIO undertake a further round of public consultations on the draft legislative text of the AI Governance Bill prior to its tabling in Parliament. Such consultations would enable stakeholders to examine the proposed provisions in greater detail

and provide informed feedback and recommendations, thereby helping to ensure that the Bill is comprehensive, practical, and responsive to the interests of all relevant stakeholders.