Annex A

Supervisory Guidance for Responsible Adoption of Artificial Intelligence

Purpose

This document sets out the Hong Kong Monetary Authority (HKMA)'s latest supervisory expectations for responsible adoption of artificial intelligence (A.I.) by Authorized Institutions (AIs). It supersedes the "High-level Principles on Artificial Intelligence" dated 1 November 2019.

Background

The HKMA, as part of its interactive and iterative supervisory approach, supports AIs' practical experimentation within risk-controlled sandbox environments. The HKMA also prioritises collaboration with the industry to promote responsible innovation, underpinned by high-level guiding principles that establish fundamental guardrails while providing flexibility and practicality for real-life situations. This reflects an inclusive approach that accommodates different forms of A.I. adoption, balances robust risk management with pragmatism and feasibility, and supports effective, fit-for-purpose solutions.

With A.I. adoption reaching a more mature state, coupled with the successful conclusion of the second cohort of the Generative Artificial Intelligence (GenA.I.) Sandbox[1], the HKMA updates its supervisory guidance to reflect the practical implementation experience gained through the Sandbox. This document has also incorporated the latest guidance issued by international bodies including the Financial Stability Board[2].

The updated guidance is anchored around three pillars that the HKMA considers to be defining characteristics of responsible A.I. adoption, and which collectively form the acronym "**ART**":

- **Accountability** — ensuring clear ownership, governance, decision-making authority and assurance throughout the A.I. lifecycle;
- **Resilience** — enabling secure, robust and well-controlled A.I. adoption that operates safely, fails safely and recovers effectively; and
- **Trust** — prioritising fair, transparent, explainable and privacy-conscious A.I. use that is responsive to customers and other affected parties.

[1] HKMA Press Release, "HKMA announces second cohort of GenA.I. Sandbox to advance responsible A.I. innovation", issued on 15 October 2025
[2] Financial Stability Board Report, "Sound Practices for Responsible Adoption of Artificial Intelligence (AI)", issued on 15 October 2026

Application

AIs are expected to apply the "ART" Principles proportionately to their A.I. adoption and manage associated risks, drawing on existing governance and risk management arrangements as appropriate. The Principles should also be read in conjunction with the report on "Practical Insights from the Second Cohort of the GenA.I. Sandbox", which sets out some non-exhaustive examples of how the Principles may be applied in practice, as well as other prevailing supervisory and regulatory requirements.

AIs should take appropriate steps to uplift their policies, procedures and controls in line with the new guidance in a timely manner.

Any questions may be directed to the Technology Risk Team at ebanking@hkma.iclnet.hk.

Supervisory Expectations

A. Accountability

1. **Board and senior management are accountable for an AI's A.I. adoption through setting effective strategy and governance arrangements.**

   The board should approve an A.I. strategy which clearly sets out the AI's approach and risk appetite for adopting A.I. across the institution's operations. As part of this, a clear governance framework should establish roles and responsibilities across the three lines of defence for assessing and approving an A.I. use case prior to deployment, and for continually monitoring and managing the risks post-deployment.

   Senior management should regularly review the continued suitability of an AI's A.I. strategy and governance framework with regard to the prevailing risk landscape. Where warranted, senior management should make prompt adjustments with the approval of the board.

2. **Assign an "owner" for each A.I. use case deployed, and ensure effective human oversight of A.I. generated outcomes.**

   The assigned owners will be responsible for defining each system's authority, overseeing the effectiveness of controls and monitoring, and ensuring prompt escalation and remediation of issues encountered. Human oversight may include "human-in-the-loop", "human-on-the-loop" or

"human-in-command" arrangements, supported by A.I. where appropriate, with risk-based triggers for human approval or intervention[3]. Reviewers involved in such intervention should have sufficient competence, information, time and authority to challenge, override or stop the A.I. system.

AIs should obtain assurances around the effectiveness of such oversight (e.g. through testing) and ensure that staff exercise appropriate independent judgement when considering A.I. outputs. In line with their overall responsibility for an institution's A.I. adoption, the board and senior management remain accountable for the outputs generated.

3. **Maintain an up-to-date inventory of A.I. use cases approved for deployment within the institution, and restrict the use of solutions not covered in the inventory.**

The inventory should enable an AI to have full visibility of the institution's use of A.I., including solutions that are internally developed, provided by third parties or embedded. To facilitate effective stock-taking, AIs should clearly define what counts as "A.I. use", standardise the attributes recorded in the inventory, and clearly assign responsibilities for keeping the records up-to-date. The inventory should generally cover an A.I. use case's purpose, owner, provider and model, risk classification, lifecycle and approval status, dependencies, as well as key controls around risk management, access and monitoring.

AIs should clearly communicate expectations around A.I. use to staff, e.g. no adoption of unapproved tools, or the use of approved tools for unapproved purposes[4]. AIs should also implement controls to manage "shadow A.I." adoption.

[3] In the HKMA circular "Consumer Protection in respect of Use of Generative Artificial Intelligence" dated 19 August 2024, it was stated that "during the early stage of deploying customer-facing GenAI applications, authorized institutions should adopt the "human-in-the-loop" approach, i.e. having human to retain control in the decision-making process to ensure the model-generated outputs are accurate and not misleading". The HKMA recognises that as A.I. development advances, a "human-on-the-loop" or "human-in-command" arrangement could be more appropriate for some use cases. In deciding the appropriate human oversight arrangement for customer-facing A.I. applications, AIs may follow the guidance set out in this circular.

[4] For solutions provided by third-party service providers, an institution should ensure compliance with contractual and licensing terms, and implement controls against circumvention.

R. Resilience

4. Manage A.I. adoption in a way that is commensurate with the materiality and risks of a use case.

Following the institution's strategy and governance arrangements approved by its board, an AI should establish a framework to assess the materiality and risks of each A.I. use case under consideration. This initial assessment[5] should determine the rigor of review and approvals required before the use case can be deployed. Use cases that are assessed to be material and/or of higher risk should be subject to more stringent requirements, including senior management sign-off and oversight. Control and audit functions should demonstrate effective challenge.

This risk-based approach should persist through the full lifecycle of the use case's deployment, and inform how an AI approaches post-deployment processes, including performance monitoring, contingency planning and incident management. Specific expectations across the lifecycle are set out below.

- Testing, review and validation: An AI should conduct end-to-end testing and reviews of an A.I. use case prior to deployment and before material changes to A.I. systems. The outcome should be reviewed against defined acceptance criteria, and generally cover key performance metrics such as reliability, security, fairness and output variability under representative and stressed conditions. Where a material or higher-risk use case is concerned, it should be subject to independent validation and where appropriate, adversarial testing. Any material findings and limitations should be addressed through remediation or approved mitigating controls. Residual risks should be documented and accepted by the approving authority following an AI's governance framework before deployment or continued use.

- On-going monitoring: An AI should tailor its approach to monitoring based on the nature of the A.I. use case. For instance, for A.I. use cases that take actions, AIs should apply equivalent and risk-proportionate controls that enable it to oversee the entire process (e.g. intermediate

[5] The framework should support the AI to comprehensively assess and conclude the materiality and risks of the use case from end-to-end. During the process, the AI should generally take into consideration such factors as what the purpose/objective of the use case is (e.g. what types of data will be used, what types of outcomes or actions will be generated), what the impact would be if the use case fails (e.g. potential harm to customers, consequence for the AI's operations assuming heavy reliance), and the maturity of the use case (e.g. in terms of technological complexity, whether fit-for-purpose models exist and have a proven track record, novelty to the AI vis-à-vis technical competence and experience) etc.

steps leading up to final outcomes). At the same time, for A.I. systems that are used to test or monitor other A.I. systems, or "A.I. versus A.I.", institutions should cater for potential weaknesses or biases that may be exhibited by the A.I. systems involved. In line with Principle 2 above, appropriate human oversight and corroborating checks should be in place. More generally, the ongoing monitoring arrangements[6] should enable AIs to promptly identify any potential issues with the A.I. use case, and take remedial actions in a timely manner.

- Response and recovery: In the event of incidents or other material issues identified through on-going monitoring, an AI should be able to:

  (i) respond rapidly, including to recover operations in the event of disruptions. In this connection, an AI should maintain credible fallback and recovery arrangements with sufficient capacity and expertise prepared to address different severe but plausible scenarios, including third-party failures; and/or

  (ii) contain, suspend, or even decommission a given A.I. use case in the event of errant behaviour. In the case of more material or higher-risk use cases, a kill switch or equivalent should be built-in to enable prompt suspension, access revocation and prevention of further actions by the affected model.

  Significant incidents should be reported to the HKMA in accordance with prevailing supervisory requirements.

- Auditable records: An AI should maintain documentation covering the full adoption lifecycle of an A.I. use case, including the risk assessment and selection process, testing, approvals, deployment records, as well as any material changes made over the course of implementation. Where an A.I. use case involves the generation of higher-risk and/or material outputs, records[7] enabling auditing and outcome tracing should be maintained as far as practicable. These records should be kept with regard for relevant regulatory requirements, including in terms of data protection, confidentiality and retention.

[6] An effective monitoring plan should cover roles and responsibilities, indicators and escalation thresholds, response actions and review frequency. It should also highlight risks that may merit particular attention, such as data or model drift, misuse, safeguard failures, and material security and conduct risks, including inappropriate bias and unfair outcomes.

[7] Examples of records may include the sources, references, models and model versions used, as well as relevant model configurations and inputs. Where applicable, including for generative A.I. use cases, records may also include prompts and evidence supporting the reasoning underlying material outputs.

The materiality and risk assessments should be regularly revisited and refreshed in line with changes to the risk landscape.

5. **Implement effective risk management frameworks and which cater for the specific needs of A.I., recognising that A.I. adoption is associated with new risks and may also complicate existing risk domains**. As part of this, AIs are expected to pay particular attention to the following risk areas:

   - Data governance: Institutions should ensure that the data used for A.I. use cases are fit-for-purpose, and subject to appropriate safeguards, including as required by law and customer protection mandates. To this end, institutions should maintain an effective data governance framework to set out the types of data which may be used, for what specific purposes, and the corresponding controls that should be in place with regard to the materiality and risks of an A.I. use case. Data use should not go beyond these established and approved parameters. Responsibility and accountability for data management should be clearly assigned.

     During the process, institutions should pay particular regard to how personal data is used, including to ensure full compliance with the Personal Data (Privacy) Ordinance and other applicable requirements. Institutions should also establish contractual safeguards to prevent A.I. service providers from re-using customers' personal data or the institution's other confidential information for model training or improvements unless specifically authorised by the institution and permitted under applicable requirements. As far as possible, an AI should also implement controls to minimise the risk exposure of personal data, such as by masking or using pseudonymous identifiers.

   - Third-party risks: Where an AI relies on external A.I. service providers to deliver an A.I. use case, the institution remains responsible for all A.I.-generated outcomes, and the overview robustness of the arrangement, including implications for the AI's operational resilience. Accordingly, A.I. service providers should be subject to proper due diligence and risk assessment before selection and engagement, bound by formal contractual safeguards[8], as well as ongoing monitoring and performance reviews. An AI should also assess concentration, continuity, substitutability and exit risks, with particular attention to material and

[8] Subject to the use case at hand, this should generally cover the service provider's responsibilities, as well as provisions guaranteeing audit rights, data protection, notifications around material changes and incidents, as well as response and recovery.

higher-risk use cases that may depend on a small group of dominant players.

- Cyber risks: An AI should apply layered and effective cybersecurity and access controls to guard against malicious attacks, including those that may try to manipulate A.I.-generated outcomes (e.g. data poisoning) or obtain unauthorised access to data held by A.I. systems (e.g. prompt injection, safeguard bypass). These controls should be developed and continually enhanced with regard to the prevailing risk landscape, including emerging tactics and latest technology developments (e.g. frontier A.I.). Generally speaking, AIs should not rely solely on model instructions or filters to obtain assurances in this regard.

6. Ensure sufficient resources and expertise, including talent, to continually deliver A.I. use cases in a safe and responsible manner.

Senior management should commit adequate resources to deliver the institution's A.I. strategy, and adopt a forward-looking mindset when considering resource and capacity planning. Depending on the degree of A.I. use, the institution should scale up organisation-wide communication and exposure to A.I. in a commensurate manner, and also ensure specialist skills are maintained through role-specific training.

T. Trust

7. AIs remain accountable for A.I.-generated outputs, including actions, and for assuring the quality of outputs.

An AI should apply appropriate safeguards to limit the incidence of unreasonable outcomes based on how the outputs will be used, and the consequences of such applications. Institutions should stay vigilant against the risks of A.I. outputs being inaccurate or incomplete, and which may manifest as plausible but fabricated outputs ("hallucinations"). Where there may be material impact, say to a customer, the outputs should be subject to more robust review and controls (e.g. required to be anchored in reliable reference sources), and/or subject to human validation. An AI should clearly define when outputs must be withheld, referred for review or restricted from triggering actions when reliability or safety standards are not met.

8. Ensure that A.I. use cases are suitably explainable and transparent, including to internal and external stakeholders.

Institutions should understand and document the degree to which each A.I. use case is explainable, as well as any limitations in explainability. The level of understanding, documentation and risk mitigating measures (if any) should be commensurate with the assessed materiality and risks of the use case. AIs should also provide appropriate amounts of information (e.g. on the underlying A.I. system's purpose, functions, limitations and basis for material outcomes) to internal and external stakeholders, as applicable, to enable a level of understanding appropriate to stakeholders' needs. Generally speaking, where the objectives of a use case can be achieved, AIs should prioritise A.I. arrangements that are more explainable.

9. Ensure that A.I. adoption is ethical and treats customers fairly.

Institutions should ensure that A.I.-driven decisions do not discriminate or unintentionally show bias against any group of consumers. Where outputs may be used with actual consequences for customers, an AI should conduct regular monitoring, benchmark and assess the reasonableness of outputs (e.g. through sample review) and take mitigating action if anomalies are noted. Affected parties should also be able to seek clarification, review or re-assessment of the outputs, and institutions should empower human reviewers to challenge and correct the outcomes. An AI should also be satisfied that its A.I. adoption complies with its corporate values and ethical standards, and upholds consumer protection principles. Before providing a service powered by A.I. technology, an AI should clearly inform consumers of the use of A.I. and the relevant risks. Existing consumer protection expectations, including around disclosure and opt-out arrangements where applicable, continue to apply.