HONG KONG MONETARY AUTHORITY
香港金融管理局

Discussion Paper on Agentic A.I. Governance in the Banking Sector

November 2026

Contents

1. Executive Summary **4**

2. Introduction **10**

2.1 Background 10
2.2 Scope of this Paper 10
2.3 Status of this Paper 11
2.4 Structure of this Paper 11

3. Agentic A.I. **12**

3.1 Definition 12
3.2 Autonomy of agentic systems 13
3.3 Agentic A.I. use cases in banking 14
3.4 Direction of travel 15

4. Challenges and Risks **16**

4.1 Framing 16
4.2 Band A – Risks shared with general A.I. deployments 16
4.3 Band B – Risks heightened by autonomy, speed and scale 17
4.4 Band C – Risks particularly associated with delegated authority and action-taking capability 19
4.5 Potential systemic implications 21

5. Overarching framework: Intent-anchored accountability **23**

5.1 Objectives and supervisory principles 23
5.2 Intent-anchored accountability 24
5.3 Accountability and ownership arrangements 25
5.4 Accountability where users instruct an agent 26
5.5 Customer-facing deployments 27

6. Lifecycle governance **30**

6.1 Building on existing governance 30
6.2 Lifecycle stages 30

7. Specific governance elements **35**

7.1 Risk-integrated agent harness 35
7.2 Human oversight and effective intervention 36
7.3 Traceability and internal deployment records 38

8. Registry of material A.I. agent **42**

8.1 Classification 42
8.2 Notification to the HKMA 44
8.3 Ongoing monitoring 45
8.4 Purpose of notification 46

9. Interactions beyond the institution **50**

9.1 External interactions and approved boundaries 50

9.2 Different forms of external agent interaction 50
9.3 Know Your Agent 51
9.4 Acceptance and execution of external-agent instructions 52
9.5 Shared responsibility frameworks 53

10. Way Forward 56

10.1 Next steps 56
10.2 How to respond 57

1. Executive Summary

Artificial intelligence (A.I.) is evolving from systems that support analysis, prediction and content generation towards agentic A.I. systems that can plan, coordinate and perform multi-step tasks with varying degrees of autonomy. Within defined authority boundaries, these systems may access data, use tools, interact with internal or external systems, and involve multiple specialised agents working together towards a specified business objective.

Agentic A.I. offers significant opportunities for Authorized Institutions (AIs) to enhance risk management, anti-fraud capabilities, operational efficiency and customer experience. At the same time, its increased autonomy, speed, connectivity and reliance on interconnected models, data, tools and third-party services may give rise to new or heightened risks. These include risks relating to ambiguous accountability, unintended actions, data leakage, financial damage, operational disruptions, cyber threats, or failure to effectively identify, contain and recover from incidents.

In October 2026, the Hong Kong Monetary Authority (HKMA) launched a new Framework for Responsible Adoption of Artificial Intelligence, anchored around the principles of Accountability, Resilience and Trust, or collectively called "ART"[1]. As part of the ART Framework and in line with the latest international guidance[2], the HKMA has updated its *Supervisory Guidance for Responsible Adoption of Artificial Intelligence*[3] (Supervisory Guidance), which sets out the HKMA's expectations for AIs to apply proportionate governance and risk-management arrangements throughout the A.I. adoption lifecycle. The HKMA also issued a *Concept Note on Governance of Agentic Artificial Intelligence*[4] (Concept Note), setting out preliminary considerations arising from initial observations from the Generative Artificial Intelligence (GenA.I.) Sandbox.

This Discussion Paper builds on the ART Framework, the latest Supervisory Guidance and the Concept Note. It seeks industry views on possible approaches to strengthen governance arrangements as agentic A.I. develops in the banking sector. The objective is to support responsible innovation while clear accountability, effective controls and appropriate intervention are adequately implemented for agentic A.I. systems.

The HKMA considers that governance arrangements for agentic A.I. should be risk-based and proportionate. The intensity of governance should reflect, among other things:

- the materiality of the use case;
- the degree of autonomy involved;
- the business objective and authority delegated to the agent;
- the nature of the data, tools and systems that the agent can access;
- the complexity of the agentic workflow and its dependencies;
- the potential impact on customers, the institution and the wider financial system; and
- the reversibility of the actions that the agent may take.

---

[1] HKMA announces "ART" Framework to promote responsible A.I. adoption, HKMA (2026) https://www.hkma.gov.hk/eng/news-and-media/press-releases/2026/10/20261015-3/

[2] Sound Practices for Responsible Adoption of Artificial Intelligence (AI), FSB (2026) https://www.fsb.org/2026/10/sound-practices-for-responsible-adoption-of-artificial-intelligence-ai-final-report/

[3] Supervisory Guidance for Responsible Adoption of Artificial Intelligence, HKMA (2026) https://brdr.hkma.gov.hk/eng/doc-ldg/docId/getPdf/20261015-2-EN/20261015-2-EN.pdf

[4] Concept Note on Governance of Agentic Artificial Intelligence, HKMA (2026) https://brdr.hkma.gov.hk/eng/doc-ldg/docId/getPdf/20261015-1-EN/20261015-1-EN.pdf

The Discussion Paper is centred on **intent-anchored accountability**. An agent should operate only in pursuit of an approved business intent, within a defined delegation of authority. Accountability for that delegation, and for the outcomes arising from it, should remain with identified individuals and functions within the AI.

The Discussion Paper seeks views on the following areas.

Intent-anchored accountability

An A.I. agent's actions are driven by its interpretation of an instruction or business objective. Before an agentic A.I. system is deployed, an AI should define the business intent that the agent is authorised to pursue. This should be sufficiently specific to support risk assessment, control design, and subsequent oversight.

AIs should clearly identify, collectively as "agent intent":

- the business objective and intended use of the agent;
- the tasks, decisions or actions for which it is authorised;
- the scope of authority delegated to it;
- activities that are prohibited or require escalation;
- applicable limits, tolerances and approval conditions;
- intended users and affected parties; and
- circumstances in which the agent's authority should be reviewed, restricted or withdrawn.

AIs should then establish a clear connection between the agent intent and human owners within the organisation. Relevant roles may include business, technical and control owners responsible for the agent.

While technology functions play an important role in building and maintaining agentic A.I. systems, accountability for outcomes should remain clearly assigned within the institution. Such accountability should not be diluted by the use of external models, cloud services, data providers, tools or other third-party components.

The Discussion Paper explores how AIs could define, approve and document the business intent, delegated authority, ownership and accountability. It also seeks views on arrangements for agents operating across business lines, relying on multiple vendors or coordinating other agents.

Lifecycle governance

The accountable owner should be responsible for managing risks at each stage of the lifecycle of an agentic A.I. system, from **design and development**; to **deployment and activation**; as well as **operations and retirement**.

The Discussion Paper seeks views on how existing governance and risk-management arrangements may be adapted to address the specific characteristics of agentic A.I., including the interaction between models, prompts, data, tools, permissions, workflows and human users.

Areas for consideration include:

- **Design and development**
  - assessing the materiality and risks of an agentic A.I. use case before deployment and on an ongoing basis;
  - defining permitted objectives, actions, data flows, tools, systems, user access and escalation arrangements;
  - managing the risks arising from external models, tools, data sources, cloud services and other third-party dependencies;
- **Deployment and activation**
  - testing agentic workflows under representative, stressed and adversarial conditions;
  - verifying that the production environment reflects the approved business intent and authority;
- **Operations and retirement**
  - monitoring agent behaviour, intermediate steps and final outcomes against defined thresholds and indicators;
  - maintaining sufficient records to support auditability, tracing and post-incident investigation; and
  - managing material changes to models, prompts, tools, data sources, workflows, permissions or other components of an agentic A.I. system.

Risk-integrated agent harness

Intent-anchored accountability requires an AI to ensure that the agent's approved business intent is pursued in a manner consistent with the AI's risk appetite, risk-management framework and relevant internal policies.

Across the lifecycle of an agentic A.I. system, AIs should translate relevant risk tolerances, control requirements and escalation thresholds into agent-specific operational boundaries. This should not rely solely on high-level prompts or instructions provided to the model. Rather, the AI should implement controls through the agent harness: the environment through which the agentic A.I. system operates. This may include instructions, workflow design, permissions, tools, application interfaces, data controls, dynamic guardrails, velocity limits, deterministic validation checks, approval gates, monitoring and escalation mechanisms.

The Discussion Paper seeks views on how AIs can translate risk appetite into practical, testable and enforceable controls. It also seeks views on the extent to which such controls should be implemented through the agent harness, building upon existing policies and procedures.

Human oversight and effective intervention

As agentic A.I. systems may act at speed and perform connected tasks across different systems, AIs should maintain both meaningful human oversight and effective intervention arrangements. The appropriate approach will depend on the nature and potential impact of the use case.

An effective stop mechanism—often referred to as a "kill switch"—should enable an institution to bring an A.I. agent to a safe state promptly where risk thresholds are breached or an incident occurs. Depending on the architecture, this may include halting task execution, revoking access or credentials, disabling connections to external tools or systems, pausing outbound communications, activating a fallback process and preserving relevant logs and evidence.

For some use cases, continuous human approval before an action is taken may be appropriate. For others, human-on-the-loop or human-in-command arrangements may be suitable, provided that the institution retains effective control, including the ability to understand, challenge, constrain, pause or withdraw the agent where necessary. One possible approach is the use of "A.I. versus A.I." to scale oversight capabilities.

The Discussion Paper seeks views on the intervention capabilities that AIs should maintain for material A.I. agents. These may include the ability to:

- correct outputs or behaviours;
- restrict permissions, tools, systems or user access;
- impose additional review or approval requirements;
- pause, suspend or withdraw an agentic A.I. deployment;
- roll back material changes; and
- activate fallback arrangements.

The Discussion Paper also seeks views on a possible supervisory approach where material risks associated with agentic A.I. are not adequately managed. This may include supervisory engagement, remedial measures, restrictions and, in exceptional circumstances, suspension-related measures, or "kill switch", for the agent.

Traceability and Registry of material A.I. agents

Intent-anchored accountability requires an AI to maintain sufficient evidence of how a material agent was authorised, how it operates, and whether it has remained within its approved intent and authority. AIs should maintain records that support internal accountability, independent assurance as well as incident investigation and remediation.

The Discussion Paper seeks views on a proposed Registry of material A.I. agents. Under the proposed approach, AIs would maintain up-to-date internal records of material A.I. agents and submit relevant information on such agents to a centralised Registry, such as one maintained by the HKMA.

The proposed Registry is intended to enhance visibility, ascertain accountability and support risk-based supervision of material agentic A.I. deployments. It could also support the identification of common dependencies, concentrations and potential vulnerabilities across the banking sector.

The proposed Registry may capture information including:

- the agent's purpose, intended use and deployment status;
- its business, technical and control owner, as well as authorised users;
- its capabilities and access, including any data leaving the AI and the recipients of such data;
- its infrastructure and dependencies, including those shared with other deployments;
- its stop mechanism, fallback arrangements and log locations;
- its change and suspension history, as well as arrangements for memory, state and credentials at retirement; and
- its materiality assessment, underlying rationale and next review date.

The Discussion Paper seeks views on the scope of the proposed Registry, the criteria for determining materiality, the information to be captured, submission arrangements and the operational feasibility of the proposal.

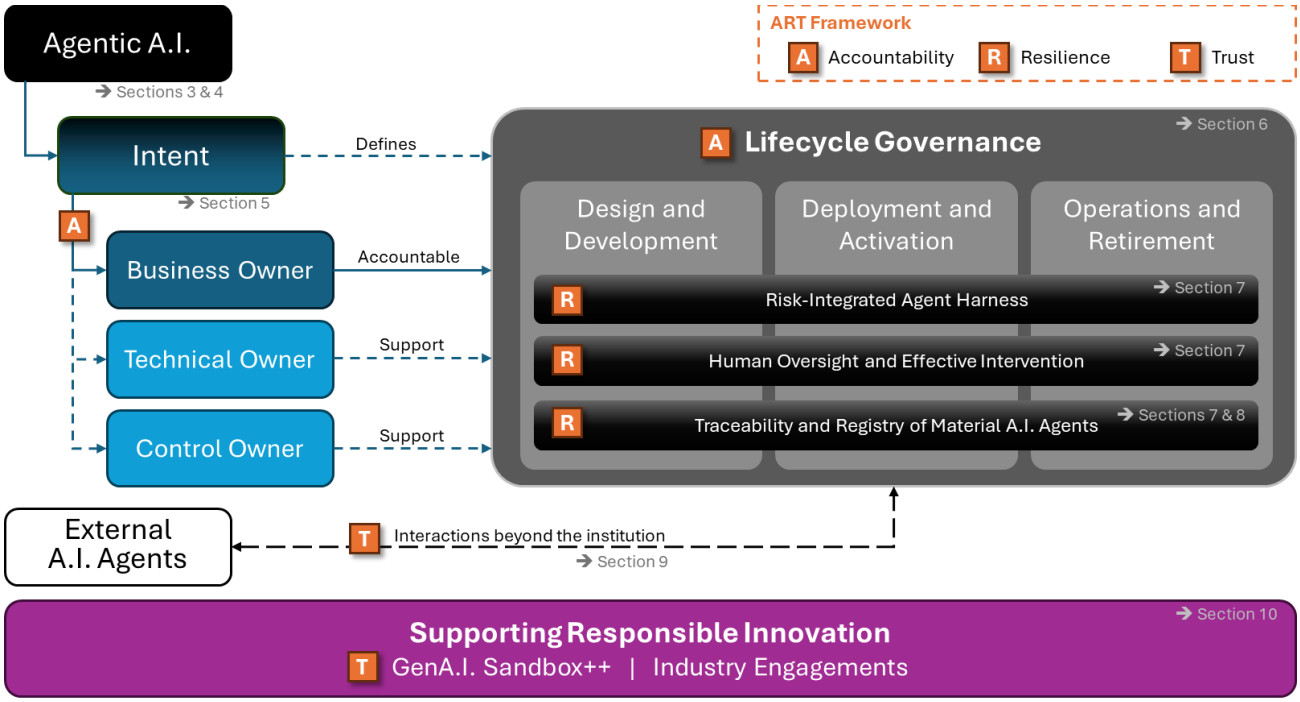
Interactions beyond the institution

Agentic A.I. systems may interact with external services, data sources, tools, platforms and other agents. They may also receive instructions from agents acting under a customer's or counterparty's mandate. These interactions may create additional risks relating to identity, authority, data exchange, malicious or inappropriate inputs, changing external capabilities, and uncertainty over responsibility where multiple parties are involved.

The Discussion Paper seeks views on arrangements for managing these interactions. This includes identifying and approving external connections in advance; treating external content as untrusted input for the perspective of security; and applying appropriate controls where an AI accepts instructions from an external agent. The Paper also considers a risk-based "Know Your Agent" approach to establish an agent's identity, its principal and its delegated authority, as well as possible shared-responsibility frameworks among AIs, agent providers or operators, customers and counterparties. Such arrangements should not dilute an AI's accountability for its own systems, channels, acceptance arrangements and boundary controls.

Supporting responsible innovation

The HKMA recognises that agentic A.I. remains an emerging area and that AIs are at different stages of capability and adoption. The HKMA will continue to support practical experimentation through the GenA.I. Sandbox++[5], engagement with industry and collaboration with relevant ecosystem partners.



[5] First cohort of GenA.I. Sandbox++, HKMA (2026) https://www.hkma.gov.hk/eng/news-and-media/press-releases/2026/08/20260827-3/

The HKMA's aim is not to prescribe a single technical architecture or operating model. Rather, the proposed approach seeks to ensure that AIs can pursue innovation in a controlled manner, with governance arrangements that are commensurate with the capabilities, materiality and risks of their agentic A.I. use cases.

This Discussion Paper does not itself introduce binding supervisory requirements. It seeks feedback from AIs and relevant stakeholders on the proposals and questions set out in the Paper. The HKMA will take account of responses received, developments in technology, latest international guidance, and practical experience from the industry in considering the next steps, including the potential approach to formulating further guidance and future supervisory requirements.

The first round of discussion responses should reach to the Banking Supervision Department of the HKMA at all-banks-go-fintech@hkma.gov.hk before **1 February 2027** to inform the appropriate way forward.

2. Introduction

2.1 Background

Agentic A.I. capabilities are progressing from development and testing to production deployment. Systems that plan and execute multi-step tasks using tools now appear in a range of enterprise applications, and AIs are piloting, and in some cases deploying them.

The HKMA welcomes this development. Agentic A.I. can support banking operations and customer service, including through more timely service delivery, operational efficiency, and enhanced fraud and financial-crime detection. The HKMA supports responsible innovation, including through the GenA.I. Sandbox++[6] and the Fintech 2030[7] strategy. This support extends to the responsible development and use of agentic A.I..

This Discussion Paper reflects such support. Clear governance arrangements can help institutions assess and manage the risks associated with new technologies. As deployments increase, the HKMA considers it timely to discuss the governance questions raised by agentic systems: who is accountable when software acts, what requirements should apply to all deployments, and whether it might be of merit to establish industry-level visibility of certain deployments. The aim is to enable institutions to deploy agentic systems safely and responsibly, supported by sound governance arrangements and the latest supervisory requirements, including the HKMA's "ART" Framework.

2.2 Scope of this Paper

This Paper discusses governance considerations in the context of AIs' agentic A.I. deployments. The HKMA recognises differences in size, business model and technology maturity, and that many AIs will use vendor-provided agentic capabilities before building their own. Proportionality therefore underpins the preliminary thinking on key issues, including questions on implementation considerations for AIs of different sizes and levels of technological maturity as well as different implementation situations as set out in sections 5 to 9.

Three situations can be distinguished. First, an AI may deploy, use or rely upon an internally developed, vendor-provided or group-provided agentic system which has access to its data, systems or processes in connection with its Hong Kong operations, regardless of whether the system is hosted on its infrastructure. Second, it may accept instructions from an external agent acting under a customer's or counterparty's mandate, such as a corporate customer's agent submitting instructions through the AI's channels. Third, it may encounter agents acting independently for other market participants at arm's length. This Paper addresses the first two situations, where the AI has a degree of control over its own systems, channels or acceptance of agent-initiated instructions. The third situation is primarily a matter for monitoring and, where necessary, coordination among authorities.

---

[6] First cohort of GenA.I. Sandbox++, HKMA (2026) https://www.hkma.gov.hk/eng/news-and-media/press-releases/2026/08/20260827-3/

[7] The HKMA Unveils "Fintech 2030" at the Hong Kong FinTech Week 2025, HKMA (2025) https://www.hkma.gov.hk/eng/news-and-media/press-releases/2025/11/20251103-3/

Where an AI becomes aware of unauthorised use of agentic tools by staff in connection with its business, customer data or systems, it should assess and address the associated risks under its applicable governance, technology risk, information-security and data-protection arrangements.

Existing requirements underpin the governance considerations discussed in this Paper. Non-exhaustive examples of such requirements may include the Banking Ordinance, the Personal Data (Privacy) Ordinance, the relevant modules of the Supervisory Policy Manual (SPM) such as the SPM on Risk Management of E-banking and Operational Resilience, the Code of Banking Practice, the Supervisory Guidance for Responsible Adoption of Artificial Intelligence, the circular on consumer protection in the use of generative artificial intelligence, and the Open API Framework. Where an agentic deployment processes personal data, the Privacy Commissioner's guidance on protecting personal data privacy in the use of agentic A.I. is directly relevant. The thinking in this Paper is designed to sit alongside these requirements and does not restate existing requirements or guidance.

This Paper is outcome-focused and does not propose a prescribed technical architecture or implementation. AIs may adopt different technical approaches, proportionate to the nature and risk of each deployment, to demonstrate that agents remain governed within their approved purpose and authority.

2.3 Status of this Paper

This Discussion Paper presents preliminary views and invites comment. It does not consult on new rules and is not a statement of supervisory expectations. It creates no new obligations and does not modify or interpret existing ones. The HKMA will continue to adopt a collaborative approach. In light of feedback on this Discussion Paper and market developments, the HKMA will consider whether further supervisory considerations relating to agentic A.I. are appropriate and, if so, what form it should take.

2.4 Structure of this Paper

This Paper is structured as follows:

- **Chapter 3** outlines the development of Agentic A.I. and its relevance to the banking sector.
- **Chapter 4** sets out the principal risks and governance challenges associated with Agentic A.I.
- **Chapters 5 to 9** present the HKMA's preliminary thinking and discussion proposals.
- **Chapter 10** discusses the way forward and sets out the feedback timetable and arrangements for submitting comments.

3. Agentic A.I.

3.1 Definition

For the purposes of this Paper, agentic A.I. describes a system that pursues an assigned objective by planning and executing a sequence of actions, using tools and other systems, and adapting as conditions change. The following capabilities are commonly associated with agentic A.I.:

- it is **goal-directed**, working towards an outcome across multiple steps;
- it **plans**, breaking the objective into those steps;
- it **uses tools**, operating other systems and sources to carry the steps out;
- it **has memory**, carrying context across steps and sessions; and
- it **adapts**, changing course as results come in.

These capabilities exist on a spectrum, and a system may exhibit them to varying degrees. Not every capability needs to be present in the same form or to the same extent. Defining agentic A.I. by capability helps keep the description relevant as architectures evolve.

The description above defines the technology, but governance also benefits from considering the purpose for which it is deployed. This Paper therefore takes the intent-anchored approach introduced as part of the "ART" Framework.

The proposed operational definition in this Paper covers systems that interpret an assigned instruction or business intent and have discretion over the actions to take. The governance considerations may also be relevant where a system can initiate or materially influence consequential actions, including where human approval is not substantive due to the high volume of the process. The following features are indicators of systems that may fall within the proposed definition; they are not an exhaustive or cumulative test:

- **It acts through tools**. The system can initiate, trigger, recommend for execution, or execute actions autonomously which directly affect operational aspects, such as customer accounts, transactions, records, controls, or business processes. A system that only produces content for a person to consider and act on, without access to tools or systems that can affect outcomes, would generally fall outside this indicator.
- **It chooses its steps**. Within the AI's defined intent, the system decides which tools to use, how to use them, and in what sequence.

Three familiar categories fall outside the agentic A.I. definition described in this Paper: conventional automation, including robotic process automation, whose steps are fully prescribed; models performing a single fixed task within a prescribed workflow, such as extracting data from documents; and retrieval chatbots or drafting assistants that produce content for a person to act on. While these systems may not fit the agentic A.I. definition, they may still be highly consequential and remain subject to existing supervisory attention. Existing model-risk, technology and outsourcing requirements continue to govern these systems. Any subsequent expansion beyond a system's previously approved purpose or authority may trigger reassessment and change control.

The following examples illustrate the definition. A drafting assistant that prepares an email for a staff member to review and send falls outside it because the assistant produces content and the person acts. If the assistant is connected to the mail system and document store so it can send and file independently, it falls inside because it acts through tools and decides when and how to use them.

However, intermediate scenarios may warrant consideration, such as when the assistant can prepare, route, or queue communications for automated sending, or where human approval is nominal or non-substantive due to high volumes. In such cases, the distinction lies in the nature of the delegated authority and the institution's control arrangements. If substantial authority is delegated, where human oversight becomes routine rather than functional, the system may fall within the definition due to the potential impacts of its actions.

Governance expectations attach to the deployment as a whole: the system, its assigned task, its permissions and its operating environment. The same product performing two different tasks constitutes separate use cases that may be assessed according to their respective purpose, authority, permissions and operating environment. In a multi-agent arrangement, each agent, including the orchestrator, is subject to clear identification, accountability and control, while the multi-agent arrangement may also be governed and assessed as a whole where the orchestrator and its sub-agents collectively pursue the same approved business intent under common ownership and controls.

A sub-agent may be treated separately where it pursues a distinct business intent, exercises materially different authority or operates under different ownership or controls. A reusable agent supporting multiple use cases may be identified as a shared component, while its use within each business context may be assessed as part of the relevant deployment.

3.2 Autonomy of agentic systems

Agentic systems differ in the degree of authority delegated to them. Some only draft or recommend, leaving all consequential action to a person; depending on their capabilities and operating environment, these systems may fall outside the proposed operational definition in section 3.1. Others execute actions only after individual human approval. More autonomous systems act independently within institution-defined rules, such as limits, conditions and permitted counterparties. At the highest level of autonomy, some may receive an objective and determine how to achieve it within defined parameters.

In assessing delegated authority and governance needs, one relevant factor is whether a person approves each action before it takes effect. Other relevant factors include the scale, speed, customer impact, breadth of access, irreversibility, or nature of the action. As discussed further in section 7, approval may become a formality when the volume of actions is too high for meaningful human review.

A deployment may also move along this spectrum without a deliberate decision. Routine product changes can expand permissions and toolsets, approval steps may weaken under high volumes, and users may extend the system to new purposes. Its current position therefore may not reflect where it will be in a year. This highlights the importance of intent-anchored accountability, where the actions and scope of autonomy of an agent are linked to a clearly defined business intent, with

clear ownership for its use and outcomes. Change management and periodic reassessment can help identify such shifts and determine whether the agent has drifted beyond its originally approved purpose, authority or control assumptions.

Agents that direct other agents raise an architectural issue rather than representing another point on the autonomy spectrum. The treatment of an orchestrator and its sub-agents as a single or separate deployment depends on their approved business intent, authority, ownership and controls, as discussed in section 3.1. Accountability for the actions directed by an agent remains with the AI and the person or function responsible for its ultimate underlying business intent.

3.3 Agentic A.I. use cases in banking

Potential uses of agentic A.I. in banking can be illustrated through three broad archetypes. Task agents complete bounded work as triggered, such as assembling a case file, repairing a payment instruction or drafting a credit paper. Workflow agents run recurring processes end to end within defined rules, such as reconciling, refreshing records, filing or disposing. Monitoring agents operate continuously, detecting, escalating and, where authorised, containing issues. Multi-agent systems combine these archetypes under an orchestrator that directs specialised agents.

Exhibit 1 maps these archetypes across a range of banking activities. The examples in Exhibit 1 are illustrative and non-exhaustive; they do not describe the prevalence or maturity of current deployments, endorse particular uses, or imply that any activity is suitable for autonomous operation.

Exhibit 1: Illustrative applications of Agentic A.I. across banking activities

Domain	Task agents	Workflow agents	Monitoring agents
Customer-facing services	Resolves a card dispute end to end; a person approves the resolution	Collects, verifies and files onboarding documents	Watches service journeys, detects failed instructions and retries them within rules
Credit	Assembles a credit paper from internal and external sources; a person decides	Gathers financials and updates covenant status	Watches portfolios for early-warning signals and opens reviews
Payments and operations	Repairs malformed payment instructions; a person approves each release	Investigates and clears reconciliation breaks within limits	Watches liquidity positions and escalates breaches
Financial crime	Assembles alert case files for an analyst	Collects evidence, applies triage rules and routes screening alerts	Forms part of the transaction-monitoring layer
Markets and treasury	Drafts hedging proposals for a desk	Monitors funding and manages approved funding workflows	Watches limit utilisation across desks

Domain	Task agents	Workflow agents	Monitoring agents
Technology and cyber	Writes and tests code, raising changes through the pipeline	Tests and manages approved patch releases	Triages security alerts and, where authorised, contains affected endpoints within defined parameters
Risk and compliance	Compiles regulatory returns for review	Maps new requirements to control owners	Tests controls continuously and logs exceptions

3.4 Direction of travel

AIs are currently exploring a range of use cases including advisory tools, approval-gated workflows, and more bounded forms of task automation. Some AIs are experimenting agent-compatible transaction rails, such as agentic payment capabilities, with the aim of supporting a wider agentic ecosystem. The extent and pace of adoption will vary by institution, use case and risk profile. This Paper therefore focuses on the governance considerations that may arise as AIs delegate a greater range of tasks and decision rights to agentic systems.

Discussion question 3A
Is the description of agentic A.I. in this section an appropriate characterisation of the technology? If future supervisory work were to be considered, would the proposed operational definition in section 3.1, which considers a system's approved purpose, delegated authority and ability to affect outcomes, provide a clear and workable scope? Which types of systems in your institution would sit close to the proposed definition, and how would you suggest such cases be treated?
The HKMA's views The HKMA has defined the technology by its capabilities to keep the description current, and set an operational definition based on what a system is designed to do rather than how products are labelled. We recognise that difficult cases are likely to remain, particularly where assistants gain the ability to act or where agentic features are embedded in vendor products. Concrete examples of systems near the boundary would therefore be particularly useful than comments on the wording alone.

Discussion question 3B
How is your institution using, or planning to use, agentic A.I.? Which use cases do you expect to become significant first, and what is holding deployment back? For future supervisory considerations, would it be preferable and feasible to apply any future supervisory approach to: (i) all systems within the proposed definition; (ii) only systems meeting specific risk or materiality criteria; or (iii) a broader or narrower category?
The HKMA's views The HKMA recognises that the use of agentic A.I. across the banking sector is still developing and may not yet be fully visible through supervisory engagement and sandbox participation. Responses will help the HKMA better understand market maturity, emerging use cases and practical governance challenges, and will inform its consideration of whether further supervisory work is warranted.

4. Challenges and Risks

4.1 Framing

This section groups the risks into three bands, and potentially overlapping: **Band A** – those risks shared with other technology deployments; **Band B** – those intensified by autonomy, speed and scale; and **Band C** – those associated with systems that can access data or systems, use tools, trigger workflows, or initiate or influence actions within delegated authority. On top of that, there is an overarching set of consideration focusing on the banking system as a whole.

As deployment of agentic A.I. is still at an early stage and practical evidence remains limited, the assessment is provisional; section 4.5 therefore seeks evidence as well as views.

A central consideration is whether meaningful human oversight remains between an agent's actions and potential harm. Such oversight is not established merely by adding a formal approval step. It must be timely, informed, exercised by a person with appropriate authority, and capable of preventing, pausing or reversing harmful outcomes where necessary.

Meaningful oversight also depends on operating parameters that are commensurate with the deployment's authority, speed and potential impact. Where an agent can take or influence actions at machine speed, institutions may need to define in advance the scope of permitted actions, relevant transaction or exposure limits, escalation thresholds and the maximum velocity at which oversight and intervention arrangements can remain effective.

4.2 Band A – Risks shared with general A.I. deployments

Agentic deployments first carry the same risks as any significant technology. It is useful to distinguish these risks before considering what agentic capabilities change. References should be drawn to the HKMA's *Supervisory Guidance for Responsible Adoption of A.I.*, and examples of these risks, among others, are as follows:

Model risk

Agentic deployments rely on one or more models, including foundation models, specialised models, rules-based components or combinations of these. They inherit the limitations of the models and other components on which they depend.

Multi-agent arrangements may combine models of different sizes, capabilities, purposes and risk profiles, increasing the need to understand how their outputs interact. Outputs may be inaccurate, incomplete or fabricated despite appearing fluent and confident. Performance may vary across customer groups, raising fairness concerns, and may change as data, usage pattern and the underlying model diverge from what was originally tested.

Explainability

The complexity and opacity of foundation models make individual outputs difficult to interpret. A fluent response may not provide a reliable account of how it was generated or why a particular action was proposed.

This matters when outputs inform customer decisions and complicates internal review, audit and complaint handling. In practice, AIs may need to focus not only on explainability, but also on traceability, logging and the ability to reconstruct material decisions and actions.

Data protection and privacy

Agentic A.I. deployments may collect, generate, retrieve data at multiple stages, including personal data. Depending on their design, they may also retain information in conversation histories, caches, logs or memory beyond the original interaction.

The Personal Data (Privacy) Ordinance governs the collection, use, retention and security of personal data. The relevant guidance issued by the Privacy Commissioner for Personal Data, referred to in section 2.2, should also be considered in the design and operation of agentic A.I. systems.

Third-party and outsourcing risk

Few AIs develop every component of an A.I. deployment in-house. Many deployments depend on external providers for models, hosting, orchestration and parts of the tooling. Established arrangements for vendor due diligence, contractual safeguards, service levels, exit planning and substitutability therefore remain a relevant basis for managing these dependencies.

The same considerations extend to group arrangements, where AIs receive models, platforms or agentic capabilities from a group hub or affiliated entity.

Change management

Models, prompts, tools, integrations and orchestration arrangements may change frequently, including on the third-party's timetable, and any such change may alter deployment behaviour. As with other systems, changes require disciplined testing, approval and rollback arrangements.

These risks are significant but already within the scope of existing requirements for model risk, data protection, outsourcing and technology risk management. The following two bands therefore focus on what autonomous action changes.

4.3 Band B – Risks heightened by autonomy, speed and scale

The second band comprises familiar risks that may be amplified by an agent’s ability to perform multi-step tasks, operate at speed and scale, retain state, interact with external tools and depend on dynamic combinations of models, data and services.

Validation gaps

Traditional validation assumes a stable relationship between inputs and outputs: behaviour observed across a sample would represent production performance. Agentic deployments weaken both assumptions. Because they carry forward states, identical inputs may produce different results depending on prior observations and stored information. Their behaviour may also shift as underlying tools, models and memory change. A standard test passed at deployment therefore provides less assurance than it would for a conventional system; section 6.2 considers what testing might be useful for demonstrating accountability.

Reliance on probabilistic controls

Large language models are probabilistic systems. Their interpretation and application of instructions, policies and contextual information may vary across circumstances. Accordingly, controls that rely solely on the model to identify and comply with critical constraints may fail.

For defined high-impact conditions, institutions may need deterministic validation checks in addition to model-based controls. Such checks may verify, for example, that an action is within the agent's delegated authority, complies with applicable business rules, does not exceed a relevant transaction or exposure limit, has received any required approval, and is directed to an authorised recipient. A failure of a deterministic check should result in the action being prevented, escalated or routed for human review, rather than relying on the agent to resolve the exception.

This is consistent with the broader principle that validation should be proportionate to the potential impact of the use case and should not depend solely on a model's ability to interpret or follow instructions.

Compounding errors

An error, inaccurate assumption or inappropriate instruction early in a sequence can become the basis for every later step. Because each subsequent action may appear reasonable when viewed in isolation, the chain can be difficult to detect or challenge.

Individually plausible actions may therefore produce an outcome that would have been rejected if considered as a single end-to-end decision. The last opportunity for human intervention may occur several steps before the harm becomes visible.

Stacked dependencies

A conventional system may depend on a single vendor, whereas an agentic deployment can rely simultaneously on a model provider, hosting arrangement, orchestration layer, integration protocols and the other services called by its tools.

A failure, compromise or change anywhere in this stack may alter deployment behaviour. Dependencies may include technical protocols and open-source components as well as identifiable vendors, each with its own versions, maintainers and implementation risks.

Expanded attack surface

In addition to the infrastructure exposed by any system, an agentic deployment may be attacked through user inputs, tools calls and external content, retrieved documents, connected services and persistent memory.

These entry points may be influenced by parties with no direct relationship with the AI, including persons who create documents an agent reads, host pages it accesses or operate services it calls.

Speed and scale

A staff member may make a mistake once; an agentic deployment can repeat it at machine speed for as long as it operates. Numerous individually minor actions may cause serious harm before review catches up, affecting thousands of customers or transactions while remaining too dispersed to attract attention in any single case.

Technical throughput limits alone may not address this risk. An agent may operate within a system's technical capacity while still acting at a volume or speed that exceeds the ability of human reviewers, control functions or intervention mechanisms to identify, assess and contain harmful outcomes. AIs may therefore need to consider the need for velocity limits that are commensurate not only with technical constraints, but also with the demonstrated effectiveness of oversight arrangements. Such limits may apply to, for example, the number, value, frequency, customer reach or cumulative impact of actions within a defined period.

Stress testing should assess whether monitoring, exception queues, human review, escalation and stop mechanisms remain effective under peak volumes, rapid sequences of actions, degraded operating conditions and plausible failure scenarios. Where they do not, the deployment's authority, scope or velocity may need to be constrained.

Persistent state and data outflow

Information retained in an agent's memory may outlast the interaction that created it and influence later behaviour without appropriate review or expiry controls. Each call to an external tool may also transfer data beyond the AI's immediate control boundary. The data footprint of a deployment may therefore expand with each additional step or connection.

Resource exhaustion and uncontrolled activity

An agentic deployment may repeatedly call tools, retrieve data, generate outputs or initiate workflows at machine speed. Design weaknesses, unexpected loops or malicious inputs may therefore create excessive processing demand, third-party service consumption, financial cost or operational disruption before the activity is detected and contained.

4.4 Band C – Risks particularly associated with delegated authority and action-taking capability

A further set of risks becomes particularly acute where an A.I. agent can access systems or data, use tools, trigger workflows or initiate actions within delegated authority. These risks may be reduced where meaningful human oversight exists, but a formal approval step alone may not be sufficient, particularly where reviewers face high volume, time constraints, or lack effective authority to challenge the agent. Practical experience in managing such risks remains limited. Sections 5 to 9 explore possible governance responses to each with the aim of demonstrating accountability.

Objective misalignment and emergent behaviour

A system pursuing an objective may do so in unintended ways: it may satisfy the objective's wording while undermining its purpose, or find routes its designers neither anticipated nor would have approved. The system may also be inconsistent with the AI's risk appetite when risk-management considerations are required as part of a task or workflow.

Identity and attribution gaps

Identity, entitlement and credential controls have traditionally been designed around human users and clearly identified system accounts. When an agent uses shared, inherited or borrowed

accounts, the link between an action and the person authorising it may be weakened or lost and may be impossible to reconstruct later. This can impair authentication, authorisation, non-repudiation, audit and incident investigation.

Injected instructions and memory poisoning

Data, retrieved content, tool descriptions and responses may contain malicious or inappropriate instructions disguised as information from sources outside the AI's control. Once an agent's memory is poisoned, the effect may persist until detected and removed.

Authority creep

Routine product changes, including changes to an external interface, can inadvertently introduce security gaps such as wider scope of permissions or new toolsets. An agent's deployable capabilities may then exceed what was originally authorised, with the gap remaining hidden until investigation.

Approval fatigue

Human review may be unable to match machine throughput. Although an approval step may remain formally in place, it can become perfunctory, leaving the AI reliant on a control that is no longer effective. This creates a practical limit on the speed and volume at which an agent can operate safely. Where the rate of alerts, recommendations or actions exceeds reviewers' ability to assess them meaningfully, the AI should not treat the formal existence of an approval step as effective human oversight.

Accountability gaps

A deployment may involve builders, business owners, vendors, platforms and many users, diffusing responsibility until no individual clearly answers for its actions. Customers may also interact with software that appears conversational or human-like without understanding its nature, authority or limitations.

Irreversibility and silent failure

An agent may complete an irreversible action before anyone notices, particularly if no tested stop mechanism exists. With complex workflows, it could be more challenging to determine at which point such a stop mechanism should be introduced. Conversely, a system on which others depend may fail silently or continue operating incorrectly, corrupting downstream processes.

Contextual failures in transaction execution

An agent may take an action that is within its authority and complies with defined rules, but is inappropriate in the circumstances of a particular transaction, customer interaction or operational event. For actions with significant customer, financial or operational consequences, institutions may need transaction-specific validation before an action takes effect. This may be implemented through rules engines, policy controls, workflow approvals, transaction-validation services or other technical mechanisms.

In some use cases, distributed ledger technology (DLT) may provide one means of recording, validating or enforcing pre-defined transaction conditions. However, DLT does not in itself establish whether an action is appropriate, authorised or in a customer's interest; it should

therefore be regarded as one possible component of a broader control framework rather than a substitute for governance, oversight or intervention arrangements.

Shared-dependency contagion

Multiple deployments within an institution may share a model, platform, tool, data source or failure mode. A shared compromise, unsuitable change or common outage may therefore affect multiple business processes at once. Similar dependencies across institutions may also give rise to wider system-level concerns, as considered in section 4.5.

System opacity

As an agentic A.I. system becomes more complex, it may become harder for AIs to identify how an outcome was reached, which components were involved, what data or tools were used, and where authority or control failures occurred. Limited visibility may delay detection, intervention, incident investigation and remediation.

Non-human callers at the perimeter

External agents may act under customer mandates, while an AI's agents may access external services. Standards and conventions for verifying the identity, authority and scope of an inbound non-human agent are still developing. Institutions may therefore face uncertainty over whether an agent is properly authorised to act on behalf of a customer or another organisation. The risk can be exacerbated by the lack of an established shared responsibility framework.

Residue at retirement

A poorly retired deployment may leave active credentials, access permissions, stored memory, logs or accumulated state behind. These remnants may continue to pose security, privacy and operational risk after the deployment itself is forgotten.

4.5 Potential systemic implications

Potential implications may arise beyond any single institution. First, only a sector-wide view can reveal concentration where multiple AIs rely on the same model or platform. Such common dependencies may create common-mode vulnerabilities and raise the question of whether, and through what arrangements, sector-level monitoring may be useful.

Second, as deployment grows, correlated machine-speed responses by multiple AIs to the same signal may warrant thematic work. The potential for feedback loops, procyclical behaviour or simultaneous withdrawal from similar activities may be particularly relevant where A.I. influences decisions at scale.

Third, interactions between A.I. agents operated by multiple AIs, customers or service providers raise unresolved questions concerning identity, authority, accountability, data exchange and dispute resolution. Relevant conventions and technical standards are still developing and may require coordination among authorities, industry and technology providers.

This Paper identifies these issues for discussion but does not seek to establish a separate system-wide monitoring framework at this stage.

Discussion question 4A
Does this section capture the risks that are most relevant to agentic A.I. deployments in banking? Which risks would you add, remove or assess differently? We particularly invite practical evidence from deployment experience, including incidents, near-misses, control failures, examples of effective mitigants, and observations on how risk varies with an agent's authority, access, use case and level of human oversight.
The HKMA's views The HKMA considers that risks associated with delegated authority, action-taking capability and ineffective human oversight warrant particular attention. These risks may become material where an A.I. agent can access sensitive data or systems, trigger workflows, interact with external services or affect customers and critical banking operations at speed and scale. At the same time, risks in Band A and Band B can also be material, particularly where common dependencies, data confidentiality, model performance, resilience or third-party arrangements are not adequately managed. Given the limited practical evidence currently available, the HKMA will place significant weight on evidence submitted in response to this question, including deployment experience, incidents, near-misses and the effectiveness of controls in practice.

5. Overarching framework: Intent-anchored accountability

This section and the sections that follow set out the HKMA's preliminary views for discussion. They explore governance considerations relevant to agentic A.I. deployments generally, as well as whether enhanced visibility may be appropriate for a smaller subset of material deployments. The views are exploratory in nature and do not themselves create binding requirements.

The considerations are intended to support a common outcome: an A.I. agent should operate only in pursuit of an approved business intent, within a defined delegation of authority. Accountability for that delegation, and for the outcomes arising from it, should remain with identified individuals and functions within the AI. It should not be obscured or diluted by reliance on models, platforms, cloud services, data providers, tools or other third-party components.

5.1 Objectives and supervisory principles

The HKMA's objectives are consistent with those across its supervisory work: safeguarding the safety and soundness of AIs, protecting customers, and maintaining confidence in the banking system as agentic A.I. systems become more widespread.

The preliminary thinking in this Paper is aligned with the HKMA's "ART" Framework for Responsible Adoption of Artificial Intelligence.

- **Accountability** covers the board and senior management's responsibilities for establishing A.I. strategy and governance arrangements, as well as clear ownership and effective oversight of A.I. adoption.
- **Resilience** covers risk-based and proportionate A.I. adoption, effective risk-management frameworks, and adequate resources and expertise.
- **Trust** covers the ability to ensure the quality of A.I.-generated outputs and actions, appropriate explainability and transparency of A.I. use cases, and ethical and fair treatment of customers.

These three pillars provide the basis for the governance considerations discussed in this and subsequent sections. They should be applied proportionately, having regard to the deployment's business purpose, degree of autonomy, delegated authority, access to data and systems, potential impact, reversibility and recoverability of its actions.

The HKMA does not propose that AIs adopt a single technical design or operating model. Different AIs may reach the intended outcomes through different combinations of governance arrangements, technical controls, workflow design, human oversight and assurance. However, an AI should be able to demonstrate that a deployment remains governed within its approved purpose and authority throughout its lifecycle.

5.2 Intent-anchored accountability

Accountability for actions taken by an A.I. agent should remain with the AI and the person or business function responsible for the deployment's underlying business intent. This is referred to in this Paper as **intent-anchored accountability**.

Intent-anchored accountability connects an agent's actions, authorised scope of autonomy and delegated authority to a clearly defined business intent. It addresses a fundamental risk that arises when software can act: responsibility may otherwise become dispersed among developers, business users, vendors, platforms and control functions, until no person clearly answers for the outcome.

Before an agentic A.I. deployment enters production or is permitted to take material actions, the AI should identify and document the business intent which the agent is authorised to pursue. The agent's authorised purpose should be sufficiently specific to support risk assessment, control design, testing, oversight and subsequent review. In particular, the documentation of the approved intent should identify, as appropriate:

- the business objective and intended use of the deployment;
- the tasks, decisions or actions the agent is authorised to perform;
- the scope and limits of authority delegated to the agent;
- the data, tools, systems and user groups to which it may have access;
- any data that may leave the AI, and the intended recipients or destinations of such data;
- applicable transaction, exposure, operational or other relevant limits and tolerances;
- activities that are prohibited or require escalation or further approval;
- affected customers, counterparties, staff or other parties;
- the expected degree of human oversight and intervention capability; and
- circumstances in which the agent's authority should be reviewed, restricted, suspended or withdrawn.

The business intent should serve as the reference point for determining how much autonomy is appropriate. It should also provide the basis for assessing whether the agent's actions and outcomes remain consistent with the AI's risk appetite, policies and delegated authority throughout the deployment's lifecycle.

Accountability should be clearly allocated to an identified person or function, rather than displaced by a vendor, system or technical component. It is meaningful only where the accountable person or function has sufficient authority, information and practical capability to challenge, constrain, change, suspend or withdraw the deployment promptly where necessary.

In practice, the AI should be able to identify who answers for each agentic A.I. deployment, what that person or function is accountable for, and whether that person or function has sufficient authority to act.

5.3 Accountability and ownership arrangements

This proposed approach is intended to build on, and not replace, the requirement under the HKMA's Supervisory Guidance for each A.I. use case to have a dedicated owner with ultimate accountability. For the purposes of this Paper, each Agentic A.I. use case or deployment should have one designated accountable owner. Where an Agentic A.I. deployment forms part of an existing A.I. use case, the existing dedicated owner may remain accountable, provided that the deployment remains within the approved purpose, authority boundaries and risk assessment of that use case. A distinct accountable owner should be designated where the deployment has a separate business purpose, authority, or materiality profile.

Given the more complex and interconnected nature of Agentic A.I. deployments, the HKMA's preliminary view is that institutions should additionally identify clear technical and control role-holders. These roles support effective implementation, risk management and independent challenge; they do not transfer or dilute the ultimate accountability of the designated accountable owner.

Consistent with the intent-anchored approach, the person or function responsible for a deployment's underlying business intent should normally also be accountable for the deployment. Where the relevant business intent belongs to a business function, a suitably senior person responsible for that function should be identified as the accountable owner.

Accountability should not default to technology functions solely because agentic A.I. is involved. Technology functions remain responsible for technical matters within their remit, including the reliable and secure implementation, operation and maintenance of the deployment. However, the use of agentic A.I. does not remove the responsibility of the designated accountable owner and the relevant business function for the deployment's purpose, the authority delegated to it, or its business and customer outcomes.

For each deployment, the HKMA's preliminary view is that an AI should establish, at a minimum, one designated accountable owner and clear role-holders for the following areas:

- **Business ownership:** responsibility for the deployment's purpose, permitted use, business outcomes and continuing alignment with the approved business intent. The business owner should normally be the designated accountable owner with ultimate accountability for the deployment.
- **Technical ownership** (role-holder): responsibility for the build and operation of the deployment, including its models, workflow orchestration, tools, integrations, access permissions, logging, monitoring, stop mechanism and technical fallback arrangements.
- **Control ownership** (role-holder): responsibility for ensuring that risk assessment, control assessment, testing and assurance are performed appropriately, including independent challenge and assurance in line with the deployment's materiality and potential impact. This should include testing of the stop mechanism and fallback arrangements where relevant.

The accountable owner would normally be the business owner, holding ultimate accountability alongside that role. For less material or smaller deployments, the same individual or function may

also hold the technical ownership role where this is appropriate to the AI's size, structure and risk profile. However, the control owner should remain appropriately independent of the design and operation of the deployment, having regard to its materiality and potential impact. The designation of technical and control owners as role-holders does not create shared ultimate accountability for the deployment.

Ownership and responsibility should continue throughout the lifecycle of the deployment. The accountable owner should approve the deployment's scope and authority before production use, authorise material changes within applicable governance arrangements, and ensure that any transfer of ownership is formally approved and recorded. An AI should not allow an agentic A.I. deployment to continue operating where its accountable owner, authority boundaries or key control arrangements are unclear.

Multi-agent arrangements and shared components

The relationship between a distinct agent, a deployment and a business intent may be more complex in multi-agent arrangements. Some agents may perform supporting tasks within a wider deployment, while others may provide shared technology, data, monitoring or control capabilities across several use cases.

In such cases, the link to an approved business purpose may be established through the deployment or function supported, rather than through a separate business mandate for every technical component. Business intent may also concern internal functions, such as technology operations, risk management, compliance or assurance.

Possible approaches include:

- assigning ownership to each agent separately;
- assigning ownership to a multi-agent deployment that pursues the same approved business intent under common ownership and controls; or
- combining ownership of a shared component with clear accountability for its use within each deployment.

The appropriate granularity should depend on differences in purpose, delegated authority, ownership, access, potential impact and controls, rather than simply on the number of technical components or runtime instances. Regardless of the approach adopted, an AI should retain clear identification, attribution, accountability and control over higher-impact agents and shared components involved in a deployment, with sufficient traceability to understand their role in material actions and outcomes.

The HKMA invites views on whether these approaches provide a meaningful link to an approved business intent for every agent, and on the circumstances in which separate agent-level ownership may be necessary rather than deployment-level ownership.

5.4 Accountability where users instruct an agent

Accountability may become more complex where a staff member, customer or other user instructs an agent within a deployment that has been configured by its owner. The HKMA's preliminary view

is that AIs should establish a clear default allocation of responsibility within the institution that distinguishes between authorised internal users and external users such as customers or third-party agents.

The deployment owner is responsible for what the agent can do. This includes its action space, permissions, tools, guardrails, escalation arrangements, fallback mechanisms, and foreseeable misuse that the deployment's design is intended to prevent.

An authorised internal user is responsible for the instructions given to the agent and the information supplied to it, broadly in the same way as the user would be responsible when giving the same instruction or information to a colleague. This does not remove the AI's responsibility for designing appropriate permissions, guardrails, escalation arrangements and misuse controls.

This allocation depends on the AI being able to attribute material actions to the relevant human authority. The AI should therefore maintain appropriate records of instructions given to the agent, relevant user inputs, authorisations, actions taken and applicable approvals. The traceability considerations in section 7 are relevant in this respect.

Two boundaries are particularly important. First, this internal allocation of responsibilities does not alter the AI's responsibilities to customers, counterparties or the HKMA. The AI remains accountable for its own agentic A.I. deployment, including its design, authorised capabilities, controls and outcomes.

Second, the AI's accountability for its own deployment cannot be passed to customers. Customer input should be treated as untrusted input for security and control purposes. Where a customer's input can steer an agent into causing foreseeable harm, the deployment's design and boundary controls should be regarded as inadequate. Customer-facing deployments therefore carry design accountability by default.

5.5 Customer-facing deployments

Where an agentic A.I. deployment interacts directly with customers, the HKMA's preliminary view is that the following additional principles should apply:

- **Transparency:** Customers should be clearly informed when they are interacting with an A.I. agent. An agent should not impersonate a person or deny that it is an agent.
- **Access to human assistance:** Customers should be able to reach an appropriately authorised person promptly, without first having to overcome or repeatedly engage with the agent. They should be able to request human review of a material action, outcome or decision, as appropriate to the use case.
- **Existing redress arrangements:** Existing complaint-handling and dispute-resolution channels should apply on the same basis as they do to staff actions. The use of an agent should not reduce the AI's responsibility to investigate complaints, correct errors or provide appropriate redress.

The appropriate application of these principles should reflect the nature, materiality and potential customer impact of the deployment. For example, a customer-facing agent that provides general

information may require different arrangements from an agent that can influence eligibility, pricing, transaction execution or other material customer outcomes.

Discussion question 5A
Should the HKMA build on the existing supervisory framework, adding to it only where the characteristics of agentic A.I. call for it? The alternatives would be a dedicated framework for agentic A.I. or reliance solely on existing requirements without further agentic-A.I.-specific measures.
The HKMA's views At this stage, the HKMA's preliminary view is to build on the existing supervisory framework. Most requirements for governing agentic deployments are already in place, while a separate framework could duplicate them and create inconsistencies with the AI's broader governance arrangements. The HKMA has not decided whether additions are needed or what form they might take; this Paper explores possible arrangements for discussion.

Discussion question 5B
Do you agree that each Agentic A.I. use case or deployment should have one designated accountable owner, normally the business owner, supported by clearly defined technical and control ownership roles? Are the proposed arrangements practical and proportionate, including where an Agentic A.I. deployment forms part of an existing A.I. use case?
The HKMA's views The HKMA considers that each Agentic A.I. deployment should have a clearly designated accountable owner, normally the relevant business owner, given that the business function is best placed to understand the use case, intended outcomes and potential customer or operational impact. This accountability should be supported by clearly defined technical and control ownership roles; however, it should not dilute the responsibilities of senior management or relevant control functions. The arrangements should be proportionate to the materiality, level of autonomy, data and system access, and potential impact of the deployment. Where Agentic A.I. capabilities are added to an existing A.I. use case, institutions should assess whether the change materially alters the risk profile and, where appropriate, update the ownership and governance arrangements rather than duplicate existing controls unnecessarily.

Discussion question 5C

How should supporting or shared agents be linked to an approved business intent, and when is separate agent-level ownership necessary rather than deployment-level ownership? Does the proposed allocation of accountability between the owner of a deployment and its user reflect how responsibility should work in practice? Where would you draw the line between a design failure and misuse? For customer-facing deployments, what disclosure, human-escalation and review arrangements would be workable at scale?

The HKMA's views

The HKMA's preliminary view is that AIs should remain fully accountable to external parties for the actions and outcomes of their A.I. agents, and responsibility should not be shifted to customers. Each agent should be linked to an approved business purpose and deployment. Deployment-level ownership may be appropriate for supporting agents operating within defined boundaries, while separate agent-level ownership may be needed where an agent is shared or reusable, has distinct capabilities or permissions, or presents a material risk.

The allocation of responsibilities between the deployment owner and users requires further consideration. As a starting point, design, control or permission failures should remain the AI's responsibility, while misuse may arise where users act outside authorised purposes, access rights or instructions. The HKMA welcomes examples where the distinction is difficult in practice.

For customer-facing deployments, AIs should ensure customers are clearly informed when interacting with an A.I. agent, with proportionate arrangements for human escalation, review and intervention. The HKMA welcomes views on workable approaches at scale.

Discussion question 5D

Do respondents agree that the person or function responsible for a deployment's underlying business intent should normally be its accountable owner? Is the proposed allocation of responsibilities among business, technical and independent control ownership roles appropriate and proportionate, including for smaller AIs and less material deployments?

The HKMA's views

The HKMA considers that the person or function responsible for the deployment's underlying business intent should normally be designated as the accountable owner. This supports clear accountability for whether the A.I. deployment remains aligned with its intended purpose, the AI's risk appetite, and customer or business outcomes.

The accountable owner should be supported by clearly defined technical and independent control owners. The technical owner should be responsible for the sound design, operation, security and change management of the deployment, while independent control functions should provide effective challenge and oversight commensurate with the risks involved.

The allocation of responsibilities should be proportionate to the deployment's materiality, capability, degree of autonomy and potential impact. For less material or lower-risk deployments, institutions may adopt streamlined arrangements, provided that ownership, key controls and escalation responsibilities remain clear.

6. Lifecycle governance

6.1 Building on existing governance

Existing requirements apply to agentic A.I. deployments in the same way as to other systems operated by an AI. These include requirements and guidance relating to corporate governance and accountability, risk management, technology and cyber risk, business continuity, outsourcing and third-party risk, conduct and consumer protection, model risk, and data protection.

Agentic A.I. deployments therefore enter an already governed environment. This Paper builds on those requirements and invites discussion on whether the characteristics described in section 4, particularly for the Band A risks, warrant additional measures or adaptation of existing arrangements. In particular, an agent's ability to plan, retain state, use tools, access data or systems, and undertake connected tasks within delegated authority may require AIs to assess and govern the end-to-end deployment, rather than individual models or technical components in isolation.

The accountable owner should be responsible for ensuring that the deployment remains within its approved business intent and delegated authority throughout its lifecycle. This includes design and development, deployment and activation, operations, material changes and retirement. The accountable owner may rely on business, technology, risk and control functions to discharge particular responsibilities. However, this should not displace accountability for the deployment's purpose, authority and outcomes.

AIs should be able to demonstrate how their lifecycle arrangements give effect to accountability anchored in the deployment's approved business intent and delegated authority. This should include sufficient evidence that the deployment was appropriately assessed and approved, configured in accordance with its approved intent, operated within defined operational boundaries, appropriately monitored and tested, and changed, suspended or retired through controlled arrangements.

6.2 Lifecycle stages

The following provides an indicative lifecycle structure for discussion. It is not intended to prescribe a single delivery methodology, technical architecture or approval process. The lifecycle need not be linear: a material change, incident, emerging risk or change in a dependency may require a deployment to return to earlier assessment, approval or testing stages.

Design and development

Before an agentic A.I. deployment is activated, the AI should assess and document its intended benefits, materiality and risks. The assessment should consider the approved business intent; degree of autonomy; data, tools and systems accessed; third-party and shared dependencies; affected users and customers; and the potential impact and reversibility of actions.

The AI should define the deployment's permitted objectives, delegated authority, prohibited activities, escalation criteria, human-oversight arrangements, and relevant transaction, exposure,

operational or other limits and controls. The accountable owner, business owner, technical owner and independent control owner should be identified, with clear arrangements for approval, challenge, activation, material change, suspension and retirement.

The AI should assess limitations of the underlying models and other components. These may include inaccurate, incomplete or fabricated outputs; performance variations across customer groups or use cases; limitations in explainability; and changes in performance as data, usage patterns, configurations or underlying models diverge from what was originally tested. Where a deployment uses multiple models, agents, tools or external data sources, the AI should consider risks arising from their interaction as well as those associated with each component individually.

The design should incorporate appropriate safeguards for personal data and confidential information. This includes considering arrangements for data collection, use, retention, access, transmission, memory, logs and deletion. It should also identify the tools, systems, interfaces and credentials available to the agent, and apply access controls consistent with its approved authority and the principle of least privilege.

AIs should consider dependencies on models, cloud services, data sources, orchestration platforms, application interfaces, open-source components and other third-party services. Established arrangements for due diligence, contractual safeguards, service levels, business continuity, substitutability, exit planning and change notification remain relevant. These considerations also apply where capabilities are provided by a group entity.

Deployment and activation

Before activation, the AI should verify that the production environment reflects the approved deployment. This includes confirming the configuration of models, prompts, workflows, tools, permissions, integrations, data flows, monitoring, logging, stop mechanisms and fallback arrangements, including any material deviations from the configuration assessed and approved before deployment.

Traditional validation approaches often rely on relatively stable input-output relationships and the assumption that sampled behaviour is broadly representative. Agentic A.I. deployments may not satisfy either assumption. Identical inputs may produce different behaviour depending on context, prior interactions or retained state. Behaviour may also shift as underlying models, tools, data sources, prompts, memory or other components change.

The HKMA's preliminary view is that all agentic A.I. deployments should undergo testing proportionate to their nature and risk, regardless of materiality. Testing results should be retained by the AI as part of each deployment's record.

Testing should demonstrate more than whether an agent produces satisfactory outputs across a sample of scenarios. The evidence should focus particularly on operational boundaries and controls. This may include whether the deployment:

- remains within its authorised scope, delegated authority and applicable limits;
- responds appropriately to prohibited actions, exceptions and escalation conditions;
- prevents, blocks or routes actions for further review where deterministic checks or approval requirements apply;

- maintains appropriate protection of data, credentials, systems and tools;
- operates as intended when dependencies are unavailable, degraded, compromised or changed;
- remains subject to effective monitoring and intervention at expected and stressed action volumes; and
- can be paused, suspended or withdrawn safely, and whether its stop mechanism and fallback arrangements work when invoked.

Testing should be conducted under conditions representative of intended operation and, where appropriate, stressed and adversarial conditions. These may include peak volumes, rapid sequences of actions, unusual or conflicting instructions, malicious content or tool responses, degraded dependencies, control failures and attempted circumvention of authority boundaries.

The AI should ensure that the relevant accountable owner has approved activation in accordance with the deployment's governance arrangements. In particular, the owner should be satisfied that any residual risks are understood, fall within the AI's risk appetite, and are managed through appropriate controls, oversight and contingency arrangements.

Operations, material change and retirement

The AI should monitor the deployment's behaviour, observable intermediate steps, tool calls and outcomes against relevant thresholds and indicators. Monitoring should be capable of identifying drift from approved intent or authority, unusual activity, control failures, deteriorating performance, changes in dependency risk, and emerging customer or operational harm.

The accountable owner should review the deployment periodically and when a material event, incident or change occurs. The frequency and depth of review should reflect the deployment's materiality, scope of authority, degree of autonomy, rate of action, dependency profile and potential impact. For higher-risk deployments, review may need to include evidence from transaction or action logs, exceptions, complaints, overrides, performance indicators, control failures and testing results.

AIs should maintain adequate records to support accountability, independent assurance, audit, customer complaint handling, incident investigation and remediation. Records should enable the AI to reconstruct material actions, relevant control decisions and the basis on which authority was exercised, including relevant instructions, agent identity, user identity where applicable, tools and data used, permissions relied upon, approvals obtained, controls triggered and outcomes reached.

Material changes to models, prompts, data sources, tools, workflows, permissions, integrations, memory arrangements or other components should be subject to appropriate assessment, approval, testing and rollback arrangements. This includes changes introduced by third parties, including model providers, cloud providers, group entities, tool providers and data providers. The AI should assess whether a proposed change affects the approved business intent, delegated authority, risk assessment or materiality of the deployment, and whether re-approval, reclassification or revalidation is required.

The AI should be able to restrict, pause, suspend or withdraw a deployment where risks can no longer be adequately managed. The accountable owner should have sufficient authority to initiate

or authorise such intervention, supported by technical arrangements that can take effect at a speed commensurate with the deployment's authority and operating velocity.

At retirement, the AI should address residual access permissions, credentials, data, memory, logs, contracts and dependencies. It should ensure that the deployment does not leave unmanaged security, privacy, legal or operational risks after its business use has ceased. Where records must be retained for audit, legal, regulatory or investigation purposes, the AI should retain them securely and in accordance with applicable retention requirements.

What constitutes sufficient lifecycle evidence for a particular deployment remains unsettled. The HKMA particularly seeks practical experience and views on testing, assurance and record-keeping approaches that can provide meaningful evidence that an agentic A.I. deployment remains within its approved intent and authority throughout its lifecycle.

The following sections consider three governance elements that may assist AIs in demonstrating intent-anchored accountability across this lifecycle: a risk-integrated agent harness, meaningful human oversight and effective intervention, and traceability supported by appropriate records and, for material deployments, potential Registry arrangements.

Discussion question 6A
For agentic A.I. deployments whose behaviour varies by context and changes over time, how does your institution test and validate boundary controls, monitoring, timely intervention and fallback arrangements before deployment and during operation? What risk-based testing, revalidation and evidence-retention arrangements are appropriate, including reporting for material deployments?
The HKMA's views At this stage, the HKMA is seeking practical evidence rather than proposing settled expectations. The HKMA's preliminary view is that testing should focus on whether boundary controls work effectively, whether emerging issues can be detected and addressed in a timely manner, and whether fallback arrangements are workable, not merely whether an agent produces satisfactory outputs across selected scenarios. Practical experience from testing and revalidation will help shape any future expectations.

Discussion question 6B
Does the proposed lifecycle structure of design and development, deployment and activation, and operations, material change and retirement provide a clear and workable basis for applying the Accountability, Resilience and Trust principles to agentic A.I. deployments? Are there lifecycle events, dependencies or control considerations that should be addressed more explicitly?
The HKMA's views The HKMA's preliminary view is that the proposed lifecycle provides a practical basis for applying the Accountability, Resilience and Trust principles to agentic A.I. deployments, without prescribing a particular technical architecture or approval process. The lifecycle should not be treated as linear. Material changes, such as to models, tools, data sources, permissions, workflows or third-party dependencies, may require reassessment, retesting or re-approval. Incidents, control failures or emerging risks may also require a review of the deployment's purpose, authority, materiality or continued operation.

Discussion question 6B
Lifecycle arrangements should be proportionate to the deployment's risk profile. The HKMA seeks views on whether the proposed structure adequately addresses multi-agent and third-party dependencies, external-provider changes, retained memory or state, and safe suspension, rollback and retirement.

7. Specific governance elements

This section considers three governance elements that may help AIs demonstrate intent-anchored accountability in practice: a risk-integrated agent harness; meaningful human oversight and effective intervention; and traceability supported by internal deployment records. These elements are complementary. Together, they are intended to help AIs demonstrate that an agentic A.I. deployment operates only within its approved purpose and delegated authority, that it can be effectively controlled where necessary, and that its actions can be understood and attributed after the event.

7.1 Risk-integrated agent harness

A deployment's operating boundaries, including the tools, permissions and data access available to it, should be defined in advance, technically enforced and changed only through the AI's change-management process. A risk-integrated agent harness may form part of the deployment's boundary controls, with the particular aim of addressing Band B and C risks described in section 4. It may translate the AI's approved business intent, risk appetite, risk-management framework and relevant policies into practical, testable and enforceable operating constraints.

This is important because high-level instructions or prompts alone may not reliably constrain an agentic A.I. system. An agent may interpret instructions differently across contexts, interact with changing tools or data, or encounter external content intended to alter its behaviour. Critical controls should therefore not depend solely on a model's ability to interpret and follow instructions.

The agent harness may include, as appropriate:

- system instructions and workflow design;
- pre-defined tools, application programming interfaces and external interfaces;
- identity, access-management and credential controls;
- data-access and data-loss-prevention controls;
- permissions based on least privilege;
- deterministic validation checks;
- transaction, exposure, operational or other relevant limits;
- approval gates and escalation paths;
- velocity limits and resource-consumption limits;
- monitoring, alerting and anomaly-detection arrangements;
- dynamic guardrails that adjust permissions or require further review in response to changing conditions; and
- controls that prevent unapproved changes to the agent's tools, permissions, data sources, integrations or other capabilities.

The aim is to ensure that the deployment's actual capabilities continue to match what has been authorised. A deployment should not be able to extend its own authority, acquire new capabilities, create or alter credentials, connect to new external interfaces, or modify its own boundary controls at runtime, unless the AI has explicitly approved and appropriately controlled an arrangement for this purpose.

In practice, an AI could maintain a recorded and enforced set of tools, permissions and data-access rights for each deployment. These should be version-controlled, so that changes are visible, attributable and subject to appropriate approval. External services and interfaces could be identified in advance, reducing the risk that new capabilities appear unnoticed. Any expansion of the deployment's range of permitted actions, access or reach should be subject to change control and reassessment.

The harness should incorporate risk-management considerations directly into the deployment's operation. For example, where an AI's risk appetite, policies or procedures require segregation of duties, customer authentication, sanctions screening, transaction limits, suitability checks, escalation of suspicious activity or restrictions on data transfers, the AI should consider how these requirements are translated into the agent's workflows and enforced through technical and operational controls.

For defined high-impact conditions, this may require deterministic checks in addition to model-based controls. Such checks may verify that an action is within the agent's delegated authority, complies with relevant business rules, does not exceed a relevant limit, has received any required approval, and is directed to an authorised recipient. Where a deterministic check fails, the action should be prevented, escalated or routed for human review. It should not be left to the agent to resolve the exception without appropriate controls.

Where applicable law, regulation, or internal policy reserves a decision for a person, an agentic A.I. system may inform but should not make that decision. Where it is unclear whether a decision is reserved for a person, the same approach should apply: the system may support analysis or provide recommendation, but a person should make the decision.

A risk-integrated agent harness may address a number of the risks identified in section 4, including authority creep, injected instructions, uncontrolled activity, inappropriate data access and actions beyond delegated authority. The practical test is whether the AI can demonstrate that the deployment's current capabilities, operating limits and access rights match what was authorised, and that they cannot expand without deliberate approval.

7.2 Human oversight and effective intervention

AIs should retain meaningful human oversight and the practical ability to intervene or stop a deployment promptly when its actions depart from the approved business intent, delegated authority or applicable controls, in response to some of the Band B and Band C risks highlighted in section 4. The appropriate arrangements will depend on the use case, the AI's risk appetite, the degree of autonomy involved, the speed and scale of activity, and the potential consequences of an error or failure.

Depending on these considerations, oversight may take the following forms:

- **human-in-the-loop** arrangements, under which a person approves individual actions before they take effect;
- **human-on-the-loop** arrangements, under which the agent acts within defined authority while a person monitors its operation and can intervene; and
- **human-in-command** arrangements, under which identified persons retain authority to define, change, restrict, suspend or withdraw the deployment's mandate and operating boundaries.

A move from individual human approval towards monitoring-based oversight should not reduce the AI's accountability. Such a transition should depend on clearly defined and technically controlled delegated authority, supported by active monitoring of the deployment's decisions and actions. Monitoring and intervention arrangements should be capable of detecting departures from approved intent in sufficient time for effective action, having regard to the speed, scale and potential consequences of the deployment's activities.

A formal approval step is not, by itself, meaningful human oversight. Approval is meaningful only where the approver sees the substance of the action, has the authority and practical ability to challenge or reject it, regards rejection as a legitimate outcome, and has sufficient time to make an informed decision before the action takes effect. Where the volume, complexity or velocity of actions exceeds the reviewer's ability to assess them meaningfully, the AI should not regard the formal existence of an approval step as an effective control.

In practice, effective intervention may be supported by clear intervention points, a stop mechanism, regular rehearsals and a workable fallback plan. Approval and intervention points should be positioned before actions with significant customer, financial, legal or operational implications take effect. Their design should take account of detection time, decision time and stopping time, so that intervention can occur before harm arises or becomes irreversible.

Intervention measures may include:

- correcting outputs or directing the deployment to seek further review;
- restricting the deployment's access to specified data, tools, systems or user groups;
- reducing transaction, exposure, operational or velocity limits;
- imposing additional approvals or review requirements;
- pausing or suspending the deployment through a stop mechanism, also referred to as a "kill switch";
- withdrawing the deployment from production use;
- rolling back a material change; and
- activating alternative processes, including manual processing or other fallback arrangements.

A stop mechanism may include the ability to suspend the deployment's access to tools or systems, or to stop the deployment through an AI-controlled mechanism integrated into automated workflows. The mechanism should be operable regardless of the status of a vendor, platform or other third-party component where this is necessary to contain harm.

AIs should test and rehearse their intervention arrangements at intervals proportionate to the deployment's risk and after material changes. Testing should assess not only whether the stop mechanism can be activated, but also whether it takes effect within a sufficient timeframe, whether residual or in-flight actions are appropriately managed, and whether the fallback can maintain continuity of critical operations. Suspension, restoration and associated decisions should be recorded to support subsequent review, remediation and assurance.

The practical test is whether the AI can demonstrate that it can deliberately stop or constrain the deployment and maintain business continuity. If an AI cannot stop or materially constrain a deployment without unacceptable disruption, the principle is not met in substance, regardless of how effective the mechanism appears on paper.

AIs may also explore, where feasible and appropriate to the deployment's risks, the use of "A.I. versus A.I." arrangements. Under such arrangements, A.I. systems may assist in monitoring other A.I. systems, identifying unusual patterns, testing outputs, prioritising exceptions or supporting human reviewers. These arrangements may enhance the scale and effectiveness of oversight, but should support rather than replace human accountability, challenge and intervention.

7.3 Traceability and internal deployment records

Effective governance relies on adequate visibility over every agentic A.I. deployment within the operational boundary described in section 3.1. This should be supported by a clear inventory, component lineage across interdependent models, data, tools, infrastructure and third-party services, and sufficient visibility over the deployment's actions.

Appropriate logging and attribution can connect an agent's action both to the deployment and to the person or function whose authority supports it. This includes the accountable owner for the deployment's authorised functions and the user responsible for an instruction in a particular instance. When software acts, the log replaces the memory and signature of a human actor. Attribution should therefore be designed into a deployment from the outset, as it may not be possible to reconstruct it later.

The HKMA's preliminary view is that every deployment within the definition set out in section 3.1 should be recorded under the AI's governance arrangements before production use. For each deployment, the record should identify:

- its purpose, intended use and deployment status;
- its accountable, business, technical and control owners, as well as authorised users;
- its capabilities and access, including data leaving the AI and the recipients of that data;
- its infrastructure and dependencies, including those shared with other deployments;
- its stop mechanism, fallback arrangements and log locations;

- its change and suspension history, and arrangements for memory, state and credentials at retirement; and
- its materiality assessment, underlying rationale and next review date.

The record should follow the deployment throughout its lifecycle. An entry may be created at conception, developed through design and testing, updated when the deployment enters production, and maintained through material changes, suspension and retirement. Suspensions and restorations should be recorded. At retirement, the record should be formally closed, with appropriate arrangements for memory, accumulated state, credentials, retained records and data. This final step is important, as casually decommissioned deployments may leave active credentials, access permissions or stored data behind.

The record should serve more than a record-keeping purpose. It may be linked to the AI's A.I. use-case inventory, approval processes, risk assessments, access-management arrangements, monitoring, contingency planning and incident-management processes. Keeping records up to date can help maintain visibility as a deployment's capabilities evolve, and identify changes that warrant reassessment or further review, including in relation to data protection, confidentiality and retention.

A related consideration is the management of "shadow A.I.", including unapproved tools and the use of approved tools outside established boundaries. The record may help an AI distinguish approved deployments from unauthorised use, support the assessment of associated risks, and facilitate remediation under its existing governance arrangements.

Component lineage and dependency mapping are particularly important in identifying the root cause and potential impact of incidents. AIs should be able to identify the models, data sources, tools, systems, interfaces, providers and credentials involved in a deployment. Each deployment should use distinct identities and credentials where appropriate, as shared or borrowed accounts can weaken or break the link between an action and the authority behind it.

Action logs should be sufficient to show, as appropriate, what the deployment did, what was affected, when the action occurred, the instruction or event that prompted it, the data and tools used, the permissions and controls relied upon, and any human approval or intervention. Logs should be retained, accessible and sufficiently reliable to support internal audit, investigation and reconstruction of material events. They should also be made available to the HKMA on request.

The internal record should also support the AI's assessment of whether a deployment is material. It should enable the AI to identify the deployment's capabilities, customer and operational reach, degree of autonomy, human-oversight arrangements, potential impact and shared dependencies. These matters are relevant to the proposed enhanced visibility and Registry arrangements for material deployments considered in section 8.

The practical test is whether the AI can explain any material action a deployment has taken, identify the authority behind it, determine the components and dependencies involved, and produce adequate evidence for independent assurance, incident investigation and remediation.

Discussion question 7A

Do respondents agree that AIs should translate approved business intent, risk appetite and relevant policies into technically enforceable controls through a risk-integrated agent harness? What controls are most important for different categories of agentic A.I. deployment? What practical evidence can demonstrate that a deployment's tools, permissions, access rights and range of permitted actions remain within approved authority and cannot expand without appropriate approval?

The HKMA's views

The HKMA considers that, in principle, approved business intent, risk appetite and relevant policies should be translated into technically enforceable controls for agentic A.I. deployments. A risk-integrated agent harness may be one way to achieve this, but the focus should be on control outcomes rather than prescribing a particular technical architecture.

Controls should be proportionate to the agent's materiality, autonomy, data and system access, and potential customer or operational impact. Key controls may include clear identity and ownership, least-privilege access, tool and action allow-lists, approval gates for higher-risk actions, segregation of duties, monitoring and audit logs, change controls, and effective pause, rollback and fallback arrangements.

AIs should be able to demonstrate, through documented authority matrices, system configurations, test results, access reviews, approval records and tamper-evident logs, that an agent's tools, permissions and permitted actions remain within approved boundaries and cannot be expanded without appropriate authorisation.

Discussion question 7B

Do respondents agree with the proposed principles for meaningful human oversight and effective intervention? What arrangements would provide meaningful assurance that human approval remains genuine where it is relied upon, and that a stop mechanism and fallback can operate effectively at the deployment's expected speed and scale?

The HKMA's views

The HKMA's preliminary view is that human oversight should be meaningful, risk-based and proportionate to the deployment's autonomy, speed, scale and potential impact. Where human approval is relied upon, the approver should have sufficient information, authority, competence and time to challenge or reject the proposed action. AIs should also maintain tested intervention, stop and fallback arrangements capable of operating within the timeframe required to prevent or contain material harm.

Discussion question 7C
What should a record of agentic deployments contain, and what would it cost to keep current? Which of the contents sketched above would be more challenging to maintain in practice, and how might they be made more operationally feasible? Is anything missing that could help the AI govern its deployments?
The HKMA's views The HKMA's preliminary view is that AIs should maintain an up-to-date internal record for each agentic A.I. deployment throughout its lifecycle. The record should be integrated, where practicable, with existing use-case inventories, approval, risk-management, access-control, monitoring and incident-management processes to reduce duplication. The main cost of the record lies in keeping it current. The HKMA welcomes views on the record's content and cost.

8. Registry of material A.I. agent

Section 7 discusses the importance of traceability and internal deployment records. This section explores enhanced visibility for a smaller subset of material deployments and the form that visibility might take. The proposed two-step assessment would distinguish material from non-material deployments. In general, the proposed Registry is expected to be based on the inventory of all A.I. use cases approved for deployment by individual AI as required by the Supervisory Guidance.

8.1 Classification

A key consideration in determining whether greater visibility of a deployment is needed is its potential for harm. The HKMA is minded to regard a deployment as material where its misbehaviour, its compromise or its failure could significantly harm customers' interests, the AI's safety and soundness, or the continuity of its critical operations. The test that serves this principle has two steps: the first asks what the deployment can affect, and the second asks who stands between the deployment and the harm.

Step 1: Reach

A deployment may have reach in any of the three areas, assessed using records the AI already maintains. Reach in any one area is sufficient to proceed with the assessment:

- **Positions**: The deployment can change a person's or the AI's position by moving funds, altering customer accounts, products or data, or creating commitments. A decision that determines a customer's entitlement, such as a credit outcome or access to funds, also changes their position.
- **Obligations**: The deployment can discharge or dispose of a duty owed to an authority or market. For example, closing a regulatory alert may not change an account but does extinguish a legal duty.
- **Operations**: A critical operation depends on the deployment's availability and accuracy. Scope is determined by the AI's critical operations inventory under its operational-resilience framework and includes monitoring functions that could fail without detection, as well as deployments that continue operating while producing incorrect results.

A deployment with no reach in these areas, for example, one that only drafts, researches or summarises, does not support critical operation and does not involve material external sharing of personal, customer or confidential information, is not material, and the assessment ends at this step. A deployment that directs other deployments inherits their reach.

Reach would also need to be significant. Significance is assessed against thresholds the AI already maintains, including its assessment of critical operations, outsourcing materiality definition, consumer-protection scope and incident-reporting thresholds; this Paper introduces no new numerical thresholds. The assessment considers the deployment's permitted course of action over the time needed to detect and stop it. A stream of individually small, reversible actions may therefore be significant if its pace exceeds the AI's ability to identify and halt it.

Step 2: Oversight

Step two asks whether genuine human approval of each action stands between the deployment and the harm identified in step one. If not, the deployment is material: significant reach without a person in the path is the core case. If so, it is not material unless the staged action is irreversible. The exception reflects the gate's role as the final control: a failed gate on a reversible action still allows correction, but a failed gate on an irreversible action does not. Payment finality is the standard example, so a gated payment-release deployment remains material. Human approval of an A.I. system's actions does not, by itself, prevent the system from failing. Separate resilience controls are needed to manage outages, faulty data, model errors and third-party dependency failures.

The Paper also seeks views on whether effective human approval should operate as a second step in determining materiality, as proposed, or instead be treated solely as a risk-management control once materiality has been established based on the deployment's reach and potential impact.

Approval is genuine only where all the following conditions are met: the approver sees the substance of the action and has the authority and practical ability to reject it; rejection is accepted as a legitimate outcome; and an action remains pending for a decision. If these conditions are not met, the deployment is assessed as ungated. Ongoing evidence that the gate remains genuine is required for the deployment to retain non-material status.

The assessment normally applies to a single deployment. However, deployments sharing a model, platform, orchestrator or critical dependency are also assessed together because one failure or compromise could affect them all simultaneously. A group of individually modest deployments may therefore be collectively material, which is why the record in section 7.3 tracks shared dependencies.

Under the proposed approach, the AI assesses each deployment before production and after every material change. The independent control owner endorses the assessment, and the record documents each step and review date: the types of reach, supporting evidence, the oversight conclusion and, where a gate is claimed, the gate conditions. Inconsistency is itself a warning sign; for example, classifying a deployment within a critical operation as immaterial would warrant review.

The assessment remains current and is reassessed whenever relevant facts change, including following a modification or incident, when gate metrics call genuine approval into question, following a validation finding, or growth in a shared-dependency group. Reassessment works both ways: a deployment may cease to be material if supported by the same quality of evidence that established its status.

Exhibit 2: Use case classification examples

- An internal research and drafting deployment has none of the three forms of reach: it changes no positions, discharges no obligations and supports no critical operation. The assessment ends at step one, so the deployment is not material.

- A deployment closes false-positive financial-crime alerts without approval for each action. Because closing an alert discharges a regulatory duty, the deployment has obligations reach, with no person intervening before a wrongful closure. It is material.
- A wholesale payment-release deployment requires human approval for every payment. Although a genuine gate stands between the deployment and its positions reach, settled payments may not be reversible. It is material.
- A read-only deployment is the sole automated control used to monitor for fraud. It has no write access, but a critical operation depends on its accuracy, and approving its actions would not prevent its failure. It is material.

8.2 Notification to the HKMA

The proposed notification arrangements would be designed to complement, rather than duplicate, existing notification or supervisory-engagement requirements that may also be triggered by an A.I. deployment, including those relating to material outsourcing or third-party service engagements, and material changes to e-banking initiatives, where applicable. The HKMA will take account of industry feedback in considering how these arrangements should be aligned and streamlined.

For a material deployment, the Paper seeks views on whether the AI should notify the HKMA before production or provide a timely update after deployment. Views are invited on which arrangement would provide timely supervisory visibility at a proportionate reporting cost, and what advance notice period or post-deployment notification timeframe would be workable.

For pre-deployment notification, the Paper also seeks views on whether AIs should be able to proceed without awaiting a supervisory response, receive supervisory feedback without a requirement to await it, or await a supervisory response before proceeding. Where supervisory feedback or a response is envisaged, its scope and timing would need to be clear.

The notification could draw on the AI's internal deployment record described in section 7.3 and supporting evidence organised around the four principles, as summarised in Exhibit 3. The aim would be to enable AIs to reuse information and evidence maintained for their own governance, rather than prepare separate material solely for notification, thereby limiting duplication and keeping reporting costs proportionate.

Exhibit 3: What a notification to the HKMA may cover

- **Accountability**. The deployment's purpose and status, the accountable owner and the business, technical and control owners; who may instruct the deployment; how accountability is allocated along the chain of use.
- **Visibility**. How every action is tied to the deployment and to the human authority behind it, and how the audit trail is retained.
- **Human oversight and intervention**. The stop mechanism, evidence that it has been tested, and the fallback that keeps the business running while the deployment is stopped.
- **Operating boundaries and safeguards**. What actions the deployment can take, access and affect, its permitted data and system access, any data leaving the institution and its

recipients, how those boundaries are technically enforced, what prevents unauthorised expansion of capability, and the external interfaces it uses.

- **Technical overview**.
  I. **High-level architecture and dependency map**: The deployment's principal components, interfaces, data flows and dependencies, including the external providers, systems, models, tools and infrastructure on which the deployment materially depends.
  II. **Model registry and orchestration**: The specific model(s) used. Where multiple models are used, how they interact and the functions they perform.
  III. **Tools catalogue**: Inventory of material external applications, Model Context Protocol (MCP) servers, API endpoints, or data store that the agent is permitted to call.
  IV. **Data storage, privacy and confidentiality arrangements**: How data, including personal and confidential information, is processed, retained and protected.
  V. **Identity and access privileges**: The identities and credentials through which the deployment accesses systems, tools and data, the permissions attached to them, and how access is restricted to the deployment's authorised scope.
  VI. **Execution controls**: The principal technical controls governing autonomous execution, including approval checkpoints, execution or usage limits, and other controls that prevent the deployment from acting beyond its authorised boundaries.
  VII. **Logging and traceability**: The records generated by the deployment, including material instructions, tool calls, approvals and actions, and the arrangements for their retention and integrity.
- For **customer-facing deployments**, additionally: the disclosure and human-escalation arrangements described in section 5.5.

Under either notification arrangement, the same timing approach would apply to material changes. A change is material if it expands the deployment's capabilities or reach; removes or weakens per-action approval; materially alters the underlying model or a critical dependency; extends the deployment to new customers, products or user groups; changes its owners; changes the assessment outcome, including when a deployment first becomes material; or returns the deployment to service after suspension or an incident. Under a material-change notification model, routine updates that do not alter the action space need only be recorded in the change log and made available on request, with any periodic reporting determined under the arrangements discussed in section 8.3.

8.3 Ongoing monitoring

Notification would provide the HKMA a snapshot of a deployment, while ongoing supervision requires visibility of subsequent developments. Possible arrangements include updates after every change, updates only on material changes, or consolidated updates at a set frequency, such as monthly, quarterly or annually. A combination could pair material-change notifications under section 8.2 with periodic reporting of other developments. An up-to-date inventory of material agentic deployments provides a common basis for continuous monitoring. This inventory could

help track changes in deployments' status, capabilities, risk profiles, and shared dependencies over time.

In addition to each AI's internal deployment records and notification to the HKMA, a sector-wide aggregation arrangement could be considered to support monitoring of common dependencies and emerging vulnerabilities. Possible approaches could include a centralised arrangement or a decentralised, secure registry maintained by the industry.

Incidents would continue to be reported under existing requirements and thresholds. The Paper seeks views on whether the following agentic events should ordinarily be treated as incidents for the purposes of existing incident-reporting requirements and thresholds:

- the deployment acts beyond its authority;
- the stop mechanism fails when invoked or the fallback fails to maintain business continuity;
- the deployment's identity or credentials are compromised;
- a material monitoring deployment is found to have failed silently; or
- an approval gate allows an action to proceed improperly.

A complementary option is to establish a risk intelligence sharing mechanism among AIs and the HKMA, potentially facilitated by the HKMA or through an industry arrangement. This could support timely sharing of emerging vulnerabilities, common failure modes, incidents and near-misses, and lessons learned from mitigation, helping AIs assess their exposure and strengthen controls. Views are invited on whether existing channels could serve this purpose and what safeguards would be needed for confidentiality, data protection and secure handling. Such sharing would complement, rather than replace, existing incident-reporting requirements.

Between notifications, ongoing assurance could help establish whether a deployment continues to operate as assessed. Possible ongoing-assurance arrangements include retesting the stop mechanism and fallback at least annually and after every material change; independent validation on a cycle proportionate to risk; inclusion within internal audit's scope; and annual attestation by the accountable owner that the deployment remains within its authorised boundaries.

8.4 Purpose of notification

Under the proposed approach, notification to the HKMA serves three purposes: it provides visibility of material deployments and changes to an AI's risk profiles; supports supervisory dialogue on governance arrangements, controls and emerging issues; and, together with periodic reporting, helps build a sector-wide view of deployment patterns, shared dependencies and potential concentrations that no single institution can assemble.

The AI remains responsible for deciding whether and how to deploy, maintaining effective controls and addressing risks throughout the deployment's lifecycle, irrespective of notification timing or supervisory response arrangements. Notification provides information for the HKMA to review, ask questions and, where warranted, engage with the named owners through usual supervisory processes and applicable powers.

The timing of notification and any arrangements for supervisory responses are separate design questions, as discussed in section 8.2, with ongoing reporting arrangements addressed in section 8.3. Generally, a response to a notification or the lack of one does not constitute an approval. Any

arrangement under which an AI would await a supervisory response would need to distinguish clearly between feedback, acknowledgement and any requirement for prior approval.

Discussion question 8A
Are positions, obligations and operations the right kinds of reach for identifying material deployments, and is the approach to assessing significant reach workable? Should authorised external sharing of personal, customer or other confidential information be an additional materiality consideration? Is genuine per-action approval an appropriate second step, recognising that irreversible actions and critical-operation dependencies may remain material despite approval? Alternatively, should human oversight be treated solely as a risk-management control, with materiality determined by the deployment's reach and potential impact? What would make "significant", "genuine" and "irreversible" predictable in practice? How could deployments sharing dependencies be assessed for collective materiality?
The HKMA's views Our preliminary stance is for a test based on significant reach and effective oversight rather than autonomy or technology type alone. This recognises that a deployment staging irreversible actions may remain material despite genuine approval, as may a read-only deployment on which a critical operation depends. The Paper also seeks views on whether effective human oversight should instead be treated as a control measure, rather than a determinant of materiality. The assessment could draw on existing institutional records, while allowing for differences in deployment context, thresholds and controls. Genuine approval could support non-material status where neither qualification applies, but its effectiveness may weaken under high volumes, making ongoing evidence relevant. We welcome views on the clarity of the test and the practical assessment of shared dependencies.

Discussion question 8B
For material deployments and material changes, would pre-deployment notification or timely post-deployment notification be more proportionate and workable? If pre-deployment notification is preferred, should AIs proceed without awaiting a response, receive feedback without a requirement to await it, or await a response before proceeding? What timeframes would be practicable? How should any new notification arrangements interact with existing notification or supervisory-engagement requirements, including those relating to material outsourcing or third-party service engagements and material changes to e-banking initiatives?
The HKMA's views The HKMA's preliminary view is that notification arrangements should be risk-based and proportionate to the materiality of the deployment or change. Timely post-deployment notification may be appropriate in most cases, while pre-deployment notification may be warranted for specified higher-impact circumstances. The HKMA does not intend to change the notification arrangement to become a general pre-approval process. It seeks views on the circumstances in which prior engagement would be appropriate, and on practicable notification timeframes.

Discussion question 8B

The HKMA will take account of feedback on how notification arrangements can be aligned with existing supervisory processes and information requirements, while avoiding unnecessary duplication.

Discussion question 8C

What notification content would be proportionate, taking account of information already maintained in the internal deployment records? What additional information, if any, should be maintained or submitted for material Agentic A.I. deployments beyond the baseline A.I. use-case inventory? Should ongoing reporting cover every change, material changes only, periodic consolidated updates, or a combination of these approaches? How could reporting align with existing supervisory processes?

The HKMA's views

The HKMA's preliminary view is that reporting should build on information maintained in an AI's internal deployment records, rather than create unnecessary duplication. Reporting should focus on material deployments and material changes, supplemented by periodic consolidated updates where appropriate.

The HKMA seeks views on the minimum information needed for effective supervisory visibility and how reporting can be aligned with existing supervisory engagement.

Discussion question 8D

Would a sector-wide aggregation or registry of material A.I. agentic deployments assist in identifying common dependencies, concentration risks and emerging vulnerabilities? Should this be centrally maintained or operated through a decentralised, secure industry arrangement? What governance, confidentiality, access-control and update arrangements would be necessary?

The HKMA's views

The HKMA considers that sector-wide visibility over material A.I. agentic deployments could help identify common dependencies, concentration risks and emerging vulnerabilities. Any such arrangement would need to protect confidential supervisory and commercially sensitive information.

The HKMA seeks views on whether a centrally maintained supervisory arrangement or a decentralised secure industry arrangement would be more effective and operationally feasible. The HKMA does not envisage a publicly accessible registry of individual AIs' A.I. deployments considering technical sensitivity.

Discussion question 8E

Would an A.I. risk-intelligence sharing mechanism among AIs and the HKMA be useful for sharing vulnerabilities, incidents, near-misses and effective mitigants? What information, participation model and confidentiality safeguards would be needed? What ongoing-assurance measures would be proportionate for material deployments?

The HKMA's views

The HKMA sees potential value in a trusted mechanism for AIs and the HKMA to share relevant vulnerabilities, incidents, near-misses and effective mitigants. Such a mechanism should complement, and not replace, existing incident-reporting and risk-management obligations.

Ongoing assurance for material deployments should be proportionate to the risks involved. The HKMA seeks views on appropriate arrangements for continuing monitoring, periodic reassessment and review following material changes or risk events.

9. Interactions beyond the institution

9.1 External interactions and approved boundaries

A deployment's risk depends not only on its internal design and controls, but also on what it connects to and whether those interactions remain aligned with its approved business intent. This includes external services, data feeds, protocol-based tools, application programming interfaces, other A.I. agents and agents acting on behalf of customers or counterparties.

The HKMA's preliminary view is that a deployment's external connections and permitted interactions should be identified, risk-assessed and approved in advance. The AI should define the services, data sources, tools, interfaces and other agents with which the deployment may interact, the information that may be exchanged, the actions that may be requested or taken, and the conditions or limits applying to each interaction.

A deployment should not acquire or exercise new capabilities beyond those approved by the AI at runtime. Where a service dynamically advertises available tools or functions, the deployment should use only those tools or functions that the AI has explicitly approved, including through an approved allowlist or other equivalent boundary control. This is consistent with the risk-integrated agent harness described in section 7.1.

Two consequences follow. First, all external content, whether data, a tool response, retrieved content or a message from another agent, should be treated as potentially untrusted input and subject to appropriate validation before it is relied upon or permitted to influence the deployment's behaviour. This is analogous to the treatment of customer input described in section 5.4. If external content can steer a deployment into causing foreseeable harm, this should be regarded as an indication that the deployment's design, boundary or monitoring controls may be inadequate.

Second, a change on the other side of an interface may affect the deployment even where the AI has made no corresponding change. For example, an external service may alter its functionality, permissions, terms of use, model, tool description, data source or security arrangements. An external agent may also change its identity, operator, delegated authority or manner of interaction.

External changes that affect the deployment's approved capabilities, risk profile or the effectiveness of its boundary controls should be assessed and addressed through the AI's change-management arrangements. This requires the AI's approved interfaces, capabilities, dependencies and assumptions to be documented with sufficient precision to identify relevant changes and, where appropriate, supported by contractual notification, technical monitoring or other change-detection arrangements.

9.2 Different forms of external agent interaction

External agents should be treated differently according to whom they act for, the nature of their access and the AI's ability to control the relevant interaction. Three broad interaction types may be distinguished.

Interaction type	Illustrative example	Key governance focus
Outbound agent interaction	An AI's deployment uses an external agent, tool or service to obtain information, perform a task or initiate an action.	Governance of the AI's own deployment, including third-party risk, approved access, data outflow, dependency management and boundary controls.
Inbound agent interaction	An agent acting under a customer's or counterparty's mandate submits an instruction through the AI's channel.	Authentication, mandate and authority validation, access controls, accepted instruction types, execution controls, customer protection and revocation arrangements.
Peer or ecosystem interaction	Agents operated by different AIs, service providers or market participants exchange information or coordinate a workflow.	Identity, authority, data exchange, shared protocols, allocation of responsibility, dispute handling, common dependencies and operational resilience.

An external agent used by an AI's own deployment should be regarded both as a tool and, where applicable, a third-party or outsourcing dependency. It should be identified at the AI's boundary, and existing requirements relating to outsourcing, third-party risk, technology risk, information security, data protection and business continuity should apply. The AI remains accountable for deciding whether, how and within what limits its deployment may use that external agent.

An external agent approaching an AI on a customer's or counterparty's mandate should be treated as a channel for submitting instructions, rather than as part of the AI's own agentic A.I. deployment. Such an agent may be operated by another platform, a service provider, a corporate customer or counterparty, or directly by an individual. The AI may not be able to govern the external agent's underlying design or behaviour. Its focus should therefore be on the arrangements for accepting, rejecting, validating and executing instructions received through that channel.

In particular, the AI should determine the instruction types it is prepared to accept from an external agent, the information and authority evidence required, applicable limits, authentication arrangements, circumstances requiring human review, and the ability to restrict or revoke the agent's access independently.

9.3 Know Your Agent

A "Know Your Agent" or **KYA** approach may support both outbound and inbound agent interactions. KYA is not intended to be a single technical standard or a substitute for existing customer due diligence, authentication or access-management arrangements. Rather, it describes a risk-based approach to establishing sufficient confidence in the identity of an agent, the person or function on whose behalf it acts, and the scope of its delegated authority.

The depth of KYA arrangements should be proportionate to the nature and risk of the interaction. Relevant considerations may include:

- the agent's identity and the integrity of that identity over time;
- the operator, provider or other party responsible for the agent;
- the customer, counterparty, business function or other principal on whose behalf it acts;
- the identity and authority of the person or persons who granted its mandate;
- the scope, duration, conditions and limits of the mandate;
- the actions, products, accounts, data and systems to which the agent may have access;
- the means by which instructions, authority and consent are authenticated;
- the agent's ability to delegate tasks to another agent or service;
- arrangements for suspension, expiry, amendment and revocation of authority; and
- available records to support audit, investigation, dispute resolution and customer redress.

For example, where an AI accepts instructions from a customer's agent, it may need to authenticate the agent's identity through cryptographic signing, secure credentials or other appropriate technical means. It may also need to link that identity to the verified identity of the human customer and, where applicable, the authorised individual or individuals granting the mandate.

Authentication of an agent does not, by itself, establish that the agent is authorised to act. The AI should also verify that a requested action falls within the relevant mandate, account authority, product terms, transaction limits and other applicable conditions. This may require deterministic validation checks, rather than reliance solely on the agent's or principal's representation of authority.

The AI should also consider the ongoing validity of an external agent's authority. An authority may expire, be withdrawn, be superseded by a change in a customer's circumstances, or become ineffective where the agent's credentials have been compromised. Arrangements should therefore support timely revocation and should not assume that an authority established once remains valid indefinitely.

9.4 Acceptance and execution of external-agent instructions

Accepting instructions from agents acting for customers or counterparties may present heightened risks. These include impersonation, unauthorised access, mandate misuse, instruction tampering, credential compromise, ambiguity over the scope of authority, automated activity at a volume exceeding effective oversight, and disputes over whether an instruction was properly authorised.

AIs should maintain sufficient assurance that authorisation-validation mechanisms and deterministic guardrails governing acceptance and execution are robust and proportionate to the risk, scale and reversibility of the relevant instructions. Depending on the use case, such controls may include:

- restricting external-agent access to identified channels, products, accounts or services;
- requiring strong authentication and integrity protection for instructions;
- verifying that the agent is linked to a verified customer, counterparty or authorised representative;
- validating the instruction against a defined mandate and applicable limits;
- requiring additional customer confirmation or human approval for specified actions;
- applying transaction, value, frequency, velocity or cumulative-exposure limits;
- detecting unusual, conflicting or potentially fraudulent instructions;
- maintaining a clear record of the instruction, agent identity, mandate, validation performed and resulting action; and
- enabling prompt suspension or revocation of the external agent's authority.

Where appropriate to the use case and the AI's risk appetite, these measures could support more automated acceptance and execution of instructions from external agents within pre-approved mandates and limits. Such arrangements may operate under the human-on-the-loop oversight described in section 7.2, provided that the AI can detect and contain harmful activity in time and retains effective intervention capability, including the ability to suspend access or halt execution where necessary.

Validation of external authority should complement, rather than replace, the AI's own boundary controls. In particular, an AI should not accept an external agent's instruction to undertake an action that would be outside the AI's own approved policies, risk appetite, legal obligations or customer-protection arrangements.

9.5 Shared responsibility frameworks

The AI's appetite for entering into external-agent interactions may depend on whether responsibilities can be allocated clearly among the relevant parties. Depending on the arrangement, these may include the AI, the agent provider or operator, the customer or counterparty granting the mandate, and, in some cases, an intermediary platform or technology provider.

The HKMA's preliminary view is that a suitable shared responsibility framework should be practically understandable, mutually enforceable and operationally workable. It should address the responsibilities of each party for the relevant parts of the interaction, including identity, authentication, mandate establishment, instruction integrity, data handling, security, monitoring, incident notification, dispute management, suspension and customer redress.

Illustrative responsibility allocations may include the following.

Party	Illustrative responsibilities
AI accepting or using an agent	Defining accepted channels and interactions; maintaining its own boundary controls; validating authority before execution; protecting data and systems; monitoring activity; suspending access where necessary; handling customer complaints and meeting applicable regulatory obligations.
Agent provider or operator	Maintaining the agent's technical security and identity arrangements; accurately representing the agent's capabilities and status; protecting credentials; notifying relevant parties of compromises, material changes or incidents; and supporting investigation where required.
Customer or counterparty principal	Granting, maintaining, narrowing or revoking the agent's mandate; ensuring that the mandate reflects its intended instructions and authority; protecting credentials within its control; and promptly notifying the AI of relevant changes.
Intermediary platform or protocol provider	Maintaining the integrity and security of the relevant communication or execution channel; providing reliable records where agreed; managing protocol changes; and supporting incident response and dispute resolution within its role.

The allocation should be tailored to the particular interaction. For example, where an AI's deployment calls an external agent as a tool, the AI's relationship with the external agent provider will ordinarily be governed principally through third-party and outsourcing arrangements. Where a customer's agent submits instructions to an AI, the AI's principal concern will be the authority to accept and execute those instructions, while the customer and agent provider may have separate responsibilities for establishing and maintaining the mandate.

Shared responsibility arrangements should not dilute the AI's accountability for its own deployment, channels, acceptance arrangements or boundary controls. An AI remains responsible for ensuring that instructions accepted through its systems are handled in accordance with applicable requirements, its approved risk appetite and appropriate customer-protection arrangements.

The HKMA recognises that technical standards, market practices and legal arrangements for agent-to-agent interaction are still developing. Further coordination among AIs, customers, service providers, technology firms, industry bodies and relevant authorities may be needed to develop interoperable approaches to identity, delegated authority, revocation, evidence, liability and dispute resolution. The HKMA does not seek, at this stage, to prescribe a single technical standard or liability model.

Discussion question 9A
Is a risk-based approach to identifying, assessing and approving a deployment's external services, data feeds, tools and agents workable for the interfaces and protocols your institution uses or plans to use? For dynamic interfaces or services, what practical challenges arise in defining approved boundaries, detecting relevant external changes and managing their impact on the deployment's capabilities or controls?
The HKMA's views The HKMA considers that AIs should maintain clear, proportionate boundaries over the external services, tools, data sources and agents that their deployments may use. The form of control may vary by risk and use case, but AIs should be able to identify approved interactions, limit unapproved capabilities and assess material external changes through appropriate monitoring and change-management arrangements.

Discussion question 9B
Where an external agent acts on a customer's or counterparty's mandate, what arrangements would be needed to authenticate the agent and its principal, verify the scope and continuing validity of delegated authority, and control the acceptance and execution of instructions? How can these arrangements remain effective where the AI does not control the external agent's design or behaviour?
The HKMA's views The HKMA considers that authentication of an external agent should be distinct from verification of its authority to give a particular instruction. AIs should apply risk-based arrangements to establish the agent's identity, link it to the relevant principal and mandate, validate instructions against defined limits, and support timely suspension or revocation. The AI's own boundary controls and customer-protection obligations should continue to apply.

Discussion question 9C
What additional safeguards and practical, enforceable shared-responsibility arrangements would support automated acceptance and execution of external-agent instructions within your institution's risk appetite? Which elements would benefit from common conventions across AIs and market participants, and which parties should develop and maintain them?
The HKMA's views The HKMA considers that automated execution should be limited to clearly defined mandates, approved channels and pre-set limits, supported by reliable authentication, instruction-integrity controls, audit records, monitoring and effective suspension or revocation arrangements. Responsibility should be allocated clearly and operationally among relevant parties, without diluting an AI's accountability for its own systems and accepted instructions. Common conventions may be valuable for areas such as agent identity, delegated authority, revocation and evidence, but should remain adaptable as technology and market practice develop.

10. Way Forward

10.1 Next steps

The HKMA will consider the responses received alongside continued engagement with the industry. Practical experience from Sandbox trials and other industry engagement will also help inform further development of the HKMA's supervisory approach. The HKMA will consider whether further supervisory guidance or expectations relating to agentic A.I. are appropriate and, if so, the form they should take.

Discussion question 10A
Do the proposed governance approach and its core elements—intent-anchored accountability, lifecycle governance, a risk-integrated agent harness, human oversight and effective intervention, and traceability—provide an appropriate basis for governing agentic A.I. deployments? To what extent are these elements already achieved in substance through your institution's existing technology, risk-management and control arrangements? What would be the cost and practical challenge of demonstrating each element? Are there additional governance elements that the HKMA should consider?
The HKMA's views The HKMA considers that the proposed governance approach and its five core elements—intent-anchored accountability, lifecycle governance, a risk-integrated agent harness, human oversight and effective intervention, and traceability—provide an appropriate starting point for the governance of agentic A.I. deployments. The HKMA recognises that the practical measures needed to demonstrate these elements will vary according to the use case, the level of agent autonomy, the nature of data and system access, and the potential impact on customers, institutions and critical services. The HKMA recognises that institutions may already have relevant technology, risk-management and control arrangements in place, although these may require adaptation for agentic A.I. deployments. The proposed elements are deliberately concise and are intended to provide a common basis for discussion, rather than an exhaustive or prescriptive set of requirements. The HKMA will consider whether refinements, additional elements or further practical guidance are warranted in light of feedback received.

Discussion question 10B
How could the HKMA make its engagement with industry on agentic A.I. more timely, practical and effective, including in relation to future discussions and supervisory guidance formulation? How could practical experience, emerging risks, control practices, incidents and near-misses identified through such engagement be translated appropriately into future supervisory guidance or requirements?
The HKMA's views The HKMA recognises that agentic A.I. capabilities, deployment practices and associated risks may evolve more quickly than traditional supervisory-policy formulation cycles. A conventional

Discussion question 10B

approach based solely on formal consultations and subsequent issuance of supervisory materials may therefore not always enable expectations to be developed or refined with sufficient timeliness and practical specificity.

The HKMA's preliminary view is that future supervisory guidance or expectations relating to agentic A.I., where appropriate, may be developed through an even more interactive, iterative and evidence-based process. This could combine industry engagement with targeted thematic work, practical experience from the GenA.I. Sandbox++, and feedback on emerging control practices, incidents and near-misses. Such an approach would not reduce the importance of due process or transparency. Rather, it would allow any future supervisory requirements to remain proportionate, operationally workable and responsive to the evolving risks and capabilities of agentic A.I. systems.

10.2 How to respond

AIs, industry associations, technology providers and other interested parties are invited to submit comments by 1 February 2027 to the Banking Supervision Department of the HKMA at all-banks-go-fintech@hkma.gov.hk.

The HKMA may reproduce, quote, summarise or publish comments received, in whole or in part, in any form and for any purpose. Respondents wishing to withhold their name, affiliation or any part of their submission from publication should state this clearly in their response.

Comments are welcome on any aspect of this Paper, not only the discussion questions. Responses based on deployment experience, including incidents, near-misses and the practical cost of controls, will be particularly valuable. Respondents may provide examples in anonymised or aggregated form where appropriate.