Annex C

Concept Note on Governance of Agentic Artificial Intelligence

Introduction

Artificial intelligence (A.I.) is evolving from systems that support analysis and content generation towards **agentic A.I.** systems that can plan and execute multi-step tasks with varying degrees of autonomy. Agentic A.I. may access data, use tools and interact with internal or external systems in pursuing a specified objective. It may also involve multiple agents performing connected tasks.

Agentic A.I. offers opportunities for Authorized Institutions (AIs) to enhance risk management, anti-fraud measures and customer experience. The industry's interest in such systems was observed in the second cohort of the Generative Artificial Intelligence (GenA.I.) Sandbox. As agentic A.I. systems become more capable and autonomous, governance arrangements will need to evolve accordingly.

This Concept Note sets out the Hong Kong Monetary Authority's (HKMA's) preliminary considerations on the governance of agentic A.I. based on the initial observations from the GenA.I. Sandbox. It is intended to facilitate industry discussion and does not constitute specific supervisory requirements at this stage. The considerations are aligned with the HKMA's "ART" Framework, which seeks to foster an A.I.-enabled banking ecosystem with **Accountability, Resilience and Trust**.

Key governance concepts

(a) Intent-anchored Accountability

The actions of an A.I. agent are driven by its interpretation of a given instruction or business objective. Governance arrangements may therefore need to establish a clear connection between the agent's actions and the underlying business intent. Such intent defines the agent's authorised scope of autonomy as well as the person or function accountable for its use.

This concept, referred to as intent-anchored accountability, would support **clear allocation of responsibility throughout the lifecycle** of an agentic A.I. application. It may also provide a basis for assessing whether an agent's actions remain consistent with the intended business purpose.

For instance, the business function responsible for the underlying business intent would generally be expected to be accountable for the agent's outcomes. While technology functions provide necessary support, they may not always be the accountable owners.

(b) Visibility over material A.I. agents

Effective governance relies on adequate visibility over an agentic A.I. system. This may be particularly relevant for agents that perform material business functions, have a high degree of autonomy, or can access sensitive data and important systems.

A clear **inventory of material A.I. agents** could provide a common basis for continuous monitoring. Agentic AI applications may rely on interconnected models,

data, tools, infrastructure and third-party services. These interdependencies can make it more difficult to identify the root cause and potential impact of an incident. Building resilience therefore requires clear **component lineage**.

The HKMA is also exploring the possibility of an **industry-level registry for material A.I. agents** deployed by AIs. Such a registry could help ascertain accountability arrangements and enhance understanding of the supporting technologies and control measures adopted across the banking sector. It may also facilitate the identification of common dependencies, as well as the timely sharing of intelligence.

(c) Human oversight and intervention

Agentic A.I. can act at speed and may perform a series of connected actions. This highlights the importance of appropriate monitoring and human oversight. The level of oversight may depend on the nature and potential impact of the activity undertaken by the agent. The approach is moving from "human in the loop" to "human on the loop".

Active monitoring of A.I. agents' actions and decisions is essential to ensure **alignment with relevant business intent**. AIs may also need to consider the circumstances in which human approval or intervention is required, and retain the ability to respond promptly where an agent's actions depart from the business intent.

Intervention measures may include suspending the agent's access to particular tools or systems, or activating an emergency stop, such as a "**kill switch**". Supported by timely intelligence sharing, appropriate **intervention points** may need to be built into agentic A.I. workflows before actions with material customer implications are carried out.

(d) Operating boundaries and safeguards

A.I. agents should operate within clearly defined boundaries, such as the tools, systems and data that the agents are permitted to access or invoke. Such arrangements, which may be supported by "**risk-integrated agent harness"**, can help translate business intent, risk appetite and internal controls into practical operating constraints.

Appropriate safeguards help protect customer and bank data from being accessed, transmitted or misused by A.I. agents without authorisation. Clear operating boundaries are essential to maintaining customer and public trust in the use of agentic A.I.

Next steps

The HKMA will continue to monitor developments in agentic A.I. through engagement with the industry and the GenA.I. Sandbox++. Practical experience from Sandbox trials will help inform the future development of supervisory expectations.

This Concept Note presents preliminary considerations only. The HKMA will soon issue a discussion paper that makes reference to this Concept Note and set out options for possible supervisory requirements relating to the governance of agentic A.I. The HKMA will continue to adopt a collaborative approach during the process and seek views from AIs and relevant stakeholders.