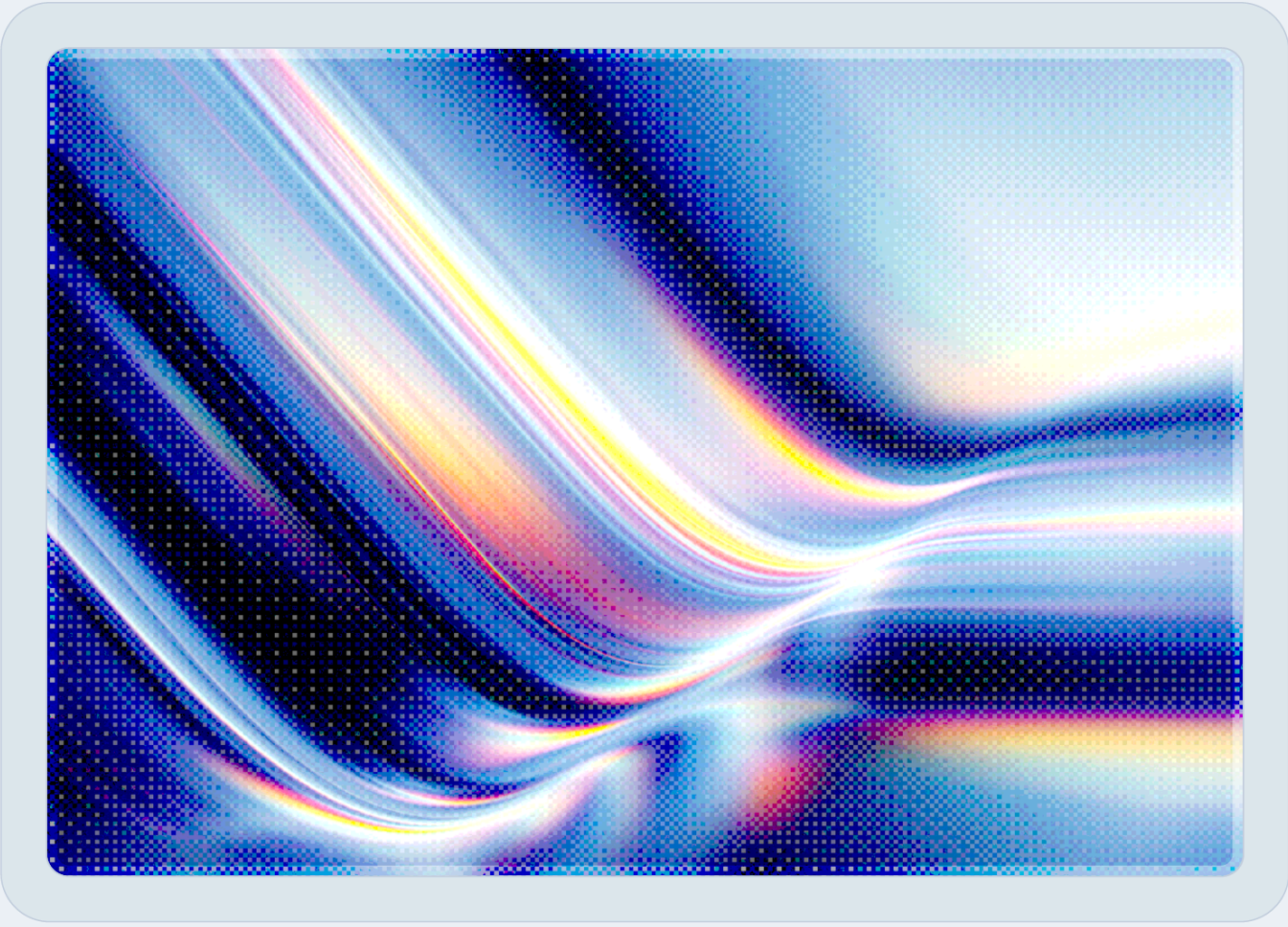
CSAI Foundation | Cloud Security Alliance

Registering the Agent Fleet: HKMA's Proposed Material AI Agent Registry

Security implications of a supervisory inventory obligation for agentic AI in banking

2026-11-28

AI-assisted Rapid Research



CSAI

CSA cloud security alliance®

© 2026 Cloud Security Alliance. Some rights reserved.

You may download, store, display, view, print, redistribute, and link to this document in its original, unmodified form, provided that attribution to the Cloud Security Alliance is maintained and all trademark and copyright notices remain intact.

This document may not be modified or altered. You may quote portions of the document as permitted by the Fair Use provisions of the United States Copyright Act, provided that attribution is given to the Cloud Security Alliance.

This document may be shared on professional and social media platforms in its original form with attribution.

This document was generated with AI assistance and has not undergone official CSA review and approval processes.

Key Takeaways

On 2 November 2026 the Hong Kong Monetary Authority (HKMA) published its *Discussion Paper on Agentic A.I. Governance in the Banking Sector*. The paper asks for views on a proposed Registry of material AI agents, which would be fed by the internal deployment records that authorized institutions keep [1][2]. It is not a rule. The HKMA says plainly that the paper "creates no new obligations and does not modify or interpret existing ones," and it invites responses by 1 February 2027 [1]. Even so, it sets out in unusual operational detail what a supervisor may soon expect a bank to show about every agent it runs. For security teams, the proposal is effectively a draft specification for an agent inventory.

Five points matter most for security and risk leaders.

First, the internal record is the real baseline. The HKMA's preliminary view is that *every* agentic deployment within the paper's definition should be recorded before production use, not only material ones. The proposed fields cover owners, capabilities and data egress, shared dependencies, the stop mechanism, log locations, and how credentials are handled at retirement [1].

Second, the materiality test is about consequences, not technology. A deployment is material when it has significant "reach" over positions, obligations, or critical operations and no genuine per-action human approval stands in the way. Irreversible actions, and read-only agents that a critical operation depends on, stay material even when a human gate exists [1].

Third, the proposed notification content reads like a security architecture review. It includes a tools catalogue that names Model Context Protocol (MCP) servers and API endpoints, an identity and privilege inventory, execution controls, and tamper-resistant logging [1].

Fourth, the paper proposes that certain agent-specific events be treated as reportable incidents: an agent acting beyond its authority, a stop mechanism that fails, a compromised agent credential, a silently failed monitoring agent, or an approval gate that lets an action through improperly [1].

Fifth, sector-wide aggregation is an explicit goal. The HKMA wants to see "common dependencies, concentrations and potential vulnerabilities across the banking sector," which brings shared model and orchestration providers into supervisory view [1].

Background

The discussion paper is the third step in a sequence the HKMA set out during the autumn of 2026. On 15 October 2026 it announced the "ART" Framework (Accountability, Resilience and Trust), together with updated Supervisory Guidance for Responsible Adoption of Artificial Intelligence and a Concept Note on Governance of Agentic Artificial Intelligence [3][4][5]. In that announcement the HKMA reported that more than 60% of the use cases in the second cohort of its Generative AI Sandbox explored agentic AI. It also said an upcoming discussion paper would cover "intent-anchored accountability, a registry of material A.I. agents, and effective intervention mechanisms" [3]. The paper itself appeared on 2 November 2026, released alongside Project SAGE (Securing Agentic Finance with Governance and Enablement) and a cross-sectoral Agentic Finance Expert Group [2][6]. The HKMA also places its approach in line with the Financial Stability Board's October 2026 sound practices on AI adoption [1][7].

The paper's organizing idea is "intent-anchored accountability." An agent should act only in pursuit of an approved business intent and within a defined delegation of authority. Accountability for that delegation must stay with named individuals inside the institution, and the HKMA says it "should not be diluted by the use of external models, cloud services, data providers, tools or other third-party components" [1]. On this foundation the paper builds five governance elements: intent-anchored accountability, lifecycle governance, a risk-integrated agent harness, human oversight with effective intervention, and traceability. The Registry proposal in Section 8 serves as the supervisory interface to the internal traceability records described in Section 7.3.

Scope deserves careful reading. The paper defines agentic AI by capability: goal-directed, planning, tool-using, carrying memory, and adaptive. For operational purposes it focuses on systems that act through tools and choose their own steps. It excludes fully prescribed robotic process automation, single fixed-task models, and drafting assistants that only produce content for a person to act on [1]. Coverage applies "regardless of whether the system is hosted on [the institution's] infrastructure." It includes vendor-provided and group-provided agents with access to the bank's Hong Kong data, systems, or processes, as well as instructions that arrive from a customer's or counterparty's agent [1]. The paper also expects banks to treat unauthorized staff use of agentic tools, or "shadow A.I.", under their existing governance arrangements [1].

Hong Kong is not alone in this direction. CSA's analysis of the UK Financial Conduct Authority's 24 November 2026 statement on agentic AI found that it likewise leaves accountability with senior managers even when an agent comes from a third party [8]. CSA's review of OCC Bulletin 2026-54 found that the U.S. examiner framework also scales control intensity with consequence and reversibility rather than technical sophistication [9]. What sets the HKMA paper apart is its concrete proposal for a regulator-facing register of individual agents, with a defined notification schema.

Security Analysis

The internal record as an agent configuration baseline

The fields the HKMA proposes for each deployment record line up closely with the artifacts a security team would need to contain an agent incident. Table 1 maps each proposed field to the security function it supports.

Table 1. Proposed deployment record fields and their security function (fields from HKMA Discussion Paper §7.3 [1])

Proposed record field	Security function it supports	Typical gap today
Purpose, intended use, deployment status	Baseline for detecting scope drift and unauthorized use	Purpose captured at approval but not linked to runtime permissions
Accountable, business, technical and control owners; authorized users	Escalation path during incidents; access recertification	Owners leave or change roles; agents become orphaned
Capabilities and access, including data leaving the institution and recipients	Data-loss prevention scoping; third-party data-flow mapping	Tool permissions granted broadly; egress unmapped
Infrastructure and dependencies, including shared ones	Blast-radius analysis; concentration risk	Shared models and orchestrators not tracked per deployment
Stop mechanism, fallback, log locations	Containment and forensic readiness	Kill switches untested; logs scattered across vendor consoles
Change and suspension history; memory, state and credentials at retirement	Change control; credential hygiene at decommissioning	Retired agents leave active tokens and stored context behind

Proposed record field	Security function it supports	Typical gap today
Materiality assessment, rationale, next review date	Risk-tiering for control intensity	No consistent tiering method across business lines

Two features of the proposal deserve emphasis. First, the HKMA tells institutions to build attribution in from the start: "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" [1]. It also expects each deployment to use "distinct identities and credentials where appropriate," because shared or borrowed accounts break the link between an action and the authority behind it [1]. Second, the paper treats retirement as a control point, warning that "casually decommissioned deployments may leave active credentials, access permissions or stored data behind" [1]. Both points match the identity-centric failure modes, orphaned agents and untraceable actions, that CSA's agentic identity work has examined [10][11].

A consequence-based materiality test

The proposed classification works in two steps. Step 1 asks whether a deployment has significant reach in any of three areas. *Positions* covers moving funds, altering accounts or data, creating commitments, or deciding a customer's entitlement. *Obligations* covers discharging a duty owed to an authority or market. *Operations* covers being something a critical operation depends on [1]. Significance is judged against thresholds the institution already uses, such as its critical-operations, outsourcing-materiality, consumer-protection and incident-reporting thresholds. The paper adds no new numerical thresholds [1]. Step 2 asks whether genuine human approval of each action stands between the agent and the harm. If it does not, the deployment is material. If it does, the deployment is still material when the gated action is irreversible [1].

Several features of the test have direct security consequences. Reach is assessed "over the time needed to detect and stop" the agent. As a result, "a stream of individually small, reversible actions may therefore be significant if its pace exceeds the AI's [authorized institution's] ability to identify and halt it" [1]. In HKMA usage, "AI" in these quotations refers to an authorized institution rather than to artificial intelligence. Detection and containment speed therefore feed directly into regulatory classification. A deployment that "directs other deployments inherits their reach," so orchestrator agents take on the materiality of their sub-agents [1]. Deployments that share "a model, platform, orchestrator or critical dependency" are assessed together, so several individually modest agents may be material as a group [1]. The paper also sets a high bar for a human gate to count. The approver must see the substance of the

action, have authority and practical ability to reject it, and face a pending action. Rejection must be accepted as a legitimate outcome, and institutions need "ongoing evidence that the gate remains genuine" [1]. Rubber-stamp approval under high volume would leave a deployment classed as ungated.

Table 2. HKMA classification examples (from Discussion Paper Exhibit 2 [1])

Deployment	Reach	Human gate	Outcome
Internal research and drafting assistant	None	n/a	Not material
Agent closing false-positive financial-crime alerts	Obligations (closes a regulatory duty)	None per action	Material
Wholesale payment-release agent	Positions	Genuine gate on every payment	Material (settlement may be irreversible)
Read-only agent that is the sole automated fraud monitor	Operations (critical dependency)	Approval would not prevent its failure	Material

The fourth example matters for security teams that classify risk mainly by write access. Under the HKMA approach, a read-only monitoring agent can be material because its silent failure would remove a control. That is why "a material monitoring deployment is found to have failed silently" appears on the proposed list of agentic incidents [1].

Notification content as a security architecture disclosure

For material deployments, Exhibit 3 of the paper lists what a notification to the HKMA may cover. Beyond accountability and oversight, it calls for a technical overview with seven parts: an architecture and dependency map; a model registry explaining how multiple models interact; a tools catalogue of "material external applications, Model Context Protocol (MCP) servers, API endpoints, or data store that the agent is permitted to call"; data storage and privacy arrangements; identity and access privileges; execution controls such as approval checkpoints and usage limits; and logging and traceability arrangements [1]. The paper also asks institutions to explain "what prevents unauthorised expansion of capability" [1].

This content is more technical than typical AI use-case inventories, which often record purpose, owner and risk rating but not the tool surface or the credential model. Naming MCP servers in a regulatory schema suggests the HKMA sees agent tool connectivity as a supervisory concern in its own right. Institutions that cannot list which tools an agent can call, or under which identity, would probably find notification hard to complete. The HKMA's stated aim is for institutions to reuse evidence they already keep for their own governance rather than produce separate material for the regulator [1]. In practice, that makes the quality of internal tooling the deciding factor in reporting cost.

The paper presents notification timing as an open question. It asks whether material deployments should be notified before production or soon after, and whether a bank could proceed without waiting for a supervisory response [1]. The HKMA's preliminary view is that "timely post-deployment notification may be appropriate in most cases," with pre-deployment notification reserved for higher-impact circumstances. It also states that it "does not intend to change the notification arrangement to become a general pre-approval process" [1]. Material changes would follow the same timing. The proposed definition of a material change is broad: expanding capabilities or reach, removing or weakening per-action approval, materially altering the underlying model or a critical dependency, extending to new customers or products, changing owners, or returning the agent to service after a suspension or incident [1].

Stop mechanisms, incident triggers and supervisory intervention

The paper sets a practical test for kill switches. An institution must show it "can deliberately stop or constrain the deployment and maintain business continuity," and "if an AI [authorized institution] 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" [1]. Possible ongoing-assurance arrangements include retesting the stop mechanism and fallback at least once a year and after every material change, risk-proportionate independent validation, internal audit coverage, and an annual attestation by the accountable owner [1]. The paper also asks for views on supervisory measures where risks are not adequately managed, up to "suspension-related measures, or 'kill switch', for the agent" in exceptional circumstances [1].

The five proposed agentic incident triggers each match a distinct security failure. Acting beyond authority points to permission or guardrail failure. A stop mechanism that fails points to containment failure. Credential compromise is an identity failure. A silent monitoring failure is a detection failure. An approval gate that lets an action through improperly is a human-in-the-loop control failure. The paper further suggests a risk-intelligence sharing mechanism among banks and the HKMA covering

vulnerabilities, failure modes, incidents and near-misses [1]. If adopted, these triggers would make agent telemetry a compliance dependency. An institution that cannot detect an out-of-authority action cannot report one.

Inbound agents and Know Your Agent

Section 9 extends governance to agents acting for customers or counterparties. The proposed risk-based "Know Your Agent" (KYA) approach would establish an agent's identity, its principal, the person who granted the mandate, the mandate's scope and duration, and how it can be revoked [1]. The paper draws a security-relevant distinction: "Authentication of an agent does not, by itself, establish that the agent is authorised to act". It recommends deterministic validation that a requested action falls within the mandate, account authority and transaction limits [1]. The paper also asks banks to treat external content as untrusted input [1]. In effect, it tells institutions to defend against prompt injection and confused-deputy attacks at the boundary rather than relying on the inbound agent's own claims. Agent identity and KYA are also one of the four focus areas of Project SAGE [6].

Recommendations

Immediate Actions

Institutions with Hong Kong operations should first find every system that meets the paper's two operational indicators, acting through tools and choosing its own steps. That includes vendor and group-provided agents and agentic features embedded in vendor products, which the HKMA names as likely boundary cases [1]. Security teams should check that each identified agent has its own non-human identity and credentials rather than a shared service account. They should also confirm that its actions can be attributed to both the deployment and the instructing human. Teams should identify any agent whose stop mechanism has never been exercised in a production-like setting, because the HKMA's "in substance" test will turn on demonstrated rather than documented capability [1]. Institutions that want to shape the outcome should prepare evidence-based responses before the 1 February 2027 deadline. The HKMA says it particularly values examples of boundary systems and the practical cost of controls [1].

Short-Term Mitigations

Over the next two to three quarters, institutions should extend existing AI use-case inventories with the Section 7.3 fields, using Table 1 as a gap checklist. Priority should go to dependency mapping, data-egress recipients, and credential disposal at retirement. Running a dry run of the two-step materiality test

against current deployments is worthwhile. It will likely surface orchestrators that inherit reach, read-only monitors on critical paths, and clusters of agents that share a model or platform. Teams should then build the Exhibit 3 tools catalogue directly from configuration, such as MCP server allow-lists, API gateway policies and IAM bindings, rather than from questionnaires, so it stays current. Detection engineering should add rules for the five proposed incident triggers, with particular attention to silent failure of monitoring agents and to approval gates whose approval rates or decision times suggest rubber-stamping.

Strategic Considerations

The Registry proposal points toward supervisors treating agents as a distinct class of regulated asset, with lifecycle obligations similar to those for material outsourcing or critical operations. Institutions should expect convergence across jurisdictions on consequence-based tiering and named human accountability. The HKMA, FCA and OCC materials reviewed here all lean that way, although their mechanisms differ [1][8][9]. A single internal agent record designed to meet the most demanding of these regimes is likely to cost less than separate per-jurisdiction inventories. Vendors that supply agentic capabilities to banks should prepare to provide the architecture, model, tool and identity information banks will need for notification, and to report material changes such as model swaps promptly. The HKMA's interest in sector-wide concentrations also suggests that shared model and orchestration providers may draw indirect supervisory attention. Institutions should model what happens if a dependency shared across many deployments is compromised.

CSA Resource Alignment

CSA's research note *Who Did That? Agent Action Attribution as an Audit Obligation* anticipates the HKMA's attribution expectations most directly [11]. It proposes a four-part evidentiary record covering which agent acted, on whose behalf, under what delegated authority, and what it did. It also cites CSA and Aembit survey data indicating that 68% of organizations cannot clearly distinguish AI agent activity from human activity [11]. The HKMA's statement that "the log replaces the memory and signature of a human actor," and its proposed notification fields on visibility and logging, are a regulatory counterpart to that record [1]. Institutions can use the attribution note's framework to structure the logging part of their Section 7.3 records.

CSA's *Agentic AI Identity & Access Management: A New Approach* addresses the identity layer the HKMA proposal relies on [10]. That includes distinct agent identities, delegated authority, and credential lifecycle through to retirement. The paper's call for distinct credentials per deployment, its Exhibit 3 identity and

privilege disclosure, and its KYA approach for inbound agents all depend on these identity primitives [1]. CSA's analyses of the FCA's agentic AI statement [8] and OCC Bulletin 2026-54 [9] give comparative context for institutions that operate across Hong Kong, the UK and the US.

For control mapping, the CSA AI Controls Matrix (AICM) v1.1 provides a vendor-neutral control set covering governance, identity and access management, logging and monitoring, supply chain, and change control [12]. Institutions can map the HKMA's record fields and notification content to it, which helps reconcile Hong Kong obligations with other frameworks. CSA's MAESTRO agentic threat-modeling framework [13] gives a structured way to assess the "reach" and failure modes the materiality test requires. It is particularly useful for orchestrated multi-agent deployments, where the HKMA proposes that reach is inherited and that shared dependencies be assessed together.

References

[1] Hong Kong Monetary Authority. "Discussion Paper on Agentic A.I. Governance in the Banking Sector." HKMA, November 2026.

[2] Hong Kong Monetary Authority. "Hong Kong FinTech Week 2026." HKMA press release, 2 November 2026.

[3] Hong Kong Monetary Authority. "HKMA announces "ART" Framework to promote responsible A.I. adoption." HKMA press release, 15 October 2026.

[4] Hong Kong Monetary Authority. "Supervisory Guidance for Responsible Adoption of Artificial Intelligence." HKMA, October 2026.

[5] Hong Kong Monetary Authority. "Concept Note on Governance of Agentic Artificial Intelligence." HKMA, October 2026.

[6] Hong Kong Monetary Authority. "HKMA launches Project SAGE and establishes Expert Group for agentic finance." HKMA press release, 2 November 2026.

[7] Financial Stability Board. "Sound Practices for Responsible Adoption of Artificial Intelligence (AI): Final report." FSB, October 2026.

[8] Cloud Security Alliance. "FCA on Agentic AI: Senior Managers Own Third-Party Agents." CSA AI Safety Initiative, November 2026.

[9] Cloud Security Alliance. "OCC Bulletin 2026-54: Examiner Framework for Agentic AI in Banking." CSA AI Safety Initiative, November 2026.

[10] Cloud Security Alliance. "Agentic AI Identity & Access Management: A New Approach." CSA, 2025.

[11] Cloud Security Alliance. "Who Did That? Agent Action Attribution as an Audit Obligation." CSA AI Safety Initiative, November 2026.

[12] Cloud Security Alliance. "AI Controls Matrix (AICM) v1.1." CSA, accessed November 2026.

[13] Cloud Security Alliance. "Agentic AI Threat Modeling Framework: MAESTRO." CSA, February 2025.