FSB FINANCIAL STABILITY BOARD

Sound Practices for Responsible Adoption of Artificial Intelligence (AI): Consultation report

Response to Consultation

Stephen T (independent response)

1. **Do you agree with the benefits and risks of AI adoption by financial institutions described in this report? Are there any substantive benefits or risks not covered?**

The report accurately captures the risk landscape, and from a fraud practitioner's seat the point it makes about non-adoption deserves emphasis: adversaries are adopting these tools faster than most institutions, so declining to use AI in fraud defense is itself a risk position. The agentic AI risk discussion in Section 1.3 (unauthorized actions, erroneous actions, data breaches, disruption to connected systems, oversight limits, memory poisoning) is accurate and useful.

Two substantive gaps and Recommendations:

(1) Agentic financial-crime typologies. The report describes agentic risks as operational and cyber phenomena. From a financial-crime perspective, the same properties - autonomy, speed, delegation, and machine-mediated authorization - produce recognizable fraud and laundering typologies that compliance functions will need to manage. Examples worth naming in the report:

- Manipulated-authorization fraud. Social engineering shifts from tricking a human into sending a payment to tricking a human into granting an agent an over-broad or misdirected authorization. This is the agentic analogue of authorized push payment fraud, and controls (and reimbursement debates) developed for that typology will need agentic equivalents.

- Agent takeover and impersonation. Account takeover, at machine speed, across many principals at once. Compromise of an agent's credentials, or impersonation of a legitimate registered agent, scales in a way individual account compromise never did.

- Synthetic and stolen principals. Agents operated under synthetic or stolen identities move KYC evasion up one layer: the agent may be technically "verified" while its accountable principal is fictitious - a verification obligation at the agent-principal binding, not only at the customer account. Attribution of agents to real, accountable principals is the KYC problem of the agentic era.

- Mule-agent networks. Agents used as money mules could execute layering and dispersion patterns continuously, including adversaries deploying disposable sub-agents at

scale, at volumes human mule networks cannot match - compressing laundering timelines from days to seconds.

- Oversight-fatigue exploitation. Where human-in-the-loop approval exists, adversaries can deliberately generate high volumes of benign-looking agent requests to condition reviewers toward reflexive approval - a weaponization of the automation bias the report already identifies.

- Cross-institution velocity abuse. Agent activity that appears low-risk to each individual institution can be high-risk in aggregate; no single institution sees the full pattern. There is also a stability dimension the FSB is well placed to monitor as adoption scales: if agent-driven layering at machine speed triggers simultaneous automated defensive responses (asset freezes, connection cut-offs) across multiple institutions, the interaction could produce localized liquidity or settlement strain. Both points argue for the information-sharing mechanisms encouraged under Sound Practice 4 to explicitly include agentic fraud typologies.

(2) Repudiation and dispute risk. The report addresses unauthorized and erroneous agent actions as they occur. The mirror-image risk arises afterward: customers, counterparties, or the institution itself disputing whether an agent's action was authorized. An institution that cannot evidence who authorized what, within which constraints, and what the agent actually did faces lookback exposure, restitution demands, and liability disputes it cannot defend. This evidentiary dimension deserves explicit treatment as a risk, and it is the basis for my recommendation under Question 2 on evidence standards.

Finally, the risk discussion treats agents primarily as tools the institution deploys. Financial institutions will also face third-party agents acting on behalf of customers, merchants, vendors, and other counterparties. Agent-payment programs and open protocols announced in 2025-2026 already use signed authorization artifacts, verifiable credentials, and agent-specific cryptographic signatures to evidence user intent and agent authority.[1][2] This raises institution-side questions the report does not yet fully address: where sanctions and counterparty screening should occur relative to agent execution; how an inbound agent is attributed to an accountable principal; and how the institution validates, retains, and reconciles the third-party authorization artifacts it relies on. An IMF note has framed the public-policy version of this shift as moving from know-your-customer toward know-your-agent requirements grounded in verifiable identities, delegated authority, and auditability. [3]

Notes cited in this answer:

[1] Google Cloud, "Powering AI Commerce with the New Agent Payments Protocol (AP2)," 16 September 2025, and the AP2 specification (https://ap2-protocol.org), describing cryptographically signed Intent, Cart, and Payment Mandates carried as verifiable credentials.

[2] Mastercard, Agent Pay announcement (29 April 2025) and Verifiable Intent materials (2025-26), describing agent-bound payment tokens and signed intent records supporting authorization and dispute evidence; Visa, "Trusted Agent Protocol" announcement (October 2025), describing agent recognition via cryptographic signatures.

[3] Davidovic, Sonja, and Hervé Tourpe. 2026. "How Agentic AI Will Reshape Payments." IMF Note 2026/004, International Monetary Fund, Washington, DC. April 2026. https://doi.org/10.5089/9781513533308.068.

2. Are the sound practices sufficiently comprehensive and clear to enable financial institutions' responsible AI adoption?

The twelve practices are well structured and the proportionality framing is right. Five additions would make them materially more operational for agentic AI. In each case the aim is the same: convert an aspiration into an examinable control.

Recommendations

(1) Tested deactivation (Sound Practices 10 and 11). The report correctly identifies that overriding or remediating agent actions "can be difficult or impossible." The final text should therefore state that, for material or high-risk agent deployments, deactivation capability is a control to be designed, measured, and tested. That means a defined time objective from decision to full deactivation; revocation that propagates to all of the agent's credentials, active sessions, tool access, and any delegated sub-agent credentials (deactivating the agent while its tokens remain live is not deactivation); and periodic exercises, analogous to business-continuity and disaster-recovery testing, which fits naturally within Sound Practice 11's direction to incorporate AI scenarios into tests and exercises. A 2026 Wolters Kluwer survey of 230 US banking professionals found that 72% identified either regulatory reporting of an AI failure (37.83%) or model kill-switch protocols (34.35%) as the area of AI-related risk for which their institution is least prepared.[4] A tested-control expectation would address that gap directly.

(2) Identity attribution and accountable principals (Sound Practices 2 and 3). The documentation case study under Sound Practice 3 - agent-level inventories capturing tools, components, guardrails, and "certification" within defined boundaries - describes what I regard as the emerging baseline, and I recommend elevating its core elements into the practice text. Each agent should be bound to a unique identity and assigned to an accountable human or organizational owner. It should operate under scoped, expiring credentials rather than shared or long-lived ones. Its authorization scope, permitted financial authority (can it move funds, commit the institution, alter customer data?), and delegation depth should be recorded as inventory attributes, and the inventory should distinguish internally deployed agents from external, third-party agents, which warrant different verification and screening expectations. Without unique identity and ownership, the accountability chain that Sound Practice 2 requires cannot reach the agent's actions.

(3) Evidence standards for agent actions (Sound Practices 3 and 8). The report's logging and documentation provisions - prompt and version history, agent activity logs, tool invocations, intermediate decisions, and lifecycle documentation - address whether records exist. For material agent use cases, especially any that initiate transactions, the guidance should also address what quality of record suffices. I suggest the final text name five properties. Records should be attributable (linked to a specific agent identity and its principal), tamper-evident, and time-synchronized. They should be independently verifiable, not solely self-attested by the agent platform or model vendor - a point that intersects with the third-party transparency limits the report already flags under Sound Practice 12. And

they should be retained and reconcilable across the parties to a transaction. The test I would put to any institution, because it is the one a supervisor or a court may eventually apply: after an incident or dispute, can you reconstruct who authorized what, within which constraints, and what the agent actually did, to a standard that survives independent examination? Signed authorization artifacts now emerging in the payments industry are useful inputs to that file but are not, by themselves, the file.

(4) Deterministic boundaries outside the model (Sound Practice 10). The report rightly notes that real-time human review does not scale with agent volume. The corollary belongs in the sound practice itself: for agents with financial authority, hard limits - spending caps, counterparty allow/deny lists, velocity limits, sanctioned-jurisdiction blocks - should be enforced by deterministic, out-of-band policy mechanisms outside the model's execution environment, such as an API gateway, policy engine, or ledger-control layer, so that a compromised, manipulated, or hallucinating agent cannot exceed them regardless of its reasoning. Human oversight then shifts to where it scales: boundary design, exception review, and risk-based sampling of agent decisions. Two further corollaries follow. First, the configuration of those enforcement mechanisms should itself be integrity-protected and subject to change control, so that an alteration of an agent's limits is as detectable, attributable, and reviewable as the agent exceeding them; a boundary that can be silently rewritten is not a boundary. Second, boundaries should apply consistently across every channel and rail through which an agent can act. Agents, and adversaries steering them, will route activity through the least-controlled channel; a limit enforced on one payment channel but absent from another is, in practice, not a limit. This is consistent with the report's compensating-controls approach under Sound Practice 8 and its "policy-aware data" concept under Sound Practice 7.

(5) Ongoing agent due diligence and point-of-execution standing checks (Sound Practices 5, 9, and 10). The report's monitoring provisions address whether an agent is performing as intended. A separate question is whether it remains authorized and trustworthy, and for agents the answer to that question decays in seconds rather than review cycles. An agent's risk state can change materially between onboarding certification and any given action: its credentials may be compromised, its code updated or mutated since certification, its authorization amended or revoked, adverse behavioral signals accumulated, or the standing of its accountable principal changed. This is not a duplicate of onboarding inventory or certification; it is the execution-time check that the inventoried identity, authority, and risk status still hold. Customer due diligence is already migrating from periodic refresh toward perpetual, event-driven review; agents require a faster version of that same shift. The guidance should therefore state that, proportionate to materiality and financial authority, an agent's standing is re-validated at or near the time of execution. At minimum: confirmation that its credentials are valid and unrevoked (not merely unexpired), that the authorization artifact under which it acts is the current, unaltered version, and that its certified integrity and risk status remain in good standing. Deeper reassessment should be triggered by events such as code changes, behavioral anomalies, elapsed time, or principal-level changes. Standing checks of this kind are lightweight and cacheable, compatible with transaction-speed requirements - a design pattern analogous to continuous access evaluation in modern identity infrastructure - so this is a re-validation expectation, not a per-transaction re-certification burden. Paired with recommendation (a), it closes the control loop: point-of-execution standing checks are the sensor; tested deactivation is the response.

Notes cited in this answer:

[4] Wolters Kluwer Compliance Solutions, "US Banking AI Risk and Governance Index," May 2026 (survey of 230 US banking professionals). https://www.wolterskluwer.com/en/news/wolters-kluwer-survey-highlights-ai-governance-and-other-ongoing-needs-for-banking-institutions.

3. **Do the sound practices strike an appropriate balance between managing risks relating to all forms of AI, and addressing some of the risks relating to emerging and new complex forms of AI, such as GenAI and agentic AI?**

Broadly yes. The technology-neutral core with agentic-specific elaboration is the right architecture, and the agentic risk description in Section 1.3 is notably detailed for a cross-sector sound-practices document.

The imbalance is between the risk section and the practice text: agentic risks are described richly, but the corresponding controls (identity attribution, tested deactivation, evidence standards, deterministic boundaries, pre-execution screening for agent-initiated payments) appear mainly in case studies, footnotes, or by implication.

Recommendation

Question 2 above lists the specific promotions I recommend. One addition specific to balance: the practices address the institution deploying agents but not the institution facing them. Agent-initiated payment capabilities are moving from protocol design into pilots and early commercial implementations, so a short subsection on managing third-party agents acting on behalf of customers and counterparties (verification of agent identity artifacts, principal attribution, screening placement, artifact retention) would make the toolkit more durable as agent-mediated payments develop. I recognize the FSB may view parts of this as sitting with payments-focused standard setters; if so, a cross-reference would still help institutions locate the perimeter.

4. **Are the sound practices sufficiently flexible to accommodate and address newer types of AI and responsible adoption over time?**

Yes, with one structural suggestion: the durability of the practices will depend on anchoring them to capability and authority attributes rather than technology labels. Sound Practice 5 already moves this way (autonomy, scope of action, access to systems and data).

Recommendation

I recommend making the attribute set explicit and extensible, adding at minimum: financial authority (whether the AI can move value or bind the institution), delegation depth (whether it can create or task other agents, and how far the chain extends), and counterparty exposure (whether it transacts with parties outside the institution's control). Practices indexed to these attributes will apply cleanly to whatever succeeds today's agent architectures; practices indexed to "GenAI" or "agentic AI" as category labels will date quickly. The external-awareness provisions of Sound Practice 4 should also explicitly include monitoring of emerging authorization and identity standards for agents, since institutions will be consuming these artifacts whether or not they help design them.

5. **Do the case studies in this report sufficiently highlight how different types of financial institutions can benefit from responsible AI adoption? Are there additional case studies for inclusion in the report? If so, please provide such case studies, particularly for nonbanks.**

The case studies are one of the report's strengths, particularly the agentic fraud-detection example with human-in-the-loop rule approval, and the agent-inventory/certification example. Three topics are currently uncovered and would add practical value:

1. A deactivation exercise. No current case study covers incident response for a misbehaving agent - detection, decision, revocation propagation, verification of full deactivation, and post-incident review. Even a stylized example would set expectations.

2. Financial-crime controls for agent-initiated payments. Public market examples in 2026 show compliance controls being placed at the autonomous transaction layer, including sanctions screening, counterparty-risk checks, policy thresholds, and pre-settlement blocking.[5] A case study on pre-execution screening placement, as distinct from post-hoc monitoring, would be directly actionable for banks and for nonbank payment providers, whose agent exposure may arrive first.

3. Dispute and evidence handling. A case study showing how an institution reconstructed and evidenced an agent-initiated transaction in a customer dispute would make the documentation practices concrete.

I would be glad to assist the FSB in developing any of these, within the limits of public information.

Notes cited in this answer:

[5] Cited as a public market example rather than a regulatory standard: "ampersend and TRM Labs Launch Real-Time Compliance Screening for AI Agents," CryptoBriefing, June 2026 (https://cryptobriefing.com/ampersend-trm-labs-ai-compliance-screening-enterprise/), describing sanctions screening, counterparty risk controls, and pre-settlement blocking applied at the agent execution layer; see also https://ampersend.ai.

6. **Do the case studies in this report provide actionable insights for financial institutions in their responsible AI adoption?**

Yes. One small improvement: each case study could close with a one-line statement of the control objective served and the evidence the practice produces (for internal audit or supervisory review). This converts illustrations into templates institutions can operationalize, and aligns with the report's own emphasis on documentation.

7. **Are the definitions in the glossary clear and aligned with industry sound practices, including recent developments in AI?**

The glossary would benefit from defining several agentic-accountability terms on which the practice text depends: AI agent identity (the unique, verifiable identifier and credentials under which an agent acts); principal (the accountable human or legal entity on whose behalf, and under whose authority, an agent acts); delegation / delegation chain (the grant

of authority from principal to agent, or agent to agent, including its scope and expiry); authorization artifact / mandate (a verifiable record - increasingly cryptographically signed in industry practice - of what a principal authorized an agent to do, under which constraints); agent standing (the current validity of an agent's identity, credentials, authorization, integrity, and risk status at the time it seeks to act); and deactivation (the verified revocation of an agent's ability to act, including all credentials, sessions, and delegated authority). Precise shared vocabulary here will matter as much as the practices themselves, because these are the terms in which supervisors, institutions, and vendors will negotiate accountability.

8. Are there any comments you would like to make on this report? Please fill in.

I am an independent practitioner with two decades of experience in fraud detection and financial crime compliance (KYC/AML) across financial services and financial technology. My current work focuses on governance, fraud, and compliance controls for agent-mediated financial activity. This response reflects my own views. It draws on public sources and general practitioner experience only.

Summary

I support the FSB's direction and the structure of the twelve sound practices. My comments concentrate on agentic AI - autonomous AI agents acting under delegated authority - where the consultation report is strongest in its risk description and can be made more operational in its practice text. Taken together, the recommendations below follow a single control lifecycle: identify the agent and its principal, define its authority, confirm its standing at execution, enforce deterministic limits, preserve evidence, and deactivate cleanly when required. I offer seven recommendations.

1. Add a financial-crime lens to the agentic AI risk discussion (Section 1.3). The report frames agentic risks primarily as operational and cyber risks. Agentic AI also creates identifiable fraud and money-laundering typologies, extensions of well-understood patterns, that belong in the risk taxonomy.

2. Treat agent-mediated transactions, including third-party agents acting on behalf of customers and counterparties, as a distinct risk surface. The report largely treats agents as internal tools. Financial institutions will increasingly face inbound agents acting for external parties, and industry authorization artifacts for such transactions already exist. The guidance should address how institutions verify, screen, and evidence these interactions.

3. Make agent deactivation an explicitly tested control (Sound Practices 10 and 11). Deactivation ("kill switch") capability should carry defined time objectives, revocation propagation across all credentials and delegations, and periodic tested exercises, analogous to business-continuity testing.

4. Add ongoing agent due diligence with point-of-execution standing checks (Sound Practices 5, 9, and 10). Onboarding-time certification of an agent decays quickly. Proportionate to its financial authority, an agent's credentials, authorization artifacts, integrity, and risk status should be re-validated at or near execution, with event-driven reassessment after code changes, behavioral anomalies, authorization changes, or principal-level changes.

5. Specify evidence standards for high-materiality agent actions (Sound Practices 3 and 8). Records of agent authorization and action should be attributable, tamper-evident, time-synchronized, independently verifiable rather than solely self-attested by the agent platform, retained in line with record-keeping obligations, and reconcilable across parties.

6. Elevate agent identity attribution from case-study material to practice text (Sound Practices 2 and 3). Every agent should be bound to a unique identity, an accountable human or organizational principal, and scoped credentials, with authorization scope recorded in the AI inventory.

7. Qualify the AI-monitoring-AI approach (agentic risks discussion and Sound Practice 10). Monitoring agents built on the same or similar foundation models as the monitored agents can share correlated failure modes; monitoring and monitored systems trained on similar data may both fail to recognize the same novel fraud pattern. Independence and diversity between monitored and monitoring systems, plus deterministic guardrails enforced outside the model, should be stated as conditions.

Closing remark:

The consultation report is timely. Institutional experimentation with agentic AI is advancing faster than the control environment around it. Strengthening the sound practices on identity attribution, ongoing agent due diligence, tested deactivation, deterministic boundaries, and evidence standards would give boards, control functions, and supervisors a practical baseline before expectations are defined mainly through losses, disputes, and incident response. I appreciate the opportunity to respond and am available to discuss any part of this submission.