FSB FINANCIAL STABILITY BOARD

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

Response to Consultation

DX Research Group (DXRG)

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?**

We agree with the report's overall account of AI benefits and risks. Three additions would make the framework more useful for transaction-capable agents.

First, **mandate integrity is a distinct operational risk**. An agent does not act from a model alone. It acts from an authenticated mandate assembled with structured controls, current market and portfolio state, memory, tool definitions, and execution policy. Failures can occur when this operating layer gives stale state too much weight, treats prior reasoning as policy, buries a material constraint, or translates an instruction into an incorrectly scoped action. The relevant control objective is to preserve the human principal's authenticated intent across the full path from mandate compilation to a typed action, independent validation, execution, settlement, and subsequent state update.

Second, **correlated behavior can emerge without coordination**. During DX Terminal Pro, 1,544 of 3,454 active vaults bought the same asset within one hour even though agents had no shared chat or agent-to-agent signalling channel. Each agent observed the same evolving market state, and earlier actions changed the state observed by later agents. The bounded setting does not establish a system-wide effect, but it demonstrates a mechanism through which similarly configured agents can generate feedback loops at machine speed. Monitoring should therefore cover population-level concentration, correlated order flow, message rates, and common model or data dependencies, not only the compliance of individual agents.

Third, **auditability is both a benefit and a control obligation**. Software-mediated decisions can preserve a richer record than many manual processes: principal, mandate, state snapshot, model and harness versions, proposed action, policy decision, venue payload, acknowledgement, fill, and portfolio effect. That record allows institutions to identify whether a failure originated in state assembly, model interpretation, policy enforcement, execution, or settlement. Trace collection should be purpose-limited and paired with data minimisation, access controls, retention schedules, integrity protection, and controlled disclosure because mandates, portfolio state, model outputs, and credentials can themselves be sensitive.

The FSB should also encourage institutions to label the environment behind every result. Simulation, replay, paper, shadow, bounded live, and current live evidence answer different questions. Clear evidence labels reduce the risk that a successful backtest, a valid tool call, or reliable settlement is mistaken for profitability, safety, or suitability.

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

The proposed practices are broadly comprehensive. For systems able to change external financial state, the FSB should make the separation between **proposal and authority** explicit. The model may propose an action, but deterministic software outside the model's reasoning path should decide whether the exact venue-ready payload is permitted under current identity, credential, asset, venue, order-type, capital, exposure, loss, price-impact, and data-freshness constraints.

A practical minimum control set for a transaction-capable agent is:

1. unique agent identity linked to a human or institutional principal and a versioned policy;

2. least-privilege tools and credentials, with capital isolated to the approved account or vault;

3. mandate compilation with explicit precedence among authenticated instructions, structured controls, current state, and memory;

4. deterministic pre-action limits and validation of the final payload;

5. payload-bound approval for actions outside standing authority;

6. complete action traces reconciled against venue records;

7. independent monitoring, credential revocation, cancellation capability, and a tested kill switch; and

8. staged recovery that requires reconciliation, regression testing, and explicit re-authorisation.

Testing and reporting should name what was measured: decision cycles, actions proposed, actions rejected, policy-valid submissions, venue acknowledgements, fills, settlement outcomes, and portfolio outcomes. These measurements can sit in one operational account, but the denominator should always be identifiable.

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?**

The report strikes the right general balance. Agentic AI warrants additional treatment when autonomy, tool access, and external-state change combine. Regulation and supervision should remain technology-neutral while becoming more specific about decision rights and enforcement points.

The central question is not whether a system uses a particular model architecture. It is what the system may observe, decide, and change; how much capital and which counterparties it can reach; which controls remain outside its authority; and whether operators can reconstruct and interrupt its actions. Human oversight should also be defined by function rather than by the mere presence of a person. An approval is meaningful only if the approver sees the exact action and relevant state, the approval is authenticated and time-limited, and the approved payload cannot change before submission.

For lower-risk uses, post-action review may be proportionate. For material or irreversible actions, institutions should use pre-action limits, explicit approval thresholds, or fully deterministic rejection rules. An agent should never be able to expand its own permissions through generated text or disable the systems that monitor and constrain it.

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

The practices will remain flexible if they anchor to evidence and outcomes rather than a fixed list of model techniques. Models, scaffolds, tools, and deployment patterns will change quickly. Stable expectations are that an institution can identify the principal and policy behind an action, show the information available at the time, reproduce the control decision, reconcile the resulting transaction, and demonstrate that interruption and recovery mechanisms work.

Material changes to the model, prompt or mandate compiler, memory policy, tool schema, execution code, venue, asset universe, or risk policy should trigger proportionate re-evaluation. Institutions should retain versioned configurations and representative traces and should test changes first in lower-risk environments. Evaluation should include action stability, repeated runs, stale and malformed state, boundary values, unavailable tools, rejected actions, partial fills, and failure of monitoring or reconciliation paths.

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.**

DXRG offers DX Terminal Pro as a nonbank case study of operating-layer controls under real capital. In a 21-day deployment beginning in February 2026, the system operated 3,505 funded vaults, with one agent per participating wallet. Owners supplied natural-language strategies and selected structured behavioral controls. They retained funding, configuration, pause, withdrawal of unallocated funds, closure, and emergency-liquidation controls, while agents selected ordinary buy and sell trades inside a bounded 12-asset market.

The deployment produced approximately 7.5 million agent invocations, 300,000 onchain actions, $20 million in transaction volume, more than 5,000 ETH deployed, and roughly 70 billion inference tokens. The system retained the invocation-level path from user mandate and current state through the model's proposed action, validation, portfolio state, and settlement. Policy-valid submitted transactions recorded 99.9% settlement success.

The operating layer mattered measurably. In a separate EVM transaction-construction evaluation, aligned successful construction rose from 87% with Claude 4 to 96% with Claude 4.6 and to 99.9% when the Terminal Pro-style harness was applied to the stronger model. Controlled pre-launch tests also showed that representation and mandate compilation changed behavior materially: moving an unchanged fee statement earlier in the rendered context increased fee citation from 3% to 74%, while a compound intervention that demoted prior reasoning from precedent to context reduced fabricated sell rules from 57% to 3% in the affected test population.

The production architecture used structured action types and deterministic validation around the model. This made failures attributable: a bad outcome could be traced to mandate compilation, state freshness, model selection, policy validation, execution, settlement, or the user's own stated objective rather than being collapsed into a single model score.

The deployment also exposed population behavior. The one-hour 1,544-vault purchase cascade described above emerged from shared market state without direct agent communication. Across the event, agents frequently took opposite sides of the same asset within the same five-minute window, showing that common model infrastructure did not eliminate behavioral diversity while still allowing local cascades.

This case establishes feasibility, not universal safety or investment performance. It covers one bounded market, one model family in production, one action space, and one 21-day period. Its policy value is that scoped authority, deterministic validation, interruption, reconciliation, and complete traces were implementable at population scale and generated evidence that conventional text-only benchmarks do not capture.

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

The report's case studies are useful, but a common disclosure template would make them more comparable and reusable. Each case should state:

- the exact evaluation subject, including model, harness, tools, memory policy, execution code, venue, dates, asset universe, and starting state;

- decision rights and the boundary between model proposals, deterministic controls, and human approvals;

- the point-in-time information available to the system;

- the action space, abstention behavior, costs, failures, partial fills, and reconciliation method;

- the evidence environment and relevant baselines;

- repeated-run, version-sensitivity, and worst-case results;

- safety controls, incident response, and recovery procedures; and

- limitations, corrections, privacy constraints, and supporting artifacts.

Where possible, institutions should publish or share versioned, machine-readable control matrices and test cases. DXRG has published a twelve-control Agentic Trading Guardrail Matrix, an Agentic Trading Benchmark Card, and a Mandate Compilation Evidence Table. These artifacts do not certify a system; they make control claims inspectable and easier to reproduce.

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

The glossary would benefit from a functional definition of an agentic AI system. We suggest: **a system that operates under a delegated objective, uses current state to select among actions, can change external state through tools or execution systems, and incorporates resulting observations into later decisions**.

The glossary should distinguish the **model** from the **agent system**. The model produces inferences. The agent system includes mandate compilation, state assembly, memory, orchestration, tools, permissions, deterministic controls, monitoring, execution, and recovery. This distinction matters because many material benefits and failures arise in the surrounding system rather than in the model alone, and because the enforceable control points generally sit outside the model.

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

DXRG consents to publication of this response and the organisational contact information supplied in the respondent-information section. Please do not publish personal data beyond that named organisational contact. The deployment evidence is historical and bounded to its stated setting; it does not claim investment performance, suitability, regulated status, universal safety, or transfer to other markets.

Supporting material:

1. Financial Stability Board, *Sound Practices for Responsible Adoption of Artificial Intelligence (AI): Consultation report*, 10 June 2026: https://www.fsb.org/uploads/P100626.pdf

2. Barton, T.J. et al., *Operating-Layer Controls for Onchain Language-Model Agents Under Real Capital*, arXiv:2604.26091: https://arxiv.org/abs/2604.26091

3. DXRG, *DX Terminal Pro*: https://www.dxrg.ai/blogs/dx-terminal-pro

4. DXRG, *Agentic Trading Guardrail Matrix v1*: https://www.dxrg.ai/research/agentic-trading-guardrail-matrix.v1.json

5. DXRG, *Agentic Trading Benchmark Card v1*: https://www.dxrg.ai/research/agentic-trading-benchmark-card.v1.json

6. DXRG, *Trading-Agent Mandate Compilation Evidence Table v1*: https://www.dxrg.ai/research/trading-agent-mandate-compilation-evidence.v1.json

7. DXRG, *The Agentic Trading Evidence Ladder*: https://www.dxrg.ai/blogs/agentic-trading-evidence-ladder