FSB FINANCIAL STABILITY BOARD

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

Response to Consultation

MSP Advisory

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?**

   Yes. The report omits correlated control failure: the financial-stability channel that runs through risk management itself, not through trading behavior.

   The report correctly flags homogeneous behaviours and correlated outcomes where institutions rely on common models for market-facing activity - herding, procyclicality, amplified stress. The same concentration exists on the control side, and the report does not name it. If most institutions validate, monitor, and red-team their foundation-model systems with evaluators and monitoring agents derived from the same small set of foundation models - often the same model family under review - then control failures correlate across the system exactly as trading behaviours do. A blind spot shared by the champion and its challenger goes undetected at every institution at once.

   The report itself supplies the ingredients. SP10 endorses "AI-in-the-loop." Section 1.3 notes that detecting erroneous agent actions "may require augmentation with another AI agent." SP11 contemplates cyber-security AI agents countering agentic attacks. SP12 documents that the supply of foundation models is highly concentrated and in places vertically integrated. Put those observations together and the conclusion writes itself: AI adoption can create not only correlated market behaviour, but correlated failure of the control architecture intended to detect it.

   Supervisory practice has long recognised the institution-level analogue of this failure: a challenge function that is formally independent but substantively captured because it shares the builder's assumptions, incentives, or constraints. Post-SR 11-7 model-risk supervision was designed in part to avoid that failure in human validation functions. The FSB draft risks rebuilding it in silicon, at sector scale, unless automated challengers are required to fail independently from the systems they oversee.

   Recommendation: Add correlated control failure to the Section 1.3 taxonomy as a distinct system-level vulnerability, separate from market correlation, and cross-reference it from SP10, SP11, and SP12.

One further omission. SP12 acknowledges that financial institutions remain accountable for AI outcomes even when the AI was not built in-house. The risk taxonomy and the practices never resolve who, by name, holds that accountability. I address this under Question 2.

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

Comprehensive in coverage; insufficient in bindingness at four points.

Element 1: envelope defined in advance. SP9 states that ante hoc performance thresholds are preferable to deciding after the fact what counted as acceptable - and then concedes this "might be harder for certain types of AI, such as agentic AI." That concedes the principle exactly where autonomy makes it indispensable. In supervision the equivalent is settled: a desk that cannot state its limits in advance is not given a wider limit; it is not permitted to trade.

Recommendation: For high-materiality agentic use cases, make ante hoc definition of the sanctioned envelope - permitted actions, prohibited actions, escalation triggers, hard limits - a condition of deployment. If the envelope cannot be written down, the use case is not ready for that materiality tier and should run at a lower one until it can be. State this as a boundary rule, not a preference.

Element 2: challenge independent of the builder. SP2 states that validation and audit should "remain independent and functionally separate from development teams," with challengers "empowered and sufficiently skilled." That is the organisational half of independence, and for AI it is no longer the whole of it. A validation team that is organisationally separate from the build team, but evaluates a GPT-derived system with a GPT-derived judge, can satisfy organisational independence while still being statistically captured.

Recommendation: Define effective challenge with the three attributes supervisory practice has long used - competence, influence, and incentives independent of the builder - and add a fourth attribute specific to AI: decorrelated challenge, meaning statistical independence of automated challengers from the systems they challenge.

Element 3: residual risk owned by name. SP5 distinguishes inherent from residual risk. SP10 calls for responsibility matrices. Nowhere across the twelve practices does a named individual accept residual risk for a high-materiality use case. Matrices distribute responsibility; they do not concentrate accountability. SP10 states that "financial institutions and individuals retain ultimate accountability" when machines do the monitoring, but does not specify the mechanism by which an individual comes to hold it. The UK's Senior Managers and Certification Regime shows the workable alternative: a named senior manager, a statement of responsibilities, and a signature.

Recommendation: For high-materiality AI use cases, require written acceptance of residual risk by a named accountable executive before deployment and at each material change, recorded in the SP3 inventory.

Element 4: evidence built for a third party. SP3 frames documentation as maintaining the institution's own "visibility over risks." That is inventory management. Documentation

discharges a duty of care only when it is engineered from the outset to be examined by someone with the power to disagree. The test is concrete: contemporaneous, attributable, tamper-evident, and sufficient for an independent examiner to reconstruct the envelope, the challenge, and the risk acceptance without the cooperation of the people who made the decisions.

Recommendation: Reframe SP3's documentation standard around third-party examinability and state that test explicitly for high-materiality use cases.

One clarity defect should be fixed regardless. SP3 should not treat chain-of-thought or model-generated reasoning traces as examiner-grade evidence. Annex 2 correctly notes that an LLM may describe one set of steps to the operator or user while operating differently. The final report should resolve this by treating such traces as unverified self-report: useful for triage and investigation, but not sufficient proof of how the system reached, used, or acted on an output.

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?**

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 imbalance runs one way. The report names agentic risks precisely, then governs them with the weakest tools available. SP9 concedes ante hoc thresholds. SP10 concedes that real-time human monitoring of agents does not scale. Section 1.3 concedes that detection may require another AI agent. Each passage identifies a point where the traditional control model breaks, and each responds by relaxing the control rather than redesigning it.

The redesign is decorrelation as a design requirement for AI overseeing AI. Wherever the report endorses automated oversight - AI-in-the-loop in SP10, agent-based monitoring in Section 1.3, AI-powered defence and counter-agent operations in SP11, automated augmentation of performance assessment in SP9 - it should require the overseeing system to be statistically independent of the overseen system. At minimum: a different base model family; different training-data lineage to the extent feasible; a different supplier where the market allows; and demonstrated disagreement. An overseer that has never disagreed with its champion is not providing challenge; it is providing cover.

This is not a theoretical concern about model behaviour, nor is it a vendor-specific criticism. It is a class-level control issue affecting preference-trained frontier assistants, including systems such as ChatGPT, Claude, Gemini and comparable tools. These systems are optimised not only for correctness, but also for helpfulness, harmlessness, user satisfaction and compliance with behavioural policies. Those objectives are not always aligned. In practice, they can produce unstable behaviour: excessive deference to the user, refusal where a truthful answer is appropriate, confident explanation after the fact, or revision of an answer when challenged.

Published research on sycophancy in language models (Sharma et al., 2023) found that state-of-the-art assistants may produce responses that match user beliefs over truthful ones, and that human preference data can reward convincingly written but incorrect or sycophantic answers. That finding matters for financial supervision because an automated challenger that yields to pressure, optimises for agreement, or rationalises its output after the fact is not an effective challenger. It is a control that can look independent in form while failing independence in substance.

The implication is direct. ChatGPT, Claude and comparable frontier assistants may be valuable components in control workflows, but they should not be treated as self-validating, self-explaining, or automatically independent. Where they are used to test, monitor, validate, or challenge another AI system, the institution should demonstrate behavioural robustness under challenge, resistance to sycophancy, calibrated disagreement, and independence from the model family or training lineage of the system being overseen.

The mapping to the report's own structure is direct. SP2 adopts the three lines of defence. A first-line agent that acts, a second-line agent that challenges, and a third-line assurance layer are three lines only if they can fail independently. Correlated lines are one line drawn three times. The report already contains the principle in embryo. SP9 notes that benchmarking is most helpful when the alternative output has been produced by a model or system that takes a different perspective, such as using different data sources or methodologies. Generalise that sentence from benchmarking to every automated oversight mechanism in the report, and state it as a requirement rather than an observation. That is the single highest-value amendment available in this consultation.

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

The practices are flexible past the point of usefulness. Three features compound: they are expressly non-binding and "not intended to establish an international standard"; they are framed as a menu; and proportionality appears in nearly every section. Each is defensible alone. Together they give any institution, of any size, a documented and defensible path to adopting almost none of the practices for almost any use case. Flexibility on that scale is the absence of a floor.

The international context sharpens this. PRA SS1/23 applies model-risk discipline across model types, including vendor models and AI/ML modelling techniques where they fall within the model-risk framework, with explicit principles for governance, independent validation and mitigants. OSFI's Guideline E-23, effective in 2027, expressly brings AI/ML methods within its enterprise-wide model-risk framework and flags dynamic self-learning and autonomous decision-making as emerging model-risk concerns. The April 2026 revised US interagency model risk guidance expressly excludes generative and agentic AI models from the scope of that guidance, while directing banking organisations to rely on broader risk-management and governance practices for tools, processes, or systems not covered.

Relative to national frameworks that impose clearer model-risk governance, validation, or accountable-executive expectations, the FSB draft is deliberately more permissive: it is non-binding, non-prescriptive, framed as a menu, and subject to proportional application. The

institutions most likely to read the menu literally are in jurisdictions with the least supplementary national guidance, which is precisely where a floor matters most.

Recommendation: Keep proportionality, but invert its structure for a small core. Designate the four accountability elements - ante hoc envelope, independent and decorrelated challenge, named residual-risk acceptance, third-party-grade evidence - as applicable to every high-materiality AI use case, with proportionality governing the intensity of each element, never its applicability. Twelve practices on the menu; four elements on the floor.

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.**

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.

With one exception, the case studies are adoption-success narratives: efficiency gained, fraud reduced, turnaround compressed. The exception is the developmental-testing case in SP9, where a bank tested an ML trading model, found it did not outperform, and declined to deploy. That is the most instructive case in the report, because the control changed the outcome. No other case study shows a challenger disagreeing with a champion and the disagreement escalating, a named executive accepting or declining to accept residual risk, an agent breaching its envelope and an independent monitor catching it, or an institution executing an exit from a third-party AI provider. Responsible adoption is demonstrated by controls that bind, not by deployments that succeed.

I offer the following illustrative case study for inclusion, presented in the report's anonymised format. It is a composite reflecting implementable practice, and shows the four accountability elements and the decorrelation requirement operating together on a GenAI deployment.

Proposed case study: Decorrelated independent challenge of a GenAI system.

A large internationally active bank deployed an LLM-based system to draft credit narratives and regulatory-document summaries for use by a control function. Before deployment, the first line and the accountable executive defined the system's sanctioned envelope in writing: permitted document types, prohibited uses including any customer-facing output, required human review points, and quantitative thresholds for factual-consistency and omission rates beyond which the use case would be suspended.

The second-line validation function built an independent evaluation harness with a deliberate decorrelation requirement: the automated evaluator used a foundation model from a different model family and different vendor than the production system, supplemented by deterministic checks - entity matching and figure reconciliation against source documents - that relied on no LLM at all. The design premise was that an evaluator sharing the champion's base model would share its blind spots, so organisational independence of the validation team was necessary but not sufficient.

The disagreement rate between champion and challenger was itself a monitored metric with two-sided thresholds. Disagreement above the upper threshold triggered sampling and root-cause review by human validators. Disagreement below the lower threshold was treated as a control failure in its own right - evidence that the challenger had stopped challenging - and triggered review of the evaluator, not the champion. In its first period of operation, the decorrelated challenger surfaced a class of omission and inconsistency errors that the production model's own confidence signals and reasoning traces had not flagged, leading to a configuration revision and a tightening of the human-review trigger.

Residual risk after these mitigants was formally accepted in writing by a named senior executive, recorded in the bank's AI inventory, and re-accepted at each material change, including vendor version changes to the underlying foundation model. The full record - envelope, challenge results, disagreement-rate history, signed acceptances - was maintained to a standard sufficient for examination by the bank's supervisor without reliance on the development team.

Three generalisable points. Decorrelation is implementable today with commercially available models, at modest cost relative to validation budgets. The disagreement rate converts effective challenge from an organisational assertion into a measured quantity, including detection of the failure mode where an automated challenger silently degrades into a rubber stamp, the same failure SP10 already checks for in human overseers. And the named-acceptance step took one signature per use case, closing the accountability gap that the report's own SP12 case study identifies and leaves open.

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

For the adoption decision, yes. For the control decision, largely no. The agentic fraud-detection case in Section 1.2 is the clearest example: it describes human-in-the-loop approval of agent-proposed rules - a real control - but never states who owns the residual risk of the rules humans approve, what the agent's sanctioned envelope is, or how the institution would detect the agent proposing rules that are individually plausible and collectively distorting. Restating each case study against a common control template - envelope, challenge, ownership, evidence - would turn illustrations into instruments.

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

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

Strong on AI terminology, incomplete on control terminology. A report whose subject is the governance of AI defines model risk, model drift, and explainability, but not:

- Validation - independent verification that a system performs as intended and is conceptually sound for its use.

- Effective challenge - critical review by parties with competence, influence, and incentives independent of the builder.

- Residual risk acceptance - the formal, attributable decision to operate with the risk remaining after mitigation.

- Decorrelation of oversight - the requirement that an automated overseer's failure modes not be correlated with those of the system it oversees, secured through distinct model families, training-data lineage, suppliers where feasible, and monitored disagreement.

- Sanctioned envelope - the ante hoc specification of permitted and prohibited actions, limits, and escalation triggers for an autonomous system.

- Sycophancy / user-pressure instability - the tendency of an AI assistant to conform to, validate, or revise answers in response to user pressure or implied preference, even where the original answer was more accurate. For high-materiality control uses, this should be evaluated as a failure mode of automated challenge.

Two smaller items: "ground truth" is used in SP9 and Annex 3 but not defined; and the report's distinction between explainability and interpretability, currently a footnote to SP8, belongs in the glossary, because industry documentation conflates the two constantly.

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

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

Section 1.3 shows the FSB understands agentic AI better than most national frameworks now in force. The remaining work is closing the distance between what Section 1.3 describes and what Sound Practices 1 through 12 require. The four elements above are not new ideas. They are fifteen years of model-risk supervision, distilled, applied to a technology that makes each element harder to deliver and more necessary to have. Decorrelation is the one genuinely new obligation this technology creates.

The control question is no longer whether a frontier assistant can produce a plausible critique. It is whether the assistant can sustain disagreement when the user, the business process, or another model pushes it toward agreement. This consultation is the right moment to establish that expectation - before AI-overseeing-AI architectures harden around correlated designs and the sector discovers, simultaneously, what all its challengers missed.