FSB FINANCIAL STABILITY BOARD

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

Response to Consultation

Sarthak Gupta (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?**

   I am responding as a data scientist who works on finance models, with a background in credit risk modeling and quantitative risk research, so my comments come from the model risk and validation side of the house.

   The report's catalogue of benefits and risks is sound, and I agree with it. The one risk I would draw out more explicitly is what I call validation debt. Financial institutions are adopting AI faster than they are building the validation capacity to match it. A traditional model risk function reviews a small number of high-materiality models on a fairly slow cadence. AI adoption multiplies the number of models and systems in production while shrinking the average time from idea to deployment, and the two trends pull in opposite directions. The problem is not that any single model is poorly validated. It is that the validation function itself becomes the bottleneck, and institutions tend to respond in one of two unhealthy ways. They slow adoption down until it defeats the purpose, or they thin out the review each model gets and quietly build up risk that surfaces later, usually all at once.

   I would stress that this is not a capability risk. For most financial use cases, current AI systems are already capable of doing useful work. The gap is one of liability and accountability. When a model's output informs a credit decision, a valuation, or a trade, and that output turns out to be wrong, the institution needs an established chain: who reviewed it, on what basis, and what the escalation path was. Without that built-in chain from the start, the sentence "the model was wrong" becomes far more expensive than it needs to be, both in direct loss and in the time it takes to figure out what actually happened.

   My suggestion is to name this in the final report as a distinct category alongside explainability and data risk. Something like validation capacity risk: the risk that oversight infrastructure fails to scale with the speed of adoption.

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

   The twelve sound practices are comprehensive as a checklist, and I do not see an obvious gap in coverage. Where the report could be clearer is in showing how the practices compose

into a single operating model rather than reading as twelve separate obligations. In particular, materiality assessment, performance management, and human oversight are presented as distinct practices, but in practice, they form a single feedback loop. Materiality determines how much oversight a system receives; performance monitoring produces the signal that oversight acts on; and the oversight decisions that follow, including overrides, should feed back into both the materiality rating and the performance thresholds set next time. As the report is currently organized, a diligent reader could implement all twelve practices as a static checklist and still miss that they are meant to move together as conditions change. A short passage making that loop explicit would help institutions treat the practices as a live system rather than a set of boxes to tick.

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?**

   This is the question I have the most to say on, and I think the report's instincts are right. The mechanism just deserves a more precise name.

   The report correctly recognises that agentic AI requires different forms of oversight than traditional machine learning. The taxonomy in the human oversight section, human-in-the-loop, AI-in-the-loop, human-on-the-loop, human-in-command, kill switch, and contestability, is genuinely useful and more precise than most guidance I have read on this. What I would add is a structural point. The right oversight form for a given system should not be a one-time label applied at deployment. It should be a layered validation architecture with explicit triggers that move a system between oversight forms as conditions change, in the same way engineering teams route production incidents through escalation tiers rather than fix severity at the moment a ticket is opened.

   In practice, that means defining threshold-based triggers, a performance metric breaching a bound, an unusual pattern of human overrides, a stability check failing, a shift in the underlying data distribution, that automatically tighten oversight. A system might move from AI-in-the-loop to human-in-the-loop, or from human-on-the-loop to a temporary kill switch, without needing a fresh governance decision each time. The report already gestures at this where it mentions triggers for transitioning between forms of human oversight based on performance metrics or user requests. I would encourage the final report to lift that from a supporting detail into a named practice of its own, something like dynamic oversight escalation, kept distinct from the static classification in the human oversight practice.

   The reason this matters more for agentic AI than for traditional models is the failure mode. An agent's mistake is not just a wrong number; it is a wrong action, and it can compound across a sequence of tool calls before a human ever sees an output to review. A validation approach built around checking final outputs, which is adequate for most traditional machine learning, undercovers an agent, where the risk lives in the intermediate steps. The report's point about monitoring and documenting the process of AI agents, including the key intermediate steps, is exactly right, and I would make it a first-class requirement rather than an implementation note for any agentic system that can execute transactions or access accounts.

Specifically on generative AI, the report is right to flag the absence of ground truth as a core evaluation problem. The most reliable compensating control I have seen discussed is triangulation against an independent, non-AI baseline wherever one exists, as the explainability example in Annex 2 describes: it runs the machine learning model alongside the previous non-AI model to benchmark the two and, in the report's own words, triangulate. I would suggest generalising that triangulation principle beyond the single example into a named practice for any use case where ground truth is unavailable, since it is one of the few methods that holds up regardless of whether the system underneath is a traditional model, a large language model, or an agent.

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

   The practices are principles-based rather than tied to any specific architecture, which is the right choice given how fast AI system design is moving. The one place I would push for more explicit future-proofing is the third-party AI risk practice. As agentic systems increasingly call other agents, including third-party agents outside the institution's direct control, the chain of accountability gets harder to trace. An institution's own validation architecture is only as strong as its visibility into what a vendor's agent is actually doing, and there is no standardised market practice yet for how vendors disclose agent-to-agent interaction. I would flag this as a specific area for near-term follow-up work, because I expect it to be the fastest-moving of the twelve risk surfaces between now and the next review.

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 distributed through the report, covering machine learning for credit risk, agentic AI for fraud detection, relationship management, and third-party AI risk, among others, are well chosen and map cleanly onto the report's own risk taxonomy, and the explainability examples in Annex 2 complement them.

   If the FSB is looking to add one, I would suggest a case study built specifically around dynamic oversight escalation: a use case that begins under a lighter oversight form, AI-in-the-loop or human-on-the-loop, and is shown transitioning to a stricter form, human-in-the-loop or a kill switch, in response to a defined trigger, with the institution's decision log preserved. That would give readers a concrete template for the escalation mechanism I described under Question 3, which the current case studies illustrate only implicitly.

   On the request for nonbank examples in particular, I would encourage at least one drawn from an asset manager or a fintech lender rather than a bank. Their model governance often runs leaner and less centralised than a large bank's, and the validation-capacity strain I described under Question 1 tends to show up there first. A nonbank case would also make the point that these practices are meant to scale down to institutions without a large dedicated model risk function, not only up to the largest banks.

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

The case studies work well as illustrations of what good practice looks like at a single point in time, but they read more like static snapshots than processes. What would make them more actionable is a short account of what changed and why on each one, showing how the institution's controls evolved after the initial deployment being described. Institutions that try to build their own oversight process, rather than simply benchmark themselves against a fixed end state, tend to learn more from the path than from a snapshot.

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

The glossary definitions of generative AI and agentic AI are clear and consistent with how the terms are used in practice. The one addition I would suggest is a definition for something like a validation trigger or an escalation trigger as a term of art. The concept sits at the centre of the human oversight section and of my answer to Question 3, but the term itself is currently used descriptively rather than defined. Giving it a shared definition would let institutions and supervisors compare oversight architectures across firms using consistent language, much like human-in-the-loop and its variants already provide a shared vocabulary for the forms of oversight.

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

I want to thank the FSB for the consultation and for the openness of the process. This is a rare piece of AI guidance written at the right altitude, specific enough to act on, general enough to survive the next couple of years of change in model architecture, and it resists the temptation to over-fit to today's systems.

If I had to reduce my comments to a single throughline, it is this. Capability is not the binding constraint on the adoption of responsible AI in finance. The binding constraints are how liability is assigned when a model is confidently wrong, and how the validation layer is designed to catch it before capital moves. A model that looks impressive in a demonstration and is unaccountable in production is a liability problem before it is a capability problem, and the strongest parts of this report, the human oversight taxonomy and the emphasis on monitoring intermediate steps, are the ones that take that seriously. My suggestions above are mostly about making that instinct more explicit and more dynamic. I am glad to expand on any of these points if it would help the Secretariat.