FSB FINANCIAL STABILITY BOARD

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

Response to Consultation

Aura Causal Systems

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 broadly agree with the benefits and risks described in the report. There is, however, one risk that I consider the most urgent in investment management today, and the report does not name it directly: the widespread use of correlation-based statistics and machine learning models without any grounding with underlying economic mechanism.

   Nearly all AI deployed in finance today, from credit scoring to trading signals, works the same way. It finds patterns in historical data and projects them forward. These patterns are correlations. The models have no representation of what causes what. They cannot tell whether a pattern reflects a genuine economic mechanism or a coincidence of the sample period. As long as market conditions resemble the training data, this works. When the regime changes, it fails, and it fails across the industry at the same time, because every institution's models share the same blind spot. The report attributes correlated outcomes to common vendors and common models.

   The report touches this twice, listing "distinguishing causality from correlation" as an explainability challenge and noting that models trained on historical data may underperform in new conditions, but treats both as technical side issues rather than the central risk.

   There is an alternative and a solution. Causal modelling, built on the structural causal models and do-calculus represents the mechanisms that generate market outcomes rather than the correlations they leave behind.

   This makes models explainable by construction and testable against regime change. The report should recognize this approach, both as a mitigant to the risk above and as a distinct category of AI adoption.

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

   The sound practices are clear and well organized, but they contain a structural gap, most visible in Sound Practice 8 on explainability.

The sound practice treats limited explainability as a fixed property of high-performing AI, something to be compensated for after the fact. The recommended responses are post-hoc explanation methods, such as feature importance analysis, and compensating controls, such as stricter guardrails, benchmarking, and increased human oversight. Annex 2 illustrates the approach: a bank cannot assess the conceptual soundness of its machine learning lending model, so it adds post-hoc explainers while acknowledging those explainers become unreliable when input features are correlated, which in financial data they almost always are.

This framing accepts a false premise: that explainability must be retrofitted because the model itself is a black box. That is true only for correlation-based models. A model built on explicit causal structure is explainable by construction. The mechanism connecting inputs to outputs is stated in the model's architecture, not reconstructed afterwards by a second model approximating the first. The factors, logic, and processes contributing to outputs, which is exactly how this report defines explainability, are directly readable.

The practical consequence: Sound Practice 8's section titled "Build explainability into model development" currently offers only one option, choosing simpler models and accepting lower performance. It should also recognise structural causal modelling as a method for building explainability in without that trade-off. Similarly, Sound Practice 6 on selection presents performance and explainability as opposing ends of a spectrum. For institutions whose use cases demand both, the sound practices should acknowledge that model architectures exist where explanation is the design rather than the patch.

Without this, the sound practices guide institutions toward managing the symptoms of unexplainable AI rather than asking whether the unexplainability was necessary at all.

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?**

In my view, the balance leans the wrong way. The report devotes its newest and most detailed material to GenAI and agentic AI, and that material is current and useful. But the biggest risks in investment management today do not come from emerging forms of AI. They come from the established ones: the correlation-based statistics and machine learning models that have been embedded in trading, portfolio construction, and risk systems for years.

These models are trusted precisely because they are familiar. They sit inside critical functions at scale, they share a common methodological blind spot to regime change, and they fail simultaneously across institutions when conditions shift. A framework calibrated by novelty will overweight the agentic AI pilot and underweight the legacy model that has been quietly mispricing risk for a decade. Risk should be weighted by exposure and methodology, not by how recent the technology is.

On the emerging forms, the guidance needs a clearer principle for the human role. The sound practices present human oversight as a menu of options, ranging from human-in-the-loop to largely autonomous operation with periodic intervention. What is missing is a stated direction: AI should be human-centric by design. The goal of agentic systems in finance

should not be to remove the need for human input, but to put the human at the centre as the orchestrator of AI agents, reviewing, challenging, and directing their work from the middle of the process rather than auditing it from the end. An architecture where agents propose and a human disposes is not a transitional arrangement to be relaxed as comfort grows. For decisions that move capital or affect customers, it is the destination.

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

   Yes, if the sound practices stay at the level of principle. Technology-specific guidance ages quickly. This report already needed dedicated subsections for GenAI and agentic AI only two years after the FSB's previous assessment, and each new form of AI will require the same unless the framework rests on properties that do not change with the technology.

   Four principles would carry the sound practices through any future generation of AI: transparency, explainability, human-centricity, and causality.

   The sound practices already contain the first three in partial form. The fourth is absent, and its absence is the framework's main exposure to the future. Models trained only on historical correlation will fail under regime change in every technology generation, however novel.

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 useful but share a limitation: every one illustrates correlation-based AI, whether predictive (credit alerts, fraud detection) or generative (credit narratives, preliminary assessments). There is no case study of causal AI, and nonbank coverage is thin. I offer the following from the asset management and trading side, where the need to migrate from correlation-based models to causal modelling is, in my view, the most urgent issue the report could address."

   Case study: Migrating from correlation-based to causal risk modelling in investment management

   A causal AI firm serving investment professionals has addressed this by filtering returns by causal regime before estimating risk: keeping the observations that share a cause and conditioning out the rest. The same structural approach drives its decision-support for bond and commodity markets, where outputs trace to stated economic mechanisms rather than historical patterns.

   The human remains the decision maker; the system explains why.

   This matters for the sound practices because fiduciary duty demands explanations. A manager who buys a stock because it is well run and cheap can explain the decision. A model that buys because of a multidimensional pattern in past data cannot.

   The report should present causal modelling not as one option among many, but as the direction of migration for investment management AI.

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

Partially. Most case studies are success stories: they describe outcomes (fraud losses down 20%, preparation time halved) but not the decision criteria that produced them. An institution reading them learns what others achieved, not how to judge whether its own use case will work or when to stop. Results without decision logic are inspiration, not actionable insight.

The most actionable case study in the report is the exception: the developmental testing example in Section 3.6, where a bank tested an ML trading model against its existing non-AI model, found it did not consistently outperform, found the training data unrepresentative, and postponed deployment. This is the only case study showing an institution deciding against AI on evidence, and it teaches more than the successes, because it demonstrates the discipline the sound practices are meant to instil. More case studies of this kind would help: models rejected at validation, deployments rolled back, agents constrained after testing.

I would also repeat the observation from my response to Question 5: every case study illustrates correlation-based or generative AI. Institutions evaluating causal approaches, which directly address the explainability and regime-change problems the report itself raises, have no reference point in the document.

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

The definitions are clear, but the glossary has a significant omission. Section 1.3 of the report lists "distinguishing causality from correlation" among the explainability challenges of AI adoption. Yet the glossary contains no entry for causality, causal inference, structural causal model, or counterfactual. The report names the problem and provides no vocabulary for it. The terms "conceptual soundness" and "economic mechanism" cannot be operationalised without one.

I propose four additions, grounded in the established literature:

Causal inference: The discipline of determining cause-and-effect relationships from data, as distinct from identifying statistical associations. Built on the formal framework of structural causal models and the do-calculus developed by Judea Pearl.

Structural causal model (SCM): A model that represents the mechanisms generating outcomes as explicit cause-and-effect relationships, rather than learning statistical patterns from historical data. Because the mechanism is stated in the model's structure, SCMs are explainable by construction and can be assessed for conceptual soundness directly.

Counterfactual analysis: Assessing how outcomes would change under interventions or alternative conditions, such as a different policy decision or market shock. Counterfactual questions cannot be answered by correlation-based models, however accurate their predictions, because the answer depends on the causal structure of the system.

Causal AI: AI systems built on structural causal models, designed to answer not only what will happen, but why, and what would happen under intervention. In production use in investment management today.

These additions would give the sound practices the vocabulary to address the causality-versus-correlation challenge the report itself identifies, and would align the glossary with a recognised development in AI: the 2011 Turing Award was given for this body of work, and its adoption in finance is underway.

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

This report arrives at the right moment, and its sound practices are a solid foundation. My responses converge on one message: the most urgent risk in investment management is not the newest AI but the oldest, the correlation-based models embedded in critical functions across the industry, sharing one blind spot to regime change.

The remedy exists. Causal modelling delivers what the report asks of AI: explainable by construction, testable for conceptual soundness, and supportive of human judgement rather than a substitute for it. I encourage the FSB to name it in the final version. AI for finance should be transparent, explainable, human-centric, and causal.