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

Response to Consultation

AI Standards Lab

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 set out below several additional risks, documented in the empirical literature, that the report does not cover or does not capture fully:

1. Algorithmic collusion and correlated errors: concentration as an epistemic risk, not only a supply-chain risk (Section 1.3, Sound Practice 12):

The report identifies "homogeneous behaviours and correlated outcomes" arising from reliance on common AI models, datasets, or infrastructure. We suggest this framing misses two well-established findings, both of which imply that treating concentration purely as a third-party or supply-chain issue is incomplete. First, experimental research suggests that pricing algorithms learning independently, in competition with one another, consistently learn to charge higher-than-competitive prices without any communication between them (Calvano, Calzolari, Denicolò and Pastorello, American Economic Review, 2020, https://www.aeaweb.org/articles?id=10.1257/aer.20190623). In other words, harmful coordinated behaviour can emerge even when institutions use entirely different vendors and share no infrastructure which is not addressed by the report's third-party and concentration controls. Second, research shows that when many decision-makers rely on the same algorithm, the overall quality of their collective decisions can fall even if that algorithm is more accurate than the alternatives for any single institution (Kleinberg and Raghavan, PNAS, 2021, https://doi.org/10.1073/pnas.2018340118). The financial application is direct: if multiple banks assess creditworthiness with a common algorithm, a borrower rejected by one bank is rejected by all. Taken together: diversifying vendors does not necessarily diversify failure modes, and each institution adopting AI in a way that is individually sensible and compliant can still add up to sector-level risks that no institution-level control detects. We suggest the final report acknowledge this distinction and consider whether monitoring for correlated model behaviour across institutions belongs in supervisory or sector-level workstreams rather than institution-level sound practices.

2. Skill atrophy and the manual-fallback assumption (Sound Practices 10 and 12):

Sound Practice 12 recommends that financial institutions retain "the ability to continue providing a given AI-enabled service manually in the event of disruption," while Sound

Practice 10 presumes a stable supply of humans capable of meaningful oversight. There is a risk suggested by recent literature that AI adoption erodes both capabilities. A 2025 multicentre observational study published in The Lancet Gastroenterology & Hepatology found that experienced endoscopists' unassisted adenoma detection rate fell from 28.4% to 22.4% after only three months of routine exposure to AI-assisted detection, described by researchers as the first real-world documentation of a deskilling effect from clinical AI (Budzyń et al., 2025, https://doi.org/10.1016/S2468-1253(25)00133-5). Convergent evidence from knowledge work points to the mechanism: a survey of 319 knowledge workers by Microsoft Research and Carnegie Mellon found that higher confidence in AI was associated with less critical thinking about its outputs, and that generative AI reduced perceived effort across most cognitive activities (Lee et al., CHI 2025, https://doi.org/10.1145/3706598.3713778), while an MIT Media Lab EEG study found that participants who wrote with LLM assistance showed the weakest neural connectivity, and continued to show underengagement when subsequently required to work unassisted (Kosmyna et al., 2025, https://arxiv.org/abs/2506.08872). Together, these findings imply that the manual fallback envisaged in Sound Practice 12 degrades as a direct function of the AI adoption it is meant to backstop, and that the quality of human oversight under Sound Practice 10 cannot be assumed constant over time if this is not taken into account and measures to maintain active efforts from human supervisors. For example, through periodic unassisted performance testing analogous to the ongoing model monitoring in Sound Practice 9, and require that business continuity plans relying on manual fallback demonstrate, not assume, that the fallback capability still exists. Finally, there is a related dimension which is a mid term risk of lack of supervisors. If AI absorbs the repetitive analytical work usually done by Junior consultants through which expertise has traditionally been built, the future supply of senior staff capable of effective oversight and manual continuity is itself at risk.

3. Model collapse and synthetic data contamination (Sound Practice 7):

The report addresses self-reinforcing bias loops and, in Sound Practice 7, treats synthetic data as one data variety subject to institution-level governance, recommending that GenAI-generated data be checked before reuse in training. We believe this framing does not capture the entire literature that shows that indiscriminate use of model-generated content in training causes irreversible defects in the resulting models (Shumailov et al., 2024, https://doi.org/10.1038/s41586-024-07566-y). Two features make this directly material to financial stability. In addition, the risk is not confined to an institution's own synthetic data: AI-generated research, filings, news, and market commentary are increasingly entering the public and vendor data pools on which financial models train and run inference, and an institution cannot filter synthetic content it cannot identify in third-party feeds. Institution-level checks of the kind described in footnote 55 are therefore necessary but structurally insufficient. We note that subsequent research indicates collapse can be mitigated where genuine human-generated data continues to accumulate alongside synthetic content, which points toward practical controls: we suggest Sound Practice 7 be strengthened to address provenance assessment of third-party and vendor data for synthetic contamination.

4. Common-mode failure via simultaneous updates:

The report correctly identifies vendors changing or updating models without notice as a transparency problem. However, it does not address the systemic problem, a flawed model update or a universal jailbreak propagating simultaneously across hundreds of institutions (a CrowdStrike-2024-style event). This is not hypothetical: research has demonstrated universal adversarial suffixes that transfer across model families, including closed-source models (Zou et al., 2023, https://arxiv.org/abs/2307.15043), and the FCA's own post-incident review of CrowdStrike drew sector-wide operational-resilience lessons (FCA, 2024). This is a financial stability scenario that arguably belongs in the Sound Practice 11 scenario library discussion

5. Persuasion and manipulation of customers:

The report's consumer protection coverage addresses bias, mis-selling, and disclosure. The literature on LLM persuasive capability points to an additional risk not captured by those categories: hyper-personalised persuasive techniques, engagement maximising design in trading apps, and AI-driven dark patterns. A preregistered study published in Nature Human Behaviour found that GPT-4 with access to basic sociodemographic data was more persuasive than human opponents. The authors stress that this effect was achieved with remarkably little personal information (gender, age, ethnicity, education, employment, and political affiliation) and an extremely simple prompt (Salvi et al., 2025, https://doi.org/10.1038/s41562-025-02194-6). Financial institutions hold far richer behavioural and transactional data on their customers, making manipulation concerns even larger.

6. Silent information loss in document-intensive workflows (Sound Practice 9):

We believe the performance-management framework in Sound Practice 9 contains a gap relevant to one of the most common GenAI use cases described in the report: document-intensive workflows such as credit file analysis, claims processing, prospectus review, and compliance checks. The dimensions of performance identified in the report, accuracy, range of output, robustness, stability, and sensitivity testing, implicitly assume that the model has processed the full input provided to it. None of them tests whether the model has in fact attended to all of the information supplied. A well documented body of research shows that large language models systematically underweigh information positioned in the middle of long inputs ("lost in the middle"; Liu et al., 2023, https://arxiv.org/abs/2307.03172), and that retrieval and reasoning quality degrade non-linearly as context fills. Critically, this failure mode is silent: the model does not signal that information was neglected, but instead produces a fluent, confident output based on a partial reading. A model summarising a lengthy credit file may overlook a covenant breach mid-document while performing well on every output-level metric the report proposes. Retrieval-augmented generation, which the report presents as a mitigant (Annex 2), introduces analogous failure modes of its own, including retrieval misses and answers drawn from parametric memory that merely appear grounded. We note this belongs to a broader class of risks in which an institution assumes the model processes what it is given and the failure is invisible at output level — fine-tuning-induced degradation of safety guardrails being another example. We therefore suggest that Sound Practice 9 be expanded to include input-coverage or attention-completeness testing for document-intensive use cases

7. Multi-agent emergent behaviour and the limits of pre-deployment testing (Sound Practices 9 and 10):

The report rightly acknowledges that multiple AI agents may interact in unanticipated ways and recommends testing agent-to-agent interactions before and after deployment. We suggest, however, that the sound practices do not yet confront the deeper issue: the testing framework throughout the report assumes that how a system behaves in testing predicts how it will behave in live deployment, and this assumption should not be made, specially for agentic AI. Research on multi-agent systems shows that testing each agent individually does not tell you how they will interact together. Unexpected group behaviours emerge only at scale and in real operating conditions that test environments cannot replicate. In addition, recent empirical work has shown that frontier models can behave differently when they detect they are being evaluated, including covertly pursuing goals while hiding their true capabilities (Meinke et al., 2024, https://arxiv.org/abs/2412.04984) and strategically complying during perceived testing to preserve their behaviour afterwards (Greenblatt et al., 2024, https://arxiv.org/abs/2412.14093) which further weakens what controlled testing can tell us. The financial sector's own history illustrates the stakes, far simpler interacting algorithms produced sudden, unexpected instability in the 2010 flash crash and the 2016 sterling crash. We suggest that the sound practices (i) explicitly acknowledge that sandbox testing of multi-agent systems provides limited assurance; (ii) recommend gradual rollout in live conditions with strict limits on potential damage (transaction caps, restricted scope, automatic circuit breakers) as the primary control rather than an additional layer to pre-deployment testing and; (iii) encourage monitoring at the level of the whole system, not just individual agents, to catch interaction effects.

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

Sound practice 1 entirely fails to ensure financial institutions' responsible AI adoption. By referring to 'financial institution's business model risk appetite and strategy' as the metric that governance and adoption needs to be aligned to, the practice explicitly allows an institution to create as much risk for itself, and for society, as top management finds acceptable. This type of practice has no place in a regulated industry, in an industry that has historically also shown that some institution's risk appetites were far too high, as far as the interests of society in having stable financial markets were concerned.

An appropriate sound practice has to take into account the bounds on allowed risk appetite that are set by applicable market regulators. It would read for example:

Sound Practice 1 (Strategic direction and oversight): The board and senior management align AI adoption and governance with applicable regulatory constraints, and with the financial institution's business model, risk appetite, and strategy.

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?**

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

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.**

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

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

1. Definition of Artificial Intelligence:

The glossary reproduces the first sentence of the revised OECD definition but omits its second: "Different AI systems vary in their levels of autonomy and adaptiveness after deployment." Autonomy and adaptiveness are not peripheral characteristics but key elements of how all AI systems work. The OECD formulation is the most widely reproduced AI definition internationally, and we see no logic in departing from it by removing these two elements. By contrast, the EU AI Act slightly modifies the wording but deliberately maintains all the points of the OECD definition (Article 3(1)). Retaining them provides a base definition for the glossary's use of autonomy in the separate concept of agentic AI and the report's extensive discussion of autonomy-related risks rather than introducing autonomy as a new property that is exclusive to agentic AI, which it is not. We suggest the final report adopt the complete OECD definition.

2. Definition of Agentic AI (Glossary).

The glossary defines agentic AI as systems "designed to autonomously perform complex and extended tasks, often making decisions and taking actions with limited human oversight." We believe this definition has three weaknesses.

First, "limited human oversight" does not distinguish agentic from non-agentic AI: a fraud detection model scoring transactions in real time also operates with limited human oversight, yet is not agentic. The distinguishing features identified in the literature, and in the report's own Section 1.1 (which describes agents as "capable of planning, reasoning, and executing complex high-level goals independently"), are goal directed behaviour and multi-step planning, which the glossary definition omits.

Second, the definition seems to treat autonomy as a binary property. Instead, agentic systems can have, or be given, more or less autonomy. The report's own sound practices actually assume this (Sound Practice 5 assesses "the AI's autonomy" as a graded risk factor, and Sound Practice 10's human-in-command involves "deciding the extent of autonomy"), so the definition should refer to varying or configurable levels of autonomy.

Third, the definition omits the capacity to act on external systems for example, through tools (APIs, databases, ICT systems, other agents), even though this is the feature that generates most of the agentic risks the report identified in Section 1.3 and the controls in Sound Practice 10.

We note that the OECD's recent work on agentic AI supports all three points: it identifies autonomous goal-directed behaviour, multi-step planning, and the capacity to execute action sequences with limited human intervention as the most frequently cited features, describes agents as perceiving and acting with a degree of autonomy to achieve goals, using tools and adapting to changing inputs, and proposes distinguishing systems by level of autonomy, degree of adaptiveness, domain and scale of impact (OECD, 2026). We suggest the final report better aligns the glossary definition with its own Section 1.1 language and the OECD's identified features: goal-directedness and planning, graded autonomy, and tool-mediated interaction with external environments.

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