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

Response to Consultation

EthicAI

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?

EthicAI is a UK-based independent AI assurance provider, assessing how organisations govern, deploy, and manage the risks of AI deployment against a multi-dimensional assessment framework.

Our independent assurance perspective is complementary to the financial institution and vendor input the FSB has gathered. An independent assessor's vantage point is distinctive to vendors and financial institutions - we observe the gap between what organisations attest to and what their evidence supports. Our assurance methodology is directed at closing that gap, making sound practices verifiable rather than aspirational.

EthicAI provides the kind of independent AI assurance that parts of this report touch on. We have framed our recommendations so that they can be implemented and verified by institutions, management, or internal audit functions irrespective of whether external providers are used.

We support the approach the FSB has taken, namely a menu of sound practices, applied proportionately, and explicitly open to revision as the technology evolves. Our comments are additive, aiming to make the practices more testable, and more resilient to the failure modes we find most often in our work.

We also want to acknowledge where the report advances the field. Its treatment of system-level risk - correlated behaviour, herding and procyclicality from common models and infrastructure, and supply chain concentration, together go beyond most published AI governance frameworks. The treatment of agentic systems is similarly advanced.

We broadly agree with Section 1 - the benefits are generally described with appropriate caution, the four vulnerability categories carried over from the FSB's 2024 analysis remain the right systemic anchors, and the agentic risk treatment is exceptionally well developed. Our comments propose six additions to the risk inventory in Section 1.3.

1.1. Synthetic media and information integrity risk to market and funding stability

The report treats deepfakes and generative AI-enabled phishing solely as inbound attack vectors. Disinformation appears once in Section 1.3, but not as a market-integrity or funding-stability channel in its own right. While the glossary defines disinformation, no sound practice explicitly addresses it - three gaps follow from this.

Firstly, outbound content integrity: no practice addresses provenance, watermarking, or AI-generation disclosure for content the institution itself generates (marketing, research, customer communications), even though binding labelling obligations arrive imminently in some jurisdictions (EU AI Act Article 50 applies from 2nd August 2026).

Secondly, coordinated inauthentic behaviour: synthetic media at scale can potentially amplify misinformation, accelerate deposit runs, and drive disorderly market moves. This can be included in section 1.3, with practice-level points - detection and monitoring capability, provenance for institutional content, and inclusion in scenario modelling alongside the CMORG example the report already cites.

Thirdly, the contamination of training pipelines by synthetic content. Sound Practice 7 names synthetic data as a data type and data provenance as a collection stage control, and footnote 55 contemplates checking synthetic content before it is used to train another model, but the underlying risk is not named. AI-generated content enters training and grounding data through two routes - externally, as it saturates the public and third-party sources the institution ingests, and internally, as institutions reuse model outputs in downstream pipelines.

A growing research base suggests that repeated training on synthetic content can reduce data diversity, distort tail distributions, and amplify distributional shifts where synthetic data become significant components of future training corpora. Although carefully curated synthetic data can improve model performance, uncontrolled contamination of training pipelines may degrade robustness and calibration, while weakening the representation of rare or under-represented patterns.

The report itself describes an adjacent loop in which biased outputs generate new data that entrench biased patterns over time. This is particularly relevant in financial applications, where accurate representation of infrequent but consequential events is critical for risk modelling, fraud detection, stress testing, and market surveillance. We recommend that data lineage expectations in Sound Practice 7 ask institutions to identify and track AI-generated content in training and grounding data from both routes.

1.2. Skill erosion and weakening of manual fallback capacity.

Sound Practice 12 assumes exit strategies that can fall back to manual service delivery; nothing in Sound Practices 10 or 12 asks whether the human capacity to do so still exists after sustained reliance on AI. We recommend naming skill erosion - and workflow-level AI dependency more broadly - as a distinct risk in Section 1.3, adjacent to automation bias, with practices that are testable. These include periodic ‘AI-off’ drills for critical workflows, dependency registers, and capability-retention checks.

Dependency can concentrate at the level of the workflow, even where providers are diversified - substitutability of vendors does not guarantee substitutability of the capability underwriting the workflow. Because reliance patterns are common across institutions, with the same providers, copilots, and redesigned workflows, fallback capacity can erode simultaneously across the sector, failing in correlated ways at the moment a concentrated provider fails. At the sector level, skill erosion could compound the concentration vulnerabilities the report already identifies along models and infrastructure, extended to human capacity.

1.3. Consumer impact of AI-driven security/fraud controls

The case study in Section 1.2 extending facial recognition with image-background analysis illustrates a defensive benefit, but is presented without any accompanying guardrails. The case study mentions no demographic performance testing, proportionality assessment, or scope-creep controls.

We recommend adding to the consumer protection and market conduct list in Section 1.3 the harms that AI security tooling can itself create: false positive exclusion of legitimate customers, biased biometric performance, and drift from threat detection to unwarranted surveillance. We propose the corresponding deployment side guardrails in our response to Question 2 (section 2.5.), and return to the same case study in our response to Question 6.

1.4. Sycophancy and related generative AI reliability failures

Section 1.3 names hallucination but omits sycophancy - agreement-seeking outputs that mirror and reinforce the user's framing rather than correcting it. This failure mode is directly relevant to the advisory, research assistant, and relationship manager copilot use cases the report showcases in Section 1.2. An LLM that corroborates rather than challenges a trader or adviser's preconceptions is a risk in and of itself, and it compounds the automation bias the report already identifies. Automation bias and sycophancy together can close a feedback loop, removing the challenge oversight is meant to provide.

1.5 A benefit the report omits - public trust and the associated adoption risk

The report treats responsible adoption as reducing risk, but does not name the converse benefit, that credible governance is itself an enabler of adoption. The 2025 University of Melbourne-KPMG global study (over 48,000 people across 47 countries) found only 46% willing to trust AI, and a clear public mandate for governance, with 70% considering AI regulation necessary.

The adoption risk that follows is absent from Section 1: customer hesitancy towards AI-mediated services, demand for opt-outs, rising complaint volumes, and reputational exposure where an institution's use of AI outstrips its customers' trust in it. The public appetite for better governed AI stresses the importance of demonstrable, tested controls as a route to adoption, not a brake on it.

1.6. The responsible AI expertise bottleneck as an implementation risk

The gap between the growing volume of responsible AI guidance and its application to a specific use case is fundamentally a skills gap. In our assessment experience, the population who can operationalise a framework against a concrete deployment, through selection of the right tests, judgment of the sufficiency of evidence, and challenging a materiality classification, is small.

Smaller institutions may not be able to hire this expertise in house, which strengthens the case for proportionality mechanisms that lighten intensity without diluting coverage (see our response to Question 4). The concentration of scarce expertise in a small number of firms and functions also may have a systemic dimension of its own, parallel to the provider concentration logic the report applies elsewhere - we suggest Section 1.3's implementation challenges name the expertise constraint explicitly.

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

The practices are comprehensive in coverage. Our comments concern verifiability - several practices state the right standard without supplying the means to verify it. We address Sound Practices 5, 8, 9, 10, 11 and 12.

2.1. Sound Practice 10: assurance of oversight

Sound Practice 10 is, in several respects, ahead of current practice. It states that oversight must be meaningful rather than nominal or 'tick-the-box'; it asks institutions to assess whether humans-in-the-loop are actually providing effective oversight, to align incentives so that thorough review is rewarded, and to analyse override patterns for rubber-stamping and pervasive misalignment. These are strong provisions and we would not dilute them.

The gap is that the practice names the standard but does not supply a test to meet it. It doesn't distinguish in an assessable way meaningful oversight from nominal oversight, resting on the assumption that inserting a human into an automated system automatically improves it. A badly specified human role can degrade the performance of both the human and the system while creating the appearance of control.

Consistent with the report's non-prescriptive menu format, we recommend that Sound Practice 10 identifies the conditions that institutions could evidence to demonstrate that oversight is genuine. Five in particular:

1. Role definition before authority. Specify the function of each human-in-the-loop: error correction, accountability provider, introducing friction, resilience; before authority is assigned. This prevents nominal 'warm body' oversight that satisfies a requirement while serving no clear, defined, function.

2. Organisational conditions. Structural independence from human-in-the-loop teams and commercial targets, and protection from adverse consequences for disagreement.

3. Interface design: in critical contexts, reviewers form an independent assessment before the AI recommendation is shown, confidence displays are well calibrated to model performance, override paths are not artificially laborious, and the combined human and AI arrangement is evaluated as a single system.

4. Integrity of classification (links to Sound Practice 5). Independent internal challenge of the materiality and risk classifications before deployment - the incentive to under-tier a use

case in order to reduce oversight is the failure mode Sound Practice 5 does not currently cover.

5. Outcome validated oversight: measure oversight quality against realised outcomes for affected customers and counterparties, not process metrics alone.

These additions converge with, rather than run ahead of, the direction of regulatory travel. MAS's proposed AI Risk Management Guidelines (consultation, November 2025) name over-reliance on automated outputs without meaningful oversight as a distinct risk and require oversight to account for automation bias, and the HKMA's GenAI circular (19th August 2024) expects institutions to retain human control over customer-facing generative AI decisions.

The ICO draft guidance on automated decision making on 31st March 2026 - its first interpretation of the Data (Use and Access) Act 2025 rewrite of UK GDPR Article 22 (now Articles 22A-22D) - sets criteria for when human involvement counts as meaningful: involvement must be active, not tokenistic - the reviewer needs the authority, discretion, and information to change the outcome, not merely endorse it; reviewers must be trained and qualified to understand the system's logic, outputs, limitations, and risks (aligned with Sound Practice 10 as drafted); review must occur before the decision takes effect, on every decision - ad hoc checks do not qualify, since they leave some decisions entirely unchecked (not currently aligned with the report's annex examples); involvement in designing or building the system does not count as involvement in the decision; and how human involvement functioned must be documented per decision.

Therefore, where the 'meaningful involvement' test fails, the decision is treated as solely automated regardless of appearance - which, for credit and similar decisions, triggers the automated-decision-making safeguard obligations (Article 22C: representations, contestation, human intervention, information about the decision). Nominal oversight is therefore not just weak risk management; it also changes the legal character of the decision in certain jurisdictions. Institutions can therefore face two compliant options: make the oversight genuinely meaningful, or accept the decision is solely automated and apply the Article 22C safeguards.

The ICO's accompanying research on recruitment tools found organisations routinely labelling tools as decision support while, in practice, no meaningful human oversight existed, and the tools were nonetheless producing legally significant decisions. This is the exact middle path Sound Practice 10 warns against but its worked examples do not foreclose: the annex illustrates sampled and spot check oversight without distinguishing the contexts in which per-decision review is required - which is the distinction we recommend Sound Practice 10 draws. While the ICO guidance is in draft and scoped to a single jurisdiction, we cite it alongside the in-force EU AI Act Article 14, HKMA circular, and MAS Guidelines as the direction of regulatory travel.

2.2. Sound Practices 8 and 10 - contestability, redress, and disclosure.

The report names contestability, appeal, opt-out channels, and human review, but supplies no mechanisms for implementing these processes. We suggest the practices indicate response time expectations, named accountability with timelines for acting on the findings of appeals, and logging and pattern review of appeals such that systematic issues surface. There is regulatory precedent - revised UK GDPR Article 22C already makes representations, contestation, human intervention, and information about decisions mandatory safeguards for solely automated decisions. Contestability and opt-out also appear in the HKMA GenAI circular (19th August 2024). This is also the safeguard regime an institution adopts if it takes the sampled oversight route described at 2.1, supporting the inclusion of these mechanisms as practice-level treatments.

Separately, Sound Practice 8's statement that appropriate transparency does not necessarily entail disclosing the use of AI to end users is sound as a principle, but is now qualified by binding law in some jurisdictions - EU AI Act Article 50 makes interaction disclosure and content labelling mandatory in scope from 2 August 2026. A cross reference would prevent the draft reading as a lower bar than applicable law.

2.3. Sound Practices 5 and 9 - reversibility and cascades

The report already acknowledges that recovery and rollback can be impossible for some AI, and that multiple AI agents may interact in unanticipated ways. We recommend structuring these acknowledgments into the assessment practices; (a) reversibility classification - classify action types by whether their outcomes can be undone, with differentiated controls (approval gates, staged execution) for irreversible actions; and (b) cascade analysis - for agentic and interconnected deployments, assess how an erroneous action could propagate across connected systems before it is contained. Both can be assessed at inception under Sound Practice 5, and testable under Sound Practice 9.

2.4. Sound Practice 11 - deployment guardrails for AI security tooling

Sound Practice 11 lists AI defensive tools without deployment-side guardrails, and the biometric case study in Section 1.2 demonstrates the resulting gap in practice (see 1.3 above). We recommend complementing the practice with compensatory controls: a proportionality assessment before deployment; demographic performance testing for biometric and identity tools, whose differential failure modes are well established; a documented boundary between threat detection and staff or customer surveillance; and scope-creep audits of the purpose/downstream use of monitoring outputs.

2.5 Sound Practice 9 - bias definition

Sound Practice 9 treats bias as a monolithic concept - 'unacceptable bias' - with examples of statistical skew and discrimination, to be tested for in 'various parts' of the model. In our experience this is where practice most often falls short of intent: technical teams typically operationalise one or two forms of bias (commonly subgroup performance/regional disparity across at most a handful of outcome variables) without addressing different forms of bias, or how different forms of bias interact and compound. Established taxonomies exist - e.g. ISO/IEC TR 24027:2021 (three bias sources across the lifecycle), and NIST Special Publication 1270 on identifying and managing bias in AI (NIST also define three types) - referencing such taxonomies would raise practice standards. We recommend that Sound Practice 9 expects bias testing to be specified by source, lifecycle stage, be proportionate to materiality, and that the glossary defines the term (see 7.3)

2.6 Sound Practice 12 - taking up the report’s invitation on sector specific assurance tools

Sound Practice 12 makes an observation we think is among the most important in the report: current AI transparency and assurance tools may not be tailored to the financial sector, and may not cater to financial institution specific risks and regulatory requirements. We agree with the diagnosis laid out in Sound Practice 12 and respond to the invitation.

A financial sector AI assurance artefact should have four properties:

1. Use-case level, not only management system level. ISO/IEC 42001 certifies the management system, but does not test whether a specific model’s oversight actually functions. Sector assurance should reach the unit at which the report’s practices operate ie the AI use case.

2. Evidence-grounded, not attestation based. Each conclusion should be anchored to evidence, so that a supervisor reading the assurance output can see what was examined, not only what was concluded.

3. Mapping to the twelve practices, so that results can be read directly against the menu and compared across institutions.

4. Graded by maturity, so that the same evidence supports the proportionality mechanism we recommend in our response to Question 4; lighter intensity for smaller institutions and lower materiality use cases, without lighter coverage.

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?**

Broadly, yes. The agentic treatment is particularly strong - memory poisoning, agent identifiers, least privilege, human-in-command, controls on agent-executed transactions, and the framing of agents as synthetic employees are all present. The ‘AI-in-the-loop’ oversight form and the capability-reliability gap framing, to the best of our knowledge, are genuine advances on existing supervisory guidance.

The structural design is also right, with twelve type-agnostic practices, and type-specific guidance layered where the risk profile demands it. We would encourage the report to make this design choice explicit - assessment dimensions stay constant across AI types, while methods and evidence differ by type.

We suggest three refinements.

3.1. Autonomy creep beyond tool-access scope creep

Sound Practice 10 monitors tool-access patterns for scope creep. A residual gap is drift from the original deployment context: the report’s documentation case study describes agent ‘certification’ with a permitted scope recorded in the inventory, and modular agents reused across use cases. Certified scope versus actual use over time warrants its own monitoring practice, otherwise certification becomes a snapshot - a potentially stale label rather than an actual operating boundary.

3.2. A named accountable human per agent action class.

Sound Practice 10 documents responsible staff per lifecycle stage, including through responsibility matrices. For agents, we recommend extending the matrix to the action class level- which classes of autonomous action map to which named accountable role, with escalation criteria, ie a RACI at the level of what the agent can do, not only where it sits in the lifecycle. When an agent acts erroneously at speed, 'who owns this class of action' should be answerable in minutes. This is where a micro practice can serve the macro mandate. Accountability mapping per action class, combined with the graduated kill switch mechanisms Sound Practice 10 already describes, makes those containment controls operable under time pressure before actions propagate.

3.3. Generative AI evaluation

Sound Practice 9's Generative AI discussion would benefit from two concrete expectations: hallucination rates measured against a baseline, with the effectiveness of mitigations (grounding, retrieval augmented generation) evaluated rather than merely deployed, and sycophancy testing for advisory use cases, per 1.4 above. Sound Practice 9 already frames Generative AI performance around coherence, relevance, and the difficulty of assessing output without a ground truth, but stops short of asking institutions to quantify the named failure modes. Naming these two expectations closes this gap and keeps the practice consistent with the threshold-setting the same section recommends for other performance dimensions.

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

Yes - the menu framing helps with durability, and the practices are largely outcomes based rather than technique bound. We propose a graduated maturity progression as a mechanism that would secure that flexibility over time, converting the report's proportionality principle from an exemption from coverage into a pathway.

4.1 A principles-based approach is the right framework for future proofing

The menu is flexible over time because it is principles-based, which we support. The principle layer states what must be true (effective oversight, performance evidenced proportionately to materiality) and is stable across AI types and over time - a new type of AI does not necessarily change whether a principle applies, only how it is satisfied. The next layer is operationalising those principles in practice - much of our response to Question 2 covers suggestions as to how this could be achieved. Operationalising the principles requires some degree of sacrifice of durability, given that the implementations are tied to a specific AI type or use case. However, there is a necessary balance to be struck, where the principles are forward-looking and well defined, yet grounded enough to provide actual guidance to institutions in implementing these principles. Without tying the principles to processes associated with certain types of AI, risks can be missed - a performance management principle applied without generative AI grounding may be operationalised as accuracy testing without surfacing sycophancy. The techniques associated with controls and processes are more volatile, and correctly stay outside of the instrument.

4.2 Maturity progression as a mechanism for proportionality.

The proportionality principle currently reads as binary - smaller or less complex institutions ‘may warrant applications of only relevant sound practices’, meaning some practices are skipped. The alternative offered is an unspecified ‘modification’. From our perspective working directly in AI assurance, a small institution deploying a credit risk assessment model still needs some kind of materiality assessment, human oversight, and performance monitoring. What should scale with size, complexity, and materiality is depth, rather than coverage.

A maturity progression per practice achieves this. Every institution applies every practice, and what deepens as the materiality rises is the rigour and independence of evidence behind it. A materiality assessment at the lowest maturity might be a documented classification held in a use case register, while at a higher maturity the same assessment is independently challenged before deployment and calibrated against realised outcomes. The practice is present throughout the maturity levels; what changes is the evidentiary bar, not whether the practice applies.

Progression through the maturity ladder matches the adoption dynamics the report describes (deploying in non-critical operations first and scaling into critical operations later - the ladder tells institutions what ‘more’ looks like), and, since each level is defined by evidence, maturity is evidenced rather than attested. This mirrors the direction of supervisory travel - MAS’s proposed Guidelines would require higher-risk AI to be reviewed pre-deployment by parties not involved in its development. This is essentially independent assessment attaching to higher materiality use cases, which is exactly the gradient a maturity ladder makes explicit.

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 report’s case studies would be strengthened in two ways. Firstly they cluster at large, internationally active institutions and G-SIBs; how the practices apply proportionately at smaller and digital-native banks is under-illustrated, despite the fact that this is exactly where the report’s proportionality principle does the most work (see our response to Question 4).

Secondly, and more importantly, the case studies are almost uniformly success stories. The developmental testing case study, in which a bank postponed an ML model that failed to outperform its non-AI incumbent, is currently the report’s only negative decision story and is perhaps one of its most instructive, as it shows the practices producing a business decision. Responsible adoption is further clarified by what institutions decide to decline, postpone, or reverse. We recommend the final report adds rollback and failed deployment examples alongside successes, stating not just the decision but the evidence and reasoning that drove it.

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

6.1. State controls and residual risk, not only outcomes and benefits

Each case study should name the controls applied and the residual risks accepted alongside the benefits. The agentic fraud detection case study does this well, including human-in-the-loop approval of all new detection rules, and can serve as the template. The facial recognition case study is the clearest counter-example, presenting only the benefits, with no mention of demographic performance testing, proportionality assessment, or limits on the downstream use of what the system observes - the omissions we identify at 1.3 as consumer risks, and at 2.5 as missing deployment guardrails.

6.2. Consistent structure across case studies, and anchoring to evidence

We suggest a consistent case study structure - context, risk, control, evidence/assurance, and outcome - so that readers can extract the lesson, and so that practices are anchored to the evidence that demonstrate them (override logs, agent action audit trails, consent and lawful basis registers, privacy impact assessments).

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

The glossary is clear and current in what it covers. We propose additions and four consistency points, as well as drafted definitions ready to adopt or adapt.

7.1. Terms the sound practices rely on but the glossary does not define

Described in Section 1.3 but not in the glossary: prompt injection, jailbreaking, and shadow AI (also used in Sound Practice 11).

Defined in footnotes only: red-teaming (footnote 74), automation bias (footnote 26), memory poisoning (footnote 31), and interpretability (footnote 57).

Defined only within Sound Practice body text: human-in-the-loop, human-on-the-loop, human-in-command, and AI-in-the-loop, together with kill switch and contestability (all Sound Practice 10), and inherent and residual risk (Sound Practice 5).

Used but not defined: guardrail and assurance.

Absent: misinformation (only disinformation is defined) and sycophancy.

7.2. Consistency points

Materiality. Criticality is defined but materiality is not, despite materiality being the central assessment concept across Sound Practices 5 to 10. The definition in the body of Sound Practice 5 should be promoted to the glossary.

AI use case. This term is the unit of assessment for half of the sound practices and warrants a definition.

Bias. The report relies on 'unacceptable bias' without defining bias or distinguishing its sources (see 2.6.). A glossary entry, with a pointer to an established lifecycle taxonomy of bias sources, would support it. A proposed definition is at 7.3.

The boundary of 'AI'. The glossary's definition of AI is deliberately broad and, read literally, captures conventional statistical techniques (a linear or logistic regression "infers, from the input it receives, how to generate outputs") that institutions have run under model risk frameworks for decades. This boundary determines which use cases enter AI inventories and become subject to these practices. We suggest the report acknowledge the boundary question and treat scoping as a documented decision under Sound Practices 3 and 5, rather than leaving each institution to draw the line unaided.

7.3. Proposed definitions

Drafted for adoption or adaptation, each aiming for consistency with the report's style and with established usage.

AI-in-the-loop. The integration of AI into the oversight process itself, using AI to augment human monitoring and performance management while humans retain control and ultimate accountability.

AI use case. A defined application of one or more AI models or systems to a specific business purpose; the unit at which materiality and risk assessment, guardrails, and human oversight are applied.

Assurance. A structured, evidence-based evaluation of whether an AI model or system, or the governance surrounding it, meets defined criteria, producing a conclusion on which others (boards, supervisors, customers, counterparties) can rely.

Automation bias. The human tendency to over-trust and under-scrutinise the outputs of automated systems, accepting recommendations without the critical review that the oversight role assumes.

Contestability. Mechanisms enabling affected individuals to challenge, appeal, and seek redress for outputs or decisions influenced by AI, including access to human review and, where feasible, intervention before consequences become irreversible.

Guardrail. A technical or procedural control that constrains the inputs to, behaviour of, or outputs of an AI model or system so that its operation remains within defined boundaries (for example, content filters, prohibited-action lists, human approval checkpoints, and rate limits).

Human-in-command. High-level human oversight of highly autonomous AI in which humans determine the extent of autonomy, set guardrails, and manage the system's overall impact rather than reviewing individual outputs.

Human-in-the-loop. A form of human oversight in which a human reviews and approves each decision or output of an AI model or system before it takes effect.

Human-on-the-loop. A form of human oversight in which the AI operates autonomously while humans monitor its operation and intervene periodically or on defined triggers.

Inherent and residual risk. Inherent risk is the risk posed by an AI use case before risk management and mitigation are applied; residual risk is the risk that remains after they are in place.

Interpretability. The degree to which a human can predict what an AI model or system will do and understand the cause of its outputs; a related but narrower concept than explainability.

Jailbreaking. The deliberate use of prompts or interaction techniques to circumvent an AI model's safety features or guardrails so that it produces outputs its developer or deployer intended to prevent.

Kill switch. A mechanism enabling human intervention to halt or constrain an AI model or system, ranging from complete shutdown to graduated degradation from autonomous to human operation.

Materiality (of an AI use case). The importance of an AI use case to the financial institution's viability, critical operations, and ability to meet key legal and regulatory obligations, including towards customers (drawn from the body of Sound Practice 5).

Memory poisoning. An attack in which malicious data are injected into an AI agent's memory or knowledge base (including retrieval-augmented generation stores) to influence its behaviour over time.

Misinformation. False or misleading information disseminated without deliberate intent to deceive; distinct from disinformation, which is deliberately created and disseminated to deceive or manipulate.

Prompt injection. An attack in which content supplied to an AI model or system - directly by a user, or indirectly through retrieved documents, web pages, or tool outputs - is crafted to override the model's instructions or safeguards and alter its behaviour.

Red-teaming. A structured adversarial testing effort to identify flaws, vulnerabilities, and unsafe behaviours in an AI model or system, typically conducted in a controlled environment by testers independent of the development team.

Shadow AI. The use of AI tools by staff without the financial institution's authorisation, or for unapproved purposes, reducing the institution's visibility of AI-related risks and its ability to monitor and manage them.

Sycophancy. Agreement-seeking outputs that conform to the user's stated or implied view rather than providing an independent or corrective response; distinct from hallucination in that each output may be individually plausible yet systematically biased toward the user's framing.

Unacceptable bias. Systematic error or unjustified differential treatment in an AI model or system, or in its data, that is inconsistent with the institution's risk appetite, its legal and regulatory obligations, or the fair treatment of customers. Bias can enter at multiple points in the lifecycle, and different forms can interact.

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

Two themes cut across our responses and, in our view, would most strengthen the final report.

The first is verifiability. Much of the draft's attention is on the content of the principles, which is sound. What is largely absent is the layer beneath them: the mechanisms, tests, and evidence by which an institution (or its supervisor) could distinguish a practice that is genuinely performed from one that is merely adopted. We are conscious that specifying this layer risks tying durable principles to the faster-moving churn of tools and techniques. Our recommendations (notably on Sound Practices 9 and 10) are therefore framed at the level of the conditions an institution should be able to evidence, rather than the specific methods used to meet them. This keeps the verifiability element as durable as the principles it supports, leaving technique outside the instrument and not disturbing the menu's non-prescriptive character.

The second is harmonisation. Where a sound practice reads as a lower bar than binding law already in force in major jurisdictions, the menu risks anchoring institutions below their existing obligations. We identify specific instances against the EU AI Act and the UK's revised Article 22 regime in our response to Question 2. Cross-references, rather than restatement, would prevent this while keeping the report technology- and jurisdiction-neutral.

We commend the FSB for a report that already advances the field in its treatment of systemic and agentic risk, and we would welcome the opportunity to contribute further as the practices develop.