FSB FINANCIAL STABILITY BOARD

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

## Response to Consultation

ZioSec, Inc.

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 broadly agree with the benefits and risks described. The agentic AI risk taxonomy in Section 1.3 is among the strongest published by an international body to date. We recommend the final report add or sharpen three risks:

- Compromised agent capability supply chains. The report's supply chain discussion focuses on cloud, hardware, foundation models, and data providers. Agentic systems introduce an additional layer: the tools, skills, plugins, connectors, and protocol-based integrations (for example, Model Context Protocol servers) that define what an agent can do. These artifacts are frequently sourced from public registries, updated outside the institution's change management process, and executed with the agent's privileges. A poisoned or trojanized skill file or tool definition is functionally equivalent to malicious code running with the agent's permissions, yet it typically evades both traditional software composition analysis and model-level evaluations. We believe compromised agent capability artifacts will be one of the dominant near-term attack vectors against agentic deployments and merit explicit treatment in Sections 1.3 and 3.8.

- Configuration and version drift as a distinct risk. The report notes that vendors may change underlying models without notice. We would generalize the point: an agent that passed validation at deployment can silently become a different system through model version changes, prompt or guardrail edits, new tool grants, or changes in retrieved data. The risk is not only degraded performance but a divergence between the agent's certified, documented state and its actual runtime behavior. This argues for treating revalidation triggers (any change to model, prompt, tools, or permissions) as a first-class control, not merely periodic re-evaluation.

- Shadow agents, not just shadow AI. Section 1.3 rightly flags shadow AI. With low-code agent builders and embedded agentic features inside SaaS products, employees can now stand up autonomous agents with tool access and credentials, not merely use unsanctioned chat tools. Surveys indicate a substantial share of organizations already have agent deployments that are not authorized or not visible to security and risk functions. Discovery of agents across the estate, including agents embedded in third-party products, is therefore a precondition for every other sound practice and could usefully be stated as such.

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

The twelve practices are comprehensive in scope. We offer two recommendations to improve their operability.

First, elevate continuous adversarial testing from a technique to an expectation for high-autonomy systems. Red-teaming currently appears as a bullet under ongoing monitoring in Sound Practice 9, a testing step in Sound Practice 10, and a testing method in Sound Practice 11. For traditional models, periodic validation against realized outcomes is often adequate. For agentic AI, the report itself observes that assessment must consider not only whether the agent achieved its objective but whether the actions it took were appropriate, that erroneous actions may only manifest in live environments, and that effective monitoring may require augmentation with another AI agent. The logical consequence is that adversarial validation of agentic use cases should be continuous and automated, with cadence tied to materiality and to change events, rather than an annual or ad hoc exercise. We recommend the final report state this expectation directly for high materiality or high risk agentic use cases: validation at inception, revalidation on every material change to the model, prompt, tool set, or permission scope, and standing automated adversarial testing in between, with results logged as evidence.

Second, connect testing evidence to governance and attestation. The practices describe documentation (Sound Practice 3), performance evidence (Sound Practice 9), and third-party assurance (Sound Practice 12) somewhat separately. In mature implementations these form a single evidence chain: discovery of agents and their capabilities, empirical validation of their behavior under adversarial conditions, governance decisions and guardrails derived from those findings, and attestation artifacts that can be shared with boards, supervisors, auditors, counterparties, and insurers. Making this lifecycle linkage explicit would help institutions avoid building parallel, disconnected processes and would directly support the management information needs the report describes for boards under Sound Practice 1.

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?**

The balance is broadly appropriate, and we support the technology-neutral foundation. We would, however, caution against extending one assumption from traditional model risk management to agentic systems: the assumption that a system validated once remains the system that was validated. Traditional models change when the institution retrains them. Agentic systems change when anyone in the supply chain changes anything. The final report could make this asymmetry explicit, for example in Section 3.2, by noting that for agentic use cases the interval between assessments is itself a risk parameter, and that risk assessments should define the change events that automatically trigger reassessment.

We also encourage the FSB to retain and strengthen the guidance in Sound Practice 10 on monitoring intermediate agent steps, tool access patterns, and scope creep. In our testing work, the majority of consequential agent failures are visible first in intermediate behavior (unexpected tool invocations, queries beyond task scope, privilege accumulation) well

before they are visible in final outputs. Behavioral telemetry at the step level is the practical foundation for the kill switches, contestability, and override mechanisms the report describes.

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

   Yes, with one suggestion. Flexibility over time is best preserved when guidance anchors to observable adversary behavior rather than to named technologies. The report already points institutions to OWASP, MITRE ATLAS, and NIST adversarial ML taxonomies. We recommend the final report encourage institutions to adopt a technique-level taxonomy for agentic threats specifically (covering, for example, prompt and tool injection, memory poisoning, capability supply chain compromise, privilege escalation through tool chaining, and inter-agent manipulation) and to map their testing coverage against it. Technique-level mapping has two durable benefits: it remains valid as model architectures change, and it allows the same underlying test evidence to be expressed in the language of multiple frameworks and jurisdictions, reducing duplicative compliance work for internationally active institutions. ZioSec maintains and uses such a taxonomy, the Agentic AI Offensive Security Framework (A2OSF), in production testing, and we would be glad to share our experience with technique-level coverage mapping.

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 weighted toward banks and toward adoption benefits. We suggest two additions, both relevant to nonbanks:

   • Third-party risk management evidence exchange. A case study showing a financial institution requiring standardized, machine-readable adversarial testing evidence from an AI vendor as part of onboarding and continuous monitoring, in place of or alongside questionnaire-based assessment. Risk and TPRM practitioners we work with increasingly view static questionnaires as insufficient for AI-enabled vendors; a worked example of evidence-based assessment would give Sound Practice 12 practical teeth and would be directly relevant to insurers, asset managers, and FMIs as evidence consumers.

   • Insurance underwriting of AI liability. A case study from the insurance sector using empirical agent testing results, including blast radius analysis of what a compromised or malfunctioning agent could reach, as underwriting inputs for AI-related liability coverage. This illustrates a nonbank consuming the same evidence chain for a different purpose and shows how responsible adoption practices create transferable value.

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

   Largely yes. The agentic fraud detection case study and the agent certification discussion in Section 2.3 are particularly actionable. To increase actionability further, we suggest case studies state, where possible, the revalidation cadence and triggers used, how testing

evidence was recorded and surfaced to governance bodies, and what guardrails were derived from findings. These operational details are what implementing teams most need and most rarely find in published guidance.

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

The glossary is clear and well sourced. Several terms used substantively in the body of the report do not appear in it. We recommend adding definitions for: prompt injection; jailbreaking; memory poisoning (including poisoning of retrieval-augmented generation stores); AI red-teaming; guardrails; agent identity (the assignment of unique, manageable identifiers and credentials to AI agents); tool or skill (a capability artifact granted to an agent, including protocol-based integrations); blast radius (the scope of systems, data, and actions reachable by an agent if compromised or malfunctioning); and attestation (portable evidence that a system has been assessed against defined criteria). Defining guardrails would be especially valuable: in our experience the term is used inconsistently across the industry to mean everything from static content filters to inferred, runtime-enforced policy, and supervisory clarity here would improve the quality of vendor disclosures under Sound Practice 12.

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

The report is a substantial and timely contribution. We particularly commend three elements. First, the explicit treatment of agentic AI as a distinct risk category, including unauthorized and erroneous actions, memory poisoning, agent-to-agent interaction effects, and the impracticality of real-time human monitoring at scale. Second, the recognition in Section 2.3 that agent-level inventory attributes, including tool access, guardrails, and certification within defined boundaries, are emerging as governance practice. Third, the candid acknowledgment in Sound Practice 12 that current AI transparency and assurance tools are generally not tailored to the financial sector and do not tell a deploying institution how to integrate a model, system, or agent safely into its own environment.

Our overarching recommendation, developed in the responses above, is that the final report should more clearly distinguish between two complementary forms of assurance: point-in-time, documentation-based assurance (model cards, certifications, vendor disclosures) and continuous, empirical, adversarial validation of the deployed system in its operating context. The report references red-teaming in several places, but treats it as one monitoring technique among many. In our experience, for agentic AI the empirical leg is not optional. Agent behavior is a function of the model version, the system prompt, the tools and data the agent can reach, and the environment it operates in. Any of these can change without the deploying institution's knowledge. Documentation describes what a system was; only continuous adversarial testing establishes what it is today. Industry estimates suggest that the large majority of the deployed agentic attack surface in enterprises remains untested, even as most large enterprises move agents into production, and unauthorized or unrecorded agent deployments are common. The sound practices are well positioned to close this gap.

A note on sector-specific assurance tools:

We strongly support the report's invitation for financial institutions, authorities, and third-party AI providers to collaborate on targeted AI transparency and assurance tools, such as financial sector-specific model or system cards. We would add one design principle: such artifacts should carry empirical, adversarial test results with timestamps and version identifiers, not only descriptive disclosures, so that a consuming institution can verify that the evidence corresponds to the system version it is actually deploying. ZioSec would welcome the opportunity to contribute to any FSB, standard-setting body, or public-private workstream developing these artifacts, and to share anonymized findings from production agent testing across enterprise environments.

We thank the FSB for a rigorous and practical consultation report and would be pleased to discuss any of the points above.