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

Response to Consultation

Brazilian Federation of Banks (FEBRABAN)

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?**

FEBRABAN agrees that the report provides a consistent and balanced overview of the principal benefits and risks associated with the adoption of AI by financial institutions. The report generally reflects developments observed in the Brazilian banking sector, including gains in operational efficiency, stronger fraud prevention, predictive analysis of cyber threats, improvements in the customer experience, and enhanced operational resilience and risk mitigation. It also appropriately recognizes the associated model, data, cyber, third-party, and governance risks.

The experience of Brazilian financial institutions provides practical evidence of these benefits. One reported use case concerned a Client Onboarding Agent, in which AI was used to support operational processes while remaining subject to an assessment covering business risk, information security, data use, explainability, reliability, bias, auditability, traceability, human oversight, performance testing, continuous monitoring, operational contingency arrangements, and formal documentation. In another case, an AI agent operated strictly as a decision-support tool, without autonomy to make critical decisions, while user feedback was used to improve classifications and optimize prompts.

At the same time, the experience of Brazilian institutions suggests that certain risks are becoming increasingly material and warrant greater emphasis in the final report. These include risks specific to generative and agentic AI, such as hallucinations, prompt injection, data exfiltration through prompts, autonomous chains of decisions, and limited traceability of agentic reasoning. AI supply-chain risks, information asymmetry involving proprietary model providers, and cultural risks associated with excessive reliance on automated outputs also warrant further consideration.

These risks are illustrated by reported cases. In the Client Onboarding Agent use case, the assessment led the institution to strengthen code-level safeguards and conduct regular testing against prompt injection. Another institution identified excessive access to data and functionalities arising from unnecessarily broad technical integrations and responded by restricting permissions, strengthening identity and authentication controls, limiting calls to authorized tools, and increasing the traceability of interactions with internal environments. A further case involved a supply-chain attack affecting the LiteLLM tool, which required

threat-hunting measures, the preventive blocking of exposed installations, and enhancements to vulnerability management, asset inventories, and component-update procedures.

These examples suggest that the final report could place greater emphasis on vulnerabilities in AI components and software libraries, excessive technical integrations, unauthorized tools, prompt-based data exposure, the limited substitutability of critical providers, and institutions' capacity to rapidly contain AI-related incidents.

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

FEBRABAN considers the proposed Sound Practices comprehensive and clear. Their principles-based, nonprescriptive, and proportionate nature is one of the proposal's principal strengths, as it provides flexibility to accommodate different business models, levels of complexity, use-case materiality, stages of institutional maturity, and technological developments over time.

Reported use cases demonstrate that Sound Practices can be applied through existing institutional frameworks. For example, one institution incorporated AI-specific considerations into its Third-Party Risk Management process. Before contracting or integrating an external AI tool, the vendor was assessed for AI-specific risks and, where feasible, the solution was tested through proof of concept in a controlled environment using synthetic or minimized data. In one instance, the institution discontinued the procurement before deployment after identifying critical gaps in the solution's security controls.

Another institution established an AI Risk Management function within a multidisciplinary structure comprising Model Risk, the Data Protection Officer, and IT Risk. This function coordinates the identification, assessment, and management of AI-related risks with other governance functions. A separate institution approved a corporate data and AI strategy, established a dedicated governance unit and an AI Center of Excellence, and integrated AI oversight into its broader corporate governance framework, including risk, audit, and statutory audit committees.

Nevertheless, the final report would benefit from further operational guidance. The cases suggest that additional clarity could be provided on how institutions should structure assessments at different stages of the AI lifecycle, including:

- initial screening and materiality classification;
- controlled proof-of-concept testing;
- the decision to deploy, suspend, or discontinue a solution;
- the definition and documentation of mitigating controls;
- the assessment of acceptable residual risk; and
- post-deployment monitoring and reassessment.

Further guidance could also address organizational adoption and culture building, leadership capability development, coordination between business areas and control functions, and the integration of model risk, data governance, privacy, cybersecurity, and third-party risk management. Such additions would make the Sound Practices more operational and actionable, particularly for institutions at different stages of maturity, without changing their principles-based nature.

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?**

Overall, the Sound Practices strike an appropriate balance between managing risks applicable to all forms of AI and recognizing risks associated with emerging and more complex forms of AI. Their technology-neutral and proportionate approach allows institutions to draw on existing frameworks for model risk, data governance, information security, operational resilience, compliance, and third-party risk management, while incorporating AI-specific considerations where necessary.

The reported use cases illustrate this combination. In the Client Onboarding Agent case, the institution applied controls commonly used across model and operational risk frameworks—including data quality, fitness for purpose, bias testing, auditability, traceability, human oversight, performance testing, contingency arrangements, and continuous monitoring—while also introducing safeguards addressing risks more specific to GenAI, such as prompt injection.

Similarly, in a reported agentic-AI case, the solution was limited to a decision-support function and had no autonomy to make critical decisions. Users could correct inaccurate classifications, and the resulting feedback was used to refine criteria and optimize prompts. This demonstrates how traditional principles of accountability, oversight, and model monitoring may be combined with controls specifically designed for agentic and generative systems.

A further use case demonstrates the relevance of this distinction in higher-impact applications. In areas such as credit, AI models do not operate on a fully autonomous basis. Instead, AI outputs are combined with business rules, approval, blocking or prioritization thresholds, and clearly defined manual override mechanisms. After deployment, continuous monitoring is maintained, with alerts triggering human review when material deviations occur.

Nevertheless, the final report would benefit from further elaboration on:

- permissible scopes and levels of AI autonomy;
- autonomous chains of decisions and actions;
- controls over training data and fine-tuning;
- the treatment of prompts as sensitive assets;
- hallucination and prompt-injection safeguards;

- the traceability of agentic reasoning;
- access by AI agents to internal systems and external tools;
- rapid containment and deactivation mechanisms; and
- hybrid forms of human and AI oversight.

The report could also provide examples of triggers for human review and distinguish among human-in-the-loop, human-on-the-loop, and AI-in-the-loop arrangements, including criteria for assessing whether oversight is effective in practice rather than merely formal.

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

Yes. The principles-based, nonprescriptive, and proportionate nature of the Sound Practices provides sufficient flexibility to accommodate different business models, levels of complexity, stages of institutional maturity, and technological developments over time. This flexibility is particularly valuable in a field characterized by rapid technological change.

The reported cases demonstrate how institutions may adapt existing frameworks as technologies evolve. For example, one institution updated its traditional third-party risk management process to incorporate AI-specific considerations. Another developed a technical guide covering the vendor lifecycle from needs identification through post-deployment monitoring, including technical feasibility, information security, data governance, regulatory compliance, solution architecture, proofs of concept, service-level agreements, audit rights, explainability, and performance monitoring.

Institutions have also adapted cybersecurity controls in response to emerging risks. In one case, unnecessarily broad technical integrations led the institution to review the solution's access architecture, restrict permissions, strengthen identity and authentication controls, limit interactions to authorized tools, and improve traceability. In the LiteLLM supply-chain incident, the institution enhanced threat hunting, vulnerability management, asset inventories, and component-update procedures in response to a newly identified risk.

These cases indicate that flexibility should not mean that the Sound Practices remain static. To preserve their adaptability while maintaining effectiveness, FEBRABAN recommends that the FSB consider establishing mechanisms for their periodic review. Such reviews could incorporate lessons from:

- emerging AI architectures and use cases;
- institutional implementation experience;
- AI-related incidents and near misses;
- developments in agent autonomy and agentic reasoning;
- changes in human-oversight arrangements;

- AI model and software supply-chain risks;
- evolving training and fine-tuning practices; and
- new approaches to monitoring, containment, and deactivation.

The FSB could also encourage continued convergence with relevant sectoral and multilateral initiatives to avoid regulatory fragmentation and reduce unnecessary compliance costs for institutions operating across multiple jurisdictions.

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 provide relevant illustrations of how financial institutions can benefit from responsible AI adoption and how the Sound Practices may be implemented in practice. However, they should not be interpreted as exhaustive or representative of the full diversity of business models, risk profiles, use cases, governance arrangements, and levels of AI maturity across the financial sector.

The following cases reported by Brazilian financial institutions could complement the FSB report:

AI-specific third-party assessment

One institution updated its Third-Party Risk Management process to address AI-specific risks. Before contracting or integrating an external AI tool, the vendor was subject to an enhanced assessment. Where feasible, procurement was preceded by a proof of concept in a controlled environment using synthetic or minimized data. Final approval depended on a favorable assessment by the relevant lines of defense and could be subject to specific mitigation measures. In one case, the institution discontinued procurement before deployment after identifying critical gaps in the tool's security controls.

Client Onboarding Agent

An institution assessed a Client Onboarding Agent against business risk, information security, data use, explainability, reliability, data quality, fitness for purpose, bias, auditability, traceability, human oversight, performance testing, continuous monitoring, operational contingency arrangements, and lifecycle documentation. Based on the assessment, the project team strengthened code-level safeguards against prompt injection and introduced regular testing to verify their effectiveness.

AI-supported decision-making with limited autonomy

In higher-impact areas such as credit, one institution reported that AI models do not operate on a fully autonomous basis. AI outputs are combined with business rules, approval, blocking or prioritization thresholds, and manual override mechanisms. Continuous monitoring is maintained after deployment, and alerts trigger human review when material deviations occur.

Agentic AI as a decision-support tool

In another case, an AI agent operated strictly as a decision-support tool and had no autonomy to make critical decisions. Users were encouraged to correct classifications they considered inaccurate. This feedback supported continuous improvement and resulted in refined criteria and optimized prompts.

Cybersecurity and least-privilege controls

An institution identified excessive access to data and functionalities arising from unnecessarily broad technical integrations. It subsequently reviewed the access architecture, restricted permissions according to the solution's scope, strengthened identity and authentication controls, limited calls to authorized tools, and increased the traceability of interactions with internal environments. These measures reduced the attack surface and aligned the architecture more closely with the principle of least privilege.

AI supply-chain incident

An institution reported a supply-chain attack involving the LiteLLM tool. Its response included threat-hunting measures, the preventive blocking of exposed installations, and enhancements to vulnerability management processes, asset inventories, and procedures for updating components.

Vendor-lifecycle governance

One institution developed a technical guide establishing structured criteria for evaluating AI vendors throughout the contractual lifecycle. The guide covers the process from needs identification through post-deployment monitoring and includes technical feasibility, information security, data governance, regulatory compliance, solution architecture, proofs of concept, service-level agreements, audit rights, explainability, and performance monitoring.

These cases are drawn primarily from Brazilian banking institutions. The final report could therefore be further strengthened by including additional cases from nonbanks, particularly insurers, payment institutions, asset managers, financial market infrastructures, and other regulated financial service providers. These examples would help demonstrate how proportionality may be applied across institutions with different business models, systemic relevance, technological capacity, and risk profiles.

Drafting caveat: the attached submission does not contain a sufficiently detailed nonbank case study. Accordingly, FEBRABAN may indicate its support for broader nonbank representation without presenting an unverified case as evidence.

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

Yes. The case studies provide useful and practical insights into responsible AI adoption, particularly by demonstrating how existing governance and risk-management frameworks can be adapted to address AI-specific considerations.

For example, the third-party procurement case illustrates a concrete sequence of controls: enhanced vendor due diligence, AI-specific risk assessment, controlled proof-of-concept testing using synthetic or minimized data, review by the relevant lines of defense, definition of mitigation measures, and the possibility of discontinuing procurement where critical security gaps are identified.

The Client Onboarding Agent case demonstrates how institutions may evaluate a solution across multiple risk dimensions before deployment and translate the assessment into technical measures. In that case, the assessment led directly to stronger code-level safeguards and recurring testing against prompt injection.

The higher-impact decision-making case provides actionable insight into proportional human oversight. It combines AI outputs with business rules, approval or blocking thresholds, prioritization criteria, manual override mechanisms, continuous monitoring, and alerts that trigger human review when material deviations occur.

The cybersecurity case illustrates how the principle of least privilege may be applied to AI solutions by restricting permissions, strengthening identity and authentication controls, limiting interactions to authorized tools, and improving the traceability of interactions with internal environments. The LiteLLM incident also provides practical lessons on threat hunting, preventive containment, asset inventories, vulnerability management, and software-component updates.

However, the actionability of the report's case studies could be further enhanced by explicitly identifying:

- the risk or problem identified;
- the stage of the AI lifecycle at which it was identified;
- the materiality or risk criteria applied;
- the governance functions involved;
- the mitigating controls adopted;
- the criteria used to approve, suspend, or discontinue the solution;
- the monitoring and escalation mechanisms established; and
- the lessons that may be transferable to other institutions.

Additional examples could also address acceptable residual risk, triggers for human intervention, documentation requirements, post-deployment reassessment, operational continuity, withdrawal plans, and the substitutability of critical AI providers. These additions would make the cases easier to translate into implementation pathways while preserving the report's nonprescriptive and proportionate approach.

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

The definitions in the glossary are generally clear and broadly aligned with current industry terminology and responsible-AI practices. However, given the rapid evolution of AI, the glossary would benefit from periodic review and further clarification of concepts associated with emerging technologies, particularly agentic AI, scopes and levels of autonomy, agentic reasoning, AI supply chains, prompts as sensitive assets, and different forms of human oversight.

The reported use cases demonstrate why these distinctions matter in practice. For example, the Client Onboarding Agent required controls addressing traceability, human oversight, output reliability, prompt injection, and continuous monitoring. In another agentic-AI case, the solution was expressly limited to decision support and was not authorized to make critical decisions. In higher-impact applications such as credit, institutions combined AI outputs with business rules, thresholds, manual override mechanisms, and alerts triggering human review.

Accordingly, the glossary could clarify how concepts such as autonomy, decision support, automated decision-making, human intervention, monitoring, validation, and override relate to one another. It could also distinguish among the following oversight arrangements:

- Human-in-the-loop: human participation is embedded directly into AI operations, requiring human review, validation, or approval at critical stages before actions are finalized.

- Human-on-the-loop: the AI system operates autonomously but remains subject to human monitoring and intervention.

- AI-in-the-loop: AI systems support the monitoring, validation, or oversight of other automated processes, while accountability remains assigned to human decision-makers.

The glossary should emphasize that these approaches are not interchangeable and that the appropriate oversight arrangement should be calibrated to the materiality, risk, autonomy, complexity, and criticality of each use case. It should also make clear that the use of AI to support monitoring or oversight does not transfer ultimate accountability away from human decision-makers.

The glossary could further define or clarify concepts evidenced by the reported cases, including:

- agentic AI and agentic reasoning;

- levels or scopes of autonomy;

- prompt injection;

- prompts as sensitive assets;

- AI supply-chain risk;

- AI-supported oversight;

- rapid containment or deactivation; and

- traceability of interactions between AI systems, internal environments, and external tools.

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

The responses provided herein are non-exhaustive and are intended to highlight selected considerations and illustrative use cases. FEBRABAN's complete submission, containing its detailed assessment and supporting evidence, has been submitted separately by email.

Ultimately, FEBRABAN stands ready to further discuss any of the matters addressed in this submission, provide additional case studies from its associated institutions, and participate in future discussions on the responsible adoption of AI in the global financial system. FEBRABAN looks forward to continuing its engagement with the Financial Stability Board in support of convergent, proportionate, risk-based practices that promote financial stability while accommodating the financial sector's continued technological development.