



Emerging Risks and Opportunities of Generative AI for Banks

A Singapore Perspective



Emerging Risks and Opportunities of Generative AI for Bank

A Singapore Perspective

List of MindForge Consortium Members





Executive Summary

Inaugurated in 2019, the Veritas initiative is a collaboration involving key players from both the financial and technology sectors in Singapore. Led by the Monetary Authority of Singapore (MAS), the goal of the Veritas initiative is to develop approaches with financial institutions (FIs) in adopting MAS' Fairness, Ethics, Accountability and Transparency (FEAT) Principles in the use of Artificial Intelligence and Data Analytics (AIDA). The Veritas initiative has led to specific and usable methodologies and toolkits for FIs to implement the FEAT Principles.

Project MindForge is a collaboration among financial industry participants, including MAS, Citi, DBS, HSBC, OCBC, Standard Chartered, and UOB, and technology partners Accenture, Google and Microsoft. The project builds on the work of the Veritas initiative to examine the impact and potential risks of generative artificial intelligence (AI) technology on financial services. The project first aims to develop an industry-led whitepaper setting out a private sector perspective for the responsible use of generative AI. A consortium was established to explore experiments to illustrate how the proposals put forward in the whitepaper can be applied in actual use cases.

The release of accessible generative AI products and services since 2022 has radically transformed the AIDA landscape. The advancement of generative AI has also opened up new commercial, social and technological opportunities. However, this advancement is clearly double-edged. The whitepaper aims to examine specific risks posed by generative AI systems that go beyond those of "traditional" AI and how such risks have extended beyond the scope of the current FEAT Principles, first published in 2018.

Generative AI includes diverse techniques for creating content, spanning text, images and other audio-visual elements. It is driven on large machine learning models known as foundation models (FMs), with a subset called large language models (LLMs) trained on trillions of words for various natural-language tasks. The adoption of generative AI across industries, including the banking sector, contains significant potential to improve customer satisfaction, enhance employee experience while augmenting their productivity, reduce costs, enhance decision-making and mitigate risks. Specifically, within the banking industry, Accenture's research and analysis¹ using US labour data found that many day to day tasks and levels of working hours have high potential to be impacted through deployment of generative AI solutions.



While generative AI presents numerous opportunities for innovation in financial services, the associated risks must also be incorporated into frameworks for the responsible use of AI. The whitepaper will evaluate these risks at different stages of the lifecycle of a generative AI system, and map several major risks across seven dimensions of risk: Accountability and Governance, Monitoring and Stability, Transparency and Explainability, Fairness and Bias, Legal and Regulatory, Ethics and Impact, and Cyber and Data Security. This is summarised in Table 1.1.

¹ <https://www.accenture.com/content/dam/accenture/final/accenture-com/document/Accenture-A-New-Era-of-Generative-AI-for-Everyone.pdf>

Table 1.1: Mapping of select risk dimensions relevant to generative AI

Risk Dimensions for generative AI	Select Major Risks Specific to each Dimension
 <p>FAIRNESS AND BIAS</p> <p>Setting fairness objectives to help identify and address unintentional bias and discrimination.</p>	<ul style="list-style-type: none"> • Unrepresentative, under-representative or biased data inputs, especially data sourced from the Internet for foundation models. • Adverse or inappropriate impact on individuals and groups.
 <p>ETHICS AND IMPACT</p> <p>Ensuring responsible and ethical outcomes in the use of AI against clearly defined core values and practices.</p>	<ul style="list-style-type: none"> • Value misalignment. • Environmental sustainability impact. • Dark patterns, deceiving or manipulating users into certain behaviours. • Toxic and offensive outputs.
 <p>ACCOUNTABILITY AND GOVERNANCE</p> <p>Enabling accountability and governance for the outcomes and impact of data and AI systems.</p>	<ul style="list-style-type: none"> • Lack of awareness of generative AI risks. • Unclear or unenforceable third-party accountability. • Lack of use case and model governance. • Inadequate human oversight.



Risk Dimensions for generative AI	Select Major Risks Specific to each Dimension
 <p>TRANSPARENCY AND EXPLAINABILITY</p> <p>Enabling human awareness, explainability, interpretability, and auditability of data and AI systems.</p>	<ul style="list-style-type: none">• Unclear output accuracy level.• Unclear origin of training or test data, leading to potential ingestion of low-quality data.• Lack of explainability.• Anthropomorphism, deceiving or misleading users.• Inadequate feedback and recourse mechanisms.
 <p>LEGAL AND REGULATORY</p> <p>Identifying any legal or regulatory obligations that need to be met or may be breached by the use of AI, including issues with compliance, data protection and privacy rules, or related to equality laws.</p>	<ul style="list-style-type: none">• Data sovereignty: Inability to ensure location compliance for model hosting as well as data access and processing.• Unclear data ownership.• Unauthorised data transfer and storage.• Breach or misalignment to regulatory or organisational standards.• IP infringement.• Unavailability of IP protection.• Inadequate privacy protection.• Record keeping: Inability to appropriately retain or delete data associated with training and use of generative AI systems, in line with applicable regulations.

Risk Dimensions for generative AI	Select Major Risks Specific to each Dimension
<div data-bbox="229 539 416 725" data-label="Image"> </div> <p data-bbox="469 555 826 584">MONITORING AND STABILITY</p> <p data-bbox="469 611 879 712">Ensuring robustness and operational stability of the model or service and its infrastructure.</p>	<ul data-bbox="975 539 1334 1485" style="list-style-type: none"> • Hallucination / Fabrication / False memories, leading to inaccurate or misleading outputs. • Overconfidence, leading to misinterpretation of outputs. • Training data or inputs not fit for intended purpose. • Lack of monitoring. • Insufficient data quality. • Model staleness, causing untimely outputs. • Insufficient model accuracy or soundness. • Model degradation, leading to undesirable behaviours. • Inadequate operational resilience. • Unmet architectural requirements, limiting robustness and leading to inadequate governance.
<div data-bbox="229 1563 416 1749" data-label="Image"> </div> <p data-bbox="469 1579 810 1608">CYBER AND DATA SECURITY</p> <p data-bbox="469 1635 887 1809">Protecting data and AI systems from cyberattack, unauthorised access, data loss, and misuse or adversarial model manipulation by malicious actors.</p>	<ul data-bbox="975 1563 1326 2040" style="list-style-type: none"> • Inappropriate or illegal use. • Data poisoning, leading to malicious outputs. • Adversarial model manipulation. • Re-identification of personally identifiable data. • Data leakages. • Model inference attacks, revealing sensitive information.



Jurisdictions have used different combinations of guidance and regulation to mitigate the risks of generative AI. Relevant documents for FIs in Singapore include the FEAT Principles², Veritas Methodology³, The Association of Banks in Singapore (ABS) Cloud Computing Implementation Guide⁴, MAS Guidelines on Technology Risk Management (2021)⁵, and MAS Guidelines on Outsourcing⁶. These references continue to apply to generative AI use cases, but may need updating to better reflect the new and amplified risks arising from the use of this technology.

For example, FIs face growing challenges to meet the “Fairness” FEAT principles because of the growing difficulty of identifying and responding to bias or prejudice in the inputs and outputs characteristic of generative AI systems. The use of generative AI also challenges the “Ethics” principles given its potential to create content that contravenes the values of an organisation or the norms and laws of society. The nature of generative AI technology, which is expected to encourage a greater dependence on a growing number of third-party providers of large foundation models, adds complexity to implementing the existing principles of “Accountability”. “Transparency” is more difficult to achieve given the large volumes of unstructured, openly sourced data that is used to develop the models.

Generative AI presents new governance challenges, such as its potential to create content that violates the intellectual property rights of third parties, its vulnerability to new kinds of attacks, and its need for increased monitoring for unexpected behaviour. These challenges lead us to conclude that while the FEAT Principles are enduring and have broad application to generative AI, we should consider specific augmentations and extensions to the principles to address the specific implementation challenges faced by FIs in using generative AI. In the same vein, the Veritas toolkit may require updates to accommodate new design criteria and in the evaluation of generative AI systems for conformity with the FEAT Principles.

In addition to core governance considerations, there are a number of new technological considerations posed by generative AI systems. The whitepaper will summarise the key decisions about architecture and infrastructure that an FI needs to consider in the adoption of generative AI. Enterprise-level IT capabilities must be sufficiently robust across seven dimensions (as shown in Figure 1.1) of technology consideration. A continuous feedback loop of improvement throughout the lifecycle of a system is key to long-term success. These dimensions are:

1. **Foundation Model & Infrastructure:** Foundation model selection, accessibility and model hosting infrastructure.
2. **Data Architecture:** Appropriately managing data and providing the foundation model with data access.

² <https://www.mas.gov.sg/~media/MAS/News and Publications/Monographs and Information Papers/FEAT Principles Final.pdf>

³ <https://www.mas.gov.sg/~media/MAS/News and Publications/Monographs and Information Papers/FEAT Principles Final.pdf>

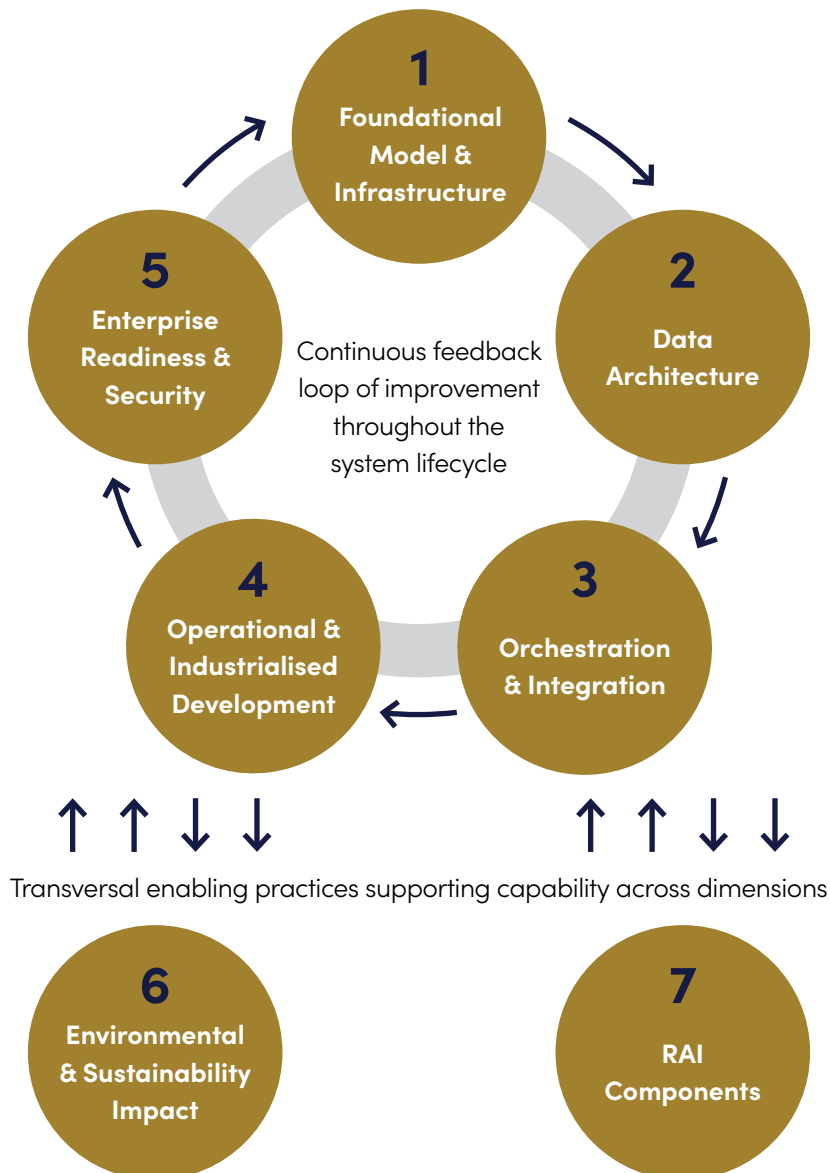
⁴ <https://abs.org.sg/docs/library/abs-cloud-computing-implementation-guide.pdf>

⁵ <https://www.mas.gov.sg/regulation/guidelines/technology-risk-management-guidelines>

⁶ <https://www.mas.gov.sg/regulation/guidelines/guidelines-on-outsourcing>

3. **Orchestration & Integration:** Connecting the model to existing enterprise systems.
4. **Operations & Industrialised Development:** Operating generative AI systems at scale through streamlined development, deployment management, continuous monitoring, and analysis and improvement.
5. **Enterprise Readiness & Security:** Standards on scalability, security and compliance.
6. **Environmental & Sustainability Impact:** Considering the environmental impact of generative AI adoption.
7. **RAI Components:** Adopting responsible AI practices across the enterprise.

Figure 1.1: Seven dimensions of generative AI technology consideration for enterprise-wide adoption





The whitepaper will present a platform-agnostic generative AI reference architecture for FIs, underpinned by these seven dimensions. The reference architecture highlights the importance of contextualising generative AI as a component within a larger technology system. Generative AI is underpinned by high-performance computing and effective data supply chains, and is able to interface with its users through a set of end user applications and, sometimes, existing technology infrastructure. A crucial component of this architecture is the inclusion, at every step, of effective security, meaningful guardrails against unwanted system behaviour, and some degree of human oversight to backstop technical safety measures. Continuous monitoring and evaluation of system performance, and measurement of enterprise value enabled by the system, help ensure that it is safely serving the purpose it was designed for.

Illustrative use cases can help the industry to better understand the impact of generative AI on cybersecurity, sustainability, business, society and other human factors. Within the whitepaper, an illustrative use case will be mapped against the industry risk framework to identify relevant use case-level risks. These risks are then assessed using the current Veritas Methodology, to highlight risks which are not adequately covered in the current framework. The detailed outcomes, which will be highlighted in the full publication, will assess the risk impact of generative AI-powered applications. With considerations to the issues covered in the whitepaper and relevant enhancements to governance, we believe generative AI can be responsibly employed.

The full whitepaper is expected to be published in early 2024.

Disclaimer

This content is provided for general information purposes and is not intended to be used in place of consultation with our professional advisors. This document may refer to marks owned by third parties. All such third-party marks are the property of their respective owners. No sponsorship, endorsement or approval of this content by the owners of such marks is intended, expressed or implied.