



Emerging Risks and Opportunities of Generative AI for Banks

A Singapore Perspective



Emerging Risks and Opportunities of Generative AI for Banks

A Singapore Perspective

List of MindForge Consortium Members



Table of Contents

Foreword	02	D: Challenges for Banks Operating Across Multiple Jurisdictions	89
Executive Summary	03	E: Architecture and Infrastructure	91
1 Introduction	05	E.1 Generative AI Deployment and Adoption Approach	91
1.1 Background of Project MindForge	05	E.2 Key Considerations in Setting Up Private Infrastructure	94
1.2 Navigating the Paper	06	E.3 Seven Dimensions of Generative AI Considerations	96
2 Opportunities and Risks of Generative AI	07	E.3.1 Foundation Model and Infrastructure	97
2.1 Generative AI Overview and Opportunities	07	E.3.2 Data Architecture	99
2.2 Risk Framework	09	E.3.3 Orchestration and Integration	100
2.3 Regulatory Landscape	15	E.3.4 Operations and Industrialised Development	101
3 Risk Assessment of FEAT Principles and Veritas Methodology	16	E.3.5 Enterprise Readiness and Security	102
3.1 Analysis Approach	16	E.3.6 Environmental and Sustainability Impact	103
3.2 Fairness-Related Assessment	19	E.3.7 Responsible AI Components	104
3.3 Ethics and Accountability-Related Assessment	21	E.4 Key Aspects of Data Architecture with Generative AI	106
3.4 Transparency-Related Assessment	24	E.5 Sample of Platform-Agnostic Reference Architecture for Generative AI	107
3.5 Gaps Beyond FEAT	26	E.6 Key Measurements to Monitor, Evaluate and Analyse Technology Stacks Including Generative AI	109
3.6 High-Level Mitigation Approach to Generative AI-Related Risks	28	Bibliography	110
3.7 Evaluation of Current Cloud Implementation, TRM and Outsourcing Guidelines	29	Acknowledgements	116
4 Use Case Implication	32		
4.1 Need for Industry Use Cases	32		
4.2 Description of Industry Use Cases	32		
5 Next Steps	43		
Appendix A: Glossary	44		
Practitioner Section	46		
B.1: Risk Definitions	46		
B.2: Implications of Select Risks to FEAT	61		
B.3: Risk Assessment with Two Additional Sample Use Cases	73		
C: Risk Assessment of the Veritas Methodology	76		



Foreword

Generative artificial intelligence (AI) presents significant potential for the banking sector, promising transformative advantages across the various banking functions. However, alongside the plethora of opportunities, it is crucial to recognise the heightened risks as well as ethical and legal issues introduced by Generative AI which demand meticulous scrutiny and thoughtful deliberation. Such analysis would enable Generative AI to be used in a safe, robust, and responsible manner by all financial institutions (FIs).

The MindForge consortium formed in 2023 comprises of Monetary Authority of Singapore (MAS), Accenture, financial institutions and technology partners to understand the risks and opportunities of Generative AI technology specifically for financial services industry. The MindForge consortium developed this whitepaper setting out a private sector perspective for the responsible use of Generative AI. The consortium also experimented with potential industry use cases and will conduct further work to explore their value and viability.

The whitepaper provides a detailed but non-exhaustive list of Generative AI risks identified by the consortium and includes the risk framework assessing the existing Fairness, Ethics, Accountability and Transparency (FEAT) principles and Veritas methodology as well as suggested areas for extension of FEAT and architecture considerations for the deployment. The whitepaper also highlights the global regulatory landscape which would help in providing advice to the FIs for responsible adoption of the technology.

As Generative AI technology is fast evolving, the industry would need to be mindful that there will be new areas of opportunities as well as risks rising from this. This phase of work focused on the banking industry and for the next phase, the consortium will shift its focus to expanding the coverage to other industries within financial services sector. The consortium will carry out detailed study to provide mitigations and guardrails to systematically mitigate the risks posed by the deployment of Generative AI in the financial services sector. The consortium will also work on exploring additional use cases with potential for developing into industry pilots to benefit the industry. Through these efforts, we aim to advise the industry in the responsible adoption of this technology.

We would like to express our appreciation to all the members of the consortium for their active participation and generous support in the development of the white paper and industry pilots. We would also like to acknowledge the contributions by our industry partners – Accenture, Citi, DBS, Google, HSBC, Microsoft, OCBC, Standard Chartered, The Association of Banks in Singapore and UOB – in this remarkable endeavour.

Sopenendu Mohanty

Chief Fintech Officer, Monetary Authority of Singapore

Executive Summary

Project MindForge is a collaboration among financial industry participants, including the Monetary Authority of Singapore (MAS), Citi, DBS, HSBC, OCBC, Standard Chartered, The Association of Banks in Singapore (ABS) 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) on financial services. The MindForge consortium developed this whitepaper setting out a private sector perspective for the responsible use of Generative AI. The consortium also experimented with potential industry use cases and will conduct further work to explore their value and viability.

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, has a significant potential to improve customer satisfaction, enhance employee experience while augmenting their productivity, reduce costs, enhance decision-making, and mitigate risks. This paper draws primarily on consortium members' experience with language-based Generative AI systems (supported by LLMs), the earliest forms of Generative AI to gain widespread adoption among financial institutions (FIs).

The advancement of Generative AI has opened up new commercial, social, and technological opportunities. However, this advancement is clearly double-edged. The whitepaper aims to examine risks posed by Generative AI systems that go beyond those of predictive, "traditional" AI and how such risks extend beyond the scope of current Fairness, Ethics, Accountability and Transparency (FEAT) Principles, published in 2018.

This paper enumerates these risks across seven dimensions: Fairness and Bias, Ethics and Impact, Accountability and Governance, Transparency and Explainability, Legal and Regulatory, Monitoring and Stability, and Cyber and Data Security.

The technological considerations for adopting Generative AI effectively, securely, and responsibly are also crucial. To support this goal, the paper introduces a platform-agnostic reference architecture. It highlights principal components and underlines the significance of guardrails, continuous monitoring, and human involvement throughout the development and deployment lifecycle.

Developing industry use cases can help the industry better understand this technology's impact. Use cases are intended to provide examples on how risk assessment can be conducted for Generative AI solutions. As the industry's use of Generative AI solutions evolves, these solutions can better position FIs to thrive in a rapidly changing environment.



Generative AI is a relatively new technology, of which its full range of risks and impacts are not fully known, controllable or capable of being deployed responsibly. But it is a promising technology that, when used correctly, can improve commercial, social, and governance outcomes for FIs around the world.

1 Introduction

1.1. Background of Project MindForge

In the realm of artificial intelligence (AI), few advancements have captured the imagination of industry, policymakers, researchers and the public as the emergence of Generative AI. The journey of Generative AI can be traced back to the early days of AI research, where pioneers laid the groundwork for intelligent machines capable of generating original content. However, it was the release of ChatGPT in November 2022 that propelled Generative AI to the forefront, marking a new era by capturing global attention and sparking a wave of creativity rarely seen before.

Generative AI refers to the use of AI to create new content, such as text, images, music, audio, and videos, through machine learning models that have been pre-trained on vast amounts of data. These models are commonly referred to as foundation models (FMs). A subset of FMs called large language models (LLMs) are trained on trillions of words to perform a plethora of natural language tasks. These LLMs can learn, generate text, engage in interactive conversations, provide responses and recommendations, answer questions, and summarise content in human-like ways.

With this technology now mainstream, it is creating value across industries at an accelerated speed. Specific to the banking industry, Generative AI tools can enhance customer satisfaction, lower costs, improve decision-making and employee experience, and decrease risks through better fraud and risk monitoring. Accenture's research and analysis using US labour data found that deploying Generative AI solutions can have high potential impact for many day-to-day tasks and working hours. By 2028, the industry could see a 30% employee productivity gain across front-and back-office operations.¹

In light of the transformative potential and associated risks of Generative AI in the financial sector, the Monetary Authority of Singapore (MAS) has embarked on a Generative AI initiative (Project MindForge) to facilitate responsible use of Generative AI by financial institutions (FIs). As part of the Project MindForge initiative, the banking industry partners Citi, DBS, HSBC, OCBC, Standard Chartered and UOB, and technology industry partners Accenture, Google and Microsoft, supported by MAS, leveraged on the existing Fairness, Ethics, Accountability and Transparency (FEAT) Principles and Veritas Methodology published by MAS, and the experience and expertise of the consortium members to deliver:

- **Industry-Led Whitepaper:** Explores Generative AI benefits in the banking context and outlines a Generative AI risk framework for the financial services industry. This includes assessing the application of existing FEAT Principles and Veritas Methodology to address Generative AI risks and providing suggestions on infrastructure considerations for implementing Generative AI-enabled solutions.

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



- **Industry Use Case:** Enable Generative AI innovation through a centralised experimental use case. An experimental use case illustrates how proposals put forward in the whitepaper can be applied in an actual use case while incorporating responsible AI (RAI) considerations.

1.2. Navigating the Paper

The structure of this document is as follows:

- Section 2 introduces Generative AI, its potential opportunities, and a detailed risk framework to identify Generative AI risks. This section also includes considerations for scaling and adopting Generative AI across organisations, with an outline of the global regulatory landscape around Generative AI.
- Section 3 provides a detailed assessment of existing FEAT Principles and Veritas Methodology and proposes recommendations of additional principles and checklist questions to manage Generative AI risks. This section also includes a high-level evaluation of existing cloud adoption guidelines, addressing risks arising from Generative AI.
- Section 4 details benefits of developing industry experiments and provides an overview of an industry use case developed as part of Project MindForge.
- Finally, section 5 outlines areas for consideration in future phases.

2 Opportunities and Risks of Generative AI

2.1 Generative AI Overview and Opportunities

Generative AI models are deep machine learning models that recognise patterns and learn from large, unstructured datasets, including text, images, audio, visual content or other forms of data. FMs, the core of a Generative AI system (Figure 2.1), learn from these datasets to create new outputs in response to user prompts.

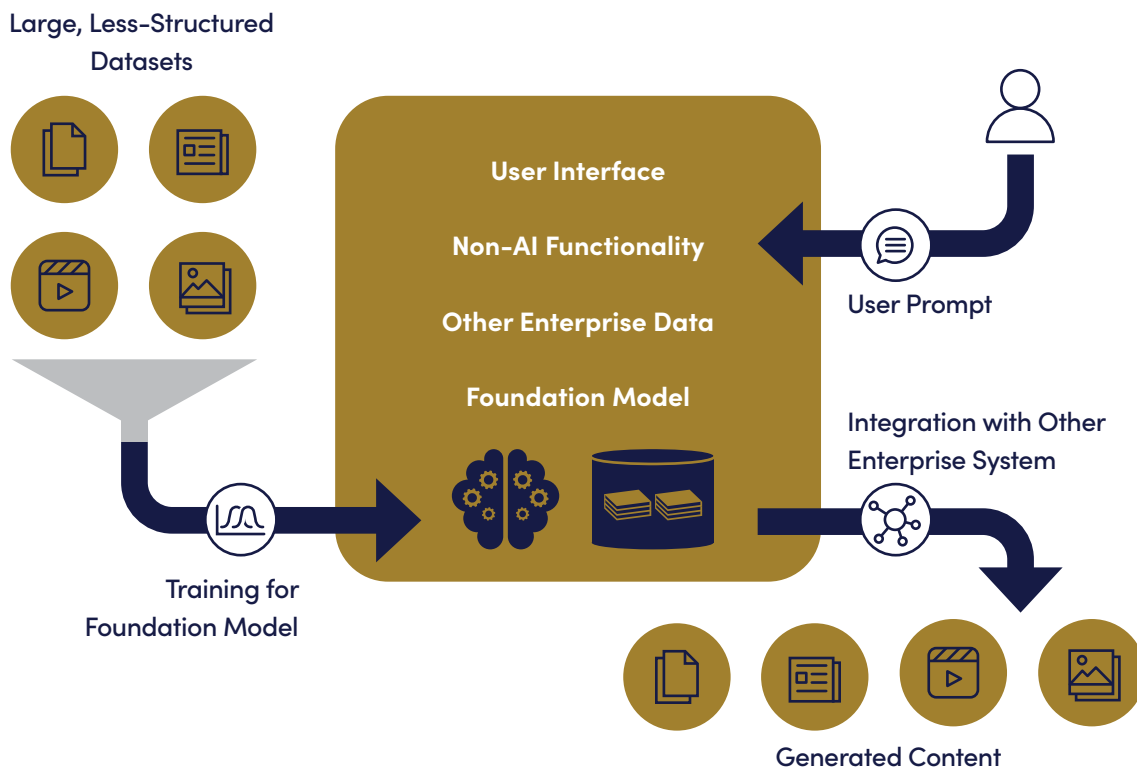


Figure 2.1: Stylised system architecture for a representative Generative AI use case

The ongoing development of Generative AI capabilities is already beginning to improve model outputs. Besides learning from improved data and context, models fine-tune outcomes through human interaction and reinforced learning. Model outputs have also been further refined by improving how humans pose questions to these systems (prompting).

The list of use cases in the financial services industry grows as new opportunities are identified, FIs continue experimenting, and Generative AI capabilities mature. FIs are still in the early stages of implementing these use cases in practice or at scale. Most FIs remain focused on internal and low-risk implementations built around productivity gains, insights, or risk management instead of higher-risk, external-facing use cases.



Nevertheless, FIs are exploring a range of Generative AI applications, with early use cases focused principally on internal productivity, risk reduction, improving insights, and emerging use cases exploring revenue generation and improving customer interaction. The list below includes a sample of common use cases explored by FIs for productivity gains and customer service:



DEVELOPMENT ASSISTANT:

Generative AI can generate user stories, code, test scenarios, and help with code documentation or other technical inputs for software development specific to the financial sector and FI.



RISK IDENTIFICATION:

To manage risks, Generative AI can synthesise risk-related information like adverse news. This will aid risk assessments and surveillance of emerging risks in the external domain, provide early warning for counterparty risks, and assess the impact of market developments on an FI's risk profile.



KNOWLEDGE MANAGEMENT:

Generative AI can retrieve and summarise information, improving employee productivity and increasing automation in customer relations.



MARKET RESEARCH:

Generative AI can develop enhanced insights through summaries of news media and other public-domain developments that are challenging for a human analyst to monitor in real time and at a sufficient scale.



SALES EFFICIENCY:

Generative AI can automate labour-intensive tasks required for generating leads, creating draft communication strategies, and providing recommendations to customers or relationship managers. For example, Generative AI can recommend or even create products for a potential lead, such as an insurance product tailored to a client's risk profile.



HYPER-PERSONALISED MARKETING:

Generative AI can create marketing and communication content for personalised multi-messages or even personalised product offerings, such as a line of credit tailored to a client's needs and creditworthiness.



PERSONALISED EXPLANATION FOR DENIAL:

Generative AI can synthesise bank policies and user-specific information to identify why a customer was declined for a particular product (such as a loan or credit increase) to provide fairer, more transparent financial services.

2.2 Risk Framework

While Generative AI offers numerous opportunities for innovation and advancement in FIs, its value can only be realised through responsible utilisation of the technology. Several factors could challenge its adoption in the financial sector. This includes the scale of FIs, evolving societal and regulatory expectations, disruptions to critical business processes, increased opportunities for cybersecurity threats to emerge, and the preservation of customer and social trust. By examining the risks posed by Generative AI, the paper aims to contribute to developing frameworks and guidelines that will enable its responsible integration within the financial sector.

It should be noted that the risks discussed in this paper will continue to evolve, and some may become less critical as the technology develops. The consortium recommends continuous risk monitoring and evaluation to ensure that effective guardrails are placed whilst maintaining the potential to effectively harness the technology. Furthermore, it is important to recognise the increasingly challenging external risk landscape in a world where Generative AI models are readily available and customisable. A detailed assessment of external risks remains out of scope for this paper and is an area for future study.

With Generative AI, existing risks associated with traditional AI are often amplified. This includes the perpetuation of biases from training data, ethical concerns related to misinformation and deepfakes, resource-intensive computational requirements, and difficulties in explaining model decisions. Other risks are particularly associated with generated content, including a potential lack of control, privacy issues from reproducing personal information, and potential misinformation. Cybersecurity and data protection risks also emerge from the increased attack surface in new AI systems. Risks may be in the form of attacks on enterprise information technology through these new systems, extraction of confidential data through the manipulation of Generative AI systems, or manipulation or damage of Generative AI systems through novel attacks.



Table 2.1 identifies a non-exhaustive list of risks amplified or heightened by Generative AI (see Practitioner Section B.1 for a detailed view) as shortlisted based on consultations and workshops with consortium members. These risks are mapped to key dimensions of



Generative AI risk. These risk dimensions and their associated major risks, identified by the consortium at this stage, provides an initial framework to support a responsible by-design approach to develop, implement and govern Generative AI confidently and in alignment with each FI's values, principles and legal obligations.

Table 2.1: Mapping of risks to risk dimensions of Generative AI

Risk Dimensions of 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 FMs • Adverse or inappropriate impact on individuals and groups
 <p>ETHICS AND IMPACT</p> <p>Ensuring responsible and ethical outcomes in AI use against a clearly defined set of 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 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 accountability within and outside the FI, including third-party accountability • Lack of use and model governance • Inadequate human oversight

Risk Dimensions of 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 • 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.</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 with regulatory or organisational standards • Intellectual Property (IP) infringement • Lack 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 of Generative AI	Select Major Risks Specific to Each Dimension
<div data-bbox="225 539 416 728" data-label="Image"> </div> <p data-bbox="469 555 826 584">MONITORING AND STABILITY</p> <p data-bbox="469 611 882 712">Ensuring the robustness and operational stability of the model or service and its infrastructure.</p>	<ul data-bbox="975 539 1342 1489" style="list-style-type: none"> • Hallucination, fabrication, or 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="225 1592 416 1780" data-label="Image"> </div> <p data-bbox="469 1608 807 1637">CYBER AND DATA SECURITY</p> <p data-bbox="469 1664 858 1877">Protecting data, AI models and systems, and other enterprise information technology (IT) assets from unauthorised access, data loss or leakage, and misuse by malicious actors.</p>	<ul data-bbox="975 1592 1318 2094" 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 (including training data, personal/ company sensitive data) • Model inference attacks, revealing sensitive information

Identifying Risks Across a Generative AI System Lifecycle

It is important that organisations consider risks holistically across the system’s lifecycle, depending on the specificities of each use case. This allows the introduction of early controls and stage gates while solutions are being developed, in line with the previous MAS recommendations on assessing FEAT Principles across the development lifecycle.

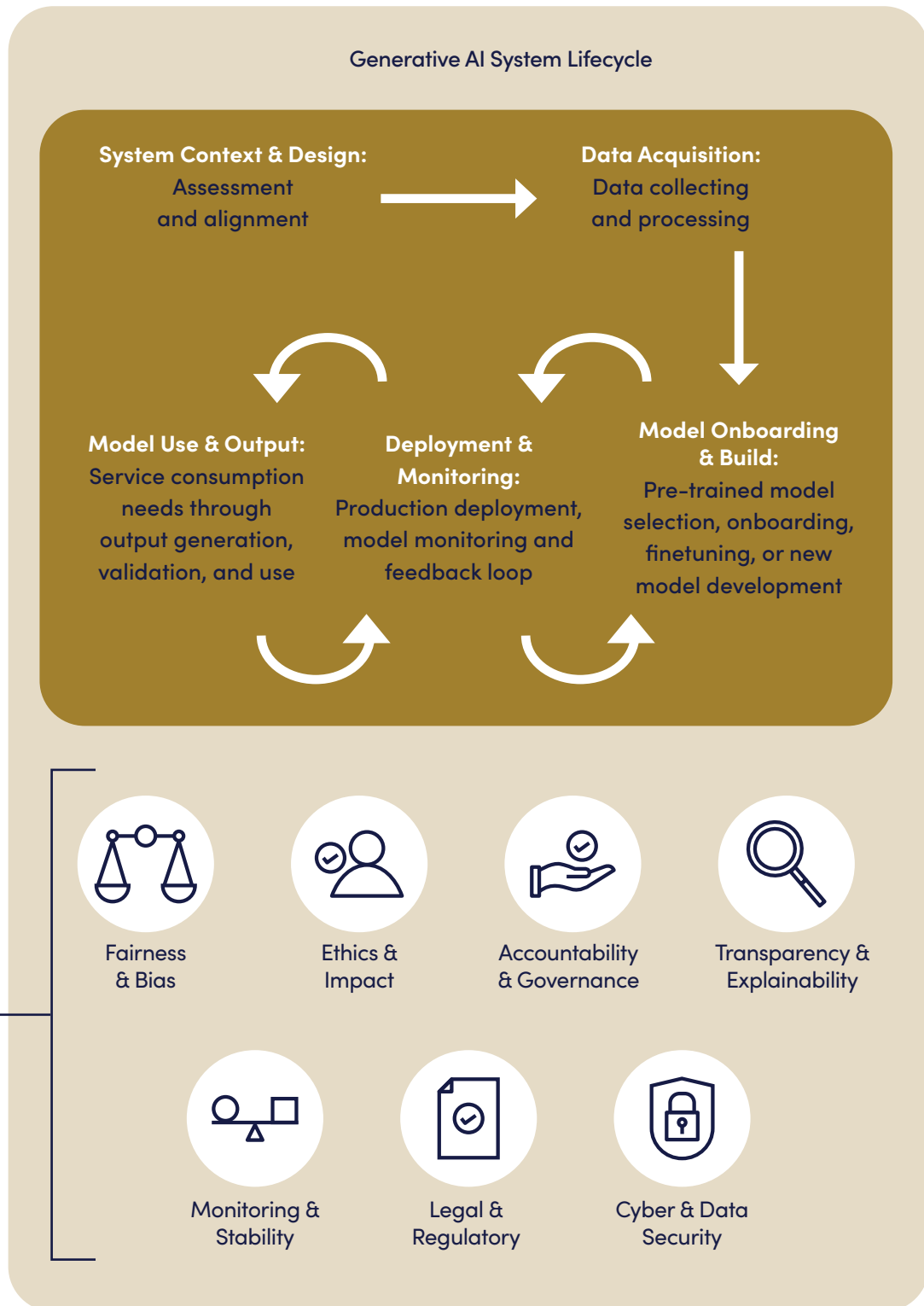
Below are broad considerations that guide risk identification and mitigation across the lifecycle of traditional AI models and Generative AI systems alike.

- **Use Case Context:** This includes clarifying the purpose and expected outcomes of the use case, assessing if the use case is aligned with organisational and social values, and defining the limits of the use case.
- **Use Case Materiality:** This includes the dimensions that determine the materiality of a use case. The FEAT Principles Assessment Methodology (2021)² includes a set of parameters to consider when determining materiality; these considerations remain pertinent in the context of Generative AI.
- **Deployment Pattern:** This includes the infrastructure and ecosystem used for deployment – whether on-prem, cloud-based, procured LLMs, internally developed or from a third-party model – and the implications for the system’s behaviour and security. This consideration is discussed in detail in Section E.1.

The lifecycle depicted in Figure 2.2 is similar to that documented in the MAS Veritas whitepapers, except with the introduction of the stages “Model Onboarding and Build” (choosing an existing model or pretraining your own) and “Model Use and Output” (deploying models for inference and building LLM powered applications). These are specifically labelled as stages because risks unique to Generative AI tend to manifest during development, deployment and use, especially with models sourced from third-party providers, where design decisions may not be fully known.

It should also be noted that the stages “Deployment and Monitoring” and “Model Use and Output” are depicted as a loop; effective Generative AI deployment involves an iterative dialogue between lessons learned during model use and its deployment in the ecosystem.

² MAS FEAT Principles Assessment Methodology. <https://www.mas.gov.sg/-/media/MAS-Media-Library/news/media-releases/2022/Veritas-Document-3---FEAT-Principles-Assessment-Methodology.pdf>



(The list of Generative AI risks included in Practitioner Section B.1 expands on the simplified view.)

Figure 2.2: Illustration of a Generative AI system's lifecycle, with risk dimensions of Generative AI applied throughout

2.3 Regulatory Landscape

When Generative AI exploded on the world stage in late 2022 or early 2023, its potential benefits and capabilities were met with great enthusiasm, particularly as these Generative AI solutions were directly and immediately accessible to the public and enterprises.

However, it was soon followed by a raft of concerns about the risks and negative impacts Generative AI was beginning to display. To date, Generative AI has precipitated a slew of well-publicised lawsuits and general concerns pertaining to IP, privacy, copyright, misinformation or disinformation, toxicity, and information security, to name a few.

Governments and regulators around the world have started investigating the risks of Generative AI and considering required forms of governance and risk management. This is accompanied by an acceleration in policy development and consideration of enforceable regulation and compliance measures. For instance, China and the European Union have quickly responded with steps towards a binding regulatory framework, while the United States has laid out plans for a series of administrative rules and regulations. Singapore, the United Kingdom, Canada, and several subnational jurisdictions have also taken steps towards governing Generative AI through regulation or voluntary guidance.

Compliance requirements around Generative AI risks are expected to grow in the coming years as frameworks specific to the technology are put into practice and new regulatory policies develop. Major considerations for policymakers and regulators appear to be threefold:

- The need to safeguard state and citizen interests, the rule of law, human rights and social values must be effectively balanced with the need for continued investment in innovation to reap the benefits of Generative AI. Too much weight on either consideration can detrimentally implicate other dimensions underpinning effective policy.
- With many Generative AI risks still emerging, the nature and timing of regulation need to be carefully considered. Implementing regulations too early could lead to frequent changes and stifle innovation while implementing them too late could result in adverse social and economic impacts.
- The level of requirements and conditions for any regulation also need to be considered. If it is too low, one could question its value or effectiveness; if it is too high, it could pose a significant barrier to entry for smaller and less mature organisations in the adoption of Generative AI.

In a rapidly changing technological landscape, governance based on consensus, collaboration, and information sharing between public and private sectors has proven effective and reinforced Singapore's position as a leading global market in the responsible use of AI. Regulators also need to facilitate progressive implementation by organisations with open lines of communication and consultation where essential. Inter-regulatory dialogues to harmonise requirements as well as continued open communications and consultations between regulators and organisations, would be instrumental in facilitating this adoption.



3 Risk Assessment of FEAT Principles and Veritas Methodology

3.1 Analysis Approach

As an overall approach, this paper leverages existing AI governance frameworks as it is expected that most FIs will begin their Generative AI governance journey by building on existing governance and risk assessment processes. Existing frameworks include FEAT Principles, Veritas Methodology and guidelines previously published by MAS and the Association of Banks in Singapore, such as Technology Risk Management (TRM), Cloud Implementation, and Outsourcing. This section assesses whether the FEAT Principles are sufficient in addressing the risks in Generative AI, and proposes updates to the Veritas checklist questions where applicable.

In 2018, MAS released a set of 14 principles to promote fairness, ethics, accountability and transparency (Principles to Promote Fairness, Ethics, Accountability and Transparency (FEAT) in the Use of Artificial Intelligence and Data Analytics in Singapore’s Financial Sector).³ Figure 3.1 shows how the principles are mapped to FEAT.



Fairness

- P1:** Individuals or groups of individuals are not systematically disadvantaged through AIDA-driven decisions, unless these decisions can be justified.
- P2:** Use of personal attributes as input factors for AIDA-driven decisions is justified.
- P3:** Data and models used for AIDA-driven decisions are regularly reviewed and validated for accuracy and relevance, and to minimise unintentional bias.
- P4:** AIDA-driven decisions are regularly reviewed so that models behave as designed and intended.



Ethics

- P5:** Use of AIDA is aligned with the firm’s ethical standards, values, and codes of conduct.
- P6:** AIDA-driven decisions are held to at least the same ethical standards as human-driven decisions.

³ FEAT Principles 2018. <https://www.mas.gov.sg/~/-/media/MAS/News%20and%20Publications/Monographs%20and%20Information%20Papers/FEAT%20Principles%20Final.pdf>

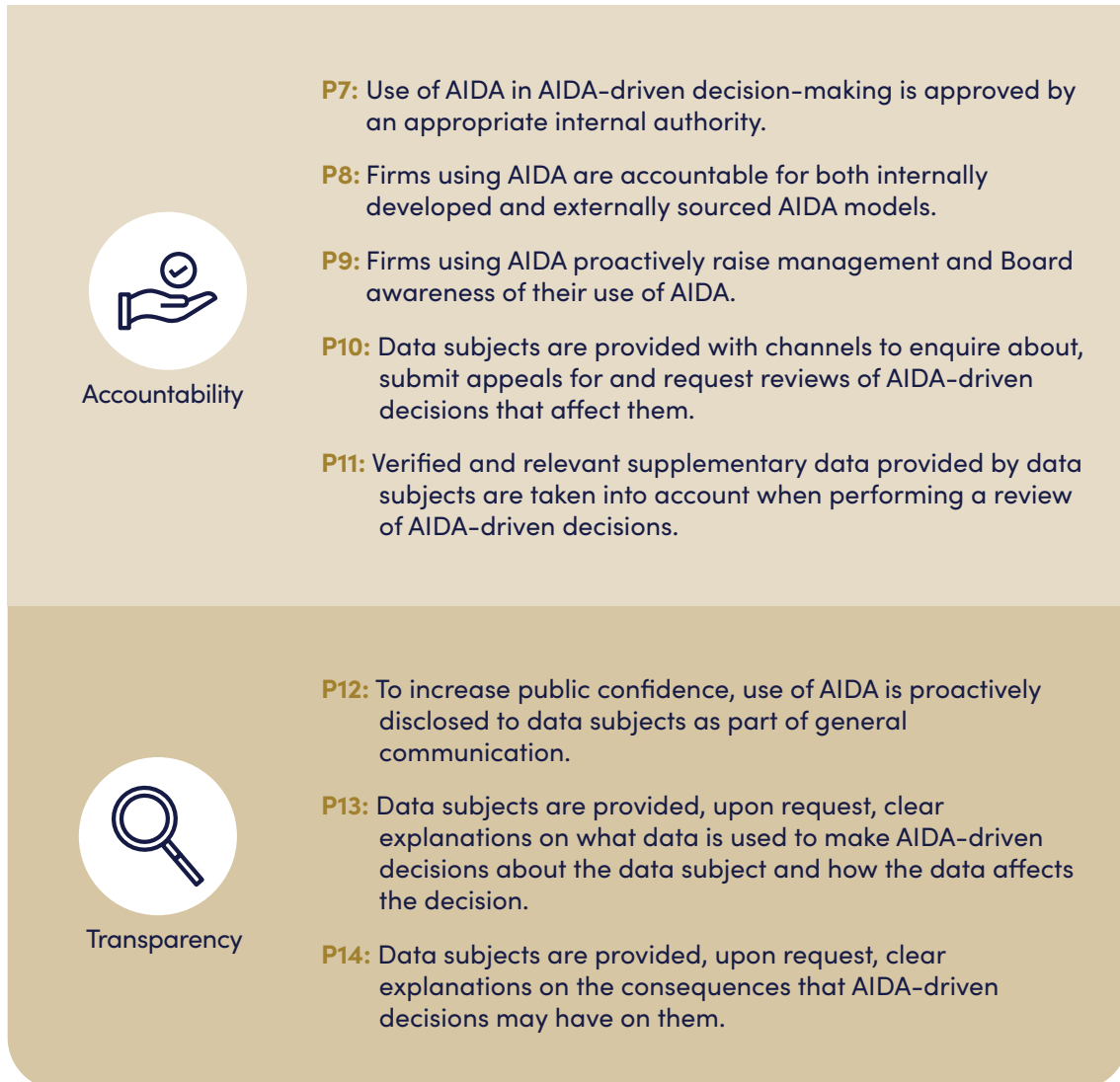
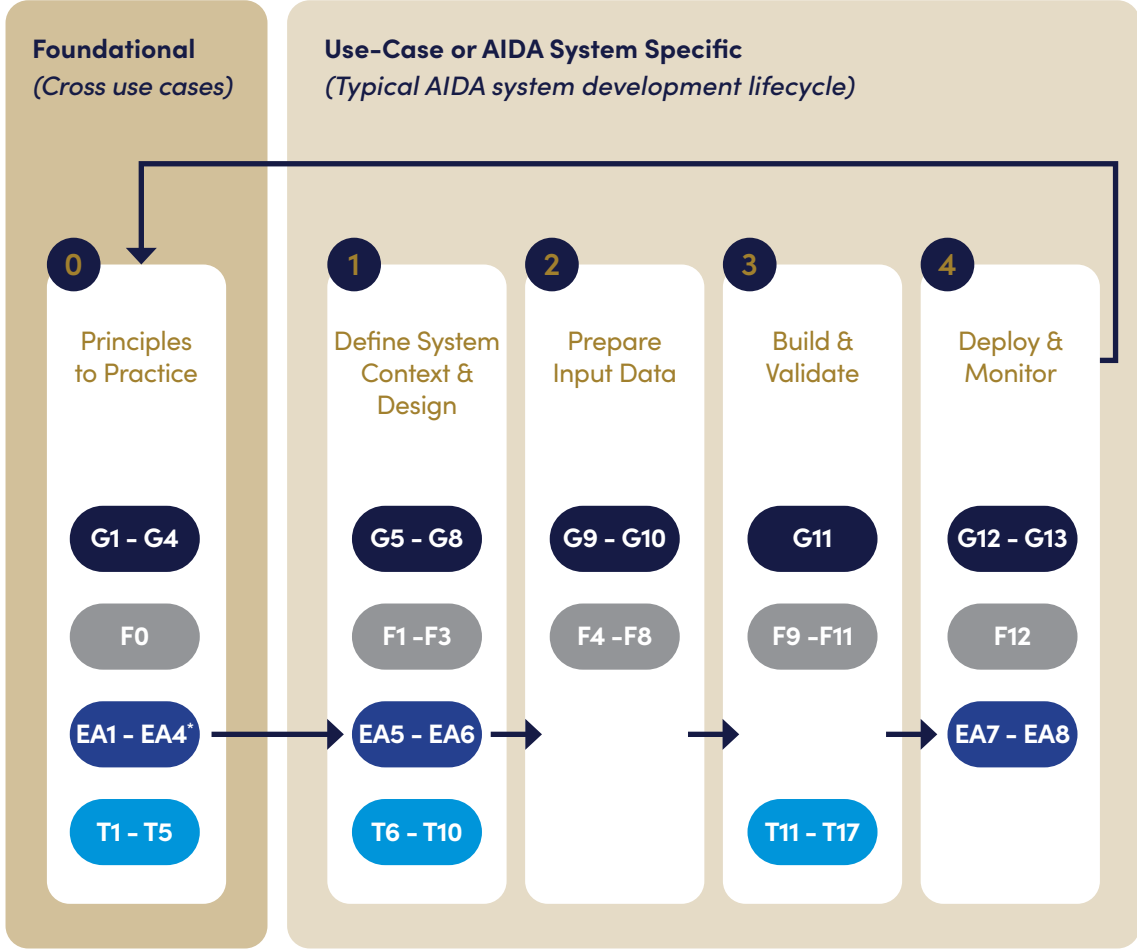


Figure 3.1: 14 principles mapped to FEAT

MAS subsequently released the Veritas Methodology (FEAT Principles Assessment Methodology),⁴ which provides a practical set of steps to support the operationalisation of FEAT Principles. The methodology broadly consists of a high-level view of the AIDA system lifecycle, as shown in Figure 3.2. The methodology includes detailed checklist questions to consider at each stage of the lifecycle.

⁴ MAS FEAT Principles Assessment Methodology. <https://www.mas.gov.sg/-/media/MAS-Media-Library/news/media-releases/2022/Veritas-Documents-3---FEAT-Principles-Assessment-Methodology.pdf>



*Note: EA questions in Step 0 need to be answered before others



Figure 3.2: Outline of Veritas Methodology

3.2 Fairness-Related Assessment

This section assesses the adequacy of fairness principles in addressing Generative AI risk drivers under the fairness and bias risk dimension. For example, Generative AI systems learn from diverse datasets. Without proper evaluation, they may perpetuate or amplify existing social biases in the data. While input-related issues such as data bias also exist in traditional AIDA systems, they are amplified in Generative AI systems, which require large amounts of training data and have the ability to create new content.

Table 3.1: Assessment of fairness principles in addressing Generative AI risks

Fairness Principles	Assessment of Fairness Principles in Addressing Generative AI Risks
<p>P1: Individuals or groups of individuals are not systemically disadvantaged through AIDA-driven decisions unless justifiable.</p>	<p>Considering this principle in the context of Generative AI raises additional challenges. For example, “disadvantage” may arise not only from decisions but the inherent systemic design (including model parameters) and its lack of transparency. Where Generative AI systems are sourced from third parties, there may not be full control, disclosure, or oversight over the nature of training data and system development processes. In particular, if the training data is already biased or contains unauthorised copyrighted material, the system may disadvantage individuals or groups of individuals.</p>
<p>P2: Use of personal attributes as input factors for AIDA-driven decisions is justified.</p>	<p>In traditional AIDA systems, FIs often own training data. Therefore, the process of identifying personal and proxy attributes is straightforward.</p> <p>This principle is still applicable to Generative AI. However, if Generative AI systems are procured from third parties, FIs may not have full control over the training data used. This makes the process of identifying and justifying the use of personal attributes more challenging. Even Generative AI systems refined with data from FIs would have been trained on third-party data unless developed in-house. While this is potentially an issue for all third-party AIDA systems, the scale of datasets used in Generative AI amplifies the concern.</p>



Fairness Principles	Assessment of Fairness Principles in Addressing Generative AI Risks
<p>P3: Data and models used for AIDA-driven decisions are regularly reviewed and validated for accuracy and relevance, and to minimise unintentional bias.</p>	<p>This principle is conceptually adequate but may not be practically effective in addressing Generative AI-specific risks of fairness and bias. It will be challenging for FIs to review data and models developed by third parties due not only to a lack of access but also the enormous scale of datasets to be validated.</p> <p>There are also several types of direct and related fairness risks amplified by Generative AI. One of them is “hallucination”, where a model generates untrue information that could potentially contravene fairness principles. Identifying such risks can be challenging due to a lack of clarity over the source of information used to generate the output.</p>
<p>P4: AIDA-driven decisions are regularly reviewed so that models behave as designed and intended.</p>	<p>This principle is conceptually adequate but may not be practically effective in addressing Generative AI-specific risks of fairness and bias.</p> <p>Generative AI systems are new technologies, and their limitations and risks are not well understood yet. Risk mitigation methodologies are also still an active area of research among AI researchers and practitioners.</p> <p>Importantly, while FIs can review AIDA-driven decisions and identify problematic behaviours, their ability to change the models themselves may be limited, particularly in the case of third-party solutions.</p>

Principles P1 through P4 should still hold when addressing fairness-related risks arising from Generative AI systems. However, P1 may need to be enhanced to ensure its application across the full lifecycle. Additional considerations for FIs using Generative AI systems should also be included, along with a set of extra guardrails and rules designed to ensure the risks of these systems are fully evaluated and properly mitigated.

3.3 Ethics and Accountability-Related Assessment

Ethics and accountability principles seek to ensure AI technology is used responsibly, aligning with social values, upholding privacy norms, and avoiding unintended consequences. This section looks at ethics and accountability principles and their applicability to Generative AI systems under the ethics and impact, and accountability and governance risk dimensions respectively.

Table 3.2: Assessment of ethics principles in addressing Generative AI risks

Ethics Principles	Assessment of Ethics Principles in Addressing Generative AI Risks
<p>P5: Use of AIDA is aligned with the firm’s ethical standards, values and codes of conduct.</p>	<p>It may be challenging to ensure the use of Generative AI within the organisation or by its customers is aligned with its ethical standards and values for several reasons. Firstly, due to limited visibility and control over FM development and sourcing, FIs may not be able to perform the work required (e.g., data audits, code inspection) to ensure the system is built in alignment with its organisational values.</p> <p>Additionally, the variability of created content drives risk, especially in automated use cases. The model cannot independently reason or arbitrate against ethical standards and, therefore, may produce content misaligned with organisational values.</p> <p>Lastly, while FIs can put in place reasonable governance, measurement and monitoring mechanisms to promote alignment, ultimately, errors or confabulations may be difficult to detect. Explaining the rationale of an output would also remain challenging.</p>
<p>P6: AIDA-driven decisions are held to at least the same ethical standards as human-driven decisions.</p>	<p>For human-driven decisions, a lineage of context and reasoning can be documented and relied upon to explain the rationale behind a decision and course of action in support of organisational values. In AIDA implementations involving less complex models, ground truth and predictable output can be used to measure and manage accuracy, while model parameters and weightage can be used as levers to ensure ethical standards are consistently maintained.</p>



Ethics Principles	Assessment of Ethics Principles in Addressing Generative AI Risks
	<p>Achieving the same standards as human-driven decisions in Generative AI may be difficult. For instance, levers required to measure and adjust the models along their lifecycle may not be within an FI's control. The model may inadvertently produce content misaligned with the organisation's values, and the rationale or root cause may be difficult to pinpoint, much less addressed.</p> <p>The lack of independent reasoning and ground truth makes it difficult for models to assess their output against ethical standards. Thus, FI users of Generative AI may have less transparency over the rationale of an output, and therefore unable to hold the Generative AI models to the same standards expected from human-driven decisions.</p>

The ethics principles P5 and P6 broadly cover the ethical risks expanded or amplified by Generative AI. While there is no immediate requirement for updated or additional ethics principles to address Generative AI-specific risks, it is imperative for organisations to monitor ethics-related risks and maintain updated ethical standards, values and codes of conduct. Ensuring that the use of Generative AI within an organisation or its customers aligns with its ethical standards and values may be challenging due to limited visibility and control over FMs. Hence, FIs may need to adapt their governance and procurement processes.

Table 3.3: Assessment of accountability principles in addressing Generative AI risks

Accountability Principles	Assessment of Accountability Principles in Addressing Generative AI Risks
<p>P7: Use of AIDA in AIDA-driven decision-making is approved by an appropriate internal authority.</p> <p>P8: Firms using AIDA proactively raise management and board awareness of their use of AIDA.</p>	<p>While approval and awareness of use may seem straightforward to achieve, there are a number of challenges when implementing this principle.</p> <p>Explainability and transparency are fundamental components for determining and obtaining accountability. Without them, internal authorities cannot approve or assume accountability for outcomes produced by Generative AI use cases. However, this is challenging to achieve when considering third-party Generative AI services where explainability and transparency are limited.</p>

Accountability Principles	Assessment of Accountability Principles in Addressing Generative AI Risks
	<p>While existing outsourcing arrangements, contractual terms and conditions, service level agreements, and other risk mitigation or transfer mechanisms can help alleviate the impact of such risks, they may not adequately address the risks themselves. FIs, as the ultimate accountable party, bears reputational risk and will likely see its stakeholders impacted by irresponsible outputs.</p>
<p>P9: Firms using AIDA are accountable for both internally developed and externally sourced AIDA models.</p>	<p>FIs are ultimately accountable for all Generative AI systems deployed, whether developed in-house or through a vendor. Ensuring that FIs deliver the necessary protection and governance for Generative AI is challenging when vendors do not provide the explainability and transparency needed, as mentioned earlier.</p> <p>Regulatory bodies and customers may also expect FIs to provide clear explanations for AI-driven decisions, especially in critical areas like loan approvals and risk assessments. If FIs are accountable for Generative AI models sourced from external developers, including its responsible deployment and appropriate utilisation, it must be able to validate and explain the model. This may include working with vendors to understand and possibly change the code and data attributes used. FIs are at risk if they assume accountability without controlling parts of the Generative AI development.</p>
<p>P10: Data subjects are provided channels to enquire about, submit appeals for and request reviews of AIDA-driven decisions that affect them.</p>	<p>Enquiry, support and feedback channels would be available regardless of AIDA implementation, its decisions and impacted data subjects. However, FIs may face a challenge explaining and understanding the basis of decisions made by Generative AI models, particularly with closed-source external solutions. In addition, review of Generative AI-driven decisions would be complicated by the lack of visibility over model data, development and generation of output. FI may not be able to explain why a Generative AI model came to a particular conclusion or decision, and therefore may not be able to provide justifications to data subjects or amend the issue.</p>



Accountability Principles	Assessment of Accountability Principles in Addressing Generative AI Risks
<p>P11: Verified and relevant supplementary data provided by data subjects are taken into account when reviewing AIDA-driven decisions.</p>	<p>FI may also be unable to adjust FMs based on a data subject’s request or with supplementary information provided. The models may therefore continue to generate content that is counter to the data subject’s rights which requires human intervention in the final decision making.</p>

Principles P7, P8, P9, P10 and P11 broadly cover accountability-related risks of Generative AI. There is no immediate requirement for updated or additional accountability principles to address Generative AI-specific risks. However, implementing these principles will be challenging if external, closed source FMs are considered for Generative AI use cases within FIs. It is clear that FIs are ultimately accountable for all deployed Generative AI systems, whether developed in-house or through a third-party provider. It is therefore the FIs’ responsibility to ensure that the key vendors provide the needed transparency and explanation on the deployed Generative AI systems. Without a sufficient level of these, internal authorities and management will not be able to approve or assume accountability for Generative AI use cases, especially in critical business services.

3.4 Transparency-Related Assessment

Transparency principles require disclosing the use of AIDA, explaining outcomes, and understanding what data is used to drive the outcome. This is necessary to improve trust in AI technology. This section looks at transparency principles and their applicability to Generative AI systems under the transparency and explainability risk dimension.

Table 3.4: Assessment of transparency principles in addressing Generative AI risks

Transparency Principles	Assessment of Transparency Principles in Addressing Generative AI Risks
<p>P12: Use of AIDA is proactively disclosed to data subjects as part of general communication to increase public confidence.</p>	<p>Typically, FIs use notices on websites and the Terms and Conditions section in application forms to communicate the use of AI to customers. Generative AI falls under AI as a new technology, with one important differentiator – Generative AI generates new content based on prompts supplied. FIs may proactively disclose if communication or recommendations were created using Generative AI and if it was reviewed by a human.</p>

Transparency Principles	Assessment of Transparency Principles in Addressing Generative AI Risks
	<p>For example, FIs may include an explicit consent statement to the use of AI when a customer signs up or applies for a new product or service. If this should include a component built on Generative AI, these statements may need to include additional information. This includes emphasising that end-to-end interaction is done with a machine and that customers must ask to speak to a relationship manager if they feel uncomfortable or unsatisfied. This reassures customers that they can ask for human intervention at any time during the interaction with a machine.</p>
<p>P13: Data subjects are provided, upon request, clear explanations of what data is used to make AIDA-driven decisions about the data subject, and how the data affects the decision.</p>	<p>Explainability of AI is key to implementing this principle. Generative AI, by design, can create new content (including code, texts, images and videos) that may not be comparable to existing precedent. Therefore, incorporating explainability may be a significant challenge.</p> <p>For reasons described earlier when assessing fairness and accountability, it is likely that, where third-party Generative AI services are employed, understanding of the provenance of data used to make decisions may be limited unless comprehensive disclosure by the third party is agreed. FIs may be able to support customers with explanations where the decision is determined using data specific to the FI, but may struggle to disentangle how the underlying FM arrived at an outcome.</p>
<p>P14: Data subjects are provided, upon request, clear explanations of the consequences that AIDA-driven decisions may have on them.</p>	<p>With such a nascent technology, outputs from Generative AI systems may not be predictable or consistent. Thus, providing explanations for its decisions and consequences may be undermined. Until such time that there is a more concrete and consistent understanding of the reliability of Generative AI system outputs, it may be necessary to ensure additional human oversight.</p>



Principles P12, P13 and P14 broadly cover transparency requirements for Generative AI. Current transparency principles address expanded Generative AI risks, including hallucination and anthropomorphism. Hence, there is no requirement for updated or additional transparency principles to address Generative AI risks. However, it is recommended that the elaboration of principles and associated illustrations around transparency be updated to cover Generative AI-specific considerations for reasons outlined in the table above.

3.5 Gaps Beyond FEAT

Generative AI introduces heightened risks, prompting necessary enhancements to FEAT Principles. This section recommends enhancements to existing FEAT Principles and new principles to broaden and update its scope to address emerging risks associated with Generative AI use.

Existing FEAT Principles

The overall assessment indicates that the current set of 14 principles is sufficient to address Generative AI outcomes.

- **Fairness:** The existing principles (P1 to P4) should still hold when addressing fairness-related risks arising from Generative AI systems. However, P1 may require enhancement to cover the full lifecycle – AIDA-driven decisions are its only current focus.
- **Ethics:** There is no immediate requirement for updated or additional ethics principles to address Generative AI-specific risks. However, organisations must monitor ethics-related risks and maintain updated ethical standards, values and codes of conduct.
- **Accountability:** There is no immediate requirement for updated or additional accountability principles to address Generative AI-specific risks. However, challenges posed by using third-party Generative AI systems should be carefully considered.
- **Transparency:** Principles P12 to P14 sufficiently cover requirements related to transparency, including heightened Generative AI risks such as hallucination, copyright, and anthropomorphism. While updated or additional transparency principles are not needed to address these risks immediately, methods to improve transparency to end users should be considered.

New Principles Beyond FEAT for Consideration

Key risk dimensions that need to be considered beyond the FEAT Principles are:

- **Copyright or Intellectual Property (IP) and Privacy (Legal and Regulatory):** The nature of Generative AI entails use of data in public domains for training. In many cases, models are developed by third parties. This presents several legal and regulatory risks, such as copyright or IP infringement, and third-party risks, including accountability and contractual obligations. Risks can be attributed to (a) unclear ownership or copyright of training data

if sourced from the public domain, (b) undefined or unclear rules regarding ownership of new content produced by Generative AI or (c) amplified privacy risks with the use of personal data (e.g., from social media) in Generative AI. Such legal and regulatory risks of Generative AI go beyond the financial sector and will require appropriate cross-sectoral guidance to clarify the copyright, IP and privacy standards that must be met.

- **Monitoring and Stability:** Section 2 identifies several new Generative AI risks under model monitoring and stability. The lack of model robustness can lead to fabricated output or hallucination. The risks from data used in training Generative AI, including data representativeness, quality, and drift, must be properly addressed. Some of these aspects could also apply to non-Generative AI models and merit consideration when the FEAT principles are reviewed.
- **Cyber and Data Security:** Generative AI amplifies security risks such as data poisoning, model manipulation, and prompt injection. This may be attributed to the large amount of data it requires for training, the reliance on third-party suppliers for model and data, and the free-format input accepted and output generated. Further, Generative AI may be used for malicious purposes. They can amplify the scale, speed and sophistication of cyberattacks, scams and information, or influence operations. With a principles-based approach to AI-driven outcomes and their implication for data subjects, FEAT does not explicitly address these risks.

Other Considerations for FEAT

This section identifies opportunities to enhance the remaining FEAT-related publications to ensure their relevance to Generative AI.

- **AIDA Definition and Scope:** FEAT publications define AIDA as “technology that assists or replaces human decision-making”. This definition serves traditional AIDA well. However, it has limitations in addressing the content-generation capabilities of Generative AI and LLMs which may not directly assist in human decision-making. An enhanced definition for AIDA that includes Generative AI characteristics is important to ensure Generative AI systems are covered by the principles. Related considerations include common definitions for Generative AI and other key terms such as LLMs and FMs.
- **Third-Party FM Providers:** FEAT applies to participants in the financial sector who offer financial products and services. With the emergence of FMs and LLMs, third-party technology firms play a greater role. Clarifying the role of such third-party providers, especially in the context of FMs, will help FIs in their adoption of Generative AI-driven solutions or even innovative use of such technologies. This also increases accountability and assurance to end users, regulators, and other stakeholders. Many of the risks regarding the use of Generative AI are amplified due to the lack of transparency. FIs are not assured that systems have been developed to meet and enable the requirements put forward by regulatory authorities. For example, FIs may not have visibility on the suitability and ownership of training data used unless the LLM provider chooses to share them. It is



recommended that additional roles, such as “AIDA providers”, are defined. This clarifies the FIs’ responsibility to set out the necessary service level agreements and contractual terms with third-party vendors to meet regulatory expectations including the need for regulatory access when required, to offer assurance to customers and other critical stakeholders.

- **FEAT Advice and Illustrations:** Illustrations help explain how FIs can adopt the underlying principles. These illustrations cover use cases involving structured data and traditional decision-making applications but do not include the use of Generative AI and its associated risks. FEAT publications should include use case illustrations involving Generative AI and advice on alignment with FEAT Principles.

3.6 High-Level Mitigation Approach to Generative AI-Related Risks

To mitigate Generative AI-related risks, FIs may:

- revise and refine their AI ethical principles to ensure they are still appropriate for the use of Generative AI across the organisation,
- set out internal guidance of the “Dos and Don’ts” when using Generative AI,
- evolve existing AI governance processes to adequately address risks amplified by Generative AI, and
- build the capacity or capability within FIs to appreciate the deployment.

In addition, FIs may take into consideration the following guardrails to mitigate Generative AI-related risks.

1. FIs may conduct thorough due diligence on third-party Generative AI systems, pushing for high-quality and relevant data with justification (if applicable), compliance with legal and IP regulations, customer consent (for the collection, use, disclosure, transfer and storage of data), query abilities, and fair representation in training data. If the third party is unable or unwilling to provide such transparency, the FI can consider including in its contract appropriate representations or warranties about bias, accuracy and other fairness-related aspects.
2. FIs may adopt risk-based approaches such as human-in-the-loop and in-depth analysis of model output to identify potential disparities in model accuracy based on their risk appetite. It is recommended to involve human oversight and intervention. This should not be a once-off but an ongoing process to periodically validate the accuracy of outputs.
3. FIs may implement controls over Generative AI model outputs. They should be evaluated against ethical standards, values, and codes of conduct, with consideration for independent evaluations based on materiality.
4. FIs could also have adequate and reasonable transparency on data provenance, legality and quality, along with model development methods. They should align with ethical

standards, values and codes of conduct, ensuring stakeholder approval and protecting data subject privacy.

5. While FIs are accountable for externally and internally sourced models, responsibilities of vendors and external parties could be explicitly agreed upon and periodically inspected for conformity. Additionally, appropriate contractual clauses within legal contracts with third-party providers could include their responsibility to share information with FIs, so FIs can meet ethics and accountability standards.
6. Ethical vendor onboarding is crucial for deploying Generative AI within FIs. This involves due diligence in ensuring product features align with ethics, transparency, and bias mitigation. FIs can leverage the Ethics and Accountability Framework to identify specifications aligned with their core values. FIs can then ensure vendors provide measurable data, which they can observe and monitor throughout their engagement.
7. FIs may conduct thorough analyses of the performance of deployed AI models through iterative testing and continuous monitoring, and take the appropriate action to address gaps that are found (e.g., discrepancies between AI outputs and ethical standards, gaps between expected and actual model performance). FIs must demonstrate their commitment to responsible AI deployment.
8. To address ethical and accountability gaps, a multidisciplinary approach involving AI experts, ethics professionals, legal advisers and business stakeholders could be deployed.
9. FIs could draw on leading industry expertise and common standards to identify risks and evaluate exposure of AI systems using Generative AI Evaluation Sandbox developed by the AI Verify Foundation and Singapore’s Infocomm Media Development Authority.

3.7 Evaluation of Current Cloud Implementation, TRM and Outsourcing Guidelines

The ABS Cloud Computing Implementation Guide,⁵ MAS Guidelines on Technology Risk Management,⁶ and MAS Guidelines on Outsourcing⁷ were in place or developed prior to the major leaps in Generative AI and its adoption pace from 2022. Generative AI adoption presents a challenge to FI governance in Singapore. It potentially introduces additional complexity into vendor–FI relationships, presents additional security and data protection considerations, and increases certain risks to FIs. Particularly, autonomous outputs

5 ABS Cloud Computing Implementation Guide 2.0. <https://abs.org.sg/docs/library/abs-cloud-computing-implementation-guide.pdf>

6 MAS Technology Risk Management. <https://www.mas.gov.sg/-/media/mas/regulations-and-financial-stability/regulatory-and-supervisory-framework/risk-management/trm-guidelines-18-january-2021.pdf>

7 MAS Guidelines on Outsourcing. https://www.mas.gov.sg/-/media/mas/regulations-and-financial-stability/regulatory-and-supervisory-framework/risk-management/outsourcing-guidelines_jul-2016-revised-on-5-oct-2018.pdf

The consortium’s study of the MAS Guidelines on Outsourcing (2016, rev. 2018) was conducted prior to the issuing of the new Notices and Guidelines on Third-Party Risk Management, which take effect on 11 December 2024. This section’s conclusions refer only to the MAS Guidelines on Outsourcing (2016, rev. 2018) and not to the Notice and Guideline taking effect in 2024. Refer to their text at <https://www.mas.gov.sg/regulation/third-party-risk-management>



by Generative AI systems increase risks to FIs as they may violate IP rules and spread misinformation (see above for detailed discussion on risks exacerbated by Generative AI).

While specific challenges accompany Generative AI, it is not, by nature, fundamentally different from other technologies governed by rules and guidelines on cloud implementation, TRM and outsourcing. The consortium expects TRM to be pertinent to all Generative AI implementation. Many, but not all, Generative AI systems implemented by FIs are also expected to utilise outsourcing or cloud arrangements. FIs implementing Generative AI systems in Singapore will likely be able to continue complying with requirements set out by the three instruments, where they are applicable, within the framework of a careful and responsible Generative AI implementation.

Cloud computing, specifically private cloud, will be, for most FIs, the infrastructure of choice for their Generative AI system. The requirements of implementing a Generative AI system in the cloud are not expected to be materially different from other cloud computing deployments. Data guardianship will be particularly relevant to Generative AI systems to protect confidentiality in training data. Considerations around materiality would also be highly relevant since the current list provided by ABS was not devised with Generative AI in mind. Its extensive discussion on governance may need to be further supplemented in light of Generative AI – it should reflect that models co-developed with FIs have additional governance considerations.

TRM may be where the greatest adaptation will be required to address the changing demands of Generative AI. The risks identified in the report are some examples. As Generative AI evolves, more risks will be identified. Please refer to MAS 2021 TRM guidelines for technology risks. Furthermore, Generative AI could impose additional considerations around shared responsibility for vendor-licensed FMs. FIs licensing a model may not have full visibility of the model's design. In fact, the design may be the vendor's trade secret. As such, responsibility for the model's outputs must be adequately attributed between the two.

Outsourcing, whether in cloud computing or procurement of Generative AI FMs or whole systems on a software-as-a-service basis, will be relevant to many Generative AI implementations. Provisions in the Notice on Management of Outsourced Relevant Services as well as Guidelines on Outsourcing will likely require minimal updates to accommodate the outsourcing of Generative AI systems. Outsourcing Generative AI systems is not qualitatively different from outsourcing other technologies. The current guidelines cover principal Generative AI risk considerations. This includes data sovereignty, availability of critical business processes, protection of shared confidential information, and weighing public cloud arrangements against private cloud or other hosting options depending on the materiality of the use case.

The consortium expects the procurement of nearly all Generative AI systems in the foreseeable future to involve third parties. However, some arrangements will not qualify as outsourcing. Existing guidelines around TRM and business continuity management (see MAS' 2022

Business Continuity Management Guidelines⁸) address many aspects of such relationships. The additional complexity introduced by the licensing of Generative AI FMs may require additional consideration.

The growing adoption of Generative AI systems and the growing role of third-party technology vendors in developing, providing, and maintaining AI systems highlight the importance of a broad, integrated approach to governance. This section only discusses guidelines that apply directly to FIs. They should be considered holistically alongside regulations, guidelines and other requirements that lie outside the realm of financial sector regulation.

The impact of Generative AI on the broader regulatory framework for FIs in Singapore is not yet fully understood, particularly as it pertains to instruments beyond the three assessed in this section. The MindForge consortium undertakes the continued study of the issue, acknowledging that more work must be done to determine what changes, if any, will be required to adopt MAS' regulatory framework or supervisory approach to accommodate challenges posed by Generative AI.

8 Business Continuity Management Guidelines. <https://www.mas.gov.sg/-/media/mas/regulations-and-financial-stability/regulatory-and-supervisory-framework/risk-management/bcm-guidelines/bcm-guidelines-june-2022.pdf>



4 Use Case Implication

4.1 Need for Industry Use Cases

It is evident that Generative AI possesses substantial potential to reshape the banking industry and drive its value chain, but not without introducing novel risks.

The importance of industry use cases is underscored by their role as testing grounds for real-world applications of Generative AI within the banking sector. By fostering collaboration among FIs, these initiatives will enable the pooling of resources, expertise, and insights to assess the viability and impact of Generative AI on their operations. Such collaborative efforts can be instructive in shaping the principles guiding the technology's adoption, promoting a unified and responsible approach to innovation. A key measure of success for these experimental use cases lies in their capacity to propose actionable solutions and best practices.

Industry use cases empower FIs to make informed decisions regarding Generative AI adoption, ensuring alignment with industry standards and regulatory requirements. In essence, they pave the way for the banking sector to navigate the transformative potential of Generative AI with confidence, resilience, and a shared commitment to responsible innovation.

4.2 Description of Industry Use Cases

Compliance Co-Pilot

The dynamic nature of regulatory changes demands dedication from banks, necessitating constant vigilance, resource allocation, and strategic adaptation to ensure ongoing compliance. In 2020, a third of financial institutions reported spending 5% or more of their annual budget on regulatory compliance⁹ and in 2020, MAS issued 55 articles on Guidance and Notices¹⁰ while other leading regulators published 850 new pages of AI regulations.¹¹ A majority of compliance professionals reportedly spent at least 10% of their time on the review and analysis of regulatory developments, with 28% spending more than 20% of their time on it.¹² These statistics emphasise the magnitude of resources and manual effort invested by the banking industry to remain compliant, particularly for banks whose operations span jurisdictions, increasing the complexity of their compliance operations.

The Compliance Co-Pilot, an intelligent assistant powered by Generative AI, aims to assist FIs in managing complex tasks that are effort-intensive throughout the policy lifecycle in a

⁹ Kroll's Global Regulatory Outlook 2020. <https://www.kroll.com/en/insights/publications/financial-compliance-regulation/global-regulatory-outlook-2020>.

¹⁰ A 2020 analysis of MAS' page on Regulations and Guidance. <https://www.mas.gov.sg/regulation/regulations-and-guidance>

¹¹ A 2022 analysis by Statista Research Department. <https://www.statista.com/statistics/656873/time-spent-per-week-by-compliance-teams-updating-procedures/>.

¹² Thompson Reuters Regulatory Intelligence survey in 2023. <https://legal.thomsonreuters.com/en/insights/reports/cost-of-compliance-2023/>.

context-specific manner. The Compliance Co-Pilot can be a potential asset in enhancing compliance via its capacity to swiftly process huge volumes of unstructured regulatory data. The Compliance Co-Pilot is not intended to replace human analysis; rather, it is a tool to expedite manual work so that compliance officers can focus on tasks that require critical thinking and analysis.

The use case is developed as an experimental proof of concept and is jointly developed by UOB, Accenture, Standard Chartered, HSBC, Citi, and Microsoft, utilising selected anti-money laundering (AML) policies from regulatory bodies in Singapore, Indonesia, Malaysia, Thailand, the United States, along with selected credit risk policies from Singapore and Malaysia. The use case serves to assess Generative AI capabilities in compliance management, and design a security framework that helps facilitate secure and responsible sharing, storage and processing of proprietary/confidential data from banks.

The Compliance Co-Pilot leverages on Generative AI capabilities, such as summarisation, persona-based explanation, comparison, citation, and communication drafting. These capabilities enhance efficiency, drive consistency in policy interpretation, and improve overall policy lifecycle management.

System Design and Testing

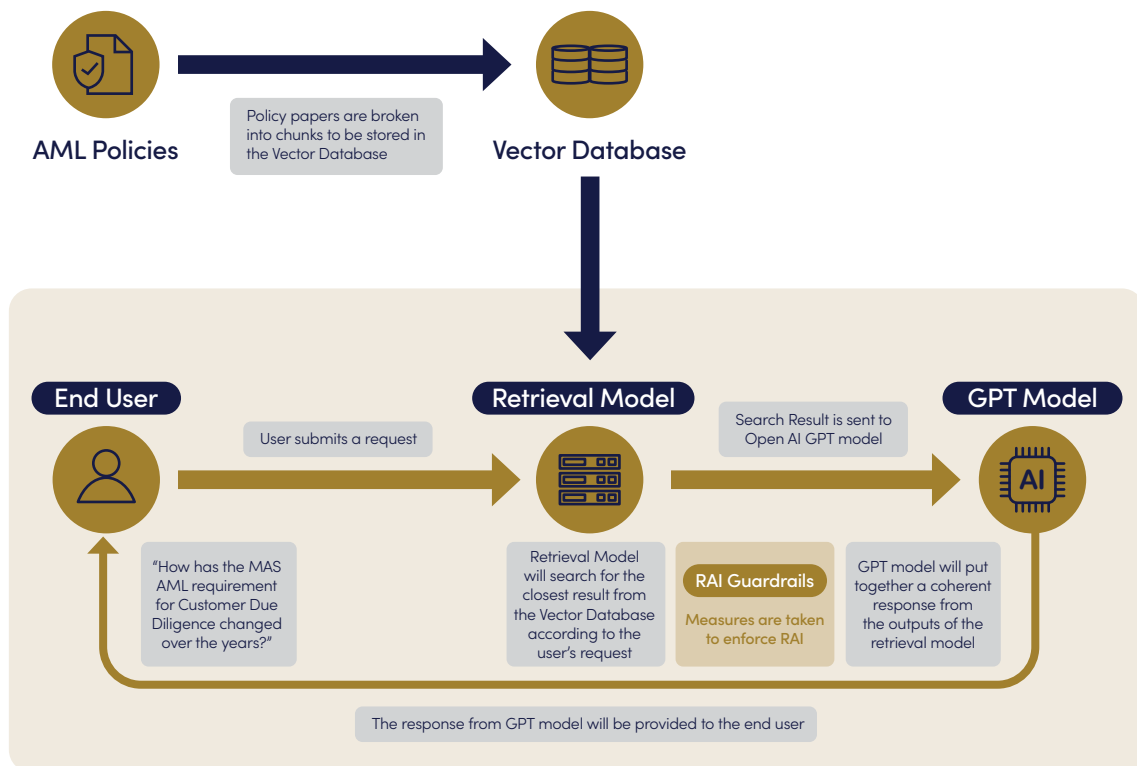


Figure 4.1: Illustrative working process of Compliance Co-Pilot



In the system design, the consortium has adopted a retrieval augmented generation (RAG) approach to enhance the quality of model output. This approach enables the consortium to integrate domain-specific data during model inferencing without the need to retrain or fine-tune LLMs. Moreover, this design has several advantages, such as more accurate and relevant outputs as well as reduced hallucinations. In addition, appropriate guardrails have been put in place to tackle RAI concerns related to bias, toxicity and security. As illustrated in the technology reference architecture in Figure E.5, the FM in the system is built on appropriately managed data and integrated effectively into a broader architecture, including a retrieval model, underpinned by effective technical guardrails and security measures.

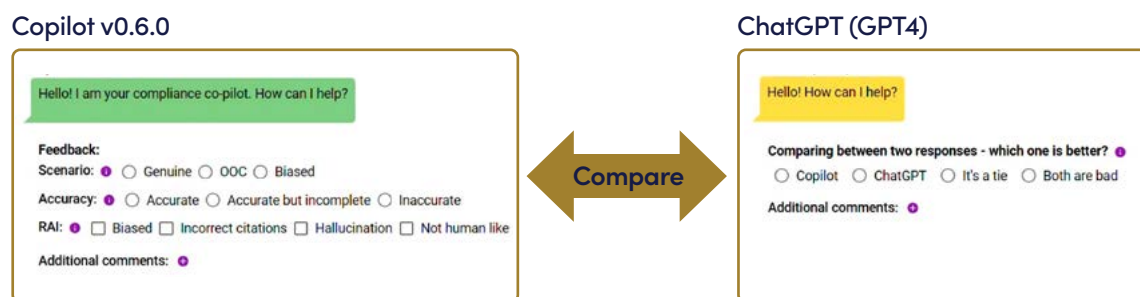


Figure 4.2: Illustrative output testing from Compliance Co-Pilot and Open AI GPT by subject matter experts



Extensive testing involving subject matter experts from multiple banks has been conducted to evaluate relevant RAI risk elements in the pilot use case. The subject matter experts underwent several rounds of testing to determine if the pilot use case outperforms the Open AI GPT model in areas such as accuracy, bias mitigation, and hallucination reduction (see above figure).

Risk Dimension Assessment



This section assesses the Compliance Co-Pilot using the risk framework specified in Section 2.2 to understand the risk elements applicable to the pilot use case. The following Table 4.1 provides a comprehensive assessment of the different risk elements, with an emphasis on key risks. The risk level in this section helps to demonstrate the effectiveness of the mitigation measures in addressing inherent risks. The risk level is defined by two dimensions – severity of the harm and the probability of harm happening.



For example, the severity of harm due to hallucination is high because this could lead to wrongful interpretation of the policy. As the probability of this happening is likely, the inherent risk level is high. The adoption of RAG approach has reduced the probability of this happening. As such, the residual risk level has dropped to moderate. The residual risk level is the remaining risk level after the application of relevant mitigation measures.

Table 4.1: Risk framework assessment using Compliance Co-Pilot use case



Risk Dimension	Risk Elements	Inherent Risk Level	Risk/ Implication	Residual Risk Level	Mitigation Measures
 Fairness and Bias	Unrepresentative or biased data inputs	High	Since the FM is provided by a third party, FIs lack the ability to determine and scope the relevant data that will be retrieved and processed by the FM, and whether the data used in developing the FM exhibit bias against any individual or groups of individuals.	Moderate	Extensive testing by industry experts and safeguards like the use of Azure built-in toxicity checks and prompt engineering have been put in place to reduce risks associated with fairness and bias to a certain extent.
	Adverse or inappropriate impact to individuals and groups	High		Moderate	
 Ethics and Impact	Dark patterns	High	There is genuine concern about dark patterns and toxic outputs, as any misinformation, unintended responses or hallucination originating from the Compliance Co-Pilot can result in misguided actions if the output from the pilot is used without proper human oversight/ supervision.	Moderate	The RAG approach and RAI guardrails such as toxicity checks have demonstrated effectiveness in reducing hallucination & toxicity. Using the RAI guardrails, we can prevent and manage dark patterns and toxicity in both input and output of the Compliance Co-Pilot. In addition, associated risks will be largely reduced if the usage of this portal is guided by appropriate subject matter experts.
	Toxic and offensive outputs	High		Moderate	





Risk Dimension	Risk Elements	Inherent Risk Level	Risk/ Implication	Residual Risk Level	Mitigation Measures
 Accountability and Governance	Lack of Generative AI risk awareness	High	Since Generative AI is relatively new, there is a limited level of awareness when addressing potential risks. We see a need to educate users on the impact of this technology.	Moderate	Given the critical nature of the policies (i.e., AML and credit risk) involved, it is important to ensure that the usage of this portal is guided by subject matter experts. Additionally, appropriate disclaimers should be put in place to alert users about the potential risks (such as hallucination) of Generative AI.
	Lack of third-party accountability	High		High	
	Inadequate human oversight	High	In addition, with third-party FMs, it is important for FIs to consider the risks and benefits when assuming accountability for the usage of this policy portal.	Moderate	
 Transparency and Explainability	Unclear output accuracy	High	A lack of clear output accuracy and data traceability poses significant risks when utilising this use case portal. Failure to address this issue potentially renders the model unusable.	Moderate	Extensive testing has been performed to evaluate model performance. Multiple rounds of human testing were carried out with compliance subject matter experts. A specialised testing user interface was designed to collect subject matter experts' feedback around accuracy, relevancy and risk compliance (like hallucination, bias and toxicity). From close to 200 "In-context" queries, responses with "accurate" and "accurate but incomplete" accounted for 76% of the responses. After several rounds of refinement, we were able to achieve an accuracy rate of 76% based on the
	Unclear provenance for training/ test data	High		High	

Risk Dimension	Risk Elements	Inherent Risk Level	Risk/ Implication	Residual Risk Level	Mitigation Measures
 <p>Transparency and Explainability</p>					human-in-the-loop approach and assessment across the phases of development and deployment.
	Lack of explainability	High	Given that the model explanation method for Generative AI is currently the subject of extensive research, it will be a topic for future discussion and analysis. As such, the inherent risk level is high.	Moderate	In order to meet the necessary level of external transparency, citations and information sources are featured in the output, allowing users to easily identify the origin of the results, and better understand if the answer provided by the model is accurate.
	Anthropomorphism	High	This risk is closely linked to the lack of Generative AI risk awareness.	Moderate	Appropriate disclaimers should be put in place to alert the users about the potential risks of Generative AI and how users should use and interpret the results.
 <p>Legal and Regulatory</p>	Breach or misalignment with regulatory or organisational standards	High	This risk is closely linked to the lack of Generative AI risk awareness.	Moderate	It is important to ensure that this use case pilot is used under the guidance of subject matter experts with the necessary domain knowledge.
	IP infringement	Low	The regulatory policies are public data meant for consumption by banks to ensure compliance with credit risk and AML policies.	Low	



Risk Dimension	Risk Elements	Inherent Risk Level	Risk/ Implication	Residual Risk Level	Mitigation Measures
 <p>Legal and Regulatory</p>	Lack of IP protection	Low	The regulatory policies are public data meant for consumption by banks to ensure compliance with credit risk and AML policies.	Low	It is important to ensure that this use case pilot is used under the guidance of subject matter experts with the necessary domain knowledge.
 <p>Monitoring and Stability</p>	Hallucination/ Fabrication/ Confabulation	High	These are genuine concerns that can result in misguided actions.	Moderate	Although hallucination is a common problem with Generative AI models, the RAG approach has demonstrated effectiveness in addressing this challenge.
	Overconfidence	High		Moderate	
	Insufficient model accuracy/ soundness	High		Moderate	By incorporating relevant external knowledge into the generation process, RAG can help mitigate hallucinations by encouraging the model to produce responses that are more aligned with factual information and contextual relevance. However, it's important to note that RAG alone may not completely eliminate hallucinations, as they can also arise from limitations in the model's understanding of context and the complexity of the input prompt. Additional techniques, such as fine-tuning on specific tasks or filtering generated responses based on their coherence and factual accuracy, may be

Risk Dimension	Risk Elements	Inherent Risk Level	Risk/ Implication	Residual Risk Level	Mitigation Measures
 Monitoring and Stability					necessary to further address hallucinations in LLMs. In addition, under the guidance of subject matter experts, the risks in using this use case portal can be reduced.
 Cyber and Data Security	Cyber and data security	Low	In the pilot phase, as data used is limited to publicly available information, this risk is immaterial.	Low	For future phases to ensure that cyber threats do not percolate from foundation/source models into FIs' internal systems, a secure architecture is being worked on.

Application of Veritas Methodology

This section aims to showcase how the Veritas Methodology is used to evaluate the Compliance Co-Pilot against the MAS FEAT Principles in a verifiable manner. While there are 51 checklist questions in the Veritas Methodology, we focus solely on those relevant to this use case. We will omit foundational questions, namely G1–G4, EA1–EA4, T1–T5 and F0, from this assessment as their purpose is to advise FIs in establishing the foundation for proper implementation of FEAT Principles across all AIDA use cases. General questions, namely G5–G8 and G12–G13, are also excluded from considerations as they are not relevant to the use case study.

Table 4.2: Assessment of Compliance Co-Pilot use case against Veritas Methodology

FEAT Principles	Checklist Question	Assessment
Fairness	F1–F12	Currently, there is a lack of research material regarding the fairness assessment for unsupervised learning including Generative AI models. This is an area that will require further examination in the future.








FEAT Principles	Checklist Question	Assessment
Ethics and Accountability	EA5–EA8	<p>We have identified several key areas where the Veritas Methodology for assessing ethics and accountability is applicable. These include IP infringement, toxicity and dark patterns.</p> <p>To conduct the evaluation, subject matter experts submit a variety of questions to test the model with the aim of identifying unintended model outputs or results that do not conform to the responsible use of the model.</p>
Transparency	T6–T17	<p>Given the critical nature of these policies, the need for transparency requirements and standards is of utmost importance. However, as the explanation method for Generative AI is currently under extensive research, we may need to revisit this in the future.</p> <p>To address external transparency requirements, we have integrated citations and information sources into the output, allowing users to easily identify the origin of the results.</p>
General	G9–G10	<p>Since these are third-party FMs, there is a lack of transparency regarding data and processing. Therefore, the assessment will have to be re-evaluated in the context of Generative AI.</p>

Risks Addressed/Not Addressed in Veritas Methodology




While the risk framework aims to highlight risks associated with Generative AI, the Veritas Methodology has been developed to assess the use case in alignment with the FEAT Principles. By combining the Veritas Methodology with risk dimensions in the following table, we can pinpoint areas of improvement in the methodology to address emerging risks.

Table 4.3: Evaluation of Compliance Co-Pilot risks elements against Veritas Methodology

Risk Dimension	Risk Elements	Generics	Fairness	Ethics and Accountability	Transparency	New Principles	Remarks
 Fairness and Bias	Unrepresentative or biased training data		N				Due to the residual risk and lack of research material in accessing fairness in Generative AI models, further discussion and examination are warranted.

Risk Dimension	Risk Elements	Generics	Fairness	Ethics and Accountability	Transparency	New Principles	Remarks
 Fairness and Bias	Adverse or inappropriate impact to individuals and groups		N				Due to the residual risk and lack of research material in accessing fairness in Generative AI models, further discussion and examination are warranted.
 Ethics and Impact	Dark patterns			C			The Veritas Methodology for ethics and accountability can aid in breaking down the management of these risks into commitments and specifications, allowing for a qualifiable way to assess the mitigation of these risks.
	Toxic and offensive outputs			C			
 Accountability and Governance	Lack of Generative AI risk awareness			C			
	Lack of third-party accountability			N			
	Inadequate human oversight			C			We can apply the ethics and accountability methodology for this risk.
 Transparency and Explainability	Unclear output accuracy				N		This is a topic for future discussion.
	Unclear provenance for training/test data				N		
	Lack of explainability				N		
	Anthropomorphism				N		



Risk Dimension	Risk Elements	Generics	Fairness	Ethics and Accountability	Transparency	New Principles	Remarks
 Legal and Regulatory	Breach or misalignment with regulatory or organisational standards			C			We can apply the ethics and accountability methodology for this.
	IP infringement					N	
	Lack of IP protection					N	
 Monitoring and Stability	Hallucination/ Fabrication/ Confabulation					N	
	Overconfidence					N	
	Insufficient model accuracy/ Soundness					N	
 Cyber and Data Security	Cyber and data security				N		This is a topic for future discussion.

N: Areas that require more research as well as broadening of FEAT Principles and Veritas Methodology
 C: Areas that are already covered by FEAT Principles and Veritas Methodology

5 Next Steps

This paper provides an initial overview of considerations, risks, leading practices, and options for the responsible implementation of Generative AI. It has analysed the suitability of the existing FEAT Principles and Veritas Methodology for Generative AI applications and provided perspective on best practices for architecture and enterprise capabilities to deliver these systems.

With this foundation, future phases can enhance the following key areas based on consortium recommendations:

1. Extension and expansion of FEAT and Veritas

One of the key outcomes of Project MindForge is to assess the completeness of the FEAT Principles in their application to Generative AI use cases. In this regard, the consortium recommends that MAS consider addressing all of the key principles FIs should consider in implementing Generative AI. However, to address specific risks and challenges posed by Generative AI, the consortium recommends enhancing fairness principle P1 and also supplementing the FEAT Principles with new domains such as copyright/IP and privacy, monitoring and stability, and cyber and data security as suggested in Section 3.5.

2. Detailed advice and risk mitigations

The risks identified in this paper (and detailed in Practitioner Section B1) are numerous, and represent a mix of risks already identified or controlled as part of frameworks. A renewed set of principles and methodology should include a detailed perspective on mitigations and guardrails for each of these risks. Generative AI, as an emerging technology that is just entering widespread societal use, presents “unknown unknowns” in addition to the risks enumerated in this paper. Only further experience and careful study will make it possible to approach these risks with specific and appropriate mitigations.

3. Detailed guidance on managing unstructured data

The growing importance of unstructured data – text, images and audio-visual content on a scale that is extremely challenging for humans to review – will challenge the data governance capabilities of FIs. Additional efforts are required to ensure that the governance of data in FIs is fit-for-purpose as they embark on the use of Generative AI. As new and existing unstructured data is harnessed, FIs must assess whether these can power the next generation of Generative AI systems. Governance policies and procedures may need to be updated, technologies may need to be modified, and skills may need to be developed in order to hold unstructured datasets to the same standards as structured data currently.

4. Continue to consider the implications of Generative AI for TRM, outsourcing, and cloud services implementation

Besides TRM, which has straightforward relevance to all technology systems, it is typical for FIs to outsource all or part of a Generative AI system’s deployment, and to do so via a cloud service arrangement. FIs in Singapore will turn to these guidelines as part of their Generative AI journey and, with the new notice and guideline on outsourcing now released, will also refer to those documents. This paper’s initial perspective identified several high-level subject areas of the guidelines that will be impacted by innovations in Generative AI. This perspective is a first step in an exploration of these impacts.



Appendix A: Glossary

Generative AI: A technology using deep learning that produces outputs in one or several forms (text, image, video, audio, etc.). Generative AI is based on a foundation model, which is a large neural network (consisting of at least several million parameters) that is trained on a large dataset that may be unstructured and will, in some respect, resemble the desired outputs. Generative AI is distinct from traditional AI in its ability to produce complex and novel outputs, sometimes in unstructured media like text, in response to a user prompt.

Large language model: A deep learning algorithm that can perform a variety of natural language processing (NLP) tasks. Large language models use transformer models and are trained using massive datasets. This enables them to recognise, translate, predict, and generate text or other content.

Foundation model: A deep learning algorithm that has been pre-trained on a broad spectrum of generalised and unlabelled data and capable of performing a wide variety of general tasks such as understanding language, generating text and images, and conversing in natural language.

Training data: A large dataset used to train the foundation model and consisting of labelled (annotated with the correct output for the given input) and unlabelled (data consists only of inputs) data.

Self-supervised learning: Unsupervised learning that involves training a foundation model to predict some parts of the input data without using explicit labels.

Overfitting: A situation where the foundation model is well trained for specific patterns or tasks and performs poorly on new, unseen data. Typically occurs when the foundation model has too many parameters or when the training data for large-scale models is too small or biased.

Deep learning neural networks: Types of foundation models that are based on the structure and function of neurons in the human brain. They contain layers of interconnected nodes and are capable of learning more complex features and patterns to process input data and make predictions or decisions.

Weight parameters: Values that a foundation model learns from the training data to make predictions or decisions. Typically, initial weight parameters are values assigned during the initial foundation model training, and are adjustable to give optimal performance and minimise the error between the foundation model's predictions and the actual outputs.

Knowledge base: A form of repository designed to capture the knowledge of human experts and serve as reference for foundation model to support decision-making and response generation to complete the task.

Fine-tuning: Adaptation activities that involve taking a pre-trained foundation model and further training it on a new dataset to perform better on a new task within a similar domain.

Prompt engineering: Adaptation activities that involve designing prompts or questions to guide the behaviour output of foundation models.

Evaluation metrics: Used to evaluate the performance of foundation models. Common evaluation metrics for language modelling include accuracy, precision, recall, and F1 score.

Application-layer capabilities: Capabilities which extend, supplement, and orchestrate inputs from Generative AI systems in order to integrate them with a user interface and with a broader application ecosystem in an FI.



Practitioner Section



B.1: Risk Definitions

This section provides a non-exhaustive list as shown in Table B.1.1 comprising dimensions of RAI, risks pertinent to each dimension and its respective definitions to help FIs augment their existing AI governance approaches and frameworks. It is important for FIs to take note that there are certain risks that can manifest across multiple risk dimensions of Generative AI. For example, hallucination/fabrication/confabulation risks can be mapped across the following dimensions of RAI: fairness and bias, legal and regulatory, and monitoring and stability. Similarly, third-party accountability can be mapped across accountability and governance, and legal and regulatory dimensions. In this paper, detailed discussion on each risk is only in reference to their primary risk drivers to provide readers with a focused and pragmatic perspective that facilitates a deeper understanding of the risk dimensions.



The lifecycle stages included below are:


- System context and design: assessment and alignment
- Input data acquisition and preparation: data collection and processing
- Model onboarding or build: pre-trained model selection, onboarding, finetuning, or new model development
- Deployment and monitoring: production deployment, model monitoring and feedback loop
- Model use and output: servicing consumption needs through output generation, validation, and use

Table B.1.1: List of Generative AI risk dimensions, risks pertaining to each dimension and its description



Risk Dimension	Risks Pertinent to Each Dimension	Risk Definition	Secondary Dimensions Impacted by Risk	Lifecycle Stages Implicated				
				System Context and Design	Data Acquisition	Model Onboarding and Build	Deployment and Monitoring	Model Use and Output
 Fairness and Bias	Unrepresentative or biased data inputs	Data is biased against, or unevenly represents, certain individuals or groups of individuals, which can produce biased model outputs.	Monitoring and Stability		✓	✓		
	Adverse or inappropriate impact to individuals and groups	Models generate outputs that can be detrimental or inappropriate for individuals or groups.					✓	✓
 Ethics and Impact	Value misalignment	Generative AI service, output or its use does not align with corporate or social values.		✓				
	Environmental Sustainability Impact	Environmental impact of running LLMs, especially increased carbon emissions which impact the corporate social responsibility and ESG outcomes for the organisation.		✓		✓		✓





Risk Dimension	Risks Pertinent to Each Dimension	Risk Definition	Secondary Dimensions Impacted by Risk	Lifecycle Stages Implicated				
				System Context and Design	Data Acquisition	Model Onboarding and Build	Deployment and Monitoring	Model Use and Output
 Ethics and Impact	Dark patterns	Generation of synthetically created deceptive or manipulative content that may trick or mislead users into taking certain actions without fully understanding the consequences (example, nudging children towards certain content or services).					✓	✓
	Toxic and offensive outputs	Outputs produced contain harmful, offensive, hateful, discriminatory, violent, racist, sexist or nudity-related information.	Legal and Regulatory Security and Access				✓	✓
 Accountability and Governance	Lack of Generative AI risk awareness	Insufficient education or reskilling resulting in undertrained resources lacking awareness of the unique risks involved with Generative AI.		✓	✓	✓	✓	✓


Risk Dimension	Risks Pertinent to Each Dimension	Risk Definition	Secondary Dimensions Impacted by Risk	Lifecycle Stages Implicated				
				System Context and Design	Data Acquisition	Model Onboarding and Build	Deployment and Monitoring	Model Use and Output
 Accountability and Governance	Lack of third-party accountability	Organisation has limited control or oversight over the development, modification and decision-making process for Generative AI models/services from third-party providers.				✓		
	Lack of use case and model governance	Failure to implement and enforce principles, guidelines, protocols and controls to proactively manage risks, and ensure traceability and responsibility in cases of undesirable outcomes.		✓	✓	✓	✓	✓
	Inadequate human oversight	Insufficient human-in-the-loop or oversight, limiting recourse to human correction or intervention in the event of a failure or when generating content with risk levels requiring human validation.					✓	✓




Risk Dimension	Risks Pertinent to Each Dimension	Risk Definition	Secondary Dimensions Impacted by Risk	Lifecycle Stages Implicated				
				System Context and Design	Data Acquisition	Model Onboarding and Build	Deployment and Monitoring	Model Use and Output
 Accountability and Governance	Inadequate feedback and recourse mechanisms	No mechanism to provide feedback or seek recourse for those impacted by harmful or biased outputs, and no consequence for the system's developers or owners for any negative outcomes.						✓
	Unclear output accuracy	The level of accuracy needed for the proposed Generative AI use case outcome is not clear and cannot be validated.				✓	✓	✓
 Transparency and Explainability	Unclear provenance for training/test data	The data used to train and test the model cannot be convincingly and comprehensively traced, presenting challenges for audit, disclosure, and potentially compliance, as well as posing the risk of the FI not having the right to use the data.	Legal and Regulatory		✓			


Risk Dimension	Risks Pertinent to Each Dimension	Risk Definition	Secondary Dimensions Impacted by Risk	Lifecycle Stages Implicated				
				System Context and Design	Data Acquisition	Model Onboarding and Build	Deployment and Monitoring	Model Use and Output
 Transparency and Explainability	Lack of explainability	Explanation unavailable for how the model works and derives outputs due to its black box nature.			✓	✓	✓	
	Anthropomorphism	The characteristic of Generative AI to mimic human characteristics in its output, enhancing the risk that users may find the outputs of Generative AI inappropriately convincing or may easily come under the impression that they are interacting with a human instead of a machine.	Security and Access				✓	✓
 Legal and Regulatory	Inability to ensure location compliance for model hosting and data processing	Inability to ensure adherence to FM hosting and data processing regulations that mandate the storage and processing of data within specific geographic boundaries or jurisdictions.		✓	✓	✓	✓	✓




Risk Dimension	Risks Pertinent to Each Dimension	Risk Definition	Secondary Dimensions Impacted by Risk	Lifecycle Stages Implicated				
				System Context and Design	Data Acquisition	Model Onboarding and Build	Deployment and Monitoring	Model Use and Output
 Legal and Regulatory	Unclear data ownership	Ownership of data used to train the Generative AI model and data created by the Generative AI model is unclear, leading to additional legal, commercial and privacy risks.		✓				
	Unauthorised data transfer and storage	Data is transported and stored on unauthorised systems as per the licensing terms or organisational policies.				✓		✓
	Breach or misalignment with regulatory or organisational standards	The model and its outputs fail to meet legal or regulatory requirements, organisational practices or values in how the business operates.	Fairness and Bias Ethics and Impact		✓	✓	✓	✓
	IP infringement	Data provided as input to a Generative AI system or product is used to create an output/content that violates IP rights owned by another individual, organisation, or entity.					✓	✓


Risk Dimension	Risks Pertinent to Each Dimension	Risk Definition	Secondary Dimensions Impacted by Risk	Lifecycle Stages Implicated				
				System Context and Design	Data Acquisition	Model Onboarding and Build	Deployment and Monitoring	Model Use and Output
 Legal and Regulatory	Unavailability of IP protection	The outputs of Generative AI built on FMs are not afforded IP protection such as copyright or trademarks due to a lack of legal clarity over IP protection for AI-generated content.						✓
	Inadequate privacy protection	Inadequate protection of or originally misclassified data that can result in the processing and use of personal or sensitive data, which lacks legal or ethical justification.	Fairness and Bias Ethics and Impact		✓	✓	✓	✓
	Unclear data retention and deletion	Lack of clarity on the policy around retention of personal, sensitive, or confidential data of data subjects.	Ethics and Impact		✓	✓		





Risk Dimension	Risks Pertinent to Each Dimension	Risk Definition	Secondary Dimensions Impacted by Risk	Lifecycle Stages Implicated				
				System Context and Design	Data Acquisition	Model Onboarding and Build	Deployment and Monitoring	Model Use and Output
 Monitoring and Stability	Hallucination/ Fabrication/ Confabulation	The models produce outputs that are not grounded on any source content or convincingly contradict the source content due to lack of understanding of real-world views. This can have an adverse impact on social groups or may constitute grounds for libel. They may also misinform, mislead, or negatively impact users and reduce user or public faith in the reliability of AI systems.	Ethics and Impact Legal and Regulatory Fairness and Bias					✓
	Overconfidence	The characteristic of Generative AI models to produce convincing outputs that do not properly account for the complexity, uncertainty, or contradiction in their sources. This leads to the potential to present false information as factual, or uncertain information as clear. Presenting this information in such a way interferes with the ability of users to review using their judgement.	Fairness and Bias Transparency and Explainability				✓	✓


Risk Dimension	Risks Pertinent to Each Dimension	Risk Definition	Secondary Dimensions Impacted by Risk	Lifecycle Stages Implicated				
				System Context and Design	Data Acquisition	Model Onboarding and Build	Deployment and Monitoring	Model Use and Output
 Monitoring and Stability	Training data or inputs not fit for purpose	Training data used in model are not representative of the geographical and cultural context where the model will be used or not aligned to the system's intended goal, leading to incorrect outputs or conclusion.	Fairness and Bias			✓		
	Lack of continuous monitoring	Absence of ongoing and systematic surveillance on how Generative AI systems are performing, how they are utilised, and on various parties to ensure they are in accordance with intended purposes, ethical guidelines and regulatory requirements.				✓	✓	✓
	Insufficient data quality	Low-quality or noisy data used for training could result in poor model performance, increased debugging efforts and higher development costs.			✓	✓		




Risk Dimension	Risks Pertinent to Each Dimension	Risk Definition	Secondary Dimensions Impacted by Risk	Lifecycle Stages Implicated				
				System Context and Design	Data Acquisition	Model Onboarding and Build	Deployment and Monitoring	Model Use and Output
 Monitoring and Stability	Model staleness	Data used to train the model becomes outdated and irrelevant due to changes in its statistical properties over time, leading to the model developing ingrained biases, reduced accuracy and performance.				✓		
	Insufficient model accuracy/ Soundness	The model outputs are inaccurate or does not meet the performance thresholds required to ensure fit for purpose.			✓	✓	✓	
	Model degradation from unexpected use	A wider range of unexpected usage patterns due to the broad capabilities of generative models create outcome instability or unexpected failure modes.			✓	✓	✓	
	Inadequate operational resilience	Operational resilience or service continuity plans increase in complexity due to the broad set of services and capabilities of Generative AI.				✓	✓	


Risk Dimension	Risks Pertinent to Each Dimension	Risk Definition	Secondary Dimensions Impacted by Risk	Lifecycle Stages Implicated				
				System Context and Design	Data Acquisition	Model Onboarding and Build	Deployment and Monitoring	Model Use and Output
 Monitoring and Stability	Unmet architectural requirements	Requirements to govern the model and its outputs based on its deployment pattern (e.g., on prem, cloud etc.) are unmet due to technology, cost or people constraints.			✓	✓	✓	
	Unintentional, inappropriate or illegal use	Consumers or employees use Generative AI for inappropriate or illegal activities unintentionally with liability remaining with the FI.				✓	✓	
 Cyber and Data Security	Data poisoning	Deliberate manipulation of the model by a malicious actor, either through the introduction of malicious data at the point of initial training or during the course of use. This can lead to security vulnerabilities or inaccurate and harmful outputs.		✓	✓	✓	✓	



Risk Dimension	Risks Pertinent to Each Dimension	Risk Definition	Secondary Dimensions Impacted by Risk	Lifecycle Stages Implicated				
				System Context and Design	Data Acquisition	Model Onboarding and Build	Deployment and Monitoring	Model Use and Output
 Cyber and Data Security	Adversarial model manipulation	Deliberate manipulation of a Generative AI system's behaviour by a malicious party with access to its FM. This can lead to undesirable or unpredictable behaviour, including inaccurate or harmful outputs.				✓	✓	
	Prompt injection	The use of carefully designed prompts to encourage a Generative AI system to circumvent its programmed guardrails or filters. This type of attack, if successful, allows malicious actors to generate content that an FI explicitly sought to disallow. Prompt injection attacks designed to reveal sensitive or confidential information fall under "model inference attacks" below.						

Risk Dimension	Risks Pertinent to Each Dimension	Risk Definition	Secondary Dimensions Impacted by Risk	Lifecycle Stages Implicated				
				System Context and Design	Data Acquisition	Model Onboarding and Build	Deployment and Monitoring	Model Use and Output
 Cyber and Data Security	Re-identification	Possibility of de-identified records/ data being able to be re-identified mostly with malicious intent. This risk is related to “model inference attacks” (below) but is distinct in that it refers to data released in the normal course of operations, whereas model inference attacks imply the use of deliberately designed inputs.					✓	✓
	Data leakage	Model outputs or the model development/ training/fine-tuning process inadvertently reveal sensitive, confidential or personal data to an unauthorised user. This can occur unwittingly – when innocuous prompts produce sensitive outputs – or through prompt injection, where malicious prompts deliberately seek to evade controls and force the release of sensitive information.			✓	✓	✓	✓



Risk Dimension	Risks Pertinent to Each Dimension	Risk Definition	Secondary Dimensions Impacted by Risk	Lifecycle Stages Implicated				
				System Context and Design	Data Acquisition	Model Onboarding and Build	Deployment and Monitoring	Model Use and Output
 Cyber and Data Security	Model inference attacks	Inference attacks including submitting carefully crafted input and analysing the corresponding output to reveal the membership, attributes or features about individuals in the training datasets increase in severity due to model's ability to respond to natural language prompts and the fact that Generative AI models often have larger attack surfaces.				✓	✓	

B.2: Implications of Select Risks to FEAT

Fairness-Related Risks Amplified or Introduced by Generative AI

Generative AI introduces a unique set of challenges, notably in the realm of fairness. Fairness-related risks increase with the deployment of Generative AI systems due to their inherent characteristics and capabilities. These systems, often trained on vast datasets, inherit biases present in the data, which can inadvertently perpetuate and even amplify existing social biases. The generative nature of AI means it can output content, such as text, images, and videos, that may reinforce stereotypes or discriminate against certain groups, thus perpetuating social inequalities. While these risks are present in all AIDA systems, they are amplified by Generative AI.

It is essential to address these risks to fulfil the main objective of the FEAT Principles. The FEAT Principles primarily aim to protect users of AI systems from unintentional harm, particularly harms associated with systemic disadvantage, bias, or the propagation of disadvantage to vulnerable groups in society. These principles are closely aligned with fairness concerns arising from Generative AI. This section covers fairness-related risks, assesses the adequacy of the fairness principles and assessment questions in addressing fairness-related challenges arising from Generative AI.

Table B.2.1 presents a high-level list of risks and explains their relevance to Generative AI.

Table B.2.1: Inventory of key risks related to fairness

Risks Related to Fairness and Bias	Implication for Generative AI
<p>Unrepresentative or biased data inputs: data are biased against, or unevenly represent, certain individuals or groups of individuals, which can produce biased model outputs.</p>	<p>While this risk is present in all AIDA-driven decisions, it is heightened in Generative AI systems because of (i) the significantly greater breadth and depth of data used to train the systems and (ii) the wider range of applications for these systems in the delivery of financial services and products, resulting in greater potential for Generative AI-based decisions to impact individuals or groups negatively.</p> <p>If input data is affected by ingrained bias against individuals or groups, or are not sufficiently representative of the full spectrum of people accessing relevant financial services and products, there is a risk of systemically disadvantaging those individuals or groups.</p>



Risks Related to Fairness and Bias	Implication for Generative AI
	<p>Examples of the types of bias that may affect training data include:</p> <ul style="list-style-type: none">- Cultural bias: this could arise if the language, text, pictures, videos and other data that have been used to train the Generative AI system are biased against certain cultures, nationalities, religions or other social groups, for example by portraying them in an unfavourable light or stereotyping them into particular roles.- Gender bias: where the data favour or disfavour, or assign particular (negative) characteristics to, a particular gender.- Organisational bias: if the majority of the data used to train Generative AI systems are collected from the internet, those systems could reflect the agendas of major corporations, governments and other organisations with the means to generate online content to the exclusion of other perspectives.- Temporal/historical bias: where perspectives from a particular time are favoured or disfavoured.- Political bias: where politicised viewpoints or opinions in the training data may unduly emphasise a predominant political ideology, resulting in overwhelmingly partisan output. Generated content may therefore exclude other valid viewpoints that contradict the predominant ideology. <p>These biases can build on each other and create a systemic risk – so-called “model collapse”. As more models are derived from other (base) models, and more data on which models are trained are themselves derived from outputs of Generative AI systems, systemic risk is introduced since bias in the base models can be propagated or inherited by descendant models.</p>

Risks Related to Fairness and Bias	Implication for Generative AI
	<p>The quality of training data also has the potential to impact the fairness of decisions made by Generative AI systems. Content can be posted on the internet by anyone, often without fact checking or other verification, and such content may be inaccurate, misleading, or harmful. If developers of Generative AI systems collect large quantities of data from the internet, without undertaking proper due diligence on the data and their sources, it may include such harmful content. Systems trained on low-quality data risk creating unfair outcomes.</p>
<p>Adverse or inappropriate impact to individuals and groups: models generate outputs that can be detrimental or inappropriate for individuals or groups.</p>	<p>Examples of these types of risks are:</p> <ul style="list-style-type: none"> - Personal data protection risk: given the manner in which Generative AI systems ingest large amounts of data (potentially including personal data), it may be difficult to ensure that such data are gathered and processed fairly, and that such data processing does not lead to systemic disadvantage and discrimination or other risks. Due to the opacity of Generative AI systems and the data on which they are trained, it may be difficult for FIs to be confident that individuals will not be treated unfairly or disadvantaged, or that their data will not be at risk (such as through unauthorised disclosure or data breach), when it is not clear how to ascertain if their data has been used in the development of the system. This could result in individuals being harmed without justification such as in cases where personal data was collected without their consent or notice. - Toxicity: Generative AI may produce harmful, offensive or malicious content if trained on large amounts of data from the internet which include toxic content. Here, the term toxicity includes discrimination, hate speech, and abusive pictures or language. This is a fairness-related risk since such toxic output could systemically disadvantage groups of individuals by reinforcing prejudices against them.



Ethics and Accountability-Related Risks Amplified or Introduced by Generative AI

In light of the transformative and powerful potential of Generative AI, it is important to give due attention to ethical risks and concerns posed by the technology. One primary ethical concern associated with Generative AI lies in its potential to produce highly convincing deepfake content which can lead to misinformation. As defined in FEAT, ethics is “the work performed to specify and satisfy values, according to normative content, through governance”. Ethical risks are therefore regarded as risks that limit, prevent or challenge an FI’s ability to satisfy its ethical standards, values and codes of conduct in its deployment of Generative AI or its ability to ensure that Generative AI-driven decision-making is held to the same ethical standards as human-driven decisions.

Accountability is the vital link that operationalises an FI’s values and ethical standards in a sustainable and demonstrable way. In this context, it refers to the state of being responsible to internal and external stakeholders for outcomes, including actions, products, services, or decisions, driven by Generative AI models, and in responding to potentially harmful outcomes.

Table B.2.2 and Table B.2.3 present a list of high-level risks related to ethics and accountability, respectively, and explain their specific relevance to Generative AI.

Table B.2.2: Inventory of key risks related to ethics

Risks Related to Ethics and Impact	Implication for Generative AI
<p>Value misalignment: failure to incorporate organisational principles in the design and operation of Generative AI.</p>	<p>This risk occurs as a result of a gap or multiple gaps along the process of organisational value identification, communication or application, which could lead to the AI system generating misaligned outputs or taking unexpected and undesirable routes to achieve a designated objective. This is distinct from other ethical risks because it encompasses the full range of issues stemming from any direct or indirect violation of an FI’s values or desired social impact, leading to various negative outcomes for the bank, its customers and broader society.</p> <p>Though this risk is shared with non-Generative AI systems, it is exacerbated by the characteristic of Generative AI systems to involve foundation models, for which FIs may lack visibility and control over development and data inputs. The implication is that by default FIs are less capable of performing the work required to ensure the design and development process</p>

Risks Related to Ethics and Impact	Implication for Generative AI
	<p>of the Generative AI system reflects its own values, including managing the issue of “Do as I say not as I do”, for instance where bias in training data inherently contradicts the organisation’s values. Specifically for FIs, where there is a duty of care to ensure decisions are taken with proper deliberation to advance customers’ interests and organisational values, being able to explain how an output was achieved is important, and the opaque nature of Generative AI foundation models inherently complicates that.</p> <p>Additionally, the mutability of generated output and lack of autonomous reasoning or self-assessment mean that, on its own, the Generative AI system has no mechanism to measure its output against an organisation’s values, and would be just as likely to generate convincing content that contradicts the organisation’s values, even fabricating the rationale for doing so, as it would generate content that is ethically sound. FIs will have more levers to control parameters and data for non-Generative AI models, and those developed in-house, where objective rules or ground truth could be codified and used by the model to assess against (for example, refer to the Veritas Toolkit v2.0).</p>
<p>Dark patterns: deceptive or manipulative content or user interfaces trick or mislead users into taking certain actions without fully understanding the consequences.</p>	<p>Dark patterns are an example of value misalignment, in which the process an AI system takes to achieve an intended output is harmful or discordant with an organisation’s values. This includes methods to achieve outcomes in a manner that may not reflect the organisation’s values, such as through behavioural manipulation.</p>
<p>Toxic and offensive outputs: malicious or harmful patterns are propagated in results.</p>	<p>This risk refers to the presence of harmful or malicious content in data or AI-generated responses, which is an important consideration for Generative AI systems, where content generated is unique and mutable. The model may produce inappropriate or offensive content, for example in generating customer service responses, damaging the FI’s reputation and customer</p>



Risks Related to Ethics and Impact	Implication for Generative AI
	relationships. Organisations will need to evaluate whether they are able to perform the work required to ensure that the model and its outputs remain aligned with their values, detect transgressions and address the root cause of those issues in a Generative AI context where they may be dependent on third-party vendors.
<p>Environmental sustainability impact: FIs' ESG commitments are compromised by Generative AI.</p>	<p>This risk refers to the potential negative externalities that occur particularly during the development of Generative AI foundation models and, to some extent, in the decisions an organisation takes to procure and deploy Generative AI systems. For example, significant use of natural resources for the physical hardware to create and run Generative AI foundation models can put an organisation's environmental commitments at risk. The significant energy requirements to train a foundation model can impact CO2 emissions in cases where clean energy is not available. These organisations will need to conduct due diligence on Generative AI vendors to assess whether such externalities are present and material, as ESG risks are readily obfuscated through complex supply chains.</p>

Table B.2.3: Inventory of key risks related to accountability

Risks Related to Accountability and Governance	Implication for Generative AI
<p>Third-party accountability: FIs do not have control over important characteristics of their Generative AI system.</p>	<p>Typically, due to cost and effort required to build foundation models used in Generative AI, they are developed by established technology companies with the scale and resources to engage in such work. These companies make decisions including which data sources to use for model pre-training, levels of representation required, which filters and corresponding thresholds to "cleanse" unacceptable or undesirable data, and the extent of human supervision and intervention as opposed to automated techniques. The model may then be fine-tuned to a particular user group or purpose by the developer, deploying organisation, or both, through the ingestion of more relevant data and supervised</p>

Risks Related to Accountability and Governance	Implication for Generative AI
	<p>machine learning techniques such as supervised fine-tuning, reward modelling and reinforcement learning.</p> <p>Once the FI procures the technology, it becomes responsible for the rollout to its staff and customers, including the provision of guidance, policies and training, governance and monitoring. Internally, a chain of accountability flows from users of the technology issuing prompts to the FI's senior management. Depending on the characteristics of the setup, it is dependent on the technology provider for hosting the models and prompts, and maintenance of updates to the model.</p> <p>The design decisions and human supervision along this development and procurement process become critical to the model's final form, contingent on the knowledge, skills and intentions of the parties involved. Given that a significant number of important decisions may be made independently by a third party, for which the FI may not have the authority to dictate its requirements, the issue is then the availability and reliability of information and attestation for FIs to be comfortable making decisions to acquire, deploy and assume accountability for the Generative AI model.</p> <p>Then, returning to the essence of accountability, should there be a negative outcome for the bank, its customers, or broader society, the initial challenge may be to determine the root cause – did this occur as a result of the model itself, a prompt, or both? This is important not only for remediation, prevention and reparation, but also for ownership and accountability. Understanding and agreeing on the lines and layers of accountability between FIs and third-party vendors with regard to new or amplified risks of Generative AI is a key factor without which risk exposure for the FI may be too significant to tolerate.</p> <p>To summarise, FIs may have limited control or oversight over the development and decision-making process</p>



Risks Related to Accountability and Governance	Implication for Generative AI
	<p>for Generative AI systems that are wholly or partially produced by third-party providers. If FIs are to be held accountable for decisions made regarding the Generative AI model and be held responsible for how the model is deployed and utilised, FIs must be able to control, validate and explain the model, including data sources used. If FIs assume accountability without being able to control parts of the process that led to the Generative AI model’s outcomes, they become exposed to risks for outputs they are not directly positioned to prevent in spite of internal governance and surveillance, particularly when one considers known challenges in detecting risks such as confabulations.</p>
<p>Internal accountability: failure to implement and enforce principles, guidelines, protocols and controls to proactively manage risks, and ensure traceability and responsibility in cases of undesirable outcomes.</p>	<p>This refers to the risk faced when the chain of accountability/ownership within FIs is not clearly defined during the design stage. The FI may not be able to identify the chain of authority due to internal stakeholders’ lack of accountability in aspects of the system that are out of their control.</p> <p>Generative AI is a nascent and emerging technology, for which organisations are still developing an understanding of its risks and impacts. However, adoption of the technology is expanding rapidly and employees may be experimenting independently. Though FIs have established risk management practices, existing policies, governance structures, protocols, controls and guidelines, they may not be able to adequately manage new or amplified risks. Similarly, mechanisms to monitor compliance and prevent incidents may not be tailored to Generative AI scenarios. This can include decisions on permissible use cases, materiality assessments, and establishment of internal authorities for review and approval. Even if these mechanisms are in place, characteristics of Generative AI such as its black box nature can also preclude or challenge FIs’ ability to fully address and reduce risks, and provide complete decision lineage.</p>
<p>Feedback and recourse mechanism: those harmed by the consequences of the system are not provided recourse.</p>	<p>In the event of an incident (e.g., data privacy violation), the FI may be limited in its capacity to fix any issues or provide reparations. The FI may not be able to adjust foundation models based on a particular data subject’s</p>

Risks Related to Accountability and Governance	Implication for Generative AI
	<p>request, and the model may continue to generate content that is counter to the data subject's rights.</p> <p>This risk is amplified by particular challenges of Generative AI, where violations resulting from a foundation model's behaviour can result in significant technical workarounds or changes to the model itself. FIs may not have access to a model sourced from a vendor, and may lack the skills or computing power to modify a model that they do have access to.</p>
<p>Generative AI risk awareness: insufficient education or reskilling efforts results in undertrained resources lacking awareness of risks involved.</p>	<p>While certain applications of Generative AI are well publicised and FI staff may be aware of the technology, considerations around its usage and implications may not be adequately communicated across relevant internal stakeholders. This includes decision makers, existing internal authorities, and lines of risk management not limited to end users. This can be in part due to the rapid emergence of the technology, lack of internal or external expertise, and evolving risk and legal landscape.</p>
<p>Human oversight: sufficient human-in-the-loop or oversight is not available or possible at scale.</p>	<p>Unlike traditional rules and statistical models, where there are ground truths and expected results, Generative AI outputs can vary based on prompt, context, phrasing, and other inputs. This can occur even when identical prompts are provided. This characteristic is a key rationale for maintaining human oversight over models and their outputs, for example in verifying accuracy, compliance or ethics, and restricting or eliminating use cases where the models act or decide independently. However, maintaining human oversight can be challenging. If users are required to validate most if not all model outputs, time and effort spent in validation may reduce or negate expected productivity gains, rendering the benefits of adopting the technology untenable. Organisations may not have sufficient domain experts to effectively oversee all possible applications of the technology, particularly given that general purpose models can be applied across multiple use cases and scenarios.</p>



Transparency-Related Risks Amplified or Introduced by Generative AI

One of the primary objectives of the set of principles published under FEAT is to “build public confidence and trust in the use of AIDA”. In order to build confidence and trust, FIs should understand how AI systems work, what data is used to drive outcomes, and why a particular outcome materialised. FEAT also advocates (with a few exclusions) for FIs to communicate the use of AIDA to customers, including data used, how outcomes/decisions impact them, providing redress/recourse channels to impacted data subjects where unfavourable outcomes can be reviewed and remediated. The transparency principles under FEAT guide this direction.

The ability to explain how AI works and why it makes the decisions it does – “explainability” – is a key enabler for implementing the transparency principles. Generative AI, unlike traditional AI, generates new content autonomously including but not limited to text, images, voice and videos. Generative AI systems are inherently complex, therefore explainability is a challenge especially when FIs use third-party foundation models. Explainability being a key enabler for transparency, the challenge extends to an FI’s ability to achieve transparency with its internal and external stakeholders on the use of Generative AI.

Table B.2.4 presents a high-level list of risks related to transparency and explainability and explains their relevance to Generative AI.

Table B.2.4: Inventory of key risks related to transparency

Risk Related to Transparency and Explainability	Implication for Generative AI
<p>Overconfidence: untrue information is presented as if factual, referred to as hallucination or fabrication.</p>	<p>One of the top risks posed by Generative AI is its ability to hallucinate or fabricate information that appears genuine and authentic but is in fact entirely fictitious.</p> <p>The use of such fabricated outputs produced by Generative AI within FIs could be minimal, but they undermine public confidence in Generative AI and AI in general.</p> <p>FIs may need to enhance their transparency practices including establishing appropriate guardrails to manage the risk and impact to data subjects and the public in general to build and sustain their confidence in the new technology.</p>

Risk Related to Transparency and Explainability	Implication for Generative AI
<p>Anthropomorphism: a Generative AI system is mistaken for human.</p>	<p>Anthropomorphism is defined as “attribution of human traits, emotions or intention to non-human entities” (in this case, AI or a machine). Generative AI is capable of interacting using natural language, and its capability to have a seemingly “real” conversation could lead users to lose sight of the fact that they are interacting with a machine. Institutions will have to find the right balance between providing personalised services and making machine interactions distinct from human interactions.</p> <p>Anthropomorphism of the outputs of Generative AI could, in combination with other risk factors like hallucination, pose severe challenges (e.g., potential overreliance on AI, psychological attachment, and wellbeing concerns) and actively undermine public trust in AI.</p> <p>In such cases, FIs should adopt a more proactive approach to transparency. For e.g., in a use case involving chatbot interactions, FIs should proactively communicate that interactions do not involve humans and to refrain from attaching human-like names or images to AI.</p>
<p>Output accuracy: the level of accuracy needed for the proposed Generative AI use case outcome is not clear and cannot be validated.</p>	<p>Explainability of the model is one of the key enablers for achieving transparency with both internal and external stakeholders.</p> <p>Having appropriate explanation methods, accuracy standards and associated thresholds is essential to manage model explainability. Currently, examples of Generative AI use cases are mostly employing unstructured data like natural language, images or videos, where accuracy of newly generated content is difficult to define due to the absence of a comparable baseline, let alone threshold. The lack of a threshold has a direct impact on our ability to evaluate output and provide transparency around it.</p>



Risk Related to Transparency and Explainability	Implication for Generative AI
<p>Provenance for training/test data: data used to train and test models cannot be traced, or its lineage is unestablished.</p>	<p>Transparency principle 13 requires FIs to provide data subjects with explanations on data used in decision-making.</p> <p>In order to improve confidence in such explanations, it is important to know where data comes from. Data provenance has an impact on an FI's ability to be transparent.</p> <p>For e.g., if an FI used information from their client's social media as one of the inputs to assess their risk profile and offer investment advice, the FI should know what drove the risk profile along with where the underlying data came from. Did the recommendation to include crypto assets arise because the customer indicated their preference while submitting information to the bank or was it because the risk profile had been updated based on "liking" crypto-related articles on social media?</p> <p>Knowing the provenance of training data is important. When training is performed with data of unclear provenance, the quality and accuracy/correctness of the data cannot be established. This eventually affects model output.</p> <p>This risk is amplified in Generative AI, as models such as GPTx use publicly available data from the internet. Any model that uses GPTx will inherit this risk.</p> <p>This can also have an impact on model explainability, a key aspect of transparency.</p>
<p>Model and output explainability: complex models have limited traceability of features, which can influence output and how these outputs are derived from its black box nature.</p>	<p>Generative AI models are complex by design and are driven mostly by unstructured data like natural language, voice and video, as opposed to tabular data that is commonly used.</p> <p>Black box risks inherent in AI models are amplified in Generative AI.</p>

B.3: Risk Assessments with Two Additional Sample Use Cases

Use Case 1 – Generative AI as Co-Pilot for Customer Service Teams

- **Use Case Description:** Customer service officers spend approximately 30–35% of their time per call to check information from various sources and log service request along with the call summary. Generative AI capabilities can be used for productivity gains by acting as a co-pilot to the officers in checking information for customer queries and logging query details, thereby reducing average call handling times and improving the quality of responses to customers.
- **Generative AI Capabilities Used:** Text, speech and video for content summarisation, query response.
- **Deployment Pattern:** Third-party AI assistant powered by Generative AI and organisation’s internal Generative AI suite of applications.
- **Impacted Parties:** Customer service personnel, customers, organisation.
- **Use Case Materiality:** Materiality is considered as ‘medium,’ as the use case can have undesirable customer impact based on the accuracy of outputs generated, only if the human-in-the-loop does not function as expected. These can further lead to loss of customer satisfaction and reputational damage. The severity and probability of such an impact is considered low as the process is also governed by existing customer service processes and requirements.
- **Illustrative Sample of Risks Identified Across the Lifecycle:**

Risks to consider during System Context & Design	Risks to consider during Data Acquisition	Risks to consider during Model Onboarding & Build	Risks to consider during Deployment & Monitoring	Risks to consider during Model Use & Output
Considerations around workforce implications	Risks: <ul style="list-style-type: none"> • Data poisoning 	Risks: <ul style="list-style-type: none"> • Insufficient model accuracy/soundness • Adversarial model manipulation 	Risks: <ul style="list-style-type: none"> • Inadequate Operational Resilience • Data leakage • Adversarial model manipulation 	Risks: <ul style="list-style-type: none"> • Hallucination/ Fabrication/ Confabulation • Insufficient model accuracy/soundness

Figure B.3.1: Illustrative sample of risks identified in use case 1



Use Case 2 – Generative AI for Code Generation

- **Use Case Description:** To use Generative AI for productivity gains by helping in code generation or completion by suggesting individual lines or whole functions instantly, and debugging of errors by drawing context from comments.
- **Generative AI Capability Used:** Code generation, revision or debugging.
- **Deployment Pattern:** Closed source third-party AI code generators, and third- party code hosting platform.
- **Impacted Parties:** Developers.
- **Use Case Materiality:** Materiality is considered as ‘medium’. This is a code generation use case with the severity and probability of any adverse impact offset by thorough human review and adherence to the organisation’s SDLC requirements. The developers can edit the code as per their requirements before it is submitted for further review. Other aspects of materiality such as the use of personal data, options for recourse, etc., are not applicable for this use case. Nevertheless, caution is advised due to the emerging nature of Generative AI and lack of awareness of its limitations.
- **Illustrative Sample of Risks Identified Across Identified Across the Lifecycle:**

Risks to consider during System Context & Design	Risks to consider during Data Acquisition	Risks to consider during Model Onboarding & Build	Risks to consider during Deployment & Monitoring	Risks to consider during Model Use & Output
<p>Considerations around workforce implications</p>	<p>Risks:</p> <ul style="list-style-type: none"> • Training data or inputs not fit for purpose 	<p>Risks:</p> <ul style="list-style-type: none"> • Unclear Data Ownership • IP infringement • Lack of 3rd party accountability • Unintentional Inappropriate or illegal use 	<p>Risks:</p> <ul style="list-style-type: none"> • Data leakage • Unintentional Inappropriate or illegal use 	<p>Risks:</p> <ul style="list-style-type: none"> • Hallucination/ Fabrication/ Confabulation • Unintentional Inappropriate or illegal use

Figure B.3.2: Illustrative sample of risks identified in use case 2

The above two use cases utilise quite different Generative AI capabilities and content types. Materiality is accorded based on any adverse impact it has on customers, employees, or the organisation. The environmental impact is not calculated due to the unavailability of organisational capabilities at this point. While the above risks are identified by specific use case teams responsible for the development and delivery of the use case, cross-functional teams comprising analytics, technology and various control/risk functions are still needed to review risks at all stages and the appropriate guardrails and risk controls established across the entire spectrum of risks.



C: Risk Assessment of the Veritas Methodology

This section evaluates the Veritas Methodology to assess the gaps and limitations of existing questions in adequately addressing Generative AI-related risks. The focus is to only include questions that have limitations in Generative AI coverage. The methodology is discussed in detail in MAS' 2022 publication: FEAT Principles Assessment Methodology.

Evaluation of Fairness-Related Risks Against Veritas Methodology

In Veritas Phase 2, the FEAT Assessment Methodology documents provided a comprehensive checklist that spans the entire lifecycle of the AIDA framework. This checklist serves as an advisory tool for FIs to navigate the FEAT assessment process for traditional AIDA systems, and provides considerations to assist FIs in responding to assessment questions. Within the checklist, a set of 12 questions specifically address fairness considerations (as shown in Figure C.1).

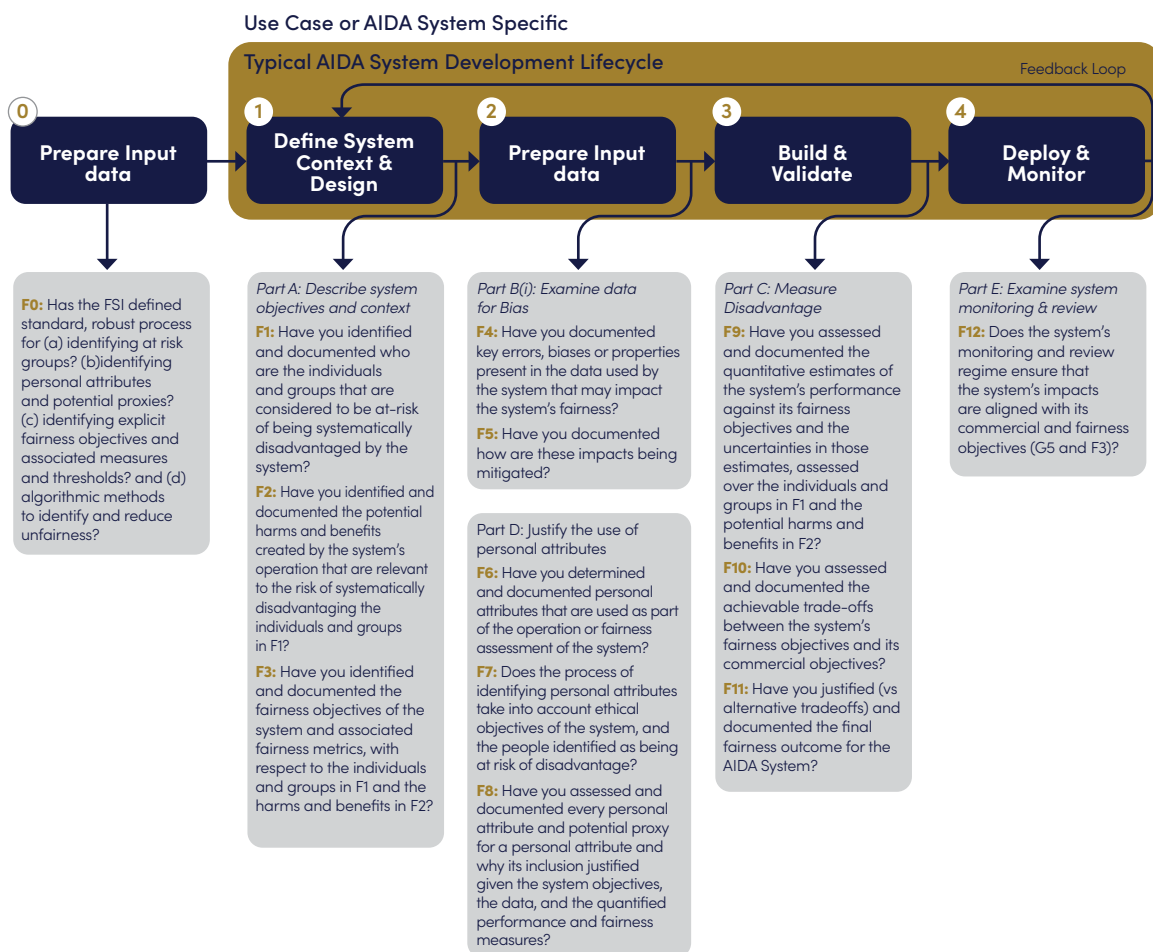


Figure C.1: Fairness checkpoints for AIDA systems

While the fairness checklist questions are still applicable for assessing fairness, these high-level analyses offer ample opportunities for further evaluation of the Veritas Methodology on fairness.

(F2) Have you identified and documented the potential harms and benefits created by the system's operation that are relevant to the risk of systematically disadvantaging the individuals and groups in F1?

Where the FI does not own or provide data used to train a Generative AI system, it may be difficult to identify individuals and groups that are at risk of being systemically disadvantaged by characteristics of the underlying dataset. In addition, the ability of Generative AI to create new content makes it more difficult to identify individuals and groups who could be systemically disadvantaged by the outputs of the system.

The same characteristics of Generative AI systems made it difficult to identify the harms and benefits relevant to this potential systemic disadvantage.

(F3) Have you identified and documented the fairness objectives of the system and associated fairness metrics, with respect to the individuals and groups in F1 and the harms and benefits in F2?

Since Generative AI systems are relatively new, a lot of fairness metrics cannot be accurately calculated, such as fairness metrics in image generation categories. Therefore, the practitioner may propose fairness objectives, but they may not be able to actually calculate fairness metrics at this point.

(F4) Have you documented key errors, biases or properties present in the data used by the system that may impact the system's fairness?

If FIs develop Generative AI in-house, they will be able to identify the types of bias that are present in the training data. However, if they buy a Generative AI system from a third party, it will be difficult to know what types of bias are embedded in the training data, e.g., temporal bias, confirmation bias, etc. Therefore, FIs are encouraged to evaluate the system throughout the fine-tuning process to ensure the system produces appropriate, non-offensive content.

(F5) Have you documented how are these impacts being mitigated?

If bias is present in the training data and there is no such technical toolkit to remove the bias risk, we may not be able to remove it from the model's lifecycle (i.e., input, model, and output).



(F6) Have you determined and documented personal attributes that are used as part of the operation or fairness assessment of the system?

If FIs are buying Generative AI systems from a third party, they may fine-tune the system with their curated set of validation data to mitigate any fairness risks and achieve their needs. However, if FIs are developing the Generative AI system in-house, and not using a third-party system as a base, then AI practitioners have full control on data, and they can determine which personal attributes should be used in the decision-making process. Nevertheless, appropriate explainability assessments should be performed to ensure that outputs of the system and the use of personal attributes to drive those outputs are understandable.

(F7) Does the process of identifying personal attributes take into account ethical objectives of the system, and the people identified as being at risk of disadvantage?

AI practitioners must consider ethical principles and guardrails of their FI when identifying personal attributes in the training set.

(F8) Have you assessed and documented every personal attribute and potential proxy for a personal attribute and why its inclusion is justified given the system objectives, the data, and the quantified performance and fairness measures?

There are three scenarios to consider:

- 1) If the Generative AI use case is low-risk and does not have any kind of bias or fairness risk, then there is no need for this checkpoint.
- 2) If the Generative AI system has been developed fully in-house, so that the FI has control and oversight of the data used to train it, but the proposed use case has some bias or fairness-related risks, then a technical toolkit shall be provided to the practitioner to assess the impact of personal attributes on the system's performance, fairness objectives and trade-offs (e.g., the challenge at present is that the Veritas toolkit cannot be utilised to assess fairness outcomes of Generative AI systems).
- 3) The most challenging scenario is where the FI has bought a third-party system. In that case, in addition to technical and explainability assessments, FIs are recommended to obtain a transparency report on training data from the third party to demonstrate their due diligence.

(F11) Have you justified and documented why the fairness outcomes observed in the system are preferable to these alternative trade-offs?

If the AI practitioner is able to measure and assess the fairness outcome in F9 and F10, then they should be able to justify and document the results. However, if they are unable to measure and assess F6 and F7, then it will be difficult to give any justification.

Evaluation of Ethics and Accountability-Related Risks Against Veritas Methodology

In Veritas Phase 2, the FEAT Ethics and Accountability Principles Assessment and Methodology, including the Ethics and Accountability Framework (as shown in Figure C.2) and accompanying workbook Operationalising Ethics and Accountability, helps FIs govern ethical decision-making and generate internal and external accountability in their AIDA practices. The framework, in particular, helps organisations adhere to their core values and specific commitments around a particular AIDA use case. This framework is agnostic to the type of AIDA deployment and would be equally meaningful and useful to firms assessing the use of Generative AI models in their organisations.

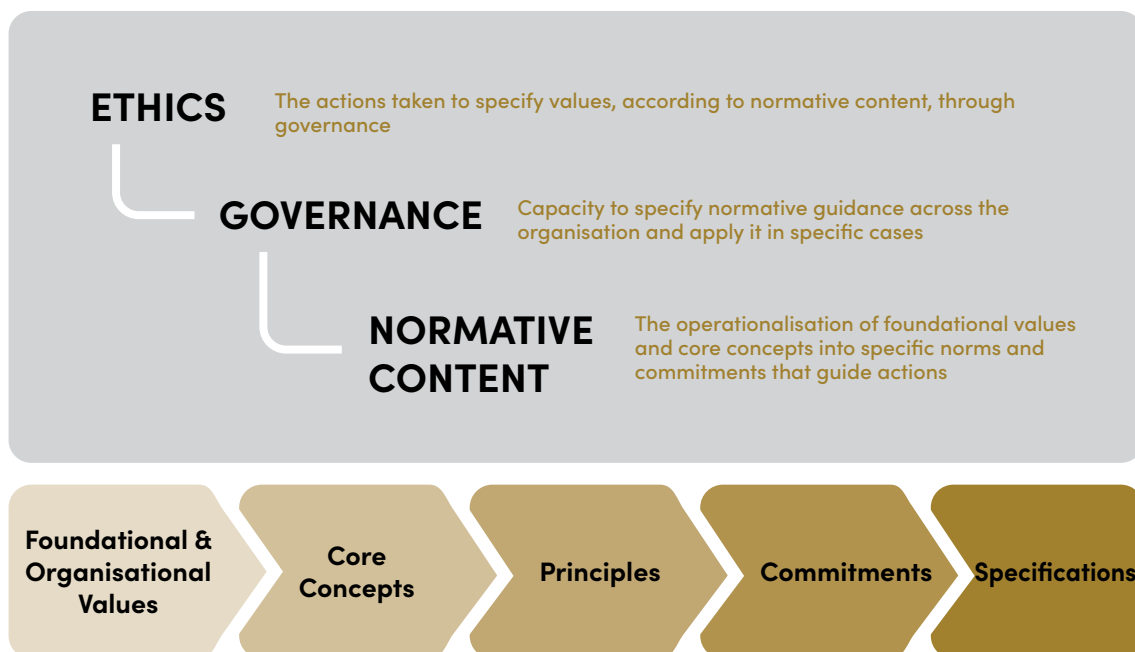


Figure C.2: Ethics and Accountability Framework

However, as is expected as part of the framework, Generative AI use cases may raise ethical and accountability risks not currently covered by the organisation's core concepts, principles or commitments, and may require the addition of new commitments, specifications, and possibly principles or core concepts. The predominant characteristics of Generative AI such as the closed-box nature of foundation models also require specific commitments to be adapted to suit the realities of Generative AI models and may also illustrate possible gaps where commitments are harder to achieve. This brings into question whether the use case will sufficiently meet the organisation's ethics and accountability expectations to be implementable.



Due to the framework’s holistic nature and wide applicability, the focus of this evaluation centres around assessing and expanding the Veritas Methodology checklist questions to encompass considerations which are amplified by Generative AI use cases. Within the checklist, a set of eight questions specifically address ethics and accountability considerations (as shown in Figure C.3).

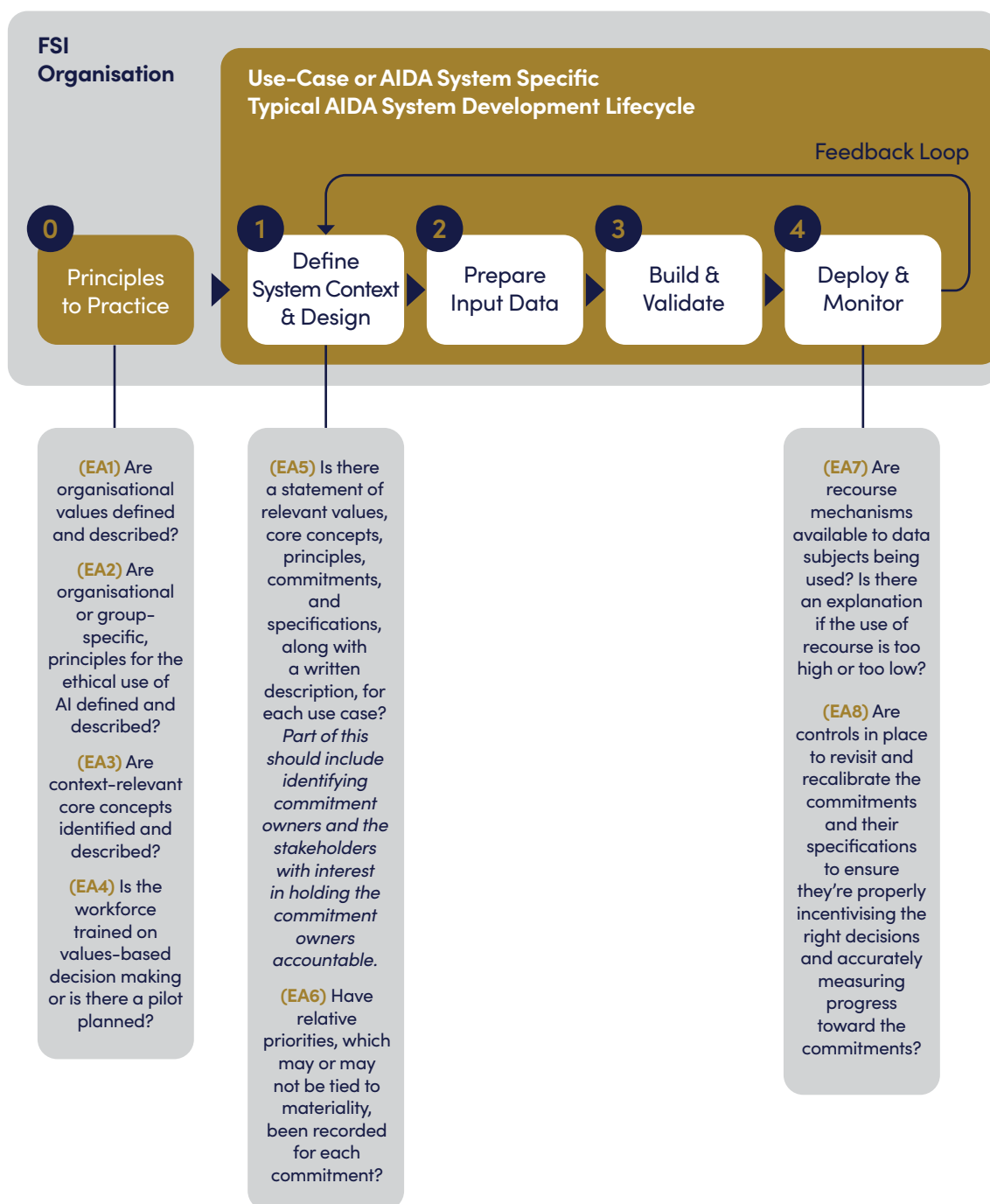


Figure C.3: Ethics and accountability checkpoints for AIDA systems

The following part of this section evaluates select checklist questions for ethics and accountability for their ability to comprehensively address risk implications identified in Section 2, individually as well as in totality.

(EA1) Are organisational values defined and described?

Organisational values play a crucial role in guiding AI development in alignment with the bank's mission and principles, and when successfully embedded in company culture, reduce risks AIDA applications have on the organisation's values or ethics.

Recommendation:

However, with the rapid transformation, evolution and emergence of new technologies such as Generative AI, there should be processes in place to periodically assess whether values adequately guide organisations in navigating against new or amplified risks brought about by these technologies.

(EA2) Are organisational, or group-specific, principles for the ethical use of AI defined and described?

Recommendation:

Similar to the above suggestion, principles for ethical use of AI needs to be monitored and re-evaluated periodically to ensure they sufficiently consider additional risks of emerging technologies. For example, principles may not adequately address confabulation and/or IP risks.

(EA4) Is the workforce trained on values-based decision-making or is there a pilot planned?

Training the workforce to practise values-based decision-making is critical to ensure employees understand the bank's ethical considerations and apply them appropriately in AI-related tasks. However, risk management should also be factored into decision-making to reduce the likelihood of AI-generated outputs conflicting with the bank's values and to help manage potential risks arising from unintended consequences.

(EA5) Is there a statement of relevant values, core concepts, principles, commitments, and specifications, along with a written description, for each use case? Part of this should include identifying commitment owners and the stakeholders with interest in holding the commitment owners accountable.

This question emphasises the need for comprehensive documentation of values, core concepts, ethical principles, and commitments for each AI use case. Having clear documentation reduces ambiguity, ensures consistency in AI decision-making, and ensures stakeholders are accountable for adhering to defined guidelines, thereby mitigating risks related to biased or unethical AI outcomes. However, this question also needs some expansion for Generative AI use cases as there can be challenges in identifying, documenting and measuring the appropriate specifications for all risks.



(EA6) Have relative priorities, which may or may not be tied to materiality, been recorded for each commitment?

Assigning relative priorities to commitments allows the bank to focus on addressing the most critical risks first. By understanding the potential impact of each commitment, the bank can allocate resources accordingly to tackle the highest-risk aspects of AI deployment effectively.

(EA7) Are recourse mechanisms available to data subjects being used? Is there an explanation if the use of recourse is too high or too low?

Recourse mechanisms provide a means for individuals affected by AI decisions to seek redress in case of adverse outcomes. Evaluating the use of these mechanisms helps the bank identify whether they are effective in addressing issues and managing risks related to AI-generated outputs. However, given the nature of Generative AI and the use of foundation models, recourse may be difficult or impossible to provide, as the bank may not be able to permanently adjust the foundation models used based on a particular data subject's request or situation.

(EA8) Are controls in place to revisit and recalibrate the commitments and their specifications to ensure they're properly incentivising the right decisions and accurately measuring progress toward the commitments?

Continuous monitoring and recalibration of commitments and their specifications is essential to adapt to evolving risks and challenges related to AI. By having controls in place, the bank can make necessary adjustments to align AI decisions with desired outcomes, reducing potential risks associated with Generative AI systems.

Evaluation of Transparency-Related Risks Against Veritas Methodology

The original Veritas Methodology proposed a set of 17 questions (as shown in Figure C.4) across a typical AIDA development lifecycle to implement the transparency principle of FEAT for an AIDA initiative. This section provides an assessment of transparency questions and whether they adequately address risks of Generative AI.

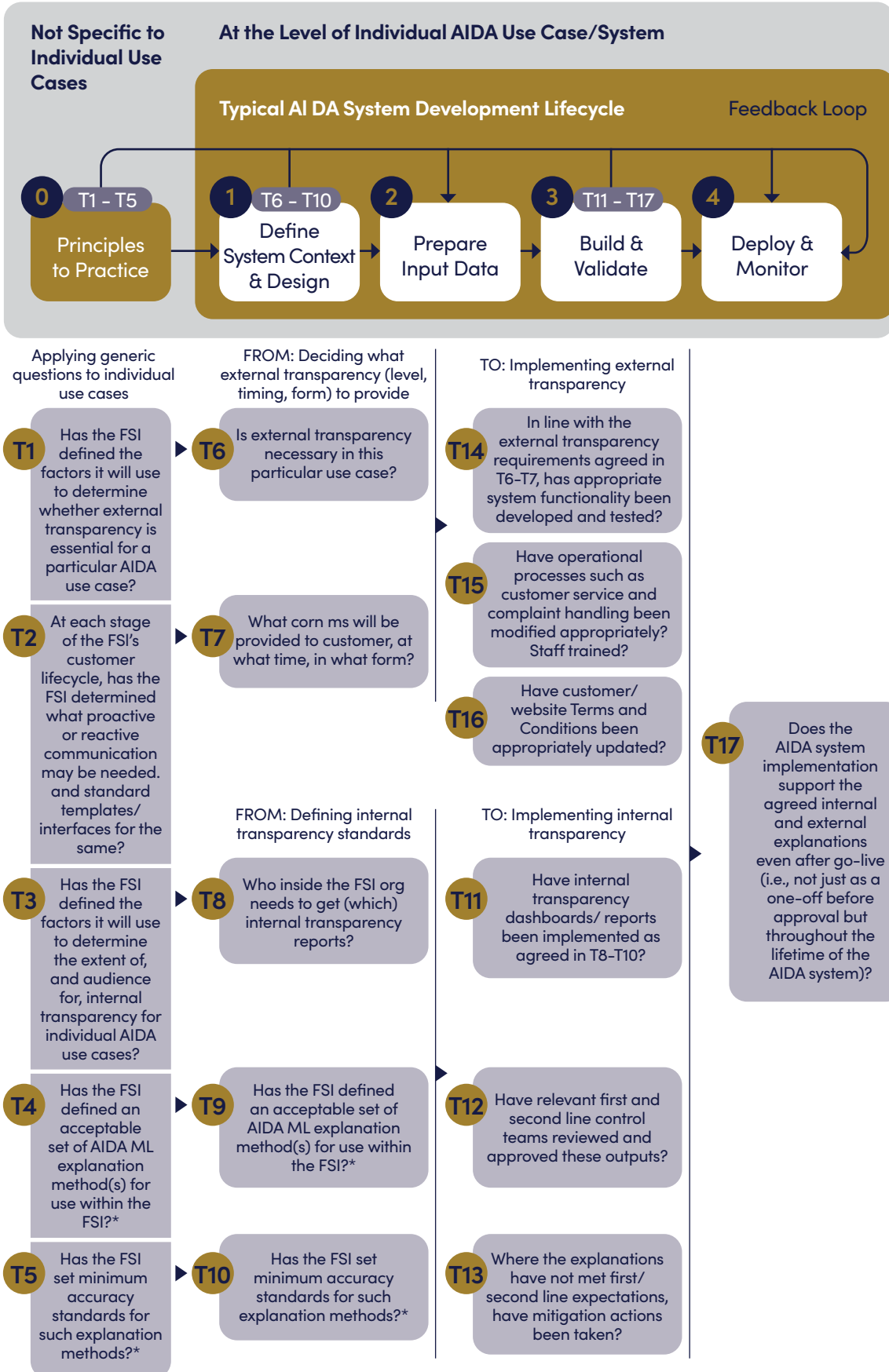


Figure C.4: Transparency checkpoints for AIDA systems



(T1) Has the FI defined the factors it will use to determine whether external transparency is essential for a particular AIDA use case?

The Veritas Methodology provides advice on factors to consider for providing explanation to data subjects. These factors apply to Generative AI as well and could be amplified due to an expanded risk environment with Generative AI.

(T2) (Where an FI has chosen to provide external transparency) At each stage of the FI's customer lifecycle, has the FI determined what proactive or reactive communication may be needed, and the standard templates/interfaces for the same?

The Veritas Methodology provides three different categories of external transparency:

- 1) Proactive vs Reactive
- 2) Generic vs Specific
- 3) Informational vs Action-Oriented

All three categories of external transparency apply to Generative AI. Proactive transparency plays a vital role in mitigating Generative AI risks. Traditional AI use cases typically take a set of input values and provides outputs/decisions. In such cases, it may be sufficient to communicate the use of AI once, before the data subject/customer applies for the product or service. Generative AI, unlike traditional AI, can be interactive, involving multiple iterations over time. This is a key consideration in deciding how to implement proactive communication for Generative AI.

(T3) Has the FI defined the factors it will use to determine the extent of, and audience for, internal transparency for individual AIDA use cases?

FIs should use factors from the Veritas Methodology to determine internal transparency requirements. Considering the risks of Generative AI, it is also required to consider copyright risks. For example, consider a use case where an FI's communications team uses Generative AI to generate pictures of students and sportspersons as part of launching a campaign to increase customer base. In this case, the communications team needs to ensure that the generated pictures do not result in copyright infringement to the original picture used in training the model.

The level of internal transparency is determined by the materiality of the use case. Also, different internal stakeholders may require different levels of explanations. Generative AI systems, due to their inherent complexity, present explainability challenges.

(T4) Has the FI defined an acceptable set of AIDA ML explanation method(s) for use within the firm?

Internal transparency explains current state of Generative AI/ML methods:

Post Hoc Interpretability: Model-Agnostic Methods

The focus is to provide explanation as to why a Generative AI model generated a certain output and not another. This explanation method is the model and output agnostic (whether content is image or text). Detailed explanation is provided in the following section specifically for use cases. While Generative AI can generate any type of unstructured content, we discuss the use case of text-to-text generation in detail as this represents a significant portion of application in FIs in current context.

FIs should determine the level of explainability needed when using Generative AI. These can include explaining what inputs (prompt) led to the output.

The purpose of this part is to provide insight into **“What are the inputs given to the model that drive/influence output generation?”**

While the generative model can also be used to provide binary answers, we focus mainly on unstructured generation for this part. Traditional ML mainly generates structured output while Generative AI mainly generates unstructured outputs.

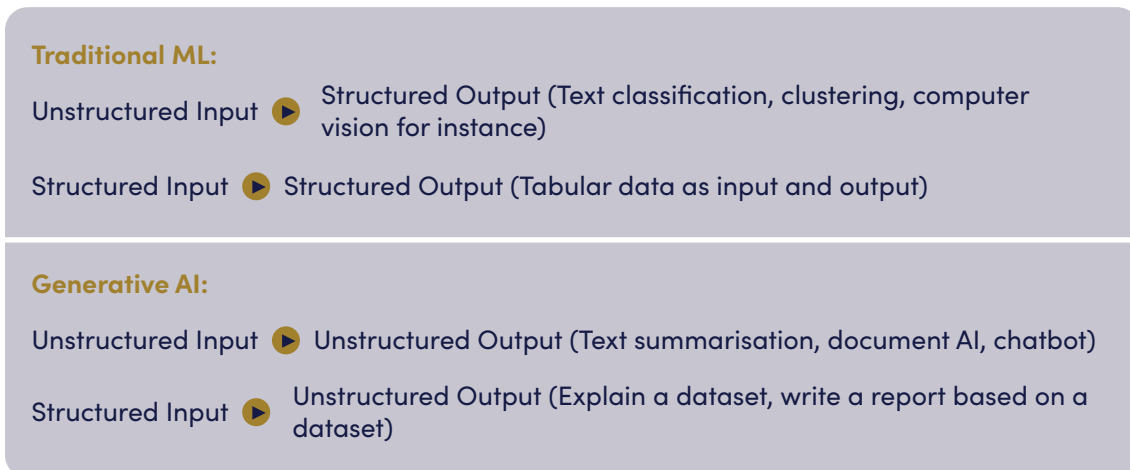


Figure C.5: Traditional ML versus Generative AI

As general transparency methods, institutions should try to provide transparency on elements of the input influencing output. In addition, most of these generations are random by nature. It means that results could differ between two generations with the same input. In addition, slight modifications of the input can lead to different results. Most of these models are therefore sensitive to prompt formulation. Note: sharing specifics about prompt management aspects to the wider public can compromise the security aspects of the system.



a) Prompt sensitivity

The objective is to address the randomness of the generation. Simulation should be done to address the variation in generation depending on slight modifications of the prompt. Modification can be as simple as adding space, changing punctuation, rephrasing sentences to passive form or any way that does not alter the meaning. The same holds true for prompt lineage.

b) Initialisation sensitivity

Generative AI can be sensitive to initialisation or past prompts. In many cases, models have some initialisation where prompts are added before the user prompt. Models can be sensitive to these settings and this sensitivity should be monitored and addressed.

Post Hoc Interpretability: Model-Specific Methods

Model-specific explanation is at an early research stage. Most of the ongoing work is an attempt to explain what is happening in the model (mainly transformers) and whether generation is independent of words that could influence fairness. Most of these methods do not try to explain the model or its interim steps but focus on explaining outcomes and mapping them to inputs.

a) Text generation

Text generation is one of the main use cases of Generative AI. It can be used for customer interaction, preparing drafts for response, translation and many other use cases. It is expected to drive FIs' productivity and operations. The challenge remains in explaining generations made by models. The input of text generation is usually assumed to be text (although it could be images, sounds or others for multimodal models).

b) Input to output visualisation

As a key consideration, we try to explain the model output based on input and try to use attention maps for models.

c) Visualisation of transformers

These methods attempt to visualise the attention of the neural network and could be reproduced for Generative AI models. Scores and input-output association are probabilistic and not causal by the nature of transformer models.

d) Integrated gradients

As mentioned, this method is generic but can be computationally intensive. For translation via LSTM (actually a generative model), attention can be mapped almost 1-1 with words. The following is the result of integrated gradients on an LSTM used for translation. For instance, the word “Gentlemen” maps most to Herren although it is split into three. LLMs are essentially very large conditional probability models for which such mapping can be mimicked at a high computational cost.

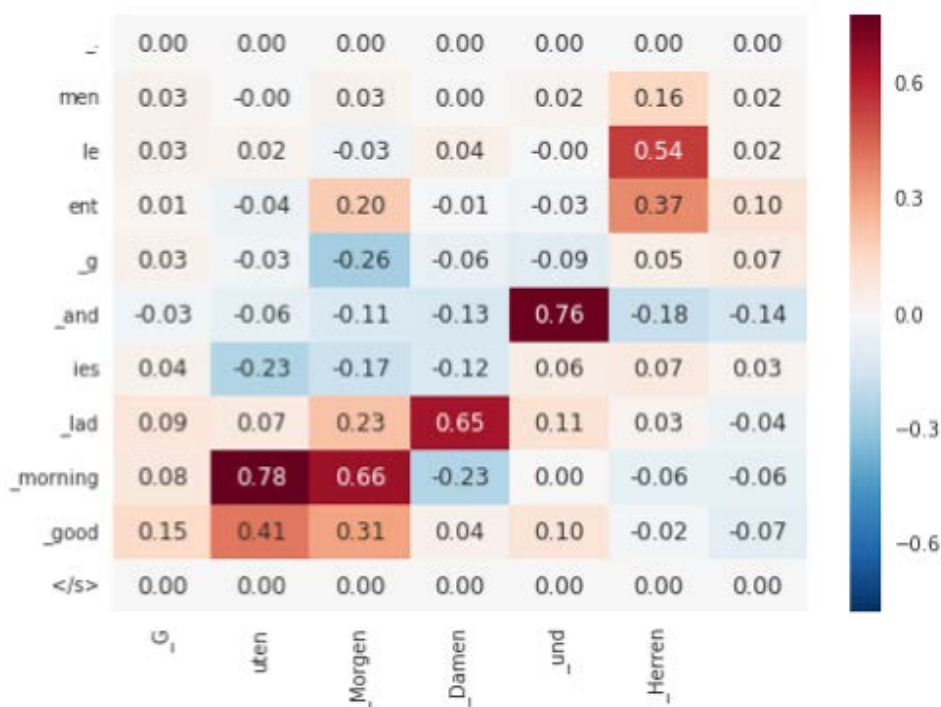


Figure C.6: Integrated gradient of “Gentlemen”¹³

e) Counterfactual explanations

By providing the opposite information to the Generative AI model, the generation is impacted. The counterfactual will focus on assessing the opposite sentence and ensure that the outcome is in line with the expectation. If the LLM is also involved in decision-making, we should aim to assess its accuracy.

f) Order sensitivity

Where a model is used to summarise or retrieve information from a vector database, for instance, the model developer should ensure that potential randomness from the retrieval is considered. Chunking used in LLMs should be investigated.

¹³ Mukund Sundararajan, Ankur Taly, and Qiqi Yan, “Axiomatic attribution for deep networks,” in *International conference on machine learning (PLMR: 2017)*, 3319–3328.



(T5) Has the FI set minimum accuracy standards for such explanation methods?

It is important to note that the concept of accuracy and minimum standard is not well defined for Generative AI explanation methods. Confidence intervals will not exist and the difference between generations will not be very objective. Further, Generative AI systems produce output in the form of natural language, images, videos, etc., which are challenging to assess for accuracy, let alone for explainability. Hence it is premature to mandate minimum accuracy standards for explanation methods related to Generative AI systems.

(T7) If yes, has the team identified the proactive and reactive communication needed at each stage of the customer lifecycle, and the form of such customer-facing communication? Apply standards from T2 to help answer the question.

In line with the use case and factors identified in T6, use additional advice provided under T2 along with the original advice from Veritas to determine the transparent communication requirements for the use case.

(T8) Has the team determined the level of internal transparency needed, and the audiences for the same? Apply standards from T3 to help answer the question.

Use Veritas to determine internal stakeholder and transparency requirements. Additionally, determine copyright-related transparency requirements, if applicable. There is no Generative AI-specific update to this step.

(T9) Has the team selected a suitable explanation method for this specific use case from the approved list in T4?

Use additional advice from T4 to select a suitable explanation method for the use case. There is no further Generative AI-specific update to this step.

(T12.1) Have relevant legal teams reviewed and approved the outputs to be in line with liability apportionment, copyrights?

FIs should engage their legal teams to confirm if the transparency implementation for the use case is appropriate.

(T16) Have customer/website Terms & Conditions been appropriately updated (i.e., to explain how data is used, how benefits and risks to individuals are associated with the processing, and how individuals may participate and object where appropriate)?

The implementation may extend beyond terms and conditions. FIs should implement requirements as identified in T7.

D: Challenges for Banks Operating Across Multiple Jurisdictions

Financial services as a global industry is certainly no stranger to regulatory policies. While there is no empirical evidence to suggest it is the most regulated industry, many would agree that it is still highly regulated and becoming increasingly so. Regulation impacting financial services broadly comes from three main areas:

- **International:** in the form of common standards, policies, laws, and guidelines, which are collaboratively developed globally and generally enforced locally (e.g., by country regulators). An example of this would be the Basel Accords, which ensure that firms operating in the industry are prudently managed, by establishing rules to ensure institutions hold enough capital to secure continuation of a safe and efficient market and are able to withstand any foreseeable problems.
- **Regional/National:** in the form of common standards, policies, laws, and guidelines, which apply to many or all industry sectors. An example of this could be the General Data Protection Regulation in the European Union, or the Personal Data Protection Act (PDPA) in Singapore.
- **Industry:** in the form of common standards, policies, laws, and guidelines, which apply to FIs operating in a specific country.

Below are a few examples of AI regulations across the different areas:

- International bodies (e.g., OECD framework for the classification of AI systems)
- Regional and national bodies (e.g., Personal Data Protection Commission Singapore – A Proposed Model Artificial Intelligence Governance Framework)
- Industry (e.g., Hong Kong Monetary Authority – high-level principles on AI)

There is little argument that such a vast body of work has helped raise awareness of AI, its potential benefits, and key risk considerations. Many organisations will have also spent much time researching this area to inform internal perspectives and have actively leveraged such frames of reference to develop organisation-specific AI governance frameworks. Despite the development of such a vast body of work aimed at governing the use of AI systems, the challenges involved in keeping pace with the adoption of AI solutions across industry sectors are clearly evident.

While this report itself does not elaborate on specific reasons for the challenges involved, as a consortium, we may be able to put forward some insights.

1. Across the AI governance body of information, there is a level of convergence in articulating why AI governance is important and there are similarities in some of the key principles and areas of focus. Conversely, there is noticeable divergence in guidance



provided on how to turn principles into practice across jurisdictions. With such significant divergence in approaches at a policy level across jurisdictions, for organisations operating across multiple locations (commonly larger organisations that are more likely to be driving AI adoption at scale), complying with all relevant policies and guidelines is challenging. Developing a “one size fits all” framework or customising an AI governance approach for each jurisdiction would prove to be highly challenging tasks for financial institutions.

2. Many organisations may still be fairly nascent in their adoption of AI and still experimenting and piloting without much awareness of global regulatory requirements as opposed to moving forward with full-scale deployment. As such, many organisations in this phase will be using these pilots to concurrently develop and test their AI governance controls and processes in a small and contained environment.
3. Current frameworks and approaches are either too theoretical or over-engineered for organisations to adopt at scale. Where efficiency and effectiveness gains expected from AI-driven systems are offset by increased governance efforts, organisations may slow down adoption or cut corners on governance.
4. Experienced and qualified AI governance resources are understandably very scarce. AI governance requires experience in both data science and governance with regulatory requirements, two distinct skill sets which are rare to find in a single resource. Even across large organisations, it is hard to find governance resources with the aptitude to learn data science and even more unlikely to find data science resources to perform governance roles.
5. In practice, many of the anticipated risks associated with AI have failed to materialise or manifest in any significant way or are being effectively mitigated through existing business processes and proactive human intervention.
6. While regulation or compliance to AI guidelines remains optional across many industries, there may be some level of inertia or procrastination in adopting and operationalising suggested governance controls and processes. This is likely to change as more statutory requirements emerge and are enforced.

Despite the apparent inertia in organisations mitigating risks associated with AI systems, there are of course other examples where this is not the case. In a recent whitepaper published by MAS Veritas – From Methodologies to Integration, a few FIs shared in detail the progress they had made in operationalising more robust AI governance frameworks and controls through the adoption of the MAS FEAT Principles and MAS Veritas Framework and Toolkit.

E: Architecture and Infrastructure

E.1 Generative AI Deployment and Adoption Approach

Generative AI FMs vary in size and complexity. Some of the smallest in commercial use today will have 300 million parameters, while some of the most complex will have hundreds of billions. Larger models are more sophisticated and capable of more complex tasks. However, they also require much more data and computing power to train. Further, the added complexity may pose risks.

Organisations can choose from a broad and growing set of FM providers, and their choice will be a key factor in determining the model hosting approach. This includes models GPT4 and DALL-E from OpenAI, Cohere, Claude by Anthropic, LLaMA by Meta AI, StabilityAI, MosaicML, and Inflection AI as leading models and players in the market.

Companies such as Salesforce, Stability AI and Hugging Face have open-source models with code relatively free for use. Research institutes and other open-source organisations like LAION and Eleuther also offer open-source FMs.

Additionally, FIs can source FMs from hyperscale cloud provider partners. They provide a one-stop-service in building, training, deploying and accessing the model. Such providers even include infrastructure for model hosting, such as AWS with Amazon Bedrock, Google Cloud with Vertex AI, and Microsoft with Azure AI. The considerations when selecting a deployment pattern are discussed in more detail below.

Before the selection of deployment pattern, many organisations have taken a three-step adoption approach to address the desired use case:

1. Run an initial minimum viable product (MVP) implementation project, a production pilot, or both. The production pilot should involve prompting a pre-trained model as a gateway process before fine-tuning or training a model.
2. Follow a structured approach where key results are identified upfront, ideally with metrics that quantify the success achieved by the MVP or pilot.
3. Scale up after successful justification of the case, its benefits, expense and model fine-tuning and training complexity.

In addition, considering ways to integrate the Generative AI platform within existing architecture is crucial. Figure E.1 illustrates how Generative AI platforms can impact FIs' security and interaction models or application ecosystem.



Position of Gen AI Platform within Enterprise Architecture

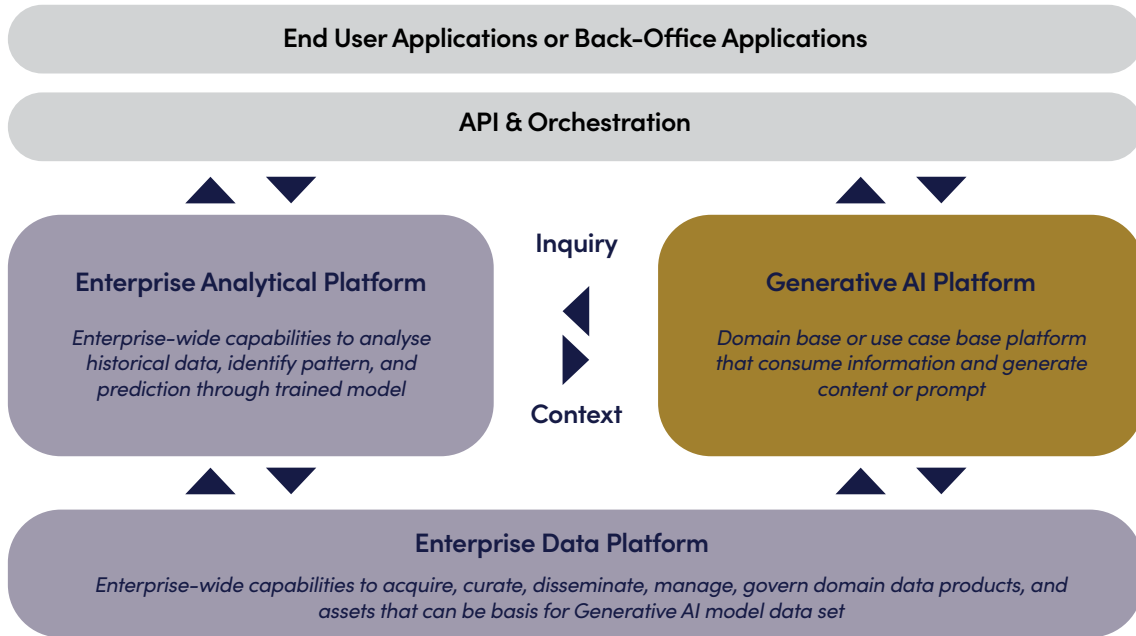


Figure E.1: Illustration of Generative AI positioning within an enterprise architecture

The architecture choice can help mitigate risks, aid risk appetite or threshold setting, and assess and monitor goals with FEAT Principles.

Given the selection of FMs, the deployment pattern of Generative AI can be described in three archetypes: buy, boost or build.¹⁴ The spectrum from buy to build represents a trade-off between the unique features, effort required, time to market, cost, and risk profile of each deployment pattern, as illustrated in Figure E.2 below:

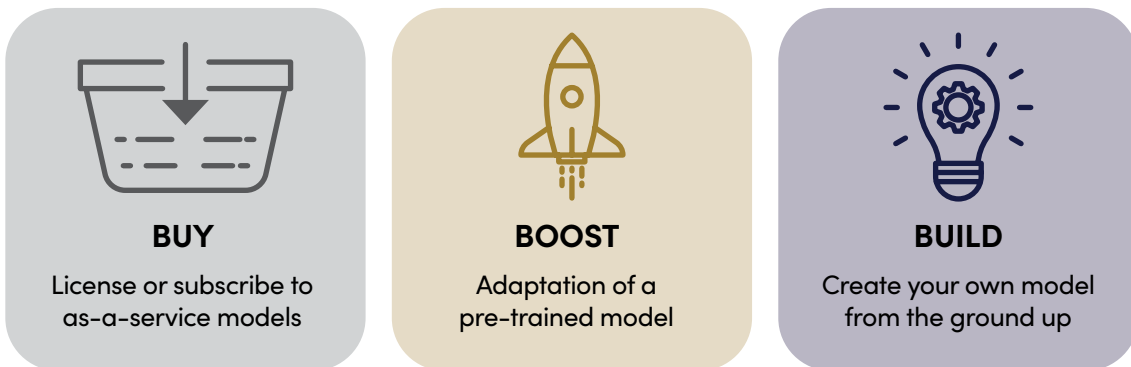


Figure E.2: Three archetypes of Generative AI deployment patterns

¹⁴ From Accenture blog on 7 architecture considerations for Generative AI - <https://www.accenture.com/us-en/blogs/cloud-computing/7-generative-ai-architecture-considerations>

The choice of deployment pattern is more than an architectural choice; it is an important business decision crucial to the organisation's effective, secure and responsible adoption of Generative AI.

Deployment pattern choices broadly impact the sourcing of FMs (open or proprietary), their functionality, technology architecture, resources required, skill and competencies required, and governance. Business considerations such as auditability, data requirements, cyber and data security and compliance, and associated risks will also be affected.

Buy: License or Subscribe to As-a-Service Model

This deployment pattern refers to the broad use case of an organisation that works with a vendor to license or leverage as-a-service FMs with minimal changes to the model itself. Buy-type deployment reduces deployment time but also presents trade-offs: the organisation will not have control over sources or data on which the model was trained. FMs, by nature, create outputs based on training data and inputs they receive during use. Without any control over the data or internal parameters of the model, organisations will have limited ability to control the model's output or to explain why it produced that output.

With buy-type, in-terms of output, an organisation focuses on prompting and grounding to make the FM relevant and contextual to business situations. A key advantage with this is that data access takes place at the application layer. Hence, it is possible to securely control or limit the data access of the foundation model.

In cases where prompting and grounding are not considered sufficient, an organisation may well choose to contextualise their models using a closed knowledge repository, assigned specifically to a model for decision-making.

Boost: Adjustment or Adaptation of a Pre-Trained Model

This deployment pattern refers to a broad set of cases where an organisation uses an FM – either licensed from a vendor or available through an open-source arrangement – and modify it to change its behaviour or to integrate their domain data. Boost-type deployment patterns can be further divided into two categories:

- *Adjust-type trains the model with additional datasets with the objective to fine-tune weight and parameter to improve model accuracy.*
- *Adapt-type trains the model with new datasets with the objective to perform unique tasks for a new domain that result in weight and parameter.*

With boost-type, an organisation's focus is to train the model with a dataset. It is recommended that data used in fine-tuning or training a model do not contain confidential material that cannot be disclosed to the model's intended users. This deployment pattern ensures the model is contextualised to address specific tasks. However, to adopt this deployment pattern, more technical expertise and computing power is required.



Build: Create Your Own Model from the Ground Up

This deployment pattern refers to a use case where an organisation chooses to own the full process of designing, developing, and training a new proprietary FM. This deployment pattern requires significant time and resources, and typically requires access to high levels of AI skills, either in-house or through a third-party arrangement.

Build types provide organisations with the ability to fully control data, inputs, design of the model (to ensure data quality), and relevancy of data for training. This reduces the risk of model inaccuracy compared to a pre-trained model. This accords additional control to mitigate associated risks with Generative AI adoption.

Practical aspects of adopting Generative AI from a research perspective include advancing an organisation’s reputation, technological capabilities, and competitive advantage, particularly if it also develops its own proprietary system.

E.2: Key Considerations in Setting Up Private Infrastructure

Although private infrastructure provides greater control and confidence for data residency, security, and privacy, it often comes with higher demands of internal knowledge and capabilities, in particular IT capabilities, and in-house expertise in hosting and supporting AI models. Organisations seeking to deploy a Generative AI system on their own private infrastructure must take into account several key considerations.

Table E.1: Key Considerations in setting up private infrastructure

Key Consideration	Description
1. FM infrastructure requirements	Evaluate existing infrastructure or determine the necessary infrastructure to support the FM software. Consider factors like server capacity, storage, network bandwidth, and scalability requirements to accommodate expected user base or transaction volume.
2. Ensure FM’s compatibility to infrastructure	Evaluate the compatibility requirements of selected FMs to available infrastructure needs. Verify your operating system, web server, database, and other software components to ensure they are compatible and properly configured.
3. Safeguard data and prioritise security	Identify and implement robust security measures to protect sensitive user data and prevent unauthorised access. Use SSL/TLS encryption to secure both data transmission and data at rest, implement user authentication and authorisation protocols, and regularly update and patch the FM software to address security vulnerabilities.

Key Consideration	Description
4. Prepare for growth and scale	Evaluate scalability of existing infrastructure to handle increasing user loads. Ensure that server setup and network architecture are capable of handling concurrent connections and high volumes of data.
5. Protect data with backup and recovery	Implement a comprehensive backup and disaster recovery plan to protect FM data. Regularly back up database and content files and test the restoration process to ensure data integrity. Store backups off-site or in secure cloud storage to mitigate the risk of data loss.
6. Ensure high-quality and reliable connectivity, including sufficient bandwidth	Evaluate and assess internet connectivity to ensure sufficient bandwidth for smooth Generative AI operation. High-quality and reliable internet access is essential for fast content delivery, video streaming, and seamless user experiences. Consider redundancy options, such as multiple internet service providers, to minimise downtime.
7. Ensure continuous monitoring and analysing	Adopt and implement monitoring tools to track system performance, identify bottlenecks, and proactively address issues. Utilise analytics to gain insights into user behaviour, course effectiveness, and engagement levels. These can help optimise the model and make data-driven decisions.
8. Evaluate financial implications and associated costs	Consider the financial aspects of hosting FMs on private infrastructure. Hosting a Generative AI model on private infrastructure can be expensive, which includes cost of hardware, software, and electricity.
9. Regular maintenance and updates on the FM	Regularly maintain and update the FM software, including bug fixes, security patches, and feature enhancements. Stay informed about new releases, security advisories, and community forums to keep the model up to date and secure.
10. Ensure there are necessary skills and competencies	Consider internal resources, skills and competencies needed in managing a Generative AI FM. Assess whether the necessary technical skills and experience to handle hosting and maintenance are available, including alternative options such as seeking professional assistance. Engaging an experienced consultant can provide valuable guidance, troubleshooting, and support throughout the process.



E.3 Seven Dimensions of Generative AI Considerations

One key factor FIs need to consider in Generative AI adoption is the choice of deployment pattern. However, to adopt Generative AI effectively, securely and responsibly for enterprise-grade usage, FIs should also consider the seven dimensions of technology considerations, which were developed by the consortium following consultation with its technical experts, as shown in Figure E.3 below.

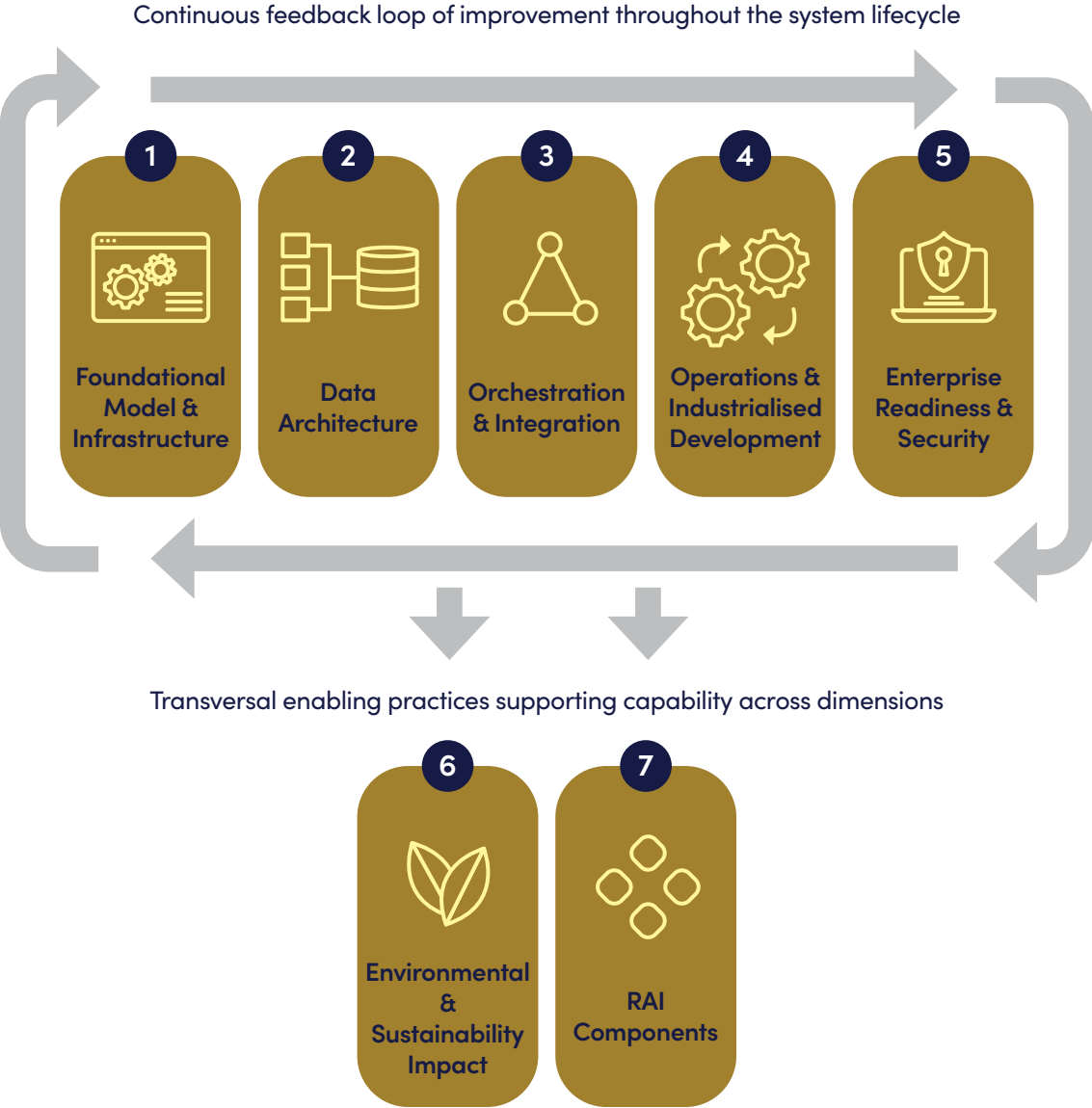


Figure E.3: Seven dimensions of Generative AI considerations for enterprise-wide adoption

1. **FM and Infrastructure:** Selecting FMs, their accessibility, and their model hosting infrastructure.
2. **Data Architecture:** Appropriately managing data and providing FMs with data access.
3. **Orchestration and Integration:** Connecting the model to existing enterprise systems.
4. **Operations and Industrialised Development:** Operating Generative AI systems at scale through streamlined development, deployment management, continuous monitoring, and analysis and improvement.
5. **Enterprise Readiness:** Establishing standards on scalability, security, and compliance.
6. **Environmental and Sustainability Impact:** Considering the environmental impact of Generative AI adoption.
7. **RAI Components:** Adopting responsible AI practices, such as the Veritas Methodology, across the enterprise.

E.3.1 Foundation Model and Infrastructure

The consortium conducted workshops and expert consultations to identify the key criteria (shown in Table E.2) to consider when choosing FMs that fit the FI's needs. These criteria are underpinned by the level of control the FI requires. The first option is to go for full control by deploying the models on the organisation's public cloud or private infrastructure. The second option is to focus on speed and simplicity by accessing Generative AI as a managed cloud service from an external vendor. Both options have pros and cons.

FIs choosing full control need to be aware of additional factors to consider. This often requires strong internal knowledge and IT capabilities, such as identifying and managing the right infrastructure, version controlling the models, developing associated talent and skills, developing full-stack services, and innovating specialised infrastructure (refer to Appendix D.2 for further elaboration).

FIs in Singapore should consider the MAS Technology Risk Management Guidelines, ABS Cloud Computing Implementation Guide, and MAS Guidelines on Outsourcing (superseded as of 11 December 2024 by the new Notice and Guideline on Third-Party Risk Management) when making decisions about deployment and accessibility (see Section 3.7 for a detailed discussion).



Table E.2: Key criteria in FM selection

Key Driver	Key Criteria	Description
Technical driver	Ease of adaptation	Ability to fine tune the FM with weights
	Integration features	Ease of integration with enterprise data and systems
	Suite of models	Spectrum of available models to meet different task patterns
	Model features	Built-in features of the model, e.g., context window, prompt UI, etc.
	Built-in governance	Governance capabilities provided to meet enterprise standards
	Scalability	Ability for the service to scale on the volume of API calls, users, performance, response, latency, etc.
	Usage cost	Total cost of ownership
Functional driver	Domain awareness	Visibility on domain information that has been used to pre-train the model
	Task specificity	Accuracy of the model's performance against specific use cases
	Speed of adoption	Time to complete experiments for specific tasks to get the expected outcome
	Language base	Gauge the need for multilingualism or a specific language for business
Policies and principles driver	Information security	Data residency, data sharing, movement, protection policies
	Model security	Robustness against security risks and security-by-design approach
	Deployment pattern and model source	When to use open source or proprietary for the choice of deployment pattern
	Model hosting (private vs public)	When to use public cloud infrastructure vs in-house

E.3.2 Data Architecture

Generative AI requires a wider range of context to accurately generate novel content. The adoption mostly leverages vast amounts of semi-structured or unstructured data, such as text documents, social media feeds, chat streams, images, video files, etc.

FMs need vast amounts of curated data to learn, which makes solving the data challenge an urgent priority for every business. Customising FMs also requires access to good quality domain-specific organisational data, semantics, knowledge, and methodologies to ensure quality of the model output.

This highlights the need for a modern enterprise-grade analytics and data platform, built with a trusted, reusable set of data products that is cross functional. The platform will allow data to break free from organisational silos and be democratised for use across an organisation.

Companies need a strategic and disciplined approach to acquiring, growing, refining, safeguarding and deploying data for Generative AI adoption. Broadly, there are two areas (domain data risk assessment and mitigation, and domain data supply chain) that organisations must pay attention to and work together to successfully establish data environments as illustrated in Figure E.4 below.

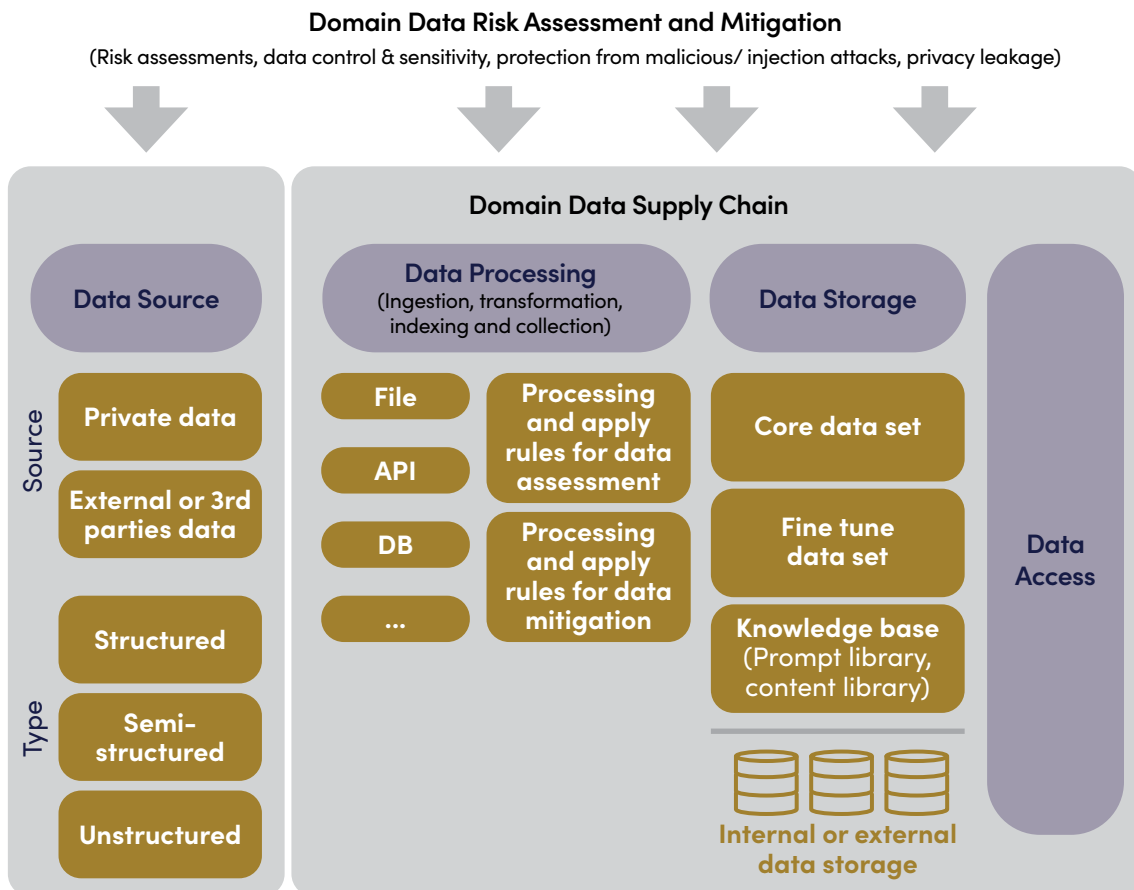


Figure E.4: Data architecture focus areas for Generative AI adoption



A combination of these two areas constitutes a good data architecture for Generative AI adoption. It is also crucial to consider how the people creating the business case and running the operating model may operationalise the implementation and enforcement of the data architecture.

Domain data risk assessment and mitigation focuses on governance and policies applied to data collection and catalogue management, data quality, data lineage, data security, compliance and risk control, and data storage. Data supply chain focuses on adopting the right technology and tools that will allow organisations to easily access high-quality, domain-specific data for Generative AI adoption (refer to Appendix D.3 for further aspects to consider).

Organisations may need to augment existing computing infrastructure and tooling stacks depending on their choice of deployment pattern. This primarily relates to the source provider of the foundation model and infrastructure for the model hosting approach, for example, buy and boost. By leveraging available framework and tools from cloud providers, organisations can bridge the gap in modern data platforms and accelerate the adoption of Generative AI.

E.3.3 Orchestration and Integration

Organisations will use FMs of varying size, complexity, and capability when adopting Generative AI. These models need to be able to work together with existing systems or applications. To generate value in interacting with the user or customer through existing systems or applications, the model needs to access multiple applications and data sources within the organisation's application ecosystem. Building a separate tech stack for Generative AI creates significant additional complexities. However, integrating the models into existing systems can be challenging due to several reasons:

1. The underlying model architecture is often complex and may require specialised knowledge to understand and work with.
2. Integrating the models may require working with multiple programming languages and frameworks.
3. Integrating into a system with a different programming language or framework may require specialised knowledge and experience.
4. Integrating with legacy systems may require significant modifications to the existing codebase, which is complex and difficult to modify without causing undesired consequences.
5. Legacy systems are often written in outdated programming languages or use old technologies, making it difficult to integrate with modern Generative AI models.

As such, organisations need to establish enterprise capabilities to enable standard orchestration and integration patterns, such as standard APIs. There are two approaches for orchestration and integration that can be leveraged and explored:

- Adopt workflow and orchestration tools that natively support multiple Generative AI model calls.
- Adopt separate technology to extend model capabilities to interface with other Generative AI models, and deal with Generative AI data to be generated and consumed.

Regardless of how Generative AI is procured or deployed, orchestration and integration will be crucial elements for enterprise-wide adoption. Generative AI cannot exist in isolation, and will serve as consumers, producers, or both.

E.3.4 Operations and Industrialised Development

Organisations with robust, mature processes to manage AI technologies and/or cloud-hosted services should transition well into in-house Generative AI adoption. Equally, this applies to the existing approach to ensuring the effectiveness, efficiency, consistency, and quality of technology solutions. It is important to assess and onboard solutions for any gaps in tooling in line with the organisation's existing processes.

Generative AI adoption introduces a new paradigm called Generative AIOps to accommodate its unique traits, for example, the ability and flexibility to be contextualised to completely different downstream tasks with efficiency and scalability while keeping in place risk identification and mitigation throughout their development lifecycle. Organisations need to thoroughly review and update their machine learning operations (MLOps) framework to productise machine learning applications based on the new technical capabilities required to effectively manage Generative AI.

To industrialise the development process, other considerations must be taken into account. The objective is to identify and fill any gaps in new frameworks, tools, or technology capabilities that may not be part of the existing internal offering, such as vectorisation capabilities, knowledge graphs, prompt engineering, human-in-the-loop adoption, etc.

Organisations need to conduct thorough reviews on operability strategy and industrialised development application approach to ensure alignment for Generative AI adoption. Working and learning with tech vendors are initial steps to consider, as they have developed technology stacks with frameworks and tools to work effectively with FMs.

Establishing enterprise-wide monitoring, evaluation and analysis capabilities is equally crucial, with key measurements to adopt and apply across technology solution layers (refer to Appendix D.4 for considerations of key measurements). Combined with effective governance, making use of the lessons learned in this emerging area of practice, organisations can systematise leading practices and apply them across their operations and industrial development ecosystem.



E.3.5 Enterprise Readiness and Security

The emergence of Generative AI comes with opportunities and risks as outlined in Section 2. Generative AI adoption allows organisations to maximise efficiency and competitive advantage. Organisations need to assess their readiness to adopt enterprise-grade Generative AI through trust building. However, it will be a journey, and not just a one-time effort.

Trust building in Generative AI requires user and stakeholder education; culture development; useful, reliable and trustworthy systems; continuous improvements; and the adoption of standards or best practices. FIs also need to continuously assess if risks are appropriately mitigated, given the ever-evolving AI threat landscape.

Specifically for cybersecurity, Generative AI is a double-edged sword. On one hand, organisations can leverage AI to strengthen cybersecurity capabilities (e.g., AI-enabled threat detection and response). On the other hand, bad actors can leverage Generative AI to launch more sophisticated cyberattacks. Examples include polymorphic malware that bypasses signature-based detection, advanced persistent threats (APTs) with AI to avoid detection, AI-powered malware, phishing with NLP and machine learning, and deepfake attacks.

Generative AI will rapidly advance, and it is essential that organisations continuously update their specialised knowledge and strategy to protect against cybersecurity threats in lockstep with Generative AI adoption.

FIs must consider several practices such as updating organisational standards to adequately adopt and enable enterprise-grade Generative AI through technology implementation.

There are three areas of technology considerations for enterprise readiness are: 1) Scalability, 2) Security, 3) Compliance. They are cross-cutting, extending far beyond an individual Generative AI system, and crucial for enterprise-grade Generative AI, as depicted in a non-exhaustive list (Table E.3).

Table E.3: Sample key practices for enterprise-grade Generative AI

Scalability	Security	Compliance
<ul style="list-style-type: none"> • Model Inference • Model Hosting • Model Suites Coverage • Model Modularity • ... 	<ul style="list-style-type: none"> • Cyberattacks and Mitigations • Data Access Control • Data Breach • Data Loss Prevention • Data Residency • Model Architecture • ... 	<ul style="list-style-type: none"> • Regulatory Requirements

Once key practices are identified and applied, organisations need to have robust technology capabilities and processes to monitor, analyse and evaluate risks and enforce safeguards. These technology capabilities are not limited to Generative AI and must be aligned with architectural principles.

FIs might consider establishing an AI Centre of Excellence to ensure enterprise readiness for Generative AI adoption and that it is used responsibly across the organisation. It also ensures effective governance mechanisms are established, existing policies are periodically reviewed, and standards and guidance documents are established to manage AI/ML adoption. This enhances internal processes and further strengthens the risk and control functions.

E.3.6 Environmental and Sustainability Impact

At a macro level, one-quarter of global carbon emissions come from electricity generation. Currently, the data centre industry comprises some 2% of electricity consumption (US Department of Energy). Technology energy consumption is expected to rise with high-density computing such as AI/ML.

Simply reducing the amount of consumption should not be the only solution. It is also important to have a holistic view which looks at the sustainability of a business process or function, technical infrastructure, architecture, operating model and governance structure rather than at any specific technology in use.

To drive strategic environmental and sustainability improvements, it is important to adopt an approach that combines consumption reduction, process improvement or innovation, and supplier engagement. This approach will ensure that both internal and external stakeholders comply with environmental and sustainability policies, social and governance regulations, or any other commitments to operate and procure in a sustainable manner. Generative AI use may drive up consumption whilst simultaneously enabling process improvements that reduce consumption. In particular, Generative AI has high potential to automate manual tasks. Therefore, this requires a balanced end-to-end view.

When considering where to host Generative AI solutions, one environmental factor to note is that hyperscale data centres and cloud providers are building facilities with best-in-class Power Usage Effectiveness (PUE <1.2) which may exceed the sustainability performance of an organisation's data centre. This may present an opportunity for organisation to improve, not worsen, its environmental footprint. Most cloud providers are committed to achieve carbon neutrality by 2030 and adopt 100% renewable energy. To achieve this, providers are investing in renewable energy, making data centres more efficient, developing new technologies, sharing progress with the public, and educating employees and customers.



There are several options available to mitigate the environmental impact of Generative AI. Organisations should continuously and periodically assess the adoption of Generative AI to make it greener and more sustainable. There are several ways to achieve this:

1. Use existing FMs instead of generating a new model.
2. Optimise the FM architecture and processor for efficiency and compactness.
3. Use a greener data centre with renewable energy sources and efficient cooling systems.
4. Reduce data size by using compression and augmentation techniques.
5. Limit the frequency and volume of inference requests using caching and filtering techniques.
6. Evaluate the level of resource consumption and promote the adoption of models consuming less resources for a given task.
7. Monitor and report the environmental impact of your model by using tools and standards.
8. Offset the environmental impact of Generative AI models by investing in projects or initiatives that reduce emissions or enhance sequestration.

Organisations should have a robust system in place for acquiring standardised data on total environmental impact from internal providers or vendors. With this visibility, organisations are empowered to identify opportunities to improve the environmental impact of their Generative AI use.

E.3.7 Responsible AI Components

Responsible AI (RAI) is a term that refers to a broad family of practices that address concerns around the impacts of AI on companies, individuals and society. The Veritas Methodology is one set of RAI practices. Mitigations to the Generative AI risks identified in Section 2 could also represent RAI practices. From an architecture perspective, organisations must put in place tailored RAI mitigations across the technology stack and enterprise operating model, as illustrated in Table E.4.

Table E.4: Responsible AI across technology solution layer

Technology Solution Layer	What Stakeholders Should Do	Control Ownership
Applications The UI/UX layer used to access FM models	<ul style="list-style-type: none"> • Assess the accountability and transparency of generated outputs • Address potential IP, confidentiality and privacy risks from model inputs 	Within organisation's control

Technology Solution Layer	What Stakeholders Should Do	Control Ownership
Orchestration and Integration The orchestration and integration layer used to access FMs	<ul style="list-style-type: none"> Assess and implement the accessibility control of data and data systems within an organisation Address potential security and privacy risks from application to model and vice versa 	Within organisation's control
Contextualisation The additional tuning executed on FMs; uses custom domain data and/or human feedback	<ul style="list-style-type: none"> Explore how fine-tuning, embedding and prompting can enhance the soundness, fairness, explainability and robustness of model outputs 	Within organisation's control
Foundation Models The complex AI/ML systems trained on core data including text, image, audio, etc.	<ul style="list-style-type: none"> Consider the risk and impact of undesirable bias in model outputs, as well as data protection concerns 	Shared control between organisation and vendor
Data The core data used to initially train and tune FMs	<ul style="list-style-type: none"> Consider whether core data used to train and tune FMs may introduce bias, privacy or IP concerns 	Shared control between organisation and vendor

The RAI considerations described above should be considered in combination with any shared responsibility with partners/vendors throughout the lifecycle of Generative AI adoption, from initial conception to operation and eventual decommissioning as described in Section 3.



E.4 Key Aspects of Data Architecture with Generative AI

The quality of a model output is directly dependant on the quality of accessible data. Any organisation that wish to build a successful data environment needs to have these aspects properly integrated:

Table E.5: Key aspects of data architecture with Generative AI

Aspects	What Is It?	Why Is It Important?	What to Consider?
Data collection and catalogue management	Curated data that consist of internal and external data sources, including set of data used for Generative AI	Provide accessible data of good quality that are properly managed and governed for specific or generic use cases for Generative AI.	<ul style="list-style-type: none"> • Enable indexing and searchability for semi/unstructured data with the right technology, e.g., vector databases. • Enable a sustainable data supply chain for batch and real-time data with automation to apply governance rules in mitigating risks. • Manage dataset for pre-training, synthetic data, etc. • Manage knowledge repository, prompt library, content generation template, etc. • Manage dataset for human feedback.
Data quality	Accuracy, completeness, and consistency of data	Ensure that data is reliable and can be used for decision-making.	<ul style="list-style-type: none"> • Define data quality standards, implement data quality checks and monitor data quality. • Leverage automation and technologies such as AI, ML, and NLP. • Build specialised skills in data management and analysis for semi/unstructured data. • Apply data validation checks to limit the entry of semi/unstructured data .
Data lineage	History of how data is created, used and transformed	Understand where data comes from, how it is used, and how it is transformed.	<ul style="list-style-type: none"> • Capture data lineage information. • Store data lineage information. • Make data lineage information.

Aspects	What is it?	Why is it important?	What to consider?
Data security, compliance and risk control	Label management, tagging and classification for data security, privacy and compliance, including data cleansing	Enable data security and compliance through labelling, tagging and classification to ensure sensitive and valuable data are properly managed, while mitigating associated risks that may occur with Generative AI.	<ul style="list-style-type: none"> • Define data privacy classification policies and tagging rules. • Implement data classification tools. • Train employees in data classification. • Define data security access control. • Define data cleansing and filtering rules. • Adopt the right tools and embed the rules applied to processes as part of data supply chain.
Data storage	Storing data in a secure and accessible location	Ensure data is protected from unauthorised access and loss.	<ul style="list-style-type: none"> • Choose the right data storage solution for external knowledge, internal and external domain dataset. • Implement data backup and recovery procedures. • Secure data storage systems. • Adopt scalable data storage systems.

E.5 Sample of Platform-Agnostic Reference Architecture for Generative AI

This section introduces a platform-agnostic reference architecture for Generative AI, which provides a list of building blocks and components for technology capabilities organisations can consider.

Acknowledging the dynamic tech environment and rapidly changing technology landscape, this reference architecture is intended to serve as a sample framework instead of norms to comply to. It allows organisations to jumpstart the pilot phase or implementation planning throughout the adoption journey, from use case to enterprise-grade, as illustrated in Figure E.5.

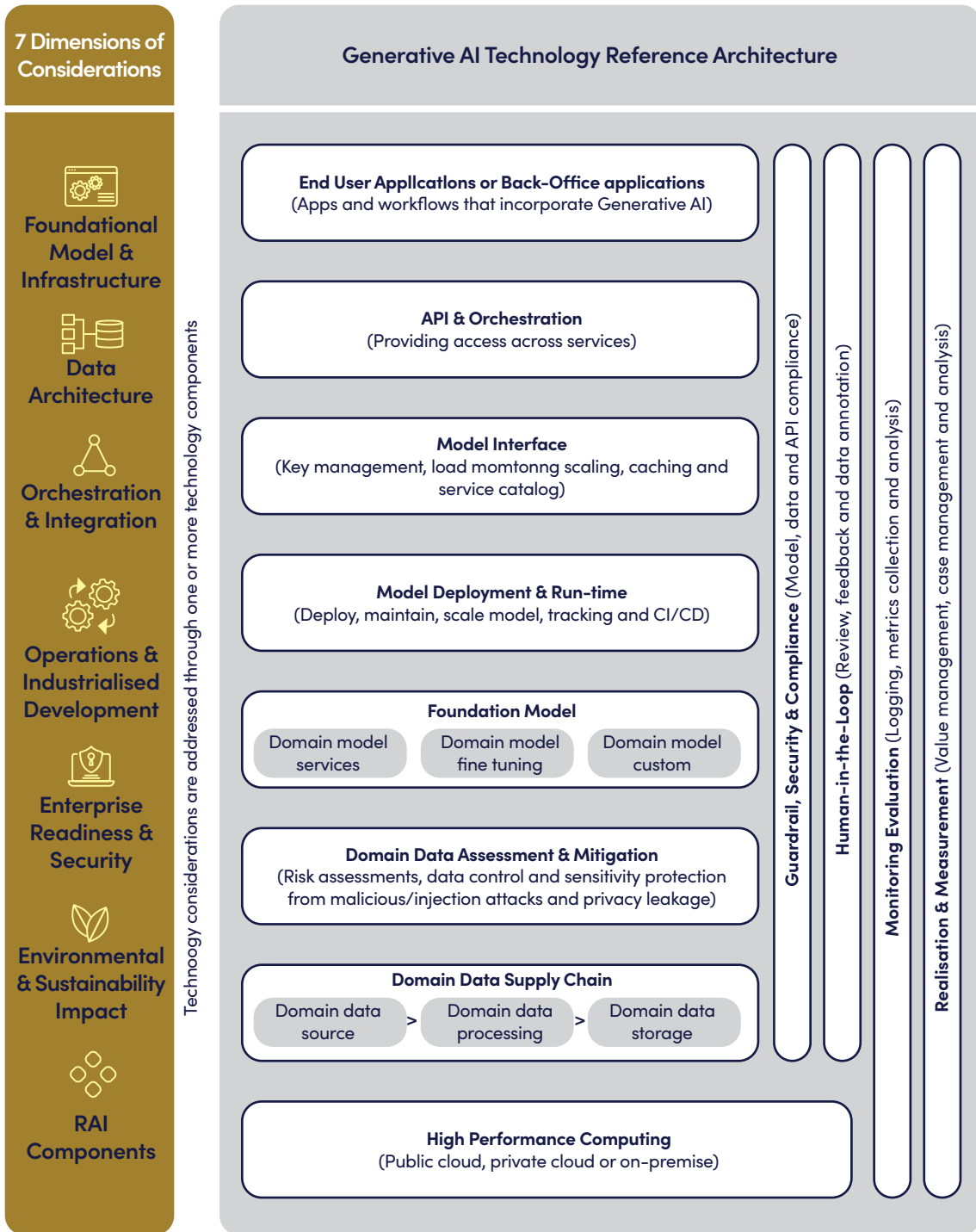


Figure E.5: Sample of Generative AI platform-agnostic reference architecture

Organisations are free to adapt or adopt a different reference architecture based on each organisation’s approach, context (small to large organisations) and needs, upon defining robust enterprise-level technology capabilities across various technology consideration dimensions.

E.6 Key Measurements to Monitor, Evaluate and Analyse Technology Stacks Including Generative AI

Several key measurements that FIs should consider in the monitoring, evaluation and analysis of the technology stack when integrating Generative AI components include:

- Validate and ensure consistency in FMs' response to known prompts, both continually as well as ad-hoc, when modifying prompts or changing models.
- Evaluate and compare the behaviour of multiple models using the prompt library over a range of metrics.
- Observe prompts and responses at scale by extracting key telemetry data and comparing them against smart baselines over time to ensure quality signals when debugging or fine-tuning the Generative AI FM application experience.
- Choose the right model-monitoring metrics, such as quality, relevance, sentiment, and security.
- Use alerting systems to identify and respond to potential issues such as threshold, frequency and escalation.
- Monitor and automate Generative AI reliability and scalability.
- Enable human feedback and reinforcement learning to train an FM through frameworks where humans can review and allow task definition, feedback collection, fine-tuning and evaluation.
- Enable guardrails, security, and compliance for FM applications in real-time and across large-scale deployments, instead of relying on embedded elements.
- Measure business value by enabling a multidisciplinary approach and collaboration between technology, business, and finance teams to balance cost, usage and organisational needs in optimising value.
 - o Getting visibility of components, associated costs and owner
 - o Optimise forecasting and planning to reduce costs and maximise business value
 - o Implement continuous improvements to day-to-day activities in using, managing and governing the model



Bibliography

1. Accenture (2023). 7 architecture considerations for Generative AI. Retrieved from <https://www.accenture.com/us-en/blogs/cloud-computing/7-generative-ai-architecture-considerations>
2. Accenture (2019). AI: Built to Scale. Retrieved from <https://www.accenture.com/content/dam/accenture/final/a-com-migration/thought-leadership-assets/accenture-built-to-scale-pdf-report.pdf>
3. Accenture (2023). AI for everyone. Retrieved from <https://www.accenture.com/us-en/insights/technology/generative-ai>
4. Accenture (2023). A new era of Generative AI for everyone. Retrieved from <https://www.accenture.com/content/dam/accenture/final/accenture-com/document/Accenture-A-New-Era-of-Generative-AI-for-Everyone.pdf>
5. Accenture (2023). Breaking barriers: Exploring how banks scale Generative AI for growth. Retrieved from <https://bankingblog.accenture.com/how-banks-scale-generative-ai-for-growth>
6. Accenture (2023). How do we build Generative AI we can trust?. Retrieved from <https://www.accenture.com/us-en/blogs/cloud-computing/building-generative-ai-we-can-trust>
7. Accenture (2021). Responsible AI: From principles to practice. Retrieved from <https://www.accenture.com/us-en/insights/artificial-intelligence/responsible-ai-principles-practice>
8. Accenture. Ready. Set. Scale. Retrieved from <https://www.accenture.com/content/dam/accenture/final/a-com-migration/manual/r3/pdf/pdf-122/Accenture-Ready-Set-Scale.pdf>
9. Accenture (2021). Responsible AI: From principles to practice. Retrieved from <https://www.accenture.com/us-en/insights/artificial-intelligence/responsible-ai-principles-practice>
Accenture (2023). What is Generative AI?. Retrieved from <https://www.accenture.com/id-en/insights/generative-ai>
10. Aicadium (2023). Generative AI: Implications for Trust and Governance. Retrieved from https://aiverifyfoundation.sg/downloads/Discussion_Paper.pdf
11. AWS (2023). Generative AI with Large Language Models – New Hands-on Course by DeepLearning.AI and AWS. Retrieved from <https://aws.amazon.com/blogs/aws/generative-ai-with-large-language-models-new-hands-on-course-by-deeplearning-ai-and-aws/>

12. AWS. What is Generative AI?. Retrieved from <https://aws.amazon.com/what-is/generative-ai/>
13. Board. ESG Compliance 2023 Guide. Retrieved from <https://idealsboard.com/esg-compliance/>
14. China Briefing (2023). Understanding China's New Regulations on Generative AI. Retrieved from <https://www.china-briefing.com/news/understanding-chinas-new-regulations-on-generative-ai-draft-measures/>
15. Ferrara, E. (2023). Should ChatGPT be Biased? Challenges and Risks of Bias in Large Language Models. Retrieved from <https://arxiv.org/abs/2304.03738>
16. Geelal, J., Khalil, M., Samko, O. & Chung R. (2023). An Overview of Regulations and Ethics of Artificial Intelligence in the Financial Services: Recent Developments, Current Challenges and Future Perspectives. Retrieved from https://www.researchgate.net/publication/374062322_An_Overview_of_Regulations_and_Ethics_of_Artificial_Intelligence_in_the_Financial_Services_Recent_Developments_Current_Challenges_and_Future_Perspectives
17. Google. Safety & Fairness Considerations for Generative Models. Retrieved from <https://developers.google.com/machine-learning/resources/safety-gen-ai>
18. Google Cloud. Generative AI examples. Retrieved from <https://cloud.google.com/use-cases/generative-ai>
19. Google Cloud. Introduction to Vertex AI. Retrieved from <https://cloud.google.com/vertex-ai/docs/start/introduction-unified-platform>
20. Google Cloud. Model Garden on Vertex AI. Retrieved from <https://cloud.google.com/model-garden>
21. Google Cloud. Principles and best practices for data governance in the cloud. Retrieved from https://services.google.com/fh/files/misc/principles_best_practices_for_data-governance.pdf
22. Google Research. Responsible AI. Retrieved from <https://research.google/research-areas/responsible-ai/>
23. Google Sustainability. Net-zero carbon. Retrieved from <https://sustainability.google/operating-sustainably/net-zero-carbon/>
24. Hacker, P., Engel, A. & Mauer, M. (2023). Regulating ChatGPT and other Large Generative AI Models. Retrieved from <https://dl.acm.org/doi/abs/10.1145/3593013.3594067>



25. Hong Kong Monetary Authority (2019). High-level Principles on Artificial Intelligence. Retrieved from <https://www.hkma.gov.hk/media/eng/doc/key-information/guidelines-and-circular/2019/20191101e1.pdf>
26. IFLR (2021). A closer look at Singapore's mandatory corporate ESG disclosures and associated legal risks. Retrieved from <https://www.iflr.com/article/2a646mpixhlq5fi7f5wqo/a-closer-look-at-singapores-mandatory-corporate-esg-disclosures-and-associated-legal-risks>
27. IMF Blog (2020). The Great Lockdown: Worst Economic Downturn Since the Great Depression. Retrieved from <https://www.imf.org/en/Blogs/Articles/2020/04/14/blog-weo-the-great-lockdown-worst-economic-downturn-since-the-great-depression>
28. Infocomm Media Development Authority (2023). About Artificial Intelligence Singapore (AI SG). Retrieved from <https://www.imda.gov.sg/about-imda/research-and-statistics/sgdigital/tech-pillars/artificial-intelligence>
29. Kroll (2020). Global Regulatory Outlook 2020. Retrieved from <https://www.kroll.com/en/insights/publications/financial-compliance-regulation/global-regulatory-outlook-2020>
30. Liang, P., et al. (2023). Holistic Evaluation of Language Models. Retrieved from <https://arxiv.org/abs/2211.09110>
31. McKinsey & Company (2020). Meeting the future: Dynamic risk management for uncertain times. Retrieved from <https://www.mckinsey.com/capabilities/risk-and-resilience/our-insights/meeting-the-future-dynamic-risk-management-for-uncertain-times>
32. McKinsey & Company (2023). The next frontier in risk efficiency. Retrieved from <https://www.mckinsey.com/capabilities/risk-and-resilience/our-insights/the-next-frontier-in-risk-efficiency>
33. McKinsey & Company (2023). The state of AI in 2023: Generative AI's breakout year. Retrieved from <https://www.mckinsey.com/capabilities/quantumblack/our-insights/the-state-of-ai-in-2023-generative-ais-breakout-year>
34. McKinsey & Company (2023). What is Generative AI?. Retrieved from <https://www.mckinsey.com/featured-insights/mckinsey-explainers/what-is-generative-ai>
35. Microsoft. Empowering responsible AI practices. Retrieved from <https://www.microsoft.com/en-us/ai/responsible-ai>
36. Microsoft. Microsoft Solutions Playbook: Artificial Intelligence. Retrieved from <https://playbook.microsoft.com/code-with-mlops/>
37. Microsoft. Working with Large Language Models. Retrieved from <https://playbook.microsoft.com/code-with-mlops/technology-guidance/generative-ai/working-with-llms/>

38. Microsoft Azure. Azure OpenAI Service Documentation. Retrieved from <https://learn.microsoft.com/en-us/azure/ai-services/openai/>
39. Microsoft Azure. Content filtering. Retrieved from <https://learn.microsoft.com/en-us/azure/ai-services/openai/concepts/content-filter?source=recommendations>
40. Microsoft Azure. Microsoft Purview. Retrieved from <https://azure.microsoft.com/en-us/products/purview>
41. Microsoft Research. Reducing AI's Carbon Footprint. Retrieved from <https://www.microsoft.com/en-us/research/project/reducing-ais-carbon-footprint/>
42. Monetary Authority of Singapore (2021). Advisory on addressing the technology and cyber security risks associated with public cloud adoption. Retrieved from <https://www.mas.gov.sg/-/media/MAS/Regulations-and-Financial-Stability/Regulatory-and-Supervisory-Framework/Risk-Management/Cloud-Advisory.pdf>
43. Monetary Authority of Singapore (2016). Guidelines on Outsourcing. Retrieved from https://www.mas.gov.sg/-/media/MAS/Regulations-and-Financial-Stability/Regulatory-and-Supervisory-Framework/Risk-Management/Outsourcing-Guidelines_Jul-2016-revised-on-5-Oct-2018.pdf
44. Monetary Authority of Singapore (2018). Principles to Promote Fairness, Ethics, Accountability and Transparency (FEAT) in the Use of AI and Data Analytics in Singapore's Financial Sector. Retrieved from <https://www.mas.gov.sg/publications/monographs-or-information-paper/2018/FEAT>
45. Monetary Authority of Singapore. Regulations and Guidance. Retrieved from <https://www.mas.gov.sg/regulation/regulations-and-guidance?sectors=Banking&page=1&date=2020-01-01T00%3A00%3A00.000Z%2C2020-12-31T00%3A00%3A00.000Z>
46. Monetary Authority of Singapore (2021). Technology Risks Management Guidelines. Retrieved from <https://www.mas.gov.sg/-/media/mas/regulations-and-financial-stability/regulatory-and-supervisory-framework/risk-management/trm-guidelines-18-january-2021.pdf>
47. Monetary Authority of Singapore (2020). Veritas Document 2: FEAT Fairness Principles Assessment Case Studies. Retrieved from <https://www.mas.gov.sg/-/media/mas/news/media-releases/2021/veritas-document-2-feat-fairness-principles-assessment-case-studies.pdf>
48. Monetary Authority of Singapore. Veritas Document 3: FEAT Principles Assessment Methodology. Retrieved from <https://www.mas.gov.sg/-/media/MAS-Media-Library/news/media-releases/2022/Veritas-Document-3---FEAT-Principles-Assessment-Methodology.pdf>



49. Monetary Authority of Singapore. Veritas Document 5: From Methodologies to Integration. Retrieved from <https://www.mas.gov.sg/-/media/mas/news/media-releases/veritas-document-5---from-methodologies-to-integration.pdf>
50. Monetary Authority of Singapore. Veritas Document 6: FEAT Principles Assessment Case Studies. Retrieved from <https://www.mas.gov.sg/-/media/mas/news/media-releases/veritas-document-6---feat-principles-assessment-case-studies.pdf>
51. OECD (2022). OECD Framework for the Classification of AI systems. Retrieved from <https://www.oecd.org/publications/oecd-framework-for-the-classification-of-ai-systems-cb6d9eca-en.htm>
52. OECD. Regulatory policy. Retrieved from <https://www.oecd.org/gov/regulatory-policy/>
53. Putri, M., Xu, C. & Akwetteh, L. (2020). Financial Behavior during COVID-19: Cognitive Errors That Can Define Financial Future. Retrieved from <https://www.scirp.org/journal/paperinformation.aspx?paperid=103763>
54. Ray, P. P. (2023). ChatGPT: A comprehensive review on background, applications, key challenges, bias, ethics, limitations and future scope. Retrieved from <https://www.sciencedirect.com/science/article/pii/S266734522300024X>
55. Singapore Legal Advice (2021). Copyright Law in Singapore: Your Rights and What It Protects. Retrieved from <https://singaporelegaladvice.com/law-articles/copyright-law-in-singapore/>
56. Smits, J. & Borghuis, T. (2022). Generative AI and Intellectual Property Rights. Retrieved from https://link.springer.com/chapter/10.1007/978-94-6265-523-2_17
57. Sobabe, A., Djara, T. & Vianou, A. (2020). Biometric System Vulnerabilities: A Typology of Metadata. Retrieved from https://www.astesj.com/publications/ASTESJ_050125.pdf
58. TechTarget (2023). What is Generative AI? Everything you need to know. Retrieved from <https://www.techtarget.com/searchenterpriseai/definition/generative-AI>
59. The Association of Banks in Singapore. ABS Cloud Computing Implementation Guide 2.0. Retrieved from <https://abs.org.sg/docs/library/abs-cloud-computing-implementation-guide.pdf>
60. The Banker (2023). Generative AI could save banks billions. Retrieved from <https://www.thebanker.com/Generative-AI-could-save-banks-billions-1688025535>
61. The Straits Times (2017). Global banks have paid \$453b in fines. Retrieved from <https://www.straitstimes.com/business/global-banks-have-paid-453b-in-fines>

62. The Washington Post (2018). A guide to the financial crisis – 10 years later. Retrieved from https://www.washingtonpost.com/business/economy/a-guide-to-the-financial-crisis--10-years-later/2018/09/10/114b76ba-af10-11e8-a20b-5f4f84429666_story.html
63. World Economic Forum (2023). The European Union's Artificial Intelligence Act – explained. Retrieved from <https://www.weforum.org/agenda/2023/06/european-union-ai-act-explained/>
64. Zhuo, T. Y., Huang, Y., Chen, C. & Xing, Z. (2023). Red teaming ChatGPT via Jailbreaking: Bias, Robustness, Reliability and Toxicity. Retrieved from <https://arxiv.org/abs/2301.12867>



Acknowledgements

MAS

Sopnendu Mohanty, Project Sponsor, Chief FinTech Officer

Dr Xuchun Li, Project Director, Head of AI Development Office

Qiang Zhang, Project Lead, Deputy Director AI Development Office

Accenture

James Gan, Managing Director, Financial Services

Bhavna Rawley, Data Science Principal Director, Responsible AI

Ali Shah, Global Principal Director, Responsible AI

Septian Livan, Senior Technical Architect

Tianyou Zhang, Data Science Lead

Chayanika P. Narula, Data Science Lead

Kai Kathy Shen, Visualisation & Interactive Science Lead

Dr Deepa Rajagopalan, Data Science Manager

Rupini S. Pandian, Data Science Consultant

Garialdi Salim, Business Analyst

Hunter McGuire, Responsible AI Analyst

Citi

Prag Sharma, Director, Global Head - Artificial Intelligence Centre of Excellence (AI CoE)

Jack Ow, Assistant General Counsel, Intellectual Property and Operations & Technology

Richard Lomas, Senior Vice President, Global Government Affairs

Ian Micallef, Managing Director, Technology

Raja Sudalaimuthu Sudalaiandi, Senior Vice President, Engineering & Architecture Practice – Citi Global Wealth Technology

Shankar Rao, Director, Engineering & Architecture – Citi Global Wealth Technology

Ronnie Hugo, Senior Vice President, Technology Infrastructure – Data Science Group Manager

DBS

Sameer Gupta, Chief Analytics Officer

Simon Chater, Lead - Data Management

Sandhya Jilson, Lead - AI Governance

Florent Juglair, Lead - Data Architecture & Platform Evolution

Eric Han Siong Chua, Lead - Cybersecurity

Wee Peng Teo, Lead - Data, Support Units and Operations, Legal & Compliance

Genevieve Low, Data, Support Units Legal & Compliance

Juon Qiang Yin, Head Group Investigations

Google

Chandrika Kadirvel Mani, AI/ML Tech Practice Lead, Southeast Asia, Google Cloud

Chester Chua, Head of APAC Financial Services Policy; Head of APAC AI Policy, Google Cloud

Derrick Yeo, Partner Customer Engineer, Google Cloud

Vikas Desai, Principal Architect, Google Cloud

HSBC

Mohamad Khalil, AI Research Analyst

Jai Geelal, AI Research Analyst

Ronnie Chung, Data and AI Ethics Lead

Bruno M De Oliveira, Data Scientist

Priyak Bandyopadhyay, Data Scientist

Jorja Jane Kirk, Technology Graduate

Dan T. Dixon, Head of Data and AI Innovation

Michael Argy, Senior Legal Counsel

Li-En Wee, Associate General Counsel

Jen Yan Loy, Resilience Risk, Senior Manager

Adam Matthew Weaver, Singapore Head of Data & Analytics Office

Amrita Raje, Data Governance

Joshua Woo, Senior Legal Counsel

Microsoft

Dr Julia Gusakova, Senior Cloud Solution Architect

Poonam Brijesh Sampat, Senior Cloud Solution Architect

Mavis Yee, Account Director (Public Sector)

Juan Madera Jimenez, FI Sales Leader

Prasanna Nandhakumar, Account Technology Strategist

Matthew Blume, Client Technology Leader

Anil Mathur, Director Partner Development

Jeth Lee, Chief Legal Officer



OCBC

Adrien Chenailier, Head of Data Science

Donald MacDonald, Head of Group Data Office

Andrea Pisoni, Head of Data

Standard Chartered

Adhinarayan Nammalvar, Director, Data Conduct: RAI & Data Ethics

Vijay Jairaj, Exec Director, Data Conduct: RAI & Data Ethics

Martin Kay, Global Head, Data Quality and Responsible AI - CDO

Carlos Queiroz, Global Head Data Science Engineering, DCDA, CCIB

Emily Yang, AI & Innovation Lead, Human Resources

Dor Kedem, Head of Analytics and AI, CPBB

Siddhartha T., Executive Director, Head - Artificial Intelligence

Grace Foo, Director, CFCC Governance

UOB

Richard Lowe, Chief Data Officer

Alvin Han Wen Eng, Head, Enterprise AI

Robert Tong Ngee Cheah, Enterprise AI

Rishiraj Singh, Head, Enterprise Data Governance

Davina Tauro, Enterprise Data Governance

Ryan Hansen-Reeder, Enterprise Data Governance

Sruthi Basani, Enterprise Data Governance

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.