MINDFORGE

AI Risk Management: Operationalisation Handbook

January 2026

MINDFORGE

The MindForge Consortium

MAS Monetary Authority of Singapore

AIA HEALTHIER, LONGER, BETTER LIVES · BlackRock · citi · DBS · eastspring investments A Prudential plc company

Great Eastern An OCBC Company · GIC · GXS · HSBC · HSBC Life

income made yours · Julius Bär · Manulife · Maybank · MSIG

MUFG · Munich RE · OCBC · PRUDENTIAL · SMBC

standard chartered · STATE STREET · UBS · UOB

aws · Google Cloud · Microsoft · NVIDIA

Supported by: accenture

With participation from:

abs The Association of Banks in Singapore · General Insurance Association · imas Investment Management Association of Singapore · Life Insurance Association Singapore Life is worth protecting. Invest in it. · SFA Singapore FinTech Association

Contents

Contents

Figures

Tables

Illustrations from Financial Institutions

Opening Points

Introduction

Artificial Intelligence (AI) is one of the most impactful technologies to be adopted in the financial services industry in recent years. It presents a range of opportunities for efficiency and better experiences for employees and customers, but also presents a range of risks if not well-governed. Project MindForge, of which this Handbook is a part, was launched in 2023 as the continuation of a multi-year legacy of proactive industry collaboration to address the responsible use of AI with the long-term leadership and support of the Monetary Authority of Singapore (MAS).

MAS issued the 14 FEAT (Fairness, Ethics, Accountability, Transparency) Principles for responsible AI use in the financial services industry in 2018.[1] Following the introduction of the FEAT principles, the Veritas Initiative was established by MAS and a consortium of financial institutions (FIs), consultancies, and technology companies to operationalise those principles.[2] Between 2020 and 2023, the Veritas Initiative co-created guidance to help FIs evaluate real-world solutions using AI and data analytics against the FEAT Principles, producing a Methodology and Toolkit. This Methodology is now used by FIs to implement AI and data analytics solutions responsibly.

Following significant advancements in Generative AI (Gen AI) technology in late 2022, the industry felt it necessary to assess the risks of this new technology, examine the FEAT Principles and Veritas Methodology to confirm their applicability to these risks, and adapt them where needed. This was the focus of Phase 1 of Project MindForge, which had four key outcomes. The first was to produce the first financial industry-specific taxonomy of Gen AI risks. The second was to review the FEAT Principles and the Veritas Methodology against these risks, identifying areas where additional considerations may be required to address the unique characteristics of Gen AI. The third was to provide an overview, based on the state of the art at that time, of the architectural and infrastructure considerations around responsibly developing and deploying Gen AI. The fourth was to develop two practical Gen AI use cases on risk management and compliance that demonstrated the application of responsible principles to new Gen AI tools. Phase 1 concluded with the publication of a whitepaper on the emerging risks and opportunities of Gen AI for banks in May 2024.[3] FIs are using the MindForge whitepaper to adapt their AI governance and risk management approaches to the challenges of Gen AI.

The Association of Banks in Singapore (ABS) elaborated on the MindForge whitepaper to publish a Handbook on Generative AI Guardrails in Banking in May 2025. This Handbook focuses on a selection of risks highlighted in the MindForge whitepaper and proposed specific, tangible guardrails for addressing them. Its work was an important input in the development of this Handbook.

Phase 2 of Project MindForge formally kicked off in November 2024. Its mission is to enable and facilitate FIs, at different levels of AI maturity, to scale AI with trust by adopting and operationalising AI governance and risk management across the enterprise: supporting industry AI use that is rapid, but responsible.

This Handbook, the first part of Phase 2 of Project MindForge, draws on the FEAT Principles, the work of the Veritas Initiative, and the risks identified in Phase 1 of Project MindForge to create a comprehensive guide to AI governance and risk management for the industry. It extends the work of MindForge Phase 1 to include traditional AI, Gen AI, and more recent technologies such as Agentic AI. Harmonising the range of good practices across the ecosystem into one Handbook will help make AI governance and risk

[1] Read the FEAT Principles at https://www.mas.gov.sg/publications/monographs-or-information-paper/2018/feat
[2] Read about the Veritas Initiative, and see its publications, at https://www.mas.gov.sg/schemes-and-initiatives/veritas
[3] Read the story of Phase 1 of Project MindForge, and the 2024 industry whitepaper, at https://www.mas.gov.sg/schemes-and-initiatives/project-mindforge

management straightforward, systematic, and universal. This Handbook is intended to accompany and support the implementation of the proposed MAS Guidelines on Artificial Intelligence Risk Management.

Industry alignment on the Handbook began in advance of the public kick-off in September 2024, when the consortium's 24 primary members endorsed the Handbook's scope and structure. The consortium convened three Working Groups made up of the primary members, tech partners, and a consulting partner, which drafted the text of the Handbook between February and July 2025. The scope and draft text were also reviewed and supported by the members of financial industry associations in Singapore through four engagement sessions in February, June, August, and November 2025, and two rounds of open feedback in November-December 2024 and August-September 2025. Over 100 financial organisations outside the primary consortium were engaged through these activities. After taking that feedback into account, this Handbook was formally launched at the Singapore Fintech Festival in November 2025.

The Handbook consists of three documents, of which this is the second. These three documents are:

AI Risk Management Executive Handbook. This document provides Considerations and Implementation Practices for governing AI across each Section in the Handbook's scope. It is intended as a resource for executives in the financial services industry.

AI Risk Management Operationalisation Handbook (This Document).
This document provides detailed guidance on the operationalisation of each of the Implementation Practices recommended under each of the Handbook's Consideration. It includes illustrations of good practices from primary members, appendices, and other supporting materials.

AI Risk Management Handbook Implementation Examples. This document provides detailed case studies on individual financial institutions' experiences implementing AI governance and risk management.

These three documents are meant to be used in conjunction, and together make up the MindForge AI Risk Management Handbook.

MindForge is founded on a commitment to using AI responsibly, in a manner that manages its risks while leveraging its benefits. Governance and adoption are not, as concepts, in tension; in fact, widespread, rapid, and useful innovation in AI requires robust risk management and good governance. FIs that responsibly manage the risks of AI will be able to transform their businesses with the confidence that new technologies will behave as intended, follow the law, be secure, and protect their (and their employees and customers') data. It accelerates innovation and supports value realisation through measures that support observability, controllability, and oversight. It allows customers, employees, stakeholders, and society to trust FIs that use AI because they are confident that the technology's use will be fair, ethical, accountable, and transparent. Rather than impeding AI innovation, the practices outlined in this Handbook will enable it.

Acknowledgements

We would like to extend our thanks and recognition to the participants in the development of this Handbook.

Project MindForge is a collaborative industry initiative led by the Monetary Authority of Singapore (MAS) and delivered cooperatively by a consortium of FIs across the banking, insurance and capital market sectors, supported by consulting and technology partners and industry associations. Several members volunteered to take on additional responsibilities as leads and co-leads of the MindForge consortium's three working groups, which wrote the text of the Handbook.

The primary members of the MindForge consortium are (in alphabetical order):

- AIA
- BlackRock (Co-Lead for the "Data & AI" and "Enterprise" Working Groups)
- Citi
- DBS (Lead for the "Data & AI" Working Group)
- Eastspring Investments
- GIC
- Great Eastern Life
- GXS Bank
- HSBC
- HSBC Life
- Income Insurance (Co-Lead for the "Data & AI" Working Group)
- Julius Baer (Co-Lead for the "Data & AI" Working Group)
- Manulife
- Maybank
- MSIG
- MUFG Bank
- Munich Re
- OCBC
- Prudential (Co-Lead for the "Risk & Compliance" Working Group)
- SMBC
- Standard Chartered Bank (Lead for the "Enterprise" Working Group)
- State Street
- UBS
- UOB (Lead for the "Risk & Compliance" Working Group)

The development of the handbook was supported by the consortium's consulting partner:

- Accenture

The consortium was advised by its technology partners, which are:

- Amazon Web Services
- Google Cloud
- Microsoft
- Nvidia

The consortium would like to acknowledge the contributions made by the approximately 200 individuals from the aforementioned institutions who participated in the Handbook's drafting.

Several financial industry associations and their membership provided guidance, inputs, feedback, and support throughout the development process. The five associations that were members of the consortium are:

- The Association of Banks in Singapore
- General Insurance Association of Singapore
- Investment Management Association of Singapore
- Life Insurance Association Singapore
- Singapore FinTech Association

The consortium would like to acknowledge the contributions of these participating organisations and the numerous peers in the industry who contributed to this Handbook's successful development and refinement.

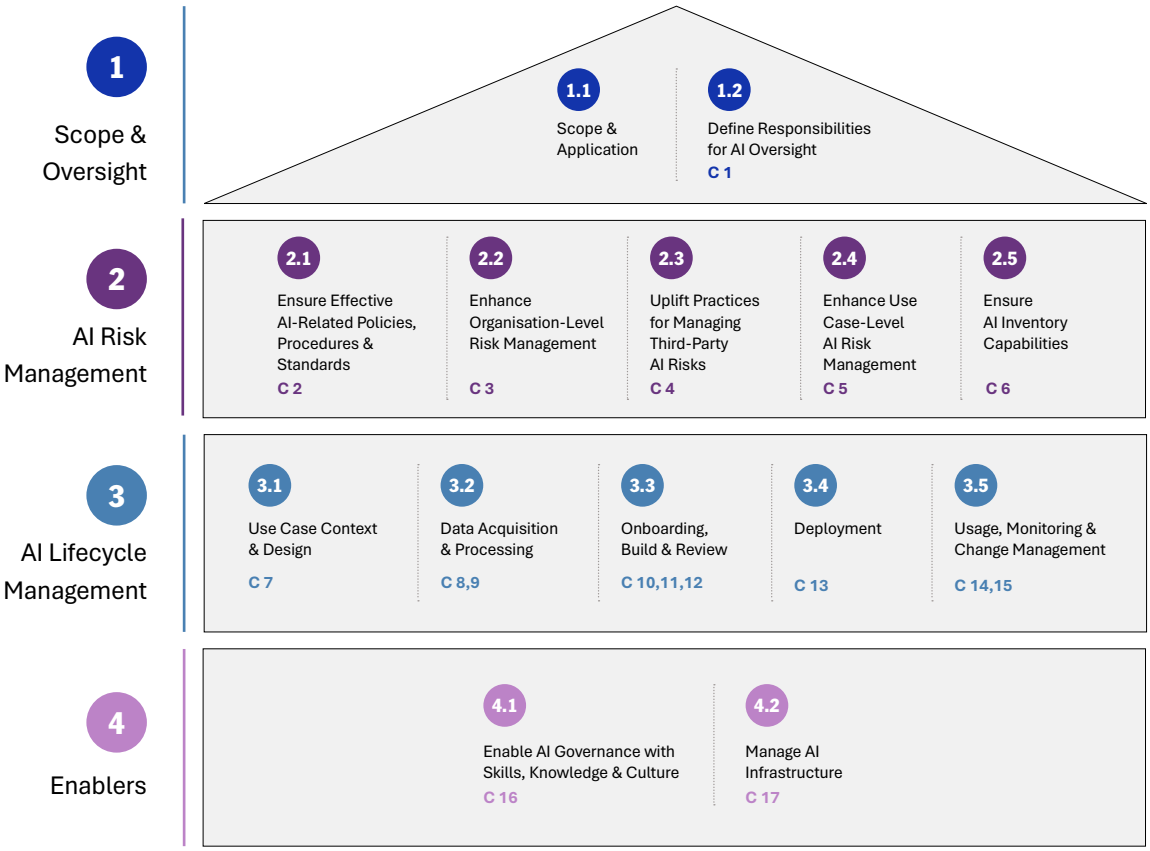
Part 2

MindForge AI Risk Management and Governance Framework

MindForge AI Risk Management and Governance Framework

Each Section in this Handbook corresponds to a component in the framework below, which represents a logical model to support FIs in implementing the Handbook.

Figure 0.0.1: MindForge AI Risk Management and Governance Framework



C – Considerations included in that Subsection

Section 1, **Scope and Oversight**, discusses some of the foundational concepts underpinning this Handbook. It also discusses how AI is overseen in an FI's operating model.

Section 2, **AI Risk Management**, discusses how FIs can measure, monitor, and mitigate the risks of AI by establishing policies, procedures, processes, and systems in their organisation.

Section 3, **AI Lifecycle Management**, discusses the key activities and considerations that can be applied to manage risks at each stage of the lifecycle of an individual AI use case.

Section 4, **Enablers**, discusses the foundational capabilities that can support AI risk management.

1. Scope and Oversight

1.1 Scope and Application

Organisations in Scope

This Handbook is addressed to all Financial Institutions (FIs) that use AI in their businesses.

This Handbook is addressed to FIs of all sizes. Although it makes specific references to Singapore's context, it is also intended to be globally relevant. As such, it aims to make recommendations that are aligned to global good practices beyond Singapore.

The Handbook is intended to be relevant to all functions within those organisations.

AI-Specific Risks in Scope

This Handbook defines practices for addressing AI-specific risks – new or enhanced risks arising from AI use that go beyond those of traditional software use in the context of an FI. The overall approach of this Handbook is to describe AI governance and risk management that is risk-based and proportionate, and that continuously improves as lessons are learned and as risks evolve. This Handbook addresses the risks posed to FIs by the use of AI in their value chain – in their business, their vendors, and their service providers, such as contractors or consultants. It does not consider risks posed to FIs by the use of AI by unrelated parties or external malicious actors.

This Handbook is designed to supplement and function in tandem with existing non-AI-specific risk management practices that FIs already have in place. These include practices for managing technology risk, cybersecurity, and risk management, many of which are governed by instruments listed in Appendix C of the Operationalisation Handbook. Where relevant, these non-AI-specific risks are referred to here, but are not replicated in this Handbook.

This Handbook aims to address governance and risk management for all types of AI, including traditional AI, Gen AI, and Agentic AI. In addition to the well-known risks of traditional AI, Gen AI and Agentic AI may introduce new or additional risks or challenges.

The Handbook took the work of MindForge Phase 1, which developed an AI risk taxonomy, as its starting point; an updated version of this taxonomy is provided in Appendix B of the Operationalisation Handbook. From this list of risks, the Association of Banks in Singapore (ABS) Handbook on Generative AI Guardrails in Banking highlighted ten "top risks", which were a particular focus:

- Unrepresentative or biased data inputs.
- Toxic and offensive outputs.
- Lack of AI risk awareness.
- Lack of use case, data and model governance.
- Inadequate human oversight.
- Inadequate feedback and recourse mechanisms.
- Hallucination/ Fabrication/ Confabulation.
- Overconfidence.
- Insufficient model accuracy/ soundness.
- Model degradation from unexpected use.

The landscape of AI risk is rapidly evolving, however, and in addition to the AI risk taxonomy in this Handbook, it is important that FIs and the industry overall can continue to consider new, emerging, or diminished AI-specific risks as they apply the Considerations in this Handbook.

Intended Audience

Within an FI, the Handbook is addressed specifically to:

- **Executives:** Decision-makers and leaders.
- **Builders:** Software developers, data engineers, data scientists, AI practitioners, systems integrators, and other technical specialists involved in the development, deployment, and use of AI.
- **Custodians:** Employees in oversight, governance, enablement, and risk management roles in an FI who apply AI governance and risk management policies and procedures and manage AI risks, either directly or in an enabling capacity such as talent, legal, or technology.
- **Use Case Owners:** Employees who are accountable for an AI use case.
- **Business Users:** Employees who use or apply AI use cases in the course of their business responsibilities.

Each FI may organise these functions differently, and under different titles, depending on its structure and needs (such as the concepts of the first, second, and third lines of defence). This handbook uses the above terms as generic terms of reference to common enterprise activities and aims to be relevant to all FIs irrespective of their internal organisation.

Structure of the Handbook

The Handbook is made up of 17 Considerations. These are the operative unit of the Handbook; each Consideration is a thematic recommendation that will support an FI in operationalising AI governance and risk management. Together, they form a checklist of actions that FIs can take to align with the approach in this Handbook. These Considerations are grouped thematically into several Subsections, which each begin with a brief overview that introduces its key concepts.

Under each Consideration are one or more Practices. These are specific actions that, when taken collectively and appropriately to the FI's context, can implement the Consideration. Each Practice is accompanied by a long-form text – the "Operationalisation Guidelines" – which describes that Practice in more detail. That text is included in this document, but omitted in the Executive Handbook.

Several Subsections are also enriched with an illustration contributed by a consortium member describing how the Practices in that Subsection are implemented in a real-world setting. These illustrations appear at the end of the Subsections where they are included.

Figure 1.1.1: Structure of Each Handbook Section

Define Responsibilities for AI Oversight

Consideration 1: Ensure that an AI governance operating model is clearly defined by leveraging and, as needed, uplifting the roles and capabilities of existing enterprise functions including the relevant roles from the Board, Senior Management, and operational governance, with sufficient operating effectiveness measures in place to support them.

Practice 1: Embed additional responsibilities for AI governance and risk management, as required, in relevant Board and Senior Management roles.

Operationalisation Guidelines:

Approach:

- Extend the roles and responsibilities of relevant Board members or bodies to include relevant AI-related actions, including the endorsement of key AI governance documents, ensuring that AI-specific skills are in place, and ensuring that AI risks are managed.
- Extend relevant existing Senior Management roles and responsibilities to include the implementation of effective AI governance and keeping the Board well-informed.

The roles of the Board and Senior Management are already well-defined in each FI; their core responsibilities for managing risk are described in Principle 11...

...

Relationship to Existing Enterprise Functions

AI is one of many technologies that FIs use, and FIs already have extensive enterprise functions in place for providing oversight, governance, and risk management beyond AI. Continuing to apply existing good practices in areas like risk management, data governance, procurement, software lifecycle management, talent, and cybersecurity is a key foundation for this Handbook – and for responsibly and effectively governing AI.

This Handbook was designed with the following goals in mind:

1. To build on, but not duplicate or replace, existing industry frameworks and good practices that apply to AI but are not AI-specific.
2. To describe only those considerations that are unique or additional to AI governance and risk management.

As a result, this Handbook does not describe activities or practices that are not specific or additional to the governance of AI. This should not be taken to imply that existing non-AI-specific practices are not also important to AI governance and risk management. FIs can continue to apply industry norms and good practices in risk management, data governance, procurement, software development, and cybersecurity, and can consult Appendix C for a list of other frameworks that were referenced in the development of this Handbook.

Relationship and Proportionality to Risk

This Handbook is based on the principle of proportionality between governance measures and AI risk. This risk-based approach underpins all of the Considerations in this Handbook and emphasises scaling risk management activities based on factors such as the nature of the FI's business, the scale and nature of its AI use, and its appetite for risk. This balanced approach encourages innovation and experimentation while focusing resources on managing the most significant risks.

The Handbook's Considerations are also based on the principle of relevance. Not all practices for AI governance and risk management will be feasible for all AI use cases, depending on the technologies used, the deployment pattern, or other situational factors. FIs can determine in context how to make this Handbook's recommendations relevant to their use of AI.

Relationship to FEAT and Veritas

The MindForge Handbook supports and builds upon the foundation of the FEAT Principles. These fourteen principles (logically grouped into Fairness, Ethics, Accountability, and Transparency) were issued in 2018 and continue to serve as a valuable reference to the industry. The overall direction set by FEAT remains an underlying philosophy underpinning the Considerations in this Handbook, and an indicative mapping to the individual principles of FEAT is provided in Appendix D. This Handbook is written to facilitate adherence to the FEAT Principles when its Considerations are applied.

The methodology developed by the Veritas Initiative is a detailed, widely accepted framework for implementing the FEAT Principles in practice. This methodology remains pertinent and effective in managing AI risk today; like FEAT, however, the fast-moving nature of the field means that practices have evolved substantially since it was written. Veritas, most notably, was not drafted with Gen AI or Agentic AI in mind and is focused on the governance of AI models, not AI use cases or the enterprise overall.

Major advances in Gen AI and Agentic AI since these frameworks were developed, however, have made it challenging to apply them as written to modern AI use cases. This Handbook is written with the intention of going beyond FEAT and Veritas to manage the risks of advanced AI, such as by extensively addressing the governance of the enterprise; viewing AI use cases holistically, rather than at the model level; and by considering a wider range of risks. The seven risk dimensions highlighted in Appendix B reflect this broader view.

Figure 1.1.2: Diagram of FEAT, Veritas, and MindForge

FEAT Principles

Foundational guidance that underpins Veritas and MindForge (2018)

Veritas Initiative

Detailed methodology for implementing FEAT (2020-23)

Project MindForge

Updated and expanded approach across the enterprise (2023-)

Relationship to MAS Guidelines on AI Risk Management

This Handbook is intended to accompany and support the implementation of the proposed MAS Guidelines on Artificial Intelligence Risk Management. A mapping of this Handbook's Considerations to the proposed Guidelines is provided in Appendix D. A non-exhaustive list of other MAS frameworks that may be relevant in applying this Handbook is provided in Appendix C.

Relationship to Other Regulations

Leading global frameworks played an important role in the development of this Handbook, and are enumerated in Appendix C. The practices in this Handbook are closely aligned to global AI governance and risk management norms, and the Handbook overall can serve as a useful repository of industry leading practices that FIs around the world can consider. Adopting the Considerations in this Handbook can also support, but may not be sufficient to address, compliance requirements in other jurisdictions where AI is highly regulated. FIs may refer to existing data protection, risk management, and market conduct rules for non-AI-specific obligations that will continue to apply when using AI.

Scope of Artificial Intelligence (AI)

MAS' proposed Guidelines on Artificial Intelligence Risk Management ("Guidelines") defines AI as follows:

> "AI includes use cases involving models or systems that learn and/or infer from inputs to generate outputs such as estimates, predictions, content, summaries, recommendations, or decisions that may influence physical or virtual environments, and vary in their levels of autonomy and adaptiveness after deployment. Calculators or tools whose outputs are solely based on predefined programming logic or rules would not be regarded as AI for the purpose of these Guidelines."[4]

The proposed Guidelines also highlight that the definition of AI "would generally include AI based on machine learning, deep learning, reinforcement learning, as well as Generative AI, AI agents and any newer AI technologies."

For the purposes of this Handbook, the MindForge consortium focuses on two key characteristics to help practitioners consistently identify and manage AI:[5]

- AI learns and/or infers from inputs[6] to generate output such as estimates, predictions, content, summaries, recommendations, or decisions.
- AI excludes calculators or tools whose outputs are solely based on predefined programming or rules.

These characteristics are intended to help generalise the definition in the proposed Guidelines for easy and practical operationalisation.

A non-exhaustive list of examples of technologies that are included in this scope of AI are:

- Linear or logistic regression models.
- Models using machine learning and its derivative techniques such as deep learning, transformers (LLMs), and diffusion models.
- Computer vision models, including optical character recognition features that use computer vision models.
- Virtual assistants based on language models.
- Agent and agentic systems that use language models and other AI tools.

A non-exhaustive list of examples of technologies that would not be included are:

- Traditional rule-based software.
- Macros and other deterministic automations, including some types of Robotic Process Automation (RPA).
- Chatbots that are menu-based or rules-based, such as those using keyword identification.
- Pre-defined data processing logic.

[4] MAS' definition is broadly aligned with the OECD's definition of AI: "An AI system is a machine-based system that, for explicit or implicit objectives, infers, from the input it receives, how to generate outputs such as predictions, content, recommendations, or decisions that can influence physical or virtual environments. Different AI systems vary in their levels of autonomy and adaptiveness after deployment."

[5] This scope is intended to be broad, and would also include systems with such characteristics that may influence physical or virtual environments, and vary in their levels of autonomy and adaptiveness after deployment.

[6] Inputs, in this context, can refer to inputs or data received during training or fine-tuning and/or to inputs received at runtime (such as user prompts).

Definition of AI Governance and Risk Management

AI governance and risk management are, together, a set of interdisciplinary activities for:

1. Managing AI-specific risks posed to the enterprise, its stakeholders, and society.
2. Ensuring that the use of AI adheres to an FI's principles, especially Fairness, Ethics, Accountability, and Transparency.
3. Ensuring that the use of AI conforms to relevant laws and regulations.

AI governance and risk management supplement, and do not replace, existing domains of enterprise management. These typically includes (but is not limited to) model governance, data governance, technology risk management, third party risk management, and risk culture.

Other terms – Responsible AI, Trustworthy AI, and AI Safety – are also commonly used in the industry to refer to many of the activities involved in AI governance and risk management. These terms have a variety of definitions currently in use in the ecosystem. Where this Handbook uses the term AI governance and risk management, FIs can understand it to refer to a broadly similar set of considerations and activities as these other common terms in use throughout the industry.

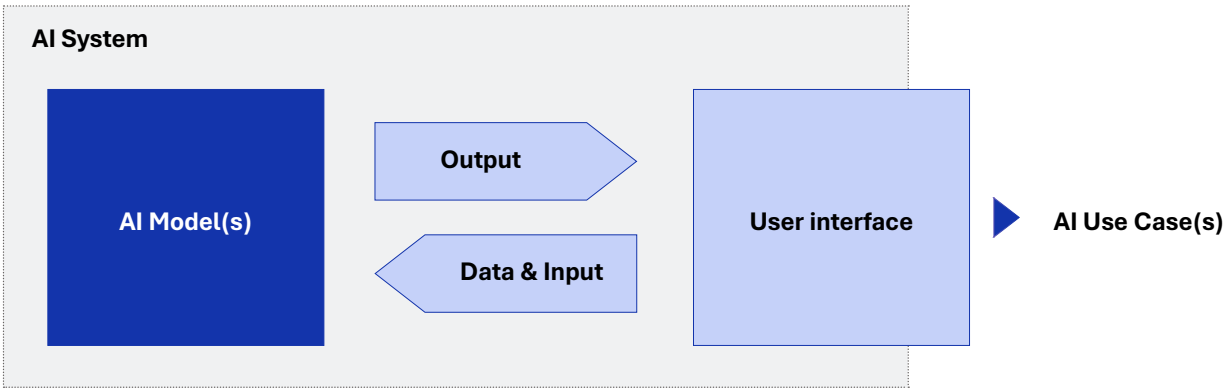
Some FIs will establish dedicated AI governance teams or AI governance or risk management processes. Others will make changes to existing teams and processes to ensure that they can perform the activities of AI governance and risk management, or may adopt a hybrid approach, where a dedicated team handles some AI governance and risk management activities (or all AI governance and risk management activities for a subset of AI), and existing teams handle others. This Handbook does not prescribe a specific approach.

AI continues to be subject to the same principles, regulations, and good practices as existing technologies; the role of AI governance and risk management, in that context, is to address incremental AI-specific risks and practices for managing them. AI governance and risk management are additional to, and not a substitute for, existing technology governance.

Definition of AI Model, System, and Use Case

The practical implementation of AI involves three elements: a model, a system, and a use case.

Figure 1.1.3: AI Model, System, and Use Case



Adapted from the Explanatory memorandum on the updated OECD definition of an AI system [19]

In a real business context, implementations of AI can be considerably more complex, sometimes involving multiple models and a range of software components operating as a single system or as an orchestrated collection of systems. The glossary (Appendix A) includes a definition of Traditional AI, Gen AI, and Agentic AI. Agentic AI, which includes complexity beyond the diagram above, is discussed further under Future Perspectives.

Model

An AI model, which is a mathematical or logical representation mapping inputs to outputs, is the foundation of modern AI technology. An AI model is any model which meets the definition of AI.

Models output estimates, forecasts, or projections related to real-world phenomena.

> E.g. A function predicting the likelihood of a borrower to default based on their credit history is a model.
>
> E.g. A Large Language Model (LLM) that predicts word tokens is a model.

System

An AI system includes an AI model as well as other software components that allow models to be used in real-world applications. A system can contain many software features, and potentially many models that interact via software.

> E.g. A spreadsheet containing user data and a statistical model saved as a JSON file, when accessed through an integrated development environment, are together a system.
>
> E.g. An AI chatbot that a user interacts with using a message box, and which uses an AI search engine to retrieve files based on user prompts, is a system.
>
> E.g. An agentic workflow consisting of multiple orchestrated AI "agents" that collaborate to interact with the user and influence the environment, is a system.

Existing Information Technology (IT) practices in the financial services industry define what is considered to be a single "system". In some cases, a large software system may have multiple functionalities, only a subset of which are supported by AI – such as a customer relationship management system that includes a Gen AI feature for summarising notes and emails. For the purposes of this Handbook, AI governance and risk management would apply only to those outputs or behaviours of the system which are affected or influenced by its AI components or functionalities.

Use Case

An AI use case is the specific, real-world context in which an AI system is intentionally used. Use cases are crucial for understanding the impact, risks, and functioning of systems.

> E.g. Using AI to consider several data points about a customer to underwrite loan risk is a use case.
>
> E.g. Searching for enterprise risk management documents based on user queries and then referring to those documents to answer employee questions using a natural language processing technique is a use case.

Use of Model, System, and Use Case in Governance

The model, system, and use case are each integral considerations in identifying, assessing, managing, and monitoring AI risks. AI governance and risk management traditionally focused on models, which was sufficient for governing traditional AI and the range of use cases that were common at the time. With the advent of Gen AI and Agentic AI, however, AI has become increasingly general-purpose; more than ever, the risks associated with an AI model are contingent on the nature of its chosen use case and the guardrails built into its system. Gen AI and Agentic AI systems may also involve several models, each highly complex, working in concert. The risks of those models may be impossible to assess or mitigate in isolation from the broader architecture of the system and parameters of the business use case.

This Handbook refers to "AI use cases" as the basic unit of AI governance and risk management, reflecting the increased importance of considering factors beyond the model alone. Readers should understand this to include associated AI systems and models, which for the purposes of governance are inseparable from their use case. While the term "AI use case" is used throughout, this Handbook notes that FIs may each choose the basic unit of AI governance and risk management in their own context, so long as when doing so they effectively consider the relationship between models, systems, and use cases.

When necessary to distinguish them, the Handbook specifically refers to AI "models" or "systems".

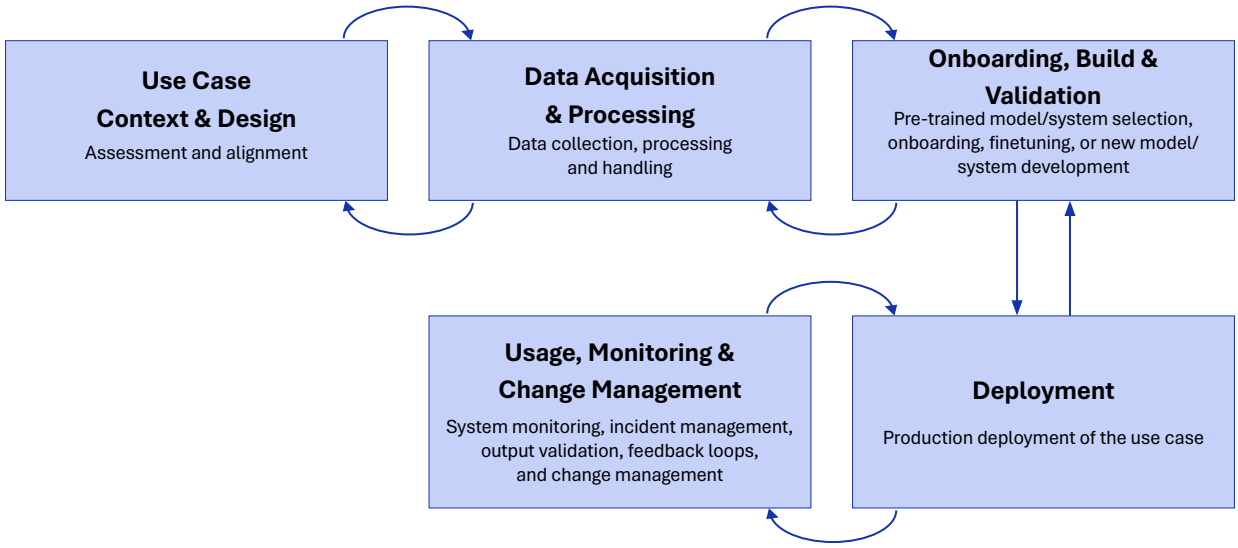
MindForge AI Lifecycle

The AI lifecycle consists of five stages:

- Use Case Context and Design.
- Data Acquisition and Processing.
- Onboarding, Build & Validation.
- Deployment.
- Usage, Monitoring & Change Management.

These stages do not always proceed in a linear fashion, and the progression between them will frequently be an iterative process of continuous improvement. The relationship between these stages is illustrated in Figure 1.1.4.

Figure 1.1.4: AI Lifecycle



1.2. Define Responsibilities for AI Oversight

In FIs that use AI, AI governance and risk management are integral components of their overall management of risk, which includes technology, security, and data risks. It is generally not a standalone capability. Responsibility for the governance and oversight of AI is often integrated with existing governance functions, where it forms a part of the FI's overall operating model. Each FI typically determines, in its own business context, the best way to uplift that operating model to oversee the unique risks and challenges of AI.

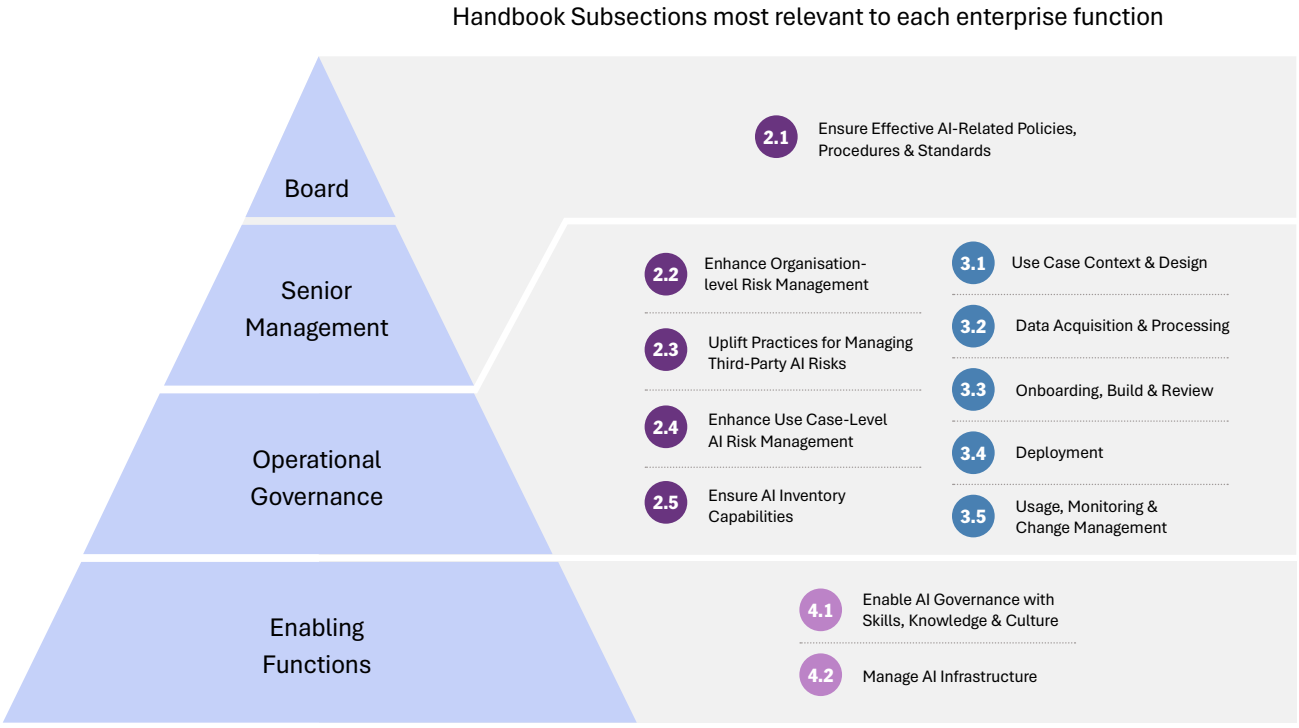
The oversight functions described in this Subsection are responsible for operationalising the FI's overall operating model, which also includes AI risk management processes (discussed in Section 2), AI lifecycle management (discussed in Section 3), and the culture, talent, and infrastructure that support AI governance and risk management (discussed in Section 4).

AI oversight begins with the existing roles within an FI; these are typically defined by industry standards like those documented in Appendix C. This Subsection considers the following roles:

- The Board and Senior Management.
- Operational governance, which is typically organised into three lines of defence (3LOD) including "1) business units responsible for daily risk management, 2) an independent corporate risk management function providing oversight and support, and 3) independent assurance functions like internal audit."[7]

These existing roles already have well-defined responsibilities and structures in place. FIs can assess their effectiveness and augment them with additional organisation or responsibilities, as needed, to ensure that they can effectively oversee AI.

Figure 1.2.1: Stylised View of the Enterprise and Related Handbook Subsections



[7] The three lines of defence framework is outlined in the latest version of the Basel Committee on Banking Supervision's Principles for the Sound Management of Operational Risk.

While not every FI has a need for employees dedicated full-time to AI governance and risk management or dedicated AI-governance forums, every FI that uses AI can benefit from having key oversight functions in place.

Consideration 1

Ensure that an AI governance and risk management operating model is clearly defined by leveraging and, as needed, uplifting the roles and capabilities of existing enterprise functions including relevant roles from the Board, Senior Management, and operational governance, with sufficient operating effectiveness measures in place to support them.

Practice 1: Embed additional responsibilities for AI governance and risk management, as required, in relevant Board and Senior Management roles.

Approach:

- Extend the roles and responsibilities of relevant Board members or bodies to include relevant AI-related actions, including the endorsement of key AI governance and risk management documents, ensuring that AI-specific skills are in place, and ensuring that AI risks are managed.
- Extend relevant existing Senior Management roles and responsibilities to include the implementation of effective AI governance and risk management and keeping the Board well-informed.

The roles of the Board and Senior Management are already well-defined in each FI; their core responsibilities for managing risk are described in Principle 11 of the Singapore Corporate Governance Code. For further guidance on these non-AI-specific elements of enterprise governance, FIs can consult expectations published by MAS and by industry groups.[8] It is important that FIs involve relevant individuals or bodies from their Boards and Senior Management in AI governance and risk management; as AI becomes more deeply embedded into their businesses, its proper functioning will be increasingly integral to the FI's ability to achieve its results across business domains.

In the context of AI governance and risk management, the role of the Board or its designates include responsibility for the overall adequacy of AI-related governance measures. This could include, directly or indirectly, the oversight and endorsement of AI-related policies and AI-specific processes like AI identification (discussed in detail in Subsection Section 2), and for the establishment and functioning of a robust AI risk management framework (discussed in detail in Subsection 2.1). The Board, or a subordinate body thereof, typically endorses the FI's AI-related principles (Subsection Section 2).

As a fast-moving field characterised by complex new and emerging technologies, relevant executives in the FI may themselves require regular upskilling in additional AI-specific knowledge; this is discussed in more detail in Subsection 4.1. The role of relevant Board members or bodies to periodically oversee and review risk governance may likewise need to be reviewed considering the rapid pace of change in the field of AI. FIs benefit from an agile, responsive approach to the design of the Board's role.

Senior Management's existing roles include ensuring that policies for risk management are sound and prudent; the impact of AI on these policies is discussed in this Handbook (see Subsections 2.1, 2.2, and 2.4). As part of their role in keeping the Board apprised of technology developments, they can consider monitoring changes to the AI landscape and the emergence of new AI risks. Senior

[8] MAS expectations include those defined in Appendix C. These industry resources may include open frameworks like TOGAF (The Open Group Architecture Framework), which is available at: https://www.opengroup.org/togaf.

Management is typically responsible for AI-related enablers like talent, culture, and infrastructure (see Subsections 4.1 and 4.2). As with the Board, Senior Management has a key role to play in the continuous improvement and evolution of AI governance and risk management.

Most FIs have a range of Board committees in place to manage key enterprise risks; these committees may include roles for Senior Management as well, especially the CEO, CIO, CDO, and CRO, where those roles exist. Typical committees with a role related to AI are committees on "Risk and Compliance", "Enterprise Architecture, and "Data Governance".

Some FIs elect to include AI within the mandate of bodies on Data Governance, reflecting the extent to which the effective management of data is an essential and inseparable part of governing AI and that a range of activities and processes are shared between the two subjects. In determining whether to do so, FIs can consider if the teams and forums involved in Data Governance have the capacity and skills to address AI-specific governance and risk management topics, or if these are best addressed in whole or in part elsewhere.

FIs could prepare these committees to address the unique challenges of AI governance and risk management by including key personnel with AI-related perspectives; they could also revise each committee's targets and mandates so that they have clearer responsibility. Effective coordination mechanisms, such as support from a Senior Management role with overall responsibility for AI governance, can ensure that each committee's work on AI governance and risk management is mutually supporting and that they are collectively comprehensive in addressing AI-specific risks.

AI's interdisciplinary nature and impacts may cut across existing organisational silos in each FI. The organisation of their Board and Senior Management may benefit from being supplemented by mechanisms for interdisciplinary coordination. The Board and Senior Management continue to hold an overall responsibility for coordinating the enterprise and may take actions to facilitate interdisciplinary AI governance and risk management.

Practice 2: Ensure that operational governance functions have clear roles and responsibilities assigned to operationalise AI governance and risk management activities across the enterprise.

Approach:

- Leverage the capabilities of existing operational governance functions like the three lines of defence, or create new bodies as needed, to ensure that responsibilities for completing all AI governance and risk management activities are clearly assigned in addition to the existing roles of governance functions.

FIs already have operational governance structures like the three Lines of Defence (3LOD) in place. Roles in the 3LOD are part of what this Handbook refers to as Custodians.[9] FIs also already have well-defined roles and responsibilities for these functions as part of their existing non-AI-specific governance of risk and technology. These include a range of governance forums and activities, such as architecture reviews, change management, and security assessment.

[9] In the terminology used in this Handbook, "Custodian" is used to refer to a range of supporting functions which can be organised differently in each FI. This includes internal audit, risk management, and technology oversight functions – typically referred to as the 3LOD – as well as supporting functions, such as an FI's technology function, talent function, or legal practice. The unifying theme in these roles is that they all provide overall, transversal support for AI governance and risk management activities.

FIs can ensure that their operational governance is prepared for all aspects of AI governance and risk management. This includes assessing end-to-end processes to determine where teams may be required to take on additional roles; they can then be provided with a clarified mandate, new procedural steps, or new capabilities to support their AI governance and risk management responsibilities. Whether FIs create new bodies within their operational governance structure (such as a "Line 1.5") to govern AI depends on their own context and the sufficiency of forums that they already have in place. In general, effective AI governance and risk management require more interdisciplinary cooperation and coordination.

Operational governance functions have a key role to play in implementing AI-related policies and procedures, as well as supporting their development and maintenance (discussed in Subsection Section 2); they also are responsible for implementing AI-related risk governance alongside their non-AI-specific responsibilities (discussed in Subsections 2.2 and 2.4). Operational governance functions maintain the FI's AI inventory (discussed in Subsection 2.5); this may include a responsibility for providing oversight of the FI's overall use of AI and the delivery of updates to Senior Management or the Board. FIs can ensure that post-deployment monitoring related to AI is well-defined in the responsibilities of the appropriate governance functions, including both overall "portfolio" monitoring and roles in the monitoring of individual use cases.

> Given AI's status as an emerging technology, Custodians in some FIs play a role in sharing expertise with and advising Builders, in addition to overseeing them. FIs that implement this approach typically balance the importance of effective organisational independence with the value of sharing AI-related expertise in organisations where it may be limited.

Practice 3: Ensure that existing governance processes, forums, assets, and tools are updated to effectively enable AI governance and risk management.

Approach:

- Leverage existing processes, assets, and tools to ensure that operational governance is equipped to manage AI-specific risks.
- Empower existing forums, or create new forums, to provide the oversight required to support AI governance and risk management.

As FIs update the roles and responsibilities of their operational governance bodies, they can also update their underlying capabilities to enable effective AI governance and risk management. These can take the form of review processes, tools like dashboards or workflow managers, and assets like templates, questionnaires, and agendas.

The management of AI use cases across the AI lifecycle is not fundamentally different from existing IT, Software Development Lifecycle (SDLC), model risk management (MRM), or business operations practices, and will leverage the processes and practices associated with each of them. This can include adding AI governance and risk management -related processes to existing workflows, adding AI-specific risk-related questions to templates or questionnaires, or incorporating AI-related subjects in the agendas or mandates of existing meetings or forums.

FIs can also consider updating their existing governance forums to review the effectiveness of their AI governance and risk management through metrics and/or qualitative indicators. In the absence of suitable forums, FIs may also create communications channels as they deem fit to internally discuss AI-related matters, including governance. FIs can periodically assess the fitness of their relevant governance processes, forums, assets, and tools as AI technologies evolve.

Practice 4: Ensure that sufficient operating effectiveness and horizon-scanning measures are in place to monitor and improve the AI governance and risk management operating model over time.

Approach:

- Monitor the performance and effectiveness of the overall AI governance and risk management operating model with appropriate metrics, issue reporting, and feedback collection. Consider the results of reporting and proportionate improvements in a periodic exercise.

To ensure continuity of AI governance and risk management, FIs can consider embedding mechanisms for the ongoing evolution and improvement of the relevant operating model. To do so, they can establish suitable reporting on the operating effectiveness of their AI governance and risk management, which can inform periodic practices for its improvement.

Holistic measurement of AI governance and risk management is important because it allows FIs to understand, end-to-end, how well their operating model is functioning in promoting the FI's principles and managing AI risks across the business; where deficiencies or opportunities are identified, FIs can revise their operating model or overall approach to AI governance and risk management accordingly. This is closely aligned with the recommendations of the standard ISO-IEC 42001:2023, which emphasises ongoing improvement as a key aspect of AI management.

Each FI, based on its context, the extent of its overall exposure to AI risk, and its appetite for risk, will determine the appropriate nature and frequency of such measurement. FIs may track the effectiveness of their AI governance and risk management operating model in several ways, including but not limited to the following:

- Tracking and reporting on metrics related to end user complaints.
- Collecting information on trends in user feedback or AI outcomes more broadly.
- Detecting and documenting issues, inefficiencies, or gaps related to governance.
- Collecting feedback from stakeholders, especially from working-level employees monitoring or using AI.

Assessment of the fitness and functioning of the operating model is particularly important as AI technology continues to evolve. Horizon scanning, assessment of peers, and an ongoing outside-in approach to identifying and assessing AI risks and regulatory developments will support FIs in keeping their AI risk management operating models relevant in the long term, especially as AI continues to evolve and as it becomes implicated in an ever-greater share of he enterprise's functions. FIs may also wish to periodically review their governance practices against evolving industry norms and tools to improve the overall effectiveness of their AI governance and risk management.

2. AI Risk Management

2.1 Ensure Effective AI-Related Policies, Procedures, and Standards

FIs typically operationalise key elements of their governance through policies, procedures and standards (collectively referred to here as "governance documents") that define processes, set out rules, and offer guidelines. AI-related governance documents are the FI's authoritative position on the management of AI risk. These documents consistently define key AI-related concepts and typically document and assign responsibility for the enterprise's approach to AI governance and risk management. Doing so in a generally accessible body of documents helps to align principles and expectations among stakeholders. The governance described elsewhere in Handbook is typically operationalised by setting policies, procedures and standards.

Some FIs will create an "AI Policy" that governs most AI-specific risk management activities, whereas others will embed policies and procedures related to AI in other documents, like their MRM Policy if they have one. Each FI will choose, based on its organisational context, how best to create, update, or leverage relevant local- or group-level governance documents related to AI to support effective AI governance and risk management. FIs that are branches or subsidiaries of parent entities in other jurisdictions may choose to draw on the AI risk management frameworks of those parent entities.

Consideration 2

Ensure that governance documents define key AI-related concepts, processes, and responsibilities, and that they remain up-to-date and effective in supporting all aspects of the FI's approach to AI governance and risk management.

Practice 1: Ensure robust conceptual foundations for AI governance and risk management by establishing AI principles, defining key AI-related concepts, establishing frameworks for effective AI identification, and continuously improving these foundations over time as necessary.

Approach:

- Define the FI's AI principles.
- Set a clear definition of AI for consistent use within the FI, as well as relevant governance-related concepts.
- Establish an effective AI identification framework with clear rules, controls, responsibilities, and documentation to operationalise the FI's definition of AI.
- Ensure that AI principles, definitions, and identification frameworks are periodically reviewed and updated.

The first step in AI governance and risk management is the establishment of a conceptual foundation. Setting out the FI's principles for AI use, defining AI, and setting up a framework for effective AI identification is an essential enabler for all subsequent activities.

AI principles set an overall direction that each use case can be judged against for the purposes of review and for making deployment decisions. FIs may choose to adopt guiding principles governing AI use – with Fairness, Ethics, Accountability, and Transparency (drawn from the 14 FEAT principles issued by MAS in 2018) serving as a common industry standard and point of reference. In addition to the FEAT Principles, FIs can consider further principles corresponding to their corporate values (e.g. sustainability), globally relevant standards like the OECD AI Principles, or other sets of AI-related principles that are relevant in jurisdictions where the FI operates. FIs can consult the documents in Appendix C, many of which contain examples of AI principles.

Defining open-ended, less-prescriptive principles ensures that they will remain relevant even as AI technology continues to evolve.

AI governance and risk management can only be consistent and effective when an FI has a clear, enterprise-wide position on what technologies meet the definition of AI. There are three general considerations for FIs to address in the choice of an AI definition:

1. Whether the definition of AI is effective in identifying technologies or use cases that pose AI-specific risks, including emerging deployment patterns such as embedded AI.
2. Whether the definition is clear and useful to both technical and non-technical Business Users within the FI. An effective AI definition enables working-level staff to consistently interpret and apply the definition when assessing whether specific technologies or use cases fall within the scope of AI.
3. Whether the definition is aligned with the definition used in this Handbook, with those adopted by relevant regulators, and generally accepted definitions across the broader ecosystem.

Other AI-related concepts that the FI uses in governance may also benefit from clear definitions. The definition of AI used in the proposed MAS Guidelines on Artificial Intelligence Risk Management, as well as of other key terms related to governance and risk management like "model", "system", and "use case", is provided in Subsection 1.1. Other relevant terms that and FI could consider defining include "Traditional AI", "Gen AI", and "Agentic AI" or "AI agent".

Clear and actionable definitions of AI are essential for FIs to accurately identify AI in real-world applications and apply appropriate governance measures. To enhance the practical application of these definitions, FIs can include helpful guidance for employees, such as a list of examples of what is and is not AI, and a set of frequently asked questions.

However, definitions can be overlooked if not effectively operationalised – especially where AI is embedded in other applications and is hard to detect. Alternatively, governance personnel in different business functions may interpret key terms differently, leading to inconsistent oversight, or FIs may encounter definitional edge cases like new AI technologies which may be challenging to categorise. This presents several risks, including both false positives (use cases that do not contain AI subject to AI governance and risk management, misallocating resources away from real AI risks) and false negatives (use cases whose AI-specific risks are not managed, also known as "shadow AI").

The effective governance and risk management of AI involves a consistent and methodical implementation of a uniform AI definition across the enterprise. To do so, FIs can establish an AI identification framework. This typically involves:

- Formalising clear and consistent enterprise-wide standards for applying AI-related definitions within governance documents.
- Establishing controls to ensure consistent and robust AI identification and which limit the extent of shadow AI. They can include:
    - Establishing clear roles and responsibilities for AI identification, supported by training, incentives, and consequence management. This can help ensure that identification processes are consistently and correctly applied by their accountable owners. Currently, the most common approach to AI identification is self-declaration of AI use by application owners.
    - Formalising AI identification steps throughout the FI's SDLC, such as in governance checklists at each lifecycle stage, to facilitate self-declaration of AI use by Builders or Business Users.

    - Incorporating AI identification into existing risk and control reviews, including those applied to key risk processes. This approach leverages existing control mechanisms to enhance coverage and strengthen oversight.
    - Including AI identification in the onboarding process and periodic application reviews for third party AI products and services and the reviews of release notes from third party providers. This is discussed in detail in Subsection 2.3.
- Defining clear responsibilities for AI identification throughout the AI lifecycle, as well as accountability and response for incorrect identification where appropriate.
    - A key responsibility is the designation of a final authority or forum for applying the FI's AI definition and adjudicating differences between employees or business units, particularly in relation to new or emerging technologies. It can be particularly effective to assign this role to the forum or committee responsible for creating the AI definition.
- Ensuring that effective documentation is captured on the process of identifying AI.

FIs can also ensure that processes are in place to periodically review their AI definitions and revise them to be more effective or to account for new developments in the industry. Consulting key stakeholders – such as impacted business functions, internal users of AI, third party AI providers, and external partners or regulators – can be an important input in this process.

Practice 2: Ensure that all aspects of AI governance and risk management are effectively institutionalised throughout the FI's governance documents, and that a process is in place to periodically review and reassess them.

Approach:

- Assess the FI's end-to-end set of enterprise governance documents to identify gaps related to AI governance and risk management, and address these gaps through changes to existing documents or the creation of new ones where necessary.
- Ensure that newly created or revised governance documents are clearly communicated and effectively shared with relevant employees.
- Periodically re-assess governance documents for new gaps against evolving AI technologies and their associated risks.

Governance documents will typically define AI governance and risk management roles and responsibilities, rules, processes, and controls. They may also document key templates, metrics, and standards related to managing AI risks across the AI lifecycle. FIs can consult relevant industry frameworks, including this Handbook and relevant rules or regulations, to identify AI governance and risk management requirements, and based on these requirements can conduct an assessment of their existing governance documents to ensure that they are fit for purpose. Where gaps exist, FIs can augment those documents or create new documents, depending on which approach is most suitable to their organisation. FIs can also modify existing rules or processes to address new risks introduced by AI – such as by modifying conduct policies to address the risk of the inappropriate use of AI systems.

FIs can leverage their existing governance review processes to perform this gap assessment and remediation exercise. These processes can be used to establish or update governance documents through appropriate consultations, drawing on input from AI governance and risk management professionals as well as experts from impacted functions, and can ensure that changes are effectively communicated across the organisation.

Given the rapid evolution of AI technologies and the potential emergence of new associated risks, FIs may find it beneficial to monitor changes in AI technologies, regulations, and risks that may implicate their governance documents, and to periodically revise those documents where necessary.

AI governance and risk management through metrics and/or qualitative indicators. In the absence of suitable forums, FIs may also create communications channels as they deem fit to internally discuss AI-related matters, including governance. FIs can periodically assess the fitness of their relevant governance processes, forums, assets, and tools as AI technologies evolve.

Illustration 2.1.1: Income Insurance: Integrating AI Governance with Risk Policy

Prior to 2023, FEAT (Fairness, Ethics, Accountability and Transparency) assessments were already being conducted on AI models in line with MAS guidelines. However, this was done independently of enterprise Model Risk Management.

To better integrate AI Governance with corporate policy, an initiative was undertaken in 2023 to update Model Risk policy, making AI Governance part of the enterprise Model Risk Management framework. The Risk Management Function (referred to below as Risk) would evaluate all models and use cases to first establish their materiality. An updated scoring system was also developed for materiality assessments, which enabled Risk to effectively assess the materiality of AI models and use cases.

After materiality was determined, Risk would subsequently trigger FEAT assessments if the model or use case was deemed to be AIDA (Artificial Intelligence and Data Analytics). An enterprise definition of AI was established in order to guide Risk in identifying AIDA. This definition would also come under routine review to keep it relevant to the rapidly changing landscape of AI. As a result of this review process, the definition was further updated in late 2023 to incorporate Gen AI.

Once triggered by Risk, FEAT assessments would be conducted with model and use case owners by the AI Governance team using a risk-based approach, with a standard FEAT checklist template being used for all models/use cases. For high-materiality models and use cases, an enhanced checklist would also be used. All information captured by assessments would subsequently be stored in a centralised model register.

Apart from this integration, FEAT Principle guidelines and assessment templates were formally made part of corporate policy documentation. In addition to this, an annual AI Governance training was made mandatory for all AIDA model owners and developers. This was coupled with an attestation that FEAT Principles were understood and would be adhered to, to ensure awareness and that the right culture around AI was nurtured.

As we look to the future and onboard advancements in Gen AI, this operationalised AI Governance framework forms a solid foundation to build from.

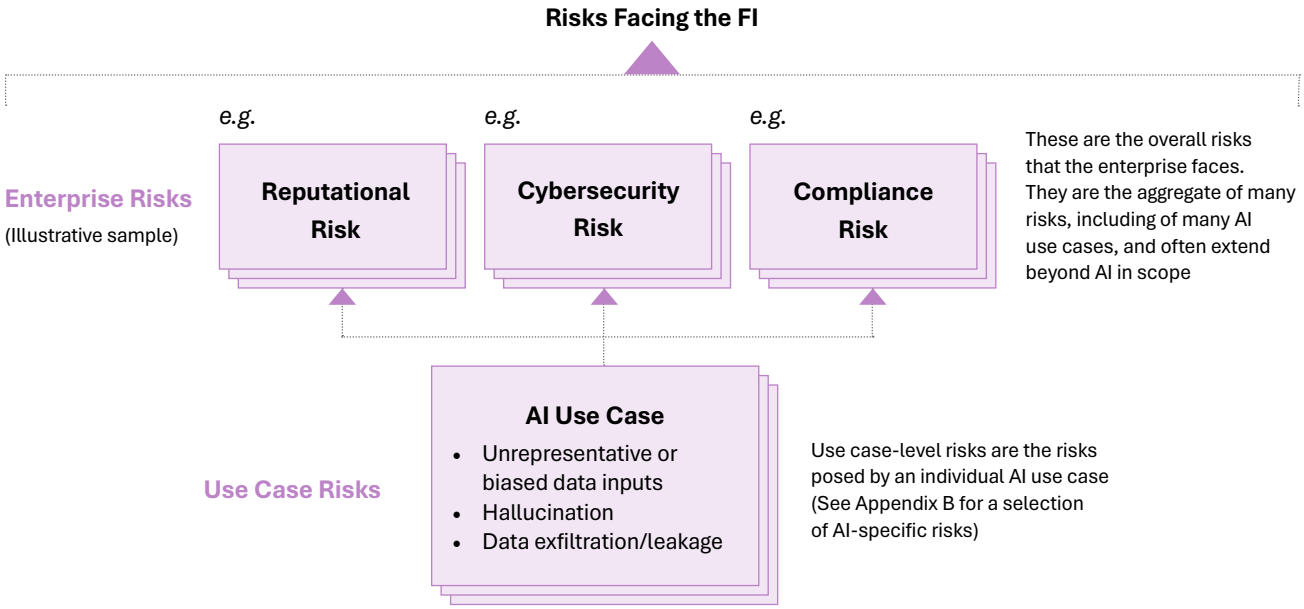
2.2 Enhance Organisation-Level Risk Management

All FIs have existing risk management practices in place to address various risks – including cybersecurity, technology, and reputational risks. This is sometimes referred to as "enterprise risk management". When adopting AI in the organisation, new or enhanced risks – referred to here as AI-specific risks – can be introduced.

Some AI-specific risks are organisation-level and can impact the FI as a whole. Other AI-specific risks can be use case-specific, including risks such as data bias, data leaks, or hallucinated outputs. This is illustrated in Figure 2.2.1 below.

Both types of risk are important for effective AI governance and risk management. This Subsection discusses how FIs can update their risk management to include the impact of AI on organisation-level risk; Subsection 2.4 discusses how FIs can update their risk management to address individual use case risks.

Figure 2.2.1: Organisation-Level and Use Case-Specific AI Risks



This Subsection is structured around the typical steps in risk management, a stylised view of which is illustrated Figure 2.2.2. FIs already have robust risk management processes and well-defined risk appetites in place and, where not otherwise specified, those practices will not change when managing AI-specific risks.

Figure 2.2.2: Illustration of the Steps in Risk Management

FIs will each categorise their risks in a way that is appropriate to their business needs and reflective of their compliance requirements.

An example of a use case-specific risk is the possibility that a Gen AI system will unexpectedly produce outputs that breach regulatory or organisational standards.

Such an incident could pose broader risks to the enterprise. Depending on the type of standard that is breaches, such an incident in a public-facing use case can reduce public trust in the FI, harming its reputation. Outputs that are serious breaches – such as recommending that risk-averse seniors invest in cryptocurrencies – could risk regulatory or legal consequences. If the FI's operations depend on accurate outputs from the AI use case, breaches in performance standards could also cause broader, enterprise-wide operational risks.

Consideration 3

Enhance the organisational risk framework and risk appetite to include enterprise risks, strategies, and key risk indicators (KRIs) that track, monitor, and mitigate AI-specific risks.

Practice 1: Identify the new or enhanced risks of AI that are relevant to the enterprise and ensure that the enterprise risk taxonomy effectively captures them.

Approach:

- Identify AI-specific risks that are relevant to the FI and its foreseeable use of AI.
- Incorporate AI-specific risks in the enterprise's risk management framework in a manner that is suitable to the FI, such as by integrating AI-specific considerations in existing risks or defining new AI-specific risks.
- Assess and manage AI-specific enterprise risks based on their likelihood and materiality.
- Develop a portfolio-level view of AI-specific enterprise risks and monitoring metrics for senior leadership decision-making.
- Periodically review the FI's identified AI-specific enterprise risks and address new risks as they arise.

Beyond existing technology or security risks, enterprise AI-specific risks could impact reputational risk, legal risk (such as around intellectual property), and regulatory risk. Each of these may require additional considerations to be adequately managed. The list of the AI-specific risks identified by the MindForge consortium is included in Appendix B.

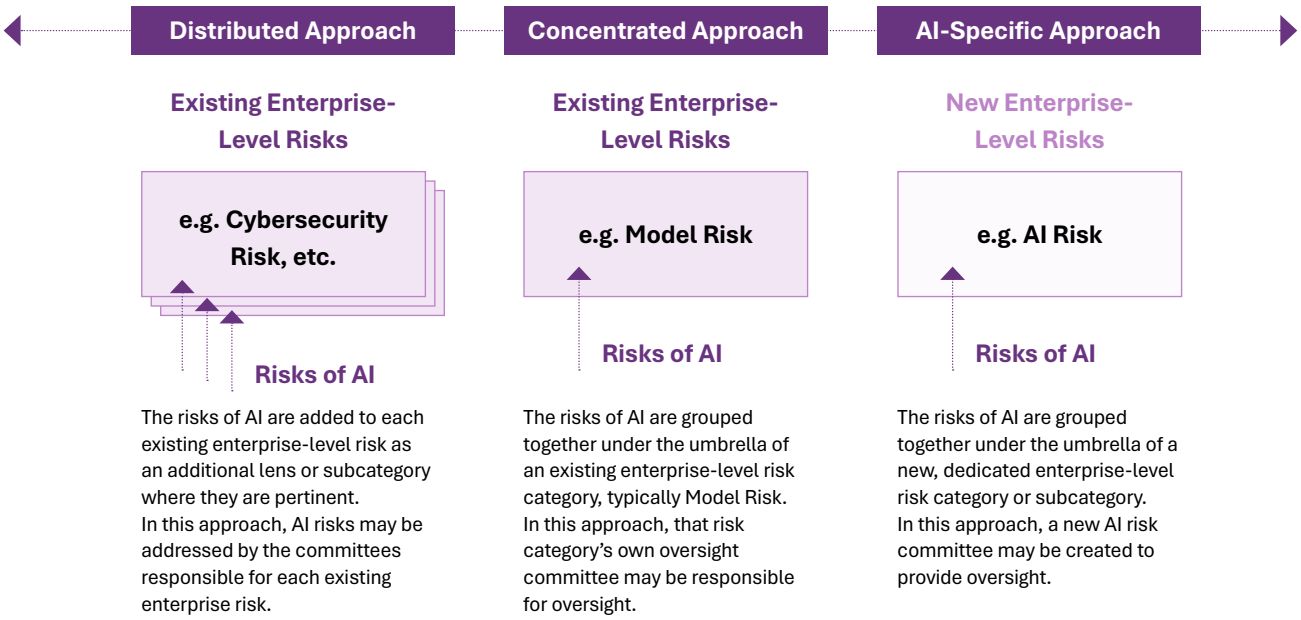
Existing risk management approaches in the FI, especially those around model and technology risk, continue to be relevant to managing AI risk.[10] As such, FIs can start the process of managing AI-specific risk by assessing their existing risk frameworks and identifying how best to adapt them.

FIs may choose to retain their existing categories of risk and to update them with the impacts of AI; this is an example of a distributed approach to AI-specific risk management. Others could expand one existing category of enterprise risk to include AI-specific risks; this is an example of a concentrated approach. FIs that select the concentrated approach typically tend to choose, but are by no means limited to, "Model" or "Technology" risk. Finally, FIs may create a new category of risk dedicated to AI-specific risks; this is an example of an AI-specific approach.

[10] For a non-exhaustive sample of those existing risk management practices, reference Appendix C.

These approaches are illustrated in Figure 2.2.3. These examples are not mutually exclusive or exhaustive; FIs may also employ a combination of approaches, and may adapt their approach over time.

Figure 2.2.3: Approaches to Taxonomising Organisation-Level AI Risk



FIs may also consider the availability of the relevant AI expertise on their existing risk governance bodies; this may influence the design of their AI risk management approach and their choice to centralise or decentralise. Each approach has its own characteristics, and the FI's risk management committee structure and process for managing AI-specific risk will differ depending on the chosen approach. It is important that FIs carefully consider, and clearly assign, accountability for enterprise risk management in their chosen approach. In each of these approaches it is important for FIs to ensure that risk management is coordinated and consistent across the enterprise, ensuring that there are no gaps in coverage between risk types or parts of the FI's organisation. FIs can consider involving internal experts and stakeholders as part of this process.

Enterprise functions that may be relevant stakeholders in defining AI-specific risks include:

- Data Management/Governance
- Security
- Privacy
- Risk & Compliance
- Legal
- Operations
- Technology & IT
- Talent & HR
- Core Business Functions

FIs will each identify the right functions to consult. This decision can include considering those who develop AI, use AI, are exposed to AI risk, or are involved in the AI risk management process.

FIs may also choose to update their enterprise risk appetite to account for the new considerations introduced by AI adoption. They will already have practices in place for assessing risk appetite, such as those laid out in common industry frameworks like those by the Committee of Sponsoring Organizations of the Treadway Commission (COSO).[1]

It may be useful for FIs to periodically consolidate an enterprise or portfolio view of the aggregated risks of their AI use for the Board or Senior Managers. This is aligned with industry norms such as the COSO Enterprise Risk Management Framework.[1]

Given AI's rapid evolution, FIs could consider periodically performing horizon scanning for new or enhanced AI-specific risks and regulatory developments on AI governance and risk management. This supplements periodic reviews of the FI's enterprise risk management to assess its effectiveness, including by reviewing the effectiveness of their AI risk management approach.

Practice 2: Assess existing enterprise risk controls for their fitness in addressing AI-specific enterprise risks, and uplift those controls where gaps exist.

Approach:

- Assess existing organisation-level controls for their suitability in managing the enterprise's AI-specific risks.
- Uplift or supplement existing controls as needed to address enterprise AI-specific risks.
- Review the adequacy of controls on an ongoing basis.

FIs can manage enterprise AI risks by assessing the adequacy of their existing controls as part of the process of uplifting their overall risk management. Existing controls such as access management, data protection, and incident management, response, and recovery continue to apply to AI use cases with some AI-specific enhancement. New controls may need to be introduced where these existing controls are not sufficient. In all cases, it is important to ensure that these controls are well-suited to the FI's AI principles. Several examples of AI-specific organisation-level risk controls are defined in the Veritas Initiative.[11]

As FIs progress in their AI journey, they benefit from continuing to re-evaluate and, if needed, adapt existing controls to capture the evolving risks that AI poses.

[11] Read the publications of the Veritas Initiative at https://www.mas.gov.sg/schemes-and-initiatives/veritas

Practice 3: Ensure that key risk indicators (KRIs) are in place to measure AI-specific risks and that relevant incidents, issues, or risk events are appropriately tracked and managed.

Approach:

- Ensure that appropriate KRIs related to the FI's AI-specific risks are in place and assigned to appropriate risk owners, leveraging or supplementing existing enterprise KRIs as needed.
- Perform organisation- or portfolio-level tracking of AI-specific risk.
- Track incidents, issues, and/or risk events related to AI-specific risk, considering their severity when doing so.
- Ensure that practices for responding to incidents, issues, risk events, and/or trends in associated risks are suitable for addressing AI-specific risks.

Each FI will select appropriate KRIs for its own AI-specific risks, whether these leverage existing KRIs or are novel to AI. FIs can ensure that KRI ownership is well-defined and that, in all cases, owners with appropriate AI competencies are designated. A non-exhaustive list of sample cross-use case KRIs is included below for reference.

- Accountability & Governance
  - Proportion of use cases not in the AI inventory, indicating potential gaps in risk governance (see Subsection 2.5).
  - Number of AI use cases in production without approval/ethics review or number of use cases that have breached conditional approval deadlines, indicating a failure to apply AI governance and risk management standards.
- Transparency & Explainability
  - Number of "black box" AI systems, which are those whose outputs cannot be explained or justified to a meaningful extent. This can potentially be indicative of an inability to deliver on the FI's commitments to transparency.
- Legal & Regulatory
  - Number of process exceptions, indicating a potential pattern of breaches of organisational governance standards.
  - Number of third-party copyright infringement claims against the FI related to AI outputs used externally.
  - Number of customer-facing use cases found to contravene legal or regulatory expectations, such as requirements to treat customers fairly.
  - Number of AI use cases found to engage in inappropriate market conduct or to breach financial services regulations.
- Monitoring & Stability
  - Anomalies detected, error rates, performance degradation metrics, or model drift metrics, indicating potential failures of use case-level controls or the deterioration of common models applied across use cases.
  - Overall data quality, indicating potential problems in the FI's data ingestion and handling practices or implications for AI performance.
- Risk Exposure
  - Aggregate number of use case KPI breaches. This can indicate a range of risk events.
  - Aggregate financial exposure of AI use cases.
  - Number of risk events, faults, or incidents. These can indicate a range of potential risks depending on their nature.
  - Qualitative user feedback on AI outputs.

The below criteria are an illustrative example of tracking AI-specific incidents by severity. For the purposes of this example, an AI incident is treated as a case where AI-related controls fail to prevent a risk's occurrence, resulting in an impact.

Severe Incident: Using AI, or recommendations generated by AI, to make financing-related decisions related to end customers that internal analysis indicates was based on racial or gender bias, or which results in legal or regulatory action on the basis of racial or gender bias.

Moderate Incident: Using AI, or recommendations generated by AI, without completing effective human review, in high-risk or medium-risk use cases where human-in-the-loop review was identified as a required control to mitigate risks related to Fairness & Bias.

Minor Incident: Using AI to generate customer marketing communications, where an AI-identified list of customers for outreach contains existed customers who no longer consent to receiving communications.

Tracking the number of AI incidents or risk events is an important element of managing AI-specific risks. FIs already maintain their own definitions of incidents or risk events and have frameworks for managing them in line with industry practices and compliance requirements in jurisdictions where they operate. FIs will each determine how best to manage AI incidents, risk events, and other failures in that context. This may include AI-specific management practices, such as longer incident management timelines.

Each FI may choose to track AI incidents separately or as a subset of other types of incidents. In all cases, it is important to track issue ageing – given the longer resolution timelines for AI incidents – to ensure that they are managed in a reasonable amount of time despite their complexity and that downstream users are kept appraised of potential impacts.

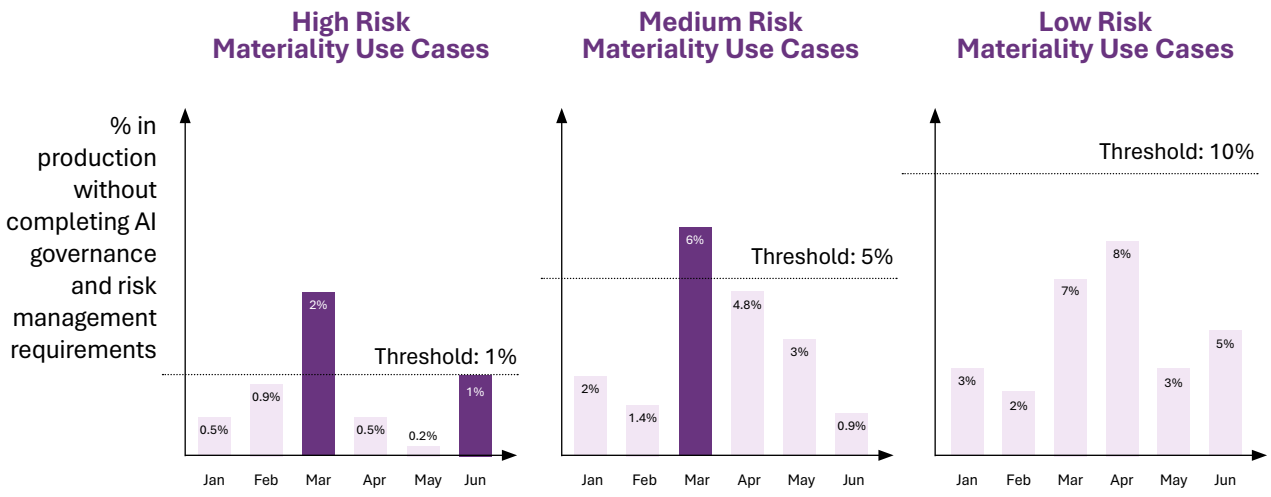
Practice 4: Ensure that effective monitoring is in place to identify AI-specific risk events or breaches of KRI thresholds to a degree proportionate to the FI's risk appetite.

Approach:

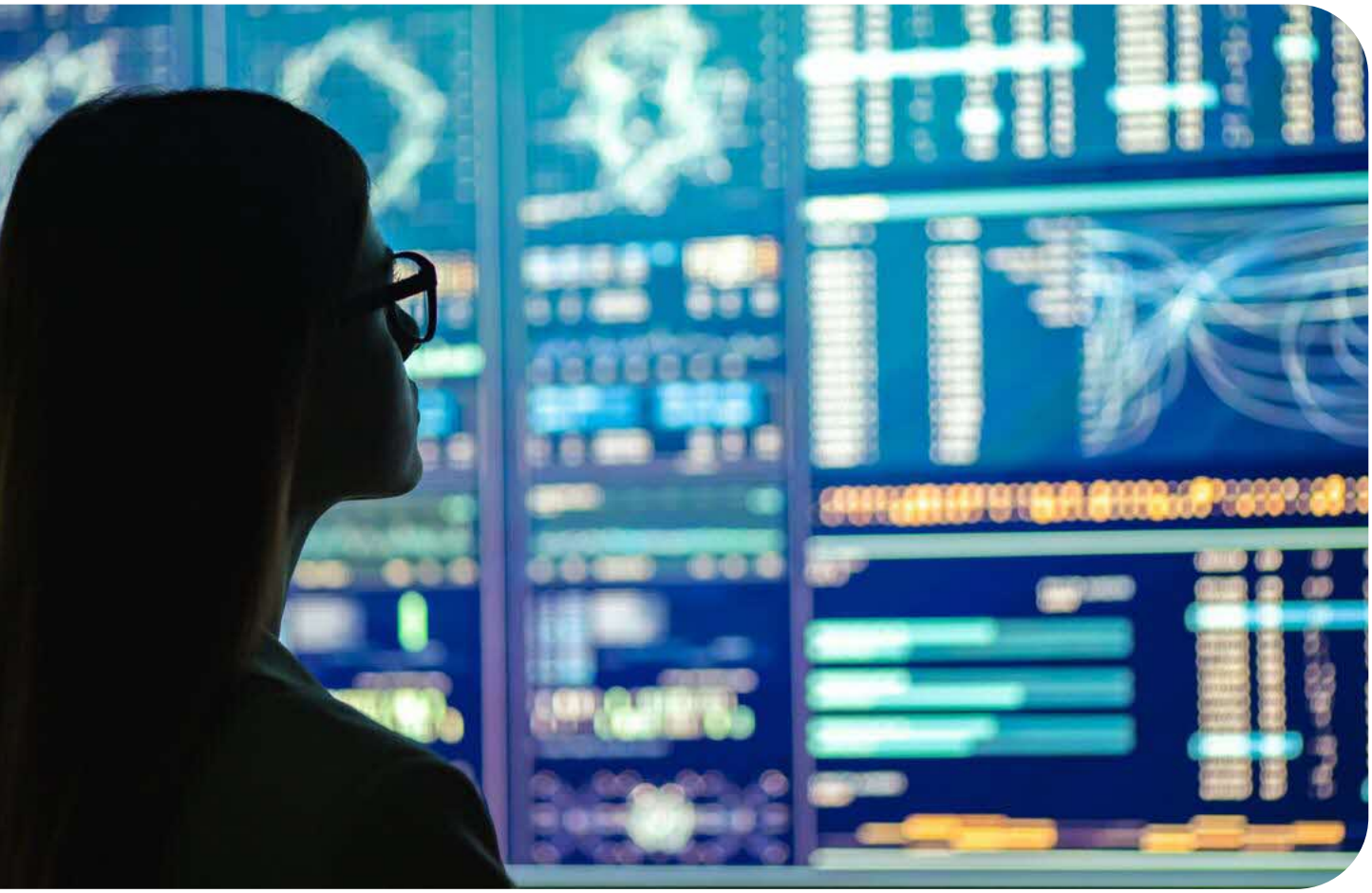
- Uplift existing practices for tracking organisation-level KRIs to ensure that AI-specific KRIs and risks are effectively monitored against appropriate thresholds.
- Ensure that practices are in place to respond to AI-specific risk events or KRI threshold breaches, updating those practices to be suitable for AI-specific risks as needed.

The scope, frequency, and intensity of AI-specific KRI monitoring and reporting are most effective when they are in line with the FI's existing risk management practices and risk appetite. Figure 2.2.4 illustrates the example of tracking an AI governance and risk management process exception, such as the proportion of AI use cases in production without completing the full set of AI governance and risk management requirements.

Figure 2.2.4: Illustration of Organisation-Level AI Process Exception Tracking for Risk Management



FIs can consider reviewing their existing risk response and escalation procedures to include AI-specific risks, such as reporting AI-related incidents in relevant forums.



2.3 Uplift Practices for Managing Third Party AI Risks

FIs are increasingly relying on AI models and systems provided by third parties, especially when adopting Gen AI and Agentic AI. Sometimes, those models or systems are accessed as SaaS, or are open source. The industry's long-term shift towards prebuilt AI reflects the increasing complexity and competency of general-purpose AI models.

The use of AI products or services from third-party vendors, service providers, and contractors (henceforth, "third parties") may introduce new AI-specific risks, especially as FIs shift towards using AI as SaaS. These risks generally relate to 1) the information asymmetries between FIs and third parties and 2) the division of responsibilities and consequences between FIs and third parties.

FIs are typically considered to be accountable for the behaviour of AI that they use irrespective of whether it was built or procured. FIs generally continue to manage the risks of their third-party AI products or services using the same AI use case lifecycle that they apply to all other use cases, including those developed in-house. Following the same steps, checks, and standards, with additional considerations added where necessary to address risks that are specific to the use of third-party AI products or services, can help FIs hold third party AI products and services to the same standard as those developed internally.

Potential AI-specific risks of third-party products and services are similar to those that FIs face when procuring any software: regulatory non-compliance, failure to perform as intended or manage risk appropriately, or misalignment with the FI's mission or principles. AI enhances some of these risks because:

- Some AI products or services may not be fully transparent to the FI, and as such, it can be more difficult to comprehensively assess them for risks. Ways in which third party AI products or services may not be transparent to FIs that procure them could include:
    - Not having access to the underlying AI model(s).
    - Not having access to or information on the data used to train the model(s).
    - Not having information on how the FI's data is used by the third party, especially after that data is used in model training.
- AI products or services often evolve after deployment.
- AI-specific risks can be challenging to comprehensively assess in a procurement or onboarding exercise because of the required AI-specific technical skills, which may not always be present in relevant teams.
- Information on tests and evaluations conducted on an AI product or service may be missing or incomplete.
- Some FIs may be over-reliant on key third party AI providers, especially for AI products and services where the market is highly concentrated. This reliance can be direct – where an FI is itself highly dependent on a single AI provider, product, or service – or indirect – where an FI uses a range of third-party products or services, but those services have upstream dependencies on a single AI provider, product, or service. This could include Gen AI third party products or services that use a common foundation model.

Existing standards and processes for managing third party risks continue to play a crucial role in managing the risks of AI models or systems from third parties. Existing third-party risk management practices designed for traditional, static software may be insufficient on their own for managing AI-specific risks, however, because they lack requirements for AI-specific disclosure or testing, contracting practices designed to manage AI risks, or procedures to involve relevant AI experts.

A range of terminologies are often used to refer to the arrangements under which FIs access third-party AI products or services. For the purposes of this Handbook, the following deployment patterns are considered typical for third party AI:

- **Onboarding with customisation.** FIs procure a third-party AI product or service and modify it before use, such as when they procure a model and fine-tune it.
- **Onboarding without customisation.** FIs use a third-party AI product or service, either on premises or as SaaS, without making changes to it.
- **Embedded AI.** FIs use a third-party product or service which contains AI components but is not principally an AI product or service.
- **Connected service.** FIs use a third-party product or service which does not contain AI, but where AI is involved in its value chain, such as where the FI is provided with AI-generated content by an outside partner.
- **Non-connected service.** FIs use a third-party product or service whose provider uses AI elsewhere in their organisation, indirectly exposing the FI to AI risk. The materiality of that risk is typically low.

Third parties implicated in AI governance and risk management are not limited to those that provide AI that is integrated into the FI's ecosystem and includes AI used by providers of outsourced services, such as contractors or consultants. Such services can also pose AI risks to the FI. In addition to existing outsourcing guidelines, FIs can assess and address AI-specific risks as part of the procurement and post-procurement lifecycles of outsourced service providers as well.

Some non-connected services may expose FIs to very little AI-specific risk; impacted risk types in these cases are typically reputational or operational in nature. FIs may find that these can be managed using existing third party/outsourcing risk management practices.

For a relevant guideline on managing outsourced service risks in the banking sector, see the ABS Guidelines on Control Objectives and Procedures for Outsourced Service Providers (OSPAR).

While the terms "open" and "closed" source are well-established in the industry when referring to traditional software, they are sometimes used inconsistently when describing AI (especially when discussing LLMs). For the purposes of this Handbook, the term "open source" can refer to three distinct approaches for distributing AI models.

Table 2.3.1: Open-Source Deployment Types for AI Models

Deployment Type		Description
Open Source	Fully Open Source	According to the Open Source Initiative (OSI), "fully" open-source AI is: "...Made available under terms and in a way that grant the freedoms to: • "Use the system [or model] for any purpose and without having to ask for permission. • "Study how the system [or model] works and inspect its components. • "Modify the system [or model] for any purpose, including to change its output. • "Share the system [or model] for others to use with or without modifications, for any purpose."[21] Most notably, this includes detailed information on the model's training data. While training data need not be shared outright, per the OSI, information on the training data for a fully open-source model is such that "a skilled person can build a substantially equivalent system". [21]
	Open Weights	The model developer or provider shares the model's parameters, usually by sharing the model file directly. They may also share other information, like architecture, code, and training data/ methodology. The model can usually be run locally and accessed, modified, re-trained, and deployed without prior permission. Open weights distributions differ from open-source distributions because they generally do not share complete information on the underlying training and data, and as such do not meet the OSI standard that "a skilled person can build a substantially equivalent system".
	Restricted Use	The model is shared according to the open weights approach above. However, unlike in open weights distributions, restricted use distributions may include further legal or contractual limitations – such as licensing terms restricting the number of users, restricting the model's use for commercial purposes, restricting certain types of use on ethical grounds, requiring attribution, or requiring that derivative software/models also be made available under specific license terms.
Closed Source	Restricted Weights	The model developer or provider does not directly share the model's parameters, architecture, or training data/methodology. They may provide disclosures describing this information in varying degrees of detail. Usually, these models are accessed as SaaS and cannot be run locally.

Consideration 4

Uplift existing procurement and third-party risk management activities to address AI-specific risks, including disclosure templates, vendor assessment and procurement practices, change detection and notification, contracting practices, and ensure that teams have access to relevant expertise in AI.

Practice 1: Define, based on relevant AI-specific risks, a proportionate level of disclosure to seek from third party providers of AI products and services, and a process for assessing disclosures.

Approach:

- Define organisational standards for the expected disclosures that can be sought from third parties that provide AI products or services, and for the types of usage that are suitable depending on the level of disclosure. This can include the designation of an expected template (see Appendix E).
- Ensure that a process is in place for making informed decisions about the adequacy of third-party disclosures in light of mitigating factors, such as indemnification, attestation, or testing results.

A common industry practice is for companies using AI products or services from a third party, irrespective of its deployment pattern, to request the disclosure of key information to enable risk management and proportionable governance. Some institutions, technology providers, and policymaking bodies use standard disclosure templates – "AI Cards" – to do so. FIs differ in the prescriptiveness of their disclosure templates, with some preferring a more open-ended information collection approach and others using fixed lists of questions. Whether using a prescriptive template or a more flexible, thematic approach, it is important for FIs to be consistent in the types and levels of information that they seek.

Disclosures are referred to here as "AI Cards" but may also be known as "model cards", in the case of disclosures related to individual models, "system cards", in the case of disclosures related to whole systems, or "nutritional labels". There is a wide variety of terminology and content for disclosures across the industry.

The ISO/IEC 42001:2023 standard suggests appropriate disclosure elements for third parties to provide, including a general description, instructions for its use, technical assumptions and limitations, and monitoring functionality, as well as some description of the model or system's development and data (ISO/IEC 42001:2023 B.6.2.7). It further suggests that third parties should provide users, such as FIs procuring AI in an "embedded" or "connected service" deployment pattern, with relevant information for using and governing the system (ISO/IEC 42001:2023 B.8.2).[15]

There are a range of other templates for AI Cards or related disclosures[12] developed by policymakers and researchers, although none of these have yet emerged as an industry standard. FIs can consider which templates are appropriate for their needs when managing risk, considering factors like regulatory requirements in jurisdictions where they operate. They can pay particular attention to the MindForge AI Card template, included in Appendix E.

FIs can consider establishing a policy or procedure that clearly defines the information that they request from third parties. This can include a procedure for requesting standard disclosure information and ensuring that a qualified and capable party within the FI reviews that documentation for whether it matches the FI's expectations. FIs may vary the depth and extent of information required based on the risk materiality of the expected use case. FIs can also vary the deployment patterns under which they request information.

[12] See, for an influential sample of AI model and system disclosure templates, Mitchell et al. (2019)[17], Golpayegani et al. (2024)[10], Arnold et al. (2019)[3], and OECD (2022)[19]. For a sample of dataset disclosure templates, see Gebru et al. (2021)[9], Hutchinson et al. (2021)[12], and Holland et al. (2018)[11]

It is common in the industry to request a similar standard of disclosure for onboarded, embedded, or connected AI products or services, but not to request this full range of information for non-connected AI services given their limited impact on the FI.

Each FI can determine the circumstances under which it would request AI-specific disclosures in its context.

Approaches for managing AI risk on the basis of disclosure are generally less resource-intensive and include assessing dataset disclosures to check for IP/privacy risks, reviewing the architecture of the model or system to identify risks associated with their design, or considering whether the AI product or service's performance on evaluations performed by the vendor meets the FI's thresholds for risk.

FIs may not always receive all of the information that they request third parties to disclose; disclosures related to training datasets, in particular, can be commercially sensitive for third parties, and requests for information on these datasets are often declined. FIs need not automatically reject incomplete disclosures; in these cases, it is important instead that FIs have a process in place for making informed decisions based on the types of disclosure that are missing and the risks of the use case. FIs can consider several mitigating factors in cases where disclosures are incomplete:

- Indemnification from related risks. For example, a third party that does not share information related to their AI model's training data may provide contractual indemnification against Intellectual Property (IP) violation. Doing so may incentivise third parties to manage risk and may entitle FIs to financial compensation for the purposes of risk remediation; overall, however, indemnities serve a narrow purpose and make a limited contribution to pre-emptive risk mitigation. FIs can note that some of AI's potential harms to their customers, their operations, and their reputations may go beyond what monetary compensation can repair.
- Credible external attestation. For example, a third party that does not wish to share information on some of its system-level guardrails with the FI could engage a trusted external body, such as an auditor, to evaluate and attest to the effectiveness of elements of their risk management, such as certifying compliance with relevant regulations or the implementation of a standard.
- Compensatory testing results. FIs can manage a range of AI risks by testing third party AI products and services on AI risk-related performance metrics as part of their procurement processes. This testing is "compensatory" in that it aims to make up for gaps in an FI's knowledge around an AI model or system by examining its performance and potential risks under various scenarios.

Whether or not they use a fixed AI Card template, FIs can vary the information that they request from third parties. For example, FIs can request that third parties provide them with additional evaluation results related to Fairness only if the intended use case has a potential risk of discrimination, or if the third party's data disclosure identified a risk that it was trained on personal data.

Requesting additional fields or a greater depth of information based on risk-related factors is an effective strategy to ensure that disclosure requirements are proportionate.

FIs can also be attentive to changing circumstances, such as new risks, new technologies, or lessons learned from the use of their disclosure templates. They can do so by ensuring that effective processes are in place to periodically review and revise their approaches to third party information disclosures related to AI. Designating a clear point of accountability for third party disclosure processes and templates can support them in doing so.

Practice 2: Ensure that processes and capabilities are in place for AI-specific risks to be evaluated at appropriate points in procurement, onboarding, and throughout the post-procurement lifecycle.

Approach:

- Address gaps in existing procurement and risk management practices by considering information disclosures, legal review, vendor assessment, compensatory testing, and relevant mitigating factors.
- Ensure that the alignment of third-party AI products or services with the FI's values and principles is included in risk-based procurement decisions.
- Continuously monitor and periodically re-assess AI products and services to ensure that their risks continue to be managed. Where AI products and services are used beyond their initially assessed use cases, consider repeating or supplementing the initial evaluation.

FIs already have procurement and third-party risk management practices in place, including cybersecurity assessments for third party providers, the definition of points of accountability, the identification of risks, and a process for reviewing, mitigating, and assuming risks. FIs can also continue to use their existing legal and cybersecurity reviews for the assessment of AI products and services. As such, before creating new AI-specific functions and processes, FIs can begin by considering how to address AI-specific gaps in their existing functions through targeted uplifts.

A key step in updating existing practices for the procurement of AI products and services is the identification of AI. FIs can consider requesting that vendors disclose whether a product includes, or is connected to, AI components. This is particularly important for "embedded" or "connected" AI products or services, where FIs may not have the ability to directly observe or detect AI components.

FIs can consider requesting AI-specific information disclosures (see Practice 1, above) as a starting point in risk management, as well as whether AI-specific legal and licensing considerations, discussed in more detail in Practice 4 below, have been appropriately addressed in procurement.

FIs can also consider the third party's expertise in the field and the effectiveness of their AI risk management and governance when assessing prospective providers in a procurement process. This supplements existing vendor assessments in areas like data handling, security, and regulatory compliance. This may include requesting and assessing documentation on the third party's processes for developing AI products or services and managing AI-specific risks, requesting that providers complete an AI governance and risk management questionnaire, requesting specific documentation on how the third party manages attributes like fairness and representativeness in data acquisition and management, or referencing relevant trusted accreditations that the third party may hold. The effectiveness of a third party's AI governance and risk management measures may need to be periodically reviewed.

The US-based Financial Services Information Sharing and Analysis Center (FS-ISAC) published an open Gen AI Vendor Evaluation & Qualitative Risk Assessment Tool in 2024. This resource provides a question list that serves as a practical starting point for FIs to leverage in updating their own procurement practices, recommending specific questions related to subjects like data protection, AI-specific security measures, and the disclosure of a provider's own upstream AI providers.

Testing third party AI products and services before procurement is an effective, but sometimes resource-intensive, risk management practice that can be used depending on the risk materiality of the intended use case. The depth, extent, and stringency of tests can also be increased to compensate proportionately based on use case risk materiality or when information disclosures are incomplete or unsatisfactory. Testing, in addition to declarations from the provider, can be used to assess whether an AI product or service conforms in its behaviours to the FI's values and principles. Tests related to fairness are particularly important to conduct; FIs can ensure that they are well-integrated into onboarding procedures by setting out clear procedures, proportionate to use case risk, for ensuring that testing for systematic disadvantage is conducted as part on onboarding.

A sample of AI risk-related performance metrics is provided in Appendix F. In addition to testing for quantitative metrics, FIs can consider other testing approaches like simulation testing and red teaming, where appropriate; these approaches are discussed further in Subsection 3.3. Testing of AI products or services for risks before or during onboarding does not substitute for the typical risk assessment and review processes described in Subsection 2.4.

The manner in which FIs conduct testing may differ slightly between deployment patterns; FIs will generally have sufficient access to perform testing in "onboarding"-type deployments where an AI product or service is deployed into the FI's technology ecosystem, but may have limited access to embedded AI and connected services. FIs generally do not have access to non-connected services to conduct testing. Where FIs have limited access for testing purposes, they can consider other options, such as requesting additional information on guardrails or requesting that a credible external attestation is provided for those AI products or services.

The decision of whether to procure a given AI product or service is ultimately a business decision informed by risk and by a judgement as to whether the product or service's characteristics and behaviours align with the FI's values and principles for AI use. AI products or services procured from third parties are acquired with one or several specific use cases in mind; FIs can vary the depth of their pre-procurement due diligence, and their tolerance for limited information disclosures, based on the risk materiality of that intended use case. For AI products or services with multiple potential use cases, FIs can also consider documenting limitations or restrictions on the range of acceptable use cases to those that were assessed as part of the procurement exercise. This may include, for example, limiting the use of a general-purpose AI language tool to personal productivity tasks, and restricting employees from using it for software development or integrating it into externally-facing business processes or applications. Depending on the expected risk materiality of the intended use case and the options available, FIs may choose to accept trade-offs related to cost and performance to select AI products or services that manage risk better.

After a third-party AI product or service is onboarded, FIs can continue to manage third party AI risks by ensuring that contingency planning is robust – covering major third-party risks like model changes or discontinued support – and conducting awareness efforts to ensure that internal stakeholders are aware of the risks of third-party AI products or services. For embedded AI and connected AI

services, FIs can consider requesting that third parties share monitoring information; where this is not possible, they can design effective solutions for monitoring the AI product or service's inputs or outputs once they reach the FI's ecosystem.

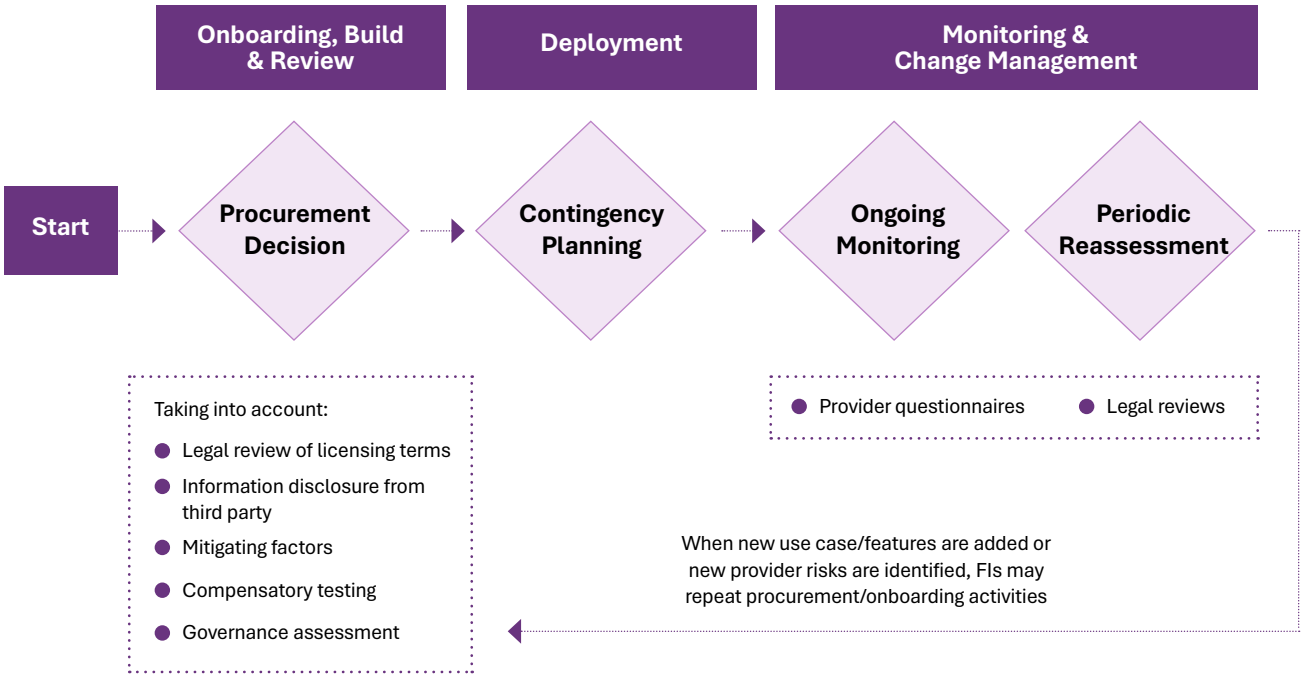
FIs that engage third parties early in an AI model or system's development process, such as when initiating a partnership to collaborate on a product, may not have a finished model or system to assess at the point of procurement. These types of partnerships are especially common with smaller third-party providers.

FIs that engage in these types of partnerships may need to accept additional flexibility in their procurement processes and decision points; this may, for example, require them to shift certain elements of information disclosure, assessment, and testing to a well-defined later stage of the lifecycle. Designing processes with appropriate degrees of flexibility, and with clear points of accountability for deviations, can allow them to take advantage of partnerships with such smaller providers.

Lifecycle monitoring is also an integral component of managing the risks of third-party AI products and services. FIs can continue to use, and can consider enhancing the stringency of, their approaches for monitoring the extent to which third party AI products or services behave as intended. Monitoring AI use cases after deployment is discussed in more detail in Subsection 3.5.

Finally, FIs can consider planning for the periodic re-assessment of third-party AI products or services at a frequency proportionate to risk – potentially leveraging existing AI-specific review processes discussed in Subsection 2.4. Additional review processes for third party AI products and services can include questionnaires on AI governance and risk management to AI providers and legal reviews of changed license terms.

Figure 2.3.1: Stylised View of Select Pre- and Post-Procurement AI Risk Mitigations



After procurement, FIs may sometimes change the use case for which an AI product or service is used, especially for general-purpose AI systems with a range of potential applications. It is important in these cases that FIs refer to the initial procurement process, legal reviews, and approvals to determine whether the new use case falls outside the scope of what was approved at that time. If so, FIs can manage the risk of this new use case by supplementing that initial procurement process – such as by requesting additional information or conducting additional tests – to seek procurement approval for a new, riskier use case in addition to existing IT change management practices.

FIs can ensure that accountabilities for third party risk management practices are clearly defined and that processes are in place for AI procurement practices to be periodically reviewed and revised, as needed.

Practice 3: Identify new or modified AI components or features in third party products and services already introduced into the FI's technology ecosystem.

Approach:

- Consider implementing legal assessments, periodic questionnaires, vendor re-assessments, or contractual provisions that can support the identification of new or modified AI components from third parties.

Changes to third party products or services initiated by those third parties can sometimes add AI components to non-AI products or services that the FI has procured or can meaningfully change AI products or services that the FI already has. This could introduce AI into the FI's ecosystem without appropriate control or governance of AI-specific risks. These AI deployments correspond to the "embedded" and "connected service" deployment patterns and represents a vector for shadow AI.

AI, by its nature, tends to change and evolve over time, and FIs should not expect to be notified of all changes. Some vendors conduct frequent retraining, which could make notifications of modifications impractical.

FIs can most practically manage their AI risks by seeking information only on major changes to an AI product or service's functionality. Determining which changes are "major" – such as changes to hyperparameters, architectures, or guardrails – is discussed in more detail in Subsection 3.5.

Monitoring for un-approved AI components in third party products or services can take several forms. FIs can consider implementing processes to assess when the terms and conditions of a product or service have changed – triggering a legal review of the new terms in those cases – or when an analysis of release or update notes identifies the potential addition or meaningful modification of AI features. They can also consider measures such as periodic questionnaires or vendor re-assessments to detect embedded or connected AI that was added without the FI's knowledge or without appropriate scrutiny and visibility to teams involved in procurement or AI risk management. FIs with large and complex software environments could consider periodically reviewing their overall ecosystem for its AI posture, further strengthening their ability to identify the introduction of AI features in major applications. They can also consider contractual provisions that require notification in cases where AI components are added to or modified in a product or service that the FI consumes.

Employee vigilance is the most effective means of monitoring for the addition of unauthorised AI components by third parties; employee reporting, however, requires a robust AI risk culture (see a discussion of AI risk culture in Subsection 4.1).

Practice 4: Consider whether contracts and licenses with third parties providing AI products and services are sufficient to clearly address AI-specific risks.

Approach:

- Consider whether contracts with third party providers and software licenses for products and services adequately define AI-specific provisions around indemnity, data protection, cybersecurity, monitoring, and AI introduction.
- Determine the FI's willingness to use open-source AI models and systems, depending on the legal risks posed by their license terms.

Effective contracting, licensing, and risk sharing mitigate risk but do not substitute for the effective oversight and monitoring of AI use cases to prevent risks from occurring.

FIs already have legal practices in place that can ensure that contracts include appropriate performance, support, and audit provisions; these contracting practices remain relevant to AI. When negotiating or reviewing contracts or licenses with third parties on AI products or services, FIs can further consider some of the below AI-specific issues. The below list is neither prescriptive nor exhaustive, and contracting will vary significantly between FIs, providers, and jurisdictions. In some cases, FIs will be unable to secure some or all of the below measures in their contracts; the unavailability of desired protections or guarantees is an important input into risk-based procurement decisions. Where desired contractual terms are not available, FIs can consider mitigations such as compensatory testing.

AI-specific issues in contracting and licensing include:

- **Indemnity or shared responsibility for AI-specific risks.** While FIs already consider general service level criteria when procuring IT, new or enhanced AI risks may require additional indemnity or clarity on shared responsibility. Responsibilities may also be distributed throughout the value chain – such as in a case where an FI procures an AI service from a provider who themselves licenses an AI model from an upstream developer. Specific risks to assess responsibilities for include:
    - Intellectual property violations, such as cases where a provider's AI model was trained on copyrighted material.
    - Clear apportionment of responsibility for damages in case of failure.
    - Warranty for damages caused by outputs, such as harms to users from errors or defamation from Gen AI hallucinations. FIs can clarify the extent to which third parties are expected to assume responsibility for mitigating these types of behaviours in their expected use cases and can consider whether the warranty's extent is compatible with their risk appetite.
- **Data retention and protection.** FIs can consider the importance of AI-specific contractual provisions that clarify the extent or duration over which third parties can use their data, especially when data is related to sensitive use cases, for purposes like quality control or AI training.
- **Cybersecurity.** FIs can consider whether contracts or licenses provide sufficient protection against AI-specific security risks.
- **Monitoring.** FIs can consider contracting for third parties to share the outputs of risk-related monitoring with the FI.
- **AI introduction or modification.** For third parties that provide products or services to the FI, FIs can consider seeking disclosure of when AI components are introduced or substantively modified, or disallowing the introduction of AI components without the FI's prior consent. This is particularly relevant for products or services that do, or which could in the future, support embedded or connected AI.

FIs may also consider negotiating contractual obligations around AI-specific risk management, such as adherence to certain thresholds on AI-specific risk metrics. They should note, however, that this is not currently a common practice in the industry and may be challenging to enforce. Other non-AI-specific contracting practices that remain relevant include requesting a right to audit.

Procurement practices can also help FIs address legal risks associated with open-source AI models or systems. In addition to traditional procurement considerations like performance, expected total lifecycle cost, and support options, FIs can consider the suitability of license terms for managing the risks associated with open-source AI products. FIs may, for example, exclude certain open-source models or systems from some high-risk use cases should their license provisions indemnify the developer against all damages. Other open-source models may be distributed with licenses that are incompatible with for-profit use, such as provisions that require any software using the model to also be made available on an open-source basis.

Practice 5: Ensure that teams with AI-specific legal, technical, and risk-management skills are involved in procurement, contracting, onboarding, or other third-party risk management activities as appropriate.

Approach:

- Ensure that processes are in place for specialists in relevant skills, such as legal skills, technical skills in the development and operation of AI systems, or AI risk management, to be involved in third party risk management activities related to AI.

When managing the third-party risks of AI use cases, and especially when conducting procurement and contracting activities, FIs may find that their existing teams do not have the AI-specific technical and cross-functional skills required. These skills may include:

- Technical skills for testing and evaluating AI products or services, or interpreting the results of tests or evaluations. This can include AI/ML engineering skillsets, data science skillsets, or data governance and privacy skillsets.
- AI risk management skills for assessing the AI-specific risks of an AI product or service.
- AI risk management skills for evaluating the effectiveness of a potential third-party provider's AI governance and risk management processes.
- AI-specific legal experience to ensure that contracting and licensing considerations related to AI are effectively managed.

Existing procurement and legal teams may have certain upskilling or right-skilling needs, especially around AI literacy and risk identification. The process of identifying and augmenting skills related to AI governance and risk management is discussed in Subsection 4.1.

FIs may also design processes for engaging relevant internal experts in each domain when needed in the procurement, contracting, or onboarding process. FIs may find that, because the procurement of AI products and services can involve discretionary judgement and close collaboration with third party providers, directly including AI experts in the process can be more effective than using checklists or written procedures.

FIs may also benefit from engaging interdisciplinary teams in the design of AI-specific procurement practices and contracting standards.

Illustration 2.3.1: Standard Chartered Bank: AI Risk Management in Third Party Solutions

Identification: We have identified the various areas of entry into the organisation and have introduced control gates that include identification tools and mandatory requirement for entries into the Global AI Inventory.

Legal coverage and indemnity: Inclusion of RAI Principles-based requirements into third-party vendor contracts, with the capability for customisation where required.

Solution Risk Assessments: Risk is weighted by vendor type and solution.

Change Management: Pre-change transparency declaration requirements are incorporated within contracts. We also scan for the inclusion of AI capabilities through patches.

Ongoing Monitoring: Mandatory requirements for periodic reviews of performance with the vendor.

2.4 Enhance Use Case-Level AI Risk Management

This Subsection focuses on practices for assessing and managing the materiality of use case-specific risks. See Subsection 2.2 for a detailed discussion of the distinction between organisation-level and use case-specific risks.

Subsection 1.1 defined the distinction between "models", which are sets of mathematical relationships, and "use cases" which are business contexts and objectives that models and systems are applied to achieve. Most FIs have historically managed the risks of AI at the "model" level, in line with existing MRM practices. This approach was appropriate for traditional AI, where purpose-trained models generally only served a single use case. The introduction of Gen AI, and the overall shift in the industry towards "general-purpose" AI models that have a broad range of capabilities has changed this paradigm.

Table 2.4.1: Comparison of Risk Management for AI Types

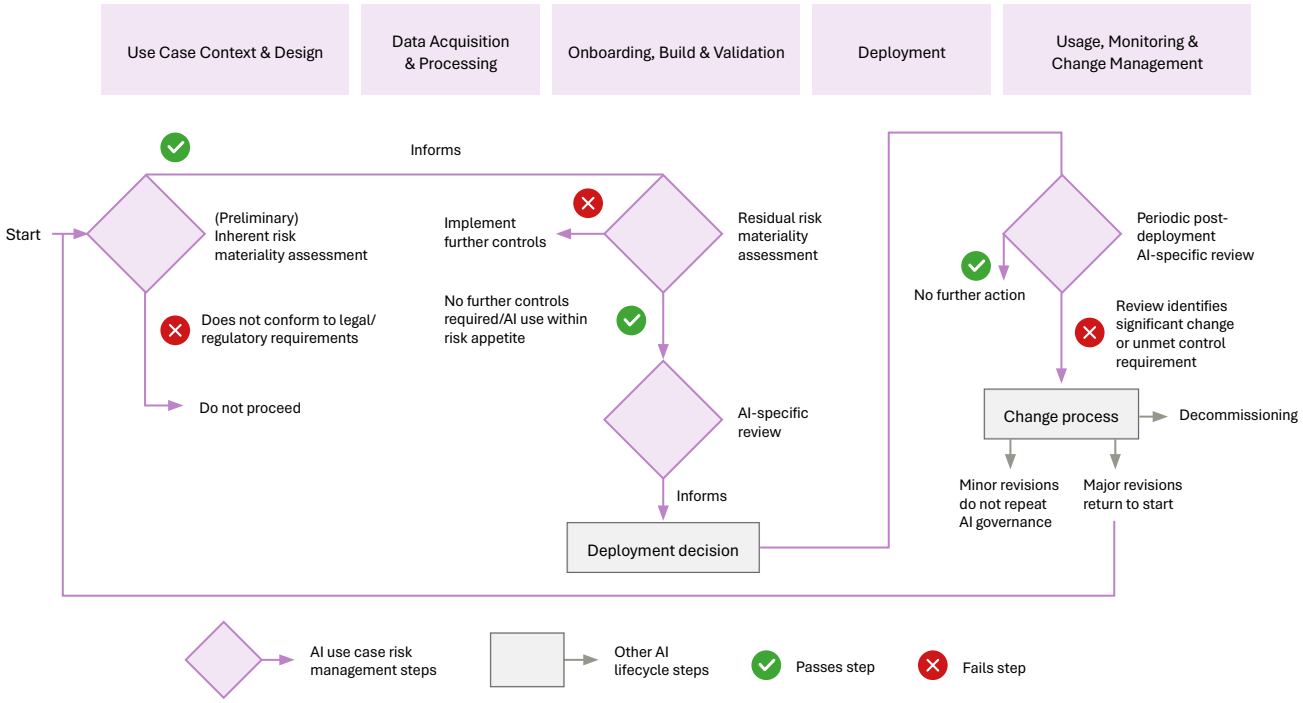
AI Type	Typical Use Cases	Implication for Risk Materiality Assessment
Traditional AI	Usually one specific business purpose.	Typically applied to a narrow use case. In such cases, "model risk" and "use case risk" are similar or identical. Complexity and operational independence are generally lower.
Gen AI	Serves broad business use cases, often implicating many processes.	Broad range of business applications increases the importance of thoroughly assessing each context of use, rather than just the model's risks. Complexity is generally higher.
Agentic AI	Serves broad business use cases, often implicating an even wider range of processes.	Large number of broad, complex business uses and complex interactions between numerous traditional AI and Gen AI models makes it essential to consider broader use case factors when assessing risk. Complexity and operational independence are both generally higher.

Some FIs continue to consider "model risk" for traditional, single-purpose AI and "use case risk" for Gen AI, Agentic AI, or other multi-purpose AI. This is a legitimate approach to addressing AI risk so long as all contextual AI-specific risks are considered.

In line with the terminology used in key industry frameworks (see a list in Appendix C), this Subsection uses the term "risk materiality" to refer to the extent to which an AI use case presents risks to an FI.

Each AI use case presents a specific level of risk materiality. Determining the specific risk materiality of an AI use case is important because it allows organisations to take a risk-based approach to govern those use cases, better allocating resources and tailoring mitigation strategies proportionately to risk. An effective AI risk management framework sets out practices that balance the stringency of governance activities against their cost, difficulty, and necessity.

Figure 2.4.1: Illustration of the AI Risk Management Approach



FIs begin the risk management process by defining a use case-level AI risk materiality assessment methodology. This is typically applied at specific points in the AI lifecycle. Suitable controls can then be applied based on the risks and degree of risk materiality identified. AI-specific review is then conducted both before deployment and periodically post-deployment. Figure 2.4.1 illustrates a high-level relationship between the (inherent and residual) risk materiality assessment and (pre- and post-deployment) AI-specific review.

An example of an AI use case with high residual risk materiality could be a credit underwriting system that, based on personal information about customers, makes key decisions about that customer's creditworthiness.

Examples of AI use cases with a medium residual risk materiality could be automated client sentiment analysis for relationship management, personal finance tools providing spending analysis and recommendations for customers, or claims automation, where there is human-in-the-loop oversight that avoids direct interactions between the AI system and the end customer.

An example of an AI use case with a low residual risk materiality could be an AI knowledge management chatbot, which interacts only with employees on an internal network and on non-critical tasks.

Each of these examples highlights residual risk materiality because it describes a case where controls are in place.

Consideration 5

Ensure that a framework is in place to manage the risks of each AI use case. This includes defining a risk materiality assessment approach, implementing a framework for inherent and residual AI risk assessments, applying controls that are commensurate with the risks identified, and conducting pre- and post-deployment AI-specific reviews as appropriate.

Practice 1: Define levels of risk materiality for AI use cases based on criteria relevant to the FI's context.

Approach:

- Define criteria for determining the risk materiality of an AI use case and establish discrete tiers of risk materiality, considering dimensions such as its impact, its complexity, and the degree of reliance on it.

Each FI is responsible for determining the criteria for assessing use case-level AI risk materiality, considering existing risk assessment approaches and the AI risk strategy discussed in Subsection 2.2. Some of the specific criteria that FIs may assess as part of their use case-level AI risk materiality assessment methodology are:[13]

- Impact of the AI use case on the FI, its customers, or its stakeholders.
    - Monetary and financial impact. The quantitative potential for an AI to impact an FI by causing losses, incurring costs, or resulting in foregone revenue (such as through lost trust or damaged relationships).
    - Severity and probability of impact on different stakeholders, including individuals. The potential for the use case to impact stakeholders – whether internal or external – in significant ways. Impact can also be affected by the volume and scope of usage, such as the number of people that interact with it or its outputs.
    - Reputational risk. The potential for AI use to generate controversy or negative publicity for an FI, even when it is legally compliant.
    - Options for recourse. The extent to which people impacted by an AI can seek effective remediation or challenge the AI's decisions. A use case with fewer options for recourse can have further impacts and be riskier.
    - Regulatory impact. The potential for AI use to engender compliance risk or regulatory penalties.
    - Use of personal data. The extent to which sensitive personal data is implicated by the AI use, which could potentially create opportunities for unjustified bias or for potential data loss or leakage.
- Complexity or novelty of the AI use case.
    - Complexity of the AI in use. The use of multiple datasets, complex architectures (e.g. LLMs), multiple models in concert (e.g. agentic AI), or other data or computational characteristics that can limit their interpretability or predictability.
- The degree of the FI's reliance on the AI use case.
    - Extent of automation of process of AI-driven decision-making. The dependency on an AI to supplement or substitute for human decision-making, which can increase the scope of that AI's impact. AI that is customer-facing without a human in the loop can represent a greater degree of process automation.

[13] Adapted from the criteria in Veritas Document 3: https://www.mas.gov.sg/-/media/MAS-Media-Library/news/media-releases/2022/Veritas-Document-3---FEAT-Principles-Assessment-Methodology.pdf

Risk materiality is often measured in a tiered approach, classified as "Low", "Medium", or "High". FIs can determine the best way to tier risk materiality within their organisation, considering which risk materiality tiers may be best suited to account for the AI risks that they face while also managing complexity. Effective risk materiality tiers are explicitly defined and unambiguous so that they can be consistently applied across the FI.

Definitions function best when they are consistent at the enterprise level and periodically revised taking into account changing AI technologies, regulations, and risks, as well as practitioner feedback.

Some FIs will consider using AI as part of their risk management. This could include using AI systems that support their risk function in AML monitoring, or to use AI systems as part of their AI governance approach.

The approach to AI use case-level risk management does not differ for AI use cases applied in risk management. In these cases, the risk management function is considered to be the business owner for the purposes of governance.

Practice 2: Define a process to assess the inherent risk materiality of AI use cases at the appropriate lifecycle stage, considering the fundamental characteristics of each use case.

Approach:

- Designate an appropriate process and methodology for assessing the inherent risk materiality of an AI use case as early as possible in the lifecycle, at latest before deployment. A preliminary version of the assessment can be followed by a full inherent risk materiality assessment in cases where development is more exploratory.
- Key steps in this process include the identification of use case-specific risks, such as those related to the choice of technology, data, or user.

The first element of an inherent risk materiality assessment is to establish the use case context and identify each of the specific risks that it may post to the FI and its stakeholders, as well as the drivers and determinants of those risks. FIs can leverage existing risk materiality assessment approaches as a starting point, with revisions as needed to address AI-specific risks. FIs can determine the significance of, and their tolerance for, the potential financial and non-financial impacts of various AI risks; some key risks like discrimination may be given a high significance. FIs will each decide whether to proceed with an AI use case based on their risk acceptance criteria.

Determining the tier or level of inherent risk materiality can take a variety of forms. Two common approaches are the flowchart method and the scoring method. The flowchart method involves Use Case Owners answering straightforward, binary questions about the use case and its design in a sequence. The scoring method involves developers completing a full questionnaire, with responses resulting in numerical scores that can correspond to tiers of risk materiality. These methods, and others, can also be combined, in accordance with existing practices and regulatory expectations for managing risk.

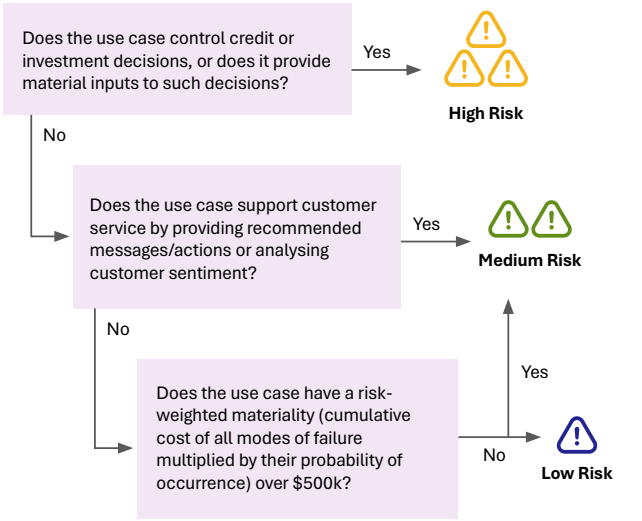
Figure 2.4.2 below illustrates the differences between these two suggested methods for assessing inherent AI risk materiality based on the characteristics of the use case. The criteria shown in this figure are simplified for the purposes of illustration and are not meant to be taken as an example of a real risk materiality assessment in use in the industry today.

Figure 2.4.2: Illustrative Inherent AI Risk Materiality Assessment Methods

Flowchart Method

Simplified for Illustration

AI Use Case



Scoring Method

Simplified for Illustration

AI Use Case

Does the use case control credit or investment decisions, or does it provide material inputs to such decisions?	10
Does the use case support customer service by providing recommended messages/actions or analysing customer sentiment?	5
Does the use case have a risk-weighted materiality (cumulative cost of all modes of failure multiplied by their probability of occurrence) over $50k?	1
$500k?	+3 *(cumulative)*
$1000k?	+3 *(cumulative)*
$5000k?	+3 *(cumulative)*

Total of

0-4	5-9	10+
Low Risk	Medium Risk	High Risk

Inherent risk materiality assessments can play an important role in shaping the development of AI use cases when they are conducted early in, or at the beginning of, the AI lifecycle. This may be impractical for some use cases, such as those that are exploratory in nature and the details of whose usage is expected to evolve as they approach deployment.

When FIs choose to delay the full inherent risk materiality assessment past the beginning of the design stage, they can consider minimally completing a light, "preliminary" version of the inherent risk materiality assessment before development. This could include:

- Assessing whether the intended use case conforms to legal and regulatory requirements
- Assessing whether the intended use case conforms to the FI's AI principles and relevant policies, such as on acceptable technology use.
- Socialising the intended use case with key stakeholders, like AI governance and risk management teams or legal professionals.
- Estimating, roughly, the tier of risk materiality that the use case is expected to fall under based on the information available.

The purpose of a preliminary inherent risk materiality assessment is to determine whether it is appropriate to proceed with further development (a simple go/no-go decision) and to set a rough expectation of the level of governance that will later be applied.

FIs can, once the exploratory phase of development is complete and the parameters of the use case are clear, proceed to conduct a complete inherent risk materiality assessment. Delaying the inherent risk materiality assessment may result in rework, additional effort, or throwaway effort.

FIs can benefit from defining clear roles and responsibilities in the risk assessment process. In addition to the Use Case Owner, who may hold overall accountability, other functions, like those that oversee risk, may be responsible for providing inputs, or ensuring that policies are followed. Use Case Owners or Builders benefit from involving functions like risk and compliance, legal, and cybersecurity in the assessment process.

Practice 3: Define a process to evaluate the residual risk materiality of AI use cases prior to deployment, considering the established controls and guardrails.

Approach:

- Designate a process at the end of the build or onboarding stage for qualitatively and/or quantitatively assessing whether a use case performs adequately on its AI risk-related metrics and benchmarks.
- Define a methodology for considering both the results of this assessment and the inherent risk materiality to assign a tier of residual risk materiality.

A residual risk materiality assessment typically takes the form of qualitative or quantitative assessments of an AI use case's behaviour in a setting representative of its intended real use. It takes place, by its nature, at a point in the lifecycle after controls have been implemented, which is typically at the end of development and prior to deployment. A residual AI risk materiality assessment assesses an AI use case in the real conditions of its deployment, with all guardrails applied. As FIs consider the results of the residual risk assessment, they may revise the use case or add controls to further reduce residual risk; a use case may pass iteratively through the addition of controls and the assessment of residual risk as many times as are required. Residual risk materiality can be periodically re-assessed after deployment as part of the AI-specific review.

The outcomes of an AI use case can be assessed through both qualitative and quantitative evaluations. It can be valuable for FIs to consider standardising their residual risk assessment methodology in an AI evaluation framework, which is a documented standard that can support the consistent analysis of AI use cases and the assessment of controls. An evaluation framework can be used as a tool in determining a use case's residual risk materiality and can be an input into the application of controls. Such a framework can include:

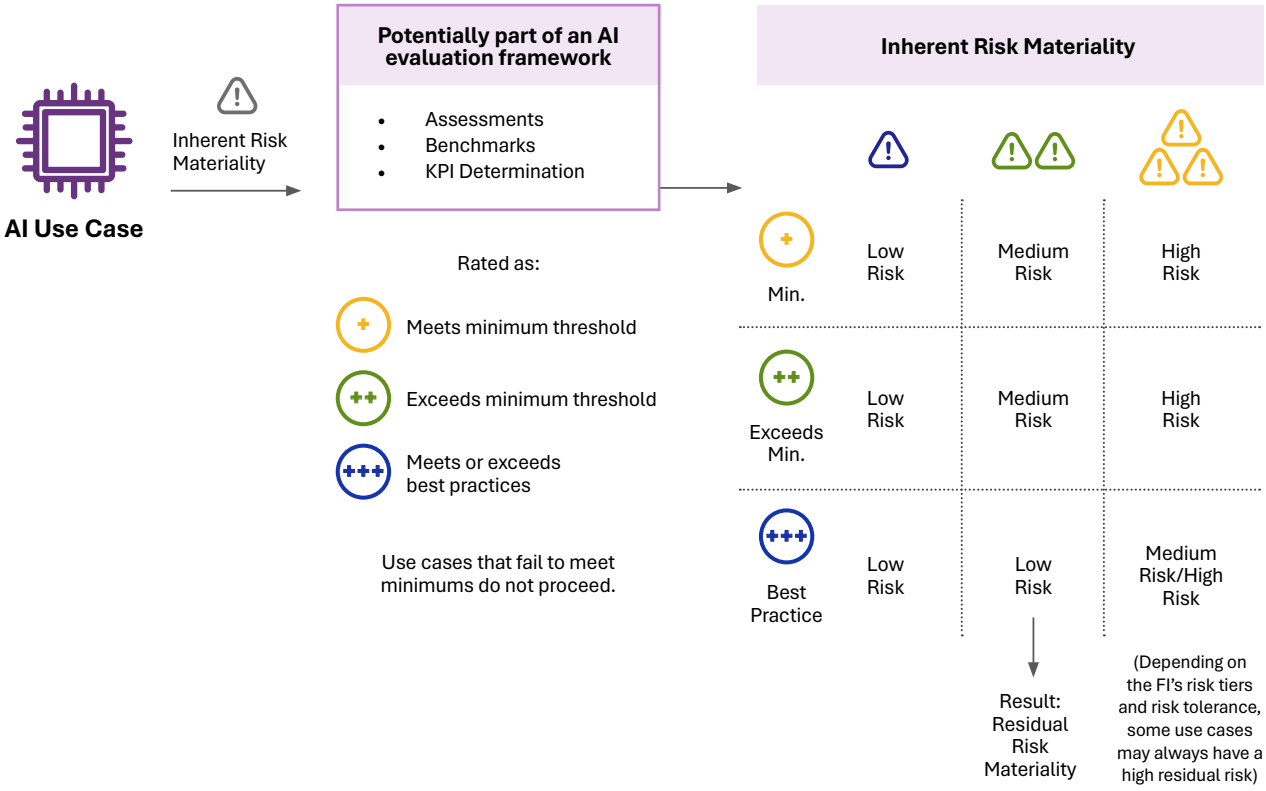
- **Recommended steps and workflows.** FIs can consider documenting a consistent approach to calculating metrics and for assessing performance.
- **Standard qualitative and quantitative evaluations for each AI risk and type of use case.** FIs can consider defining templates for evaluating certain risks or use cases.
- **Standard evaluation methodologies to use across the enterprise.** FIs can consider setting out standard, structured methodologies for testing – like simulation testing, scenario testing, and red teaming – based on AI risks and use case types.
- **Tools to use in evaluation.** FIs that have preferred evaluation tools can specify when these should be used in their evaluation frameworks.

FIs benefit from taking an iterative approach to the design and implementation of their evaluation frameworks, prioritising technologies for planned AI use cases. Designating clear ownership and review responsibilities for an evaluation framework can help to ensure that it remains current.

To calculate residual risk materiality, the results of the assessment can be rated, according to the use case's KPIs, as meeting minimal acceptance thresholds, exceeding them, or meeting a high, industry-leading standard of performance. The acceptable performance in a quantitative or qualitative risk materiality assessment is set on a case-by-case basis. For example, an AI use case for lending approval may require a very high minimum accuracy threshold due to the severe consequences of inaccuracies. In contrast, a use case for sorting employment applications, with appropriate human supervision and where sorting will not lead to the denial of employment, may have a different accuracy threshold.

Figure 2.4.3 illustrates one potential approach to doing so. In this example, for a use case with low inherent risk materiality, achieving the minimum evaluation results does not elevate that use case to overall "high" risk materiality. This is because the potential for this use case to cause harm is limited. Such a use case may be an internal knowledge management chatbot, which has few opportunities to cause serious harm. Conversely, a higher-risk materiality use case, such as one used for credit decisioning, may be given a "high" residual risk materiality rating unless it achieves a particularly high standard of fairness, accuracy, and reliability on its residual risk materiality assessment.

Figure 2.4.3: Illustrative Approach to Residual Risk Materiality Assessment Based on Inherent Risk Materiality



The approach illustrated in Figure 2.4.3 is notional and does not represent a specific recommendation on how to aggregate the results of the inherent and residual AI risk materiality assessments. The approach to doing so depends on each FI's risk tolerance, specific assessment methodology, and definitions of the tiers of risk materiality in its context. A key post-deployment control is to revisit, at an appropriate frequency, the risk materiality of AI use cases through AI-specific review.

As when implementing inherent risk materiality assessments, FIs can benefit from defining specific responsibilities for residual risk materiality assessments. While Use Case Owners may hold overall accountability, for example, Builders or other technical roles may be responsible for performing assessments or consulting other relevant experts. Risk materiality assessments may lead to disagreements between stakeholders or functions on the ratings of specific use cases. FIs can improve the robustness of their risk management by making risk materiality assessments consistent across the enterprise and by designating a final authority that can resolve differences or escalations. This final authority can be particularly effective when it is the same body or function that was responsible for creating and defining the risk materiality assessment framework, which gives the function insight into that framework's intended interpretation.

Practice 4: Identify, uplift, or create controls to be applied to each AI use case based on its risks and risk materiality.

Approach:

- Identify enterprise controls relevant to managing AI risk and, where gaps exist, supplement existing controls or add new controls.
- Assign controls to relevant AI types and risks, based on tiers of risk materiality.

FIs have existing libraries of core internal risk controls to manage technology, operations, and security risks. These existing controls may benefit from customisation to manage different types of AI-specific risks and risk materialities. When appropriate, FIs can document the fitness of their existing controls against the risks of AI, taking into account these use cases and potential risk materialities, and determine if enhancements or additional controls are required.

Figure 2.4.4: Illustration of an AI Control Library

Control	Modalities	Relevant Risks	Applied at Risk Materiality
Control 1	All AI	Lack of explainability	 Low - High Risk
Control 2	Gen AI	Hallucination/ Fabrication/ Confabulation	 Medium - High Risk
Control 3	Traditional AI	Adverse or inappropriate impact to individuals and groups	 High Risk Only

Controls can be specific to deployment patterns, modalities, lifecycle stages, or risks. Not all controls are relevant to every type of AI or use case – red-teaming, for example, could be relevant to Gen AI, but less so to conventional analytics models. Controls function best when designed with contextual specificity in mind – differing, for example, to accommodate the different nature and risks of use cases used for general productivity, for customer relations, or for investment management.

It is important to apply controls consistently throughout the AI lifecycle, including during risk assessment, AI-specific reviews, and other testing and monitoring activities. FIs can also incorporate AI testing and monitoring controls into the FI's SDLC practices.

Some controls can be applied at different levels of intensity depending on the use case's risk materiality. For example, human-in-the-loop moderation could be applied on a sampling basis for less material use cases, and for AI use cases with a higher residual risk materiality, it could be applied more frequently or made mandatory for all outputs (see a further discussion of this technique in Subsection Section 3). Tiering of control intensity based on use case risk materiality ensures that governance is proportionate and effective.

Practice 5: Define an approach for conducting an AI-specific review of AI use cases prior to deployment, confirming the risks identified, the use case's risk materiality, and the appropriateness of risk mitigations.

Approach:

- Define a pre-deployment review methodology that assesses key functional (technique justifiability, risk assessment, guardrails, data suitability) and technical criteria (metric appropriateness, metric results, vulnerabilities).
- Define a methodology for varying the depth and/or autonomy of the AI-specific review based on use case risk materiality.

AI-specific review is a pre-deployment assessment by an autonomous party. AI-specific review can be conducted for each AI use case based on its residual risk materiality as a pre-deployment activity as an input to a deployment aligned with the FI's risk appetite.

Where feasible and appropriate, FIs can leverage or uplift existing risk and governance processes – such as those that may exist under MRM or their SDLC practices – with AI-specific considerations to serve as the AI-specific review. Traditional MRM, for example, may sufficiently validate traditional AI but can require enhancements for newer advances in AI such as Gen AI to address their specific risks.

When responsibilities for AI-specific reviews are divided between teams (e.g. MRM where models are accessible, and existing IT asset management activities for third party black box systems), FIs can ensure that AI-specific review does not duplicate existing controls. An overall point of accountability can be defined to ensure that this de-duplication is effective.

The use of Gen AI by FIs usually involves taking third-party LLMs which are "black boxes" and applying them to various use cases. Unlike traditional AI – where FIs have generally focused on validating the "model" – for Gen AI the focus shifts to validating the application of the foundation model to those use cases.

This "use case" validation typically focuses on areas such as:

- **Fit for purpose:** Ensuring that the Gen AI system is fit for purpose, such as by verifying the accuracy and relevance of its outputs in context.
- **Gen AI-specific risks:** Identifying and addressing new risks introduced by Gen AI, including hallucinations, toxicity, and offensive content.
- **Risk-based guardrails:** Assessing the effectiveness and completeness of the implementation of guardrails or controls that address Gen AI-specific risks, such as adversarial testing, output filtering, or human oversight.

AI-specific reviews are most effective when conducted by parties not directly involved in the development, deployment, or operation of the use case. This ensures objective questioning of decisions made during the design, development, or deployment of an AI use case and provides an impartial perspective on its risks.

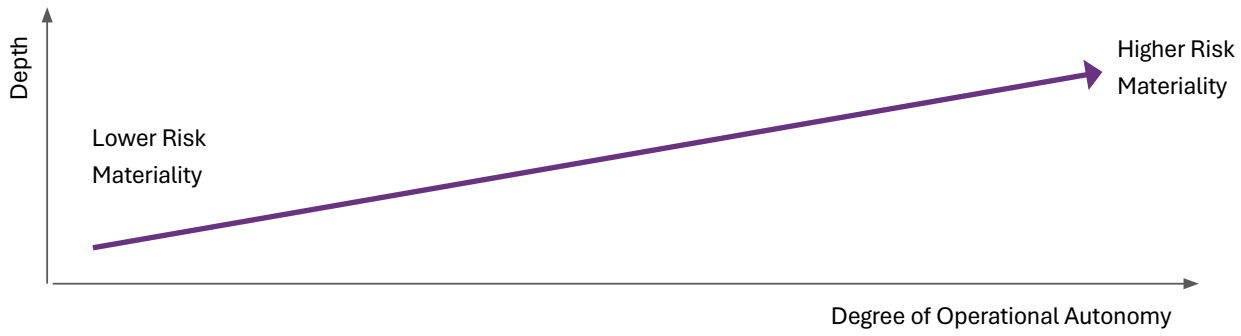
The list below is a non-exhaustive set of criteria for FIs to consider in AI-specific reviews.

- Functional criteria:
  - Justifiability and conceptual soundness of the chosen technique or algorithm for the use case where the FI has visibility on the choice of technique or algorithm. This is typically only assessed in an initial review, not post-deployment.
  - Use case AI risks and the correctness of the use case's risk materiality tiering (such as the MindForge AI risks enumerated in Appendix B). These include assumptions, limitations, the implications of the use case on upstream or downstream risks, such as other linked AI use cases.
  - The appropriate application of AI-specific guardrails and adherence to procedures related to use case risks, especially as relate to cybersecurity, ethics, transparency or explainability, and the FI's values.
  - The suitability of the dataset(s) used by the AI model(s) – specifically, whether appropriate risk mitigations are in place around the dataset, and whether the use of Personally Identifiable Information (PII) has been appropriately justified.
- Technical criteria:
  - Appropriateness of the selected performance metrics related to AI risk, including whether appropriate thresholds have been set for each metric (see Subsection 3.1). Reviewers may, in particular, focus on assessing the correctness of the protected attributes identified for the purposes of assessing fairness.
  - Results of performance metrics related to AI risk. FIs can leverage their AI-specific risk evaluation frameworks in doing so, where appropriate, and can focus on the performance of the use case on its protected attributes for the purposes of identifying fairness.
  - Vulnerabilities to key AI risks, such as by performing security-related benchmark testing, stress testing, or adversarial testing if required and applicable to the use case.

In addition to the AI-specific criteria above, FIs may also choose to integrate AI-specific reviews in generic IT reviews. These may also consider criteria such as performance, data quality, security, and system integrity.

FIs are responsible for defining a detailed and replicable list of criteria for AI-specific review, as well as for continuously evolving this list to address feedback or changes in AI risks. FIs may have a limited ability to review third party AI products and services to the level of detail that they may assess AI developed or customised in-house. Where this limitation exists, FIs can instead rely on compensatory testing (for Gen AI use cases, especially using techniques such as simulation, scenario, and sensitivity testing and red teaming), and on third party disclosures like "AI cards", discussed in Subsection 2.3.

Figure 2.4.5: Varying Depth and Autonomy of AI-Specific Review



FIs may vary the depth of AI-specific reviews based on the residual risk materiality of that use case and their risk appetite. For lower-risk materiality AI use cases, FIs can use simple review questionnaires that address the above criteria and review the metric calculations completed by Builders rather than conducting them again. For higher-risk materiality use cases, FIs might require more detailed inputs from Builders, while also requiring reviewers to rigorously verify those submissions – including by gathering required information themselves, where applicable, and by performing their own tests and calculations of relevant performance metrics related to AI risk. As the risk materiality of the use case increases, reviewers typically shift from reviewing the results of risk-related checks to conducting those checks themselves.

AI-specific reviews are always conducted by individuals or teams with a degree of autonomy – those that are not involved in the development, deployment, or operation of the AI use case. An example of a less-autonomous reviewer may be peers to the Builders or Use Case Owner who were not directly involved in the use case; more autonomous reviewers may be drawn from other teams with less prior exposure to or involvement in the use case or from the risk management function. In general, increased autonomy can improve the stringency of risk management by ensuring that reviewers are able to effectively question the decisions made by the use case team.

Gen AI can be particularly challenging to review using existing model/AI validation techniques. Models used in Gen AI tend to:

- Be trained on large proprietary datasets for which FIs may not be able to perform complete data quality checks.
- Have deep and complex architectures for which existing explainability techniques are insufficient.
- Be provided as SaaS, where FIs have limited direct access to underlying models/systems.

FIs could manage the risks of Gen AI products or services by treating their underlying models as black boxes for which direct validation is not feasible. Reviews may instead focus on compensatory testing, deploying guardrails correctly, and analysing the use case's outputs.

For use cases with a lower risk materiality, review can focus on the conceptual soundness of the grounding prompt/few-shot training examples provided by the FI. Review can also focus on outcomes and sensitivity analysis to ensure that the use case performs as intended, even under extreme scenarios.

For use cases with a higher risk materiality, more extensive input/output controls (e.g. filtering, pseudonymisation) are typically in place, and review can focus on the correct functioning of these controls. A range of compensatory controls exist for doing so, including scenario testing, adversarial testing, and benchmarks or evaluations. In each case, proportionately to the use case's risk materiality, FIs can consider combining industry-standard prompts or benchmarks with customised ones that take into account the use case's context.

Testing for Gen AI technologies is a rapidly evolving field. The AI Verify Foundation's Global AI Assurance Sandbox continues to innovate technologically in this space. In May 2025 it produced a Starter Kit for Safety Testing of LLM-Based Applications that FIs can draw on for guidance.

FIs can vary the degree of autonomy when conducting AI-specific reviews depending on the use case's residual risk materiality. Use cases with a lower risk materiality may only require a peer review. When risk materiality is high, FIs can benefit from designating independent reviewers, such as an internal audit function or an external reviewer.

FIs may also consider defining and documenting clear responsibilities both for the reviewer and for ensuring that the review takes place.

Practice 6: Ensure that AI-specific reviews of AI use cases are conducted periodically post-deployment, with their frequency based on factors including the risk materiality of the AI use case.

Approach:

- Define a process for conducting post-deployment AI-specific reviews at a frequency proportionate to the risk of the use case. In addition to risk materiality, risk factors in determining frequency can include incidents or risk events, scope changes, re-certifications, or changes to the AI model or system.

AI-specific reviews can also be conducted periodically after deployment. These include the same criteria, accountabilities, and depth/autonomy considerations as pre-deployment AI-specific reviews (see Practice 5). They benefit from a focus on metrics and criteria related to post-deployment performance, such as data and model drift. They may also focus on post-deployment events like incidents and risk events.

Each FI can determine the appropriate frequency and depth for these post-deployment reviews by considering factors such as the following:

1. Risk materiality tier of an AI use case, which includes considerations like scope and impact.
2. Incidents, issues, risk events, or other potential causes for concern identified in the course of the use case's regular monitoring or in enterprise tracking of KRIs (if applicable).
3. Jurisdictional regulatory requirements that apply to AI, such as requirements to be re-assessed or re-certified. In some jurisdictions, for example, MRM rules may require independent post-deployment validation of all AI models.[1]
4. Time since the AI use case's last review (if applicable).
5. Changes to the AI model or system, or changes in scope or usage, since deployment or since the last review (if applicable). These include technical changes initiated by the FI, changes initiated by third-party vendors like model providers, or changes initiated by end users.
6. The cost of AI-specific review and the FI's risk appetite.
7. External risk factors, such as enhanced public scrutiny on a type of use case or known issues with similar use cases or technologies.

The occurrence of some of these factors may be identified by the FI during its review of AI risk and AI-specific KRIs across use cases, described in Subsection 2.2.

Some FIs perform post-deployment AI-specific reviews on a weighted sampling basis, considering some of the above criteria to select a subset of use cases at different levels of risk materiality to review in each period.

> FIs can consider monitoring adverse external developments related to their AI techniques, models, and systems, especially those sourced from third parties. When a third-party model that the FI uses has been reported as implicated in incidents in other organisations, a similar incident impacting the FI could occur. A reasonable degree of external monitoring, such as periodically assessing media reporting, can serve as a form of early warning and can trigger a proportionate AI-specific review.

The results of post-deployment AI-specific reviews can be assessed to determine if an action is required (such as change, pause, or decommissioning) or if there has been a material change in the AI use case's risk.

Illustration 2.4.1: Risk Materiality Assessment and AI Specific Governance: Prudential Perspective

Prudential's AI governance framework is built on a risk-based model aligned with its 8 AI Ethics Principles, applying risk-based assessments for inhouse-built and 3rd party solutions.

The structured AI governance framework ensures responsible development, deployment, and monitoring of AI use cases. This framework aligns with the organization's risk tolerance and regulatory obligations.

All AI use cases undergo a pre-deployment review based on risk materiality, conducted by independent domain experts. Each use case is minimally assessed for legal, security, privacy, regulatory, and ethical compliance, with higher-risk cases subject to more rigorous evaluations. Key pre-deployment reviews align largely with the Handbook, including considering the need for additional customization or enhancement of controls.

Post-deployment reviews are conducted regularly, with frequency and depth based on risk materiality. These reviews consider the time since the last review, regulatory changes, system or use case updates, and incidents. Similar evaluations are applied as per pre-deployment assessments, covering functionality, performance, data quality, and risk management.

This risk-based approach ensures AI systems remain compliant, effective, and aligned with Prudential's goals.

(Continued on next page)

(Continued from previous page)

Example Fictional Use Case: AI-Assisted Insurance Claims

In this fictional scenario, Prudential has deployed an AI system to support the assessment and approval of insurance claims for selected health products. This use case is rated as moderate risk due to its supplementary role.

Pre-deployment evaluations addressed key risks including, accuracy, bias, data quality, and explainability. Additional controls were implemented, including:

- Human oversight for manual intervention.
- Decision logging for auditability, and explainability.
- Protection of sensitive data (including personal data) via Data Loss Protection, masking, and encryption.
- Ensuring Quality of claims through periodic model validation and retraining.

Given its potential impact on customers and the company's reputation, the fictional use case undergoes regular performance reviews. Post-deployment review is scheduled based on risk materiality to ensure continued effectiveness and relevance of controls.

The review process includes but not limited to:

- Reassessing relevance and risk materiality.
- Reviewing model performance (e.g. false positives/negatives).
- Validating results against agreed thresholds.
- Evaluating guardrail effectiveness and the need for adjustments.
- Periodic recertifications (frequency of recertifications is determined by the AI system's risk tier) to ensure ongoing compliance with AI ethics and fit-for-purpose.

Illustration 2.4.2: UOB's Approach to Risk Materiality Assessment and AI-Specific Review

UOB adopts a structured, risk-based framework for AI governance, anchored by a risk materiality assessment that informs the level of oversight and review required for AI models. This framework evaluates both the likelihood and potential severity of harm that an AI model could pose to the bank and its stakeholders including customers, employees, and broader society.

(Continued on next page)

(Continued from previous page)

Each AI model is assigned a materiality rating, either high or low, based on a weighted scoring methodology that incorporates quantitative thresholds and qualitative judgment. In addition, AI models are categorised by their intended use (e.g. risk management, regulatory reporting), and this classification, together with the materiality rating, determines the depth of review, the independence of the review process, and the responsible governance team. Where a model is used across multiple scenarios, each use case is separately assessed and tiered according to its specific context and impact, ensuring that oversight is appropriately calibrated to the model's application.

For example, high-materiality AI models used for regulatory purposes are subject to independent validation by specialised teams. In contrast, lower-materiality systems may undergo proportionate peer review. Regardless of the review pathway, all reviewers should not be involved in the development process to ensure objectivity and effective challenge.

The review scope typically includes both technical and ethical assessments, covering areas such as conceptual soundness, regulatory compliance, fairness, explainability, and performance. These assessments are documented using structured templates and recorded in a centralised AI registry, promoting consistency, traceability, and transparency across the AI landscape.

UOB's governance framework is designed to be adaptive. With the rapid evolution of technologies such as Gen AI, the bank maintains a continuous review of its governance practices to ensure they remain robust and fit for purpose. This evolving approach reflects UOB's commitment to responsible AI governance, balancing innovation with accountability to ensure that AI technologies are deployed safely, ethically, and with appropriate oversight.

2.5 Ensure AI Inventory Capabilities

An AI inventory is a repository of certain core attributes related to an FI's AI use cases. This inventory is an important component of effective AI governance and risk management; by establishing a timely and complete inventory, FIs can enable the effective risk management of their use cases across the enterprise.

An AI inventory serves two primary purposes:

1. To "ensure that AI are only used within the scope in which they have been approved for use, e.g., the purpose, jurisdiction, use case, application, system, and other conditions for which they have been developed, validated and deployed."[16]
2. To support strategic decision-making on AI, the fulfilment of regulatory requirements and disclosures related to AI, the monitoring of use case and enterprise risks in conjunction with non-AI specific risk management activities, and overall portfolio-level AI governance and risk management.

Each FI will define which information is appropriate to track in their AI inventory. They will also determine how best to structure the AI inventory: whether it will be one system or several, and whether it will be dedicated to storing AI-specific information or not. FIs may leverage existing inventories in the enterprise to do so, such as by uplifting them to record additional AI-specific attributes or by noting links between existing inventories that, collectively, provide the functionality of an AI inventory. Using an existing inventory, creating a new inventory, or any hybrid approach is legitimate where it fulfils the two above purposes.

An AI inventory can link to and help the FI track documentation on their AI use cases, which is described throughout Section 5 of this Handbook. The AI inventory typically does not include the extent of information captured in AI documentation, instead serving as a curated subset of information that the FI documents on AI use cases.

AI inventories continue to be useful to FIs that use Agentic AI. Some FIs may leverage their existing system and use case inventories to do so; others may also find it useful to retain an inventory of AI "agents". This ensures that the risks of agents that are used across use cases are appropriately tracked while minimising duplicated effort in governance.

A discussion of practices for maintaining effective inventories of Agentic AI is provided below in Future Perspectives.

Consideration 6

Ensure that core AI-specific information on AI use cases is recorded in an inventory and ensure that a process is in place to maintain the AI inventory, so that information about new, updated, or decommissioned AI use cases is reflected accurately.

Practice 1: Ensure that a form of AI inventory, designed in consideration of existing inventory systems and practices to be suitable and proportionate for the FI's context, is in place to capture a core set of AI-specific information on AI use cases.

Approach:

- Provide the functionalities of an AI inventory by leveraging one or several existing non-AI-specific inventories or creating additional inventory systems as required.
- Use AI inventory system(s) to track the purpose and scope of AI use cases, the type(s) of AI employed, the data used in each use case, AI-specific risks and associated mitigations associated with each use case, the status of development or deployment, and governance of each use case.
- Leverage information from the inventory for enterprise-wide strategic decision-making on AI.

There are a range of approaches to providing the functionality of an AI inventory. Each is appropriate where it meets the needs of the FI's context, such as by enabling relevant employees to access the information they need to govern AI. When deciding on an approach to defining an inventory, FIs can consider whether prospective inventory options have the technical capabilities required to track AI-related information and interact with AI-related processes and teams, and the completeness with which existing inventories track information on AI use cases.

Approaches to providing AI inventory functionality may include:

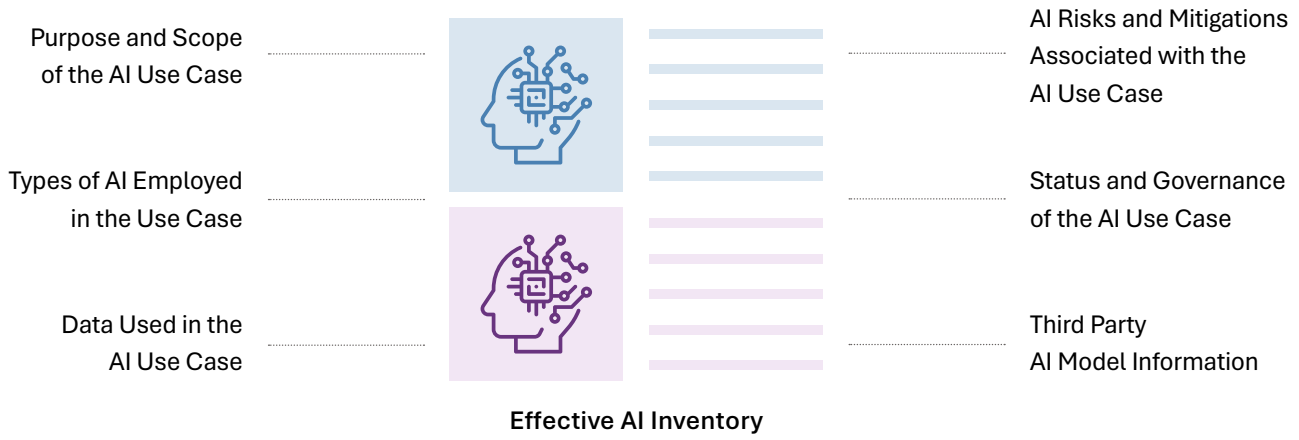
- Using an existing non-AI-specific inventory system to track AI-specific information.
- Linking several different existing inventories (such as business unit-specific inventories or subject-specific inventories like asset libraries and risk registers) to function, collectively, as the AI inventory. Doing so may increase the importance of data integrity measures and standard approaches to data collection so that those inventories can integrate effectively and meaningfully.
- Creating, if appropriate, a new inventory system to track some or all AI-specific information.

Relevant non-AI-specific information – such as the impact of the use case on broader enterprise risks or information on its security – can continue to be recorded per existing good practices. Recording AI-specific attributes does not replace, and should be considered as supplemental to, the recording of non-AI-specific information that FIs already track.

An effective AI inventory would minimally record six general attributes about the AI use cases in an FI. These six attributes are illustrated in Figure 2.5.1. These six attributes do not represent all of the information that is important to capture on AI use cases; other crucial details for effective management and governance of AI include attributes like responsible and accountable stakeholders, upstream and downstream dependencies, and the architecture and components of the system. Such general attributes are not, however, unique to AI.

FIs can best manage AI-specific use case risks by continuing to use their existing inventory and asset management practices to track standard IT attributes for AI use cases, and in addition, ensuring that AI-specific information is tracked in an appropriate way. These six AI-specific attributes are broad and can be captured in a range of ways and variety of fields, depending on the FI and its particular needs.[14]

Figure 2.5.1: Core Attributes of an Effective AI Inventory



[14] The MAS Information Paper on AI Risk Management gave several examples of inventory fields in use in the industry today. These include: the AI's purpose and description, scope of use, jurisdiction, model type, model output, upstream and downstream dependencies, model status, risk materiality rating, approvals obtained for validation and deployment, responsible AI requirements, waiver or dispensation details, use of personally identifiable information (PII), personnel responsible such as owners, sponsors, users, developers, and validators; and for third-party AI, attributes such as the AI provider, model version, and endpoints utilised.

These attributes represent a broad baseline of information to be tracked and should not limit FIs from tracking additional information in their AI inventory where it is relevant to do so. Each attribute may, depending on the FI, correspond to several distinct inventory fields. They are described in Table 2.5.1 below.

Table 2.5.1: Key Attributes and Justifications for an AI Inventory

Attribute	Description	Justification
Purpose and Scope of the AI Use Case	The intended use(s), user(s), and jurisdictions of the AI use case, and whether the use case interacts with or impacts external stakeholders such as regulators or customers.	The central function of the AI inventory is to ensure that AI is used for its intended/approved purpose. Besides intended business functionality, key elements of the purpose and intended scope of an AI use case are the people who will use it or be impacted by it, the jurisdictions where it will be used, and whether usage is internal or external – each of which can have significant ramifications for governance and control.
Type(s) of AI Employed in the Use Case	The type of AI used (e.g. regression, Gen AI, Agentic AI, computer vision) and the input/output modality (text, images, etc.).	The AI inventory can clearly indicate which kind of AI is applied to a given use case, allowing FIs to select controls that are meaningful for their own AI use. Many of the AI-specific risks are unique to certain types or modalities of AI use.
Data Used in the AI Use Case	The specific types of data used in the AI use case, their sources or provenance, and any associated sensitivities (e.g. personal or confidential data).	A key vector of AI-specific use case risk is data exposure, particularly when data is sensitive or biased. FIs can benefit from distinguishing between the use case's training data and operational data (e.g., user input or enterprise data in a Retrieval-Augmented Generation (RAG) architecture). The depth of information captured on use case data can vary based on risk appetite and use case materiality. When third party AI use cases have limited information available on data types, capturing them to the extent that is possible helps prepare FIs to assess their riskiness.

(Continued on next page)

(Continued from previous page)

Attribute	Description	Justification
AI-Specific Use Case Risks and Associated Mitigations	The risk materiality of the AI use case, as well as relevant information on the AI-specific use case risks and their mitigations.	By recording the risk materiality rating, AI-specific use case risks, and related controls/mitigations, FIs can ensure that that use case is appropriately governed in the future, that information on its risks is accessible to all relevant parties, and that risk mitigations can be assessed for inclusion. Many FIs track use case risks in a risk register, and if this register is sufficient to track AI-specific risks, FIs can consider linking to and leveraging that register rather than replicating it. The depth of documentation of AI-specific use case risks, controls, and mitigations that were identified in the risk materiality assessment can vary based on the use case's risk materiality.
Status and Governance of the AI Use Case	The Use Case Owner, status of the AI use case (e.g. proof of concept, in development, in deployment, decommissioned), as well as, where pertinent, the AI-specific approvals or exemptions received, and AI-specific evaluations conducted.	This Handbook describes having at least one AI-specific decision point, the deployment decision, for approving AI use cases. FIs may add more decision points if needed. The AI Use Case Owner (which is different from a data owner) needs to be identified and clearly documented in the inventory to ensure that this process is completed appropriately. The inventory can help FIs determine if a use case has been deployed and if it has the necessary approvals. By also tracking key governance information – dates of pre-deployment, post deployment, etc. – it helps to identify non-compliant use cases.
Third Party AI Model Information	Basic information on third party models or systems used in the use case, such as their license type, providers, and third-party disclosures like AI Cards.	To enable appropriate risk management and AI governance and risk management, FIs can explicitly document their usage of third-party AI in their inventory as well as the necessary information for governing them, such as version numbers for third party foundation models. This includes, in particular, their licenses and any information disclosed by the third party in the course of procurement (such as by linking to the third party's "AI Card").

FIs typically record all of their AI use cases in the AI inventory, including those that have very low risk materiality. This is important to ensure that FIs have a comprehensive picture of the riskiness of their overall portfolio of AI use and that they can monitor and govern their AI use at a strategic level. For some AI use cases, especially those sourced from outside vendors, only a limited amount of information on the use case may be available.

To fulfil the second objective of an AI inventory – to facilitate strategic decision-making on AI and AI-specific use case risks – AI inventories are typically able to provide enterprise-wide information on AI use across business units and geographies. This objective does not necessarily require that FIs create new inventories, have a single AI inventory, or adopt new software specific to AI. Each FI can, therefore, determine how best to integrate enterprise-wide AI-related information in their context.

Practice 2: Ensure that processes are in place and that roles and responsibilities are defined such that the AI inventory is well-maintained and kept up to date.

Approach:

- Define clear roles and responsibilities for overall AI inventory ownership, such as by designating a relevant control function, and for managing inventory entries for individual use cases.
- Create effective policies and processes to keep inventories accurate and up to date.

Creating an AI inventory is not a one-time exercise; an effective inventory is consistent, reliable, and up to date. This makes it possible for an FI to promptly identify and respond to AI-specific use case risks where they occur. An up-to-date AI inventory will also provide the FI with a comprehensive and accurate view of its AI-specific use case risks, allowing management to better understand the interdependencies of each use case and to make informed decisions on managing them.

FIs can define clear roles and responsibilities to maintain an accurate and up-to-date AI inventory. This can include designating an inventory owner to provide oversight of the inventory's overall function, including by establishing processes and standards to ensure its integrity. Typically, this ownership will sit with a control function – such as model or technology risk management – which will oversee the inventory itself, liaise with related business and technology functions, and conduct periodic reviews as needed. Primary accountability for the accuracy and completeness of individual use case entries typically rests with the respective Use Case Owners, who will generally be responsible for ensuring that their entries are updated and maintained as changes occur. Clearly defined roles and responsibilities for periodic reviews help ensure that any errors or inconsistencies are identified, escalated, and addressed. FIs can also consider assigning responsibilities for regularly assessing the overall effectiveness of the inventory and recommending improvements where necessary. Ensuring that the responsible parties are of appropriate seniority can reinforce the effectiveness of the FI's AI risk management and inventory practices.

FIs use a range of tools, sometimes existing and sometimes AI-specific, to fulfil the functions of the AI inventory. Sometimes FIs elect to include some or all AI inventory information in an existing enterprise asset inventory, such as by creating additional fields that are triggered when an asset is tagged as containing AI.

Other FIs use a spreadsheet software to maintain their AI inventories, usually in conjunction with a traditional asset inventory. This approach, while functional for small organisations, has important drawbacks: spreadsheet software typically does not have the comprehensive access control, input validation, and version management that a dedicated inventory software would. Inventories maintained in spreadsheet software, especially when maintained by multiple individuals, can be prone to operational issues.

FIs may also establish dedicated AI inventories, usually in parallel to existing non-AI-specific asset inventories. Most dedicated AI governance tools include AI inventory and metadata tracking capabilities by default; these can sometimes include dedicated registries for risks and controls.

Inventory tools may offer features such as the automated tracking of model characteristics and issues, as well as the identification of interdependencies between systems. Some tools also include AI detection capabilities that can scan for AI systems operating within the FI's on premises or cloud environments. While these features can be helpful in supporting AI inventory management, they are still evolving and currently have limitations. For example, most tools are not yet able to reliably detect embedded AI components in third-party products or services that are not hosted directly by the FI. As the market for AI inventory tools continues to evolve, these capabilities are expected to improve.

For this reason, most FIs record some basic information on AI use cases at the beginning of the AI lifecycle, and record all other attributes as early as possible, in all cases prior to deployment. Recording relevant information early in the AI lifecycle can facilitate collaboration between enterprise functions and can support the identification of risks and application of AI guardrails. Recording relevant information prior to deployment ensures that deployment approvals and post-deployment risk monitoring can use a complete set of AI inventory information to make informed, prudent decisions. FIs may choose to vary the depth and completeness of information recorded in the inventory based on an AI use case's risk materiality.

To ensure that the inventory is up to date, control processes may specify periodic check-ins with the Use Case Owner to ensure that they update the information in the inventory. Regular updates to attributes such as the use case's scope of use or model can help FIs identify cases where post-deployment AI-specific evaluation is required. Post-deployment AI-specific evaluations also function as a control on the inventory, with deviations from the attributes described in the AI inventory identified and escalated by evaluators as needed.

To ensure that their AI inventories remain effective and relevant, FIs can designate responsibilities for periodically reviewing their design and the types of information that they track. This can allow the FI to improve the design of the inventory based on feedback taken, lessons learned, and in response to new technologies.

Roles involved in managing the AI inventory will vary by organisation. A non-exhaustive set of examples of responsibilities around the AI inventory may include:

- Use Case Owner: Responsible for entering the use case's information into the inventory and periodically updating it as required.
- AI-specific reviewer: Responsible for checking that the information entered in the inventory is accurate and complete for a specific use case.
- Inventory owner: Responsible for the overall health and management of the inventory and for the database schema, which defines the specific fields and information to be tracked. May also be responsible for periodic monitoring or spot checks to identify gaps in inventory documentation.

The inventory owner, depending on the organisation, may be part of the AI governance and risk management function, or may be the owner of an existing inventory asset.

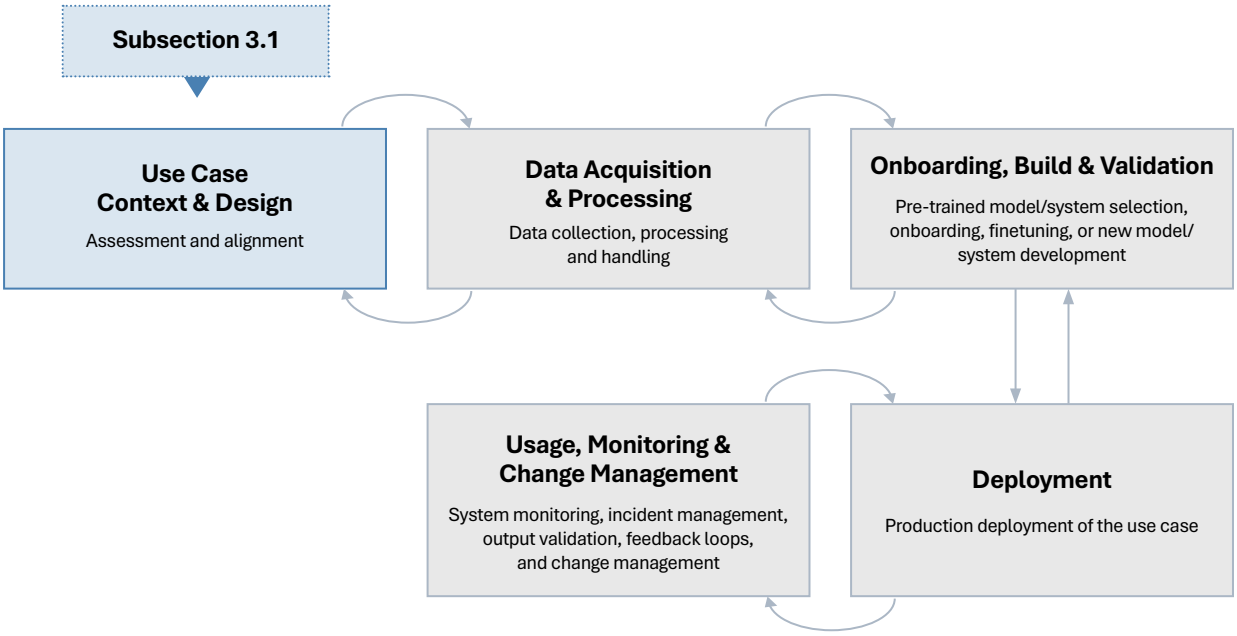
3. AI Lifecycle Management

3.1 Use Case Context and Design

An effective commencement to the AI lifecycle is foundational to ensuring the responsible and effective deployment of an AI use case. This step involves defining the business context and objectives of the AI use case as well as performing an initial assessment to determine whether the proposed AI use case aligns with the organisation's mission, values, priorities, and appetite for risk. FIs with a robust existing SDLC in place may find that the Practices below are already well-addressed in, or can be easily retrofitted to, their existing processes.

This Subsection describes considerations that make up part of the AI lifecycle (see Figure 3.1.1). It is intended to be applied to each AI use case.

Figure 3.1.1: Use Case Context and Design in the AI Lifecycle



Consideration 7

Assess the AI use case to ensure that the intended use is compatible with ethical, regulatory, and organisational standards, and determine the level of governance to be applied to the use case based on its inherent or expected risk materiality.

Practice 1: Establish ownership for the AI use case and ensure alignment with organisational standards and values for ethical and responsible AI use.

Approach:

- Identify a Use Case Owner to provide accountability and oversight for the AI use case.
- Perform a preliminary inherent risk materiality assessment of the use case to identify stakeholders at risk, assess regulatory and legal requirements, and ensure alignment with organisational values.

FIs typically have a process in place for identifying AI and distinguishing AI use cases from technologies that are not AI. This AI identification process is discussed in Subsection Section 2; AI governance and risk management process typically begins when AI is identified.

It is important to begin the use case lifecycle by assigning a Use Case Owner to ensure accountability and oversight. This is the first and foundational step for AI use case management; while the individual in the role of owner may change over time, well-defined ownership ensures that governance and risk management tasks will be effectively and consistently completed. This applies equally to use cases based on third party AI products or services, where the Use Case Owner has a critical role in governance.

To prevent misdirected development resources, it can be helpful to establish a go/no-go checkpoint in the preliminary stage of AI use case development to ensure alignment with legal and organisational policy requirements, such as technology policies or risk-based controls. This preliminary inherent risk materiality assessment is discussed in Subsection 2.4. A use case that is clearly not aligned with the law or with the FI's principles may not be worthwhile to pursue.

This preliminary inherent risk materiality assessment can include:

- Identifying stakeholders at risk: Determine whether there are individuals or groups who could be systematically disadvantaged due to the AI use case. Document potential harms and benefits while outlining fairness objectives, metrics, and relevant mitigation strategies.
- Assessing regulatory and legal requirements: Verify that the AI use case will adhere to applicable laws and regulations, such as those on data collection, processing, and AI-generated outputs. This includes ensuring that the risks of non-compliance can and will be mitigated to avoid legal or regulatory consequences.
- Aligning with organisational values: Ensure that the AI use case adheres to the FI's core values, AI principles, and other relevant commitments. Ensuring alignment with organisational values for AI use is particularly important for use cases that use, or will use, third party AI products and services, which may have been designed or developed without those values in mind.

Practice 2: Perform an inherent risk materiality assessment to determine the risk tiering of the AI use case and to guide proportionate governance efforts.

Approach:

- Perform an inherent risk materiality assessment on the AI use case to identify risks related to the use case and inform the application of controls and governance.

When planning for an AI use case, it is important to conduct an inherent risk materiality assessment; this will identify potential risks and will an appropriate level of governance based on their potential to impact the FI and its stakeholders. The process of conducting an inherent risk materiality assessment is discussed in Subsection 2.4.

The product of this assessment is the identification of discrete AI risks that could be engendered by the chosen use case and the assignment of a tier of inherent risk materiality. This risk tiering system enables the organisation to implement a proportionate level of governance based on the potential risks associated with the use case.

To ensure effective risk management and ensure that appropriate guardrails are applied as early as possible and to avoid rework, the inherent risk materiality assessment benefits from being conducted as early as possible in the AI development lifecycle and reviewed if the use case's scope, assumptions, or risk profile evolve throughout the lifecycle. In line with their existing IT management processes, FIs typically require that a risk assessment, including an identification of both inherent and residual risks, be completed and approved by the relevant governance body before deployment.

Practice 3: Capture AI use case-related information in an AI inventory to enable transparency and support risk management.

Approach:

- Register AI use case-related information within an AI inventory to enhance transparency, support risk management and facilitate governance.

FIs can ensure that AI use cases are tracked, and their risks mitigated, by registering them in an AI inventory as early in the lifecycle as is feasible, and updating that inventory entry as appropriate as the use case evolves throughout its lifecycle. This inventory provides a structured approach for relevant stakeholders to document, assess, and manage AI use cases, ensuring they align with governance, operational, and strategic objectives. It enhances transparency by offering a clear view of where and how AI is used, supports risk management by identifying risky use cases, and facilitates governance by ensuring that each use case is monitored and maintained throughout its lifecycle. The information tracked in an AI inventory is discussed in more detail in Subsection 2.5.

Practice 4: Design the AI use case to operate with a proportionate and practical level of human oversight.

Approach:

- Consider whether, and to what extent, the AI use case requires human oversight, based on its nature and its risk materiality.
- Ensure that the desired level of human oversight is enabled throughout the lifecycle of the AI use case.

In the context of an AI use case, human oversight is the role that designated employees play in the course of the use case's typical operation – controlling, supervising, and making decisions based on inputs from the AI use case. FIs typically decide at the use case design stage how they will best integrate human oversight into the AI use case. This decision can be based on the use case's inherent risk materiality, its intended modality, and the feasibility of human oversight.

Human oversight complements the monitoring of an AI use case by integrating human judgement in the operation of the use case to help mitigate AI-specific risks. It may not be suitable for all use cases; for AI that handles high volumes of tasks or high throughput, such as autonomous chatbots or fraud detection engines, real-time human oversight may not be feasible and may instead take the form of post-event monitoring or exception handling to detect anomalies. FIs can make an informed decision during the design of the use case, and can revisit this decision throughout the lifecycle, as to the degree of human oversight that is desirable and practical, balancing operational considerations with risk.

Human oversight can take several different forms, as outlined in the IMDA/PDPC Model AI Governance Framework[9]:

- Human in the Loop: Where an employee retains full control of the use case and either approves or executes actions at crucial points in operation. This may take the form of an employee reviewing AI-generated customer service emails before approving them for sending, or an AI advisor that suggests actions to a relationship manager.
- Human over the Loop: The AI use case produces outputs or actions where an employee has the option to modify or overrule its outputs or intervene in its operations but is not required to take any action in the course of normal operation. This includes situations where an AI use case escalates errors or anomalies to an employee for review, or where an employee reviews some or all outputs after the fact.

- Human out of the Loop: The AI use case produces outputs without direct involvement from an employee, such as in an autonomous chatbot. While this means that the AI can take actions autonomously, it does not imply the absence of other forms of oversight, such as post-deployment testing, monitoring, and review.

Considering whether a use case requires human oversight, and in what form, is an important first step because enabling it can have a range of implications for the rest of the use case's design. It is important to clearly define and document relevant roles, responsibilities, and escalations, where appropriate, and to ensure that employees in those roles have sufficiency competencies and authority to perform that role. FIs may have risk-related controls in place that specify the conditions under which human oversight of AI is expected.

Illustration 3.1.1: Julius Baer's Two-Stage Toll Gate Process for AI Use Case Governance

Julius Bär

Julius Baer has established a two-stage tollgate process to ensure safe, trustworthy and effective development, deployment and use of AI across Julius Baer. This process, overseen by Julius Baer's cross-functional Responsible Artificial Intelligence Council (RAIC), provides a robust framework for risk management and accountability.

The first tollgate occurs at the conclusion of the ideation phase, prior to the development phase. At this stage, a short assessment identifies and eliminates any possible cases that are prohibited by the EU AI Act and identifies potential high-risk use cases. This is critical, given Julius Baer's global operations. Furthermore, the EU AI Act is widely regarded as a pioneering regulatory framework which is referenced by other emerging standards for alignment.

The second tollgate takes place at the end of the development stage at the latest. During this step, the RAIC conducts a holistic AI risk assessment of AI use cases prior to production deployment. Julius Baer is developing a risk-based approach classifying the use cases according to their associated risks. Based on this classification, the RAIC will determine the (i) stringency of the AI Risk Assessment Process, (ii) degree of controls, (iii) oversight requirements, and (iv) reassessment frequencies. To ensure comprehensive risk management capabilities for AI and Gen AI, Julius Baer has integrated AI-specific risks into its existing risk definitions. As part of the second toll gate, individual risk type owners evaluate AI-relevant risks within their areas, recommend mitigation strategies as needed, or reject a given use case.

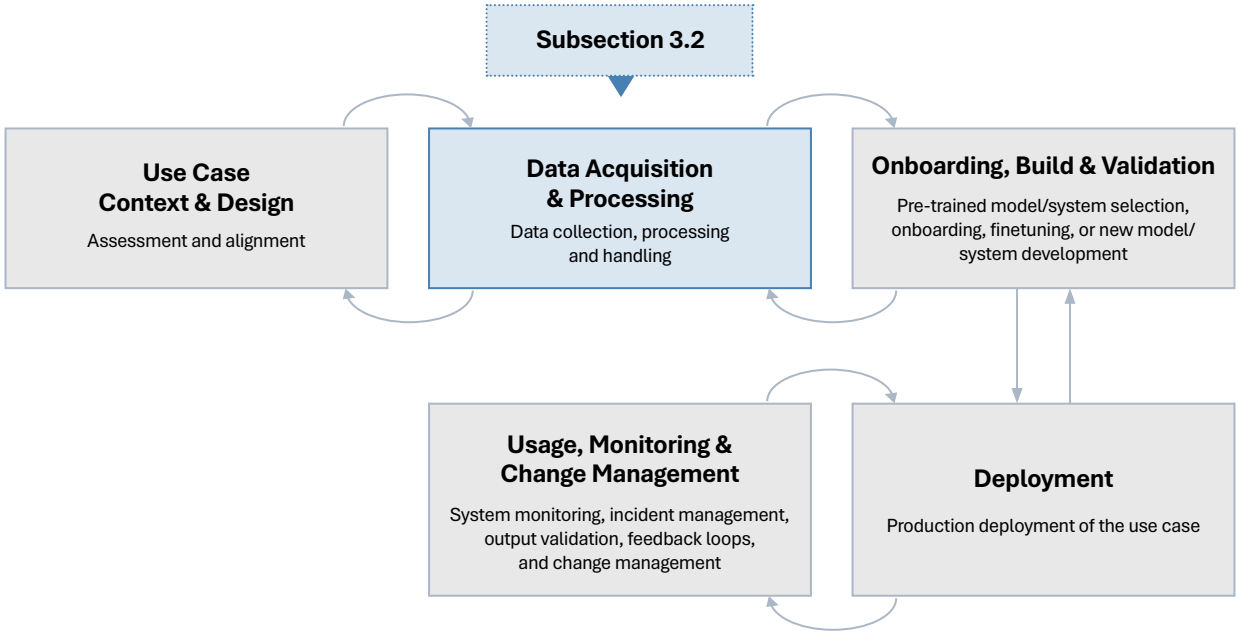
The aim of this process is to stop use cases that might be outside of Julius Baer's risk appetite as early as possible and to govern those Julius Baer wants to go ahead with as efficiently as possible, optimising time to deployment and driving re-use.

3.2 Data Acquisition and Processing

Data is a crucial component of artificial intelligence and plays a key role in is required for model development, training, evaluation, and operation. As FIs scale AI adoption, they may encounter increasing challenges in handling vast and diverse datasets, grappling with the volume, veracity, variety, and velocity of data. These characteristics often lead to business challenges related to data that need to be addressed, such as risks of bias, privacy breaches, data leaks, unauthorised manipulation, and unauthorised access.

This Subsection describes considerations that make up part of the AI lifecycle (see Figure 3.2.1). It is intended to be applied to each AI use case, and describes data management practices that are relevant to both traditional AI, Gen AI, and Agentic AI, whether built in-house or onboarded from a third party provider.

Figure 3.2.1: Data Acquisition and Processing in the AI Lifecycle



It is important that FIs continue to apply existing data governance practices and norms, and that they comply with relevant regulatory expectations on data protection and privacy. Key points of reference include the frameworks listed in Appendix C. This Subsection will focus specifically on additional AI-specific practices related to data (both first- and third-party data) used in AI use cases, and will be relevant both to AI training data and to data used in the course of the AI use case's operation.

Consideration 8

Evaluate whether the intended use of data in the AI use case is compatible with ethical, regulatory, and organisational standards.

Practice 1: Ensure that the use of data complies with ethical standards, regulatory requirements, and organisational policies or standards.

Approach:

- Assess the data intended for use in the use case for key risks and compliance requirements.
- Align the use of data with all relevant regulations, guidelines, and organisational policies and standards.

FIs will each face a range of data- and AI-related regulations and guidelines relevant to the industries and jurisdictions in which they intend to operate the AI use case. In identifying their data treatment obligations, FIs can consider factors such as their business operations and regional presence, the location of data subjects, data collection and processing activities, cross-border data transfers, and contractual obligations.

Some high-risk AI use cases or data types may require additional controls. FIs can ensure that their risks are addressed by continuing to rigorously apply existing policies, standards, and data management or governance practices to all data used in the AI use case, including supplementary ethics reviews and escalations, when a dataset is especially sensitive or impactful in nature. Examples may include health data (such as may be used in some fields of insurance), biometrics, or other forms of private customer or employee information.

FIs using customer data for the purposes of AI training may wish to ensure that they have specific consent for doing so. Where customer data is used, FIs can also ensure that the intended use case will include appropriate protections and will be privacy-preserving.

Some of the key regulations and guidelines for reference include, but are not limited to, those listed in Appendix C.

Practice 2: Ensure that the use of any third-party data complies with intellectual property rules, contractual obligations, and licensing rights.

Approach:

- Assess the data intended for the use case to ensure that it adheres to intellectual property rules and third-party contractual and licensing terms.

It is important for FIs to ensure that data used for AI use cases aligns with relevant third-party requirements by verifying specific consent restrictions, licensing terms, and compliance requirements. Third-party data used in AI is generally expected to align with the purposes for which it was obtained or as specified in the contractual/licensing agreement. In other cases, such as when using a Gen AI system that was trained on externally sourced creative content, FIs can ensure that the use of that content was compliant with relevant rules in the jurisdictions in which it operates.

FIs can manage their data risks by ensuring that third party data is appropriately verified and is used for its intended purposes.

Consideration 9

Adopt appropriate data management practices that address risks and limitations when processing data for AI use cases.

Practice 1: Ensure that data used for the AI use case is fit for purpose.

Approach:

- Assess, proportionately to the risk materiality of the use case, the fitness of datasets used. This can include identifying sources, assessing ethics, assessing quality, and assessing representativeness and balance.
- Document the results of the data fitness assessment.

It is important that that data used in AI use case is fit for purpose, high-quality, and relevant, which is essential for generating accurate and useful insights, predictions, and content. Poor data quality, which can take the form of incomplete, inaccurate, unrepresentative, or off-topic data, could result in unreliable AI outputs, impacting decision-making and business outcomes.

Key considerations for ensuring data fitness include:

- **Identifying data sources:** Engage with data owners, stewards, subject matter experts, and third party data or model providers, as appropriate, to ensure that data sources are clearly documented and that they are relevant to and representative of the business problem and AI use case objectives. This identification can support the FI in ensuring that only necessary data is collected.
- **Assessing data ethics:** Ensure that the use of the identified data sources aligns with the FI's ethical principles, such as non-discrimination. Data whose use would not be ethical is generally not considered to be fit for purpose. For third party data or models, FIs can perform due diligence, where feasible and proportionate to risk.
- **Assessing data quality:** Ensure that identified data sources align with key data quality dimensions outlined in the FI's existing data management/quality framework. While FIs may not control third-party model training data, they can continue to apply relevant data quality checks when their organisation's data is used, such as in the case of Retrieval Augmented Generation (RAG). For example, when leveraging Gen AI for question answering based on organisational knowledge repositories, FIs may need to verify that the underlying data comes from verified sources and meets relevant quality standards. FIs may request information on data quality from third party providers; this is discussed in more detail in Subsection 2.3.
- **Assessing representativeness and balance:** Ensure that identified data sources are sufficiently representative of the AI use case's intended context and range of users. Representativeness in AI includes fairness-related characteristics, such as ensuring that the dataset contains a reasonable and balanced mix of examples across attributes like race, gender, and language, but can also include ensuring that data includes a range of tasks, queries, and challenges that could reasonably be foreseen in operation, including edge cases.

In addition to AI-specific considerations, FIs can continue to leverage their existing data quality controls, which assess factors like relevance, accuracy, completeness, recency, and appropriate documentation.

When using synthetic data, FIs can assess and manage associated risks, including data representativeness, regulatory compliance, data provenance and model generalisation.

FIs considering using synthetic data for AI can consider the following principles:

1. *Is it statistically representative?*
   Synthetic data is only effective if it reflects the statistical properties, distributions, correlations, and completeness of the original dataset. Preserve data boundaries and mimic patterns of missing information.
2. *Is it fit for purpose?*
   Ensure that models trained on synthetic data perform similarly to those trained on real data, preserving key metrics like feature importance and predictive accuracy.
3. *Does it protect sensitive information?*
   Safeguard privacy by avoiding exact record matches, ensuring row novelty, and minimising risks of re-identification, singling out, or linkage to external datasets.

FIs using generated data can also consider, where appropriate, whether that data violates copyrights or other intellectual property rules.

For more details on guardrails for synthetic data generation, refer to the Proposed Guide on Synthetic Data Generation by the Personal Data Protection Commission Singapore.

Practice 2: Justify the use of personal attributes in the AI use case.

Approach:

- Assess and justify the usage of personal attributes and relevant guardrails within the AI use case.
- Appropriately document the results of that assessment.

FIs can mitigate some of the most serious AI-specific risks by assessing and justifying the inclusion of personal attributes in the AI use case, balancing benefits like improved performance or accuracy against fairness concerns (such as discriminatory treatment) and privacy risks (such as re-identification or data leakage). They can do so by systematically identifying the personal attributes used in the AI use case and clearly documenting the justification for including each.

Where personal attributes are used with risk mitigations in place – such as pseudonymisation or other masking techniques – FIs can document these mitigations as part of their justification exercise. The presence of effective privacy-preserving techniques can improve the strength of the justification for using personal attributes. FIs can also consider documenting justifications for the absence of privacy preserving techniques – such as where data is only useful where it contains protected attributes and where the risk of leakage is within the risk appetite.

A thorough justification exercise may aid Builders in identifying unnecessary attributes and removing them from the use case.

"Personal attributes" are data features that include, but are not limited to, information commonly referred to as "personal data". In many jurisdictions, personal data refers only to data that is re-identifiable to an individual. Thoroughly anonymised data would not meet this standard, but nevertheless contains personal attributes.

In addition to any privacy obligations resulting from the use of personal data, this Practice recommends that FIs be attentive to their use of both personal data and personal attributes that do qualify as personal data; even when data is not re-identifiable, its use can have a significant impact on an AI use case's fairness. This conforms to the approach and terminology used in both the FEAT Principles and Veritas Methodology; see Appendix A for the definition of "personal attributes".

Practice 3: Document metadata and data sources related to the AI use case in accordance with organisational data management policies and regulatory expectations.

Approach:

- Ensure sufficient documentation of metadata and data sources used in the AI use case, where doing so is feasible and where the FI has sufficient access to and ownership of that data.

Typical good data management practices include the adequate documentation of metadata and sources for both the data used to train AI model(s) and other data used in the system. This latter category can include datasets that are accessed at the retrieval and inference stage in a RAG architecture. Documenting key characteristics of that data and its lineage, and where synthetic data was used, ensures transparency, facilitates model validation and audits, enables root cause analysis, and supports reusability.

The depth of that data documentation depends on the FI's existing standards and regulatory expectations around data management; FIs can document the data used in AI use cases to the extent that doing so is feasible. Documentation can be retained according to non-AI-specific standards and regulatory expectations. Managing third party datasets is discussed in more detail in Subsection 2.3.

Practice 4: Ensure that appropriate data access controls are implemented based on the nature of selected AI use case.

Approach:

- Implement appropriate data access control to ensure that data associated with the use case remains secure.
- Designate responsible individuals to approve access requests for newly created data and conduct periodic reviews to assess whether data access associated with the use case remains valid and aligned with current requirements.

To safeguard data and maintain privacy, it is important that FIs implement robust data access controls for the AI use case in accordance with data management norms, regulatory expectations, and internal policies on access control that are already in place. These controls can address both input data used in the AI use case and any derived output data to prevent unauthorised access and misuse. Controls may include appointing employees to approve access requests for newly created data in the context of an AI use case and conducting periodic reviews to validate the necessity and appropriateness of granted access.

Access management extends to data sources, data pipelines, configuration files, and model versions that play a role in an AI use case's development. In addition to controlling access to these files and environments based on the F's data and information access policies, FIs can manage the risk of data poisoning or other types of unauthorised modification by implementing robust logging and change management systems at these intermediate stages.

Data access controls become critical when leveraging RAG for Gen AI use cases, as it relies on integrating external data sources with Gen AI models. This introduces distinct data access risks, such as unauthorised access, data leakage, data poisoning, and model manipulation.

To mitigate these risks effectively, it is advisable to implement robust user authentication and authorisation mechanisms. This ensures that only authorised individuals have access to the system, with their access levels being restricted according to their roles. It is essential to establish a comprehensive system and data access permission framework that ensures that the AI use case does not permit users to access data that would otherwise be restricted to them. A recommended method to achieve this is to develop role-based access controls that regulate retrieval and modification capabilities based on user credentials.

Practice 5: Establish clear ownership of any derived or transformed data to be used in the AI use case.

Approach:

- Designate clear data stewardship, including for transformed or derived data used in the AI use case, to maintain a proportionate level of oversight, quality control, access management, and security.

FIs may consider establishing clear ownership of transformed data used in AI use case, such as data in feature marts for analytical AI and vector/indexes for Gen AI, to ensure accountability. As source data often undergoes transformation in the context of AI development and deployment, such as feature engineering, data processing or anonymisation, ownership may need to extend beyond raw data to include transformed or derived data. The designated owner can be responsible for maintaining its quality and governance.

FIs can refer to their existing, non-AI-specific data management practices, as well as industry norms and relevant regulatory expectations, when doing so. This can include the definition of clear accountability for data lifecycle management practices for derived or transformed data – notably for the secure management, sanitisation, and destruction of training data, unnecessary model versions, and model outputs.

Practice 6: Identify and mitigate bias in training and test datasets.

Approach:

- Apply bias mitigation measures to ensure fairness along key protected attributes in training and testing datasets used for AI use cases. These may include, insofar as it is proportionate to use case risk, bias-aware data collection, dataset evaluation, technical tests for bias, and data processing strategies for correcting imbalances.

It is important, especially for higher-risk materiality AI use cases or those with material stakeholder impact (such as to customers or employees), that FIs carefully address bias in training and testing datasets to prevent unfair, discriminatory, or inaccurate outcomes. This requires implementing robust bias mitigation strategies throughout the data collection, model training, and testing phases.

Key bias mitigation approaches for structured data include:

- **Bias-Aware Data Collection:** FIs may consider adopting inclusive data collection using techniques like stratified sampling to balance demographic representation. This may include designing sampling methodologies around inclusion and assessing the completeness of data collection at a population level.
- **Systematically Evaluate Datasets to Identify and Mitigate Bias:** FIs could systematically evaluate training and test datasets for disparities in representation across key demographic groups (such as gender, race, age).
- **Applying Technical Tests for Bias Detection:** FIs can apply tests like Disparate Impact Analysis and Chi-Square tests for independence to detect bias. These tests are conducted during the AI model evaluation stage and highlight potential model behaviour issues related to data; results can prompt changes to the dataset for improved representation.
- **Strategies to Correct Imbalances:** FIs can mitigate bias by re-weighting the training data to correct imbalances or re-sampling underrepresented groups to ensure fair outcomes.

Bias mitigation for unstructured data remains an emerging field. FIs can reference industry norms to identify good practices when collecting and transforming unstructured data for AI use cases.

Establishing clear ownership for data – including transformed attributes and features – is an important part of bias management. Doing so ensures appropriate accountability for bias mitigation strategies and improves the traceability of data management decisions.

Illustration 3.2.1: AI Data Acquisition and Processing at DBS

As part of DBS' holistic Responsible Data Use (RDU) framework, appropriate safeguards must be taken during data acquisition and processing to support lawful, ethical, and fair AI use in the bank. This includes comprehensive approval checks and data management controls to ensure regulatory considerations and data standards are upheld.

For example, when deploying DBS-GPT, an internal Gen AI assistant accessible to staff to support content generation, information retrieval, and workflow automation, the project was thoroughly evaluated by a cross-functional Responsible AI (RAI) taskforce and approved by our RDU Committee. This comprehensive review process, involving senior management and experienced subject matter experts, ensures holistic assessment of key risks and mitigants before deployment. Rollout to DBS' core markets outside of Singapore was also subjected to approvals from respective location's legal & compliance team and data council so that location-specific requirements are met. Extensive data management controls were also implemented during the development and deployment of DBS-GPT. Some of the key controls implemented are as follows:

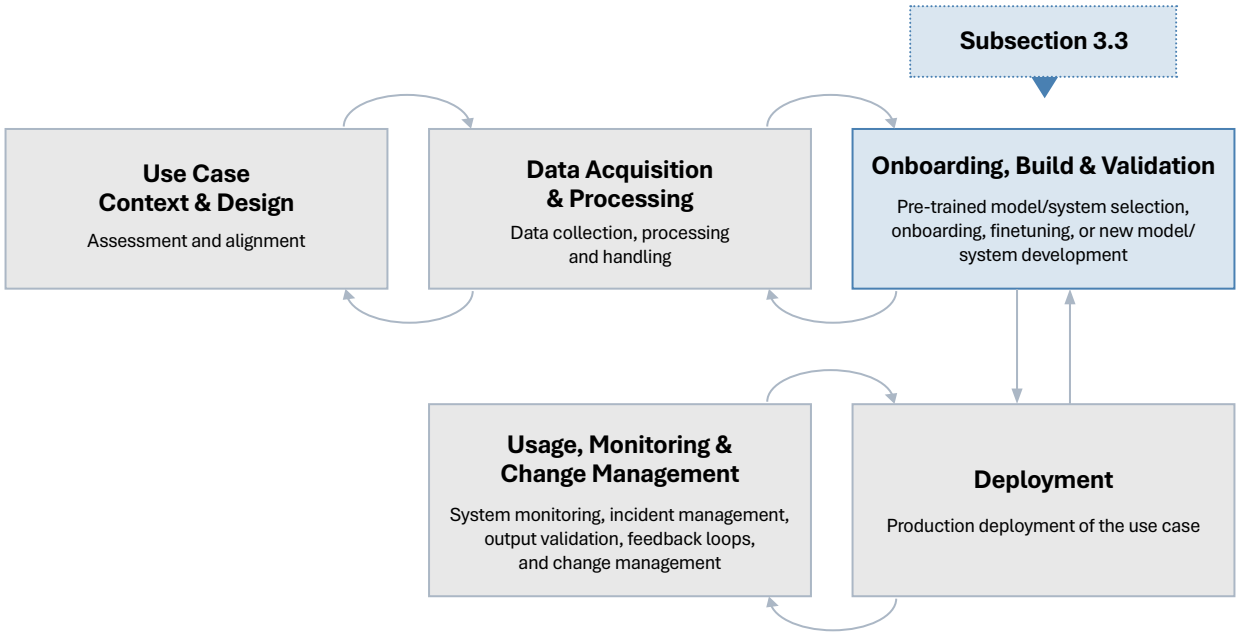
- **Data Security & Access:** To prevent data leakages, users were reminded not to enter information that may violate the bank's policies, standards and contractual obligations, such as secret data or data licensed to the bank by third parties. Data access controls were also implemented to minimise the risk of unauthorised access and misuse. Clear ownership of derived data, such as the index used for data retrieval and question-answering, was also maintained. This accountability is crucial for maintaining data integrity and governance.
- **Transparency & Traceability:** Documentation of metadata and data sources were maintained through established data onboarding processes. These documentation record key information such as data domain, ownership and source system, ensuring transparency and traceability.
- **Data Quality:** DBS-GPT's data quality was validated rigorously through performance evaluations using a golden set of questions and answers, ensuring that the data used was fit for its intended purpose. The use case team also worked closely with domain experts during performance testing to proactively mitigate unintended bias in the system's output, ensuring alignment with business needs.

3.3 Onboarding, Build, and Review

Governance and risk management activities at the build and onboarding stage ensure that AI use cases are built to function appropriately. They also include reviewing AI cases prior to deployment to confirm that they perform reliably and ethically in real-world scenarios.

This Subsection describes considerations that make up part of the AI lifecycle (see Figure 3.3.1). It is intended to be applied to each AI use case.

Figure 3.3.1: Onboarding, Build, and Review in the AI Lifecycle



This Subsection discusses three closely interlinked steps in the AI lifecycle:

- **Onboarding** involves vendor due diligence and pre-deployment reviews that ensure compliance and risk alignment.
- **Build** involves the development or customisation of an AI use case, as well as the implementation of guardrails. These are followed by testing completed by Builders: failures typically result in revisions before proceeding.
- **Review** involves an assessment of the AI use case by an autonomous party prior to deployment. The type of review depends on the materiality of the use case, ensuring an appropriate level of safety and compliance; successful review leads to Deployment (Subsection 3.4).

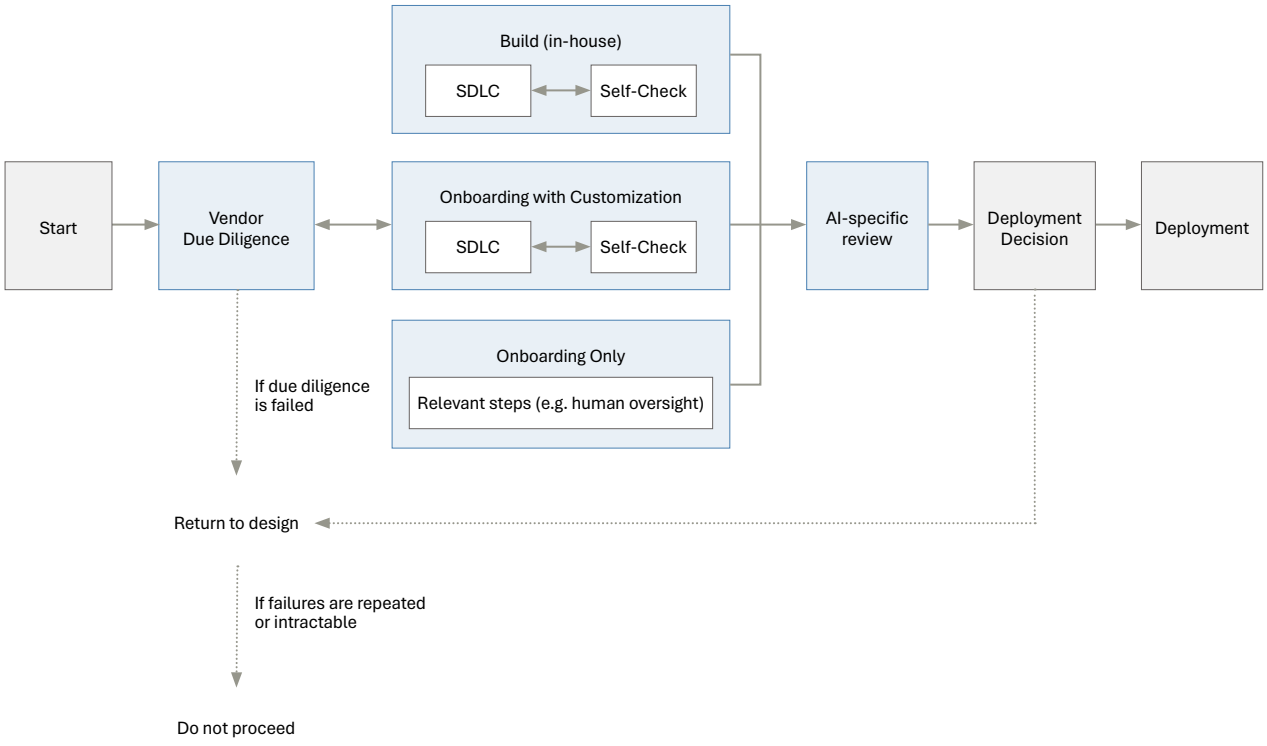
The relationship between these steps is diagrammed in Figure 3.3.2.

An AI use case's progression through these steps depends on its deployment pattern, of which three are discussed in this Handbook:

- **Build (in-house):** Developing an AI use case within an FI, including training the model and creating essential software components.
- **Onboarding Only:** The integration and/or use of an AI model or system from a third party without changing or adding components. This includes SaaS products.
- **Onboarding with Customisation:** The modification of an AI model or system from a third party, including model retraining, fine-tuning, or reinforcement learning, and the modification or development of components in or related to an AI system that contains models or components from third parties.

Each Practice below is tagged as being typically relevant to one or more of these deployment patterns. In general, FIs can manage AI risks by applying the same standard of risk management to all AI use cases, whether they are from third parties or developed in-house. The exceptions below relate to practices that manage risks unique to third party AI models or systems, or to practices that generally cannot be applied to third party AI use cases from a technical standpoint. The same standards of review remain relevant to AI use cases irrespective of deployment pattern.

Figure 3.3.2: Illustrative Relationship Between Onboarding, Build, and Review



The Considerations described below follow a typical SDLC; FIs can consider whether their existing SDLC may benefit from modification to address the unique risks of AI.

Consideration 10

Evaluate incremental AI-specific risks as part of the onboarding of third-party AI products and services within an AI use case.

Practice 1: Conduct relevant use case-specific relevant due diligence during third-party AI onboarding, in line with organisational standards, to manage the risks of a third-party AI product or service.

Applicable to deployment pattern: Onboarding with customisation, Onboarding only

Approach:

- Review third-party AI for alignment with use case-specific risk management requirements.
- Perform testing, as required, to assess the fitness and suitability of third party AI products or services in the FI's context.

When onboarding AI components, products, or services developed by third parties as part of a use case, FIs typically begin by adhering to existing procurement practices, both enterprise-specific and non-AI-specific, which have been established at the enterprise level to manage AI risks. These are discussed in detail in Subsection 2.3. In all cases, it is important to ensure that onboarding activities are conducted proportionately to the risk materiality of the intended use case; the use of third-party AI products or services could increase that use case's risk materiality.

Builders may work to ensure that third parties have taken appropriate risk management measures that are contextual to the use case and the identified risks. This can include ensuring that the third party has:

- Performed robust testing on appropriate AI-specific risk-related metrics that are relevant to the intended use case.
- Incorporated risk guardrails that are functioning adequately in the context of the intended use case.
- Established appropriate AI risk monitoring measures (where monitoring by a third party is applicable, such as in the case of a connected service).

When using synthetic data, FIs can assess and manage associated risks, including data representativeness, regulatory compliance, data provenance and model generalisation.

Additionally, Builders may ensure that appropriate use case-contextual third-party communication, such as communications on incidents, changes, and outages, has been planned for. This is important for mitigating risks associated with new third-party AI features or changes to existing ones.

Gen AI systems are particularly complex, probabilistic, and capable of exhibiting emergent behaviours that are difficult to predict. This creates a unique set of risks that traditional cybersecurity testing alone cannot fully address. Standard unit tests, integration tests, and even penetration tests might miss the nuanced, context-dependent vulnerabilities of Gen AI systems, especially in multi-turn interactions.

Many Gen AI applications involve direct user interaction. Red teaming and simulation testing help assess how users, both benign and malicious, might interact with the system and exploit its capabilities in unintended ways. This includes adversarial "jailbreaking" techniques where users try to bypass the model's intended safety boundaries. Simulation testing also makes it possible to assess how an FI accommodates cultural, linguistic, and age-related diversity in interactions.

Red teaming simulates real-world adversarial attacks before the Gen AI application is onboarded and deployed. This proactive approach allows FIs to discover vulnerabilities in a controlled environment, preventing potential harm, reputational damage, or financial losses that could occur if these weaknesses were exploited in production or supporting pre-onboarding due diligence of third-party AI products and services. Red teaming, particularly with human creativity involved, is designed to "think like the enemy" and uncover these less obvious flaws. Prompts identified via red teaming as causing inappropriate behaviours can lead to changes to grounding prompts or may simply be added to prompt blacklists which will be used by input filters to block similar attempts.

Where appropriate and where doing so is proportionate to risk materiality, FIs may also consider performing their own tests and checks when onboarding third party products and services, which are discussed in detail in Subsection 2.3. FIs may also consider other specific checks when onboarding complex third-party AI based on the AI's risk materiality and degree of interpretability. These checks can, where appropriate, cover data risks (such as data bias and leakages), performance risks (such as hallucinations), security risks (such as adversarial attacks and data leakages) and legal and compliance risks (such as IP violations).

To ensure that tests are effective and representative of the FI's intended use, FIs can consider – where feasible and proportionate to risk materiality – how best to use their own data as a reference. Third parties may, in good faith, optimise their AI products or services to perform well on known benchmarks, making representative context-specific tests more indicative of actual expected performance. Representativeness can also be improved by leveraging human feedback from domain experts, if doing so is justified by the use case's risk materiality. It is important to continue to consider privacy and data protection when defining these internal test datasets or sharing them with third parties.

Consideration 11

Ensure that the AI use case is built with appropriate guardrails and relevant metrics for effective performance and risk management.

Practice 1: Assess and select algorithms or features for the AI use case by considering its objectives and risks, including fairness, explainability, performance objectives, implementation complexity, and computational efficiency.

Applicable to deployment pattern: Build (in-house), Onboarding with customisation

Approach:

- Ensure that FEAT considerations are addressed by selecting algorithms and features that promote fairness and transparency, align with use case objectives, and are validated through domain expertise.
- In addition to FEAT requirements, evaluate algorithm choices on criteria such as performance objectives, overfitting, complexity, and computational efficiency, considering trade-offs appropriate to the AI use case.

When selecting algorithms and features for the AI use case, FIs may consider the alignment of each option with the use case's objectives, regulatory requirements, and risk considerations. FIs may consider embedding FEAT-related considerations by integrating fairness and explainability into the design of the AI use case, as the level of explainability and the methods used to assess fairness may vary based on the nature of the use cases.

For higher-risk materiality use cases, FIs may prioritise algorithms that offer inherent explainability or supplement them with explainability tools or techniques (such as SHAP and LIME) and interpretable features. These practices can support explainability as well as fairness, trust, and accountability: fairness considerations are often closely linked to explainability, as clearer insights into how features contribute to outcomes can help identify and mitigate potential biases.

As part of algorithm selection, FIs may assess model outcomes for fairness by evaluating each algorithm's performance across relevant subgroups using appropriate fairness metrics. This allows for a like-like comparison of fairness trade-offs between models. FIs may also conduct a domain review during algorithm selection to ensure that algorithm outputs are aligned with the business context and to identify patterns that may lack business relevance or could lead to unintended interpretations.

In Gen AI use cases where explainability can often be challenging, FIs can address explainability-related risks by ensuring traceability and facilitating the user verification of outputs. Examples of relevant techniques include citing sources – especially when using techniques like RAG that ground responses in relevant source documents. Other techniques include documenting data lineage to help uncover potential fairness issues stemming from data sources.

> Explainability in AI, particularly for Gen AI and Agentic AI systems, presents significant challenges due to the inherent complexity and sophisticated decision-making processes of these architectures. Explainability techniques like LIME or SHAP that highlight features in a prediction are helpful, but do not fully explain LLMs.
>
> Explainability in Agentic AI systems relies on two key pillars: observability and traceability. Observability provides real-time insights into an AI agent's internal working, behaviour, reasoning, and decision-making. This allows for a dynamic view of how the AI operates. Traceability, on the other hand, focuses on meticulously documenting and reconstructing the AI's lifecycle and internal states, encompassing data lineage, model evolution, and the detailed path of individual decisions and processes.
>
> Some novel approaches to explaining complex LLM outputs include tracing training data influences and linking outputs to verifiable sources, providing mechanistic interpretability at the "circuit" level, conducting behavioural and probing-based system-level analyses, and using simpler surrogate models that mimic complex models' behaviour.

In addition to FEAT considerations, it is important for FIs to consider other factors such as performance objectives, complexity, sustainability, and other trade-offs in light of the intended use case's risk materiality and objectives when selecting an algorithm.

Relevant considerations include:

- **Performance objectives:** The selected algorithm is typically expected to meet performance thresholds appropriate to the use case, such as accuracy. Improvements in performance may come with trade-offs, such as reduced transparency, increased model complexity, or fairness impacts. FIs can carefully quantify and document these trade-offs in algorithm selection.
- **Overfitting:** It is important to mitigate overfitting during model development. FIs can do so by favouring less complex models where appropriate, by assessing the selection of features to ensure a robust and representational distribution, and by testing for overfitting against representative real-world data.
- **Complexity:** It is important to weigh the complexity of the algorithm against the demands of the use case that it is designed to address. While complex algorithms can offer high performance, they often come with challenges related to transparency, interpretability, and computational resources. To capture the benefits of AI use, FIs can balance the need for sophisticated or less-understood approaches proportionately against operational constraints, regulatory requirements, and the necessity for clear audit trails.
- **Computational efficiency and cost:** When selecting an algorithm, a key consideration is the balance between computational efficiency and associated costs. The cost of using a Gen AI model is primarily determined by the number of tokens processed. In some cases, leveraging a more powerful LLM with a shorter prompt may be more cost-effective than using a smaller LLM with a longer prompt. FIs may also encounter complex trade-offs between the size of a Gen AI model and various fairness and security risks. To optimise for efficiency, FIs can leverage infrastructure such as hardware customised for AI efficiency, adopt techniques like mixture of experts or model distillation, and use fine-tuned or compressed models for targeted tasks. Efficiency can also be enhanced through in-memory caching during inference, sparse attention mechanisms to reduce memory load, and hybrid setups using large models offline and smaller ones for real-time tasks. Efficiency is best managed by right-sizing the choice of model to the intended task.

FIs may require Builders to justify and document their selection of algorithms or features, especially when selecting more complex algorithms or less understood features. It is important for the selected algorithm or features to balance performance against fairness, explainability, complexity, data availability, and use case needs, and be supported by theory, research, or industry practice. When selecting newer or less understood algorithms, such as complex Gen AI models, it is important that FIs weigh potential benefits against heightened risks while also assessing its capabilities to manage these risks. Builders may do so by consulting experts, users, or other stakeholders to support algorithm selection.

Practice 2: Identify and implement appropriate guardrails and controls during the development of AI use cases proportionately to the level and nature of the associated risks, to effectively manage and mitigate potential risks.

Applicable to all deployment patterns.

> **Approach:**
>
> - Apply proportionate guardrails and controls to mitigate the risks associated with the AI use case.
>
> FIs can mitigate the risks that they identified for the AI use case in a fashion proportionate to its risk materiality by applying appropriate guardrails and controls.
>
> For more details on the appropriate guardrails and mitigation strategies applicable to addressing risks in each of traditional AI, Gen AI, and Agentic AI, FIs can refer to existing industry documents, such as those documented in Appendix C of this Handbook, and the guardrails proposed in Appendix G of this Handbook.

Practice 3: Define use case-specific risk-related metrics for assessing the AI use case for risks.

Applicable to all deployment patterns.

> **Approach:**
>
> - Select use case-specific performance metrics that enable the FI to assess effectiveness, quality, and risk associated with the AI use case.
>
> To ensure that AI use cases function effectively and responsibly, FIs typically define and track use case-specific performance metrics tailored to the business expectations and risks associated with the use case. These metrics help assess the effectiveness of the use case's risk management as well as its behaviour in real-world conditions.
>
> The selection of metrics may vary significantly between traditional AI and Gen AI. Unlike traditional AI, which relies on structured data and objective evaluation metrics, Gen AI systems can produce outputs like unstructured text and image data that are hard to quantify, making performance assessment more complex. In addition to accuracy, Builders can consider dimensions such as evaluating the quality, coherence, fairness/biasedness, security, and ethical implications of generated content where applicable. A further list of metrics and benchmarks for AI, including some metrics specific to Gen AI, is included in Appendix F.
>
> FIs can improve the effectiveness of their overall portfolio-level AI risk management by using, where possible, more consistent metrics for use cases with similar modalities (such as text-based Gen AI, image-based Gen AI, or traditional AI) and objectives. Comparable AI risk-related performance metrics can improve the robustness of AI-specific risk tracking, the depth of risk awareness, and the comparability of risks across use cases.
>
> Performance evaluation for traditional AI is centred on key areas such as accuracy (precision, recall, F1-score), fairness and bias (demographic parity, disparate impact ratio), robustness and stability (Characteristic Stability Index, Population Stability Index), and transparency (SHAP, or Shapley Additive Explanations; LIME, or Local Interpretable Model-agnostic Explanations).

Effective risk-related performance metrics for Gen AI include those for fairness and bias (Toxicity Score, Statistical Parity Difference), robustness and stability (input feature data drift, output prediction data drift, factual accuracy/groundedness, topic and task adherence), and cyber and data security (Prompt Sanitisation Rate, Total Injection Vulnerability Score).

Practice 4: Evaluate and calibrate transparency measures based on the use case's risk materiality, degree of autonomy, and intended users, implementing proportionate design features and disclosures to support responsible and informed use.

Applicable to deployment pattern: Build (in-house), Onboarding with customisation.

Approach:

- Implement transparency measures proportionate to the AI use case's risk materiality, level of autonomy, and user type to ensure clarity in how AI decisions are made and communicated.
- Adopt an appropriate degree of AI-specific disclosure to end users and customers.

Transparency is one of the key principles in the responsible use of AI. Transparency can be promoted through appropriate disclosures and explanations on the use of AI, in particular to those parties who might be impacted by the outputs, decisions, or recommendations of the AI use case. One approach to determining the appropriate level of transparency is by establishing it proportionately to the use case's degree of autonomy and degree of risk materiality. Transparency measures can include:

- **Human-directed use cases:** If an employee is responsible for making decisions that cause impacts and the role of the AI use case is only to provide inputs or recommendations, AI-specific disclosures may not be necessary. Transparency efforts can instead focus on supporting internal decision-making, such as by appropriately communicating key factors in decisions and the degree of uncertainty associated with them, rather than external AI disclosures.
- **Highly Autonomous Use Cases:** If AI-generated outputs are to be used without a human in the loop, AI-specific disclosures become increasingly important to ensure that impacted parties and users understand how the system functions and makes decisions. This also supports the exercise of their right to recourse.

In cases where FIs choose to provide customer-facing explanations, these explanations may be most useful when they focus on the key factors influencing the decision rather than the underlying modelling process or technique, which may not be understandable to a layperson. This may include a description of the data used by the system (such as income or credit history) and how it impacts the outcome (such as a lack of credit history negatively affecting an applicant's creditworthiness). The explanations are most effective when they are clear, concise, and tailored to the intended audience, avoiding technical jargon.

FIs can also consider sharing clear policies on explanation, appeal and recourse with customers or data subjects.

For Gen AI use cases, FIs can consider displaying the sources used to generate the output, as well as the fact that the user is interacting with a Gen AI, where doing so would improve user trust and interpretability and proportionate to risk materiality.

Figure 3.3.3: Illustration of AI Specific Disclosure & User Transparency Messages

AI-Generated Investment Advisor

What are some financial products I can consider investing in to support myself in retirement?

1 System searching...

Your age

Your income

Your account's financial goals and risk tolerance

Results of your financial awareness questionnaire

You have 23 years left until retirement, which is a long time horizon, and you told us in your onboarding that you are an experienced investor.

We typically suggest that customers with your profile invest in our Unit Trust Risk Level 1, which balances equity and fixed-income products.

2

3 This response is generated by AI. It is based on the information you provided to us, but AI-generated responses sometimes contain inaccuracies. Check any information that AI provides and if in doubt, consult a financial advisor.

AI-Generated Mortgage Rate Simulation

1 **You told us:**

- Your profession and number of dependents
- Your household income and assets/ liabilities
- The price of the home you would like to buy and the LTV ratio you are targeting

We also consider:

- General market conditions
- The SORA interest rate benchmark

Your estimated rate:	Your estimated monthly payment:
2.5%	$2,200

4

You don't have much debt, which lets us offer a lower rate.

Your LTV is higher than average, which increases your rate and monthly payment.

3 We use AI to estimate your rate and monthly payment. This is not a final decision; we will manually review your account before the rate is finalised. If you think there has been an error or you want more information, speak to a mortgage advisor.

1 Indicate to the user the kinds of information that the system considers

2 Use visual cues to remind the user that they are interacting with AI.

3 Remind users to check AI-generated information and give them an option for recourse

4 Explain what features were important in the decision, and suggest what users can do to receive a more favourable result.

Practice 5: Document key aspects of the AI build process, including data handling, model training and selection, and evaluation decisions to enable auditability and reproducibility.

Applicable to all deployment patterns.

Approach:

- Capture relevant documentation related to AI build activities and evaluations to support pre-deployment reviews and post-deployment governance.

The auditability and reproducibility of the AI build process is essential for ensuring accountability and building trust in AI use cases. Achieving full reproducibility can be challenging, particularly when use cases include third-party AI products or services where information about the underlying model may be unavailable. In such cases, FIs can focus on capturing as much relevant information (such as metrics evaluated, types of tests performed, documentation of model behaviour, key assumptions or limitations) as possible within the use case's constraints. Documentation can be retained according to non-AI-specific standards and regulatory expectations.

FIs may consider using standardised templates, where appropriate, to capture key details of the AI build process. It is important that FIs document details about the data used (where feasible), such as its sources, processing steps, quality checks, and how it was split into training, testing, and validation sets. Documenting the model training process – such as code, software environments, hyperparameters, configurations, and key settings like random seed values – can further support efforts to reproduce outcomes.

Capturing information about the model selection process, including evaluation methods, thresholds, decision rationales, comparisons of performance across multiple AI models, and justifications for selecting the final model may also be beneficial. Capturing efforts around explainability, feature selection, and fairness assessments, along with their metrics and outcomes, can help provide clarity around key decisions. FIs can also document relevant risks, assumptions, and limitations. This comprehensive documentation will be used for assessment during AI-specific review processes and for other audit and oversight purposes. For Gen AI systems, FIs may document information such as prompt versioning, inference parameters, and model versions.

Where third party AI components are used as part of the use case, FIs can consider documenting relevant information on the provider, evaluations conducted, and known risks. A full discussion of information that can be relevant in managing third party AI risk is included in Subsection 2.3.

> To learn more about the guiding principles for effective AI documentation, refer to the CLeAR Documentation Framework for AI Transparency[8]

Consideration 12

Conduct thorough testing and review prior to deployment to assess AI-specific risks and ensure that appropriate guardrails, controls, and governance have been observed.

Practice 1: Ensure that Builders conduct appropriate AI risk self-checks during development to test use case performance, verify the effectiveness of risk management activities, and identify and mitigate issues early in the development process.

Applicable to deployment pattern: Build (in-house), Onboarding with customisation

Approach:

- Ensure that builders perform self-checks to validate performance on AI risk-related metrics and the effectiveness of guardrails prior to deployment.
- Ensure that the results of self-checks are appropriately documented.

FIs already have practices in place to ensure that Builders check their own work for risks prior to deployment. Self-checks for AI-specific risk and guardrail effectiveness are supplemental to, and do not replace, existing testing practices in an FI's SDLC.

As part of an AI risk self-check, builders can define clear objectives, design targeted test cases to assess the effectiveness of guardrails, and for Gen AI use cases, consider using simulation or adversarial testing to evaluate appropriate behaviours. Self-checks can also include testing the performance of the AI use case against its AI risk-related metrics and benchmarks, in addition to other evaluations such as fairness assessment, sensitivity analysis[15], sub-population analysis[16], error analysis[17] and stress testing[18]. It is important in these situational tests to use data that is as specific as possible to the FI and its intended use case, which can ensure that the results of tests are accurate in the FI's context. For use cases including Gen AI or Agentic AI, FIs can consider testing specifically for their failure modes, such as data leakage, toxicity, or hallucinations.

Self-checks may also include process checks, such as verifying that appropriate risk management activities were implemented in accordance with the FI's policies and in line with existing good practice. Self-checks may also be extended to include non-AI-specific attributes, such as performance and scalability; FIs may consider integrating AI-specific self-checks into existing pre-deployment checklists.

These activities are crucial to mitigating AI risks and are not replaced by the AI-specific review described below. Builders, by assessing use cases that they have been involved in developing or deploying, can be particularly well-suited to identifying and addressing issues. FIs can make appropriate self-checks by Builders or Use Case Owners part of their business as usual for AI governance and risk management, such as by ensuring that the results of AI risk self-checks are documented as a precondition for deployment.

Fairness is particularly important to assess as part of self-checks. Doing so may include defining relevant protected attributes – like race, gender, or age – or proxies thereof that are relevant to the use case in question and testing both the use case itself and its underlying data, is applicable, for unfair treatment along those attributes. Where the FI does not have access to underlying training data, they may nonetheless identify the potential for discrimination through simulation testing of the use case. A selection of relevant metrics is documented in Appendix F.

Practice 2: Conduct an AI-specific review based on use case risk materiality prior to deployment to ensure that potential risks are identified and mitigated.

Applicable to all deployment patterns.

Approach:

- Ensure that AI-specific review is conducted in a manner proportionate to use case risk materiality by a party not directly involved in the development, deployment, or operation of the use case.

[15] Sensitivity analysis assesses how the predictions or outputs of AI models change under different permutations of data inputs. This also helps to identify important features that significantly influence predictions or outputs, facilitating explanations of the behaviour of AI models.
[16] Sub-population analysis is an evaluation of how AI models perform across different sub-populations or subsets within the datasets (such as between different customer segments).
[17] Error analysis identifies potential patterns in prediction errors, which helps to understand the limitations of AI models.
[18] Stress testing assesses the response of AI models to edge cases or inputs outside the typical range of values used in training.

Upon completion of use case build, onboarding, and self-checks by builders, FIs can conduct an AI-specific review to assess residual AI-specific risks and confirm the appropriate application of AI governance and risk management process. AI-specific reviews are discussed in Subsection 2.4, and can support, supplement, or be integrated with other non-AI-specific deployment checklists and with Builders' self-checks. Non-AI-specific deployment checklists and reviews that can interface with and support the AI-specific review include, in particular, reviews for privacy and cybersecurity.

This review is most effective when carried out by a party not directly involved in the development, deployment, or operation of the use case; it is also more effective when completed prior to deployment, which ensures that potential risks are mitigated before they can impact people or production data.

It is important to ensure that onboarded AI use cases containing third party components are subjected to review processes that are equivalently stringent to those applied to in-house developed use cases, proportionate to their risk materiality. FIs may consider obtaining sufficient information and assurances from the third-party providers and, where necessary, perform robust compensatory testing as part of the review process.

Sharing review findings with the relevant approval body within the FI is crucial. This includes any identified limitations, recommended remediation actions, or conditions for use. This ensures that key issues are addressed appropriately before the use case is approved for deployment.

Illustration 3.3.1: Principle-Based AI Risk Oversight at Julius Baer

Julius Bär

Julius Baer has established six guiding principles for the responsible use of AI across the organisation. These principles, which align with industry best practices and our corporate values, were developed after a thorough review of regulations applicable in Julius Baer's operating locations.

Our Responsible Artificial Intelligence Council (RAIC) has translated these principles into AI-related risks and included these into its existing risk categorisation. The RAIC, Julius Baer's cross-functional oversight committee for AI use cases, plays a key role in this process, bringing together use case owners and risk specialists to evaluate and mitigate potential risks. In regular RAIC meetings, use case owners present their proposals, and risk specialists assess the use case for their respective risk types and then either accept, request mitigation measures for, or reject these identified risks. This rigorous evaluation occurs before use cases are deployed to the production stage, ensuring alignment with Julius Baer's risk appetite.

Use cases are continuously monitored throughout their lifecycle. Additionally, use cases implemented prior to the establishment of the RAIC will undergo re-validation to ensure compliance with current standards. For use cases involving third-party technologies, a dedicated flag has been integrated into the outsourcing process, ensuring early evaluation of potential AI-related risks to identify possible red flags during the outsourcing evaluations.

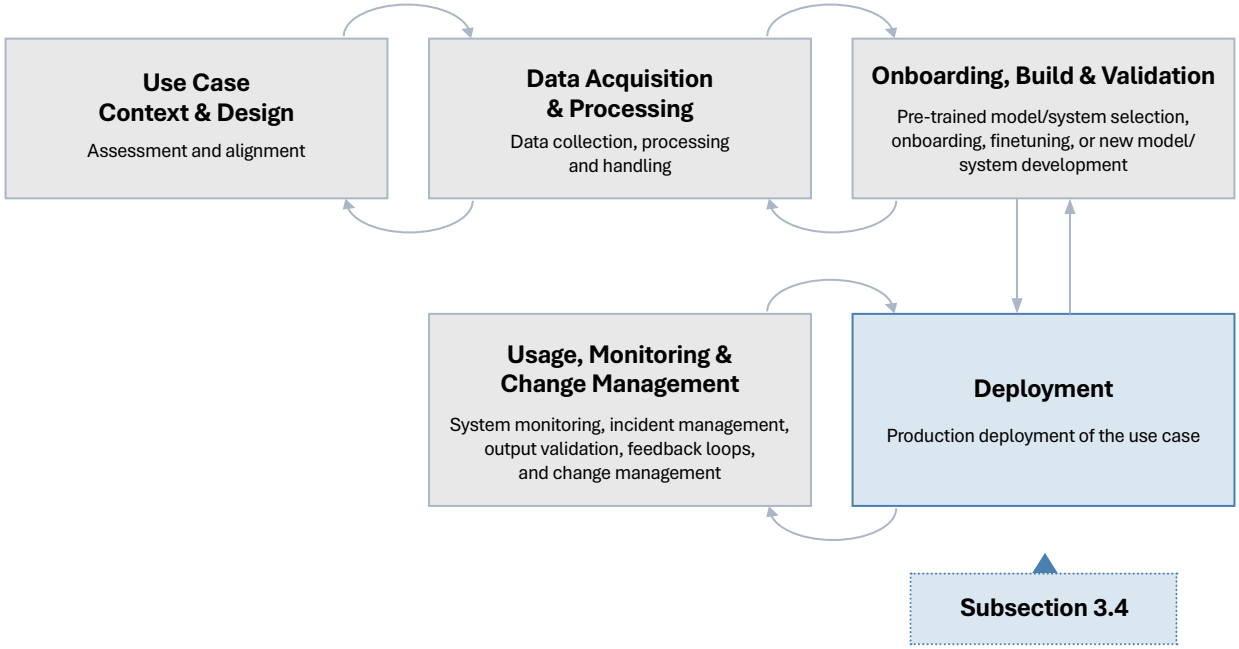
3.4 Deployment

Deployment is the process of putting an AI use case into its production environment. Deploying an AI use case involves a structured approach, sometimes in the form of a checklist, which ensures that they will consistently function as intended, align with business objectives, manage AI-specific risks, and comply with regulatory and ethical standards. Deployment activities also involve preparing the AI use case and its team for ongoing post-deployment activities, which are discussed in Subsection 3.5.

FIs can apply deployment practices, insofar as they are proportionate to risk and technically feasible, to both in-house developed and third-party onboarded AI use cases. This is in line with the overall approach that MindForge takes to third party AI products and services – that they can be held to the same standard, wherever possible, as those developed in-house. Where relevant, FIs may leverage their existing SDLC practices in doing so.

This Subsection describes considerations that make up part of the AI lifecycle (see Figure 3.4.1). It is intended to be applied to each AI use case.

Figure 3.4.1: Deployment in the AI Lifecycle



Consideration 13

Develop monitoring and contingency plans for the use case prior to its deployment, and consider risk-informed deployment options.

Practice 1: In conjunction with other monitoring activities, ensure that a monitoring plan and safeguards/ contingency measures are in place, along with the designation of an appropriate accountable person to address AI risks detected in monitoring.

Approach:

- Define a post-deployment monitoring plan performance metrics and thresholds to be tracked as part of the monitoring process, with monitoring frequency calibrated to the use case's risk materiality.
- Establish safeguards and contingency measures to address breaches or anomalies identified during monitoring.
- Designate accountabilities to individuals and teams to carry out post-deployment monitoring activities.

Before deploying an AI use case into production, FIs can ensure that risks will be thoroughly managed across the use case's post-deployment life by establishing a comprehensive AI risk monitoring plan. FIs can use these plans to document the key AI risk-related performance metrics and thresholds to be tracked, which are discussed in Subsection 3.1. The monitoring plan specifies how these metrics and their thresholds will be tracked, the frequency of monitoring, and the escalation process if thresholds are exceeded. The frequency of monitoring is usually set use case-by-use case based on factors like risk materiality, ensuring that higher-risk use cases receive closer scrutiny.

In addition to tracking AI performance against thresholds, for use cases with higher risk materiality FIs can consider tracking near misses, early warnings, and other cases where metrics approach thresholds but do not surpass them. Setting lower thresholds for activating "near miss" indications can ensure that monitoring teams are prepared to identify and rectify issues before they become serious.

To address cases where AI use cases breach their risk thresholds or otherwise exhibit undesirable behaviour, FIs can define safeguards and contingency measures, such as rollback options to revert to a previous, stable version of the model or a kill switch to deactivate the use case, ensuring that further harm will not take place while an issue is investigated. FIs can consider testing these measures prior to deployment. FIs already have incident management frameworks in place as part of their existing technology risk management; they can ensure that AI incidents related to the use case will be effectively managed by it and that AI-specific procedures are in place where relevant. Where the use case contains AI products or services provided by third parties, FIs can ensure that they apply existing contingency planning measures for third party software, such as by proportionately planning for service interruptions, performance degradations, or support discontinuation.

An important component of a monitoring plan is the assignment of accountability and operational ownership. Responsibilities for key post-deployment monitoring processes and the implementation of mitigations are most effective when they are clear and consistent; FIs can consider using standard control objects to define the individuals and teams responsible for executing the post-deployment plan. FIs can also ensure that teams assigned to conduct monitoring activities for the use case have the necessary skills and knowledge to do so; they can take into account the specific technologies used and business areas implicated to make this determination. FIs can also ensure that employees or functions involved in monitoring the use case have sufficient authority and capabilities to request relevant information and to execute the contingencies defined in the monitoring plan.

FIs could consider incorporating risk and incident monitoring and response into existing IT Service Management (ITSM) processes for Incident and Problem Management. Where this is done, it is important that the existing incident management team be trained in AI-specific monitoring and response skills. FIs with processes in place or requirements to notify customers of incidents may include AI-related incidents in those notification processes.

Practice 2: Consider the need for a phased rollout to manage the AI use case's risks and progressively validate the use case's performance prior to full deployment.

Approach:

- Implement, where doing so is proportionate to use case risk materiality, a phased rollout to limit the potential impact of AI risks, incorporating additional AI governance and risk management review stages to ensure observed risks remain within the FI's risk appetite before full deployment.
- Ensure that phased rollouts are implemented with clear guardrails and limitations to manage their potential risks.

FIs can mitigate unforeseen risks associated with an AI use case by considering, based on use case risk materiality, whether a phased rollout is appropriate. Phased rollouts help to identify system limitations and train users on interacting effectively with AI tools. AI use cases may have a unique need for phased rollouts due to their probabilistic outputs.

Phased rollouts take many forms but broadly use progressive exposure to limit the potential impact of AI risks before full production. Because phased rollouts involve real users and production data, it is important that FIs consider the risks that a use case could pose when in a small pilot or phased deployment, and that they take appropriate risk management measures. While it is important that a degree of risk identification, control, and approval takes place before a phased rollout begins, FIs may for practical reasons choose to defer certain activities – like the AI-specific review or the final residual risk assessment – until after the initial phase of the rollout is complete and feedback from it has been incorporated into the use case and its guardrails.

> Phased rollouts are particularly valuable for managing risk in Gen AI use cases due to their complexity and potential for unpredictability. It allows FIs to mitigate risks (such as hallucinations or bias) by testing in a controlled environment, gathering real-world feedback for iterative improvements, validating performance and scalability under actual load, and managing change and user adoption effectively. This strategic approach ensures responsible innovation, optimising costs and refining governance before full-scale deployment, and leads to safer and more impactful Gen AI applications.

This deviation from the "standard" lifecycle comes with risks, and it is important that FIs carefully follow their IT guidelines on managing the risks of pilots. This may include strictly limiting the scope, features, and number of users of the use case during its phased rollout, limiting the duration of each phase, and ensuring that additional monitoring is in place. Clear success and failure criteria (the latter of which could trigger deactivation) for each phase can help reduce risk. FIs can consider factors like the use case's risk materiality when determining whether, and how, to manage risk when deviating from standard AI and IT governance procedures during a phased rollout. Rather than deferring governance steps, FIs could also choose to add additional review stages during or after a phased deployment.

There are several options for implementing phased rollouts. Pilots help capture real-world data on a limited scale or scope to evaluate model performance, allowing FIs to systematically collect and analyse feedback from users, measure outputs against predefined quality and risk metrics, and assess unintended consequences. Parallel testing also provides an opportunity to compare AI-generated outputs against human or system benchmarks, helping FIs to evaluate AI effectiveness while identifying areas requiring further risk mitigations. Other common types of phased rollouts for AI include A/B deployments and canary-type deployments.

Practice 3: Engage and equip users with targeted training and use case-specific resources to support responsible use and effective oversight.

Approach:

- Conduct training for the intended users of the AI use case to raise awareness of AI-specific risks and user responsibilities, while equipping them with the necessary skills, resources, and knowledge.

Appropriate usage is an essential guardrail against risk and is applicable to all AI types. Before deployment, FIs can ensure that they are prepared to mitigate post-deployment risks by ensuring clear communication and proper training for users of AI use cases. These users may, depending on the use case, be internal or external. AI introduces several unique considerations requiring additional efforts:

- **Raising Awareness of AI-Specific Risks and User Responsibilities:** AI's probabilistic nature introduces uncertainties and has limitations; users can most effectively mitigate the risks of AI if they understand these characteristics and know what is expected of them to mitigate those risks. Internal users acting as "humans in the loop" reviewing AI-generated content or recommendations can receive specific training on their oversight responsibilities. For example, Gen AI models are known to generate plausible but incorrect or misleading outputs; an effective human reviewer is trained to identify inconsistencies and verify outputs against trusted sources.
- **Ensuring Users Have the Necessary Skills, Resources, and Knowledge:** Users can be empowered to manage post-deployment AI risks for a use case with sufficient awareness of its limitations, resources for enabling its use, and use case-specific education on how to use it responsibly. Employees operating a Gen AI use case, for example, may mitigate usage risks by having access to a prompt library and training on how to use it. Users also benefit from knowing how to identify and adhere to a use case's intended purposes. For example, users of a Gen AI tool for summarisation can best limit its risks when they know that it will not be reliable when used for Q&A. Where the use case employs AI products or services from third parties, this knowledge can include an awareness of the division of responsibilities for the use case's performance and an understanding of important information from the provider.

An appropriate awareness of risks and responsibilities can support FIs in mitigating the risk of over-reliance on a use case by its employees.

General AI risk literacy withing the FI, which is not specific to a single use case, is discussed in Subsection 4.1.

Practice 4: Ensure that the AI use case is appropriately documented, that appropriate security and governance practices are applied, that relevant data retention is provided for, and that relevant approvals are obtained before deploying to production.

Approach:

- Apply and document existing non-AI-specific deployment practices, such as pre-deployment checklists.
- Document details about the AI use case, including detailed results and findings from AI-specific review, identified issues and mitigation measures implemented.
- Provide internal governance bodies with the AI use case's documentation and supporting materials, such as confirmations of checks in technology, data, legal and compliance, and third-party/outsourcing areas, to support pre-deployment decisions.

FIs already have effective pre-deployment processes in place in line with industry norms. These are often defined in terms of a deployment checklist which sets out steps or tests to be completed before a system or application is deployed. These continue to be relevant to ensuring that AI use cases are deployed in a manner that is sufficiently robust, secure, and functional. Among the relevant existing controls that FIs can adopt are those that ensure effective access management, security, logging, and monitoring are in place; this may include industry-standard IT practices like the principle of least privilege. As a pre-deployment activity, FIs may leverage existing IT review practices to test that the AI use case's architecture, access, security and encryption measures, backups, and other controls are functional and correctly configured.

FIs can also document risk-related information about their AI use cases in addition to their existing IT documentation practices. This can involve recording the detailed results and findings from AI-specific reviews; identified issues, such as potential risks; and the corresponding mitigation measures or adjustments made to ensure that the AI use case complies with operational and regulatory standards. Other relevant documentation on risks and limitations includes the outcomes of User Acceptance Testing (UAT) and internal audits. As part of the documentation exercise, FIs may also ensure that any relevant information not already present in the AI inventory is captured prior to deployment. AI inventory practices are discussed in more detail in Subsection 2.5. Once the assessment results are compiled, FIs can ensure that this documentation is submitted to the relevant internal governance bodies.

The appropriate retention of key information and documentation on the use case is important for ensuring that it is auditable, trustworthy, traceable, and reproducible.

Illustration 3.4.1: Julius Baer's Tiered Approach to AI Literacy and Risk Awareness

Julius Bär

Julius Baer is taking a multi-faceted approach to educating employees about AI opportunities and risks. To establish a common foundation across the firm, a mandatory e-learning for all staff was introduced, which focuses on key AI and Gen AI concepts, capabilities, and limitations, Julius Baer Responsible AI principles, permissible use of AI based on regulations, and tips for effective, safe, and ethical use of AI within Julius Baer.

Additionally, the product owners offer tailored training sessions for their specific AI and Gen AI use cases, ensuring that users grasp the capabilities, limitations, and intended purpose of each solution. The scope of these trainings is also predicated on factors like use case impact, risk, and user maturity.

To further support employee development, the Julius Baer Academy curates relevant training resources on AI related topics and hosts internal and external expert-led sessions to foster key AI skills and knowledge. They also support individual business units with creating customised learning packages, leveraging internal expertise, or external content from platforms like LinkedIn Learning or Coursera.

Lastly, we promote general AI awareness across the organisation through various channels, including presentations within existing Communities of Practices (tailored to the audience, from expert Data Science groups to individuals with a general interest in the topic), intranet updates, and leadership discussions, including with Executive Board Members. By doing so, we aim to keep employees informed and up-to-date on the latest AI developments, their relevance to the Julius Baer, and the potential risks.

3.5 Usage, Monitoring, and Change Management

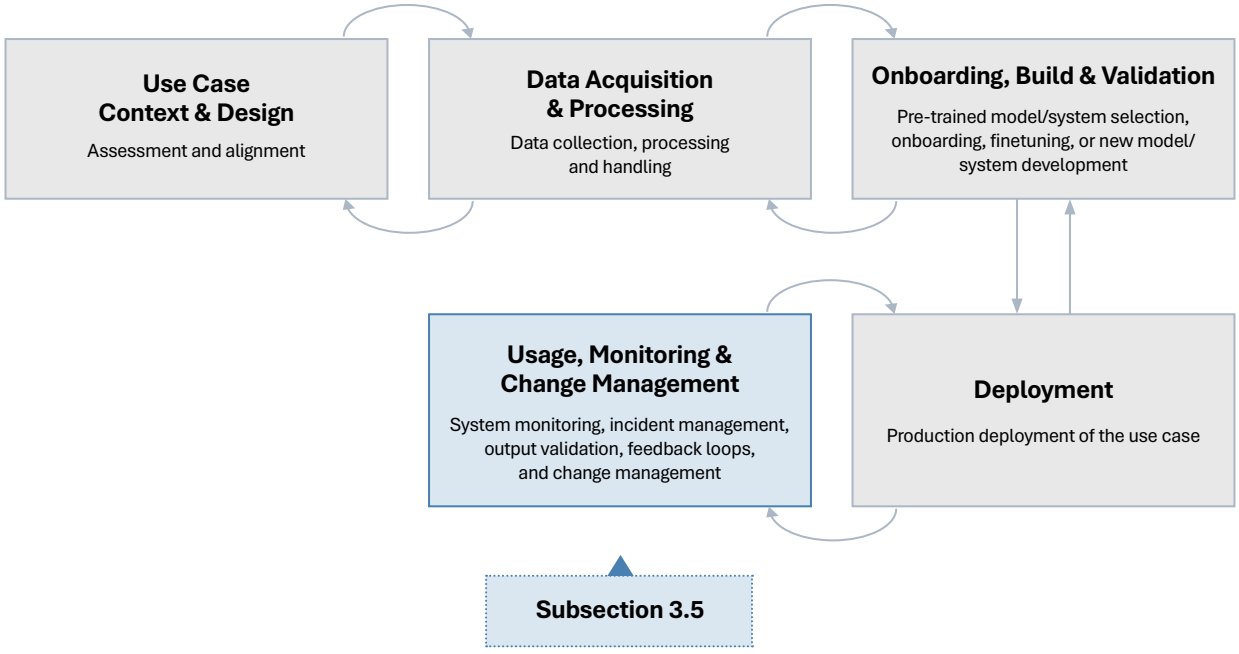
Effective AI governance and risk management includes ensuring that AI use cases are used in accordance with the FI's principles – fairly, ethically, accountably, and transparently, for instance – that risks that occur during usage are managed, and that regulations are complied with. Where relevant, FIs may leverage their existing SDLC practices in doing so.

Monitoring AI use cases involves designating accountable parties and reporting on the use case's AI risk-related performance metrics, especially when those metrics breach their thresholds. FIs can supplement monitoring by conducting post-deployment reviews of their use cases, sometimes including adversarial testing, at a proportionate frequency to identify vulnerabilities or shifts in risk exposure, with a particular focus on change over time. Managing AI risks as they arise in the post-deployment lifecycle – both those that were identified during a use case's risk assessment and those that were not – requires robust, ongoing monitoring at a reasonable frequency. This may include leveraging issue or incident management practices to ensure that clear escalation paths are in place and that concerns are surfaced and addressed appropriately.

These practices, insofar as they are proportionate to risk and technically feasible, are most effective when applied without distinction to both in-house and onboarded AI use cases. This is in line with the overall approach that MindForge takes to third party AI products and services: that they be held to the same standard, wherever possible, as those developed in-house.

This Subsection describes considerations that make up part of the AI lifecycle (see Figure 3.5.1). It is intended to be applied to each AI use case.

Figure 3.5.1: Usage, Monitoring, and Change Management in the AI Lifecycle



Consideration 14

Conduct ongoing monitoring of the AI use case and its usage to ensure that it remains fit for purpose over time.

Practice 1: Periodically monitor and report on use case metrics related to AI risks, guardrail effectiveness, and changes in the use case's operating environment, as necessary and at a proportionate intensity and frequency, and address any issues identified.

Approach:

- Ensure the correct implementation of the use case's monitoring plan as defined prior to deployment.
- Follow the incident and problem management plan defined for the use case to manage issues, and implement response mechanisms (such as roll-backs to previous versions or kill switches) when necessary.

The monitoring plan defined for each use case (see Subsection 3.4) is only effective if it is correctly implemented after deployment. FIs may monitor both performance and risk-related metrics, as well as the effectiveness of implemented guardrails in mitigating identified AI risks. This applies equally to AI developed in house and those onboarded from third parties. Each metric is typically tracked against a threshold, beyond which it is considered to be outside the FI's risk tolerance. In addition to risks, FIs may also track AI-related metrics that provide insight into the use case's uptake, cost, and return on investment.

The choice of metrics and thresholds, as well as the monitoring frequency and calculation methodology for each, benefits from a standardised approach across the FI. This may be defined in an FI's evaluation framework, which is discussed in more detail in Subsection 2.4. Customisation may be required in some situations – such as where they are impossible to calculate for a given technology, or where they may simply not be pertinent to the use case's unique characteristics.

> A common technique for visualising the results of ongoing monitoring of AI, especially with executive audiences, is a "red/amber/green" dashboard. This "traffic light" approach can straightforwardly communicate when a use case has had a breach, when it is near to breaching a threshold (a near miss), and when it is performing within tolerances.

When incidents or issues are identified, such as when the AI use case breaches the thresholds of a risk-related metric (such as the average toxicity score of the chatbot exceeding its predefined threshold), when a guardrail fails to function as intended (such as when a content filter fails to prevent the generation of sensitive outputs), or when concerns are raised in user feedback (such as a Gen AI image generator used in marketing emails being flagged by the Business User for generating racially insensitive imagery), FIs may review the cause of the incident or issue and take action to address the matter promptly. FIs may also consider implementing response mechanisms, such as model retraining, adjustments, redevelopment, rolling back to a previous version of the model, or activating a kill switch.

To proactively manage risks, especially for use cases with a higher risk materiality, FIs may also identify "near misses" where an incident was avoided. Doing so can facilitate the pre-emptive mitigation of issues.

It is important to thoroughly log the results of monitoring and to track both incidents and escalations for the use case. Effective logging supports root cause analysis and remediation; it also enables post-hoc auditability. FIs typically have existing practices in place for logging issues and metrics for IT systems and can continue to leverage those for AI. Depending on the use case, FIs may also consider logging additional information, such as AI inputs and outputs, for the purpose of monitoring and auditability.

Finally, appropriate monitoring, tracking, and results logging can support the assessment and improvement, as required, of metrics and thresholds themselves. The interpretability and actionability of past results, as well as the results of past performance, can inform adjustments to calculation methodologies, threshold and alert levels, or changes to the selection of metrics altogether. Such changes may be recommended as part of post-deployment AI-specific review.

Practice 2: Monitor and report on the quality, drift, and third-party risks associated with the use case's input and training data in an ongoing fashion, as necessary, after deployment.

Approach:

- Monitor data quality, drift, and risk indicators at a frequency and depth proportionate to the use case's risk materiality.
- Implement mitigation actions (such as retraining or adjustments) when indicators fall outside their acceptable thresholds.
- Monitor third-party datasets employed by the use case, as necessary, for key risk-related characteristics, including consent validity, usage requirements, and statistical properties.

Monitoring the quality, drift, and risks posed by internal and third-party datasets – including both training data and data accessed in operation like grounding prompts or documents in a RAG architecture – is an important part of maintaining an AI use case. FIs can conduct this monitoring at a frequency and depth proportionate to the risk materiality of the use case. Where any of these indicators fall outside thresholds set based on the FI's risk appetite, appropriate re-training or other mitigation steps can be taken.

Indicators for dataset health can include metrics on accuracy, completeness, and anomaly rates. FIs can also consider establishing ongoing statistical monitoring for data drift indicators. Where third-party data is used and where FIs have the capability to do so, FIs can consider ongoing monitoring of that dataset's key risk-related characteristics, such as its data quality, relevance and accuracy over time, consent, validity and usage requirements, other predefined KPIs, and statistical properties. This can ensure that FIs identify risks related to those third-party relationships promptly.

In addition to monitoring datasets for AI-specific risks like drift, FIs can continue to apply existing data management practices, where it is relevant and proportionate to do so. This includes monitoring access and ongoing security to datasets, environments, and pipelines related to the AI use case; FIs can draw on standard data lifecycle management practices, such as timely deletion, to support these activities.

Practice 3: Conduct periodic checks for changes to key aspects of the AI use case over time, including risk materiality, scope of usage, and key risks.

Approach:

- Ensure that the use case team periodically performs self-checks to monitor for changes to key aspects of the use case over time.
- Adjust the depth and frequency of the use case team's self-checks based on the use case's risk materiality.

In addition to monitoring use-case performance and data risks, the use case team – typically, the employees responsible for the day-to-day operation of the use case – may also periodically perform self-checks to monitor if key aspects of the use case (such as its risk materiality, scope of use, or key risks) have changed over time. Where changes have occurred, they may then take the necessary steps to address the risks of those changes or may escalate the issue to a relevant party.

These checks can be calibrated to the risk materiality of the use case. For less material use cases, these may be primarily qualitative and functional in nature, such as assessing whether AI continues to be used for its documented purposes, whether the user base has materially changed, or whether there are additional risks associated with the use case beyond those documented in its initial approval. For more material use cases, in addition to these factors, use case teams may also perform testing, whether in the form of simulation testing, adversarial testing, or the calculation of metrics or benchmarks. The depth and frequency of these self-checks are important to formalise in order to ensure that risks are effectively managed.

In addition to self-checks on the use case, use case teams can consider assessing whether their own team members and other stakeholders interacting with the use case continue to be appropriately skilled and supported in doing so. As teams change or employee turnover occurs, it will be important for FIs to ensure that ongoing education and knowledge-sharing remains in place for managing use case risks.

Practice 4: Conduct periodic AI-specific reviews after deployment to assess emerging post-deployment risks.

Approach:

- Conduct periodic AI-specific reviews on the use case after deployment, with the frequency and intensity of reviews based on key risk-related factors.

After the AI use case is deployed, FIs can periodically and proportionately repeat AI-specific reviews to assess key post-deployment risks, such as data drift or repurposing for unapproved use cases. The frequency and intensity of post deployment AI-specific review activities are typically based on the use case's risk materiality, as well as several other key triggering factors like major model or system changes, KPI breaches, stakeholder feedback, or external risk factors, like adverse developments identified in media reporting. Post-deployment AI-specific review is discussed in Subsection 2.4.

Practice 5: Ensure that the use case is operationalised with an appropriate degree of human oversight proportionate to its risk materiality or purpose.

Approach:

- Implement an appropriate and proportionate degree of human oversight as part of the use case's post-deployment usage and monitoring.
- Where the use case was designed to include human-over-the-loop review, implement that review at an appropriate frequency and sampling methodology.

Human oversight is a tool for ensuring that AI use cases – especially Gen AI or Agentic AI use cases where ground truth is hard to determine and accuracy is hard to measure statistically – are used as intended and that their outputs remain aligned with defined objectives. Doing so supports the detection of errors, misuse, or unintended outcomes that may not be captured through automated controls alone; alternatively, when automated monitoring detects a potential breach, human review can be triggered to determine the seriousness of the issue and to remediate it accordingly.

The appropriate degree of human oversight is typically defined at the design stage and reaffirmed throughout the lifecycle based on factors like technical and operational feasibility and risk materiality. After deployment, FIs can operationalise this initial design, ensuring that human overseers are enabled with effective tools and relevant training. Human oversight is more capable when AI use cases are enabled with measures that support transparency and explainability (e.g., counterfactual justifications and references to ground truth) to support effective review and ensure that outputs are assessed accurately and consistently. Where human oversight is conducted, FIs can benefit from documenting any issues identified or interventions taken.

Where the use case's design calls for human-over-the-loop review (one of the three modes of human oversight discussed in Subsection 3.1), FIs can consider the appropriate sampling methodology for operationalising it. The frequency and intensity may be applied proportionately to the use case's risk materiality and operational feasibility (including factors such as volume of outputs). This is outlined in Table 3.5.1.

Table 3.5.1: Sampling Methodologies for Post-Deployment Human Over-The-Loop Oversight

Human Oversight Type	Sample Risk	Description and Rationale
Targeted High-Frequency Review	Higher-risk use cases (e.g. AI-generated customer interactions for financial advice).	High-frequency reviews are conducted more frequently, using a representative sample of outputs. This ensures that outputs meet required standards and allows the use case team to take remediation action quickly if issues occur.
Lower-Frequency Review	Lower-risk use cases (e.g. internal knowledge assistants).	Human reviews can be done less frequently on a representative sample of the AI's output if the use-case has a lower risk materiality. This helps to reduce unnecessary governance effort.
Threshold-Based Review	Suitable for all use cases regardless of risk materiality.	Threshold-based reviews do not involve human oversight in the course of normal operation, but may trigger a human review when certain conditions are met (such as a breach of risk metric thresholds).

Practice 6: Provide end users with avenues to enquire, give feedback, or request a review on AI decisions, where applicable, to support continuous improvement and build user trust.

Approach:

- As applicable and appropriate, provide avenues, such as a feedback or contact form for end users to seek support, raise concerns, contest outcomes, and share feedback.
- Provide human recourse and remediation to parties impacted by AI-based decisions, and consider changes to the AI use case as needed.

FIs may provide avenues for end users to enquire or share feedback, where applicable, on external-facing AI use cases to allow users to seek help, raise concerns, or contest AI outcomes when needed. For example, a chatbot may provide an option to connect the user to a human agent if it is unable to address the user's needs. This is more important for AI use cases that interact, directly or through the decisions they make, with customers or partners, and especially where they have material impacts on the user, such as a system for automated credit decisioning. In contrast, internal tools like knowledge assistants may pose less risk and may not require such support mechanisms.

For AI services that are offered to external stakeholders like customers, FIs may also offer accessible channels to provide feedback such as online portals, customer service hotlines, or in-branch support for users to raise concerns, report issues, appeal, or request human review of a decision that impacted them. Feedback and complaints can be evaluated to determine suitable remedies, including detailed explanations or corrections via human review or system updates. Where human intervention takes place, FIs can document issues and changes to improve traceability and facilitate improvement.

Users may also be offered simple options to rate the system's output, flag inappropriate content, or connect to a human agent. Simple user rating techniques – like a "thumbs up, thumbs down" button – can be quantified and tracked as metrics, triggering alerts when the proportion of negative feedback increases beyond a threshold or trend. Feedback gathered through these channels can support model fine-tuning, monitoring, and system retraining, contributing to safer and more accountable AI deployments.

Feedback or escalations, especially those that are investigated and substantiated, can be treated as AI risk events, issues, or incidents, depending on their gravity. These can trigger review processes and, if needed, changes to the AI use case.

Practice 7: Ensure that proportionate monitoring and analysis are in place to safeguard against security risks during system usage.

Approach:

- Adopt a risk-based approach for the ongoing monitoring of unstructured inputs to or outputs from Gen AI or Agentic AI use cases to detect security risks such as manipulation attempts, prompt injection, or other adversarial attacks.
- Implement safeguards, where necessary, to mitigate detected security threats by flagging and blocking inappropriate, confidential, or adversarial inputs or outputs.

FIs may consider adopting a risk-based approach to monitoring Gen AI or Agentic AI use cases for the unique security threats that can be posed by unstructured, user-generated inputs, such as text-based prompts. Adversarial usage, which includes manipulation, prompt injection, or other techniques for inducing the system to reveal information or perform actions that it is not authorised to, can be detected both through the screening of prompts for undesirable topics or patterns (input filtering) or through the screening of outputs for unacceptable behaviours or content (output filtering). Monitoring can be conducted on an ongoing basis, rather than as a one-time exercise, to account for evolving threats and usage patterns. Filters may be based on predefined sensitive categories and may, depending on the use case's risk materiality, include simple string matching or complex semantic checks to flag disallowed content.

Security measures can be adopted proportionately to the risk that a Gen AI or Agentic AI use case poses. Internal knowledge management chatbots that cannot access sensitive data and are only available to employees, for example, may require little or no filtering, whereas a support chatbot accessible to the public may have a higher risk materiality and require extensive security measures.

FIs may implement a combination of safeguards to detect prompt injection and mitigate security risks in Gen AI or Agentic AI use; they can also consider using a combination of input and output filters to ensure that risks are robustly mitigated. Other measures may include logging and analysing prompts for suspicious patterns, validating outputs for format deviations, leakage, or safety violations.

Consideration 15

Capture changes to AI use cases or their components to maintain traceability and ensure that changes with a material impact on risk are reviewed and approved through an effective change management process.

Practice 1: Establish AI change management process to ensure that changes to in-house or third-party use cases are appropriately tracked, reviewed, and approved before implementation.

Approach:

- Ensure that an appropriate change management process is implemented to govern any modifications to in-house and third-party AI use cases.
- Track and manage changes to AI components, assess the risk materiality of modifications, evaluate changes in the use case's risk assessment, and support effective governance and risk management of use case modifications.

FIs may consider enhancing their existing ITSM change management processes to include approaches for managing changes to production AI use cases, whether those are managed by the FI, are embedded in third party systems, or are accessed as connected services. This may involve defining change types, approval workflows, and escalation thresholds to ensure that modifications, such as updates to model's architecture, training techniques, or system configurations are controlled, logged, and subject to appropriate review and approvals before implementation. Doing so reduces the risk of unintended performance degradations or misalignment with intended use.

As part of this process, version control and documentation can be used to systematically track and manage changes to AI components, including models or their internal weights, training data, or hyperparameters. Version control enables traceability and accountability, ensuring that modifications are logged, reviewed, and auditable. FIs may also impose governance by consistently defining what constitutes a material or significant modification, such as changes to the model architecture or training techniques.

Less-significant modifications, like retraining models with more recent data, are less likely to impact the use case's overall risk materiality and may only require light performance checks to ensure the results remain within expected parameters. FIs can nonetheless track those changes and consider them as an input in the frequency of AI-specific reviews. Where changes are material, FIs can leverage their existing IT change management processes to subject these to an appropriate review and approval process, such as ensuring that SDLC steps like validations and approvals take place. AI-specific steps, such as the AI risk assessment or AI-specific review, may also be repeated alongside relevant SDLC activities.

FIs may not always have visibility over, or control of, changes to third-party AI products and services.

Certain AI use cases, like fraud detection or anti-money laundering, may require frequent model updates due to the nature of their rapidly evolving subject matter. Use cases with such rapidly changing models are known as "dynamic AI"; in these cases, retraining on recent data is conducted at a high frequency, but changes to model architecture or hyperparameters are not regularly made.

Such dynamic AI may need to be subjected to enhanced risk management requirements, including clear justifications for enabling automatic updates, enhanced data quality checks, drift detection, and more stringent performance monitoring with tighter notification thresholds to ensure effective change management and risk oversight.

FIs can consider collaborating with the providers of those products or services to negotiate, or contractually formalise, a notification process for the FI to be informed of upcoming changes and given certain details that can support risk management. This may not be possible in all cases, however. Managing AI-specific risk in third party relationships is discussed in Subsection 2.3.

When notified of a change, FIs may assess its impact on the use case to ensure that it does not negatively impact performance or risk, and where the change is considered material, can conduct a proportionate degree of testing to understand its impacts. Where FIs make an informed decision to deploy a use case containing AI products or services without assurance of sufficient change notification, they can consider other controls, such as an increased frequency of reviews or an enhanced level of post-deployment monitoring, to mitigate the risks associated with changes.

Illustration 3.5.1: AI Usage, Monitoring, and Change Management at DBS

Robust usage, monitoring and change management practices are incorporated in DBS' holistic Responsible Data Use (RDU) framework to ensure that AI deployed in DBS continues to be safe and ethical during use. Examples of how usage, monitoring, and change management practices are implemented for DBS-GPT, an internal Gen AI assistant accessible to staff to support content generation, information retrieval, and workflow automation, are outlined below:

Usage and Monitoring:

- **Performance & Data Quality Monitoring:** DBS-GPT's performance and data quality are actively monitored through a multi-faceted approach. A dashboard tracks usage patterns and performance results, enabling proactive interventions. User feedback collected within the system helps identify areas for improvement and address emerging concerns. The underlying infrastructure and operating environment are also monitored for stability and security. Regular refreshes of data sources used by DBS-GPT ensures information presented to users is up to date and accurate. Users can also flag potential data quality issues to support continuous improvement.
- **Human Oversight & User Feedback:** Human oversight is integrated into key stages of DBS-GPT's lifecycle. During performance evaluation, domain experts review outputs using curated evaluation datasets. Additionally, all users can review and provide feedback on the outputs generated during their interaction with the system. These feedback mechanisms are built into the system and actively monitored and analysed to identify potential improvement areas. This continuous feedback loop ensures alignment with user needs and supports continuous improvement of the solution.
- **Security Monitoring:** Given that DBS-GPT is primarily for internal use, the risk of external manipulation or prompt injection attacks is considered lower. Nonetheless, user prompts are logged for audit and investigation purposes in case of any potential misuse.

(Continued on next page)

(Continued from previous page)

Change Management:

- **Change Control Process & Peer Review:** Changes are thoroughly tested in a dedicated testing environment before deployment to production. This process helps identify and mitigate potential risks before they impact users. Changes are also documented, reviewed, and approved by the product team and relevant senior stakeholders to ensure traceability and accountability. When major changes are made to the use case, the changes undergo a peer review process to ensure that the use case remains fit for use and key risks are not neglected.

4. Enablers

4.1 Enable AI Governance with Skills, Knowledge, and Culture

FIs can enable AI governance and risk management by ensuring that their employees are equipped with the right capabilities to manage AI risks. FIs can promote these capabilities through measures to promote AI governance and risk management -related skills, the knowledge that underpins and enables those skills, and a responsible, ethical, and safe culture of AI use.

Employees involved in AI governance and risk management play a variety of roles. This Subsection focuses on the five key roles in scope for Project MindForge: Executives, Builders, Custodians, Use Case Owners, and Business Users. Each role has unique skill and knowledge needs to manage AI-specific risks.

FIs already have a broad range of measures in place to manage their commitments around talent, and delivering on AI governance and risk management does not require FIs to supplant these existing practices. Instead, FIs can look to their existing strategies and, if necessary, uplift them with key measures – like education, training, and effective whistleblowing – that are tailored to the needs of managing AI-related conduct and control. FIs can improve the functioning of their AI governance and risk management by ensuring that the responsible teams are interdisciplinary and representative. Finally, the rapidly evolving nature of AI – with new technologies and risks continuously emerging – highlights the need for ongoing and frequent maintenance of those practices.

Effective management and engagement of employees is a key supporting capability for the functioning of AI governance and risk management. It often benefits from effective coordination between the AI governance and risk management function and appropriate non-AI specific talent management practices in the FI; this ensures that AI-specific risks and mitigations can be communicated and addressed in the context of broader talent and training activities.

Consideration 16

Ensure that practices are in place to equip employees with the necessary AI governance and risk management skills, knowledge, and AI culture, while ensuring that teams involved in AI governance and risk management function are sufficiently representative.

Practice 1: Ensure that employees in relevant roles have the skills that they require to identify, mitigate, and track AI risks throughout the AI lifecycle.

> **Approach:**
>
> - Determine the skills required in each AI governance and risk management persona – Executives, Builders, Custodians, Use Case Owners, and Business Users – to complete their additional AI governance and risk management responsibilities. This includes both role-specific technical skills and generally applicable cross-functional and behavioural skills that underpin AI governance and risk management.
> - Leverage the FI's existing skill and talent strategies to ensure that the FI is equipped with the appropriate AI governance and risk management skills.
>
> Skills, the "learned capacity to perform a task to a specified expectation",[7] are important for delivering on the FI's AI governance and risk management commitments and mitigating its AI risks.

Equipping employees in each AI governance and risk management role with the appropriate skills is an important part of realising AI governance and risk management.[19] Skills – the capability to act – are distinct from knowledge – which is the awareness that underpins it. Knowledge for AI governance and risk management is discussed in Practice 2 below.

Teams involved in AI governance and risk management typically require three main types of skills:

1. **Technical Skills:** Knowledge of AI guardrail implementation techniques and benchmarks.
2. **Cross-Functional Skills:** Capability in AI governance techniques and the management of AI risks.
3. **Behavioural Skills:** Familiarity with business ethics and responsible decision-making capability.[20]

FIs can each determine the appropriate approach to ensuring that these skills are present in their organisations depending on the level of skills currently in place and on their existing approaches to talent management. To do so, they may first conduct an assessment to identify the governance skills that they will require in the context of their regulatory environment, their internal structure, and their current usage of AI.

Table 4.1.1 illustrates some of the new, AI-specific skills that are typically associated with each AI governance and risk management role.

Table 4.1.1: Typical Additional AI Governance and Risk Management Skills Associated with Key Roles

	Executives	Builders	Custodians	Use Case Owners	Business Users
Technical Skills	-	• Deep technical skills around the selection, application, and calculation of AI-specific guardrails, metrics, and benchmarks. • Best practices related to application design specifiac to AI risk and governance, such as AI testing. • Operation of AI governance tools.	• Relevant[21] technical skills around the selection and calculation of AI-specific guardrails, metrics, and benchmarks to enable them in overseeing Builders. • Comprehension and interpretation of AI metrics and benchmarks.	• Comprehension and interpretation of AI metrics and benchmarks. • Ability to complete risk assessment and monitoring responsibilities.	• The skills required to use their tools responsibly – like good prompt design for Gen AI.
Cross-Functional Skills	• Ability to identify and respond to the risks of AI. • Risk management practices relevant to their job descriptions.				
Behavioural Skills	• Exercise of ethical judgement when using AI. • Ability to apply relevant conduct rules to the use of AI. • Ability to understand the context of local cultures, their impact on AI, and vice versa.				

[19] See the definitions of the key roles used in this Subsection in Subsection 1.1.

[20] These skill archetypes are described in more detail in the 2024 SFA Technology Talent Report. [23]

[21] Relevance is a distinction for each FI to make in its own context. In general, employees in governance roles do not require the same level of technical skills that Builders do. Relevant skills in their context will include a baseline familiarity with the key tools and techniques used by Builders such that they can meaningfully understand, interpret, and question the decisions made by Builders.

These AI-specific skills are additional to the typical capabilities that FIs require from their employees – which continue to be expressly relevant but are not specific or additional to AI.

FIs can leverage their existing talent strategies and talent management functions to ensure that they have sufficient AI governance and risk management skills, choosing an appropriate mix of talent acquisition, workforce augmentation, on-the-job learning, and upskilling that is suitable for their businesses, talent strategies, and AI risks.

In all cases, it is important that FIs continue to periodically revisit and revise AI governance- and risk management-related skill strategies, programmes, and learning opportunities as the underlying technology and risks evolve. Where changes occur or new developments are made, FIs benefit from a nimble and responsive approach to addressing them.

Practice 2: Ensure that learning and literacy activities are sufficient to equip current and future employees with knowledge on AI capabilities, risks, and responsibilities appropriate to their roles in managing AI risk.

Approach:

- Determine the specific knowledge required by employees in each AI governance and risk management persona – Executives, Builders, Custodians, Use Case Owners, and Business Users – to complete their additional AI governance and risk management responsibilities.
- Leverage the FI's existing learning and training strategies to ensure that employees in each role have the knowledge and awareness required for their AI governance and risk management responsibilities.
- Promote a baseline level of AI literacy across the enterprise that includes a basic awareness of AI's characteristics, its risks, and how to use it.

As FIs increasingly adopt AI, they can consider the benefits of implementing enterprise-wide education on responsible AI use to improve both general AI risk literacy and role-specific AI risk knowledge. Knowledge related to AI governance and risk management makes it possible to apply these skills and can inform other aspects of AI governance and risk management, like good conduct.

This education typically addresses topics like the key risks and limitations of AI, the responsibilities of individual employees in governing AI, and the FI's relevant policies on AI use. AI risk-related education can often be incorporated into existing educational initiatives that the FI has in place.

General AI risk literacy measures are most effective when they are proportionate to the FI's overall level of AI use and AI risk materiality. Where Business Users might interact directly with general-purpose AI tools, for example, fundamental AI risk literacy can equip them to identify risky situations, avoid potentially harmful use, and access relevant colleagues or resources where appropriate. AI literacy typically also includes general, foundational education on the FI's AI definition and core values around AI use, like the FEAT principles. It also plays a key role in ensuring that the FI's code on conduct on AI use is properly applied and upheld. In some FIs, building a risk-appropriate level of literacy may involve educating all employees; in others, a subset of employees whose job responsibilities involve interacting with AI may be sufficient.

FIs can engage their relevant talent management functions to deploy educational measures that impart role-specific knowledge on AI risks.

Table 4.1.2: Typical Additional AI Governance and Risk Management Knowledge Required by AI Governance and Risk Management Roles

	Executives	Builders	Custodians	Use Case Owners & Business Users
Knowledge Requirements	• FI AI strategy, key use cases, and their impacts in the enterprise. • Risk and ethical implications of their AI use. • General awareness of AI risks and trends so they can give responsible strategic direction. • Awareness of key industry AI governance and risk management frameworks and norms.	• Awareness of how specific technical decisions (e.g. data selection, dataset management) impact AI risks. • Awareness of market developments related to AI risks and risk management, especially around industry norms.	• Detailed, specific knowledge of AI risks and risk causes. • Awareness of market developments related to AI risks and risk management, especially around emerging AI governance and risk management regulations, frameworks, and norms.	• Knowledge of the specific risks of AI that their role impacts (such as third-party AI risk for procurement teams).

	• Fundamental literacy on AI and AI risks, which can include:
	  ◦ An awareness of the basic characteristics of AI (what it is, what it does, and what key terms like "Gen AI" mean), the FI's definition of AI, core values for AI use, and the general benefits and limitations of AI.
	  ◦ Fundamental knowledge of each of the dimensions of AI risk, such as Fairness & Bias, Accountability & Governance, Ethics, Transparency, Cyber & Data Security, Legal & Regulatory, and Robustness & Stability.
	  ◦ An awareness of the "do's and don'ts" of AI, in line with the FI's code of conduct.			

FIs may choose to develop such education in-house or leverage existing training providers, like course or certification providers.

As with skills, FIs benefit from carefully considering opportunities to improve training and knowledge-building programmes or materials. Periodic monitoring to identify new technologies, risks, or other developments is an effective method for doing so.

Practice 3: Ensure that practices, programmes, and policies related to culture and conduct are sufficient to foster a healthy AI culture around responsible, ethical, and safe AI use for current and future employees.

Approach:

- Assess the suitability of existing risk culture and conduct measures in managing AI-specific risks and uplift them, as necessary, with AI-specific components that support a culture of responsible AI use.
- Uplift rules and expectations around employee conduct to manage the risks of inappropriate or unauthorised AI use, as necessary.

Policies, processes, and technical guardrails are only effective if the individuals implementing them take their responsibilities seriously and execute their roles faithfully. This is equally true for AI as it is for other types of risk management in an FI. FIs are expected to already have conduct- and risk culture-related strategies, practices, and programmes in place: these foundations continue to be relevant in establishing an effective AI culture. In particular, FIs that use AI typically already have practices in place to ensure that their AI culture promotes Fairness, Ethics, Accountability, and Transparency. Where the use of AI tools is widespread, a key element of an effective AI culture is diligence against over-reliance and the maintenance of human judgement.

The new or enhanced risks and rapidly evolving nature of AI, however, mean that FIs may need to assess gaps in their practices, and if necessary, uplift them to address new or enhanced AI-specific risks or to work at the pace at which AI is evolving. FIs may also choose to uplift existing mechanisms for monitoring and assessing their risk cultures to serve that function for their AI culture.[22] Doing so can help FIs respond to culture and conduct challenges related to new technologies like Gen AI or Agentic AI.

An illustrative, and not exhaustive, list of examples of changes to existing programmes in response to new or enhanced AI risks is included below:

- Including vigilance against external AI-based attacks like deepfakes in security training.
- Incentivising employees to uphold the FI's commitments to data protection when using AI, and to use only approved AI tools.
- Including AI risk communication in senior management messaging.
- Ensuring that incentives for prudent risk-taking and ethical behaviour are designed and effective for AI use.
- Uplifting training and education measures such that employees understand the significance of AI risks and their role in managing these risks.
- Uplifting existing culture monitoring surveys to include questions on AI risks and AI use.
- Adding AI-related data to culture reporting dashboards.

Each of these approaches is most effective when AI governance and risk management teams work closely with relevant non-AI-specific enterprise talent management functions.

[22] See Appendix C for references that may impact existing culture and conduct management practices.

Figure 4.1.1: Illustration of Approved Usage of Specific AI Systems or Models

	Personal productivity use cases	Internal-facing content creation use cases	Document summarisation use cases	Production code generation use cases	Customer-facing use cases
Tool A	✓	✓	✓	✓	✓
Tool B	✓	✓	●	✗	✗
Tool C	✓	✓	✓	●	●
...					

✓ Use standard AI governance procedure ● Escalation required ✗ Not allowed

FIs can further strengthen their risk culture around AI, and guard against conduct-related risks, by revising rules around technology use. This may include new expectations that employees commit to only using AI software approved by the FI (rather than chatbots accessible over the internet), and to only use AI for its intended purposes. FIs may also prohibit the use of AI for certain types of use cases, such as in producing text used in regulatory submissions. Conduct expectations may differ between specific AI models or systems; Figure 4.1.1 illustrates how an FI could clearly communicate this to employees.[23]

A key AI risk that can be mitigated by setting robust conduct expectations is the abuse of testing sandboxes for business purposes – where employees utilise the outputs of an AI model that has not been approved for deployment and is hosted in a development sandbox. This is a serious violation of the FI's IT controls and can introduce numerous unmitigated risks into the FI's ecosystem. While fallible, clear conduct prohibitions against the misuse of testing sandboxes can improve employee compliance and can support the FI in providing effective consequences when violations occur.

Culture-related measures are important to periodically review and revise based on their effectiveness, as well as in consideration of changing patterns of AI use. This will be especially important in view of a broad shift in the industry towards user-managed applications, which will increase the risk management responsibilities on Business Users.

[23] For example, an FI may find that a specific third-party AI model includes unfavourable license terms that expose the FI to an increased degree of risk when using the model. Clearly communicating this to employees through a table of approvals can improve the degree of AI risk awareness and overall compliance. Mitigating the legal and contractual risks associated with third party AI products and services is discussed in Subsection 2.3.

Practice 4: Ensure that AI governance and risk management activities involve a sufficiently representative and interdisciplinary group of employees who can effectively represent a range of perspectives on AI's risks and impacts.

Approach:

- Consider approaches for including a relevant and appropriate range of perspectives on AI – from different technical disciplines, professional backgrounds, cultures, and social groups – in functions related to AI governance and risk management.

In cultivating the talent for AI governance and risk management in their enterprises, FIs can consider whether the teams they have assembled are sufficiently representative to appropriately complete the tasks assigned to them. Interdisciplinary diversity is a key component of effective AI governance and risk management, and it may take the form of involving employees with a range of backgrounds in relevant fields or from diverse parts of the enterprise.

Another key capability for AI governance and risk management teams to cultivate is awareness of the potential fairness implications of AI, especially when those implications go beyond the data fields that are obviously impacted, such as in an AI system trained on data where protected characteristics like gender or race have been removed. This is especially important in large or international FIs, where cross-border social and cultural awareness plays a role in managing AI risks like toxicity and discrimination that may differ between cultural contexts.

To improve this awareness, FIs may also wish to consider the importance of context-appropriate diversity in teams responsible for AI governance and risk management, whether those teams represent the range of relevant disciplinary, social, and global perspectives, and whether those perspectives are heard and respected. FIs can look to their existing talent management practices and internal commitments to diversity and inclusion, as well as any relevant guidance and governance on the subject in jurisdictions where they operate, to identify appropriate dimensions of representation, such as race or gender, to prioritise and tools for doing so.

4.2 Manage AI Infrastructure

AI infrastructure includes the computing power, data storage, and cloud access that facilitates AI use. This infrastructure is a key enabler of AI and is also a source of potential AI risk, especially related to availability, outages, data quality, and breaches of security or data privacy. These risks, and the good practices for managing them, are similar to those of infrastructure for traditional software, and correctly applying existing industry best practices is the most effective way for FIs to manage the infrastructure-related risks of AI.

Consideration 17

Support AI deployment by ensuring that supporting infrastructure is fit for purpose.

Practice 1: Ensure that the FI's AI-related infrastructure is suitable for managing scalability, availability, and security risks posed by the FI's use of AI.

Approach:

- Continue to apply existing security and governance practices to environments used for AI use cases.
- Assess availability and scalability risks relevant to the FI's chosen deployment patterns and ensure that appropriate mitigations for them are in place.
- Consider monitoring infrastructure-related AI metrics to identify risks.

Infrastructure management is important for managing AI risks related to availability, outages, data quality, environmental sustainability, and breaches of security or data privacy. These risks are generally addressed by the application of existing, non-AI specific good practices. FIs can consult the relevant guidelines (defined in Appendix C), and resources such as the Open Platform for Enterprise AI (OPEA) for information on effective, risk-appropriate infrastructure management and architecture. Applied correctly, such existing practices address nearly all AI-related infrastructure risks.

Access management to AI components and infrastructure continues to be important and can continue to be managed in line with existing infrastructure security practices. This includes both technical measures around authentication and process controls, such as an effective separation of duties.

When deploying AI use cases, FIs can also continue to leverage existing technology governance practices, especially those discussed in the frameworks referenced in Appendix C of this Handbook, that ensure that deployment is secure. This includes the provision of deployment environments that are hardened against attack, with appropriate identity and access management, and that support key cybersecurity measures like input validation, encryption, and data loss prevention. FIs can ensure that they have sufficient infrastructure in place to provision the secure environments and segregated networks that can facilitate effective AI deployment. FIs can pay particular attention to security around third party components or APIs, such as by ensuring that environments support authentication measures.

Availability and scalability are particularly important for AI use cases, whether deployed on premises or in the cloud, because of their high compute and data requirements. When measuring the usage and scaling of Gen AI use cases, FIs may need to consider metrics different from those of traditional software. A selection of relevant metrics is included in Appendix F.

When FIs use advanced AI training techniques to improve efficiency or preserve privacy, such as distributed computing, parallel computing, and federated learning, they can consider whether their infrastructure has sufficient additional capacity, meets required dependencies, and has appropriate networking and encryption capabilities. FIs using third party cloud-based AI services, especially for Gen AI, can consider whether their infrastructure can support governance-related and efficiency-boosting features like prompt gateways and prompt caching.

Part 3

Final Remarks

Future Perspectives

AI has continued to evolve over the course of Phase 2 of Project MindForge; in the time since the development of this Handbook was kicked off in November 2024, the industry has moved towards implementing AI in agentic and multi-agent architectures. Although the governance of Agentic AI is referenced throughout this Handbook where relevant, readers should note that its governance and risk management remains nascent and that well-established practices are still emerging. This section discusses some of the emerging considerations specific to Agentic AI.

The technologies, and risks, associated with AI in the financial services industry will continue to evolve. FIs will benefit from adopting a governance approach that is flexible by design – building in horizon scanning and periodic review to evolve as the FI's needs do – and from having well-defined principles like FEAT that are open-ended enough to be relevant to future technologies. FIs with robust practices in place to proportionately update governance as risks change will be positioned for success as new opportunities emerge.

Characteristics of Agentic AI

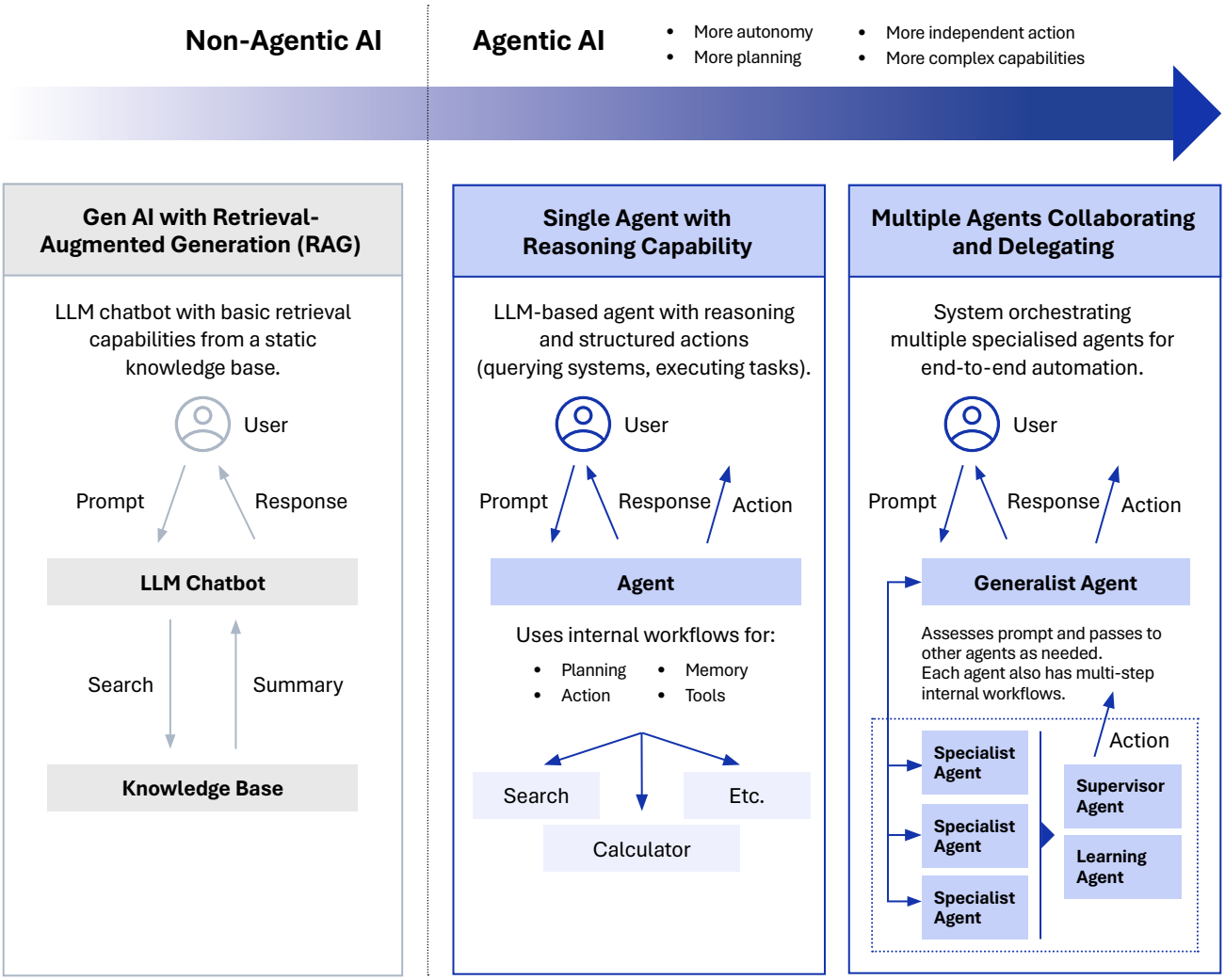
Per the International AI Safety Report, Agentic AI is a "general-purpose AI [system] which can make plans to achieve goals, adaptively perform tasks involving multiple steps and uncertain outcomes along the way, and interact with its environment – for example by creating files, taking actions on the web, or delegating tasks to other agents – with little to no human oversight".[1]

The term "AI agents" is often used to describe individual components, powered by one or more AI models, that can perform tasks and make decisions in an agentic system. An agentic system composed of multiple agents is a "multi-agent system".

In practice, there is not yet a widely-accepted and clear boundary between agentic and "non-agentic" AI. FIs can instead consider "agenticness" as a question of degree, in which several qualities commonly associated with Agentic AI are present. One early paper in the field of Agentic AI governance suggested four characteristics: "underspecification", "directness of impact", "goal-directedness", and "long-term planning". [7] A recent survey lists characteristics "autonomy and goal complexity", "environmental and operational complexity", and "independent decision-making and adaptability".[1] Another recent paper lists "goal complexity", "environmental complexity", "adaptability", and "independent execution". [22]

In general, the term "Agentic AI" is used in the industry to describe systems that include components for making plans and for taking action, often involving tools that interact with their environment. Gen AI systems like RAG-based chatbots are not typically referred to as Agentic AI due to their relatively low levels of autonomy. The Figure below illustrates a (non-agentic) Gen AI system alongside stylised depictions of two illustrative agentic architectures.

Figure 5.0.1: Stylised Comparison of Gen AI and Agentic AI Architectures



Business Potential of Agentic AI

Agentic AI has positive potential for the financial services industry in three primary areas: the creation of business value, the democratisation of AI capabilities, and the management of AI risk. The practical implementation of Agentic AI in the sector is a rapidly emerging field, and it is likely that its upside potential will continue to grow as the technology matures and as businesses gain experience with its adoption.

Agentic AI's potential to create business value lies in its ability to take complex, multi-step actions, including the creation of plans based on high-level natural language directions and the execution of those plans using connected tools. Agentic AI systems can undertake tasks that could never before be automated – in this sense it represents a substantial maturation of Gen AI, which was primarily constrained to content creation, summary, or retrieval tasks. Where Gen AI systems were primarily reactive in nature, agentic systems have the potential to be proactive, to effect real-world states directly rather than relying on users to take actions on their advice, and to offload tasks that are not well-suited to AI to other software tools. Mathematical tasks, for example, severely vexed many early Gen AI models and continue to be challenging for them to answer reliably; an agentic system could dynamically pass mathematical tasks to a traditional calculator. Agentic systems can also, through their memory and planning functions, take on broader and more complex tasks without losing context, which is otherwise a major limiting factor for LLM utility.

Secondly, Agentic AI can give Business Users more capabilities to design their own AI workflows. "Build-your-own-bot" approaches, which allow users to call and configure individual AI components in systems of their own design, have the potential to unlock a range of new efficiencies and use cases across an FI by allowing even non-technical staff to design powerful AI tools. Enterprise platforms for employees to design and run AI agents could one day also have AI guardrails built in. While the adoption of user-managed agentic systems remains at an early stage, its long-term benefits are clear.

Thirdly, agentic systems have significant potential to mitigate AI risks. A known challenge in the business adoption of Gen AI is its tendency to commit errors or to outright hallucinate information, which can sometimes make it risky to use for critical use cases. While errors and hallucinations also occur in agentic systems, the introduction of multi-step reasoning processes or built-in error checking gives agentic systems a greater potential to identify and catch their own errors. Agentic architectures, because they involve multiple steps and components, can also lend themselves well to the use of guardrail modules, whether for screening inputs or outputs for PII, detecting toxicity, or fact-checking for hallucinations.

Agentic systems represent an important frontier for research in the field, and their capabilities are rapidly evolving. They are currently well-suited to tackling more complex, higher-materiality workflows than prior AI architectures could, and can unlock significant benefits for FIs as a result.

Select New Challenges and Potential Risks of Agentic AI

As Agentic AI technologies mature, FIs may encounter challenges specific to them that they will seek to manage and overcome; they may also find that the use of Agentic AI introduces or enhances AI risks. This is an emerging field, and this description represents an initial, partial view of the broader landscape of Agentic AI's opportunities and challenges.[24]

- Challenges related to Agentic AI's complexity.
    - Interpretability. The technical complexity of agentic systems – which can orchestrate numerous AI agents, each powered by a "black box" AI model – can make it harder to interpret their behaviours and decisions.
    - Misalignment and controllability. Alignment with human intentions is a well-known existing challenge for AI. The complexity of agentic architectures, as well as their tendency to be used for much more complex tasks over longer time horizons, can make alignment more challenging.
    - Emergent behaviours. Complex interactions between components, especially between different agents, are difficult to predict. Even in cases where individual agents behave within acceptable parameters, emergent behaviours from systems of individually performant agents could lead to unintended or undesirable outputs.
    - Cascading or correlated errors. In complex agentic systems, especially those that include multiple interconnected agents, the number of points of failure increases non-linearly. Errors in one agent or component could cause unpredictable or unwanted behaviour overall.

[24] For other perspectives, FIs can consult relevant resources like the OWASP Top 10 (at https://genai.owasp.org/resource/agentic-ai-threats-and-mitigations/) and the IBM Whitepaper Agentic AI in Financial Services: Opportunities, Risks, and Responsible Implementation (at https://au.newsroom.ibm.com/IBM-Whitepaper-Accountability-and-Risk-Matter-in-Agentic-AI).

- Challenges related to Agentic AI's capability.
  - Ability to take actions. Agentic systems are characterised by their high levels of autonomy, including the ability to take actions that impact the digital and physical environment (such as executing actions in other software). Agentic systems with high degrees of autonomy pose increased risks of unintended harms. Tool access is one of the most important vectors of risk for agentic systems; the potential impact of those connected tools is a key consideration in determining the riskiness of an Agentic AI use case.
  - Cybersecurity. Agentic systems frequently interact with other services or the open internet; they also often collect large quantities of user and enterprise data. The replication of data between components or models, as well as the large attack surface of highly interconnected systems, enhances the risk of intrusions or privacy violations. The ability of Agentic AI to take actions autonomously, without a human's direct involvement, can also enhance their risk of being hijacked by malicious actors to perform actions without detection.
- Challenges related to Agentic AI's governability.
  - Accountability and ownership. Agentic AI, especially systems with extensive user-managed components, may create new challenges around identifying ownership and assigning responsibility for impacts.
  - Testing and risk assessment. Because of the additional complexity of agentic systems – which frequently include many interconnected components – it can be more challenging for FIs to comprehensively identify their risks or test their behaviours. When incidents or issues occur, root cause analysis is made more difficult by the number of independent components in the system.
  - Governance scalability. The use of Agentic AI can increase the number of individual AI models and systems in an FI's ecosystem. Without enabling infrastructure and a robust, scalable, and automated governance framework, FIs may find it challenging to effectively oversee Agentic AI over the long term.

While beyond the scope of this Handbook, the general adoption and acceleration in capabilities of Agentic AI can introduce further inbound risks, such as those introduces when customers use AI agents to interact with an FI. FIs can be attentive to this challenge in the future.

Additional AI Governance and Risk Management Considerations When Using Agentic AI

Governance considerations in managing Agentic AI remain an emerging field. FIs can consider some of the following techniques as a starting point where doing so is reasonable and proportionate to the risk materiality of the use case. Most importantly, FIs can work to remain current as the study of Agentic AI continues to evolve and mature.

As with all other practices referenced in this Handbook, FIs will each determine in their own context how best to apply governance to Agentic AI. Doing so proportionately to risk materiality – and with a view towards feasibility, given that many of the considerations below remain experimental – is important to balance risk management against the clear benefits of agentic techniques.

2.4: Enhance Use Case-Level AI Risk Management

One simple change that FIs can make to address the challenges of agentic architectures is to weight complexity or "agenticness" as a factor in their risk assessments; more complex systems can require more extensive controls to manage them. FIs can consider other Agentic AI-specific assessment factors when conducting use case risk assessments – especially the tools that the agentic system will have access to and the use case's potential attack surfaces, such as agents with internet access.

FIs can consider whether some actions may be too risky to delegate to AI under any circumstances, such as authorising financial transactions or making employment decisions.

Agentic AI accelerates a shift away from traditional model validation and towards assessments of systems in the context of their use cases. Many of the challenges associated with Agentic AI are found in the interactions between components, and in multi-agent systems, in the interactions between agents.

2.5: Ensure AI Inventory Capabilities

AI agents may require additional information to be tracked to facilitate effective risk management. These may include agentic-specific identifiers like information on the tools that the agent can access, the components it has, and what agentic-specific limitations are imposed on it to manage risk.

As FIs increasingly adopt Agentic AI, they are expected to do so in an increasingly modular fashion, where individual AI agents – such as a summarisation agent, a translation agent, or a compliance advisory agent – are developed and made available for integration into multiple agentic systems. This "marketplace" approach is highly efficient, helping to reduce duplicated work and accelerate time to deployment.

FIs that use such a modular development approach, where doing so is proportionate to risk materiality, can consider tracking information on each reusable agent, such as by documenting its permitted uses and known risks. They may also track information on the use cases that those modules are applied to. In more mature organisations, this can take the form of an agent "certification". By tracking risk-related information at the agent level, rather than solely at the use case level, FIs can minimise duplicated governance effort, can ensure consistency in how the risks of agents are managed in deployment, and can track the FI's aggregate exposure to the risks of any individual agent.

FIs can also consider how best to inventory user-defined "build-your-own-bot" systems. They may give users flexibility to create agents within acceptable parameters, only registering agents that are systematised as part of ongoing business processes. This mirrors existing practices for governing user-defined code and macros.

3.1: Use Case Context and Design

FIs can address governance issues at the design stage by clearly defining a division of accountability between parties for the behaviour of Agentic AI. In general, it is most effective to assign final accountability for an AI agent's actions to the party that had the most control over that action – creating an effective incentive to use that degree of control to ensure that the agentic system behaves as intended.

Restricting the privileges, tool and data access, and capabilities of an agentic system is one of the most important guardrails in preventing unwanted behaviours. FIs can consider giving agentic systems the lowest possible level of privilege for their use case, and within each system, can make privileges more restrictive for individual components – such as allowing only one specialised AI agent to access sensitive data or the internet. FIs can be especially attentive to risk vectors, such as scenarios where an AI agent accesses the internet and receives malicious instructions, and work to make their systems secure-by-design with robust screening of inbound data, data handed off between agents, and with frequent resets of agent sessions.

3.2: Data Acquisition and Processing

Agentic AI increases the importance of dynamic data governance – focused on real-time oversight of how Agentic AI systems access and use data – over static, point-in-time data governance.

FIs can be particularly attentive to the importance of minimising data that interacts with an Agentic AI system, and when assessing data risks can consider whether sensitive information could be passed between agents in a multi-agent system. Restrictive, carefully designed data access limits on agents can mitigate against malicious attacks or data leakages.

Where needed, FIs can also build in observability and controls that enable oversight of data used by the agentic system over time. This can include effective telemetry in data used for RAG and tracking of updates to data accessible to the system to identify the root causes of errors.

3.3: Onboarding, Build, and Review

As with other Gen AI technologies, FIs can improve alignment by grounding AI agents with useful context, such as their intended purpose or the instruction that they should avoid certain types of action (such as spending money) unless specifically prompted to do so. These techniques, as with any AI use case, are best applied proportionately to risk materiality.

In the context of Agentic AI, traceability is an important part of effective transparency because it facilitates error identification and post-incident reviews. This can include robust and searchable logging and the capability to trace erroneous or harmful outputs back through the system to their initial point of failure.

Many agentic systems today use versions of the ReAct paradigm, proposed by Yao et al. in 2023, to address a prompt through multiple steps of "reasoning", "action", and "observation". ReAct is an elaboration of existing chain-of-thought (CoT) paradigms, both of which prompt a language model to generate intermediary steps before generating a final output. Various other techniques for hierarchical task decomposition can be used to help agentic systems break down complex, incomplete, or abstract instructions into sequences of specific actions that the system can then execute, improving the groundedness of outputs. While these methods have well-documented limitations, they can form a useful foundation for ongoing monitoring.

Additional security challenges related to Agentic, and especially multi-agent, AI systems are numerous and continuously evolving. These include injections into shared memory functions, attacks on data in transit between agents, or denial-of-service-style attacks that attempt to overwhelm system resources by introducing infinite loops between agents.[25]

3.4: Deployment

Kill switches and other forms of "interruption" can be especially useful, if proportionate to risk materiality, for stopping Agentic AI systems that could take unintended or harmful actions – such as spending money, issuing communications, or other behaviours that are difficult to reverse. Interruption capabilities can be given to individual users, but especially as the use of Agentic AI scales, could also be automated to ensure that harmful actions are promptly identified and blocked. Another prominent interruption feature for agentic systems is "timeouts", which pause or restart agents that have been running for an extended period of time to limit expense and improve alignment on complex tasks.

Phased rollouts and pilot deployments are particularly useful when using Agentic AI due to the technology's complexity. Limited initial rollouts can enable FIs to identify emergent behaviours and other complex failure modes that testing may not uncover.

[25] FIs can consult up-to-date resources like OWASP's Multi-Agentic System Threat Modelling Guide (at https://genai.owasp.org/resource/multi-agentic-system-threat-modeling-guide-v1-0/) for guidance on managing emerging security risks.

3.5: Monitoring, Usage, and Change Management

Because of their complexity, monitoring is especially important in managing Agentic AI systems for unexpected behaviours; where proportionate to risk, FIs can consider the value of automated monitoring, or even monitoring performed by dedicated AI agents. FIs using AI for monitoring can note the importance of also robustly testing the monitoring agent.

FIs can largely leverage existing good practices for monitoring to proportionately manage agentic systems, such as by monitoring points of data ingress or egress for prompt injection or data leakage. FIs could also consider monitoring the usage of tools connected to agentic systems. Tools that experience anomalous or inappropriate usage, such as spikes in activity or attempts to access sensitive data, can indicate that an agentic system could be compromised or malfunctioning.

The diverse and broad capabilities of Agentic AI increase the importance of using human-in-the-loop review as a guardrail for certain types of risky actions. Where it is needed and proportionate to risk materiality, FIs can consider distributed approvals (such as by employees operating the agent) as a useful technique for making human review practical at scale.

Conclusion and Next Steps

This Handbook is an end-to-end guide to AI governance and risk management written by and for the financial services industry. Intended to be universally applicable – to FIs of all types and sizes, and to both well-established and emerging AI techniques – it provides a practical set of actions for mitigating AI-specific risks to people, businesses, and society, aligning AI to human values, and conforming to relevant laws and regulations. This foundation will support faster and better AI adoption across the financial services industry by creating user trust, supporting regulatory compliance, and facilitating more effective value realisation.

This Handbook contributes to the literature on AI governance and risk management by proposing useful definitions of core concepts, a risk-based, proportionate framework, and a flexible approach to process that is focused on uplifting practices that FIs already have in place. FIs are already using the recommendations in this Handbook to govern AI more effectively.

This Handbook is the first of several steps that the consortium will undertake to achieve its mission: to enable and facilitate FIs, at different levels of AI maturity, to scale AI with trust by adopting and operationalising AI governance and risk management across the enterprise, and supporting industry AI use that is rapid, but responsible. These next steps can include:

Ongoing Evolution

This Handbook is a living document; as AI rapidly evolves, these considerations will remain relevant only if they also continue to evolve in response to new regulations and technological developments. The consortium aims to continue to update the text of this Handbook, especially to ensure that it is aligned with the final version of the MAS Guidelines on Artificial Intelligence Risk Management.

The governance of emerging technologies like Agentic AI and other user-managed AI applications is a particular area of interest that the consortium will monitor further.

Training and Education

Effective AI governance and risk management involves building new skillsets – including training governance professionals to implement AI-specific frameworks like this Handbook. Building the talent pool for AI governance and risk management through effective training programmes will be a common good for the industry and is an area where the consortium can continue to collaborate.

Industry Toolkits

This Handbook highlighted the importance of consistent metrics and evaluation methods. Shared toolkits that implement generally acceptable metrics and methodologies can support the generalised adoption of this Handbook and can make its recommendations more implementable in practice.

Ecosystem Development

Institutionalising effective AI governance and risk management goes beyond adoption by FIs; it will involve the development of a robust ecosystem of fintech firms, technology providers, audit and assurance bodies, third party validators, and other providers of ancillary capabilities that FIs will rely on in the long term.

The consortium remains committed to engaging with the AI governance and risk management ecosystem to support and promote innovation that can support its mission.

The risk management of Agentic AI is a rapidly evolving subject. In January 2026, Singapore's Infocomm Media Development Authority (IMDA) launched the Model AI Governance Framework for Agentic AI, which is the world's first comprehensive guide for the responsible deployment of Agentic AI in the enterprise. This document builds on the previous editions of the Model AI Governance Framework developed by the AI Verify Foundation.

The framework identifies five key sources of risk to the enterprise from Agentic AI:

1. Erroneous actions.
2. Unauthorised actions.
3. Biased or unfair actions.
4. Data breaches.
5. Disruption to connected systems.

It also proposes four key dimensions for enterprises to manage Agentic AI risks:

1. Assess and bound the risks upfront.
2. Make humans meaningfully accountable.
3. Implement technical controls and processes.
4. Enable end-user responsibility.

FIs can consult this framework for emerging insights into the field of agentic AI governance.

Part 4

Appendices

A. Glossary of Terms

Accountability: The state of being responsible for a particular set of outcomes, such as behaviours, actions, products, services, or decisions. In combination with recourse and redress, it is a central component of any system of governance. Internal accountability is the ability of an organisation's governance system to hold an individual or group within the organisation responsible for unwanted outcomes. External accountability is the ability of governments, regulatory bodies, individuals, and other agencies to hold an organisation responsible for the outcomes of its operations (Veritas Document 3).

Adversarial Testing: In the context of this Handbook, a testing approach that deliberately engages in inappropriate, unexpected, or out-of-context usage in order to assess the response of an AI model or system. It includes red teaming.

Agentic AI: A general-purpose AI which can make plans to achieve goals, adaptively perform tasks involving multiple steps and uncertain outcomes along the way, and interact with its environment – for example by creating files, taking actions on the web, or delegating tasks to other agents – with little to no human oversight (International AI Safety Report 2025).

AI Lifecycle: The distinct stages of developing AI, including data collection and pre-processing, pre-training, fine-tuning, model integration, deployment, post-deployment monitoring, and downstream modifications (International AI Safety Report 2025).

AI Risk-Related Performance Metric: In the context of this Handbook, a quantitative measure that can indicate the presence or absence of one or more AI-specific risks.

AI Safety: The property of avoiding harmful outputs, such as providing dangerous information to users, being used for nefarious purposes, or having costly malfunctions in high-stakes settings (International AI Safety Report 2025).

AI-Specific Review: In the context of this Handbook, a review process conducted on an AI use case prior to deployment and then periodically after deployment which evaluates certain risk-related criteria and is conducted by a party not directly involved in the development, deployment, or operation of that AI use case.

AI-Specific Risk: In the context of this Handbook, a risk that is new to financial institutions, or a risk that financial institutions already face which is enhanced, when using AI. A selection of AI-specific risks are documented in Appendix B.

Application Programming Interface (API): A set of rules and protocols that enables integration and communication between AI systems and other software applications (International AI Safety Report 2025).

Artificial Intelligence (AI): See Subsection 1.1.

Build In-House: In the context of this Handbook, developing an AI use case within an FI, including training the model and creating essential software components.

Builders: In the context of this Handbook, software developers, data engineers, data scientists, AI practitioners, systems integrators, and other technical specialists involved in the development, deployment, and use of AI.

Business Users: In the context of this Handbook, employees who use or apply AI use cases in the course of their business responsibilities.

CI/CD, DevOps, MLOps, AIOps, LLMOps: Continuous integration/continuous deployment (CI/CD) or DevOps pipelines automate the process of building, testing, and deploying code changes. These terms are closely related to the term MLOps, which is used to describe tools and systems that help to automate the process of building, testing, deploying and monitoring the performance of machine learning systems. More recent terms such as AIOps and LLMOps have also been used to describe such tools and systems for AI in general or for LLMs (MAS AI MRM Information Paper 2024).

Cloud Computing: A paradigm for delivering computing services – including servers, data storage, software, and analytics – over the internet. Users can access these resources on demand and without local infrastructure to develop, train, deploy, and manage AI applications (International AI Safety Report 2025).

Concept Drift: This occurs when the underlying relationships between the features in input data and what the AI model is being used to predict or generate changes. For example, customer preferences for financial products may have shifted due to broad industry changes (e.g., a shift in the relationships between customer information and their preferences for financial products), and an AI model used to generate financial product recommendations may no longer perform as well due to such concept drifts (MAS AI MRM Information Paper 2024).

Consideration: In the context of this Handbook, thematic recommendations that will support an FI in operationalising AI governance and risk management.

Control: [A] measure that maintains and/or modifies risk (ISO 31000:2018).

Culture: The values, beliefs and practices that influence the conduct and behaviour of people and organisations (ISO 30400:2022).

Custodians: In the context of this Handbook, employees in oversight, governance, enablement, and risk management roles in an FI who apply AI governance and risk management policies and procedures and manage AI risks, either directly or in an enabling capacity such as talent, legal, or technology.

Data Drift: This occurs when the statistical properties of the distribution of the data changes. For example, the underlying distribution of customer data may have drifted or changed over time due to changes in the lifestyles of customers. Hence, an AI model that was trained on data from a more distant time period may not perform as well on data from a more recent time period due to data drift (MAS AI MRM Information Paper 2024).

Data Subject: An identifiable living person to whom a particular data item relates (Veritas Document 3).

Deep Learning: A machine learning technique in which large amounts of data and compute are used to train multilayered, artificial neural networks (inspired by biological brains) to automatically learn and extract high-level features from large datasets, enabling powerful pattern recognition and decision-making capabilities (International AI Safety Report 2025).

Deployment Pattern: In the context of this Handbook, the choice of deploying an AI model or system through one of the following approaches: Build In-House, Onboarding Only, or Onboarding with Customisation.

Deployment: In the context of this Handbook, using the outputs of AI for a business purpose at scale, or hosting an AI model or system in, or otherwise connecting it to, a production IT system.

Ethics: The work that is performed to satisfy a set of values, in accordance with principles, in support of governance (Veritas Document 3).

Executives: In the context of this Handbook, decision-makers and leaders in an FI.

Explainability: The ability to have explanations for decisions made (from an [AI-driven] and/or human process) that are understandable by an average adult; a higher standard than interpretability; a component of transparency (Veritas Document 3).

Fairness: Related to the concept of justice; what is right, impartial and equal, without favouritism or discrimination; often assessed through the evaluation of decision outcomes, along with the process followed to reach the decision. What is fair can mean different things in different contexts to different people. Fairness is a guiding principle that aims to prevent the uneven distribution of harms. Expressing questions of fairness as mathematical problems is referred to as “quantitative measures of fairness”. These measures tend to be formulated as metrics that assess a criteria or objective, such as equal or equitable allocation, representation, or error rates, for a particular task or problem (Veritas Document 3).

FEAT Principles: Principles to Promote Fairness, Ethics, Accountability and Transparency (FEAT) in the Use of Artificial Intelligence and Data Analytics in Singapore’s Financial Sector.

Financial Institution: In the context of this Handbook, a bank, capital markets firm, financial advisory firm, insurer, or payments operator.

Generative AI (Gen AI): AI that can create new content such as text, images, or audio by learning patterns from existing data and generating novel outputs that reflect those patterns (International AI Safety Report 2025).

Guardrails: Built-in safety constraints to ensure that an AI system operates as desired and avoids harmful outcomes (International AI Safety Report 2025).

Inherent Risk Materiality Assessment: In the context of this Handbook, an assessment based on a structured methodology for determining the basic potential of an AI use case to engender risk.

Intellectual Property (IP): Creations of the mind over which legal rights may be granted, including literary and artistic works, symbols, names and images (International AI Safety Report 2025).

Interpretability: The degree to which a trained professional can explain how a model arrived at a determination; a lower standard than explainability; a component of transparency (Veritas Document 3).

Key Risk Indicator (KRI): In the context of this Handbook, a quantifiable measure that can indicate the extent or impact of an enterprise risk related to AI.

Knowledge: A human or organisational asset enabling effective decisions and action in context (ISO 30400:2022).

Large Language Model (LLM): An AI model trained on large amounts of text data to perform language processing tasks, such as generating, translating, or summarising text (International AI Safety Report 2025).

Machine Learning (ML): A subset of AI focused on developing algorithms and models that learn from data and improve their performance on tasks over time without being explicitly programmed (International AI Safety Report 2025).

Modality: In the context of this Handbook, the type(s) of data that an AI model or system accepts as inputs or outputs, such as text, images, predictions, or recommendations.

Model: A quantitative method, system, or approach that applies statistical, economic, financial, or mathematical theories, techniques, and assumptions to process input data into quantitative estimates (SR 11-7: Guidance on Model Risk Management, 2011).

Model Risk Management (MRM): The primary objective of model risk management is to ensure that models are fit for purpose, meaning the potential for adverse consequences is within acceptable limits. Models not fit for purpose should not be used, or should be used only in a limited, controlled way while alternatives are sought. Some elements of model risk management programs are also aimed at maintaining the firm's capital stock of models (e.g., documentation requirements). (Global Association of Risk Professionals).

Model Drift: Model drift is a broader term that usually encompasses both data drift and concept drift, as well as other factors that can cause a model's performance to degrade over time. Aside from measures such as PSI and CSI, monitoring the statistical characteristics of AI predictions can also be used to detect drifts in general (MAS AI MRM Information Paper 2024).

Onboarding with Customisation: In the context of this Handbook, the modification of an AI model or system from a third party, including model retraining, fine-tuning, or reinforcement learning, and the modification or development of components in or related to an AI system that contains models or components from third parties.

Onboarding: In the context of this Handbook, the integration or use of an AI model or system from a third party.

Personal Attribute: Personal attributes are features that should not be used as the basis for decisions without reasonable justification. They are defined by FSIs in the context of each specific use case, and at a minimum cover any personal data included in the AIDA system, as defined in relevant data protection and anti-discrimination laws, and can also include non-personal data (Veritas Document 3).

Personally Identifiable Information (PII): Any representation of information that permits the identity of an individual to whom the information applies to be reasonably inferred by either direct or indirect means (NIST SP 800-79-2).

Practice: In the context of this Handbook, specific actions that can be taken to implement a Consideration, if appropriate in the FI's context.

Principles: The norms that describe how to implement values and set guardrails for what ought and ought not be done to operationalise values (Veritas Document 3).

Production: In the context of this Handbook, software or data that used for business purposes related to end users.

Recourse: The ability of a data subject to seek assistance when an outcome from [AI] is anomalous, negative, or unwarranted (Veritas Document 3).

Red Teaming: A systematic process in which dedicated individuals or teams search for vulnerabilities, limitations, or potential for misuse through various methods. Often, the red team searches for inputs that induce undesirable behaviour in a model or system to identify safety gap (International AI Safety Report 2025).

Redress: The act of rectifying the anomalous, negative, or unwarranted outcome from [AI] that has been learned about from a data subject seeking recourse (Veritas Document 3).

Reproducibility: The act of reproducing a model and its outcomes from scratch. Reproducible [AI] systems can be difficult to implement, requiring careful versioning of data, models, code, infrastructure, and even the random seeds used (Veritas Document 3).

Residual Risk Materiality Assessment: In the context of this Handbook, an assessment based on a structured methodology for determining the risk posed by an AI use case in the context of its inherent risk materiality, its observed performance characteristics, and after guardrails are in place.

Retrieval-Augmented Generation (RAG): A technique that allows LLMs to draw information from other sources during inference, such as web search results or an internal company database, enabling more accurate or personalised responses (International AI Safety Report 2025).

Risk Appetite: [The] amount and type of risk that an organisation is willing to pursue or retain (ISO 31000:2018).

Risk Materiality: In the context of this Handbook, the significance of the risks of an AI use case, as measured by their nature, likelihood of occurrence, and severity.

Risk: In the context of this Handbook, the potential negative impacts of the development, deployment, or use of AI on an FI, its employees, its customers, society, or the environment.

Safety: See AI Safety.

Security: The property of being resilient to technical interference, such as cyberattacks or leaks of the underlying model's source code (International AI Safety Report 2025).

Skill: [The] learned capacity to perform a task to a specified expectation (ISO 30400:2022).

Software Development Life Cycle (SDLC): A formal or informal methodology for designing, creating, and maintaining software (NIST SP 800-218).

Software-as-a-Service (SaaS): The capability provided to the consumer [to] use the provider's applications running on a cloud infrastructure. The applications are accessible from various client devices through either a thin client interface, such as a web browser (e.g., web-based email), or a programme interface. The consumer does not manage or control the underlying cloud infrastructure including network, servers, operating systems, storage, or even individual application capabilities, with the possible exception of limited user-specific application configuration settings (NIST SP 800-145).

Supervised Learning: Supervised learning is a machine learning approach where a model is trained on a labelled dataset. In this process, each data point includes input features paired with the corresponding output (label). The model learns to map inputs to outputs by comparing its predictions with the actual labels and updating the model parameters iteratively. Classification, which involves the prediction of classes or categories, and regression, which involves the prediction of continuous values, are common examples of supervised learning (MAS AI MRM Information Paper 2024).

System: An integrated setup that combines one or more AI models with other components, such as user interfaces or content filters, to produce an application that users can interact with (International AI Safety Report 2025).

Talent Management: The implementation of integrated strategies to develop improved processes for attracting, developing, retaining and utilising people with special skills and aptitudes to meet current and future organisational needs (ISO 30400:2022).

Talent: A person who has or can develop the knowledge, skills, abilities, and other characteristics to perform a function, job or role, as required (ISO 30400:2022).

Third Party: In the context of this Handbook, the vendor or provider of an AI product or service that is used by the FI.

Traditional AI: In the context of this Handbook, a subset of AI whose outputs are predictions, recommendations, or decisions in a specific context, and which is not Gen AI or Agentic AI.

Transparency: Visibility into the actions a system or organisation takes, and the decision making processes behind them. Where such visibility is aimed at data subjects impacted by an [AI-driven] decision, it is referred to as external transparency. Where such visibility is aimed at internal stakeholders and to the organisation's regulators, it is referred to as internal transparency (Veritas Document 3). Measures supporting Transparency include Explainability and Interpretability (see definitions above).

Unsupervised Learning: Unsupervised learning is a machine learning approach where a model discovers patterns in data without the use of labels. An example of unsupervised learning is clustering, where data points are grouped together based on their inherent similarities or dissimilarities (MAS AI MRM Information Paper 2024).

Use Case Owner: In the context of this Handbook, an employee who is accountable for an AI use case.

Use Case: In the context of this Handbook, a specific real-world purpose for which an AI system is intentionally applied. The application of AI is recurring and serves a business purpose.

Veritas Methodology: The assessment methodology defined in Veritas Document 3: FEAT Principles Assessment Methodology.

B. MindForge AI Risk Taxonomy

The consortium defined and endorsed a risk taxonomy as part of Phase 1 of Project MindForge. Minor updates and revisions were made to this taxonomy at the beginning of Phase 2 of Project MindForge to account for the perspectives of the financial industry beyond the banking sector.

The consortium also considered the extent to which the risks identified in Phase 1 of Project MindForge were distinct from existing model risk management. A new column, "AI-Specific Elements to Risk", was added to include this analysis.

The Association of Banks in Singapore (ABS) published the "Handbook on Generative AI Guardrails in Banking" in May 2025. This Handbook leveraged the risk taxonomy developed in Phase 1 of Project MindForge to focus on the ten AI risks that ABS members considered to be the most novel and pressing. Those risks are highlighted in the taxonomy below.

Changes made to the MindForge AI Risk Taxonomy in Phase 2 of Project MindForge include editorial revisions to the titles of three of the risk dimensions identified in Phase 1 of Project MindForge:

Risk Dimensions in MindForge Phase 2	Former Titles of Risk Dimensions in MindForge Phase 1, If Different
Fairness & Bias	
Ethics	Ethics & Impact
Accountability & Governance	
Transparency	Transparency & Explainability
Legal & Regulatory	
Robustness & Stability	Monitoring & Stability
Cyber & Data Security	

Six individual risks were also updated as part of Phase 2:

Risk Dimension	Risk	Change
Accountability & Governance	Lack of use case, data and model governance	Name revised from “Lack of use case and model governance”.
Accountability & Governance	Lack of AI risk awareness	Definition and name revised.
Transparency & Explainability	Lack of explainability	Definition revised.
Monitoring & Stability	Insufficient data quality	Definition revised.
Monitoring & Stability	Unmet architectural requirements	Definition revised.
Monitoring & Stability	Lack of reproducibility	Additional risk added as part of Phase 2.

Finally, the lifecycle stages described in the MindForge AI risk taxonomy have been lightly revised to match those used in Phase 2 of Project MindForge. The lifecycle stages are referred to numerically in the AI risk taxonomy; these figures correspond to the lifecycle stages as follows:

Number	Lifecycle Stage
1	Use Case Context and Design
2	Data Acquisition and Processing
3	Onboarding, Build, and Review
4	Deployment
5	Usage, Monitoring, and Change Management

The updated MindForge AI risk taxonomy is included in the table below.

Risk Dimension	Risks Pertinent to Each Dimension	Risk Definition	Secondary Dimensions Impacted	Lifecycle Stages Impacted	ABS Top Risk	AI-Specific Element to Risk
Fairness & 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.	Robustness & Stability	2, 3	ABS Top Risk	Risk applies to AI as well as non-AI models.
Fairness & Bias	Adverse or inappropriate impact to individuals and groups	Models generate outputs that can be detrimental or inappropriate for individuals or groups.	-	5	ABS Top Risk	AI has the potential to adversely impact different individuals or social groups in complex ways that go beyond traditional modelling considerations around data representativeness; this can include a range of types of differential treatment, exclusion, or the perpetuation of stereotypes.
Ethics	Value misalignment	Gen AI services, outputs and/or uses do not align with corporate or societal values.	-	1	-	AI's non-deterministic behaviour and the autonomy it often operates with significantly enhance the risk that tools may not be aligned with the FI's intentions.
Ethics	Environmental sustainability impact	Environmental impact of running LLMs, especially increased carbon emissions which impact the corporate social responsibility and ESG outcomes for the organisation.	-	1, 3, 5	-	The significant environmental impact of large-scale AI operations, especially those associated with Gen AI, enhance existing risks around the environmental footprint of IT operations.
Ethics	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).	-	3, 5	-	AI's capability to behave with a degree of autonomy and to, in some cases, interact with customers (directly or indirectly) creates new risks of unintentional behavioural manipulation.
Ethics	Toxic and offensive outputs	Outputs produced contain harmful, offensive, hateful, discriminatory, violent, racist, sexist or nudity-related information.	Legal & Regulatory Cyber & Data Security	3, 5	ABS Top Risk	The potential toxicity in AI-generated content presents new risks around harm to the FI's reputation or its clients.
Accountability & Governance	Lack of AI risk awareness	Insufficient education or reskilling resulting in undertrained resources lacking awareness of the unique risks involved with Gen AI. Also, over-reliance on Gen AI can lead to an erroneous, biased, or misleading outputs being accepted without adequate scrutiny. The disposition of Gen AI to produce outputs that are not easily interpretable or verifiable amplifies this risk. This reliance can compromise the quality and integrity of decision-making processes, as well as erode trust in institutional operations.	-	1, 2, 3, 4, 5	ABS Top Risk	AI presents unique risks to an FI and, especially in the case of Gen AI, may directly impact more employees. The potential for misuse when employees are not aware of the specific risks of AI is an enhanced risk.

Risk Dimension	Risks Pertinent to Each Dimension	Risk Definition	Secondary Dimensions Impacted	Lifecycle Stages Impacted	ABS Top Risk	AI-Specific Element to Risk
Accountability & Governance	Lack of third-party accountability	Organisation has limited control or oversight over the development, modification and decision-making process for Gen AI models/services from third-party providers.	-	3, 5	-	AI can often require a greater reliance on third-party vendors than traditional software. Especially when using Gen AI, the complexity and non-determinism of AI systems can enhance existing risks around holding vendors accountable for the performance of their models.
Accountability & Governance	Lack of use case, data 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.	-	1, 2, 3, 4, 5	ABS Top Risk	Risk applies to AI as well as non-AI software.
Accountability & Governance	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.	-	3, 5	ABS Top Risk	AI has significant capabilities for automated or independent action, beyond those of traditional models or software. The importance of enhanced human oversight of AI and scrutiny of AI-made decisions is an enhanced risk.
Accountability & 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.	-	5	ABS Top Risk	Risk applies to AI as well as non-AI software.
Transparency	Unclear output accuracy	The level of accuracy needed for the proposed Gen AI use case outcome is not clear and cannot be validated.	-	3, 4, 5	-	Traditional AI shares existing modelling risks around accuracy. Gen AI and other AI that produce unstructured outputs, however, create a new risk that output accuracy will be challenging to determine.
Transparency	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 & Regulatory	2	-	AI models, especially large ones sourced from open-source providers or third-party providers, can present an enhanced risk of having been trained on data of unclear provenance.
Transparency	Lack of explainability	Challenge of understanding how the Gen AI modelling techniques influence model behaviour and outputs.	-	3, 5	-	AI's non-deterministic outputs, especially in "black box models", can create a new risk around the inability to justify or explain outputs.
Transparency	Anthropomorphism	The characteristic of Gen AI to mimic human characteristics in its output, enhancing the risk that users may find the outputs of Gen AI inappropriately convincing or may easily come under the impression that they are interacting with a human instead of a machine.	Cyber & Data Security	5	-	The capability of some AI systems, especially Gen AI, to behave in a way that simulates human behaviour is a new risk of AI use.

Risk Dimension	Risks Pertinent to Each Dimension	Risk Definition	Secondary Dimensions Impacted	Lifecycle Stages Impacted	ABS Top Risk	AI-Specific Element to Risk
Legal & 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	-	1, 2, 3, 4, 5	-	Risk applies to AI as well as non-AI software.
Legal & Regulatory	Unclear data ownership	Ownership of data used to train the Gen AI model and data created by the Gen AI model is unclear, leading to additional legal, commercial and privacy risks.	-	2	-	AI models trained on data sourced on the internet – especially AI models trained by third parties whose training data may be ambiguous – can introduce new risks around data ownership and IP.
Legal & Regulatory	Unauthorised data transfer and storage	Data is transported and stored on unauthorised systems as per the licensing terms or organisational policies.	-	3, 5	-	Risk applies to AI as well as non-AI software.
Legal & Regulatory	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 & Bias Ethics	2, 3, 4, 5	-	Risk applies to AI as well as non-AI software.
Legal & Regulatory	IP infringement	Data provided as input to a Gen AI system or product is used to create an output/content that violates IP rights owned by another individual, organisation, or entity.	-	3, 5	-	Gen AI systems with the ability to produce content introduce new risks around infringing on third-party IP.
Legal & Regulatory	Unavailability of IP protection	The outputs of Gen 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.	-	3, 5	-	Gen AI systems with the ability to produce content face new risks around the protection of that content, which may be unavailable in some jurisdictions.
Legal & Regulatory	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 & Bias Ethics	2, 3, 4, 5	-	Risk applies to AI as well as non-AI software.
Legal & Regulatory	Unclear data retention and deletion	Lack of clarity on the policy around retention of personal, sensitive, or confidential data of data subjects.	Ethics	2, 3	-	The extent to which AI models are trained on large and diverse datasets – especially the models powering Gen AI systems – can enhance the risk of failing to observe data retention rules.

Risk Dimension	Risks Pertinent to Each Dimension	Risk Definition	Secondary Dimensions Impacted	Lifecycle Stages Impacted	ABS Top Risk	AI-Specific Element to Risk
Robustness & 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.	Fairness & Bias Ethics Legal & Regulatory	3, 5	ABS Top Risk	Gen AI's capability to generate factually incorrect information is a new risk of the technology.
Robustness & Stability	Overconfidence	The characteristic of Gen 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 & Bias Transparency	3, 5	ABS Top Risk	Gen AI's tendency, without proper controls, to present information or decisions without properly qualifying its confidence in its predictions is a new risk.
Robustness & Stability	Training data or inputs not fit for purpose	Training data used in model is 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 & Bias	3, 5	-	Risk applies to AI as well as non-AI models.
Robustness & Stability	Lack of continuous monitoring	Absence of ongoing and systematic surveillance on how Gen 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.	-	3, 4, 5	-	Risk applies to AI as well as non-AI models.
Robustness & Stability	Insufficient data quality	Low-quality or noisy data used for training could result in poor model performance, increased debugging efforts, and higher development costs. Likewise, extensive use of synthetic data could result in data sets underexposed to noise and real-world complexity which might lead to reduced model performance when exposed to real-life data.	-	1, 2	-	Risk applies to AI as well as non-AI models.
Robustness & 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.	-	5	-	While model staleness is an existing model risk, the large, unstructured, and sometimes ambiguous datasets used in AI – especially in Gen AI – make model staleness challenging to measure.

Risk Dimension	Risks Pertinent to Each Dimension	Risk Definition	Secondary Dimensions Impacted	Lifecycle Stages Impacted	ABS Top Risk	AI-Specific Element to Risk
Robustness & Stability	Insufficient model accuracy/ soundness	The model outputs are inaccurate or does not meet the performance thresholds required to ensure fit for purpose.	-	3, 4, 5	ABS Top Risk	Risk applies to AI as well as non-AI models.
Robustness & Stability	Model degradation from unexpected use	A wider range of unexpected usage patterns due to the broad capabilities of Gen AI models create outcome instability or unexpected failure modes.	-	3, 4, 5	ABS Top Risk	The use of AI systems that continuously learn or evolve, especially when those are Gen AI systems whose users have the opportunity to interact directly with the system through prompting, enhances model degradation risks if user interactions are unexpected.
Robustness & Stability	Inadequate operational resilience	Operational resilience or service continuity plans increase in complexity due to the broad set of services and capabilities of Gen AI.	-	4, 5	-	Risk applies to AI as well as non-AI software.
Robustness & Stability	Unmet architectural requirements	Inadequate architectural requirements due to technology, cost or people constraints, leading to technical debt and hindering the scalability, robustness and long-term viability of the Gen AI system.	-	3, 4, 5	-	Risk applies to AI as well as non-AI software.
Robustness & Stability	Lack of reproducibility	Models that have the same parameters and identical inputs may generate different outputs. This causes challenges to reproduce a specific output and determine the accuracy in the variations of the output.	-	5	-	AI's non-deterministic characteristics, which impact the reproduction of outputs or model states, can enhance existing risks around the accuracy or validity of results.
Cyber and Data Security	Unintentional inappropriate or illegal use	Consumers or employees use Gen AI for inappropriate or illegal activities unintentionally with liability remaining with the FI.	-	5	-	AI systems that directly interact with users, especially Gen AI systems, enhance conduct risks by presenting additional opportunities for misuse.
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.	-	2, 3, 4, 5	-	The use of AI systems that continuously learn or evolve, especially when those are Gen AI systems where users have the opportunity to interact directly with the system through prompting, enhances adversarial manipulation risks through data poisoning.
Cyber and Data Security	Adversarial model manipulation	Deliberate manipulation of a Gen 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.	-	3, 5	-	The use of AI systems that continuously learn or evolve, especially when those are Gen AI systems where users have the opportunity to interact directly with the system through prompting, enhances adversarial manipulation risks.

Risk Dimension	Risks Pertinent to Each Dimension	Risk Definition	Secondary Dimensions Impacted	Lifecycle Stages Impacted	ABS Top Risk	AI-Specific Element to Risk
Cyber and Data Security	Prompt injection	The use of carefully designed prompts to encourage a Gen 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.	-	3, 5	-	Prompt injection is a new risk for FIs that use Gen AI systems.
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.	-	5	-	Risk applies to AI as well as non-AI models.
Cyber and Data Security	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.	-	2, 3, 4, 5	-	AI systems that directly interact with users, especially Gen AI systems that take in user-generated prompts, can enhance the risk of data leakage.
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 Gen AI models often have larger attack surfaces.	-	5	-	AI systems that directly interact with users, especially Gen AI systems, enhance the risk that sophisticated users can conduct model inference attacks.

C. Relevant Global AI Governance Frameworks

This Handbook is intended to support the governance of AI in the context of other requirements, frameworks, and governance. It is meant to enhance, and not supplant, existing industry practices, regulations, and norms.

In October 2024, the consortium identified several relevant AI governance and risk management frameworks that were each taken into consideration when drafting this Handbook. They have been segmented into three categories based on the Handbook's approach to incorporating them: Build, Highlight, and Incorporate.

Two key AI governance and risk management frameworks – the FEAT Principles and the MAS Proposed Guidelines on Artificial Risk Management – were especially foundational in the development of this Handbook. A full mapping of this Handbook's Considerations to these frameworks is provided in Appendix D.

"Build" includes the pre-existing body of AI governance and risk management guidance in Singapore's financial sector. The key concepts, considerations, and provisions of these frameworks are integrated directly into this Handbook and form the foundation for this work.

The Build frameworks are:

- FEAT Principles (see Appendix D).
- MAS Proposed Guidelines on Artificial Intelligence Risk Management (see Appendix D).
- Association of Banks in Singapore (ABS) Standing Committee on Data Management (SCDM) Handbook on Generative AI Guardrails in Banking
- MAS Thematic Review on Artificial Intelligence (AI) Model Risk Management
- MindForge Phase 1 Document: Emerging Risks and Opportunities of Generative AI for Banks
- Veritas Methodology

"Highlight" includes frameworks that organisations in Singapore may be expected to follow, and which may continue to evolve independently of this Handbook. Where pertinent, this Handbook refers to these frameworks and highlights the importance of complying with them, but does not reproduce their contents in this text.

The Highlight frameworks are:

- ABS Cloud Computing Implementation Guide
- IMDA Model AI Governance Framework for Agentic AI
- MAS Guidelines on Fair Dealing – Board and Senior Management Responsibilities for Delivering Fair Dealing Outcomes to Customers
- MAS Guidelines on Outsourcing (Banks)
- MAS Guidelines on Risk Management Practices – Internal Controls
- MAS Guidelines on Risk Management Practices – Technology Risk
- MAS Information Paper on Culture and Conduct Practices
- MAS Information Paper on Cyber Risks Associated with Deepfakes
- MAS Information Paper on Cyber Risks Associated with Generative Artificial Intelligence
- MAS Notices FSM-N06 and FSM-N22 Cyber Hygiene
- MAS Technology Risk Management Guidelines
- PDPC Advisory Guidelines on use of Personal Data in AI Recommendation and Decision Systems
- PDPC Guide to Developing a Data Protection Management Programme

“Incorporate” includes frameworks related to AI that represent widely accepted, influential norms in the industry, especially globally. These frameworks are drawn upon for lessons learned, and select, relevant considerations inspired by these frameworks may be incorporated into the Handbook. These frameworks are not necessarily incorporated in their full extent, and the Handbook is not necessarily fully aligned with their provisions.

The Incorporate frameworks are:

- AI Verify Foundation Model AI Governance Framework (this framework was given special emphasis because of its importance to the AI governance ecosystem in Singapore)
- Bank of England Prudential Regulatory Authority (PRA) Supervisory statement SS1/23: Model risk management principles for banks
- Department of Industry, Science and Resources (Australia) Voluntary AI Safety Standards
- EU AI Act (and associated documents)
- ISO/IEC 42001:2023 Information technology — Artificial intelligence — Management system
- NIST AI Risk Management Framework (and associated documents)
- Board of Governors of the Federal Reserve System SR 11-7: Guidance on Model Risk Management

D. References to Key Local AI Governance Frameworks

This Handbook was designed to support two key AI governance and risk management frameworks in Singapore: the FEAT Principles and the proposed MAS Guidelines on Artificial Intelligence Risk Management. A discussion of the relationship of this Handbook to those frameworks is provided in Subsection 1.1.

This appendix provides a brief overview of the context of each framework and a table mapping their provisions to the Considerations in this Handbook.

Principles to Promote Fairness, Ethics, Accountability and Transparency (FEAT) in the Use of Artificial Intelligence and Data Analytics in Singapore's Financial Sector

The 14 FEAT Principles are the conceptual and ethical foundation of the Veritas Initiative (2020-2023) and Project MindForge (2023-). Issued by MAS in 2018, they have earned widespread adoption in FIs around the world and played an important role in the global shift towards responsible AI use.

The field of AI governance has matured significantly since the FEAT Principles were published in 2018, growing both in breadth and depth. The breadth of AI governance has grown as new subjects, such as legal compliance, risk management, and cybersecurity have been increasingly included in the scope of AI governance activities; its depth has increased as new risks and new risk mitigation approaches have emerged. FIs report that they find it challenging to apply the FEAT Principles directly to Gen AI and Agentic AI in particular, given that these technologies require an expansive, use case-oriented approach to governance that goes beyond FEAT's model-oriented view.[26]

The FEAT Principles serve as a clear foundation for all AI use in the industry. The scope of this Handbook, however, is much more expansive than those Principles, and reflects the industry's changed understanding of AI governance over the past seven years. The mapping below illustrates the Handbook's increased breadth; several of its Considerations do not directly correspond to FEAT Principles, and others address FEAT Principles in addition to other subjects.

[26] This shift was described in more detail in Emerging Risks and Opportunities of Generative AI for Banks, the outcome document from Phase 1 of Project MindForge. It is available at https://www.mas.gov.sg/schemes-and-initiatives/project-mindforge

Proposed MAS Guidelines on Artificial Intelligence Risk Management

MAS released a Consultation Paper on Guidelines on Artificial Intelligence Risk Management in November 2025. The proposed Guidelines on Artificial Risk Management as they appear in the November 2025 Consultation Paper are mapped, at a high level, to the Considerations in this Handbook per the table below.

Mapping Frameworks to the AI Risk Management Handbook

MindForge AI Risk Management Considerations	FEAT Principles	Proposed MAS Guidelines on Artificial Intelligence Risk Management
Consideration 1. Ensure that an AI governance operating model is clearly defined by leveraging and, as needed, uplifting the roles and capabilities of existing enterprise functions including relevant roles from the Board, Senior Management, and operational governance, with sufficient operating effectiveness measures in place to support them (1.2).	Principle 7: Accountability Principle 9: Accountability	2. AI Oversight
Consideration 2. Ensure that governance documents define key AI-related concepts, processes, and responsibilities, and that they remain up-to-date and effective in supporting all aspects of the FI's approach to AI governance and risk management (2.1).	Foundational to implementing all 14 Principles.	2. AI Oversight 3. Key AI Risk Management Systems, Policies and Procedures
Consideration 3. Enhance the organisational risk framework and risk appetite to include enterprise risks, strategies, and key risk indicators (KRIs) that track, monitor, and mitigate AI-specific risks (2.2).	Principle 6: Ethics	2. AI Oversight 3. Key AI Risk Management Systems, Policies and Procedures
Consideration 4. Uplift existing procurement and third-party risk management activities to address AI-specific risks, including disclosure templates, vendor assessment and procurement practices, change detection and notification, contracting practices, and ensure that teams have access to relevant expertise in AI (2.3).	Principle 8: Accountability	4. AI Life Cycle Controls
Consideration 5. Ensure that a framework is in place to manage the risks of each AI use case. This includes defining a risk materiality assessment approach, implementing a framework for inherent and residual AI risk assessments, applying controls that are commensurate with the risks identified, and conducting pre- and post-deployment AI-specific reviews as appropriate (2.4).	Principle 3: Fairness Principle 4: Fairness Principle 5: Ethics	3. Key AI Risk Management Systems, Policies and Procedures 4. AI Life Cycle Controls
Consideration 6. Ensure that core AI-specific information on AI use cases is recorded in an inventory and ensure that a process is in place to maintain the AI inventory, so that information about new, updated, or decommissioned AI use cases is reflected accurately (2.5).	Does not correspond directly to an existing FEAT Principle.	3. Key AI Risk Management Systems, Policies and Procedures

MindForge AI Risk Management Considerations	FEAT Principles	Proposed MAS Guidelines on Artificial Intelligence Risk Management
Consideration 7. Assess the AI use case to ensure that the intended use is compatible with ethical, regulatory, and organisational standards, and determine the level of governance to be applied to the use case based on its inherent or expected risk materiality (3.1).	Principle 1: Fairness Principle 5: Ethics	4. AI Life Cycle Controls
Consideration 8. Evaluate whether the intended use of data in the AI use case is compatible with ethical, regulatory, and organisational standards (3.2).	Principle 2: Fairness	4. AI Life Cycle Controls
Consideration 9. Adopt appropriate data management practices that address risks and limitations when processing data for AI use cases (3.2).	Does not correspond directly to an existing FEAT Principle.	4. AI Life Cycle Controls
Consideration 10. Evaluate incremental AI-specific risks as part of the onboarding of third-party AI products and services within an AI use case (3.3).	Principle 12: Transparency Principle 13: Transparency Principle 14: Transparency	4. AI Life Cycle Controls
Consideration 11. Ensure that the AI use case is built with appropriate guardrails and relevant metrics for effective performance and risk management (3.3).	Does not correspond directly to an existing FEAT Principle.	4. AI Life Cycle Controls
Consideration 12. Conduct thorough testing and review prior to deployment to assess AI-specific risks and ensure that appropriate guardrails, controls, and governance have been observed (3.3).	Principle 3: Fairness	4. AI Life Cycle Controls
Consideration 13. Develop monitoring and contingency plans for the use case prior to its deployment, and consider risk-informed deployment options (3.4).	Does not correspond directly to an existing FEAT Principle.	4. AI Life Cycle Controls
Consideration 14. Conduct ongoing monitoring of the AI use case and its usage to ensure that it remains fit for purpose over time (3.5).	Principle 3: Fairness Principle 10: Accountability Principle 11: Accountability	4. AI Life Cycle Controls
Consideration 15. Capture changes to AI use cases or their components to maintain traceability and ensure that changes with a material impact on risk are reviewed and approved through an effective change management process (3.5).	Does not correspond directly to an existing FEAT Principle.	4. AI Life Cycle Controls

MindForge AI Risk Management Considerations	FEAT Principles	Proposed MAS Guidelines on Artificial Intelligence Risk Management
Consideration 16. Ensure that practices are in place to equip employees with the necessary AI governance and risk management skills, knowledge, and AI culture, while ensuring that teams involved in AI governance and risk management function are sufficiently representative (4.1).	Does not correspond directly to an existing FEAT Principle.	5. AI Capability & Capacity
Consideration 17. Support AI deployment by ensuring that supporting infrastructure is fit for purpose (4.2).	Does not correspond directly to an existing FEAT Principle.	5. AI Capability & Capacity

E. AI Card Template

This AI Card Template is designed to be a helpful reference and starting point for FIs in requesting useful information on AI products or services from third parties. FIs may further customise, adapt, and specify this template for use in their contexts and that of their third-party providers.

This template supports AI disclosures, which are discussed in Subsection 2.3.

1. **General Information**
   a. Basic metadata – name, version, date of last update, and developer/provider name and contact.
   b. License
   c. AI type(s) (select one):
      (i) Diagnostic
      (ii) Predictive
      (iii) Generative
      (iv) Agentic
   d. If generative or agentic – input/output modalities (select all that apply for each of inputs and outputs):
      (i) Text
      (ii) Image
      (iii) Video
      (iv) Audio
      (v) Other (specify)

2. **Purpose and Usage**
   a. Brief explanation of the intended types of uses, users, and domains (free text).
   b. Information on how to use the AI model/system, including guidance on input/output data types, how to integrate and operate the AI model/system, intended/recommended level of human involvement in use, and guidance on the explanation/interpretation of results.
      (i) E.g. Sample input/output data.
      (ii) E.g. Technical integration instructions.
      (iii) E.g. Maximum input size and context window (for LLMs).

3. **Techniques and Development**
   a. Brief description of the main AI techniques used (free text).
   b. High-level description of architecture and components (if applicable). This should include a high-level description of the types of AI model(s) used, while not necessarily describing their exact characteristics and number (free text).
   c. High-level information on external AI service calls, including endpoints, providers, and types of data (free text).

4. **Risks**
    a. List of known risks associated with the AI model/system (select all that apply from the MindForge AI Risk Taxonomy in Appendix B).
    b. Description of mitigation strategies taken by the developer for addressing each selected risk (free text).
    c. Guidance on appropriate use in order to manage risks (free text).

5. **Datasets**
    a. Preferably, detailed dataset documentation for all training and evaluation data used for each model and for the system's operation, but, at minimum, information that may be useful for assessing key data risks, including:
        (i) A summary of the generic types of training and evaluation data and high-level types of sources used (free text).
        (ii) A summary of the pre-processing steps used (free text per dataset).
        (iii) A summary of any known private or personal information included (free text per dataset
        (iv) Details of how the training and evaluation datasets were assessed for representativeness across relevant features/demographic groups (free text per dataset).

6. **Evaluation and Testing**
    a. Metrics and benchmarks for assessing the risks of the AI model/system, and justification for their selection. If metrics and benchmarks are not industry-standard, the provider may also link to relevant papers or resources to reproduce them (free text).
    b. Performance of the model or system on suggested AI-specific risk metrics and benchmarks, with sufficient description of testing methodology (or link to description) to reproduce results if applicable (free text per metric).

7. **Cybersecurity and Data Protection**
    a. Description of data from the AI model/system that will be shared with the AI provider, including data residency/sovereignty (free text).
    b. Description of how data collected by the provider will be handled (free text).
    c. Metrics and benchmarks specifically related to cybersecurity risks, as well as performance on those benchmarks (free text per metric).
    d. Relevant security and data protection attestations, such as a SOC 2 report (free text).

8. **Pre-Determined Changes**
    a. A list of expected changes to the model, system, or dataset by the provider, including the frequency, subject, and expected impact of each change (free text).

9. **Standards and Certifications**
    a. A list of standards, certifications, or audit results that the model or system has obtained, along with supporting information such as links (free text).

Optional Addendum: Components and Architecture (System)

a. A list of components, such as models or other software, that make up the system, and a high-level description of their architecture. Where appropriate, integrated software components can include their names, version numbers, and links to their documentation or sources.
b. A data flow diagram describing the relationship between components and services accessed by the system.

F. Library of AI Metrics

Metrics are an important part of operationalising AI governance and risk management; below is a table of metrics for FIs to consider. This list is not exhaustive; it represents a snapshot, at a point in time, of the state of the industry across select dimensions. The consortium's technology partners played a leading role in compiling this list.

FIs can reference the ongoing literature and evolving industry practices, as well as ongoing work in the field of monitoring conducted by Singapore's Infocomm Media Development Agency (IMDA) and the AI Verify Foundation, to identify further metrics that may be useful managing AI-specific risks.

Metric	Description
Fairness and Bias	
Average Absolute Odds Difference (AAOD)	AAOD combines differences in false positive and true positive rates across two groups (a and b) to provide a comprehensive view of classification disparities. It is given by: 0.5 x (|FPRa - FPRb| + |TPRa - TPRb|) Where FPR is False Positive Rate and TPR is True Positive Rate.
Disparate Impact Ratio (DIR)	DIR measures the ratio of favourable outcome probabilities between groups, aligning with legal frameworks like the "80% rule" in US employment law (80% rule helps employers and EEOC determine if a selection process has a disproportionate negative impact on protected groups). It is given by: P(Y=1|A=0) / P(Y=1|A=1) Where Y is an outcome and A is a sensitive variable.
Equal Opportunity Difference (EOD)	EOD focuses on equality of opportunity by measuring the ratio of differences in true positive rates across groups. It is given by: P(Y=1|Y*=1, A=0) - P(Y=1|Y*=1, A=1) Where Y is an outcome, Y* is the true outcome, and A is a sensitive variable.
LLM-as-a-judge bias scoring	LLM bias can be assessed using an LLM "as a judge", with frameworks provided by several open-source resources. Typically, the "judge" LLM is given a set of instructions for evaluating a corpus of outputs from the LLM targeted for evaluation. LLM-as-a-judge evaluation can output a score for bias (using a technique like G-Eval), but depending on the choice of judge, LLM-generated scores can display arbitrariness. Alternative approaches include QAG scores or other methods that use binary classification to rate outputs as either "biased" or "not biased", taking the proportion of biased outputs as a score. It is given by: (Biased outputs / Total outputs) x 100%
Separation	Separation measures the ratio of an outcome given a sensitive attribute to the probability of that outcome overall. A separation value closer to 1 indicates that predictions are less influenced by the sensitive variable. It is given by: P(Y=1|A=1) / P(Y=1) Where Y is an outcome and A is a sensitive attribute.

Metric	Description
Socio-cultural bias	Socio-cultural bias is when an LLM's behaviour varies in undesired ways when the context of customer interactions changes (e.g. different tone, decisions due to language/age/race/religion of customer, including indirect information through name). Though underlying data does not exhibit inherent bias, the Gen AI tooling may. It is typically measured with conversation flows with different customer context, including language, customer name, and background. Socio-cultural bias can be measured as a proportion of biased outputs over total outputs or using measures of semantic distance like cosine similarity.
Statistical Parity Difference (SPD)	SPD measures demographic parity by comparing the probability of a positive outcome between two groups. It is given by: P(Y=1|A=0) - P(Y=1|A=1) Where Y is an outcome and A is a sensitive variable.
Sufficiency	Sufficiency measures disparities in outcomes given a sensitive attribute. A sufficiency value closer to 1 indicates that the outcome is not likely to be influenced by the sensitive variable. It is given by: P(Y) / P(Y|A) Where Y is an outcome and A is a sensitive variable.
Toxicity Score	Similar to other metrics on bias, the toxicity score is the proportion of toxic outputs. Toxicity can be quantified using an LLM as a judge, by regex matching, or by semantic scoring. It is given by: (Number of toxic opinions / Total number of opinions extracted) x 100%
Accountability and Governance	
Compliance Consistency Score (CCS)	A normalised score (0 to 1) reflecting the overall quality and consistency of policy adherence.
Policy Override Frequency (POF)	A measure of the extent to which policies are adhered to. It is given by: (Number of policy deviation instances / Total number of prompts) x 100%
Transparency	
Feature attribution	Measures the change in contribution of features to a model's prediction compared to a baseline. Common measures for doing so include: • LIME value (Local Interpretable Model-agnostic Explanations) • SHAP value (SHapley Additive exPlanations)

Metric	Description
Robustness and Stability	
Accuracy	Accuracy measures the overall correctness of predictions across all classes. It is given by: (TP + TN) / (TP + TN + FP + FN) Where TP is True Positives, TN is True Negatives, FP is False Positives, and FN is False Negatives.
BLEU (Bilingual Evaluation Understudy)	A longstanding metric for quantifying the similarity of translated text to a translation by a human expert; it is useful for quantifying a model or system's translation accuracy. BLEU can be calculated on a body of customised domain-specific translations or on a benchmark reference set of translated pairs. METEOR is an industry alternative to BLEU that aims to account for variances like word order.
Consistency check	Hallucinated facts in LLM-generated responses are less likely to be repeated between responses than facts that are supported by facts. With this fact in mind it is possible to perform a "zero-resource" check – that is, without requiring consultation of additional sources. A consistency check involves assessing each sentence of an output against several other sample outputs generated from the same prompt; assessment typically involves a similarity metric like BERTScore. A lower average similarity to sentences in the other samples indicates a higher likelihood of hallucination.
F1 Score	F1 Score is the harmonic mean of precision and recall, providing a balanced measure when classes are imbalanced. It is given by: 2 x (Precision x Recall) / (Precision + Recall)
Input feature data drift	Measures the distribution of input feature values compared to a baseline data distribution. Categorical: Boolean, string, L-Infinity, Jensen Shannon Divergence Numerical: Float, integer, Jensen Shannon Divergence
Output prediction data drift	Measures the model's predicted data distribution compared to a baseline data distribution. Categorical: Boolean, string, L-Infinity, Jensen Shannon Divergence Numerical: float, integer, Jensen Shannon Divergence
Precision	Precision measures the proportion of positive identifications that were correct. It is given by: TP / (TP + FP) Where TP is True Positives, FP is False Positives.
Prediction error percentage	A high error rate might indicate an issue with the model or with the requests to the model. It is given by: (Incorrect Predictions / Total Predictions) x100%

Metric	Description
Recall	Recall (also called sensitivity) measures the proportion of actual positives that were correctly identified. It is given by: TP / (TP + FN) Where TP is True Positives, FN is False Negatives
Relevance to prompt	The cosine distance or BERTScore of the prompt to the generated output. This gives an indication of how relevant the response is to the prompt and can be useful in detecting off-topic outputs. The number or proportion of outputs with a low relevance can also be tracked as a metric.
SARI (System output Against References and against the Input sentence)	A longstanding metric for quantifying the similarity between a simplified/summarised output and a source text. SARI can be calculated on a body of customised domain-specific summarisations or on a benchmark reference set of texts and reference summarisations.
Topic and task adherence	A metric for checking if an LLM system stays on topic. This risk metric can also be used to detect if corporate Gen AI tools are used for personal tasks provided that those tools are effective in refusing off-topic prompts. It is given by: (Prompts Refused / Total Prompts) x100%
Cyber and Data Security	
Injection Similarity	Prompt injection attempts can be detected by calculating semantic similarity measures – like cosine distance – for any given prompt against a library of templates of known prompt injection or jailbreak formats. There are several open-source resources that catalogue known prompt injection attempts; the currency of these templates is important as prompt injection formats continue to evolve. The number of prompts that are similar to a known injection format (cosine similarity >0.85 is considered likely to be harmful, and >0.5 may merit investigation) can be tracked as a metric.
Injection Success Rate (ISR)	A score indicating the success, in either a controlled environment or in production, of attempts to inject adversarial prompts into the system. It is given by: (Number of successful injection attempts / Total number of injection attempts) x 100%
PII Detection Rate	The outputs of an LLM can be assessed for the presence of PII, with the proportion of outputs containing PII quantified in the aggregate or by PII type. A range of PII detectors, both open-source and proprietary, exist in the market today. These may either use an LLM or regex matching (or a combination of the two) to detect data resembles PII categories like names, addresses, phone numbers, password, or credentials. It is given by: (Prompts containing PII / Total number of prompts) x 100%

Metric	Description
Prompt Refusal	The number or proportion of prompts that are rejected by a Gen AI system's guardrails can be detected by an LLM judge or regex expression and tracked as a metric. This can indicate both the effectiveness of the FI's guardrails, by tracking that they continue to reject prompts at a sustained rate, and can serve as an indicator of a potential attack on the system. It is given by: (Prompts rejected / Total number of prompts) x 100%
Prompt Sanitisation Rate (PSR)	This metric is a complement to ISR and indicates the success rate of guardrails against prompt injection. It is given by: (Number of successfully neutralised injection attempts / Total detected injection attempts) x 100%
Red-teaming attempts	Red-teaming attempts are intentional attempts to get LLMs to leak data, including system prompts, or output potentially harmful and embarrassing content. It is typically used with benchmark prompt injection datasets or long context prompt attacks. Specifically for our region, we can include multilingual prompts and conversations. It is given by: (Successful attacks / Total attacks) x 100%
Total Injection Vulnerability Score (TIVS)	A composite metric incorporating the above (and other) measurements to provide a comprehensive view of prompt injection vulnerability. Each organisation can select the measurements for calculating TIVS in their context.
General	
Accelerator average duty cycle.	The average fraction of time over the past sample period during which one or more accelerators were actively processing.
Accelerator memory usage	The amount of memory allocated by the deployed model replica.
CPU utilisation	The fraction of CPU allocated by the feature store and currently in use by online storage. This number can exceed 100% if the online serving storage is overloaded.
CPU utilisation – hottest node	The CPU load for the hottest node in a feature store's online storage.
Latency	The total time that an online serving or streaming ingestion request spends in the service.
Memory usage	The amount of memory allocated by the deployed model replica and currently in use.
Model latency	The time spent performing computation.
Network bytes received	The number of bytes received over the network by the deployed model replica.
Network bytes sent	The number of bytes sent over the network by the deployed model replica.

Metric	Description
Node count	The number of online serving nodes for a feature store.
Offline storage write for streaming write	The number of streaming write requests processed for the offline storage.
Online serving throughput	The throughput for online serving requests in MB/s.
Overhead latency	The total time spent processing a request, outside of computation.
Predictions per second	The number of predictions per second across both online and batch predictions.
Queries per second	The number of online serving or streaming ingestion queries that a feature store handles.
Replica count	The number of active replicas used by the deployed model.
Replica target	The number of active replicas required for the deployed model.
Request size	The request size by entity type in a feature store.
Streaming write to offline storage delay time	The time elapsed (in seconds) between calling the write API and writing to the offline storage.
Total latency duration	The total time that a request spends in the service, which is the model latency plus the overhead latency.
Total offline storage	Amount of data stored in the feature store's offline storage.
Total online storage	Amount of data stored in the feature store's online storage.

G. Library of AI Guardrails

Guardrails are built-in safety constraints to ensure that an AI system operates as desired and avoids harmful outcomes (see the International AI Safety Report 2025); they are a subset of the broader set of controls that FIs have access to and which apply across the full range of the AI lifecycle. This list is not exhaustive and represents a perspective taken at a single point in time.

FIs can reference the ongoing literature and evolving industry practices to identify guardrails as both AI technologies and AI governance and risk management continue to evolve. Relevant resources include OWASP, for the documentation of security-specific guardrails, and ongoing industry-specific work.

Guardrail	Implementation Examples
Fairness and Bias	
Algorithm re-selection	**All AI** The practice of selecting a different training algorithm that performs better, ceteris paribus, on fairness criteria.
Decision threshold adjustment	**All AI** Adjustment of the decision thresholds for different groups to ensure that each group receives fair treatment.
Hyperparameter tuning	**All AI** The practice of modifying system hyperparameters. Different types of AI have a wide variety in such parameters. In Gen AI applications that involve Large Language Models, common hyperparameters are temperature, which represents a model's "creativity"; k, which represents the range of possible responses; and p, which represents the likelihood of choosing less-likely responses. In applications where random fairness-violating responses are generated, these three values can each be reduced.
Input/output filtering	**Gen AI** The addition of filters to a Gen AI use case to detect biased or toxic content. Input filtering involves checking prompts for toxic language (often using regex matching or agentic monitoring), checking if they are on topic, or assessing whether they can induce the model to behave inappropriately. Output filtering involves checking a model's outputs to determine whether it contains toxicity (often using a word list or sentiment measure), is unrepresentative, or otherwise contains content which meets harm thresholds (often using an agentic architecture).
In-processing techniques	**Traditional AI** Manual adjustments to a model's loss function, the addition of variables, or the use of other algorithms that change the model or its decision-making process to prioritise fairness. These include adversarial debiasing, reductions grid search, automatic hyperparameter tuning, or the addition of fairness regularisers to a regression.
Model customisation	**All AI** Model customisation in pretrained models, especially Gen AI, refers to fine-tuning and reinforcement learning. Fine-tuning adds layers to the model based on data, such as additional data representing individuals or groups who may be discriminated against. Reinforcement learning uses human review to reward or punish the model for harmful outputs, modifying the model's parameters towards responses that meet fairness criteria.

Guardrail	Implementation Examples
Model separation	**All AI** The demarcation of different models for handling different types of uses, such as by splitting customers into two types and training models to address each distinctly. This is useful for cases where customer types are meaningfully distinct and a single model is less effective at fairly treating each.
Post-processing techniques	**Traditional AI** The modification of model outputs to improve AI performance on fairness characteristics. These include output adjustment to generate equal odds for different groups or constrained balanced accuracy for supervised learning.
Prompt design	**Gen AI** The addition of instructions to user prompts. These can include instructions to represent groups fairly, reject requests to create discriminatory content, or instructions to include different individuals or groups in outputs.
Ethics	
Algorithm selection for power efficiency	**All AI** AI training algorithms can have significantly different power requirements for certain types of tasks. Choosing the right algorithm for a specific task – and balancing both performance and efficiency – is an important part of the build process that can have significant impacts on the power consumption and environmental impact of the training or fine-tuning process.
Conduct an ethical design assessment in onboarding	**All AI** When onboarding systems or models, FIs can assess the extent to which their design, functioning, and use supports enterprise values and ethical principles. Third-party vendors can provide information on the ethical principles used in the system's design, and these can be assessed against the FI's own ethics.
Content Moderation	**Gen AI** FIs can assess the outputs of a Gen AI system for ethical compliance before sharing them with the user. This assessment can take the form of deterministic rules, but is typically implemented using an LLM (for text outputs) which assesses whether the output violates the FI's policies or principles. Outputs that fail to pass can be re-generated or can result in an error message. This approach often uses the same techniques as "input/output filtering", which is included under "Fairness & Bias".
Test prioritisation	**All AI** FIs can prioritise the sequence of tests that are performed on an AI to improve the efficiency of the testing process, reduce the number of re-training runs that need to be performed, and potentially reduce the number of tests that need to be performed. Carefully sequencing, de-duplicating, and triaging tests, typically by performing the most rigorous and comprehensive tests first, can minimise the need for power-intensive re-testing and re-training of an AI.

Guardrail	Implementation Examples
Use of pre-trained models	**All AI** By leveraging pre-trained or pre-tuned AI models, FIs can limit or completely avoid the requirement to perform their own training or find-tuning activities, with potentially significant power savings as a consequence.
Transparency	
Chain-of-thought prompting	**Gen AI** LLMs can, through grounding or prompting, be induced to produce outputs in a stepwise fashion. This encourages the system to illustrate reasoning steps in generating an output and can significantly improve the explainability of those outputs by clarifying the assumptions underlying it; it has the side effect of also improving the accuracy of some kinds of outputs. Chain-of-thought prompting is an area of ongoing research and is known to have limitations.
Confidence scoring	**All AI** There are a variety of measures of the degree to which an AI is "confident" in its outputs. Perplexity is a common measure of confidence for NLP and is especially useful in LLMs. Confidence levels can be used by builders to assess the usefulness and effectiveness of their AI, and AI outputs with low confidence scores can trigger different behaviours.
Counterfactual explanations	**All AI** Counterfactual explanations test how AI might behave under different input conditions, creating a form of explainability by suggesting which features or factors may have contributed to an output. Builders can include counterfactual explainability techniques in use cases that they create, particularly in traditional AI.
Inherent interpretability	**All AI** Some types of models, especially those with low dimensionality or that create outputs as a weighted sum of inputs (such as a linear regression), are inherently interpretable. Models built using decision trees, decision sets, or rule-based classifiers are also considered inherently interpretable. The selection of such a model significantly enhances explainability.
Interpretability constraints	**Traditional AI** AI with interpretability constraints are systems whose characteristics have been modified to improve their interpretability, sometimes at the cost of a penalty to performance. Interpretability constraints include rule-based reasoning or measures to sparsify neural networks.
Post hoc interpretability techniques	**All AI** Post-hoc interpretability techniques support the analysis of black box models by attempting to assess how an output was created. These techniques vary in complexity and applicability to different model architectures. Such techniques consist of building surrogate models, generating Shapley values, or assessing deep learning models using gradient-based assessments.

Guardrail	Implementation Examples
Retrieval-augmented generation	**Gen AI** Retrieval-augmented generation (RAG) architecture supplements an LLM with other components that provide it with sources on which to base its output. Some LLMs are capable of citing sources contextually in-line throughout an output, which provides a user with clear traceability to sources and provides an opportunity to validate the use of those sources. This can improve the explainability of use case outputs.
System prompt design	**Gen AI** The addition of instructions to user prompts. These can include instructions to reject prompts where the system cannot produce a satisfactory answer or where source data does not match the type of output desired.
Programmable conversation controls	**Gen AI** Rule-based frameworks can be used to define explicit boundaries for multi-turn conversations, ensuring that outputs consistently remain within approved topics and organisational policies. Such rules can support the codification of compliance requirements as transparent, enforceable rules with clear logic.
Robustness and Stability	
AI onboarding using domain data	**All AI** Testing that focuses on performance on domain data, such as by testing the AI against internal data or against a finance-specific benchmark, can provide assurance in cases where training data is not transparent.
Fine-tuning	**All AI** For AI models based on neural networks, fine-tuning is the additional of neural "layers" based on additional data without adjusting the underlying parameters. Fine-tuning is a resource-efficient method for adjusting system performance characteristics and introducing domain-specific data without fully retraining the model. It is particularly useful for preparing models for unexpected use or introducing a wider variety of data or circumstances. Common fine-tuning datasets include: • Fact-checking datasets, which consist of pairs of prompts and correct responses. • Negative examples, which suggest outputs that a model should avoid. • Domain data, which consists of data relevant to the FI's expected usage.
Input filtering	**Gen AI** Gen AI use cases can be vulnerable to the ingestion of low-quality or malicious data through unexpected use or prompt injection attacks, which can degrade the foundation model's performance. Input filtering, which uses a rule-based engine or another AI to assess inputs for quality before approving them for use, can filter out some low-quality, inaccurate, malicious, or poorly formatted inputs that could cause degradation.

Guardrail	Implementation Examples
Model calibration	**Gen AI** Calibration in Gen AI is an approach to system architecture which involves re-running the model repeatedly and sampling responses to determine whether they are sufficiently similar. Calibration may include the addition of small amounts of variability to the prompt in each run.
Modular architecture	**All AI** Stability can be improved by adopting a modular architecture with discrete tasks performed by individual models and components. A modular architecture can be more robust against drift, because of the resilience of other components in the use case, and can be easier to maintain, because each component can be independently monitored and repaired.
Reinforcement learning	**All AI** For AI models based on neural networks, reinforcement learning involves generating outputs and then scoring those outputs for desirability (scoring by a human is Reinforcement Learning with Human Feedback, RLHF, and scoping by a machine is Reinforcement Learning with AI Feedback, RLAIF). These scores are used to adjust model weights to favour higher-scoring behaviours. Reinforcement learning can be used to enforce desirable behaviours when encountering new data types or unfamiliar situations, or can be used to improve accuracy or soundness, such as by favouring truthful responses.
Robustness testing	**All AI** Testing an AI against a range of input types, including by deliberately selecting noisy input data or by testing the AI against adversarial inputs. These can include assessing the AI against potential variations in its operating context, such as different market conditions than those that applied to its training data. This can be used to determine whether an AI is sufficiently robust in the face of predictable types of data drift.
Small model selection	**Gen AI** Smaller foundation models in Gen AI use cases have been observed to demonstrate improved stability and reliability, at the potential cost of decreased performance. Selecting a smaller model where appropriate, and where that model provides sufficient performance, can increase the stability and predictability in diverse usage contexts.
Weight regularisation and normalisation	**Traditional AI** Potential causes of instability or data drift in neural network-based use cases include overfitting or low rates of convergence. These can be addressed by regularising parameters, reducing the largest weights, or by normalising the inputs between layers, which can reduce model instability and encourage a more narrow, predictable band of outputs.
Synthetic evaluation datasets	**All AI** Generating controlled datasets with known properties can facilitate testing and validation for AI use cases. In particular, the generation of synthetic edge cases or other rare scenarios may enable testing under conditions that are underrepresented in real data.

Guardrail	Implementation Examples
Cyber and Data Security	
Jailbreak detection	**Gen AI** See “input filtering” under “Robustness and Stability”.
Penetration testing	**All AI** Penetration testing, often referred to as “pen testing,” is a proactive security assessment where authorised professionals simulate a cyberattack. The primary goal of a pen test is to identify vulnerabilities, misconfigurations, or weaknesses before malicious actors can exploit them. Pen testing involves employing real-world attack techniques, including exploiting known vulnerabilities and identifying potential zero-day threats, to test the robustness of a system’s security controls. The test is performed in a controlled environment to ensure that no harm is caused to the system or its users.
Red teaming	**All AI** Red Teaming is an advanced security testing process where ethical hackers simulate real-world cyberattacks to uncover vulnerabilities, assess organisational defences, and improve security operations. Unlike traditional penetration testing, it takes a holistic approach, evaluating systems, processes, and personnel. In AI, red teaming is used to identify and mitigate risks such as data leaks, biased or toxic outputs, and adversarial manipulation, ensuring models are safe, reliable, and ethical. This proactive method enhances security readiness, operational resilience, and compliance with regulatory standards while safeguarding both cybersecurity and AI system. It is particularly useful for Gen AI systems.
Role-based access controls	**All AI** These are access management mechanisms that restrict system permissions based on a user’s job role within an organisation. They ensure individuals can only access the data and functions necessary for their responsibilities, reducing the risk of unauthorised actions or data breaches.
Vulnerability assessment	**All AI** Vulnerability assessments are conducted to evaluate and prioritise security risks in software, applications, and networks. These are typically recurring, automated scans that search for known vulnerabilities in a system and flag them for review. Security teams use vulnerability assessments to quickly check for common flaws. The following methodologies are commonly applied: • System Analysis: Both dynamic and static analysis techniques are employed to examine the application. These methods uncover coding errors in systems, whether they are operational or dormant. • Dependency Vulnerability Scans: Attackers may exploit weaknesses in external code components, such as libraries and binaries. These scans help identify vulnerabilities tied to specific versions of these dependencies, enabling proactive mitigation.

H. MindForge AI Risk Management Checklist

- [ ] **Consideration 1.** Ensure that an AI governance and risk management operating model is clearly defined by leveraging and, as needed, uplifting the roles and capabilities of existing enterprise functions including relevant roles from the Board, Senior Management, and operational governance, with sufficient operating effectiveness measures in place to support them. (1.2)

- [ ] **Implementation Practice 1.** Embed additional responsibilities for AI governance and risk management, as required, in relevant Board and Senior Management roles.

- [ ] **Implementation Practice 2.** Ensure that operational governance functions have clear roles and responsibilities assigned to operationalise AI governance and risk management activities across the enterprise.

- [ ] **Implementation Practice 3.** Ensure that existing governance processes, forums, assets, and tools are updated to effectively enable AI governance and risk management.

- [ ] **Implementation Practice 4.** Ensure that sufficient operating effectiveness and horizon-scanning measures are in place to monitor and improve the AI governance and risk management operating model over time.

- [ ] **Consideration 2.** Ensure that governance documents define key AI-related concepts, processes, and responsibilities, and that they remain up-to-date and effective in supporting all aspects of the FI's approach to AI governance and risk management (2.1).

- [ ] **Implementation Practice 1.** Ensure robust conceptual foundations for AI governance and risk management by establishing AI principles, defining key AI-related concepts, establishing frameworks for effective AI identification, and continuously improving these foundations over time as necessary.

- [ ] **Implementation Practice 2.** Ensure that all aspects of AI governance and risk management are effectively institutionalised throughout the FI's governance documents, and that a process is in place to periodically review and reassess them.

- [ ] **Consideration 3.** Enhance the organisational risk framework and risk appetite to include enterprise risks, strategies, and key risk indicators (KRIs) that track, monitor, and mitigate AI-specific risks (2.2).

- [ ] **Implementation Practice 1.** Identify the new or enhanced risks of AI that are relevant to the enterprise and ensure that the enterprise risk taxonomy effectively captures them.

- [ ] **Implementation Practice 2.** Assess existing enterprise risk controls for their fitness in addressing AI-specific enterprise risks, and uplift those controls where gaps exist.

- [ ] **Implementation Practice 3.** Ensure that key risk indicators (KRIs) are in place to measure AI-specific risks and that relevant incidents, issues, or risk events are appropriately tracked and managed.

- [ ] **Implementation Practice 4.** Ensure that effective monitoring is in place to identify AI-specific risk events or breaches of KRI thresholds to a degree proportionate to the FI's risk appetite.

- [ ] **Consideration 4.** Uplift existing procurement and third-party risk management activities to address AI-specific risks, including disclosure templates, vendor assessment and procurement practices, change detection and notification, contracting practices, and ensure that teams have access to relevant expertise in AI (2.3).

- [ ] **Implementation Practice 1.** Define, based on relevant AI-specific risks, a proportionate level of disclosure to seek from third party providers of AI products and services, and a process for assessing disclosures.

- [ ] **Implementation Practice 2.** Ensure that processes and capabilities are in place for AI-specific risks to be evaluated at appropriate points in procurement, onboarding, and throughout the post-procurement lifecycle.

- [ ] **Implementation Practice 3.** Identify new or modified AI components or features in third party products and services already introduced into the FI's technology ecosystem.

- [ ] **Implementation Practice 4.** Consider whether contracts and licenses with third parties providing AI products and services are sufficient to clearly address AI-specific risks.

- [ ] **Implementation Practice 5.** Ensure that teams with AI-specific legal, technical, and risk-management skills are involved in procurement, contracting, onboarding, or other third-party risk management activities as appropriate.

- [ ] **Consideration 5.** Ensure that a framework is in place to manage the risks of each AI use case. This includes defining a risk materiality assessment approach, implementing a framework for inherent and residual AI risk assessments, applying controls that are commensurate with the risks identified, and conducting pre- and post-deployment AI-specific reviews as appropriate (2.4).

- [ ] **Implementation Practice 1.** Define levels of risk materiality for AI use cases based on criteria relevant to the FI's context.

- [ ] **Implementation Practice 2.** Define a process to assess the inherent risk materiality of AI use cases at the appropriate lifecycle stage, considering the fundamental characteristics of each use case.

- [ ] **Implementation Practice 3.** Define a process to evaluate the residual risk materiality of AI use cases prior to deployment, considering the established controls and guardrails.

- [ ] **Implementation Practice 4.** Identify, uplift, or create controls to be applied to each AI use case based on its risks and risk materiality.

- [ ] **Implementation Practice 5.** Define an approach for conducting an AI-specific review of AI use cases prior to deployment, confirming the risks identified, the use case's risk materiality, and the appropriateness of risk mitigations.

- [ ] **Implementation Practice 6.** Ensure that AI-specific reviews of AI use cases are conducted periodically post-deployment, with their frequency based on factors including the risk materiality of the AI use case.

- [ ] **Consideration 6.** Ensure that core AI-specific information on AI use cases is recorded in an inventory and ensure that a process is in place to maintain the AI inventory, so that information about new, updated, or decommissioned AI use cases is reflected accurately (2.5).

- [ ] **Implementation Practice 1.** Ensure that a form of AI inventory, designed in consideration of existing inventory systems and practices to be suitable and proportionate for the FI's context, is in place to capture a core set of AI-specific information on AI use cases.

- [ ] **Implementation Practice 2.** Ensure that processes are in place and that roles and responsibilities are defined such that the AI inventory is well-maintained and kept up to date.

- [ ] **Consideration 7.** Assess the AI use case to ensure that the intended use is compatible with ethical, regulatory, and organisational standards, and determine the level of governance to be applied to the use case based on its inherent or expected risk materiality (3.1).

- [ ] **Implementation Practice 1.** Establish ownership for the AI use case and ensure alignment with organisational standards and values for ethical and responsible AI use.

- [ ] **Implementation Practice 2.** Perform an inherent risk materiality assessment to determine the risk tiering of the AI use case and to guide proportionate governance efforts.

- [ ] **Implementation Practice 3.** Capture AI use case-related information in an AI inventory to enable transparency and support risk management.

- [ ] **Implementation Practice 4.** Design the AI use case to operate with a proportionate and practical level of human oversight.

- [ ] **Consideration 8.** Evaluate whether the intended use of data in the AI use case is compatible with ethical, regulatory, and organisational standards (3.2).

- [ ] **Implementation Practice 1.** Ensure that the use of data complies with ethical standards, regulatory requirements, and organisational policies or standards.

- [ ] **Implementation Practice 2.** Ensure that the use of any third-party data complies with intellectual property rules, contractual obligations, and licensing rights.

- [ ] **Consideration 9.** Adopt appropriate data management practices that address risks and limitations when processing data for AI use cases (3.2).
- [ ] **Implementation Practice 1.** Ensure that data used for the AI use case is fit for purpose.
- [ ] **Implementation Practice 2.** Justify the use of personal attributes in the AI use case.
- [ ] **Implementation Practice 3.** Document metadata and data sources related to the AI use case in accordance with organisational data management policies and regulatory expectations.
- [ ] **Implementation Practice 4.** Ensure that appropriate data access controls are implemented based on the nature of selected AI use case.
- [ ] **Implementation Practice 5.** Establish clear ownership of any derived or transformed data to be used in the AI use case.
- [ ] **Implementation Practice 6.** Identify and mitigate bias in training and test datasets.
- [ ] **Consideration 10.** Evaluate incremental AI-specific risks as part of the onboarding of third-party AI products and services within an AI use case (3.3).
- [ ] **Implementation Practice 1.** Conduct relevant use case-specific relevant due diligence during third-party AI onboarding, in line with organisational standards, to manage the risks of a third-party AI product or service.
- [ ] **Consideration 11.** Ensure that the AI use case is built with appropriate guardrails and relevant metrics for effective performance and risk management (3.3).
- [ ] **Implementation Practice 1.** Assess and select algorithms or features for the AI use case by considering its objectives and risks, including fairness, explainability, performance objectives, implementation complexity, and computational efficiency.
- [ ] **Implementation Practice 2.** Identify and implement appropriate guardrails and controls during the development of AI use cases proportionately to the level and nature of the associated risks, to effectively manage and mitigate potential risks.
- [ ] **Implementation Practice 3.** Define use case-specific risk-related metrics for assessing the AI use case for risks.
- [ ] **Implementation Practice 4.** Evaluate and calibrate transparency measures based on the use case's risk materiality, degree of autonomy, and intended users, implementing proportionate design features and disclosures to support responsible and informed use.
- [ ] **Implementation Practice 5.** Document key aspects of the AI build process, including data handling, model training and selection, and evaluation decisions to enable auditability and reproducibility.

- [ ] **Consideration 12.** Conduct thorough testing and review prior to deployment to assess AI-specific risks and ensure that appropriate guardrails, controls, and governance have been observed (3.3).
  - [ ] **Implementation Practice 1.** Ensure that Builders conduct appropriate AI risk self-checks during development to test use case performance, verify the effectiveness of risk management activities, and identify and mitigate issues early in the development process.
  - [ ] **Implementation Practice 2.** Conduct an AI-specific review based on use case risk materiality prior to deployment to ensure that potential risks are identified and mitigated.
- [ ] **Consideration 13.** Develop monitoring and contingency plans for the use case prior to its deployment, and consider risk-informed deployment options (3.4).
  - [ ] **Implementation Practice 1.** In conjunction with other monitoring activities, ensure that a monitoring plan and safeguards/contingency measures are in place, along with the designation of an appropriate accountable person to address AI risks detected in monitoring.
  - [ ] **Implementation Practice 2.** Consider the need for a phased rollout to manage the AI use case's risks and progressively validate the use case's performance prior to full deployment.
  - [ ] **Implementation Practice 3.** Engage and equip users with targeted training and use case-specific resources to support responsible use and effective oversight.
  - [ ] **Implementation Practice 4.** Ensure that the AI use case is appropriately documented, that appropriate security and governance practices are applied, that relevant data retention is provided for, and that relevant approvals are obtained before deploying to production.
- [ ] **Consideration 14.** Conduct ongoing monitoring of the AI use case and its usage to ensure that it remains fit for purpose over time (3.5).
  - [ ] **Implementation Practice 1.** Periodically monitor and report on use case metrics related to AI risks, guardrail effectiveness, and changes in the use case's operating environment, as necessary and at a proportionate intensity and frequency, and address any issues identified.
  - [ ] **Implementation Practice 2.** Monitor and report on the quality, drift, and third-party risks associated with the use case's input and training data in an ongoing fashion, as necessary, after deployment.
  - [ ] **Implementation Practice 3.** Conduct periodic checks for changes to key aspects of the AI use case over time, including risk materiality, scope of usage, and key risks.
  - [ ] **Implementation Practice 4.** Conduct periodic AI-specific reviews after deployment to assess emerging post-deployment risks.
  - [ ] **Implementation Practice 5.** Ensure that the use case is operationalised with an appropriate degree of human oversight proportionate to its risk materiality or purpose.

- [ ] **Implementation Practice 6.** Provide end users with avenues to enquire, give feedback, or request a review on AI decisions, where applicable, to support continuous improvement and build user trust.

- [ ] **Implementation Practice 7.** Ensure that proportionate monitoring and analysis are in place to safeguard against security risks during system usage.

- [ ] **Consideration 15.** Capture changes to AI use cases or their components to maintain traceability and ensure that changes with a material impact on risk are reviewed and approved through an effective change management process (3.5).

- [ ] **Implementation Practice 1.** Establish AI change management process to ensure that changes to in-house or third-party use cases are appropriately tracked, reviewed, and approved before implementation.

- [ ] **Consideration 16.** Consideration 16. Ensure that practices are in place to equip employees with the necessary AI governance and risk management skills, knowledge, and AI culture, while ensuring that teams involved in AI governance and risk management function are sufficiently representative

- [ ] **Implementation Practice 1.** Ensure that employees in relevant roles have the skills that they require to identify, mitigate, and track AI risks throughout the AI lifecycle.

- [ ] **Implementation Practice 2.** Ensure that learning and literacy activities are sufficient to equip current and future employees with knowledge on AI capabilities, risks, and responsibilities appropriate to their roles in managing AI risk.

- [ ] **Implementation Practice 3.** Ensure that practices, programmes, and policies related to culture and conduct are sufficient to foster a healthy AI culture around responsible, ethical, and safe AI use for current and future employees.

- [ ] **Implementation Practice 4.** Ensure that AI governance and risk management activities involve a sufficiently representative and interdisciplinary group of employees who can effectively represent a range of perspectives on AI's risks and impacts..

- [ ] **Consideration 17.** Support AI deployment by ensuring that supporting infrastructure is fit for purpose

- [ ] **Implementation Practice 1.** Ensure that the FI's AI-related infrastructure is suitable for managing scalability, availability, and security risks posed by the FI's use of AI.

Works Cited

[1] Acharya, D.B., Kuppan, K., & Divya, B. (2025). Agentic AI: Autonomous Intelligence for Complex Goals—A Comprehensive Survey. IEEE Access. https://doi.org/10.1109/ACCESS.2025.3532853

[2] AI Action Summit. (2025). International AI Safety Report. https://www.gov.uk/government/publications/international-ai-safety-report-2025

[3] Arnold, M., Bellamy, R., Hind, M., et al. FactSheets: Increasing Trust in AI Services through Supplier's Declarations of Conformity. arXiv. https://arxiv.org/abs/1808.07261

[4] Bank of England PRA. (2023). Supervisory statement SS1/23: Model risk management principles for banks. https://www.bankofengland.co.uk/prudential-regulation/publication/2023/may/model-risk-management-principles-for-banks-ss

[5] Basel Committee on Banking Supervision. (2021). Revisions to the Principles for the Sound Management of Operational Risk. https://www.bis.org/bcbs/publ/d515.htm

[6] Calagna, K., Cassidy, B., & Park, A. (2020). Realize the Full Potential of Artificial Intelligence. Committee of Sponsoring Organizations of the Treadway Commission. https://www.coso.org/artificial-intelligence

[7] Chan, A., Salganik, R., Markelius, A., et al. (2023). Harms from Increasingly Agentic Algorithmic Systems. FAccT 2023. https://doi.org/10.48550/arXiv.2302.10329

[8] Chmielinski, K., Newman, S., et al. (2024). The CLeAR Documentation Framework for AI Transparency: Recommendations for Practitioners and Context for Policymakers. Harvard Kennedy School Shorenstein Center on Media, Politics and Public Policy. https://shorensteincenter.org/resource/clear-documentation-framework-ai-transparency-recommendations-practitioners-context-policymakers/

[9] Gebru, T., Morgenstern, J., Vecchione, B., et al. (2021). Datasheets for Datasets. arXiv. https://arxiv.org/abs/1803.09010

[10] Golpayegani, D., Hupont, I., Panigutti, C., et al. (2024). AI Cards: Towards an Applied Framework for Machine-Readable AI and Risk Documentation Inspired by the EU AI Act. arXiv. https://arxiv.org/abs/2406.18211

[11] Holland, S., Hosny, A., Newman, S., et al. (2018). The Dataset Nutrition Label: A Framework To Drive Higher Data Quality Standards. arXiv. https://arxiv.org/abs/1805.03677

[12] Hutchinson, B., Smart, A., Hanna, A., et al. (2021). Towards Accountability for Machine Learning Datasets: Practices from Software Engineering and Infrastructure. FAccT 2021. https://dl.acm.org/doi/pdf/10.1145/3442188.3445918

[13] Infocomm Media Development Authority & Personal Data Protection Commission Singapore. (2020). Model Artificial Intelligence Governance Framework. Second Edition. https://www.pdpc.gov.sg/-/media/files/pdpc/pdf-files/resource-for-organisation/ai/sgmodelaigovframework2.pdf

[14] ISO. (2022). ISO 30400:2022(en): Human resource management – Vocabulary.

[15] ISO. (2023). ISO/IEC 42001:2023(en): Information technology – Artificial intelligence – Management system.

[16] Martens, F., & Rittenberg, L. (2021). Risk Appetite – Critical to Success. Committee of Sponsoring Organizations of the Treadway Commission. https://www.coso.org/critical-to-success

[17] Mitchell, M., Wu, S., Zaldivar, A., et al. (2018). Model Cards for Model Reporting. FAT*'19: Proceedings of the Conference on Fairness, Accountability, and Transparency. https://arxiv.org/abs/1810.03993

[18] Monetary Authority of Singapore. (2020). Information Paper: Culture and Conduct Practices of Financial Institutions (FIs). https://www.mas.gov.sg/publications/monographs-or-information-paper/2020/information-paper-on-culture-and-conduct-practices-of-financial-institutions

[19] OECD. (2022). OECD Framework for the Classification of AI systems. OECD Digital Economy Papers. https://www.oecd.org/en/publications/oecd-framework-for-the-classification-of-ai-systems_cb6d9eca-en.html

[20] OECD. (2024). Explanatory memorandum on the updated OECD definition of an AI system. OECD Artificial Intelligence Papers. https://www.oecd.org/en/publications/explanatory-memorandum-on-the-updated-oecd-definition-of-an-ai-system_623da898-en.html

[21] Open Source Initiative. (2024). Open Source AI Definition 1.0. https://opensource.org/ai/open-source-ai-definition

[22] Shavit, Y., Agarwal, S., Brundage, M., et al. (2023). Practices for governing agentic AI systems. Research Paper, OpenAI. https://cdn.openai.com/papers/practices-for-governing-agentic-ai-systems.pdf

[23] Singapore Fintech Association. (2024). Singapore Technology Talent Report 2024. https://singaporefintech.org/wp-content/uploads/2024/11/SFA-Singapore-Technology-Talent-Report-2024.pdf