



Document 5:

From Methodologies to Integration

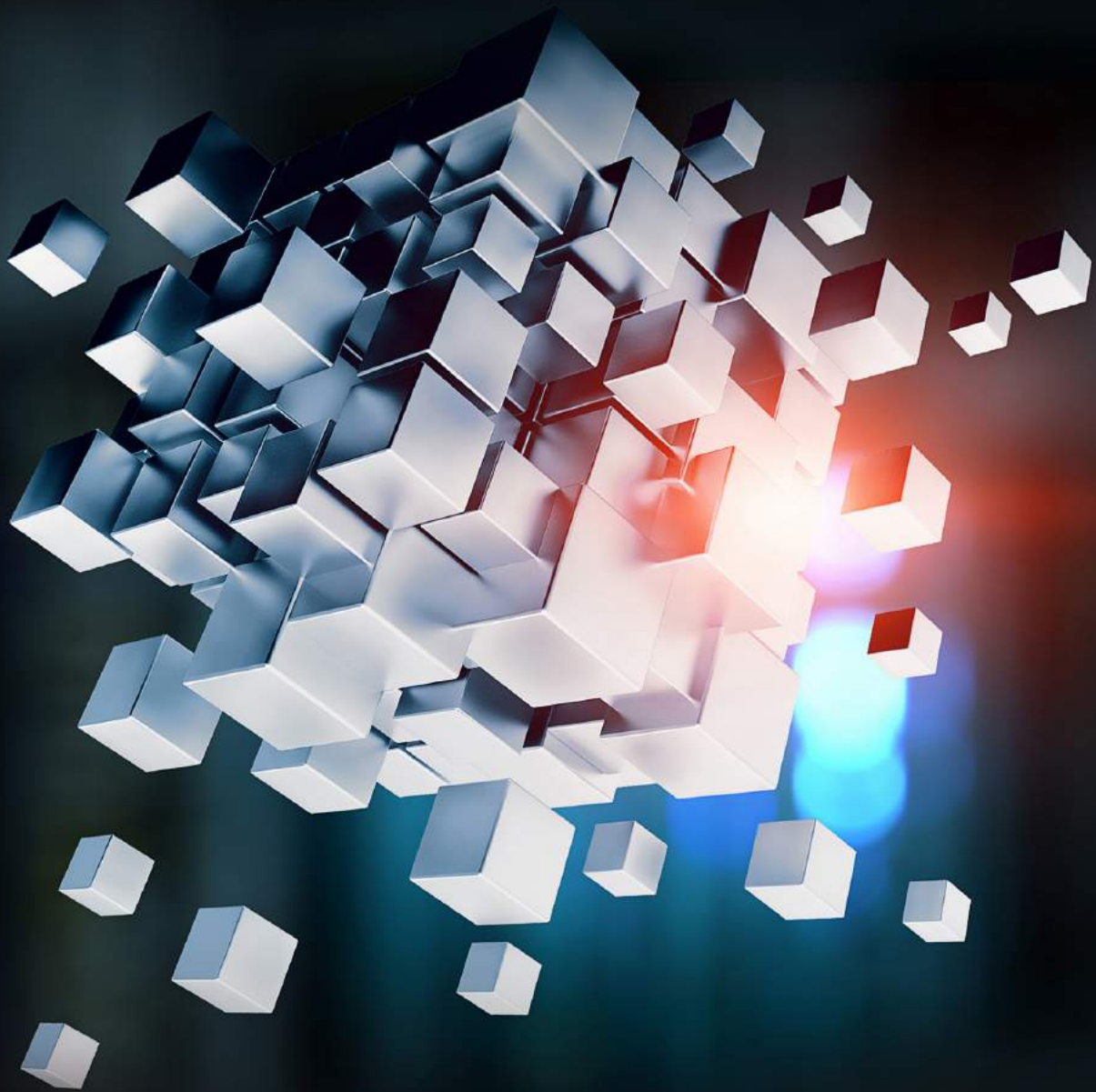


Table of Contents

01	Foreword	4
02	Executive Summary	5
03	Introduction	6
3.1	Veritas Key Objectives & Progress	6
3.2	FEAT Principles and Veritas Methodology: A Recap	7
3.3	Structure for Integration Lessons – How to Read this Whitepaper	9
3.4	Audience for this Whitepaper	11
04	Integration Achievements & Lessons Learned	12
4.1	Lessons Learned During Integration	13
4.2	Key Highlights from Financial Institutions (FIs)	19
05	Future Roadmap for Veritas	22
06	Case Studies	23
6.1	BNY Mellon	23
6.1.1	Overview	23
6.1.2	Artificial Intelligence/Machine Learning (AI/ML) Governance	25
6.1.3	AI/ML Process	29
6.1.4	Use Case Study	32
6.2	DBS	37
6.2.1	Overview	37
6.2.2	AI/ML Governance	38
6.2.3	AI/ML Process	42
6.2.4	Use Case Study	44
6.2.5	Concluding Remarks	50
6.3	HSBC	51
6.3.1	Overview	51
6.3.2	AI/ML Governance	51
6.3.2.1	Values, Principles & Governance	51
6.3.3	AI/ML Process	55
6.3.4	Use Case Study	57
6.3.5	Concluding Remarks	58
6.4	OCBC	59
6.4.1	Overview	59
6.4.2	AI/ML Governance	59
6.4.3	AI/ML Process	61
6.4.4	Use Case Study	65

6.5 Singlife	69
6.5.1 Overview	69
6.5.2 AI/ML Governance	70
6.5.3 AI/ML Process	71
6.5.4 Use Case Study	72
6.5.5 Concluding Remarks	76
6.6 Standard Chartered	77
6.6.1 Overview	77
6.6.2 AI/ML Governance	77
6.6.3 AI/ML Process	82
6.6.4 Use Case Study	83
6.6.5 Concluding Remarks	90
6.6.6 Standard Chartered – Appendix	90
6.7 UOB	101
6.7.1 Overview	101
6.7.2 AI/ML Governance	102
6.7.3 AI/ML Process	105
6.7.4 Use Case Study	106
6.7.5 Concluding Remarks	112
07 Conclusion	113
08 Acknowledgements	114
09 Bibliography	116

01 Foreword

Although Artificial Intelligence (AI) has the potential to transform financial services, it must be used in a safe and responsible manner. The Monetary Authority of Singapore (MAS) embarked on our Responsible AI (RAI) journey in 2018, when we started developing the Fairness, Ethics, Accountability and Transparency (FEAT) Principles. In November 2018, MAS introduced the FEAT Principles to firms offering financial products and services as a way of providing guidance on the responsible use of AI and Data Analytics (AIDA). The greatest challenge for the industry here, as it is elsewhere, has been to translate these high-level principles into practice.

As part of Singapore's National AI Strategy, MAS together with our financial industry partners, has created a framework called Veritas. Veritas provides a verifiable way of incorporating the FEAT Principles into AIDA solutions to address existing challenges such as the lack of industry methodologies for fairness and ethics principles, and also to work towards creating a skilled RAI talent pool. Integrating the Veritas methodology will empower financial institutions (FIs) to utilise AI confidently within their businesses and to deliver robust, trusted products to their customers. This is fundamental to our end goal which is to encourage and foster the adoption of RAI in local and global ecosystems.

The FEAT Principles were illustrated in Veritas Phases 1 and 2 through use cases such as customer marketing, predictive underwriting and credit scoring, which provide implementation methodologies for FIs to follow. This whitepaper delves into the practical challenges and lessons learned from seven lead FIs on the integration of the Veritas methodology on RAI within their organisations and their governance processes. Seven FIs have operationalised the FEAT principles by integrating the Veritas methodology within their existing risk-management framework along with other FI-specific methodologies. We took a deep dive into their approaches and drew out key lessons on how to implement the Veritas methodology in a practical manner.

Our next step is to build an open-source ecosystem for RAI that is self-sustaining and vibrant. This ecosystem will facilitate the pervasive and consistent adoption of RAI principles by FIs and tech companies, as well as support a new generation of RAI stewards who are skilled in using RAI tools. Additionally, the ecosystem will foster an active community of developers who can implement the FEAT Principles through the Veritas open-source ecosystem. Through this collaborative approach, we aim to promote RAI practices and ensure that the benefits of AI are shared by all.

We would like to thank and acknowledge the contribution of the lead members of the consortium – BNY Mellon, DBS, HSBC, OCBC, Singlife, Standard Chartered, UOB and Accenture – for the development of this whitepaper, which will serve as a resource for FIs commencing the integration journey. We thank the members of the Veritas consortium for their feedback on the integration journey and welcome any further feedback from the members as well as the public on this whitepaper. Finally, we would be happy to hear from you about how we can further enhance the Veritas methodology to ensure the responsible use of AI in Singapore and beyond.

Sopnendu Mohanty
(MAS Chief FinTech Officer)



02 Executive Summary

MAS has been working on Veritas, a multi-phased collaborative project with the financial industry, since 2018, with the project currently in its third phase. In Phases 1 and 2, MAS published the importance of the FEAT Principles for the financial services sector, and its methodology to implement FEAT (the Veritas methodology and toolkit). As part of Phase 3, MAS is working with the consortium members on integrating the Veritas methodology within existing risk management practices, and documenting potential challenges and lessons learned from these experiences which will serve as a reference for future integration cycles in an organisation.

This whitepaper provides an overview of practical challenges and lessons learned from the seven lead FIs [BNY Mellon, DBS, HSBC, Singlife, Standard Chartered, OCBC and UOB] on implementing and integrating the Veritas methodology on RAI within their organisations and governance processes. The integration journey lessons in this study have been documented under four pillars – Principles, Governance & Risk Control, Processes & Technology, Culture & Training, and Use-Case Specific Implementation. The three key takeaways that echoed with the lead FIs in Phase 3 are listed below and detailed in **Section 4** of this whitepaper.

1. There is a need for a consistent and robust RAI framework that can span geographies and industries, while also being practical to operationalise at scale.
2. All FIs that participated in this phase utilise a risk-based approach to determine the governance required for their AI use-cases.
3. While data-literacy practices are common at all FIs, with many already implementing e-certifications to increase awareness and promote responsible AI use, there is consensus on the need for training a new generation of stewards who are groomed on the qualitative and quantitative aspects of RAI specific to the Financial Services (FS) industry.

Integrating the Veritas methodology will empower FIs to utilise AIDA systems within their businesses with confidence and to deliver to their customers robust, trusted products, which is fundamental to the Veritas end-goal – i.e., to encourage and foster the adoption of RAI in local and global ecosystems. The lessons learned from this phase of Veritas will power our journey to build a self-sustaining, evolving, Responsible AI ecosystem that will be built together with the industry.

03 Introduction

3.1 Veritas Key Objectives & Progress

The MAS-led Veritas consortium developed the Veritas methodology as part of the Singapore National AI strategy together with partners from academia and industry, to implement and promote four high-level principles (Fairness, Ethics, Accountability and Transparency – the FEAT Principles) while handling AI and Data Analytics (AIDA). This helps FIs use and scale AIDA responsibly, while mitigating against unforeseen negative consequences. Veritas Phases 1 and 2 aimed to establish a methodology for the implementation of the FEAT Principles.

The Veritas methodology was illustrated in Veritas Phases 1 and 2 via use cases such as customer marketing, predictive underwriting and credit scoring which provide implementation methodologies for FIs to follow. Veritas Phase 2 also involved the development of an open-source toolkit for fairness, which enables FIs to implement FEAT on their AI/ML models. The findings and achievements from the two phases were published in 2020¹ and 2021² respectively and are guiding the industry in developing and deploying AI-based solutions to serve their customers with confidence. A summary of the FEAT Principles and assessment methodologies is provided in Section 3.2.

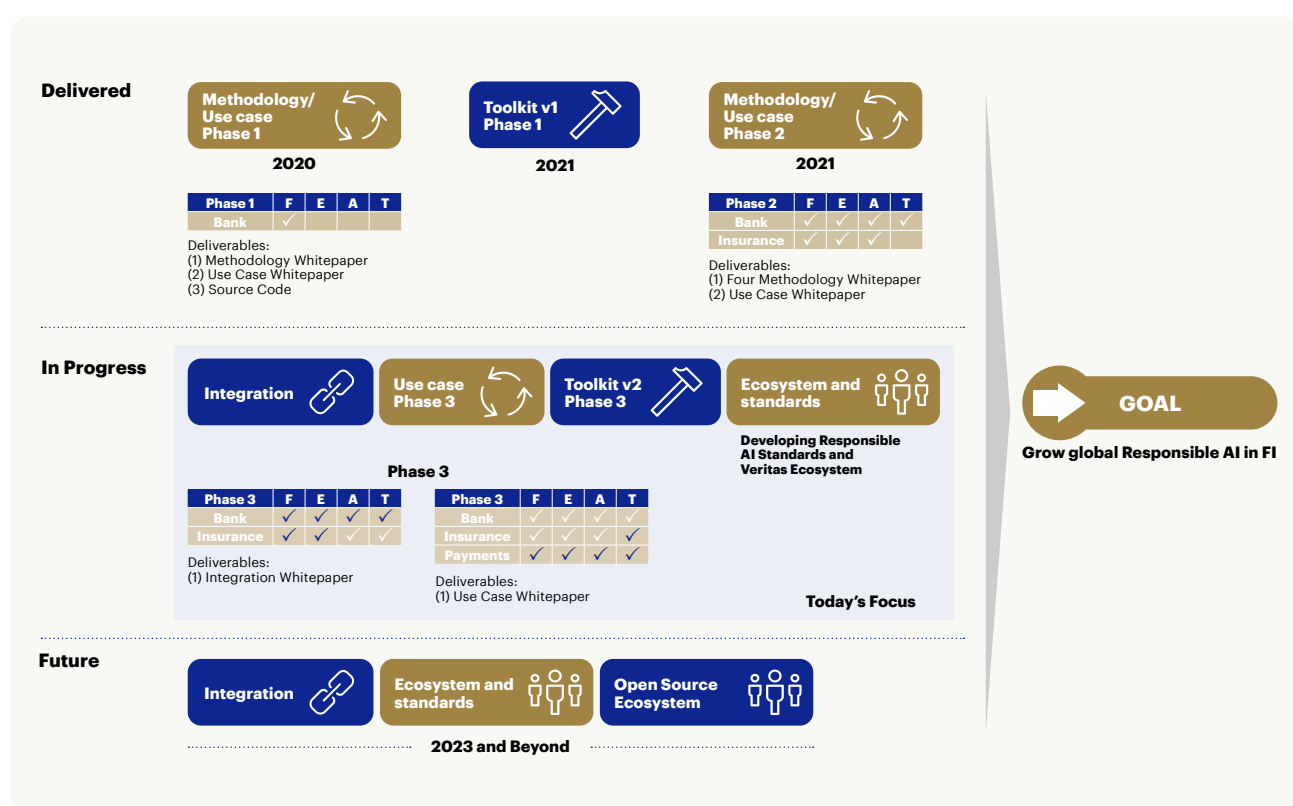


Figure 3.1.: The key objectives and progress made under Veritas.

The objectives for Veritas Phase 3 are:

- i. To document the Integration journey of the Veritas methodology by the seven participating lead FIs.

*This whitepaper will delve into the key insights and lessons learned from **the integration of the Veritas methodology** by the FIs.*

- ii. To illustrate the Veritas methodology with use cases for payment case study (Google Pay) and an insurance transparency case study (underwriting).
- iii. To develop a Veritas Toolkit 2.0 that leverages findings from Phase 2 of the Veritas methodology.

The integration workstream has seen the seven lead FIs (BNY Mellon, DBS, HSBC, OCBC, Singlife, Standard Chartered and UOB) integrate one or more of the FEAT Principles and discuss their experiences and lessons learned from this process – i.e., how does the Veritas methodology fit in with their governance, AI model development, lifecycle and the operational processes within individual organisations.

3.2 FEAT Principles and Veritas Methodology: A Recap

The FEAT Principles are summarised in the chart below.

Fairness	P1: Individuals or groups of individuals are not systematically disadvantaged through AIDA-driven decisions, unless these decisions can be justified. P2: Use of personal attributes as input factors for AIDA-driven decisions is justified. P3: Data and models used for AIDA-driven decisions are regularly reviewed and validated for accuracy and relevance, and to minimise unintentional bias. P4: AIDA-driven decisions are regularly reviewed so that models behave as designed and intended.
Ethics	P5: Use of AIDA is aligned with the firm's ethical standards, values, and codes of conduct. P6: AIDA-driven decisions are held to at least the same ethical standards as human-driven decisions.
Accountability	P7: Use of AIDA in AIDA-driven decision-making is approved by an appropriate internal authority. P8: Firms using AIDA are accountable for both internally developed and externally sourced AIDA models. P9: Firms using AIDA proactively raise management and Board awareness of their use of AIDA. P10: Data subjects are provided with channels to enquire about, submit appeals for and request reviews of AIDA-driven decisions that affect them. P11: Verified and relevant supplementary data provided by data subjects are taken into account when performing a review of AIDA-driven decisions.
Transparency	P12: To increase public confidence, use of AIDA is proactively disclosed to data subjects as part of general communication. P13: Data subjects are provided, upon request, clear explanations on what data is used to make AIDA-driven decisions about the data subject and how the data affects the decision. P14: Data subjects are provided, upon request, clear explanations on the consequences that AIDA-driven decisions may have on them.

The Veritas methodology provides FIs with a practical and detailed guide – i.e., a checklist to perform a FEAT assessment on their own AIDA models, which they can further adapt and modify to suit their organisational requirements. This covers all five steps of the AIDA life cycle – Step 0: Principles to Practice; Step 1: Defining System Context and Design; Step 2: Prepare and Input Data; Step 3: Build and Validate; and Step 4: Deploy and Monitor. The detailed checklists are published as part of the whitepapers at the end of Veritas Phase 2.³

The methodology provides recommendations on the following aspects for the FIs to consider:

- Bring FEAT assessment to scale.
- Define materiality and risk tiers.
- Adopt a risk-based governance approach for application of FEAT Principles to different use cases.
- Integrate FEAT with existing risk management practices.

The fairness checklist provides FIs with guidance to assess their AIDA system's operations with the Fairness Principle, and to define their internal foundational set of fairness assessments and customisation of assessments for a given AIDA system. The Ethics and Accountability framework provides guidance to FIs to move the organisational concepts of ethics to measurable ethical practices which can be aligned to the FEAT Principles. The checklist for Transparency covers two aspects – external transparency (i.e., what level of transparency should FIs provide to their customers/other external stakeholders) and internal transparency (i.e., guidance on how FIs can be compliant with relevant internal policies and procedures as well as any regulatory requirements).



3.3 Structure for Integration Lessons – How to Read this Whitepaper

Integrating the Veritas methodology entails incorporating the FEAT Principles within the organisation right from the strategic, governance level to the use-case specific operational level. In Phase 3, the FIs set out with the objective of applying the Veritas methodology within their organisation across these different levels. To have a consistent interpretation of the lessons learned from the various FIs' integration experiences, we have proposed four high-level pillars within which they are structured (Figure 3.2).

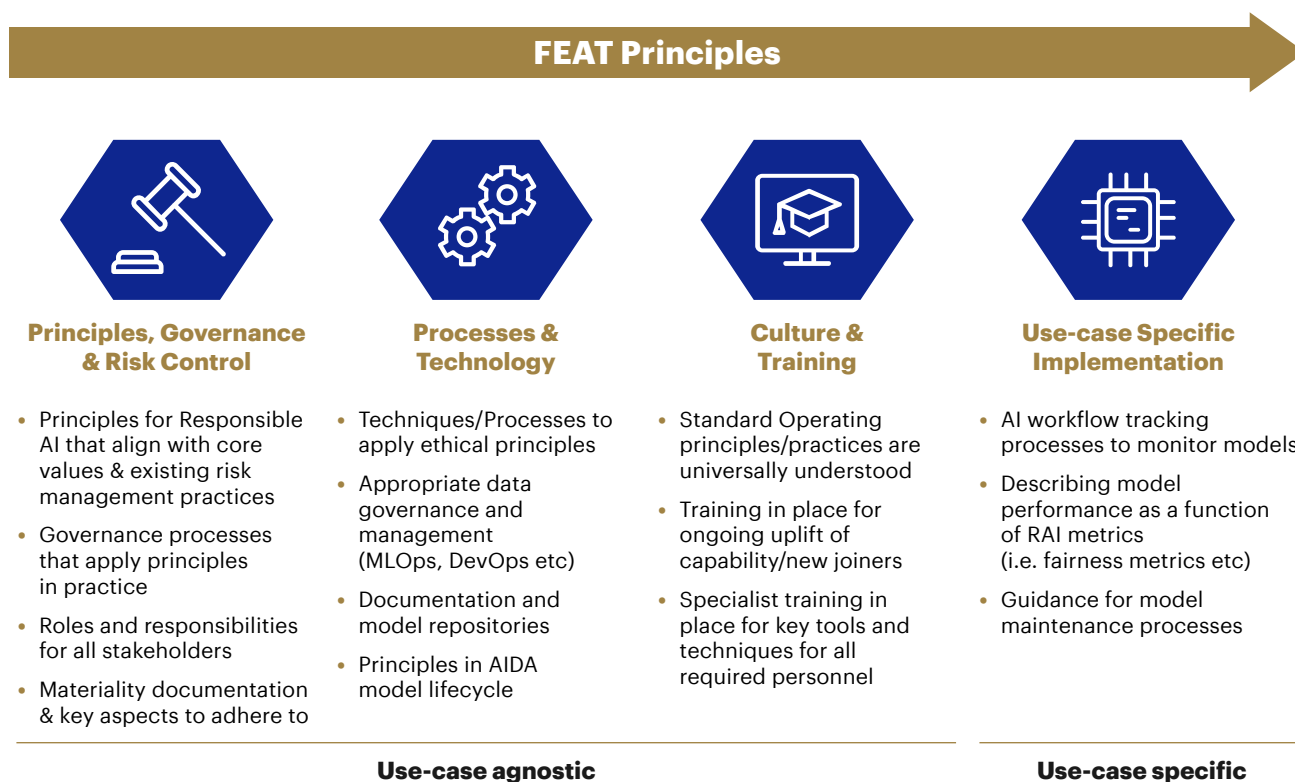


Figure 3.2.: The FEAT Principles and their relevant pillars.



i. Principles, Governance & Risk Control

This pillar aims to capture the integration insights from the alignment of an organisation's RAI framework with its own core values and principles. This is a strategic-level pillar to understand the RAI policy at the organisation-wide level, and its integration with the existing risk management measures within the organisation. The RAI Principles will dictate within the organisation its RAI policy that will in turn determine the enterprise governance frameworks. Within this pillar, FIs describe the roles, responsibilities, and accountability of all stakeholders in the organisation. In addition, this pillar captures lessons learned by FIs during risk-tiering, how materiality features into the risk-tiering as well as their considerations for making the organisation's policy and control measures commensurate with risk levels. Finally, this pillar highlights the measures of progress in rolling out AIDA governance to ensure that the organisations can keep pace with the evolving AI technologies.

ii. Processes & Technology

This is another use-case agnostic pillar to document the lessons learned and experiences of the FIs during the setup of the Operations (Ops) team – the structure, process and documentation involved. Processes flow from the risk-tiering and governance frameworks identified and established in the organisation (with lessons learned from this covered under Pillar 1). Examples of processes discussed include reporting processes and escalation processes. The processes pillar also contains documentation on the best practices within each organisation – i.e., a Responsible AI Guidebook that each organisation has started to build for itself. It also covers the experiences from the overall tech platform for data, ML, MLOps and how these are integrated with the Enterprise Risk Management Framework (ERMF).

iii. Culture & Training

Within this pillar, organisations aim to describe their journey identifying the skills required for different RAI roles. The FIs will discuss their experiences in setting up the training processes (frequency, duration, number of courses etc) as well as the measures taken to ensure that the employees are constantly learning. This section would also explore how FIs explain best practices documented in their processes to the practitioners. This encompasses an organisation-wide effort from the strategic to operational levels to understand RAI principles – i.e., training at all levels as required for them to perform their duties, and their role in contributing practically to RAI processes & technology.

iv. Use-Case Specific Implementation

This pillar focuses on highlighting insights on the RAI integration through the lens of the chosen use case. The FIs would consolidate their learnings on the AIDA model lifecycle.



The first three pillars are use-case agnostic and fall within the typical policy, process and risk control pyramids that exist in FIs in some form today. This whitepaper represents a step towards capturing the lessons learned as FIs operationalise their internal AI frameworks. To summarise, RAI Principles dictate an organisation's RAI policy, and this includes policies for risk-tiering i.e., materiality and a library of potential risk controls. The organisation's policy in turn would shape the enterprise governance and the roles and responsibilities of all stakeholders from board members to data-science practitioners. The organisation processes – i.e., escalations or day-to-day operations fall within the RAI processes layer, stem from the risk library and are commensurate with the materiality of the use cases. The best practices within each FI i.e., a FI specific RAI guidebook would be very useful for training employees on the organisation's governance frameworks and processes. This is captured in Figure 3.3.

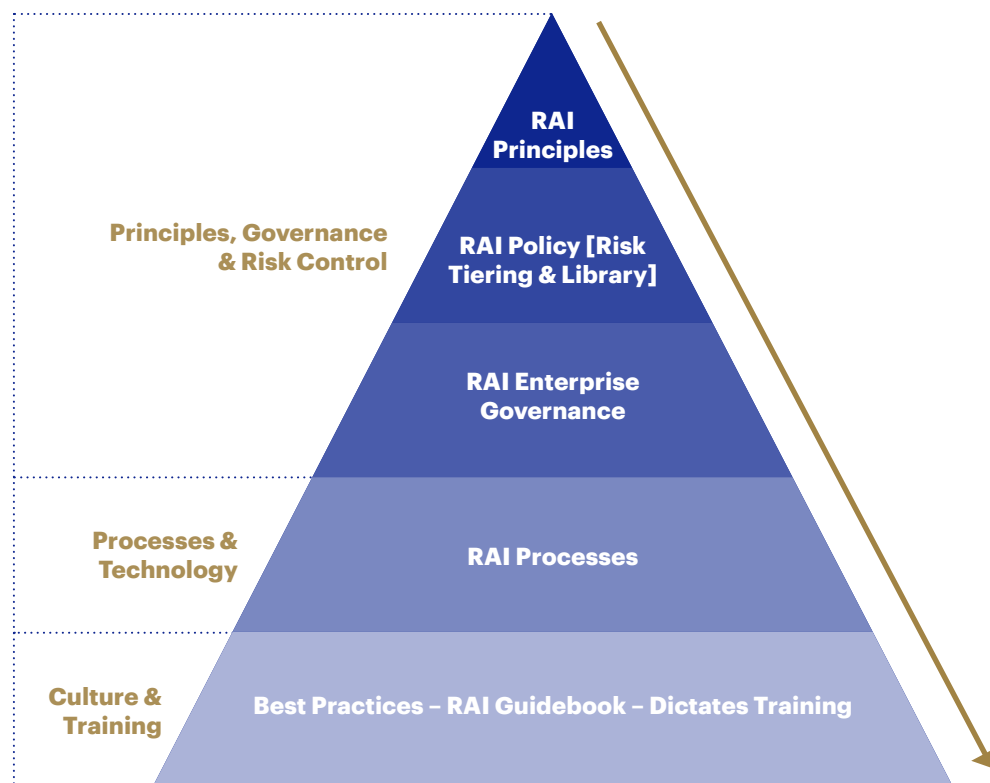


Figure 3.3.: From RAI Principles to best practices.

3.4 Audience for this Whitepaper

This whitepaper is intended to serve as a reference or baseline for audiences from FIs, tech and regulatory bodies on integrating Responsible AI within an organisation across different levels from leadership to developers. Those who will benefit from this whitepaper include representatives from FIs – C-Suite level to practitioners – such as heads of AI/data departments and data scientists/data engineers.

04 Integration Achievements & Lessons Learned

Trends across the industry show the impact of AI and Data Analytics in the financial services industry is set to increase and penetrate across all business units. Given the increase in the amount and diversity of models being deployed, it is therefore important that FIs have their frameworks established and ready. This section captures the key shared lessons learned by the seven FIs in their individual integration journeys and is followed by the detailed experience of each as a case study (**Section 6**). The case studies can serve as a useful reference for organisations that are at the start of their integration journeys.

For Veritas Phase 3, all the FIs chose the fairness principle either individually or in combination with other FEAT Principles to demonstrate integration. This whitepaper therefore shows comprehensively the processes taken to choose metrics and governance frameworks to ensure fairness in the model lifecycle and outcome. This phase of Veritas also features diverse use cases – with six FIs choosing use cases that involve structured data and one discussing the integration for a use case involving unstructured data. Although the FIs are at different stages in their integration journeys, all agreed that the exercise integrating Veritas’s methodology was highly beneficial, timely and enabled the operationalisation of AI governance frameworks at scale. The table below highlights the use cases and FEAT Principles chosen by the FIs in their integration journey.

FI	Scope of Integration	Use Case	Case Study Page Number
 BNY MELLON	Fairness	Enhancing Client Experience	23
 DBS	Fairness and Accountability	Contextual Marketing	37
 HSBC	Fairness and Transparency	Customer Marketing	51
 OCBC Bank	Fairness and Transparency	Credit Scoring	59
 Singlife	Fairness, Accountability and Transparency	Transaction Monitoring	69
 standard chartered	Fairness, Ethics, Accountability and Transparency	Signature Verification	77
 UOB	Fairness, Ethics and Accountability	Scam Analytics	101

Figure 4.1.: The seven FIs and their approaches to integrating the FEAT Principles.

4.1 Lessons Learned During Integration

The most common lessons learned by the FIs during their integration of the Veritas methodology are discussed below [Figure 4.2.] and categorised under the four pillars that were introduced in Section 3.2.

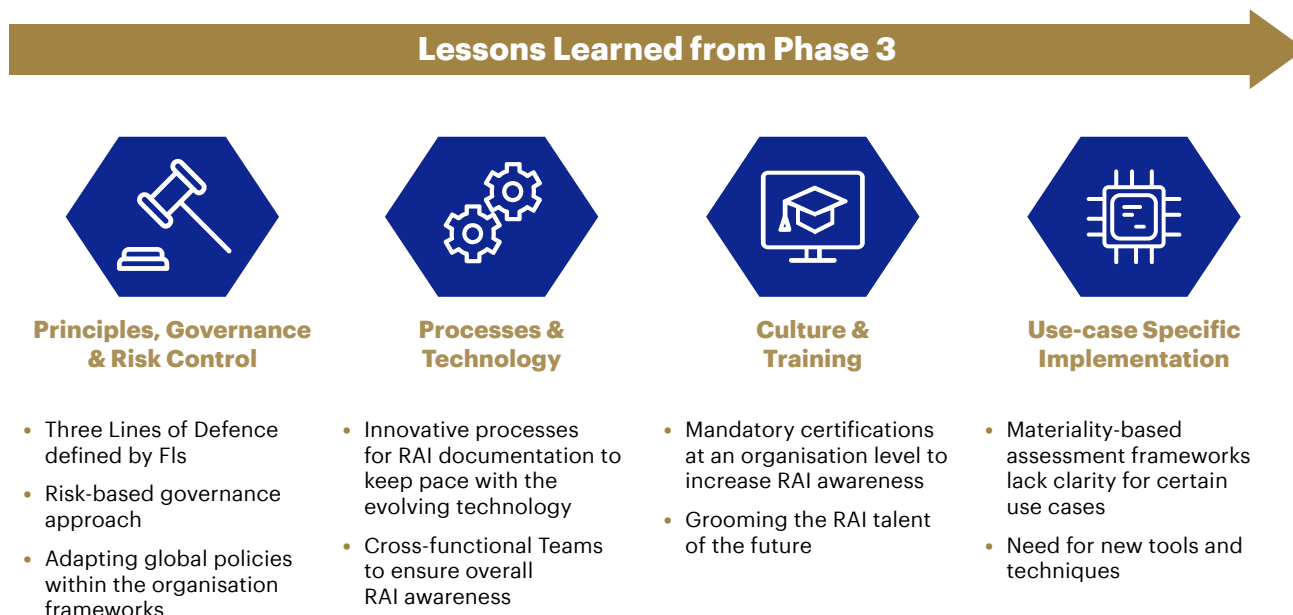
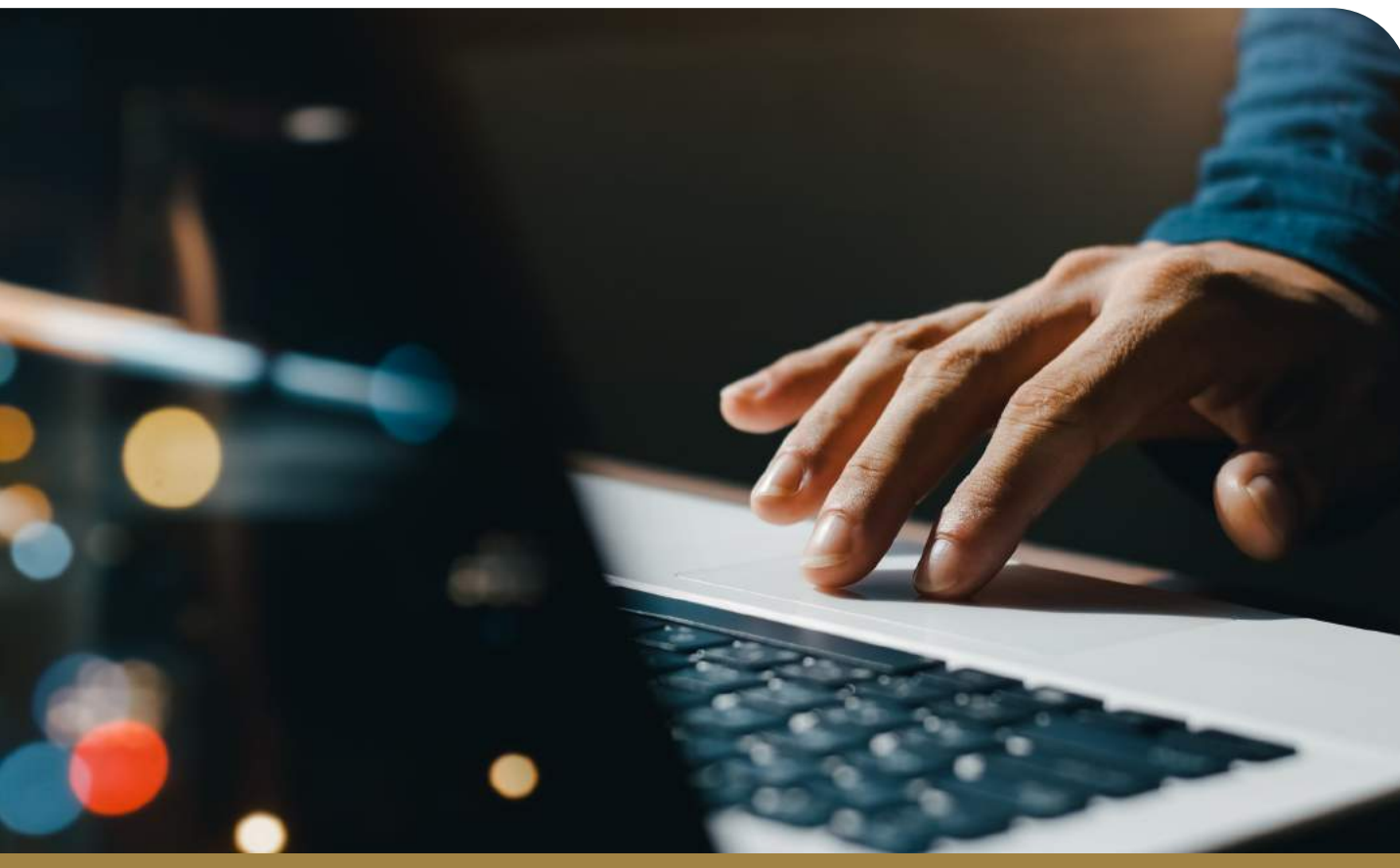


Figure 4.2.: Key lessons learned under Veritas Phase 3.



i) Principles, Governance & Risk Control:

In this pillar, we highlight the key lessons learned by FIs while integrating the Veritas methodology with their own organisation principles and how this influences the set-up of their governance framework. There are three key aspects that are consistent across all the FIs.

a) Three lines of defence (LODs) defined

FIs are aligned in delineating roles and responsibilities for all stakeholders to ensure model risk management. Typically, and as described in the case studies, most FIs have three lines of defence with delineated responsibilities associated with specific roles that are independent in nature to overcome conflicts of interest in decision-making. BNY Mellon, for example, stated the following in its case study: *“While roles and responsibilities in the process depend on the size and complexity of the organisational structure, an organisation should create three lines of defence: model-owner oversight, model risk management and internal audit.”* Similarly, DBS has its own version of the three LODs with roles and responsibilities *“defined for model owners, developers, materiality approvers, deployment approvers and independent reviewers with appropriate seniority specified.”* Standard Chartered Bank has specific roles *“defined within our RAI Standard in terms of authority to execute first-line, second-line and third-line responsibilities.”*

All the FIs stated that integrating the Veritas methodology with existing governance frameworks typically involved expanding the assessment scope to cover all AIDA models – i.e., it needs more resources whose roles are defined. They also noted that qualified resources are required to retrospectively apply governance controls to currently deployed AI models and newly deployed AI models when all are exponentially scaling the use of AI/ML at the same time as developing/operationalising the AI governance framework. As discussed in the BNY Mellon case study, the triad of people, process and technology ties very closely with data, processing and skilled teams, and leads to a functioning governance framework. This triad of people, process and technology is also central to UOB’s Data Ethics Governance Framework as described in its case study and from this paper can be broadly said to be applicable and relevant to RAI Frameworks.

b) Risk-based governance approach

There was unanimous agreement by all FIs on the importance of risk-tiering and the need to assess the materiality of a use case to determine the level of governance required to prevent harm to customers, employees and the organisation. This is exemplified in the individual FI case studies where firms adopted similar methodologies to ensure that use cases with the highest materiality were under stricter governance, and where other use cases had a baseline governance level with recommendations commensurate with risk levels. HSBC describes its model risk management strategy in detail in its case study: *“Models and model use cases are assigned a model risk materiality rating (MRMR) based on the impact. The group impact scale is used to assess the plausible impact on the business within the next 12 months.”* Similarly, OCBC also employs a risk-based approach in governance of AI use cases stating *“A key principle was to take a risk-based approach where the depth of validation of controls required for an AIDA system are commensurate with the level of potential harm arising from unintended biases from the system.”*

c) Adapting global policies within the organization frameworks

Typically, FIs operate globally and so their governance frameworks are a culmination of aspects of policies spanning different geographies, such as MAS's FEAT Principles and the Hong Kong Monetary Authority's (HKMA) High-level Principles on AI. Thus, they need to refer to multiple policies to draft a framework for their organisation to remain compliant. The study by DBS alludes to this challenge and highlights the need for a *"consistent industry-wide governance approach."* Standard Chartered Bank's case study also lists its 10 principles, including FEAT, and how those are mapped at a granular level to more than one national policy (MAS's Veritas and HKMA's BDAI guidelines). This brings out the need for the creation of robust, industry standards which can be applied at a global level and used by FIs to develop their AIDA governance frameworks. In the upcoming work with the FIs, MAS and Accenture are planning to materialise this objective and develop a consortium to build "living" RAI standards, which are global in nature.

ii) Processes & Technology:

This pillar is use-case agnostic and captures the lessons documented by the FIs on their own organisation's best practices for operational efficiencies while ensuring Responsible AI use. There are two common themes across FIs here:

a) Innovative processes for RAI documentation to keep pace with the evolving technology

A common suggestion from FIs in keeping pace with constantly evolving AI technology is that the frameworks, processes and operations need to be flexible enough to accommodate these timely updates while not stifling innovation. The processes also need to be flexible to account for the varying degrees of risk for each use case. The FIs' detailed case studies explain how they operationalise risk-tiering and determine the materiality of use cases, and thus identify processes commensurate for each risk level.

All the FIs currently have machine learning operations (MLOPs) platforms that have been developed to varying extents and innovative capabilities – i.e., either at a baseline or that are highly automated and deployed. For instance, Hydra, the OCBC's MLOPs framework, helps to automatically invoke the Veritas toolkit function that calculates fairness for a model when it is trained and submits this to the model management platform (which is a repository where every model is registered, monitored and approved). Processes are also in place for the FIs to fine-tune and monitor models at all stages, even post-deployment. As HSBC states in its case study: *"Maintenance of AI/ML solutions is more complex than maintenance of traditional software."*

b) Cross-functional team to ensure overall RAI awareness

As FIs progress in the integration journey, we note the involvement of cross-functional teams and business units. Horizontal integration across business units – including, but not limited to, legal, compliance and data science teams – needs focus and alignment to ensure all members across the organisation are aware of the processes in place for any AIDA projects. Standard Chartered Bank notes this in their case study as part of the RAI Council: *"The RAI Council is comprised of senior stakeholders across business and functions who bring a diverse set of expertise, perspectives and experiences. The council meets monthly to review, challenge, approve and, if required, reject material AI use cases."* FIs that are embarking on their integration journeys can consider involving cross-functional business teams in the early stages of their RAI process and frameworks.

iii) Culture & Training:

This pillar highlights findings from the seven FIs as they ensure RAI awareness within their organisation through training and imbibing it within the organisation's culture. Two topics best describe the journey of the FIs:

a) Mandatory certifications at an organisation level to increase RAI awareness

Data literacy practices across FIs are also consistent and include operationalising and deploying models. All FIs also have a set of mandatory e-learning certifications to increase awareness of AI usage and deployment, with those available today or in the pipeline for the near future. All FIs highlighted the importance of training across the board – i.e., from the C-Suite to the practitioners who build and deploy models. UOB described its broad, multi-pronged approach involving sessions with external industry experts, internal workshops and sharing sessions on the FEAT Principles as well as an annual FEAT self-attestation exercise, for training its employees: *“As the concept of data ethics is relatively new, a broad-based approach has been adopted to drive RAI awareness in the organisation across all levels.”* The e-learning mechanisms were useful in increasing awareness on FEAT Principles within FIs.

b) Grooming the RAI talent of the future

Several FIs commented on the scarcity of talent and training material, specifically for Responsible AI practices pertaining to the FS industry. As DBS states in its case study, there is a need for training on qualitative concepts as well as tangible, quantitative measurements across RAI for all users to gain a holistic understanding of AIDA frameworks and to ensure transparency, trust and explainability in all AIDA projects. In its case study, Singlife notes the steep learning curve for junior data scientists especially when it comes to technical implementation. This requires the capacity to develop a more tailored learning curriculum that can lay the foundations to ensure the successful adoption of RAI.

It also highlights the need for such a curriculum to educate different groups of end users (developers, programme/business managers, compliance officers) on Responsible AI practices, and is one of the key pillars of the RAI ecosystem that MAS and Accenture are currently working towards, together with partners from academia and industry.



iv) Use-Case Specific Implementation:

Under this pillar, we highlight the challenges faced by FIs and the lessons learned while navigating their use cases which range from client experience to use of unstructured data. FIs reiterated that implementing governance and checkpoints for AI/ML models, while necessary, should not slow innovation capabilities, thus highlighting the need for systematic checks at a use-case level that are commensurate with the models being deployed across all industries and risk levels. A few notable lessons are:

a) Materiality-based assessment frameworks lack clarity for certain use cases

First, the gap in having a consistent, industry-wide framework that provides guidance on the FEAT Principles' assessment and methodologies for low-materiality use cases. This is described by DBS in its contextual marketing use-case which the bank picked for integration: *"The contextual marketing use cases, by their very nature, do not restrict customers from accessing any product or service and pose no risk of unjustified systematic disadvantage. It is a time-tested and accepted way of marketing products and services across industries. An industry-wide view of the types of risks to be governed and minimum controls to be in place for such low-materiality use cases is needed, without having to apply fully fledged frameworks."* Second, the strong need to standardise fairness assessments for use cases that do not involve demographic data. This is described in Standard Chartered Bank's case study: *"Further guidance is required to assess fairness at the individual level in use cases where there is no access to demographic data."* The DBS case study also touches upon a related topic and substantiates the above statement: *"Industry-wide standards are needed on how various metrics for fairness are to be selected and thresholds set, the absence of which can show varied results across organisations due to differing interpretations and tools."*

b) Need for new tools and techniques

As highlighted in the case study by Standard Chartered Bank, there is a clear need for the development of new tools and techniques to assess unstructured data as the current methodologies extensively cover structured data formats. *"Most of the tools and techniques for assessing algorithmic fairness and system explainability that are currently available in the market are designed for structured data. In addition, they are not suitable for deep-learning architecture. For this reason, the bank uses an in-house tool to verify model fairness in these scenarios. The methodology recommends the use of differential privacy, de-identification, data anonymisation and data masking to protect personal data. These are challenging to apply for unstructured data as masked data cannot be used for training the model."* DBS also commented on the need for support for new coding languages: *"Extension of the Veritas package to support non-Python-based pipelines or models like Spark will enable wider adoption."*

c) Lessons learned specific to the FEAT Principles

All the FIs in this workstream chose Fairness as one of the principles to perform integration and demonstrated this in their various use cases. A common tendency here was to use an 'equal opportunities' approach to measure fairness of decision outcomes. Equal opportunity is a metric (either a ratio or difference) which can be defined the fraction of positive cases that were correctly predicted to be positive out of all the positive cases – i.e., the probability of positive subjects to be correctly classified as positive.⁴ The FIs typically used a threshold of 20% for this metric. This is detailed in the case studies in **Section 6:**

DBS: "Based on the above selections to the decision tree in the toolkit, 'Equal Opportunity' was proposed as the metric to determine fairness with a default threshold of 20%."

v) **Consideration for future**

Singlife notes the limitation of the current toolkit to support unsupervised learning models, which is an aspect that is being addressed by MAS in ongoing work: *“Singlife standardised the transparency assessment framework as part of its model diagnostics process, but was unable to fully utilise the functionalities of the Veritas toolkit for the fairness assessment as its chosen use case was unsupervised learning, which the toolkit currently does not support.”*

Another aspect that echoed among organisations is to support the adoption of the FEAT Principles through other methodologies, along with a framework to do so. As DBS notes in its case study: *“Organisations should assess the pragmatic aspects of any existing or emerging internal and/or external methodologies and frameworks to achieve the FEAT or other regulatory principles.”*



4.2 Key Highlights from FIs

While FIs did face some challenges during the integration of the Veritas methodology, they also highlighted the benefits of this process and their commitment to extending the integration journey to other principles and use cases. The following paragraphs summarise the highlights from the integration journey of each FI, while **Section 6** has the detailed case studies.

BNY MELLON

In this phase of Veritas, BNY Mellon set out to integrate the fairness principle and the use case to demonstrate integration is enhancing the client experience. Prior to joining the Veritas consortium, BNY Mellon was already exploring AI-related areas, including unfair prejudice, systematic bias and model explainability/interpretability, to consider how it could intentionally and systematically measure, detect and mitigate unfair bias prejudice as an enterprise.

When its Veritas project was initiated, BNY Mellon found the Veritas objectives and principles to be highly aligned with its own work, and readily joined as one of the first Veritas consortium members. The intent, objectives, approach and ethos of the Veritas project align with BNY Mellon's RAI journey. The Veritas methodology provides a common language and concepts that help to align the entire industry through systematic awareness, updated language, new methods to measure and assess bias and combat unfair discrimination, and invigorated public expectations for transparency and fairness, to the benefit of all customers and collaborators. BNY Mellon is working to incorporate the Veritas toolkit in several ways throughout the enterprise. An example of this integration was conducted on a use case to "enhance the client experience".

"The Veritas project is a trailblazer for establishing the shared industrial common language for responsible AI that our industry needs. By adopting Veritas's language and definitions, the financial services industry benefits from the synergies of collaboration and combined effort. BNY Mellon has used Veritas's best practices to help improve our internal processes (as illustrated in Section 6) while also benefitting from the confidence that our shared journey now provides in making decisions and accelerating the scalable adoption for responsible AI," BNY Mellon stated.

DBS

DBS set out to perform integration across the fairness and accountability principles, and with the contextual marketing use case. Its journey captures the need to balance the parallel scaling of both AI/ML usage and the operationalisation of AI/ML governance approaches across large and complex corporate entities – and highlights that this is arguably one of the biggest challenges ahead. It is critical therefore that organisations can put in place governance approaches which are materiality-based, seamlessly embedded into the AI development lifecycle and outcomes-focused. This allowed DBS to focus resources on governing those solutions which pose the highest risk, could embed governance as part of the way of working (rather than leaving it until the end of the process) and ensure that governance efforts were outcome-driven, not activity-driven. "The adoption and integration of core Veritas components has allowed DBS to accelerate our AI/ML governance journey and ensure alignment to the MAS FEAT Principles and outcomes," DBS said.



HSBC set out to perform integration over the fairness and transparency principles, and with a customer marketing use case. Through its integration journey, HSBC highlighted the importance of establishing a consistent and pragmatic baseline to determine the scope, materiality assessment and necessary minimum governance required to develop and deploy AI/ML solutions. It also noted how it could leverage the core values and principles to progressively enhance its existing governance process into a practical AI governance framework.



OCBC's integration journey focused on weaving the FEAT Principles into its data science workflow and MLOps infrastructure. As OCBC scales the number of AIDA use cases across the group, we anticipate the challenge of having to track and document a growing number of aspects related to model development, deployment, and governance – FEAT being just one of them. These overheads are experienced not only by data scientists, but also by management and business owners. "Our approach sought to solve this by integrating the FEAT calculations into OCBC's Hydra framework. Hydra gives data scientists a unified API interface for data querying, model training and deployment and tracking FEAT attributes. The Hydra framework also integrates with our in-house built Model Management Platform (MMP), which is a web application (complete with role-based access controls) that consolidates all project documentation, model and FEAT metrics in a single location for review and approvals from management." OCBC stated.



Singlife embarked on the integration journey choosing the fairness, ethics and accountability principles for its use case of transaction monitoring. At Singapore Life Ltd, it established the initial version of the AIDA governance framework as part of Veritas Phase 3 along with identifying key risk indicators and the baseline MLOps framework. Singlife standardised the transparency assessment framework as part of its model diagnostics process, but was unable to fully utilise the functionalities of the Veritas toolkit for the fairness assessment as its chosen use case was unsupervised learning, which the toolkit currently does not support. However, Singlife does envision great potential in the Veritas toolkit and looks forward to changes in the future which could help FIs to standardise the FEAT Principles assessment across a variety of AI/ML use cases.



Standard Chartered performed its integration across all four FEAT Principles. The bank selected “signature verification” as its use-case integration to assess the methodology in the use of unstructured data. Standard Chartered has implemented a substantial portion of the Veritas methodology in its RAI governance covering in-house AI solutions as well as onboarded third-party solutions. Its observation is that while the methodology covers several areas, it does not yet fully address the implementation of AIDA models using unstructured data. In addition, the applicability of questions in the Veritas checklists is a function of the specific use case and associated materiality; this needs to be recognised not just while implementing the methodology but also acknowledged while performing second-line and third-line reviews of the implementation to avoid over-governing the use of AI and potentially stifling innovation. The bank’s recommendation is that the industry continues to collaborate as part of Veritas and provides additional guidance in these areas. From a supervision perspective, the bank believes Veritas could be used as a valuable guideline rather than as a mandatory requirement to allow for flexibility in incorporating new techniques and metrics for managing AIDA-related risk.



UOB demonstrated its continued efforts to foster the responsible use of AIDA by integrating its learnings from its participation in Veritas Phase 1 on assessing fairness, as well as in Veritas Phase 3 on evaluating ethics and accountability, as part of an advanced scam analytics solution. The study established the importance of setting up an effective governance model with a risk-based approach commensurate with the risk profiles and materiality of AIDA models, and provided a practical application of the Veritas methodology and assessment checklist. It also recognises that data ethics is constantly evolving, and that a broad-based approach is required to drive RAI awareness across all levels in the organisation. Through Project Veritas, UOB continues to support the AI Ethics Movement and advocates for the relevance of using AIDA responsibly.



05 Future Roadmap for Veritas

AI in financial services is currently used for a wide variety of business functions, including improving the customer experience, credit risk assessments and to make available more personalised, tailored experiences. This also means responsible use of AI in the industry is particularly important since decisions regarding the granting of loans, insurance premiums and coverage as well as interest rates can significantly affect an individual – and thus the AI making these decisions must be free of bias.

The dialogue around RAI has already begun globally, and to respond to the key takeaways from the Phase 3 of Veritas, MAS, the Singapore Fintech Association (SFA) and Accenture are working with industry partners to build a RAI Ecosystem that spans regulatory and certification bodies, academia, FIs and technology firms. The goal is to catalyse the consistent, global adoption of RAI within Singapore and beyond; developing such a self-sustaining, evolving ecosystem will entail close collaboration across the industry.

To initiate the conversation, MAS and Accenture chaired a roundtable at the Singapore Fintech Festival 2022 which involved a discussion on RAI between 18 expert participants representing global banks, insurance firms, fintech, technology providers and regulators. At this event, all the participants agreed that RAI is a balancing act between fostering AI innovation while ensuring that data and privacy rights are kept intact and communities are not discriminated against while accessing financial services. There was a call for action to ensure clear regulations that are aligned amongst global authorities to increase adoption and trust in AI. The participants also called for adopting RAI by design across the entire data and AI lifecycle. Consistent with findings from the FIs embarking on the Veritas integration journey, participants at the roundtable spoke of the need for RAI education right from the level of institutes of higher learning in order to develop the talent that could accelerate AI innovation in a responsible and safe manner.

Responsible AI is a journey and one that will evolve as the technology evolves. MAS aims to work very closely with industry stakeholders to develop ‘living’ best practices that can provide organisations with the much-needed guidance to implement AI across all operations with confidence.



06 Case Studies

6.1 BNY Mellon

6.1.1 Overview

Frameworks & Foundations

Our Approach

Artificial intelligence (AI) and machine learning (ML) introduce new risks and raise new questions about how to achieve optimal results. Our teams took a detailed approach and created a prioritisation plan to execute Project Veritas efficiently. Our plan's design was driven by our approach and a framework of AI and ML based on three key pillars: data, processing and skilled teams under a functioning governance.

Data is the foundation of all AI and ML. When creating traditional technology platforms, one usually starts with a beta version, uses subsets of data, and adjusts to create the next version. However, in creating AI/ML solutions, these legacy approaches simply do not work. An accurate AI model requires all of the data at the start, or at the very least a very large and representative subset so the models will work appropriately and find the correlating features. Often, specific identifier fields can be tokenised provided the tokens can be correlated.

Due to the significant demands of data in AI/ML models, there must be sufficient processing power to process it all at speed. A distributed computing environment is required to enable productive processing speeds, and often the best option is an elastic environment that allows many specific tasks to share a majority of the processing power for short periods.

Most importantly, data and processing space rely on the expertise of the human team, and the team's competence depends on many factors: understanding the problem context, familiarity with the data resources, and experience with existing non-AI human intuition-driven approaches. Also critical is the scientific method used: the objective is to propose hypotheses (i.e., we can forecast x , given data $y..z$. $\hat{y} = b_0 + b_1x$). It is critical that the team can create an effective hypothesis, and structure its execution, data and processing so that the model will be effective.

While the team must start with a solid hypothesis, it must do so under a dual lens. Ultimately, the objective must be to attempt to disprove the hypothesis. This can feel contrary to the business objectives and requires understanding scientific principles to support the hypothesis and the anti-hypothesis. If teams are not focused on both the hypothesis and the anti-hypothesis, they waste time and effort trying to justify the often-sunk cost of creating a sub-par model. Teams must be vigilant, offsetting a natural inclination to defend a hypothesis by creating candid, unflinching documentation and requiring peer reviews of what was proven and disproven.

Finally, there must be skilled functional governance expertise. The team that creates the models must understand the law, policies, ethics and appropriateness of the goal of the model, the data that is deemed appropriate for the goal, the approvals necessary for data stewards and providers, and the records that must be maintained for future audit and review.

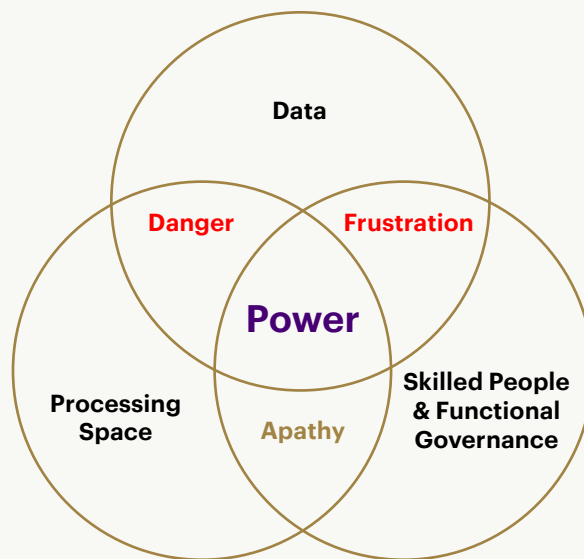


Figure 6.1.: Only at the intersection of three areas can we achieve success.

Only at the intersection of these three areas can a team achieve true success. Therefore, while sequencing may be useful, optimal results cannot be demonstrated until all three pillars are sufficiently complete, and this means working all three in parallel. However, if the team is in place before data and processing power are available, it will be hard to retain them because they will be frustrated and/or apathetic due to the inability to carry out their necessary tasks. Beyond this, if you lack competent governance and oversight, less experienced people run the risk of making serious legal or ethical mistakes.

For its part, ML requires historical data for training. The challenge is that all historical data may contain the bias that was prevalent in its time. Such bias often includes factors such as:

- Age.
- Disability.
- Ethnicity.
- Gender.
- Marital status.
- National origin.
- Race.
- Religion.
- Sexual orientation.

Bias, fairness and discrimination (BFD) are real issues in machine learning and data-driven-based decision-making. It is not just a matter of law – our ways across a spectrum of users, examples or datasets from using these poor results. For example, a car insurer's use of mileage as a factor to set premiums could discriminate against lower-income drivers who may tend to drive less.⁵ (Consumer Federation of America, n.d.)

6.1.2 (AI/ML) Governance

Vigilance & Methods for Mitigation

If vigilance is not put in place, BFD issues may creep in unnoticed. Existing data resources may reflect historical inequities, and the data may include features about people or events that can lead to biased treatment. This data can be an essential input for detecting useful signals, such as correlations (e.g., Granger causality) for predictive analytics. Yet, when this data is fed blindly into AI models, these models may yield biased results.

There are oft-practised methods to address bias or signs of discrimination such as inaction, feature prohibition, unaware modelling, demographic parity and mastering data. Firstly, a team can, and sometimes does, take no action. Blindly following the data will likely perpetuate and create a built-in bias.

Another strategy is removing prohibited features in the data, creating a fairness-through-unawareness strategy. This is largely ineffective, uninformed and fails to address redundant and/or correlated proxy attributes that are highly correlated with them. Moreover, it deletes potentially relevant and critically important signals for the analysis which prevents the use of data science itself to combat bias proactively.

Group unaware modelling is another strategy. One can implement the same threshold and/or strategy for all demographics in a data set. However, this can ignore insightful legitimate variances between groups as this strategy would price them the same. Unfortunately, even with the best group-unaware models, this approach may still discriminate between groups because underlying distribution differences are ignored.

One can also tweak algorithms to achieve demographic parity. ML models may be proactively adjusted so that different thresholds are used for different groups such that the positive rate is the same across all groups (i.e., the percentage of instances classified in a certain direction is the same).

Our team's preference is another strategy. The teams that operate and manage the AI/ML model must understand and master the data used to enforce equal opportunity – a strategy that is optimal for comprehensively managing bias. Moreover, ML algorithms may be adjusted so that different thresholds are used for different groups such that the true positive rate is the same across all groups (i.e., performance of the classifier is the same across groups). We find that this approach produces the best and most useful data science results, which then leads to more optimal decision-making. Once this is complete, it is necessary to do a materiality assessment and risk-tiering.



Materiality Assessment

Materiality assessment and risk-tiering are necessary to comply with the Model Risk Governance (MRG) processes for models. Risk related with the use of particular AIDA factors can result in both financial loss and reputational harm to an institution.

Model-tiering stems from the inherent risk of the model. Managing AIDA must first be commensurate with the materiality of the model. In general, three aspects should be taken into consideration when assessing model materiality: complexity, capital and liquidity, and model interaction:

- **Complexity:** Models can range in complexity from simple decision trees to neural networks. The more complex the model, the higher its tier. This may be due to a lack of transparency in the model's approach, a network effect whereby the model relies on many segments, or a unique approach that is not easily benchmarked.
- **Capital and liquidity:** Capital and liquidity impact a model's rate of failure, which may be due to the short – or long-term effects of a capital position or liquidity. Reputational risk falls under this category. The more significant and immediate the effect is, the higher the tier it should have.
- **Model Interaction:** Tiering must further reflect how models interact, such that if a model provides a principal input to a downstream model, its support for processes that affect a firm's capital, liquidity, profitability and reputation should be considered at least as great as the downstream model.

Additional aspects that may be taken into consideration for tiering are automatism between models' output and the decision-making process (reliance on the model's output), and the number of processes supported by the model or model exposure.



Risk-Tiering Strategies

AIDA models should follow standard model-risk policies and procedures. Structuring the process and the layers of documentation depend on the size and complexity of the firm. However, the general principle is that there should be three tiers of normative acts:

- **Tier 1: Policies** – The board or committees should define high-level key controls on model risk management. Breaches of the policy should be reported back to the approving body. Moreover, the body approving policy should also oversee models with a high-level risk appetite.
- **Tier 2: Procedures** – Documentation must define the practical implementation of policies, and this documentation should be approved at the chief risk officer (CRO) or head of model risk level. Breaches of the procedure should be reported back to the approving body. Finally, an approving procedure should also oversee KRIs/KPIs related to model risk.
- **Tier 3: Templates** – Documentation should standardise how the model document and validation reporting works. An approving body should have adequate control to promote consistent template adoption.

Our data science community is a triad comprised of people, process and technology, and this tightly aligns with the three key pillars: data, processing and skilled teams under a functioning governance. The objective is to promote the awareness of techniques and drive data-driven decisions and advanced analytics using this triad.

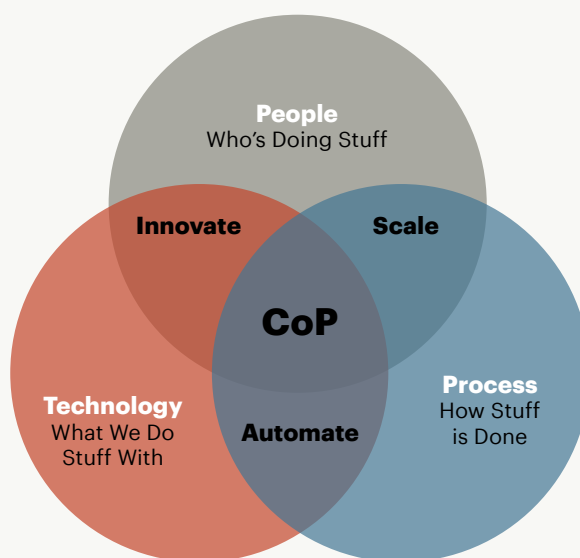


Figure 6.2.: People, process, technology – how the triad drives success.

It is necessary to make thoughtful risk-based decisions, but speed matters in business so we must choose action.

Next, we will discuss our people and their required roles, along with our process and technology.

Stakeholders & Required Roles

Roles & Responsibilities

While roles and responsibilities in the process depend on the size and complexity of the organisational structure, an organisation should create three lines of defence: model-owner oversight, model risk management and internal audit.

Model-owner oversight is required to support efficacy. Senior staff should oversee model owners, users, developers and the business risk committees, and are responsible for ensuring a business area is aware of the risk created by its models and is managing that risk. Model owners, on the other hand, should manage the risk from specific models so that proper development, validation and use takes place. For this to be effective, the model owner should have the experience to understand a model's approach and proper use. The model owner will also require supporting owners – modelers, developers and technical experts familiar with industry best practice for modelling methodologies.

Model risk management (MRM) is a second line of defence and has the primary responsibility of implementing the policies, procedures and standards on behalf of an oversight board. It should be staffed by those with the quantitative background and industry experience necessary to identify and consider options to mitigate model risk. Specifically, MRM should:

- Monitor compliance with applicable regulatory guidance, rules and regulations.
- Establish and maintain a consistent model risk-management framework across the company.
- Train and provide templates to model owners, users and developers to promote consistent application of this policy across the firm.
- Provide a process to inventory and maintain the inventory of the company's models to enable reporting on model risk including periodically reviewing the inventory for obsolete/unused models and encouraging their retirement or suspension.
- Execute the tiering process for models.
- Establish the effective challenge of models, consisting of critical analysis by objective informed parties who can identify model risks and influence appropriate mitigation options.
- Through validation activities, identify and guide the management of model risk in the company.
- Report to the board and committees.

Within the second line of defence, it may be necessary to set up a committee that oversees the MRM process. The committee might be empowered to resolve issues before escalation to the board, and acts as a check on the validation group.

Internal audit is the third line of defence. This function assesses the effectiveness of internal controls by providing independent, risk-based assurance reviews designed to identify controls, risk mitigants, control gaps and opportunities to improve efficiency. Internal audit is expected to periodically assess compliance with this policy, following its practices and methods. Internal audit does not revalidate individual models although it may sample models in its analysis of the framework.

A board may set up key model risk-management principles and document the appetite for model risk.

Team Demographics & Lessons Learned

The BNY Mellon operations team has learned many lessons throughout our Responsible Artificial Intelligence (RAI) process. First, smaller teams of experts were found to be more impactful. Smaller teams allow for quick actions and fewer coordination issues. This reduces the overall time for a model to go into production, which improves the “time to value” for a data science project.

We have found that creating “Core Teams” of one data scientist, one data engineer and one business analyst, who is a domain expert and understands the underlying transactional data, is optimal. These “Core Teams” can then be augmented by support staff comprised of other data scientists and engineers as well as cloud experts. The benefit is that there is less overhead around coordination and more focus on core engineering buildout and experimentation, though success hinges on identifying the right key talent. More importantly, team members enjoy experimentation via different models.

A multi-disciplinary team reduces the risk of ending up with a model that reflects the biases and assumptions of developers only. A well-rounded team is recommended as a requisite for an AIDA people setup.

Developing Multi-disciplinary Teams

In order to create a comprehensive multi-disciplinary team, there are specific roles to consider. First, data engineers confirm that the data is relevant and useful, while data scientists build models and help the Enterprise AI function achieve project objectives. It is also beneficial to have executive sponsors who understand the technical aspects of AI along with the broader strategic goals for the business, as well as subject matter experts (SMEs) who can provide additional context and institutional knowledge, common business practices and information about known historical biases. They also work closely with AI users and reinforce the importance of achieving the outcome for all stakeholders.

Finally, legal, data privacy and policy experts oversee data use and governance. It is also beneficial to have model-op/systems engineers to productionise and keep the models running in production and ensure the systems are robust with proper failover mechanisms in place.

To ensure processes are efficient and the setup is appropriate, we have regular interactions with business, centralised functions like technology, and clients. In addition, we leveraged cloud operations and elastic computing to enable intensive model-training requirements; this would not have been as fluid or potentially possible on an on-premises infrastructure. However, once models are in production, they are run on BNY Mellon’s on-premises infrastructure.

6.1.3 AI/ML Process

We aim to design reliable systems with a proper failover in place. It is necessary to control technical diversity, reuse what we have, buy what we need, and build for a competitive edge. Applications should be easy to use and transparent for users, as this enables them to concentrate on their tasks. The enterprise architecture is built over low-coupling, modular components that implement services that are consumed through standard interfaces, and whose engineer solutions have strong design foundations.

AI solutions and products used in the bank are integrated with an Enterprise Risk Management Framework (ERMF) which is clearly defined and documented. We use a common technology-risk platform which also serves as a line of defence in technology and unites issue identification, issue treatment, policy management and non-compliance processes for the AIDA systems used during MLOps and the model lifecycle.

These experiences from overall tech platforms for data, ML and MLOps have been more deeply integrated with ERMF, and we now have a key focus on transactional data quality. ML models can do only so much if data quality is poor, and it is important for transactional systems to invest in a robust distributed architecture.

Even if there is no immediate requirement for the data, it is imperative that we forge a data-first and data-centric approach, enabling the collection and preservation of tick-level or other granular data points in data storage system such as a data lake. We have built a robust line of business-specific, big-data warehouses that store all transactional data and subsequently helped in building predictions for frequent time periods. In hindsight, more investments are required in MLOps as we start to build more ML models across the bank.

Model Permissioning

Various control gates (such as permit-to-design, permit-to-build and permit-to-operate tollgates) are used to evaluate development initiatives throughout their lifecycles. These control gates are employed to eliminate redundancies and to align modelling efforts with business objectives. Examples of such tollgates include: A) approval to model (AtM) which may include the validation and approval of the appropriate use of data resources; and B) approval to employ (AtE) which evaluates whether developed models are ready for deployment and have been approved by the model risk-management organisation.

These model control gates are used to assess the readiness of machine learning initiatives to proceed to fully-fledged modelling and confirm that we have a clear understanding of the business objectives, scope, success criteria, impact, stakeholder sponsorship and the initial view of the solution and data resources. Additional independent reviews may be needed to evaluate whether the data that is being provisioned from the enterprise data resource (e.g., the data lake) is being used in appropriate and approved ways.

There are two outputs from the tollgates: approval to model, and/or information that requires additional clarity, in which case a follow-up session is necessary.

The purpose of AtE is to review the technical robustness of the projects before going to production. This tollgate will review the project as defined by the specs, the solutions/algorithms used and the results. It will also be reviewed by an independent board which acts as a second line of defence. During this process there will be an exploration of the algorithm and method chosen, an analysis of the data (sources, types) and, given the data, how the data will satisfy the model requirements (complexity, sampling, overfit, how to update the model and how often). There will then be an assessment of the quality of the results as well as a review of model governance, monitoring and testing.

The output from the AtE tollgate is either: approval to deploy the model into production; or additional clarity/alternative approaches required, in which case a follow-up session will be set up.

We have made the process more efficient. While initial phases of the project focused more on rapid iteration, there was less documentation; in subsequent phases there is a greater emphasis on governance and more detailed documentation.

Managing Risk with ModelOps

Machine learning in finance requires reliability and accountability in the outputs based on the data, and this has been a focus in the modelling and maintenance of AI systems. ModelOps is a set of practices to address this. The Risk Library takes a broader look at risk in ModelOps, and is available as an independent library to be used by the modelers and risk managers, and will be integrated into the Ground Truth (GT) system. It includes a broad range of metrics to measure and monitor risk, with each project using the relevant metrics. Given that projects exist over time with changes to data and the retraining of models, continuous tracking of the risk metrics for each project in the GT system is critical for the project's governance.

The Risk Library will support the following types of risk:

- Fairness.
- Stability.
- Drift.
- Uncertainty.
- Interpretability.
- Explainability.
- Causality.
- System of systems.
- Adversarial.

The proposed ML Risk Library takes a broad view of risk in MLOps including risks to customers, risks to the organisation and risks to society. The Risk Library will serve as a complement to the GT system within BNY Mellon, and will provide access to a range of metrics to measure and monitor risk. Although the assessment of risk will benefit from integration with the GT system, it is not necessarily tied or restricted to only those models or ML pipelines housed within the GT system.

Given an ML implementation, background datasets, risk considerations and performance criteria, we seek to automate the computation of risk statistics and the construction of reports that communicate the risks.

Similarly to how current ML evaluation libraries “wrap” around any generic ML algorithm, the ML Risk Library will wrap around an end-to-end MLOps pipeline. It will:

- Monitor relevant statistics at each step, including data pre-processing, model selection, model training, validation and placing in-production.
- Stress test relevant steps with respect to each “risk vector” from a family of approved perturbations and experimental tests.
- Catalogue results, rank-ordered and placed in context with respect to priority and relevance.

The decision-maker would be able to assign the percentage of weightage to different risks and receive an assessment of overall risk.

We are one of the original members of the Veritas consortium and are the pilot users of [the Veritas Toolkit Version 1](#). The Veritas toolkit is in the process of being integrated into the fairness component as part of the ML Risk Library.

6.1.4 Use Case Study

To help better serve clients, an AIDA system is developed to better anticipate clients' needs and optimise the client experience. Machine learning predictions are used to provide additional insights with explanatory reasons to help businesses make better decisions around client engagement. The final decision of a client engagement is taken based on a combination of the machine learning prediction, due diligence analysis, business rules and manual overriding.

Working with SMEs, we label time-series data to train models that obtain a better anticipation of the end user's needs and optimise the firm's services and end-user experience. The risk tier of this AIDA system follows the firm's model risk rating guidance. As a pilot-testing use case, the Veritas Fairness FEAT Assessment Methodology ("Methodology") was used to conduct the fairness assessment. The diagnosis tool under [Veritas Toolkit Version 1](#) was used for the fairness diagnosis, generating analytical results to support and answer all relevant assessment questions in the methodology.

In our use case, we had data engineers pull relevant data from multiple sources and set up a monthly pipeline. The label was not readily available, so we utilised a complex method to define the label as per the SME, utilising institutional knowledge, business practices and any historical biases. We then utilised this to train the model.

As a part of AtE (approval to employ), the suggestion was to test and compare results at different segments. A model was trained once, and the same model was reused again and again every month (with new data) for predictions. Once the model results were approved by the business, we went through the AtE process where technical details of the model were reviewed by a group of internal experts who are data scientists. Only once the model was approved by Model Risk Management Group (MRMG) and obtained the AtE could it then move forward to production.

Models also went through a rigorous review with the model risk team where the entire end-to-end process was documented in detail, and it was validated by the MRMG team from a risk perspective as well as validated for ML method correctness. The document also captures different experimentations, results and more. This model was then integrated with the GT system which helped us to monitor model results across different timelines and compare the results of different models – a very time-consuming process for the data scientists. It also enabled us to send alerts in case of the degradation of performance beyond a threshold. Ground Truth also protects against the results and underlying data being modified.



Assessing Impact

Having clear business objectives is the key to developing a fit-for-purpose AIDA system. Business objectives are measured quantitatively using model performance. Acceptance criteria are set based on iterative discussions between model developers and SMEs, who consider various types of potential risk and business benefit according to the model risk rating. The trade-off between model performance and RAI aspects such as fairness should be decided by in-depth discussion with SMEs and AIDA system-owners.

Before we discuss the trade-off between model performance and fairness, a fundamental step is to make sure that fairness is defined and assessed in a suitable and relevant context. Unlike other types of risk, fairness is a special topic which is not fully well defined, and it could be use-case specific.

There are four key considerations in performing a relevant fairness assessment:

- a. Selection of protected variable(s)** – There may be more than one protected variable with each covering different business aspects. For the current use case, three protected variables are considered, namely country, segment and class.
- b. Identify the privileged/unprivileged group(s)⁶** – Privileged and unprivileged groups can be examined based on a statistical representation or historical bias (not exclusive) in conjunction with SME knowledge and expertise. Representation bias occurs when certain groups are underrepresented in the data set. Thus, from a statistical point of view, the least-represented group(s) is prone to be the unprivileged group, and vice versa for the privileged group. On the other hand, the privileged group could be identified based on historical bias. For example, highly educated groups of individuals would have privilege in terms of credit scoring from historical records based on manual approval before utilising the AI and ML models. The identified privileged and unprivileged groups should be validated using business domain knowledge from SMEs. For this use case, the privileged group for the country-protected variable is selected based on statistical representation, while for segment and class, privileged groups are identified based on historical bias and business understanding.
- c. Group fairness** – Many fairness assessment toolkits are developed with one commonly used assumption to simplify the privileged/unprivileged groups into a binary-class grouping, meaning grouping multiple minor groups into one unprivileged group. Different fairness metrics are calculated to compare the unprivileged with the privileged in terms of benefits and harms. However, this binary-class grouping of privileged/unprivileged groups may introduce additional bias and the fairness conclusion may be misleading. In the current use case, different sub-groups are compared with the identified privileged group, so fairness metrics are calculated for different privileged/unprivileged pairs across different protected variables.

d. Primary fairness metric(s) – There are around 18 fairness metrics in the Veritas toolkit and customised fairness metrics can be developed based on use-case specific fairness objectives. Among all these fairness metrics, one or more primary fairness metric(s) are selected to make the fairness conclusion. The Veritas toolkit provides a FairnessTree to help users automatically select the primary fairness metric based on three fairness considerations, namely fairness priority, fairness concern and fairness impact. The primary fairness metric should be selected to align with the fairness objectives, and the trade-off analysis with business objectives should be handled with thoughtful considerations.

Giving a fairness assessment conclusion requires deliberate considerations. This brings several key considerations and challenging tasks:

- 1) How to synthesise different pair fairness results? Simple methods include giving different weights to pair results, prioritising based on business understanding or selecting the worst pair.
- 2) How to conduct the trade-off between model performance metric and the synthesised fairness results? Simple trade-off analysis for a privileged/unprivileged pair may not work well because this is a highly nonlinear and multi-dimensional optimisation problem.

Model-tuning can be a time-consuming exercise as we apply the GridSearchCV technique to identify the best tuning parameters, though recent improvements in the underlying infrastructure mean it has become easier to identify these parameters.

Once the model is deployed, we need to continue monitoring the incoming data and results to confirm that the model build is still relevant. For example, if the input data distribution changes for a few of the important columns, it may impact the model's performance. As results degrade beyond the predefined threshold, the model needs to be retrained.



There are mainly two risk-monitoring and governance processes that are implemented: a central AI governance system is developed as part of ERMF, while model-performance metrics, data traceability and model-training versioning are monitoring on Ground Truth and reviewed by the MRMG. The Enterprise MLRisk (Machine Learning Risk) Library is also used to detect various types of model risk apart from fairness, such as stability or drift. Results of the model risk analysis will be integrated into Ground Truth.

While statistical models follow a rigorous process to define their risk, machine learning models do not. Though the models may have good overall accuracy, they are not accurate for all data-points, and their accuracy can change with retrained models. A different approach is required to handle the risk in model errors.

The explainability of machine learning models allows for simpler models such as linear regression. On the other hand, logistics regression and decision trees have algorithm-specific explainability, and the majority of the models cannot be mathematically explained as these use a model-agnostic method.

Shapley values are also a widely used approach from cooperative game theory that come with desirable properties. The core idea behind Shapley value-based explanations of machine learning models is to use fair allocation results from cooperative game theory to allocate credit for a model's output $f(x)$ among its input features. One of the fundamental properties of Shapley values is that they always sum up to the difference between the game outcome when all players are present and the game outcome when no players are present. The "game" is the prediction task for a single instance of the dataset. The "gain" is the actual prediction for this instance minus the average prediction for all instances. The "players" are the feature values of the instance that collaborate to receive the result. The Shapley value is the average marginal contribution of a feature value across all possible coalitions.

For machine learning models this means that SHAP (Shapley Additive exPlanations) values of all the input features will always sum up to the difference between the baseline (expected) model output and the current model output for the prediction being explained. The difference between the prediction and the average prediction is fairly distributed among the feature values of the instance.

The Shapley value allows contrastive explanations. Instead of comparing a prediction to the average prediction of the entire dataset, you could compare it to a subset or even to a single data point.

There is however an impact of interpretability. In situations where the law requires explainability, the Shapley value could be a good solution as it is model-agnostic, is based on a solid theory and distributes the effects fairly.

Given a result, Shapley values can be used to explain the impact of each feature to contribute to the result. More importantly, this can also be used for a simplified what-if analysis for a specific data point, which gives an intervention solution for the model, and which must be added to the solution. Prominent subsets of data must be explained using Shapley values, giving an overall perspective of the model.

There is immense value in interpretability for the end-user. First, a user can see the result and discern if it makes sense based on the combination of the feature values. An SME who uses the model can judge whether the model's result is logical for a datapoint and can accept the result on a case-by-case basis. In this way, we transfer the responsibility of the model to an informed user with the goal that the end-user understands and accepts the responsibility of using the model and all the impacts that come with it.

Concluding Remarks

In these ways, processes for managing risk should systematically consider FEAT-based risks and employ appropriate FEAT frameworks, such as those being defined by Veritas – and this explains why collaboration with the Veritas consortium is so important for the industry. We welcome the increasingly effective outcomes-based regulatory requirements and best practices.

We want BNY Mellon to continue its leadership role in the drive for the responsible use of AI and ML. The same fundamentals continue to apply even as we look at new risk metrics and faster assessment. Risk analysis continues to be concerned with performance and its variations based on a specific dataset. The results from using the Veritas library give a new method for assessing risk, which is critical to integrating with the governance framework. This includes fairness metrics in the model evaluation which greatly expands the reach and expectations of the risk management culture. Providing our team with tools to better interpret and assess the models enhances trust in our business processes across the financial ecosystem.



6.2 DBS

6.2.1 Overview

The volume, velocity and variety of data are increasing at an unprecedented rate, with technology, regulatory views and societal norms also evolving at pace. With these advancements and newer varieties of data, numerous inferences can be derived by organisations to add value to customers and employees in the form of appropriate and tailored product/service offerings at the right place and time.

Individuals are increasingly cognisant of achieving equitable value exchange when sharing their data in return for products and services. They are also aware of their trade-offs between privacy and personalisation in relation to these products and services, and therefore able to make informed choices about what data to share and what not to share. As such, stakeholders including customers, employees, regulators and society at large, are becoming mindful of the risks involved in how organisations collect, analyse and use their data, and have less tolerance for mismanagement or irresponsible data use than before.

At DBS, we commenced our Responsible Data Use journey in 2018 and worked towards establishing a simple yet effective framework to ensure any proposed data use was responsible and aligned to our organisational values, without compromising our efficiency or use of data.

The outcome was our proprietary PURE Framework (Purposeful – Unsurprising – Respectful – Explainable) to evaluate the dimensions of responsible data use. This framework is embedded in the organisation's DNA, supported by an assessment tool, defined roles and responsibilities, comprehensive training, a bank-wide standard and appropriate senior management oversight.

As we started to leverage and scale our use of Artificial Intelligence/Machine Learning (AI/ML), we also ensured that we did so responsibly. This was the start of our AI industrialisation journey that coincided with our engagement with the Monetary Authority of Singapore (MAS) and the development of the FEAT (Fairness, Ethics, Accountability, Transparency) Principles.

A Holistic Approach to Responsible Data Use

DBS looks at responsible data use holistically to retain our stakeholders' trust and manage data risks as we continue to maximise value from our data. Our responsible data-use framework ensures that we address three core questions through our processes and practices:

1. **Can we use the data?** This is through our already established and ongoing data management, security and compliance processes, tools, practices and operating models to ensure data is owned, secured, accessible, lawful and managed.
2. **Should we use the data?** Strongly aligned to our organisational values, the PURE framework addresses the dimensions of our data use and remains sensitive to changing social norms and cultural nuances across core markets.
3. **How we use the data?** The governance approach for our AI/ML solutions ensures governance commensurate with the risks posed, and follows closely the MAS FEAT Principles of fairness, transparency and accountability. This part of our Responsible Data Use framework will be addressed in this document.

6.2.2 AI/ML Governance

After operationalising the PURE Framework, we wanted to strengthen the governance of our AI- and ML-based solutions. This was driven by our intention to scale the use of AI/ML across the bank, while also being synchronous with the increased focus on this area from various regulatory, governmental and international bodies. It would be fair to say that this is still an emergent subject at a macro level, evidenced by the ever-increasing volume of information being accumulated on this topic.

While we observe greater convergence of principles across jurisdictions and frameworks, it is apparent that intentions to move beyond principles to fully mature frameworks, toolkits and guidelines will take time, with more divergence in execution. Additionally, execution will need to ensure a balance between governance efforts and the materiality of the impact of the solutions.

We applaud the work by MAS's Veritas consortium in breaking new ground and in our opportunity to contribute to shaping the direction of this initiative. By using Veritas as a comprehensive frame of reference, while developing, contextualising and incrementally rolling out our governance framework, we can ensure we remain true to the spirit and outcomes of MAS's FEAT Principles. We will continue to keep a close eye on the evolving and emerging views from regulators and the industry to iteratively build and enhance our AI/ML governance approach at broadly the same pace as our adoption and scaling of AI/ML usage.

Sponsorship

Our strategy to ensure foundations are in place, on which we iteratively build our evolving AI/ML governance approach, is endorsed by the senior management, who are also kept updated on the internal and external developments in AI/ML governance, to ensure the bank's direction is aligned to the vision and risk management expectations.

In line with our analytics construct, the governance approach is being incrementally developed and operationalised by a collaborative team of analytics and governance representatives across the bank. This collaborative approach allows us to better balance the commercial, risk management and practical operational considerations.



Design Principles

Given the complex and evolving nature of AI/ML governance, we defined several design principles to guide the framework development and act as anchor points to ensure we remain aligned to our broader strategic imperatives.

Our fundamental focus will remain to **preserve and enhance the trust** people have in DBS through our responsible data use approach.

We will **govern our solutions based on the materiality** of their impact to customers, employees or the organisation as a whole. All AI/ML solutions will be governed with a minimum baseline, while those with a material impact use enhanced measures to ensure our governance is robust and comprehensive while balancing innovation and scaling the use of AI in the bank.

We will **iteratively develop, test and integrate** the approach **into our solutions lifecycle, leveraging existing processes, tools and ways of working** to adapt to the evolving regulatory and industry landscape and organisational growth. We solve for where we are now with AI use and not for a future that may not come to pass.

The development and delivery ethos will support **co-creation and provide learning opportunities** for teams across markets.

High Level Roadmap

We established a high-level roadmap to focus on key aspects to solve for in an iterative and agile manner to ensure we are on track in developing and operationalising our governance approach while being flexible enough to course-correct and accommodate reprioritisation where required.

Significant progress has been made in defining the scope, materiality assessment, baseline governance (including aspects of FEAT) and operating models. Additional enhanced governance for high-material solutions is currently in progress around fairness, explainability, performance and robustness, independent review, etc.



- 1 **Scope** – Determine if the models are in scope for AI/ML governance or not
- 2 **Materiality Rubric** – Assesses materiality to determine the appropriate governance level
- 3 **Baseline Governance** – Regulatory and internal considerations incorporated into a set of basic requirements for all AI/ML solutions
- 4 **Enhanced Governance Requirements** – Additional higher governance requirements for solutions with higher materiality
- 5 **AI Protocol and Inventory** – Embeds governance in solution lifecycle and inventories AI/ML solutions
- 6 **AI/ML Governance Op Model** – Roles, responsibilities and process for involved parties

Scope

There are varied views at a macro level on what constitutes an AI solution that needs to be governed. Regulatory publications consider a system consisting of multiple AI and non-AI models, human interventions and rules, etcetera, to arrive at decisions. Developing and operationalising a governance approach for such end-to-end solutions is not easy and hence it is also important that we define what will fall in and out of scope for consistent application across the bank.

MAS has also highlighted the need to have a standard way to determine the scope for governance in its recently published information paper titled 'Implementation of Fairness Principles in Financial Institutions' Use of Artificial Intelligence/Machine Learning'.

At DBS, the following factors are considered to determine the scope for AI/ML governance:

- (i) Solution is built using first-principles knowledge/business rules/logic; **AND,**
- (ii) Solution is exploratory in nature or is not intended to be used in production.

All solutions which do not meet the above criteria will be considered to be in scope for governance.

Materiality Assessment – To Determine Level of Governance

Solutions identified as in scope are assessed using a rubric to determine the materiality of its risk. The derived materiality score subsequently determines the extent of governance applied. Those identified to have a possible high adverse impact will require higher governance than solutions with a low adverse impact, ensuring governance is commensurate with impact. Thus, low-impact solutions are also not subjected to unnecessary measures that reduce their time to market. Materiality is assessed based on the below dimensions:

- 1. Degree of adverse impact:** Severity and scale of any unjustified adverse impact by the solution output, when used in decision-making, to customers, employees and/or the bank. Some of the dimensions considered are, but are not limited to:
 - Benignity of the content.
 - Possibility of withholding product/service or causing disadvantage.
 - Estimated financial loss based on a threshold.
 - Possibility of litigation, regulatory warnings, administrative penalties, etc.
- 2. Level of solution autonomy:** Extent of human involvement in the use of solution output in decision-making.

The materiality assessment approach helps to quickly, consistently and appropriately assess a solution's materiality across our analytics teams and diverse use cases. It has been developed after carefully considering different assessment dimensions, qualitative and quantitative measures, and risk impact categories. Our key learning was that adding more dimensions, measurements and granular segmentation of risk categories added complexity, subjectivity and employee toil with no measurable value in determining the level of governance required for the solution. We will iteratively review and revise the approach as we gain more experience and insights through its continuous application.

Baseline Governance – Irrespective of Materiality

All solutions in scope, irrespective of their materiality, need to apply a set of baseline governance controls to ensure minimum standards are met by every solution in the bank.

This basic set of requirements is developed considering regulatory principles and checklists such as MAS's FEAT Principles, HKMA's High-level Principles for AI, and the PDPC Model AI Governance Checklist. The baseline governance requirements cover aspects of fairness, performance, transparency and accountability.

Regional markets in which we operate can include additional requirements to ensure any jurisdiction – or context-specific requirements are covered. Below are some aspects covered in the baseline requirements:

- a)** Assessment using DBS PURE Framework (FEAT ethics).
- b)** Clear roles and responsibilities, including approving authorities (FEAT accountability).
- c)** Documentation of solution objective, analytics approach, data sources, transformation logic, assumptions, exclusions, etc. (FEAT transparency).
- b)** Materiality and justification for personal and sensitive data use (MAS fairness and transparency).
- d)** Data quality checks and remediation.
- f)** Adequate, accurate and appropriate sample representation and performance metrics.
- g)** Periodic review and remediation of data drift and model performance (FEAT fairness).
- h)** Detecting and reporting of run-time issues and contingency plan.

Enhanced Governance – High Adverse-Impact Solutions

Solutions identified as high on adverse impact to customers, employees and/or the bank will apply enhanced governance controls over and above the baseline controls with detailed and in-depth responses expected. The approving authorities for materiality score and deployment will also be more senior. Our independent validation approach is also under development for those areas not already covered.

While we incrementally develop and implement enhanced governance, the current focus is on enhanced controls around fairness and accountability in FEAT. Clear fairness requirements with detailed yet simplified guidance based on MAS Veritas methodologies are developed and Veritas's fairness toolkit tested using various use cases. The intent is to heavily leverage the overall Veritas fairness approach and implement a best-fit technical solution that can cover various aspects of FEAT for a gamut of use cases. Our considerations for tooling need to cover our longer-term functional and non-functional requirements. Our learnings from the application of the Veritas fairness toolkit are detailed in subsequent sections.

6.2.3 AI/ML Process

AI Protocol & Inventory – Embedding Governance in Solutions Lifecycle

DBS's AI industrialisation journey to reduce time-to-value across the analytics lifecycle includes aspects like: a standardised protocol for AI development and deployment; scalable, robust and stable technology infrastructure; and the right organisation structure and operating model.

DBS's AI Protocol provides a standardised process and leading practices to institutionalise the delivery of analytics use cases, supported by a knowledge hub. This tool also acts as the use case and model inventory for the bank.

All AI/ML governance requirements are embedded into the AI Protocol and used by teams across the analytics lifecycle. Use cases and related models are updated on an ongoing basis in this inventory.

AI/ML Governance Operating Model

Roles and responsibilities are defined for model owners, developers, materiality approvers, deployment approvers and independent reviewers with appropriate seniority specified. Guidelines are in place to ensure independence in roles to avoid any conflict of interest, where needed – e.g., the materiality approver should be different from the model owner. Some teams also use additional internal forums to deliberate the nuances of use cases and models. Other than the periodic updates to various senior management committees, a separate forum with senior representation from across units and markets to oversee the overall framework and controls for high material solutions is under development.

The overall operating model is in line with MAS's recommendation to have internal accountability established in the form of relevant approving parties for the use of AI/ML in decision-making, ensuring management and board oversight in the use of AI/ML, and being responsible for both internally and externally developed and/or sourced materials.

Learning & Culture

Based on DBS's transformation journey over the years, we know that the investment in awareness, training, upskilling and culture-building is not only critical but often the secret sauce to entrench change within the organisation. Our data-learning programmes aim to encourage data literacy, data governance and AI industrialisation and are designed for various audiences like data translators, data scientists, data engineers and even all employees with the depth and mode determined based on specific expected outcomes. Some examples are as below:

- Data management novice and practitioner modules for all employees on aspects like data ownership, quality, metadata, etc.
- Novice module for all employees on PURE Framework and practitioner-level content specifically for those responsible for PURE assessments.
- A workshop intended for those who need to master problem-solving with data by applying our analytics framework and collaborating with data science teams.
- A novice module and in-person practitioner training for instrumenting products and service to capture and collect data that can help to create better services for the consumer.

AI/ML governance is a new and complex area which requires significant change-management efforts to fully embed as a way of working. While testing and developing the guidance for Veritas's fairness assessment and toolkit, a strong requirement for training on qualitative concepts and quantitative methodologies and measurements across teams has emerged to ensure consistent adoption. We are also looking forward to Veritas's efforts in developing learning resources in this space. Our layered approach for learning in this space is being planned and developed:

- Novice online training specifically for data & analytics teams, but also open to all employees.
- In-depth learning for identified practitioners on quantitative and qualitative skillsets to apply governance requirements, methodologies and toolkits around fairness, explainability, performance and robustness.



6.2.4 Use Case Study

Application of AI/ML Governance to a Contextual Marketing Use Case

The bank offers multiple products and services like credit cards, mortgages, unit trusts, etcetera to consumer banking customers. Through contextual marketing, we aim to promote the bank's products and services to relevant customers via marketing channels like email, push notifications, the bank's in-app banners, etc. These communications help increase customer awareness of DBS's products and services without inundating other customers who are unlikely to be interested or eligible. Customers who are interested can then apply for the product or service through regular banking channels and will receive the same offers applicable at that given period as others who may not have received the targeted communication.

Use Case Description: We will here illustrate at a high level how the bank uses data and AI/ML to market unsecured personal loans and how AI/ML governance is applied.

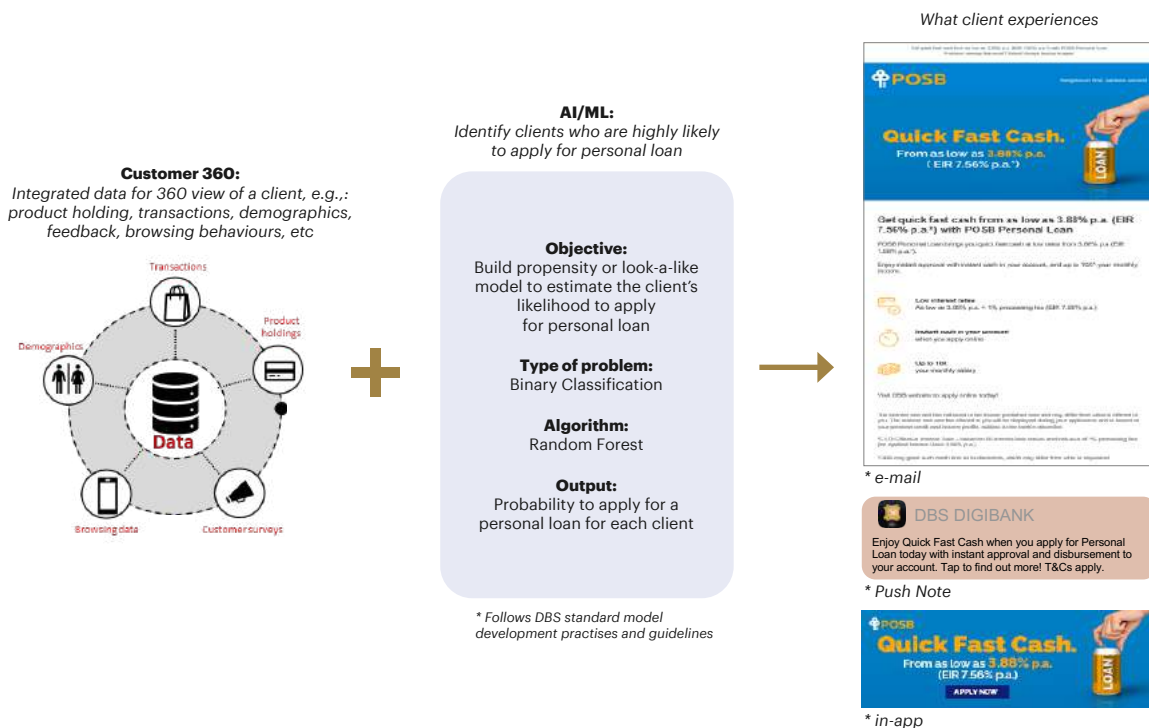


Figure 6.3.: Customer 360 – for illustrative purposes only.

- **Customer 360** as a data-mart consolidates data across multiple systems and is used to perform AI/ML and data analytics projects.
- **Personal Loan Propensity Model** assesses the clients' likelihood of taking up a personal loan. This model follows the standard analytics framework for the development and deployment of the solution.
- **Trigger Marketing Communication:** Determines a threshold by evaluating the trade-off between precision and recall to engage clients with propensity higher than the threshold through various channels.

4.2 Scope Assessment: Personal loan propensity is part of the overall contextual marketing use case and is in scope for AI/ML governance as it uses machine learning to determine who will receive marketing communications for the specific product.

4.3 Materiality Assessment: The use case is assessed to be of low adverse impact with low human involvement as below:

1. Low adverse impact:

- (a) Does not exclude or limit any customers from accessing information about or applying for the loan or providing any preferential offers.
- (b) The outcome is used only to engage clients with product information. Clients will ultimately follow the standard application process and checks.
- (c) All marketing communications follow guidelines to ensure balance between personalisation and privacy.

2. Low human involvement: Customer 360 and AI/ML pipelines are run daily in an automated manner with human involvement only on a need basis (e.g., data/pipeline issues, model enhancements, etc.)

4.4 Baseline Governance: As per DBS's AI/ML governance framework, all solutions need to meet minimum baseline governance as stated in Section 3.3. These requirements covering various stages of the development and deployment process are met by the contextual marketing use cases and models.

4.5 Fairness Assessment: As materiality is assessed to be low, the fairness assessment is conducted as per the baseline governance requirements for this use case as below:

(a) Justify use of personal/sensitive attributes and proxies: As part of the materiality assessment process, use of all personal and/or sensitive attributes and their proxies is to be justified. There are internal guidelines with a non-comprehensive list of sensitive personal attributes, but units need to assess this on a case-by-case basis. For contextual marketing use cases, based on data exploration outputs, certain demographic attributes will be included that have a high correlation to the propensity of the product's uptake.

(b) Monitoring and review of performance: Performance is reviewed during each model-scoring cycle and compared against performance during the development period. Alerts are sent to the model owner if any deterioration is observed. In case of any issues in the data pipeline, platform teams, the model owner and other business users are notified.

(c) Fall back and mitigation mechanisms: No marketing communication will be sent to clients if the model pipeline fails. Issues are raised to respective teams for resolution within internal SLAs. In case of non-resolution, the issue will be escalated to respective senior management.

For the purposes of this paper and to test and understand MAS's Veritas fairness methodology and toolkit, we assessed fairness across three attributes for identified groups using the Veritas toolkit, and after any product eligibility criteria were applied to the dataset:

1. **Gender:** Female vs. Male.
2. **Marital Status:** Single vs. Rest.
3. **Race:** Chinese vs. Rest.

The following selections were made using the toolkit decision tree to determine the fairness metric:

- Positive outcome is determined as "favourable" to the customers as those who take up the product are defined as positive labels and the model objective is to identify those who have a higher propensity to take up the product.
- "Benefit" was selected as the positive outcome as it is in no way "punitive" or "harmful" to the customers. Contextual marketing use cases, by their very nature, do not restrict customers from accessing any product or service and pose no risk of systematic disadvantage. These communications only help to increase awareness of the product or service offerings.
- Concerned group has been selected as "Eligible" as marketing communications are sent only to the customer cohort meeting the threshold and eligibility requirements set for the product.
- In terms of "Impact" to the customers, it was determined that there are "no significant consequences" nor are they considered to be "selective" as no further selection is applied on predicted positives in order to determine this specific model's outcome.

Based on the above selections to the decision tree in the toolkit, "Equal Opportunity" was proposed as the metric to determine fairness with a default threshold of 20 percent.



The toolkit results show that the solution is fair for all three attributes, as per Figure 6.4.

	Gender	Marital Status	Race
Value Distribution			
Privileged Group	53.00%	44.40%	66.83%
Unprivileged Group	47.00%	55.60%	33.17%
Equal Opportunity	-0.068 +/- 0.031	-0.128 +/- 0.035	-0.085 +/- 0.037
Fairness Conclusion			
Outcome (Equal Opportunity)	Fair*	Fair*	Fair*
Fairness Threshold	80%	80%	80%

Figure 6.4.: Results showing that the solution meets the fairness requirement.

While the model outputs are considered fair based on the above fairness outcome, the final set of customers who will receive the communication at a use case level will be a subset depending on several factors including, but not limited to, these below:

- Policies around contacting customers for marketing purposes – e.g., only one marketing email to be sent to a single customer within a prescribed number of days.
- Customers have opted out of marketing communications.
- Prioritised campaigns during a particular period.

Although the above model is fair across the various social groups, it may not be the case for all contextual marketing solutions due to the nature of targeted marketing. In such cases, it is still important to duly justify the use of personal and/or sensitive attributes and their proxies, regularly review performance and the solution's fitness for purpose, as well as the availability of fallback and mitigating mechanisms and other baseline governance requirements.

Note: The above illustration is only to show how the fairness methodology can be applied on low-materiality use cases like contextual marketing that do not involve the withholding of products or services or any preferential treatment. The internal guidance on fairness is currently being piloted across the bank and hence this should be treated as work-in-progress only.

Use Case Specific Lessons Learned and Recommendations

- The contextual marketing use cases, by their very nature, do not restrict customers from accessing any product or service and pose no risk of unjustified systematic disadvantage. It is a time-tested and accepted way of marketing products and services across industries. An industry-wide view of the types of risks to be governed and minimum controls to be in place for such low-materiality use cases is needed, without having to apply fully fledged frameworks.
- Industry-wide standards are needed on how various metrics for fairness are to be selected and thresholds set, the absence of which can show varied results across organisations due to differing interpretations and tools.
- Biases can originate from systemic factors like the behaviour of customers towards a specific product (e.g., those with higher income have a greater chance of signing up for complex investment products) or from the model development process itself (sample selection, algorithm and/or data used by the model). The current fairness metric and logic suggested for this use case measure both types of biases. Efforts should be made to determine ways to measure only the unfairness created by the model development process itself, so that targeted actions can be taken to minimise such biases.
- For most contextual marketing use cases, propensity models are not channel-specific, with a single propensity model being used to deliver marketing communication across channels. Though in the above illustration, to determine the fairness metric, the model outcome has been determined as “not selective” to predicted positives, exclusions will be applied depending on the communication channels. These exclusions will not contribute to unfairness as they can either be justified using internal policies or processes and/or involve customers’ freedom of choice to opt out.
- For the marketing use case, it was relatively easy to apply the Veritas toolkit as it is tailored for contextual marketing use cases. However, it is not clear how much effort will be required for another use case for which there is no tailored template within the toolkit.
- Clear documentation of the workflow, definitions and logics used in the tool is needed for developers to contribute to the continuous development of the open-source tool.
- Extension of the Veritas package to support non-Python-based pipelines or models like Spark will enable wider adoption.
- Detailed quantitative and qualitative trainings are needed for analytics and business teams on the full fairness methodology and toolkit, specifically on choosing metrics, thresholds, bias detection, etcetera as currently it is not straightforward for models using regression, recommendation, etc.
- Product-buying patterns vary seasonally and/or over time. Progressive development of a model based on the availability of additional data sources in an agile manner, the adoption of the latest AI/ML algorithms and frequent retraining help to improve model performance. For any solution, irrespective of its materiality, processes and platforms should support regular monitoring to identify and remediate any data drift and deterioration in model performance.

- In case the thresholds cannot be met, an assessment of the reasons, possible trade-offs and next steps should be evaluated and documented.
- Effective partnership between business, analytics and technology/platform teams in creating a roadmap and building tools to support the end-to-end AI/ML lifecycle and issue-resolution significantly improves the time to market.

Other Lessons Learned and Recommendations

- **Consistent industry-wide governance approach:** It is important to establish a consistent and pragmatic baseline across the industry on what is in scope, how materiality is assessed and necessary basic hygiene governance controls. Methodologies, tools, etcetera can further be adopted for more enhanced governance based on materiality. For example, contextual marketing is a case study used to demonstrate the Veritas Fairness Assessment Methodology in the Veritas whitepaper; the use case, by its very nature, does not restrict customers from accessing any product or service and hence has a limited systematic disadvantage. It is important to have a view on how such low-materiality use cases are to be governed without applying fully fledged frameworks.
- **Incremental vs. big bang approach:** The emerging and complex nature of AI/ML calls for an equally adaptive and nimble governance approach to consider potential risks which may have no precedent (we do not know what we do not know). A measured approach will help to avoid any future rework/restructuring due to rapid advancements in technology, regulatory views, etcetera across jurisdictions. Start small, take one step at a time and build incrementally, without using elaborate and rigid frameworks, to ensure organisations understand the risks they need to address.
- **Competitive disadvantage vis-à-vis tech organisations:** Governance needs to be balanced with the materiality of the impact of the AI-driven decision/action on the impacted party and time to market to avoid slowing down innovative applications of AI/ML. We are cognisant that much of the early guidance in this area would instruct future binding regulation. While this is only a natural progression, it is important that this is done in a manner where competitiveness is considered both at home and in the regions in which the banks operate.
- **Treatment of complex, ensemble and dynamically changing solutions or models:** Organisations are starting to adopt complex use cases and solutions where numerous, multi-layered and dynamically trained models contribute to the overall outcome. Governance of these use cases and model groups will need a non-traditional approach and guidance. This also holds true for solutions using a combination of model outputs, rules and human judgement in the final decision.
- **Talent shortage:** It is difficult to find people with both governance and AI/ML experience or interest, and to understand and translate the emerging industry or regulatory views into approaches suitable for operationalisation. Extensive options for structured learnings, discussion forums, etcetera are needed to bring the available people up to speed on various principles, requirements, methodologies and tools.

- **FEAT adoption using other methodologies:** Organisations should assess the pragmatic aspects of any existing or emerging internal and/or external methodologies and frameworks to achieve FEAT or other regulatory principles. For example, DBS's value-based PURE Framework aims to ensure the various dimensions of responsible data use are considered while planning and implementing a data use case. This framework is embedded in the bank's DNA and, while it is different from the Veritas ethics framework per se, it meets the same objectives.

6.2.5 Concluding Remarks

With the rapid increase in the number and types of AI solutions being implemented across the industry, the initiatives from MAS to provide guidance on not just “what” in terms of principles but also “how”, in the form of adoptable methodologies and tools for governance, is timely and commendable.

As organisations, it is important to keep our context, scale, vision and values in mind, while putting the principles into practice.

Industry-wide dialogues, learning and collaboration should continue to bring all organisations up to speed on the overall thinking, best practices, challenges and possible ways forward.

An industry-wide view of what is the minimum required governance would ease any confusion around how to approach the development and operationalisation of AI governance at scale for many organisations.

We look forward to continuing working with MAS to further refine or standardise the approaches in the interests of the industry and the nation.



6.3 HSBC

6.3.1 Overview

HSBC is one of the world's largest banking and financial services organisations, with operations in 65 countries, and serves more than 40 million customers. HSBC has deployed AI-based solutions across various global business and global functions with the majority of AI deployments in Wealth and Personal Banking ("WPB") enabling our Retail and Wealth Management business.

HSBC has had a holistic focus across people, process and technology. To ensure optimal benefits for its customers, while keeping them protected from the potential harm that could come with the unmanned use of emerging technologies, HSBC has put in place measures to govern the use of AI in its businesses. From establishing committees and frameworks, to training staff and taking careful considerations in the level of human involvement for its decision-making, the bank's established practices for the ethical usage of AI are progressively followed through AI model development and deployment, while adhering to appropriate global regulatory standards to stay relevant and applicable.

This paper summarises the integration of the FEAT Principles and lessons learned from our customer-selection models from our WPB line of business.

In terms of governance, we have explicitly incorporated HSBC's "Principles for the Ethical Use of Big Data and AI" (the "Principles") into our model governance process so that WPB AI models must document compliance with the Principles, check for bias at development and monitoring, and avoid restricted class attributes in models. In addition, we established the WPB AI Ethics Forum where potential ethical issues identified in AI solutions can be addressed.

From a people perspective, we have rolled out both instructor-led and web-based trainings on our Principles and the importance of ethics in AI and big data.

From the process and technology perspective, we are deploying a cloud-based solution that brings together data from WPB to roll-out AI/ML services using our Data Science Workbench.

6.3.2 AI/ML Governance

The Journey to Integrate RAI Principles, Governance & Risk Control

6.3.2.1 Values, Principles & Governance

At HSBC, Group Level AI/ML guidance and recommendations are provided to AI/ML use cases. This guidance and the recommendations are closely tied to our HSBC values that are embedded at all levels and in everything we do. The "Principles for the Ethical Use of Big Data & Artificial Intelligence" seek to:

- (1) Increase awareness of the ethical implications of big data and AI.
- (2) Bring uniformity and predictability to how we make decisions about the use of big data and AI across the bank globally.
- (3) Encourage questioning and challenges before new uses of big data and AI are adopted.

The seven core Principles are:

- Act consistently with HSBC's values.
- Protect privacy.
- Start with a clearly defined purpose.
- Address unfair bias and decision-making.
- Be responsible for AI.
- Adapt governance to meet emerging needs.
- Contribute to the development of best practices.

The above core principles are being integrated into the model risk management framework as part of the supplementary guidance to our Model Risk Policy and Standards, covering the practices to be followed for Artificial Intelligence/Machine Learning (AI/ML) models.

Our strategy is to supplement the ethical usage of AI-related minimum standards with our existing Model Risk Management Framework and incrementally enhance our AI/ML governance approach with the ever-evolving views from regulators and the wider industry in this space.

Materiality Assessment and Risk-Tiering

Due to fundamentally different design principles, AI risk is exposed to multidisciplinary risks, broader than model risk, including compliance, technology, legal, operations, data and other risks. AI risk impacts all categories of risks defined in HSBC's Risk Taxonomy and, as such, requires a horizontal risk management approach that integrates RAI governance framework elements to different areas in the risk management framework.

AI risk is evolving within the industry. It requires significant education, upskilling in practice and constantly challenging ourselves against external standards to keep on top of the risk and fully embed that knowledge into our governance processes.

HSBC has a very comprehensive Risk Management Framework that supports our global principles and guides all that we do. The global principles are embodied in our strategy, our values, how we conduct business and how we manage risk. Naturally, this approach is inculcated in every project, use case and task that we undertake.

The AI/ML models that satisfy the model attributes set out in the model risk policy fall within the governance of model risk management. Models and model use cases are assigned a model risk materiality rating (MRMR) based on the impact.

The group impact scale is used to assess the plausible impact on the business within the next 12 months. Four impact scales are assessed to arrive at the group impact rating of Minor, Moderate, Major or Extreme. The impact dimensions are: Direct Financial, Customer Detriment, Regulatory Censure and Reputational.

Organisation Policy and Control Measures and How They are Commensurate with Identified Risk Levels

A key part in managing the use of AI models involves having a set of policies and procedures. HSBC uses a layer of policies and standards for model risk management, which sets out the high-level minimum control requirements for identifying, measuring and managing model risk and supports consistent application. The model lifecycle is applied to individual models and is the core process around which the control framework is built. HSBC adopts a “risk-based” approach. The depth and rigour of the model risk management activities are commensurate with the nature of the model and the materiality of the use case. The depth and rigour also vary over time.

Additionally, ethical challenges with AI models may not line up with traditional risk materiality ratings. Simple chatbot, customer propensity models or image recognition models which do not have a material financial or operational impact on the bank may not have a high model materiality rating, but when customer-facing can have a significant risk when reviewed against our Ethical Principles. Therefore, the model developer must perform an ethical check of the model and include some form of evidence during model delivery and testing, documented in the model documentation, and prior to the model’s approval to show it is evaluated against HSBC’s Ethical Principles.

Key lessons learned from the WPB-based customer selection models:

- Clear articulation of the reasons behind the creation of the model and expectations on how the model output will be used.
- Models are logical, developed robustly and appropriate for their intended purpose.
- Models are well documented, including roles and responsibilities identified, and recorded in the Global Model Inventory System with details complete and up to date.
- Development data is evaluated for quality and bias to ensure it is representative of the model scope and complies with the data management framework.
- Models are subject to periodic assessment to ensure they remain appropriate for use, and monitoring is selectively used to track changes in model risk.
- Where required, models are selected for independent validation before implementation and on an ongoing basis.
- Models have received appropriate approval from the relevant authority or accountable individuals before implementation.

Additionally, HSBC provides model oversight through the model risk committees and model oversight forums run at the group and regional level.

Roles and Responsibilities of Stakeholders

HSBC’s values are integral to all employees, and the role of HSBC’s values in our daily operating practices is fundamental to our culture. One of the values we live by is, “Listen, treat people fairly, be inclusive and value different perspectives”. Hence, regardless of gender, race, education, skill-set, disability status, etcetera, inclusion is part of our DNA.

From the start, HSBC understood the importance of keeping a close eye on the various stages of its AI models' development to ensure reliability and accuracy. For this, the bank's Model Risk Management Framework sets out roles that are expanded across the model lifecycle where distinct responsibilities are assigned in line with each policy requirement. The Model Risk Management Framework is based on the three lines of defence approach and must be followed across HSBC:

- **First Line of Defence (LoD):** The First LoD is responsible for identifying, owning and managing the risks of the model used for their respective business and/or function. Model sponsors, model owners, model users and other supporting individuals who are managing the model's development, implementation and continued use, monitoring its performance, and controlling the potential risks of using the model, belong to the First LoD.
- **Second Line of Defence:** The Second LoD is responsible for setting the overall model risk management framework and providing an independent challenge to the model risk-management activities of the First LoD.
- **Third Line of Defence:** The Third LoD is provided by the Model Risk Audit team, which provides independent assurances to internal stakeholders and regulators about model risk controls, and performs quantitative and/or qualitative reviews of the design and operating effectiveness of the First and Second LoDs' controls, and assesses their compliance with regulations.

Our adoption of the HSBC Principles is embedded in model governance and is a frequent topic at town halls, risk committees and the model oversight committees. Senior leaders regularly review and opine on ethical use cases related to AI.

Measures of Progress in Rolling Out AIDA Governance

- Conduct model risk trainings to front-line teams to improve awareness of model risk and their adherence to the governance framework.
- Keep staff abreast of the latest regulatory developments on AI/ML and changes to the Model Risk Management Framework.
- Review model risk information sharing and ensure engagement between model stakeholders is complementary.
- Prepare model risk updates to regional forums and committees on the existing model landscape, including AI/ML models.
- Regularly review our practices with third-party experts to ensure we are not missing any key elements, and remain at or above the industry standard.

6.3.3 AI/ML Process

One of HSBC's principles on the ethical use of AI is to "Contribute to the development of best practices", and one of the ways in which this is achieved is with documentation. Documentation helps to capture best practices, experiences and knowledge from the AI solutions implementation as well as the lessons learned.

Information about the model, use cases, role-holders and model artefacts are maintained in the model inventory. As each model goes through its model lifecycle management, model artefacts (like the Terms of Reference, Model Documentation Template, Model Approval Form, Implementation Report, Model Assessment, Decommission Plan, etc.) are registered in the model inventory. These artefacts expose the risk and document the agreements and basis of decisions that were made to demonstrate compliance with the minimum control requirements and completion of each stage in the model lifecycle.

Maintenance of AI/ML solutions is more complex than maintenance of traditional software, because the simple fact that the entire AI/ML system is working does not necessarily mean it is doing what it should. Without proper monitoring and internal controls, failures can go unnoticed until the business damage is done. Hence, setting up and fine-tuning internal controls and governance around AI/ML solutions is an integral part of HSBC's AI/ML solutions.

HSBC has the following controls for model deployment:

- A model ToR (Terms of Reference) must be approved during the planning stage.
- Model reviews where required, either peer-reviewed or 2LoD-reviewed, depending on model materiality, must be completed before model approval.
- Model documentation must be completed and submitted for approval before deployment.
- Model implementation test and approval must be completed for the model's go-live deployment.

HSBC has controls for model maintaining and tuning:

- All models where required must be monitored periodically following model monitoring standards.
- Any model changes, including model tuning and demise, must follow the Model Change Management and Decommission Standards by documenting the evidence and obtaining the appropriate approvals.
- With guidance from HSBC Global, model governance continues to evolve as practices improve. We have an internal review process to comply with the group ethics principles, and have rolled out trainings to boost awareness of AI-ML ethics and governance.

The Journey to Increase RAI Culture & Training

Adoption of the Principles has been strongly encouraged from the top. HSBC's senior leadership participates in the WPB AI Ethics Review Committee where specific use cases are brought for discussion and debate. We have also rolled out both instructor-led and web-based trainings on our Principles and on the importance of ethics in AI and big data.

AI/ML model development and use is accelerating across the bank and these largely fall within the WPB business, with models used for customer selection, fraud detection, retail credit decisioning and collections, and financial crime. Model risk committees, chaired by the group of regional Chief Risk Officer and attended by the global or regional heads of business and functions, as well global and regional model oversight forums, are in place to provide model oversight and update senior management on the existing landscape of AI/ML models, as well as on enhancements to the Model Risk Management Framework and regulatory developments.

In HSBC Singapore, this is predominantly done through the Country Data Executive Council (CDEC), which provides country-level oversight on data management and governance, and is an avenue to share and discuss industry-related insights and initiatives such as RAI.

Model risk management provides training to key model stakeholders to raise awareness of model risk in the business, and to ensure their commitment to good risk management and that regulatory expectations are met. In particular, it is a requirement under the Model Risk Policy that policy-role holders (model sponsor, model owner and model users) undertake role-specific training. Additionally, in response to the increasing maturity of model risk across the Asia-Pacific region, opportunities for bespoke sessions with Asia-Pacific Markets (ASP) markets are identified to cover fundamentals of the Model Risk Management Framework, as well as model limitations, assumptions and issues of models used in house. There continues to be a collective effort to improve the AI/ML-based framework and training for model developers, along with independent model reviews on AI/ML methodologies.

A web-based training program on "Principles for the Ethical use of Big Data and Artificial Intelligence" was rolled out through HSBC's learning platform.

Additionally, HSBC took efforts to train model developers and the independent model review sub-team on AI/ML model methodologies to ensure that specific issues are addressed in a timely manner. This led to the creation of several cross-functional working groups, a noticeable upskilling of the model development and validation teams, and, more importantly, a collective effort to improve the AI-based Model Risk Management Framework.



6.3.4 Use Case Study

In our internal and external training courses, we rely on industry and internal use cases which demonstrate how AI can go wrong if not properly managed. These interactive sessions involve training participants to identify the areas where issues can emerge, and share outcomes and the impact of unmanaged AI on business and society.

During the development and use of the model, the model developer and model user respectively are responsible for ensuring adherence to HSBC's Principles for the Ethical Use of Big Data and Artificial Intelligence.

We find that two of these principles will generally cause us either to alter or change our approach to modelling:

- **Act consistently with HSBC values:** We have seen use cases where data exists to perform a specific model or analysis, but after reviewing whether that analysis will align with our values and societal good, we have decided not to progress with that analysis.
- **Address unfair bias and decision-making:** Models must avoid using restricted class variables explicitly in the model, and are checked for implicit bias against these variables. In some cases, this has led us to rework models which showed a biased outcome, leading to additional work and sometimes less predictive models.

Interpretability is also especially important for machine learning models. Many machine learning algorithms have been labelled as “black box” models because of their inscrutable inner workings. Such algorithms of machine learning include gradient-boosted ensembles (GBM), artificial neural networks (ANN) and random forests, among many others. What makes these models accurate is what makes their predictions difficult to understand: they are very complex. This is a fundamental trade-off. Unfortunately, more accuracy almost always comes at the expense of interpretability, and interpretability is crucial for business adoption, model documentation, regulatory oversight and human acceptance and trust. Interpretable, fair and transparent models are encouraged as part of the oversight of the use of AI-ML models.

If AI/ML systems make mistakes or have undesired consequences, it is not acceptable to use as an excuse that “the algorithm did it”. HSBC has a two-tier approach to dealing with this problem. In certain use cases (e.g., pricing decisions) that directly impact customers or are required by regulators, the methods with transparency and interpretability should be used. Sometimes this will mean that modellers may select not to use the most predictive deep-learning algorithms but instead settle on more basic machine learning algorithms to enable transparency. For other use cases (e.g., marketing message prioritisation) that are less material in terms of their business impact and risk, less model transparency or interpretability may be tolerated, and focus should be put more on the model accuracy and test of its “fitness for purpose”. However, even in scenarios where complex models are allowed, HSBC recommends industry-leading approaches such as Shapley values to interpret the model results, and these are embedded in the internally developed modelling tools.

6.3.5 Concluding Remarks

AI governance is a new discipline and continues to evolve. The organisations starting their AI journey may need to develop basic AI governance capabilities to ensure they get the most from their AI projects.

The growing expectation in political, regulatory and public domains is that businesses should have in place robust ethical frameworks and controls to protect consumers and maintain effective and competitive markets. The public debate is now framed around how companies should behave, rather than merely how they can behave to ensure they stay within the law.

HSBC considers AI governance as an enhancement to its existing governance process. The bank continues to keep a close eye on the evolving and emerging trends in the AI/ML space and is progressively enhancing AI/ML governance practices. HSBC ensures that AI results can still be overseen by personnel so that their commitment to customer service will not be compromised. By using AI in an ethical manner, HSBC will be able to improve its operational efficiency and provide a safer, more seamless customer experience while upholding its values of trust, accountability and fairness.

As AI is a new discipline and is continuously evolving, its learning process also needs to be continuous to strive for improvements in RAI implementation strategy.

Veritas has helped us to collaborate with leading financial institutions and MAS to carve out why Responsible AI is inevitable. From this collaboration, HSBC is looking at how to build trust-enhancing actions in the adoption of AI that translate our core values into a practical AI governance framework.



6.4 OCBC

6.4.1 Overview

OCBC prides itself on being a leading data-driven organisation. Our digital transformation push has accelerated the deployment of Artificial Intelligence & Data Analytics (AIDA) projects across OCBC Group to support both internal and customer-facing use cases.

Whilst we are keen to quickly adopt AI techniques to capture business opportunities, we recognise that models must be deployed responsibly to minimise potential harm to our customers and brand. OCBC aims to find a balance with our Responsible AI implementation so that our fairness models can be easily and consistently applied across OCBC Group – but without causing excessive administrative and coding overheads for data practitioners.

6.4.2 AI/ML Governance

As OCBC continues its digitalisation transformation, we must ensure that customers are not unfairly disadvantaged by our increased use of AIDA based decisioning.

OCBC's core values, **LIFRR** (**L**asting Value, **I**ntegrity, **F**orward-Looking, **R**espect, **R**esponsibility), influence the way we conduct business. Similarly, when implementing AIDA systems, the LIFRR values function as overarching guiding principles (see box).

OCBC CORE VALUE	MODEL PRINCIPLES
L ASTING VALUE	• Models should focus on long-term value initiatives. • We should not use scarce AI resources on low-value initiatives. • We adopt prudent risk-taking and ensure that we have a process for determining when a "human is in the loop" is required for AI decisions.
I NTEGRITY	• We assume everything we do is in full public view. We avoid building models where we would not be comfortable if the public were aware of such a model. • Customers should be aware when they are interacting directly with an AI application.
F ORWARD-LOOKING	• We embrace technological advances and should maintain our AI knowledge, techniques and platforms towards the leading edge of developments.
R ESPECT	• We treat all our customers with respect and humility and ensure that we minimise bias within our model decision-making. • We care by listening to our customers' needs. Customers should be given a channel to raise an objection if they were declined by an AIDA-based decision.
R ESPONSIBILITY	• We are reliable and give the right advice. • Our models must be rigorously tested to ensure accuracy. • We ensure model-based decision-making is approved by appropriate internal authorities.

Our focus is on the enhancement of existing governance frameworks and augmentation of the enforcement capacity to include Responsible AI Principles. This is to enable enterprise-wide consistency in the assessment of all models – regardless of whether they are developed using traditional algorithms or the latest machine-learning approaches.

In the long run, we believe this facilitates the responsible and trusted use of AIDA systems across OCBC – enhancing our ability to build trust and better serve our customers’ needs.

Adapting Existing Frameworks and Practices

OCBC has internal governance frameworks in place to assess and set standards of appropriate use of models. Rather than developing additional governance functions, existing frameworks were augmented to enable effective management of Responsible AIDA considerations.

Key enhancements being introduced to our AIDA governance frameworks include:

- Integration of FEAT considerations into the Data Management & Governance Framework, which subjects material models to higher standards for data provenance and data quality.
- Developing a standard FEAT checklist to be completed for all models, which creates a common approach for documenting and assessing FEAT compliance of all models.
- Introduction of a top-of-the-house committee comprising senior stakeholders to provide management oversight and guidance for the responsible use of AIDA systems.
- Publication of an OCBC Veritas fairness module within our Machine Learning ML-Ops environment – which is used by data scientists to automatically test models for fairness compliance.
- Enhancements to the enterprise model inventory to capture additional metadata related to FEAT compliance for each model.

In recognition of the rapidly evolving market practices on the responsible and trusted use of AIDA, we have processes in place to continually review and update the relevant frameworks, where necessary.

Risk-based Governance Approach

A key principle was to take a risk-based approach where the depth of validation of controls required for an AIDA system are commensurate with the level of potential harm arising from unintended biases from the system. This risk-based approach allows OCBC to balance time-to-market delivery of service innovations while putting in place appropriate levels of control to manage the associated risks.

Next Steps

With the customers as our focus, the responsible and trusted use of AIDA systems will require continual engagement with stakeholders within the bank, industry partners and peers to establish market best practices and manage trade-offs. Our ongoing partnership with the Veritas consortium is one of the key means towards this end.

6.4.3 AI/ML Process

Fairness Integration in OCBC

AI Lab is the Centre of Excellence for Data Science and AI within OCBC Group. AI Lab piloted OCBC's initial approach to the implementation of fairness within its modelling processing.

AI Lab's fairness integration follows a standard approach across projects.

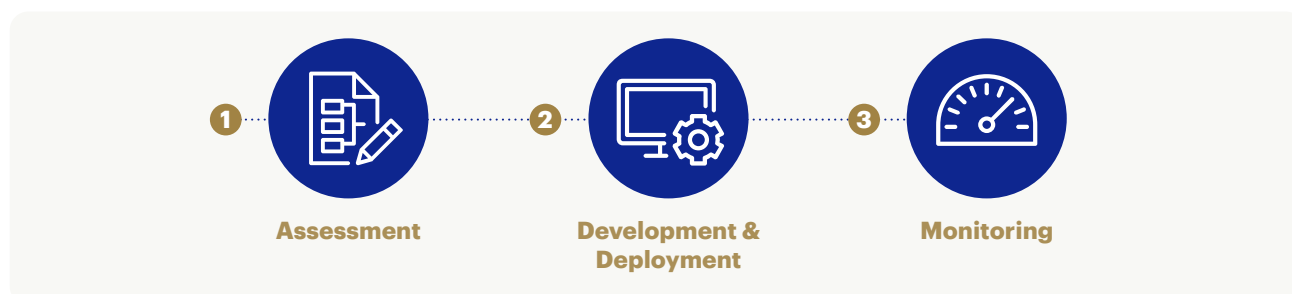


Figure 6.5.: Overview of key components in OCBC fairness assessment process.

Step 1: Assessment

Before developing an AIDA system, it is necessary to assess (1) the materiality of the system, and (2) its fairness implications.

Materiality of the System

First, the AI Lab assesses the material risks associated with an AIDA system and whether there is potential to bring harm to the bank and its customer base. A scorecard approach is used to quantitatively score first the potential severity of risk along dimensions such as financial risk, operational risk, reputational and regulatory risk, and second the likelihood of the above harms based on the estimated frequency of occurrence. Finally, the model materiality is classified as either *baseline*, *important* or *material* by considering the combination of both severity and likelihood (Figure 6.6.).

Model Materiality Matrix				
Likelihood of Harm		Severity of Harm		
		LOW	MEDIUM	HIGH
	HIGH	IMPORTANT	MATERIAL	MATERIAL
	MED	BASELINE	IMPORTANT	MATERIAL
	LOW	BASELINE	BASELINE	IMPORTANT

Figure 6.6.: Model materiality matrix.

For each level of materiality, AI Lab has defined standard control guidelines that constrain various aspects of the development process – such as the model’s algorithms that can be employed, validation techniques, the need for explainability and ongoing monitoring requirements.

Use cases that are deemed to be of baseline materiality permit greater flexibility in the use of black-box models or open source pretrained-language models, accompanied by simple validation procedures. However, material use cases will be restricted to algorithms that are explainable, coupled with robust validation and ongoing monitoring to track drifts in the input data and model performance/fairness.

Fairness Implications

Having determined the model materiality and the standard model development controls, AI Lab will examine if the model could potentially cause systematic disadvantage or be otherwise harmful to groups of individuals/corporations who are impacted by a model’s decision.

While all our models are subject to the same fairness assessment, we allow the fairness constraints to be applied with a differentiated level of rigour depending on the use case (Figure 6.7.). For example, a model used to detect instances of anti-money laundering (AML) should be optimised to produce the best possible prediction outcome regardless of the counterparties’ personal attributes. As the model outputs are used by internal AML investigators rather than by external customers, fairness considerations are less critical in this use case.

On the other hand, a model which prevents an individual from receiving employment or access to insurance could be potentially life-changing, so a formalised process of measuring and monitoring fairness would be applied.

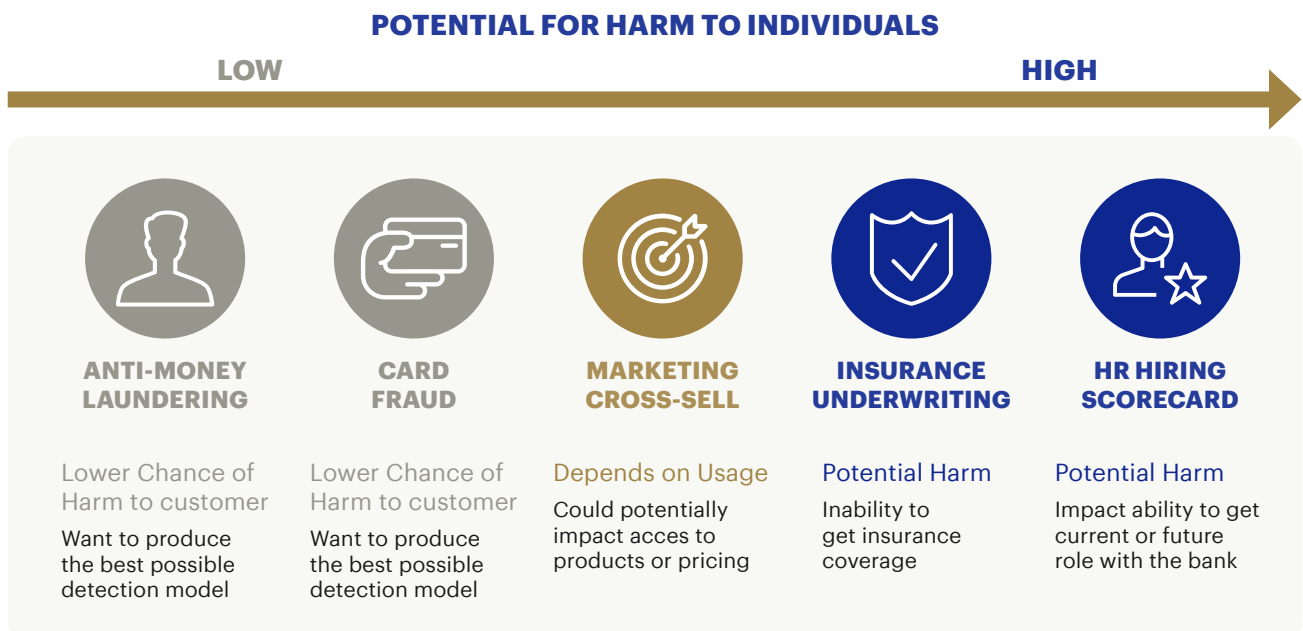


Figure 6.7.: Spectrum of use cases that receive differentiated levels of fairness controls.

To assess if systematic disadvantage could be introduced by the system, the project teams are required to consider the following vis-à-vis the use case:

- Does the model impact the customer's eligibility for products/services?
- Will a wrong prediction potentially impact the customer's long-term reputation?
- Does a wrong prediction potentially have a negative psychological effect or result in physical harm to the customer?

As a provider of economic resources, we consider the customer's ability to access our offerings as a key fairness consideration. If the model is used as an input to determine an individual's ability to access a product or service (e.g., by controlling eligibility or pricing), it is deemed a potential source of systematic disadvantage. These models should be treated more stringently than models that only support the promotion or awareness for the service (e.g., chatbots or marketing models).

If it is deemed that the model does not cause systematic disadvantage, AI Lab does not require further review of the trade-off between model performance and model fairness. Being able to maintain a lightweight process for non-harmful models is a key factor in making sure that the organisation can develop and deploy such models quickly without incurring unnecessary overheads.

If the model is assessed as having the potential to cause systematic disadvantage, then the trade-off between performance and fairness needs to be analysed. Working together with the business process-owner, data scientists first need to identify all possible protected groups. For each of these groups, they are required to determine the acceptable level of fairness (according to the selected fairness metric), and the maximum acceptable degradation in model performance. To guide this process, AI Lab requires the team to answer several questions to demonstrate a shared understanding of the fairness analysis, mitigation measures and trade-offs as part of the project documentation:

1. What are the fairness objectives for the system?

This requires the team to articulate the source of any possible systematic disadvantage.

2. Does the data used by the system contain properties that may impact the system fairness?

This requires the team to perform a comprehensive data exploration and document any inherent biases or gross imbalances in the data that could contribute to systematic disadvantages.

3. How is this impact mitigated?

The team will need to articulate the mitigation plan, possibly using a combination of measures – e.g., rebalancing the data, employing explainable models, using fairer measures of performance, and tracking a suitable fairness metric. These mitigation measures should be meaningfully demonstrated through experiments where possible.

4. What personal attributes are relevant to the use case? What are the justifications for using these personal attributes within the model?

This requires clear articulation as to why a personal attribute is used in training the model – e.g., a feature could be a strong predictor of the target variable, without which the business use case may not be compelling. This discussion should also encompass scenarios where personal attributes may be omitted from the model, but could still be needed to track fairness during model validation to avoid assuming “fairness through unawareness”.

5. Is there a trade-off between performance and fairness for the system? If so, is the current fairness-performance equilibrium acceptable?

This requires business stakeholders to recognise that a performance trade-off may need to be made to meet the fairness objectives and articulate the maximum performance trade-off that permits the use case to remain relevant to the business. A summarised output from this discussion is provided in Table 1.

Fairness Analysis			
Performance Metrics	Balanced accuracy		
Permitted performance trade-off (%)	10%		
Fairness Objectives			
Protected feature	Protected group description	Fairness Metric	Fair Threshold (%)
Gender	Female	Equal opportunity	20
Race	Minority	Equal opportunity	20

Table 1: Examples of protected features and groups, associated fairness metrics and permitted performance trade-offs.

Regardless of whether the system has potential for harm, the business process-owner is required to approve the assessment – as the decisions are often business-driven rather than being purely technical.

Step 2: Development & Deployment

Once the fairness assessment has been approved by the business process-owner, the constraints required for performance and fairness are registered within the OCBC Model Management Platform.

As data scientists develop their models iteratively, the model metrics, including performance and fairness scores (generated by the OCBC fork of the Veritas toolkit), are submitted to the platform. Based on the submitted results, the model management platform shows whether a trade-off between fairness and performance is possible without breaking the given constraints in Table 1. If a trade-off is possible, the platform calculates the optimal thresholds for each protected variable to maximise performance while maintaining a level of fairness within the provided constraints. Through this system, a model can only be deployed if it is able to produce an acceptable performance-balance equilibrium.

Step 3: Monitoring

Once a model is deployed to production, there is potential for the model performance to degrade over time as the model is retrained on newly labelled data. Similarly, model fairness may also potentially degrade over time.

Every time a model is retrained on new data, model fairness is recomputed and compared with the constraints that were previously determined. If the model is no longer able to satisfy these constraints, an automated alert is sent to the model developers and approvers. A decision is then made to determine whether the model usage should be suspended until a new version is released to fix the model fairness degradation.

6.4.4 Use Case Study

Enabling Model Fairness with Technology

OCBC has deployed several key technology components to enable streamlined operationalisation of fairness testing (see Figure 6.8.).

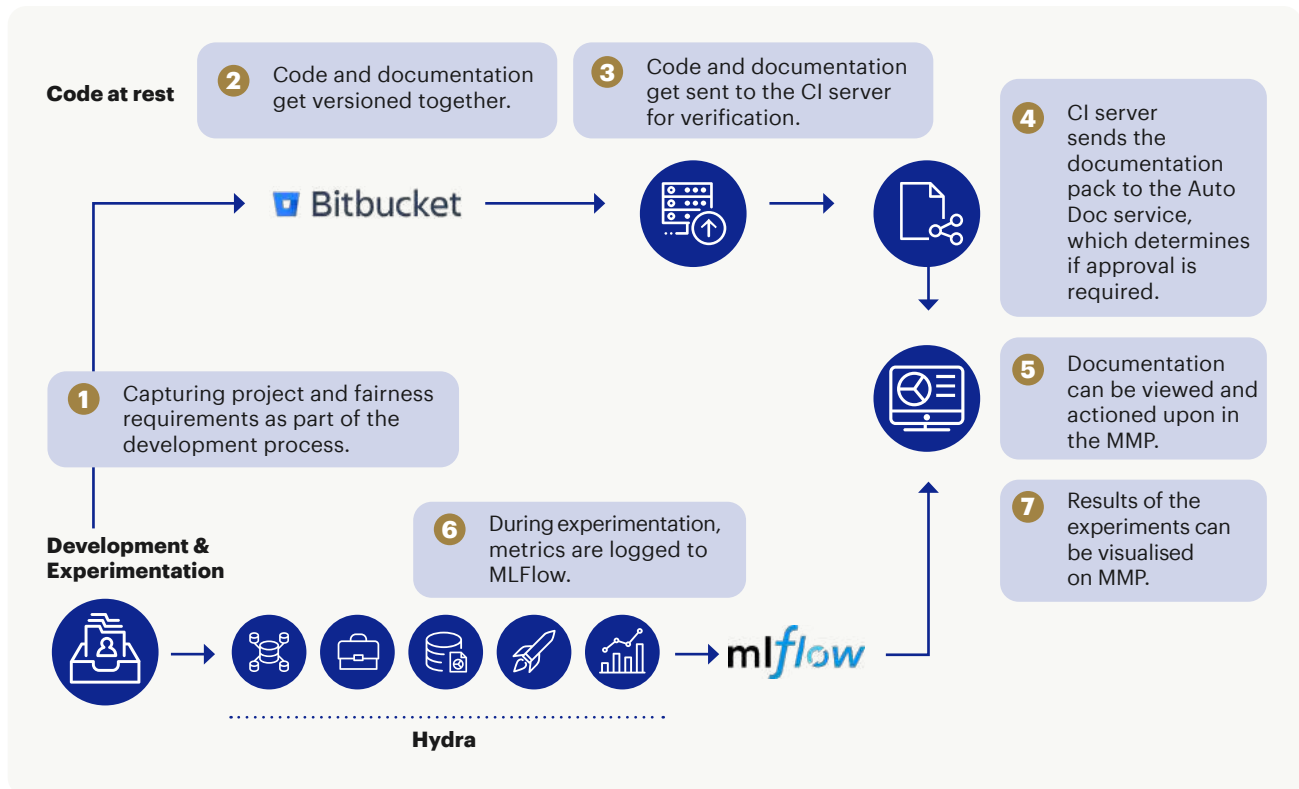


Figure 6.8.: OCBC's MLOps workflow.

1. Auto Documentation Server

A key AI Lab principle is that documentation of models should be done as part of the code – and not as a separate exercise.

Model documentation is first written in standard templatised markdown files and is stored, versioned and maintained together with model code. This is maintained within OCBC's enterprise versioning tool.

As the model and the documentation are maintained jointly in the same versioning tool, each version of the documentation is linked with a version of the model code.

Every time a change to the “documentation code” is submitted, OCBC's Enterprise Continuous Integration (CI) pipeline is triggered. As part of the CI pipeline, the documentation code is converted into an interactive HTML documentation set, which is submitted to the Auto Documentation Server (ADS). The ADS is a platform where each version of documentation for every model can be read, exported and approved.

As documentation code templates are standardised and fully machine-readable, the CI pipeline is able to programmatically capture fairness requirements and constraints described in the documentation. These fairness requirements are sent to the Model Management Platform (MMP) for further processing.

2. OCBC Veritas Toolkit

In order to integrate the Veritas toolkit into its larger ecosystem, and to remove manual steps in the fairness calculation and review, OCBC has forked the open-source Veritas toolkit code and created its own internal version of the toolkit.

Enhancements made to the toolkit include:

- Direct automated submission of fairness metrics to OCBC's Model Management Platform, instead of manually uploading the results.
- Faster fairness calculation – up to 10x performance compared to the original toolkit.
- Native integration with the MLFlow open-source framework.
- Support for the Ray-distributed computing framework, making the toolkit able to compute fairness for use cases where the data set is too large to fit in the memory of a single machine.

OCBC has contributed some of these enhancements back into the open-source toolkit, becoming the first organisation to contribute in September 2022. OCBC performance enhancements have been released as part of the Veritas Toolkit Version 1.1.

3. “Hydra” MLOps Framework

Running AIDA systems at scale requires significant automation and standardisation.

In order to face the increasing degree of complexity of maintaining a large number of systems in production, OCBC has built a Machine Learning Operations framework, “OCBC Hydra”. The purpose of the tool is to distil the bank's best practices into a standard API providing “the OCBC way” to execute a task.

Every time a model is trained, the framework automatically invokes the relevant Veritas toolkit functions to calculate model fairness and submits the results to the Model Management Platform. Using an automated framework reduces the “process overhead” for the data scientists, and at the same time ensures that fairness is continuously computed for each model.



4. Model Management Platform

The Model Management Platform (MMP) is a central repository where each model is registered, monitored and approved. The fairness requirements are captured from the model documentation and sent to the MMP.

As models get developed, the MMP compares their metrics to the fairness constraints submitted in the model documentation and computes the optimal performance-fairness trade-off for each protected variable according to said constraints. After a model is deployed in production, the MMP keeps receiving its performance and fairness metrics, and automatically alerts the relevant parties if the model is not able to fulfil the fairness requirements anymore.

The MMP serves as the central model and metrics repository, guaranteeing that model fairness can be tracked over time and audited when necessary.

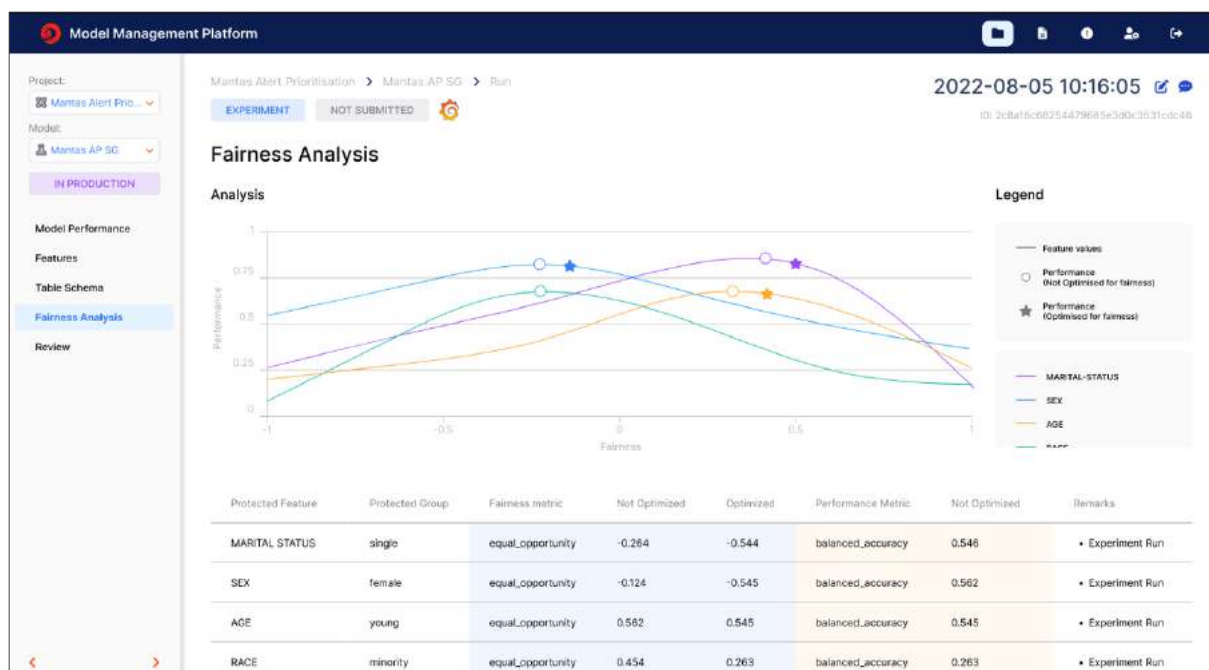


Figure 6.9.: All the metrics associated with each model's runs are also captured. Shown here are the fairness analyses for three protected features, and the associated trade-offs which are automatically computed.



Data Literacy Training

OCBC is developing a dedicated data literacy programme (“Lead with data”) to equip employees with the necessary skills to maximise use of data across the organisation. The importance of Responsible AI is a key component of the bank’s literacy training. Training is being developed for specific target personas (see box).

PERSONA	LITERACY NEEDS
General employees	Literacy Level – Awareness General employees may not need to work directly with data or AI. The focus of our training is to cultivate awareness of the potential opportunity and value of AIDA – but also of general concepts relating to the impacts of model bias. At this level, it is useful to share case studies from the literature comprising examples where outcomes with ethical concerns resulted from AI products. These concepts are touched upon within OCBC’s awareness-level online data training material.
Product owners and business units	Literacy Level – Literate For the business users, the data literacy objective is to establish a common language and understanding with their data science counterparts. We focus on ensuring our product owners understand the sources of adverse impact (such as features and models) and the best metrics used to quantify them. Business managers are given a qualitative understanding of ML concepts such as feature importance and confusion matrices to enable them to make educated decisions on whether a feature or metric is justifiable for a particular use case.
Data scientists	Literacy Level – Mastery Data scientists will also have to develop a deeper appreciation of the various FEAT metrics. In our experience, FEAT metrics are seldom comprehensively covered in the standard data science curriculum – hence the need to educate our data scientists on how these metrics are computed from the confusion matrix. Our data literacy programme also familiarises data scientists with the Hydra framework and the supporting MLOps workflow to allow them to leverage the OCBC fairness components within their model deployments. We plan to extend the “Lead with Data” literacy programme further in H1’23 by adding a specific “Fairness & AI Ethics” module.

6.5 Singlife

6.5.1 Overview

Singlife is more than just an insurance company. Merging the InsureTech capabilities of Singapore Life Pte. Ltd and the 40-year legacy of providing insurance and investment solutions, advisory capabilities and strong customer base of Aviva Ltd, Singlife is a technology-empowered financial services provider offering a comprehensive suite of insurance and investment solutions. In its mission to provide customers with “A Better Way to Financial Freedom”, Singlife will embrace disruptive technologies and innovation through the use of Artificial Intelligence and Data Analytics (AIDA).

Integration Achievements Overview

Like any new technology or idea, the adoption of AIDA carries with it unique risks that require appropriate management, such as the risks of unintended discrimination potentially leading to unfair outcomes and issues relating to consumers’ knowledge about how AIDA is involved in making decisions about them. Singlife recognizes the importance of ensuring that AI-driven decisions do not disadvantage any individual or group of individuals without justification. We have defined the initial AIDA governance process by setting up governance, process and technology alongside key risk indicators and controls. A summary of the learning and achievements worth highlighting as part of the integration process is listed below:

- Establishing AIDA governance framework with five key areas; the FEAT Principles have been incorporated into these.
- Model-approval process has been established.
- Key risk indicators for AI have been identified and included in the operational risk process.
- Baseline MLOps process has been established; further refinements are still needed in the end-to-end process.
- The Veritas toolkit has been evaluated, though we realised that the toolkit could not be used for our chosen use case as the model is an unsupervised model.
- Transparency assessment framework has been established in our model diagnostics. We have utilised a combination of SHAP/PDP along with the usual evaluation metrics to further explain the predictions by the model.

6.5.2 AI/ML Governance

There must be internal governance structures and measures to ensure oversight over Singlife's use of AI. Singlife has an existing risk management framework and internal governance structures and processes as controls within the framework. Singlife leverages on and are enhancing these governance structures and processes to include AI governance taking reference from the Model Artificial Intelligence Governance Framework published by the Personal Data Commission and integrating the FEAT Principles to take into account the challenges and considerations of responsible use of AI. As this is a developing area, Singlife may introduce additional processes, where necessary.

Key enhancements include the following:

- Integration of the AI governance processes and FEAT principles into Singlife's risk management controls, in particular in the areas of data provenance and data quality and customer disclosures and right to know how their data is being used.
- Internal Governance Structures and Measures – having a documentation and model-approval process documenting each of the FEAT principles, the model evaluation metrics, pre-deployment testing and deployment checklist proposal for approvals from the Head of Data Science, Chief Customer Officer and the respective business sponsor.
- Risk Identification and Control Measures – key risk indicators for AI have been defined in accordance with Singlife Group's risk management framework which will monitor model availability and performance tracking including the application of the FEAT Principles. We have identified the following key risks related to AIDA and have controls to mitigate them.
 - The risk that the model is unavailable for a given period.
 - The risk that the model returns a bad prediction for a given sample.
 - The risk that the model accuracy or fairness decreases over time.
 - The risk that the skills necessary to maintain the model (i.e., data science talent) are lost.

For every risk control we assess the probability of occurrence and identify contingencies to deal with it. For every model going into production, we define the complexity and the impact of the model to identify the materiality impact and then define the turnaround time and service-level agreement needed for recovery when the model is not available. When the model is not available, we define contingencies like supplementing the model output with business rules. For example, if the model API is being called from the web engine, and if the API is down or generates no response within the stipulated timeframe, then in order to reduce the waiting time on the front-end, we will replace the model output with business rules. We are currently defining materiality on a use-case to use-case basis, and it would be beneficial if the regulators could share some standard practices around defining materiality for AI/ML models.

Even though we have implemented the risk control measures for AIDA, we recognise these could be subjective. It would be useful to have guidance from the regulators, in consultation with the industry, to create commonly agreed risk identification and quantification.

6.5.3 AI/ML Process

Determining the Level of Human Involvement in AI-Augmented Decision-making

To determine the level of human involvement required in AIDA-augmented decision-making, Singlife follows the risk identification process described above and selects the level of human involvement that gives an optimal trade-off between risk and benefit. This decision may also be made with the consideration of other material factors

The three broad approaches to human involvement are:

- Human-in-the-loop: human oversight is active and involved, with the human retaining full control and the AIDA model providing only recommendations or input.
- Human-out-of-the-loop: no human oversight, and the AIDA solution has full control without the option of human override.
- Human-over-the-loop: human oversight is involved to the extent that the human is in a monitoring or supervisory role, with the ability to take over control when the AIDA model encounters unexpected or undesirable events such as model failure.

Operations Management

Data scientists are required to include the following documentation as part of operations management:

- Data documentation: To include the lineage of data used, the data quality report, bias report and experiment design. For developing AI/ML models, the data scientists use data from the data lake which is managed by Data Engineering/Tech. For this reason, data scientists often lack full visibility on how the data is created in the data lake from the source system; additionally, they do not have control over the data quality and availability.
- Model documentation: The model documentation is to be prepared by the data scientist developing or maintaining the model. The model documentation should contain key performance indicators or measures that enhance the transparency of AIDA, thus helping to establish trust and confidence among stakeholders.

Along with the AIDA framework, we are also working on setting up an MLOps process for productionising the models. As part of this process, we have incorporated a use-case agnostic deployment checklist to identify details regarding model engine availability, deployment structure, auto scaling and stress testing prior to deployment, fallback and failure planning, and a tracking mechanism. Continuous integration/continuous delivery (CI/CD) has not been fully set-up for the entire process, and as part of this integration, we also realised that having a standardised mechanism for tracking model metrics across different modelling techniques was challenging.

Stakeholder Interaction and Communication

In order to build trust among business users when deploying AI solutions, there must be constant communication related to the general disclosure about the usage of AIDA, how decisions are being facilitated by the model, and details about the FEAT Principles' assessment. As part of the framework, different business users were involved from various departments in the model-approval process. We realised that AIDA concepts are difficult to explain to business users, even though business users understand the business metrics which are derived from the AI/ML process. For this reason, it is often difficult for business users to comprehend the entire AIDA solution.

6.5.4 Use case study

Use Case (Singlife Card Transaction Monitoring)

The Singlife Account is an insurance savings plan where the customer can earn interest and have the flexibility to make top-ups and withdrawals, with no lock-in period and fees. The Singlife Account provides life insurance cover in the event of death and terminal illness. Singlife Account customers can also apply for a Singlife Visa debit card with no FX fees and no annual fees.

The purpose of Singlife Card Transaction Monitoring is to alert the compliance team by identifying potentially fraudulent transactions made on Singlife debit cards. Once the alerted transactions are validated, and if fraud is confirmed, the compliance team will take further action.

Model Details

Identifying fraudulent debit card transactions is an unsupervised business problem which in this use case has been considered as an anomaly-detection process where unexpected events in the dataset which are different from the normal events are identified. Unexpected events are rare, and their characteristics differ significantly from normal data. We have used an autoencoder model to detect anomalies in card transactions. An autoencoder is a special type of neural network that copies the input values to the output values as shown in Figure 6.10. The number of neurons in the hidden layer is less than the input layer, and the hidden layer extracts the essential information from the input layer. This condition forces the hidden layer to learn the most significant patterns of the data and ignore the "noises".

Encoding and Decoding

In the encoding process, input values are compressed to get the hidden layer. The process is a mapping from the input space into the lower-dimension latent space, also known as the bottleneck layer. The decoding process reconstructs the input layer using the information in the hidden layer. Usually the layers between the hidden-input layers and hidden-output layers are symmetrical. There is loss in the process, and the model tries to reconstruct the output close to the input. The training objective is to minimise reconstruction error.

Autoencoders are usually used for dimensionality reduction where the number of features in a dataset needs to be reduced while preserving the essence of the data. In this experiment, the algorithm has been used for anomaly detection, as this is a by-product of the dimensionality reduction. When the data is compressed and reconstructed, the main patterns are identified and those outliers that do not follow the normal pattern are revealed.

The decision scores output by the model are plotted in a histogram to identify anomalous transactions. The values are skewed to the right.

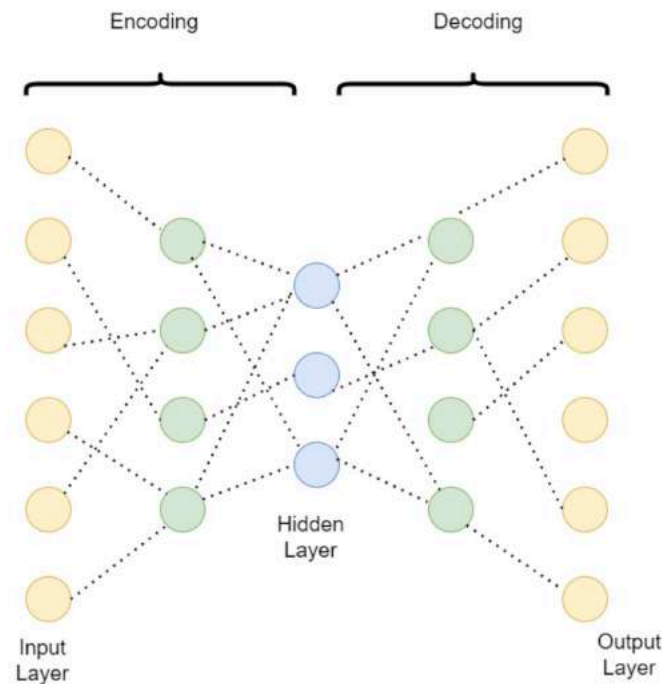


Figure 6.10.: Autoencoder representation.

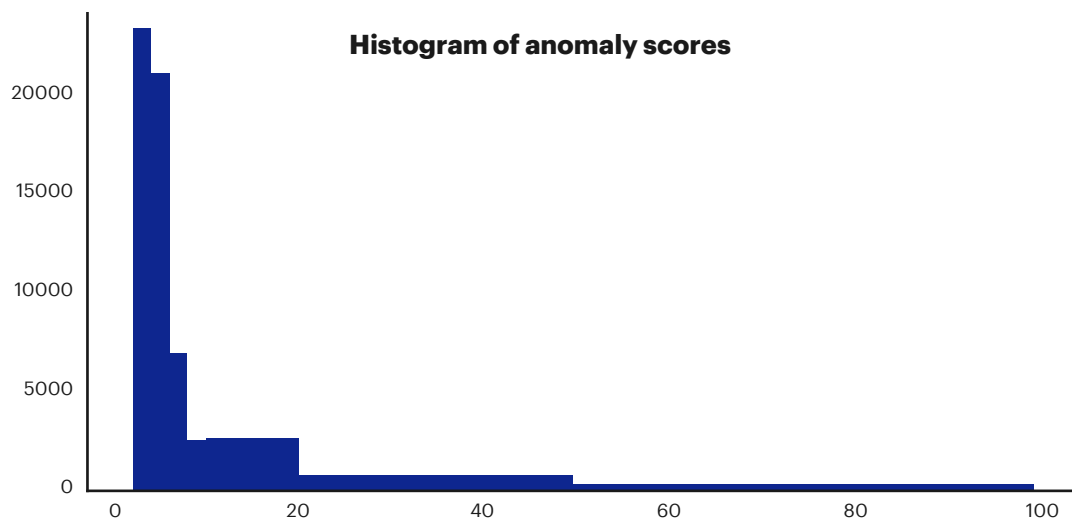


Figure 6.11.: Histogram scores from the autoencoder model to determine threshold for outliers' identification.

Fairness Assessment

We have evaluated the Veritas toolkit as part of our fairness assessment methodology. However due to the current design limitations of the toolkit, we found that we could not use it on our unsupervised model. To mitigate the risk of unfairness, we used Datasheets for Datasets, which is a tool from Microsoft's research team designed to standardise the process for documenting the datasets used for ML model-training and evaluation ([Data sheets for Datasets – Research paper](#)).

We do not use any PII information about the customer in the model. Most of the attributes used are related to the transaction time, merchant location, transaction currency, top-up/withdrawal trends, etc. As part of the fairness assessment for justifiability, the data scientists made sure that none of the individuals or groups of individuals was systematically disadvantaged. One of the findings as part of the fairness assessment is that when we overlayed the anomalous transaction results across age/gender (age and gender are not included as input features during model training) there was a higher proportion of anomalous transactions among younger males. However, in this case this finding is relevant and not unfair because when we look at our overall card-spend base, there is a significantly higher proportion of transactions from younger males.

Transparency Assessment

As AI- and ML-based models are increasingly used in a variety of applications such as recommendation systems, hyper-personalisation, fraud detection, etcetera, it is important to understand the rationale behind the prediction/output to trust the decisions of the AI/ML powered systems. However, most AI/ML systems are like black boxes, especially the deep-learning neural network. Explainable AI (XAI) is the field that aims to understand the steps involved in making the decisions, and focuses on creating tools and methods to describe how ML models make decisions. One such fundamental technique is SHAP (SHapley Additive exPlanations) values. Along with the usual model evaluation metrics like feature-importance plots, we have also utilised SHAP values to explain the “why” part of predictive models. SHAP values can be used to identify how much each has contributed to the prediction compared to the average prediction.



Figure 6.12.: Force plot for a sample prediction.

The features shown above push the SHAP value to be higher than the base value. For this record, the transaction currency is AUD. Less than 1% of the total transactions are done in AUD. Transaction_currency_5 indicates SGD currency, and this feature value is 0 for this transaction. This is also the first transaction done with the card (first yearly, monthly and daily transaction). The features – transaction currency, monthly cumulative sum, amount and daily cumulative sum – contribute to a value higher than the base value.

In this use case, the individual data subjects are indirectly impacted to the extent that their transactions are monitored by an AI, while the ultimate decision-making authority rests with the internal compliance team. Given the low materiality of the impact on individuals resulting from this compliance process, we have not implemented a procedure for customers to inquire about decisions made by the compliance team. However, there is an adhoc process in place to clarify the model to business users. If the need arises, a formal process can be established to effectively handle customer queries, ensuring their concerns are appropriately addressed.

As we move forward and create more AI/ML models with higher materiality and direct impact on customers, it becomes crucial to focus more on transparency and interpretability of the model to provide more explanations to data subjects. Therefore, we plan to establish a SOP that provides avenues for data subjects to question the transparency of decision-making for high materiality models. In case a data subject requests an explanation, we will use global and local interpretability plots such as PDP plots, feature important plots, and shapley value plots to clarify the type of data

being used, how it affects the decision-making process, and how personal attributes may impact the results. By providing clear and transparent information to the data subjects, we can ensure that they have a better understanding of how the model works and why certain decisions are being made. This will enhance the trust of the data subjects in the model and help to prevent any potential negative impact.

Accountability

There is a model-approval process in which the Head of Data Science will approve models before they go into production, along with business-user approvals. There are instances where all the data science models are not developed in-house, and SaaS solutions from external vendors are instead purchased. We have identified that, in such scenarios, we will need to ensure that these external vendors are committed to following the principles adopted by Singlife in relation to AI governance, including the FEAT Principles, in vendor contracts. Currently we do not have a central committee to oversee all AIDA development; however, we will be evaluating whether to create a central committee in future as currently we do not have many models in production. Also, there is no independent validation being done currently other than the model-approving committee where we also get approval from the business units prior to deployment.

Human-Centricity

During the model development process, we constantly involved business users in our use case, this was the compliance team and the product team. By involving business users early in the model development and validation process, we managed to supplement the model output with business rules defined by the business users.

VARIABLE	VARIABLE IMPORTANCE	CUMULATIVE
AMOUNT	9%	9%
amount_compared_to_same_txn_types_past_90D	9%	18%
amount_compared_to_RETAIL_SPEND_past_7D	9%	27%
TOTAL_AMOUNT_TOP_UP_7D	8%	36%
pct_ofTOP_UP_past_180D	8%	43%
amount_compared_to_WITHDRAW_past_7D	7%	51%
AVG_AMOUNT_WITHDRAW_90D	7%	58%
amount_compared_to_TOP_UP_past_7D	6%	64%
TOTAL_AMOUNT_WITHDRAW_7D	6%	70%
TXN_HOUR	6%	77%

Figure 6.13.: The feature-importance plot.

The Journey to Increase RAI Culture & Training

This is the first time that, as an organisation, we are documenting processes and setting up an AI Governance Framework, as the use of AIDA in our business is likely to increase. Based on the extensive work done by the Data Science and Analytics team in incorporating the AIDA framework in our processes, we have learned some lessons especially around the technical implementation of the framework.

While implementing the framework in our processes, we realised there is a steep learning curve for junior data scientists to follow the AIDA framework in the existing modelling process, and currently a lack of capacity for the Lead Data Scientist to read through the ML code thoroughly before approving the model. The Data Science and Analytics team will conduct organisation-wide roadshows to showcase AI capabilities and usage in order to increase awareness, and there will be a mandatory module for AI ethics governance in the annual essential learning pack for Singlife staff.

6.5.5 Concluding Remarks

The AIDA governance process is relatively new to the company and multiple iterations between different departments like risk and legal were needed to define the initial governance framework. We also acknowledge that this is an iterative process, and that we need to refine it further based on the lessons learned after implementing the framework in multiple use cases.

The Veritas toolkit has significant potential for standardising fairness assessment for classification problems. However, we could not use the Veritas toolkit for our fairness assessment as our use case is based on unsupervised learning, and the Veritas toolkit currently supports only binary classification algorithms. It would be useful if, going forward, the regulator could include unsupervised models in the Veritas toolkit.



6.6 Standard Chartered Bank

6.6.1 Overview

The adoption of Artificial Intelligence is changing the way Standard Chartered operates. The bank has developed in-house solutions and onboarded third-party AI solutions with an aim of providing new value propositions and distinctive customer experiences. It is essential to have appropriate governance, controls, risk management framework with related processes, and training around the development, deployment and monitoring of AI solutions so that AI is used with appropriate safeguards and does not adversely impact our stakeholders or the communities we operate in.

The objective of this whitepaper is to demonstrate the implementation of the Veritas methodology across all FEAT Principles, both at an organisational level and on the bank's selected use case of "Signature Verification". The whitepaper provides a summary of Standard Chartered's integration journey including feedback and challenges.

6.6.2 AI/ML Governance

Standard Chartered participated in co-creating the MAS FEAT Principles and has been a lead member of the Veritas consortium since inception. The bank developed its Group Responsible AI (RAI) Standard with FEAT Principles as one of the key inputs. The RAI Standard is applicable globally. The bank also participated as a lead subject matter expert to publish the FEAT Transparency Principles Assessment Methodology. The Veritas methodology provides guidance and recommendations for implementing the FEAT Principles, governance and controls. The fundamentals are already covered in the processes, policies and standards that we adhere to within the bank.

The Veritas methodology prescribes a five-step approach to the AIDA lifecycle, starting with defining organisation-wide principles and standards for the use of AI. In this context, this section describes the bank's RAI Framework.

Values, Principles & Governance (How Does the RAI Framework Align with Core Values and Principles?)

Standard Chartered's Responsible AI Governance Framework ("RAI Framework") is based on the key components in Figure 6.14, and which are described further below.

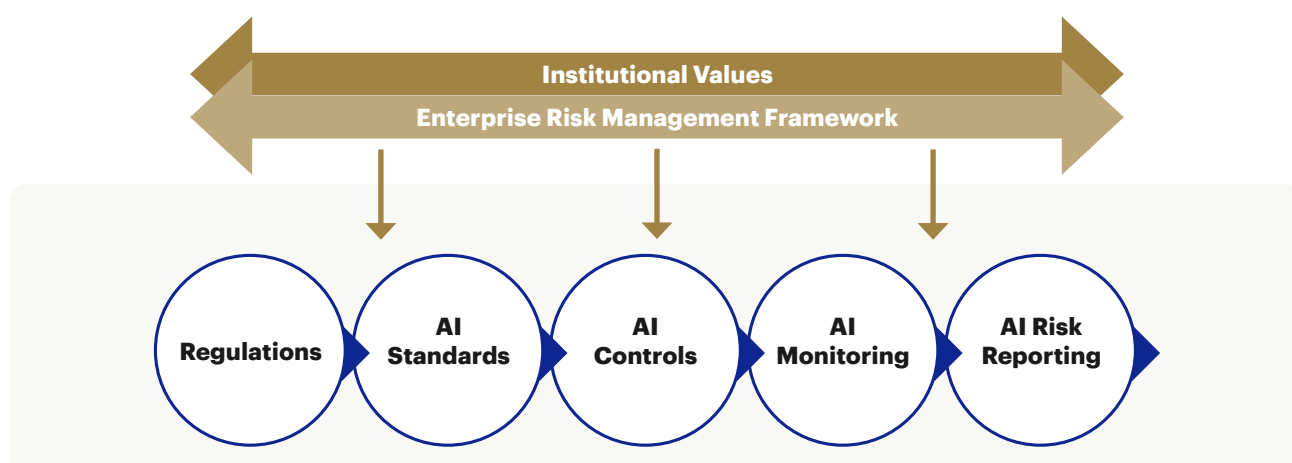


Figure 6.14.: The Responsible AI Governance Framework at Standard Chartered.

AI Regulations: Standard Chartered's Group RAI Standard ("the Standard") is guided by the bank's current regulatory environment and set of valued behaviours. The Standard is applicable globally across all geographies, and holistically incorporates the Monetary Authority of Singapore's (MAS) FEAT Principles as well as the Hong Kong Monetary Authority (HKMA) requirements holistically. The Standard is continuously assessed against new regulatory publications or for updates to existing requirements. The requirements for Responsible AI are embedded in the bank's ERMF (Enterprise Risk Management Framework) and form a key component of the bank's Data Management Policy within the Compliance Risk-Type Framework.

RAI Standard: The Standard establishes principle-based requirements for the consistent and safe use of AI in the bank. The Standard is underpinned by 10 core principles including data suitability, fairness, accountability, explainability, transparency, stability, assessment/validation, ongoing monitoring and auditability that categorise the control requirements across the AI lifecycle. These are mapped at a granular level to the MAS's FEAT Principles and the HKMA's BDAI Guidelines. The Standard also identifies interconnected policies and standards that are associated with AI.

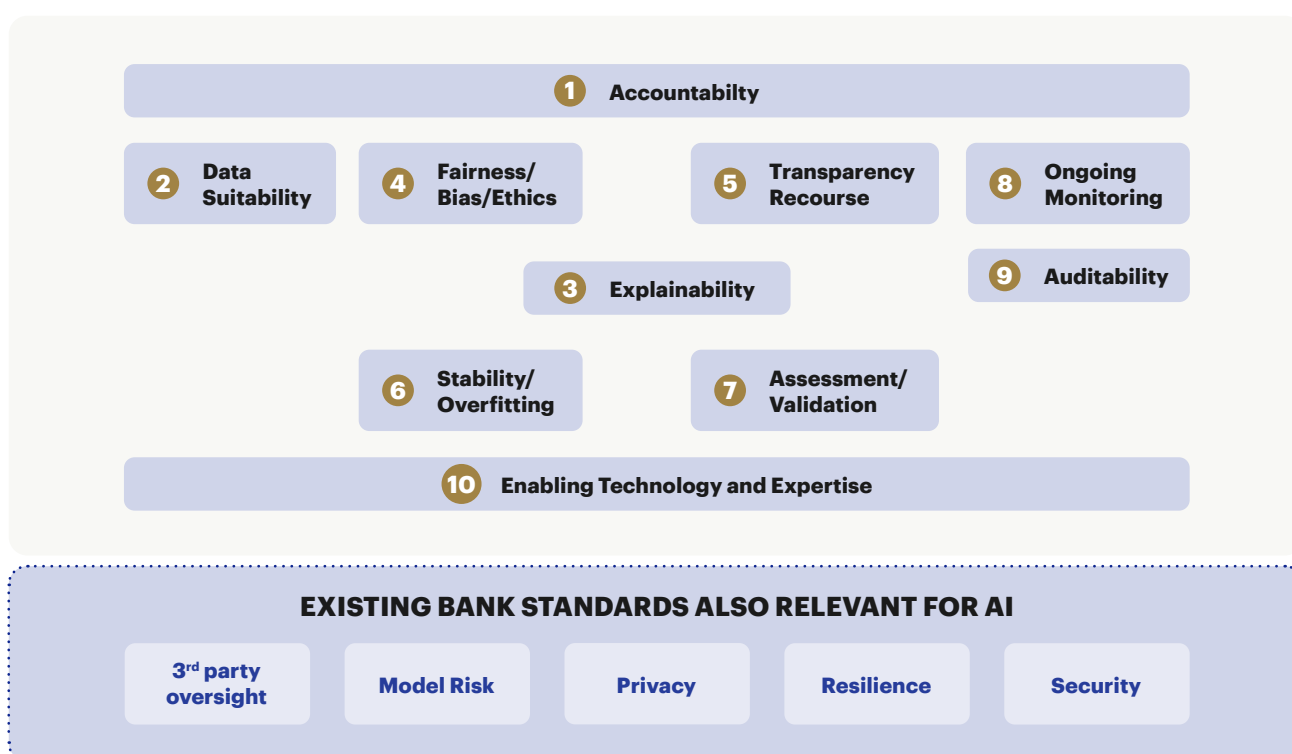


Figure 6.15: Principles for the Responsible Use of AI (RAI Principles).

AI Controls: The bank has defined an AI risk and controls library (RCL) describing the specific controls to meet the control requirements stated in the Standard to manage risk arising from the use of AI. The control requirements are aligned to the materiality of the AI use case. The identification and management of controls is operationalised through (a) the RAI governance framework and (b) respective business risk committees that govern the identified AI risks as part of the bank's ERMF.

AI Monitoring: This involves monitoring risks across the entire AI lifecycle and describes how risks introduced on account of AI usage are identified, recorded, monitored and remediated within the bank. In addition, a periodic review of the risk and control environment for the use of AI is conducted by business and functions. Any areas with elevated risks are required to create treatment plans to address and mitigate risks to an acceptable level.

Materiality Assessment and Risk Tiering

Standard Chartered follows a materiality-based approach to RAI governance. The Standard provides requirements for establishing the materiality of an AI use case as low, medium or high.

There are nine impact dimensions which closely align to the risk parameters defined under the Veritas methodology. These are designated M1 to M9 and described below:

- The first three dimensions (M1–M3) map directly to the bank’s Group Risk Assessment Matrix (GRAM) assessing incremental regulatory, financial and reputational impact to the underlying processes on account of AI.
- The remaining dimensions (M4–M9) assess specific aspects of the AI system/use – case impact and are used to drive the applicability of appropriate controls. These, include aspects like impact on customers, impact on staff and the use of personal data by the AI system/use case and level of human involvement in processing in the AI outcome.

The dimension of model complexity was originally included as an impact area but ultimately dropped due to the subjectivity inherent in defining “complexity” in an evolving area.

The overall flow of Standard Chartered’s RAI Governance is outlined in Figure 6.16., and further described below.

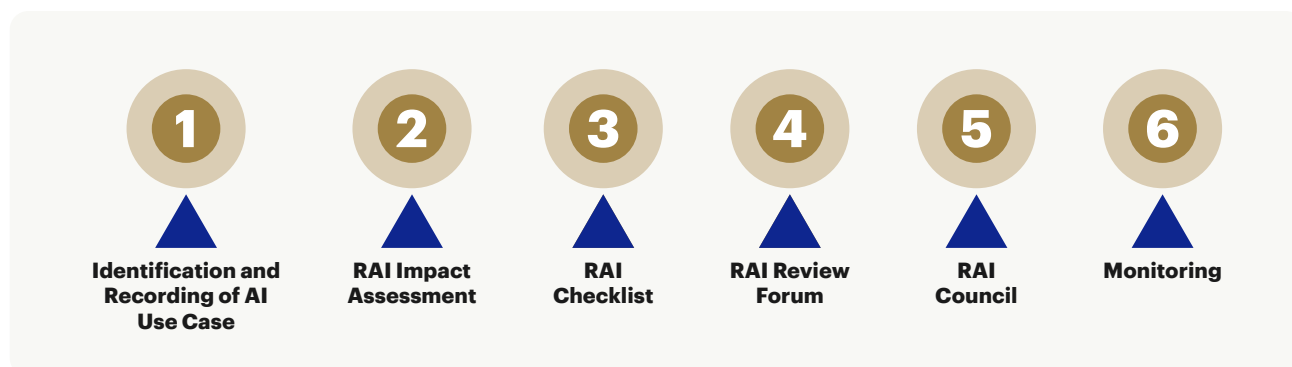


Figure 6.16: Responsible AI Governance process flow.

- 1. Identification and recording of AI use case:** All AI solutions must be recorded in the bank’s global RAI Inventory. AI solutions can be built internally, purchased from a third-party vendor or sourced through partnerships. An “AI Ready Reckoner” that asks a series of questions to establish whether a development or procurement initiative falls within the scope of the Standard forms the basis for AI identification, implemented across both – the build and buy channels. The bank has implemented controls at onboarding to ensure AI solutions purchased from third-party vendors follow the RAI governance process. Differentiating between AI and data analytics models is a challenge. The bank focuses on predictive applications of AI or those used for decision-making rather than historical reporting from a governance perspective.

- 2. The RAI Impact Assessment:** The RAI Impact Assessment determines the materiality of the use case as low, medium or high depending on the nine impact dimensions (M1-M9) mentioned earlier.
- 3. The RAI Controls Assessment:** The RAI Controls Assessment is a series of questions that must be answered by the owner of the AI use case, and ensures it meets the requirements outlined in the Group RAI Standard. The materiality of the use case determines the exact set of assessment questions that need to be completed.
- 4. The RAI Review Forum:** The RAI Review Forum is a group of subject matter experts from the bank's first- and second-line teams. The forum meets weekly to assess each AI use case against the Standard, reviews and challenges the materiality, and makes risk-based recommendations to the RAI Council.
- 5. The RAI Council:** The RAI Council is comprised of senior stakeholders across business and functions who bring a diverse set of expertise, perspectives and experiences. The council meets monthly to review, challenge, approve and, if required, reject material AI use cases.
- 6. Monitoring:** AI solutions are monitored after go-live through regular review meetings by the affected process owners and project teams. Corrective actions are taken for any breaches, and the relevant Non-Financial Risk Committee (NFRCs) are informed. There are also controls within the third-party onboarding processes that are checked for effectiveness.

Standard Chartered follows a risk-based approach where low-materiality use cases follow a more simplified governance process than medium- or high-materiality AI use cases.

Organisation Policy and Control Measures, and How They Are Commensurate with Identified Risk Levels

While risks unique to AI are covered within the Standard, incremental risks related to other areas are covered in respective risk areas and policies. These are listed below:

- 1. Information and Cyber-security risks:** The Standard identifies two broad areas related to information and cyber-security risks – data poisoning and model extraction. A security assessment has been implemented to review AI use cases for both risks and is a mandatory requirement before any AI project goes live.
- 2. Privacy risks:** A Privacy Impact Assessment (PIA) is performed for all AI solutions that involve the use of personal data. PIAs are an important privacy risk identification tool used to assess risks and identify risk-mitigating measures.
- 3.** The bank's **Group Model Risk Policy** sets out the requirements and responsibilities for the identification, measurement, mitigation and monitoring of model risk related to the AI use case. AI use cases under the scope of Model Risk Management go through an additional, independent model-validation exercise.

4. The **Group Third-Party Risk Management Policy and Standard** govern all third-party arrangements in the bank. Any AI initiative involving a third party must be declared to ensure appropriate governance in line with our RAI Standard.
5. The **Group IT Resilience Management Standard** prescribes the principles to ensure the bank is prepared for, and can respond appropriately to, operational disruptions to critical IT services.

The risk rating associated with an AI use case is driven by the initial materiality assessment, and then assessed against the effectiveness of the applicable controls specific to RAI as well as to those from the above interconnected policies and standards, not all of which may be unique to AI.

The pre-go-live process flow for Responsible AI is shown in Figure 6.17.

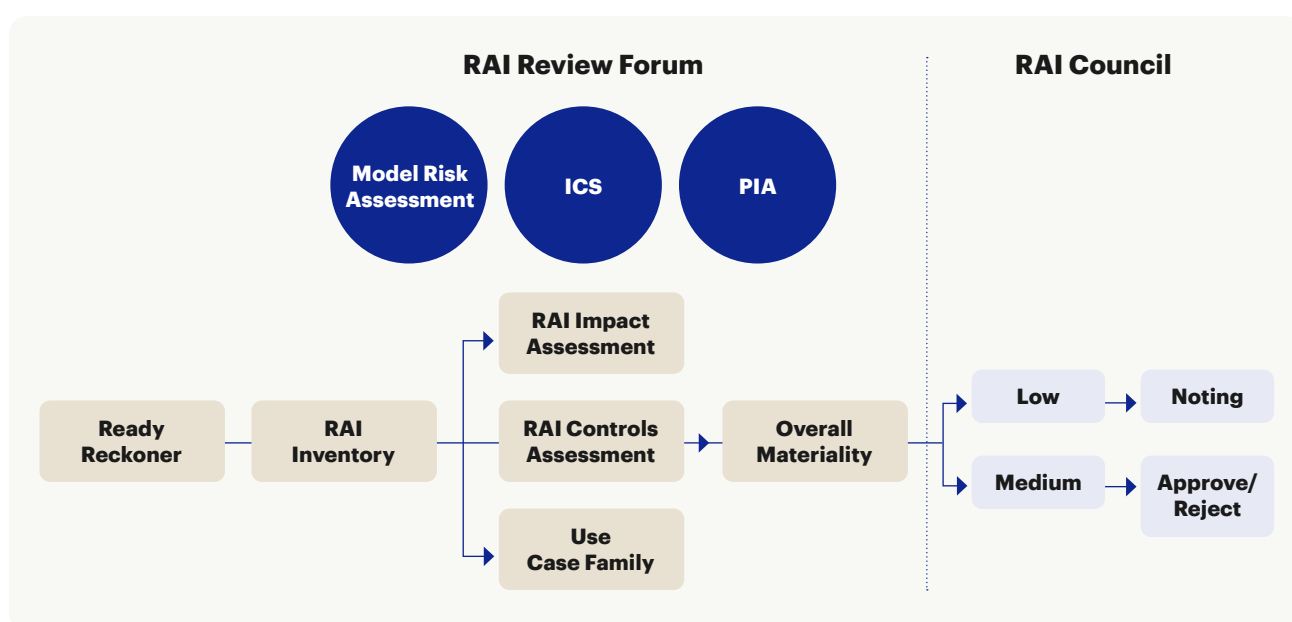


Figure 6.17.: Pre-go-live process flow for RAI

Roles and Responsibilities of Stakeholders

The following roles are defined within our RAI Standard in terms of authority to execute first-line, second-line and third-line responsibilities:

1. Group head of businesses and functions are responsible for implementing the Standard in their respective areas.
2. Process owners (PO) are business/function managers in the first line of defence who are the ultimate beneficiaries of AI solutions and the accountable party internally.

3. Chief Information Officers and Heads of Analytics/Data Scientists support the PO in first-line responsibilities and are responsible for adherence to the Standard.
4. The Group Chief Data Office is responsible for operationalising the Standard globally.
5. Subject matter experts in group compliance (data conduct) set policies/standards, provide guidance and oversee/challenge the first-line teams on managing AI-related risks.
 - Group Internal Audit within the third line provides independent assessments on the adequacy and effectiveness of the RAI risk management, controls and governance process.

Measures of progress in rolling out AIDA governance

The bank is progressing in its journey of establishing a robust AIDA governance framework. At the same time, independent reviews of AIDA governance are conducted to ensure the right level of oversight:

- The Standard owners within second-line teams conduct an effectiveness review of the RAI practice covering governance, controls, risk management, audit findings, etc. This review has been conducted on an annual basis since the rollout of the Standard.
- Second-line assurance teams will be conducting a review of the RAI practice in line with the bank's standard assurance methodology.
- In addition, an external consultancy firm with subject matter expertise in Responsible AI conducted an independent validation/review of the RAI Governance Framework. The bank is reviewing the proposed recommendations for implementation.

The bank is looking to address gaps identified through the Veritas integration exercise and will be detailing them in subsequent sections.

6.6.3 AI/ML Process

Standard Chartered has made significant progress in integrating RAI Processes and Technology within the bank's RAI Framework.

A centralised model for Responsible AI governance has been established at the bank. This enables access to a broader and deeper pool of skill-sets and a consistent approach to the review and approval process. The bank has identified existing internal roles that can be mapped to responsibilities defined within the RAI Standard.

A centralised AI workflow is under construction to digitise the existing manual governance workflow as part of our continuous improvement efforts.

Ongoing monitoring of AI systems is conducted post-implementation to ensure performance is in line with expected service levels. However, the nature of ongoing monitoring of AI systems is different to traditional systems, and additional controls and enhancements are planned.

The bank's current process to identify the use and change of AI internally has been through self-identification and certification. An "AI to detect AI" technique has been developed and piloted to identify AI usage across the bank's environment and proactively identify any gaps in the governance process. This has been developed to address the reliance on and risk from self-identification and to prepare for scaling up the use of AI across the bank.

6.6.4 Use Case Study

Description of Identified Use Case

Background

Standard Chartered selected "Signature Verification" for the Veritas framework use-case integration. This use case was selected due to its use of unstructured data in the form of images. Unstructured data is inherently difficult to analyse vis-à-vis structured or "tabular" data as it involves text and images, and cannot be evaluated using traditional tools or methodologies. Therefore, this use case is a good representation of potential lessons learned, applications and outcomes of the Veritas methodology, specifically for unstructured data.

Overview and Problem Statement

Signature Verification requires the extraction and validation of signatures from cheques as part of the cheque-clearing activity via a maker-checker-based workflow. This involves comparing the incoming signature of a customer with a specimen present in the bank's systems. This is usually performed by a "maker" role and reviewed by a "checker" role.

- The comparison is performed visually (i.e., manually) and consequently has the associated overheads of any manual process.
- Manual judgement is heuristic in nature and often lacks consistency in decision-making.
- The manual process prevents straight-through-processing of transactions and impacts the turnaround time for customers.

Solution

The solution is to build an AI model to perform signature comparison through the following tasks:

- Recognise and extract customer signatures from cheques.
- Verify the cheque signature against specimen signature samples previously collected from customers:
 - Accurately characterise each user's signature, thus offering good verification and recognition performance.
 - Compare the incoming signature (on the cheque) against the customer's specimen signature and predict a similarity score. The solution aims to automate the maker's activities, and each transaction signature will be routed to the checker for verification. In this context, there are no final business decisions or approvals made by the AI as it is overseen by a human-in-the-loop.

The benefits of automation achieved through Signature Verification are:

- Reduction in manual work by the maker.
- Improved customer experience.
- Inconsistency in decision-making involved in manual process is addressed.

Input

There are two major inputs to the Signature Verification use case:

- The customer's signature as captured on the cheque. This is sourced from the bank's Cheque Processing System.
- The customer's specimen signature used for comparison. This is extracted from the bank's Signature Repository System.

Process Flow

- Identify the customer account number from the cheque and extract the signatures from the images.
- For each extracted signature, the model will retrieve the specimen customer signature from the Signature Repository System using the customer account number as an identifier.

- The model compares the extracted signature against the specimen customer signature and provides a similarity score.
- The signature is determined as pass/fail based on the score (the account number does not influence the model and is used only for sourcing the specimen signature).

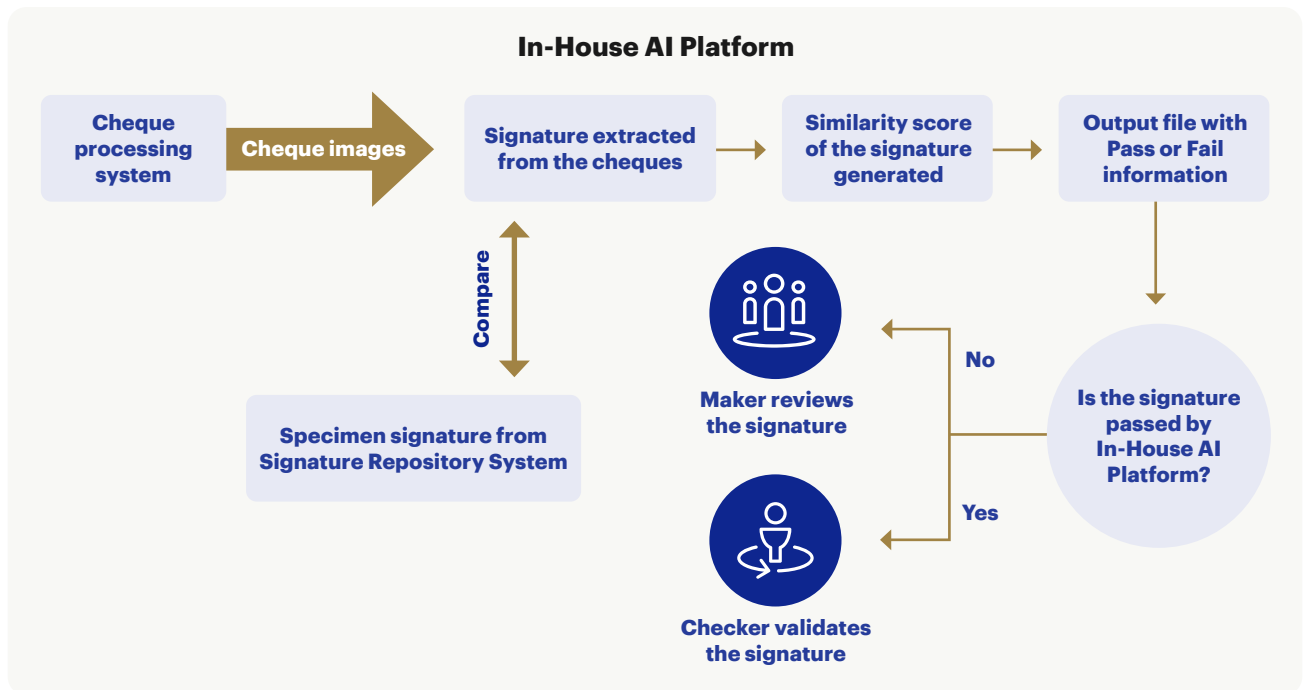


Figure 6.18.: SCB in-house AI Platform

- The verified signatures are designated as pass or fail by the model and an output file is generated.
- This is routed to the appropriate workflow in the Cheque-Processing System where all passed signatures will be validated further by a checker.
- Failed signatures will be routed for manual verification and follow-up to the maker queue.

Signature Verification has been tuned to an accuracy of approximately 92 percent and will be evaluated on stability.

Platform Used for AI-Solution Development

- The development and deployment of the solution has been done on one of the bank's internal AI platforms used for solving AI-based problems, including those involving structured data, unstructured images and text-based problems.

- The AI platform performs tasks such as data-, image- and text-extraction, document-checking, document- and email-clustering, classification and understanding of requests, computations and comparison of images such as signatures, faces and conversational chatbots, etc. The AI platform is internally referred to as SWOOSH.

Signature Verification is in the pilot phase and is expected to go live after establishing operational readiness.

How to Apply the Above Use-Case Agnostic Items on Identified Use Case?

The Foundational Step 0 of the Veritas methodology has been addressed historically within the bank through our existing Standards and Control framework.

Signature Verification has been reviewed in line with the RAI Governance lifecycle across the key stages of updating the global RAI Inventory, completing the RAI Impact Assessment, RAI Controls Assessment and obtaining approval from the RAI Review Forum.

The bank has evaluated the adherence of Signature Verification to the Responsible AI principles using an in-house tool. The tool contains specific engines for analysing data suitability, explainability, robustness, stability, fairness and bias, ongoing monitoring and model validation for AI use cases utilising unstructured data. The final report from the in-house tool incorporates elements of transparency, ethics and accountability, and auditability, and will also comprise risk-based recommendations. The in-house tool is internally referred to as FRIDAY.

The in-house tool was developed because traditional methods are more suitable to structured data and pose the below challenges:

- For semi-structured/unstructured data like images, text or graphs there are no comprehensive tools available in the market suitable for exploratory data analysis and for evaluating models against Responsible AI guidelines (manual validation lacks consistency).
- No consistent methods to check for potential biases in data.
- No suitable versioning and auditability for unstructured data. Database- or code-management software may not be appropriate due to size constraints.

Please refer to **Table 2** in the Appendix for the bank's assessment of the Veritas Generic Checklist.

Fairness and Bias Considerations

Signature Verification does not use any protected variables such as gender, race, religion, sexual orientation or nationality, and related proxies are also not utilised. Additionally, the model has been evaluated thoroughly to identify possible biases.

Signature Verification only uses signatures/account numbers and no other demographic/personal information inputs (health, age, etc.) Therefore, there is no mechanism to assess for bias, and assessing fairness at a group level is not relevant to this specific use case.

Signature Verification was evaluated on different types of fairness and bias drivers:

- **Language bias:** Signatures are primarily written in English and a small percentage are written in other languages. There is a possibility of bias due to variations in language. However, this cannot be assessed as the bank does not have the relevant Ground Truth. No language bias has been observed on the samples and there is additional reassurance through final manual validation. The model will be continuously monitored and revalidated to identify any future deviations.
- **Age-/health-related bias:** Minor variations between signatures and specimen signatures may arise due to age- or health-related issues (e.g., Parkinson's disease) which can occur in addition to other natural signature variations. The pipeline has inbuilt steps to compare against past failure predictions to avoid rejections and prevent unfair age-/health-related biases. Whenever the model continuously fails signatures from a user, but the checker passes the same signatures, the threshold for pass/fail will be realigned based on past predictions. Alternatively, if the model passes signatures that have been failed by the checker, and model performance falls below the threshold of the false acceptance rate, model retraining will be triggered since in this case the model is incorrectly passing signatures.
- **Historical bias:** Traditionally, signature verification is performed against the specimen signature provided by the customer at the time of on-boarding. This implies historical data around pass or fail is highly dependent on the specimen signature itself, which may be stale.

Quantification of the above fairness-related biases is a challenge as the use case involves unstructured data and is further compounded by the unavailability of Ground Truth related to the conditions.

Several other data-related biases have been assessed:

- **Measurement bias:** During the initial stage of data preparation, signatures are manually labelled as pass or fail. Passing/failing a signature is subjective in nature and therefore resulted in measurement bias that was identified with the help of the Gage R&R (Repeatability and Reproducibility) analysis. Mitigating actions to address this bias have been included in the Appendix under **Table 3 – Fairness Checklist**. Subjectivity in measurement is one of the challenges when working with unstructured data.
- **Pre-processing bias** was also determined to be applicable during noise removal (noise of logo, printed text, seals over and above the signature) and colour-scale conversion, thereby introducing error into the dataset. This has been addressed at each step by assessing the impact of loss in the images.

There is no impact on the client due to the above data-related biases as there are rectification measures in place and additionally there is a human-in-the-loop element where all transactions are routed through a checker for manual validation. If a scenario presents where a significant volume of signatures is failed by the model due to the above biases, they will be routed to the maker queue. This step ensures the overall resilience of the process.

There are no legal, regulatory or societal considerations based on country-specific requirements in this context, and there have been no related points highlighted by the bank's country compliance teams. Fairness and related challenges identified during the initial data collection/analysis phase have been socialised with the respective AI system owner.

Please refer to **Table 3** in the Appendix for the bank's assessment of the Veritas Fairness Checklist.

Ethics and Accountability Considerations

From an accountability perspective, the process owner has been educated on the model, scope, limitations, possible risks and constraints to ensure full ownership of associated responsibilities.

Ethics-related considerations are addressed first at an organisational level by taking into account the bank's valued behaviours in drawing up the Standard and defining its scope as global; and at the use-case level by factoring in the human-in-the-loop mechanisms for reviewing model output – with the final decision being taken by a human and complemented by existing channels for recourse.

Please refer to **Table 4** in the Appendix for the bank's assessment of the Veritas Ethics and Accountability Checklist.

Transparency Outcomes

Standard Chartered evaluated multiple tools/packages used for supporting AI explanations in Signature Verification and ultimately selected the bank's in-house tool due to its overall suitability for handling unstructured data, as explained in the previous section.

Global explanation methods such as Partial Dependence Plots (PDPs) and Accumulated Local Effects plots (ALE plots) do not seem to be appropriate for handling unstructured data and hence were not used for Signature Verification. A PDP is more suited for linear regression models, whereas Signature Verification utilises deep-neural networks (DNN). PDP can only work with two features at a time whereas the Signature Verification algorithm has features within hidden layers in the DNN. The bank will continue to evaluate other explanation tools available in the market for appropriateness.

Please refer to **Table 5** in the Appendix for the bank's assessment of the Transparency Checklist.

Specific Achievements and Lessons Learned Through the Journey of AIDA System/AI Model Lifecycle

The recommendations given in the methodology align with current practices at a fundamental level. However, there are lessons and opportunities for enhancement:

- Standard Chartered has a defined and structured AI Governance lifecycle in place.
- There is value in providing customised training with simplified explanations to the AI system owner during the onboarding process to provide guidance on the AI use-case and principles.
- The bank recommends the Veritas methodology is enhanced to include further guidance in the use of unstructured data for the reasons given below:
 - Most of the tools and techniques for assessing algorithmic fairness and system explainability that are currently available in the market are designed for structured data. In addition, they are not suitable for deep-learning architecture. For this reason, the bank uses an in-house tool to verify model fairness in these scenarios.
 - The methodology recommends the use of differential privacy, de-identification, data anonymisation and data masking to protect personal data. These are challenging to apply for unstructured data as masked data cannot be used for training the model.

- The methodology does not provide guidance on the safeguards in the use of AI for generating synthetic personal data. Such guidance could help in the usage of GANs (Generative Adversarial Networks) to generate synthetic data sets.
- Guidance on commonly used fairness metrics provided in Veritas applies mostly to structured data and does not cover scenarios where unstructured data such as images are involved.
- Further guidance is required to assess fairness at the individual level in use cases where there is no access to demographic data.

- The Veritas methodology can be applied when we have access to the AI system, training data and test data. The methodology cannot be applied to third-party systems when access is not available due to third-party concerns around intellectual property. While this is not a limitation of the Veritas methodology itself, the constraint in applying the methodology needs to be factored in when using third-party AI systems.

The Journey to Increase RAI Culture & Training

The skillsets required for different individuals within the RAI ecosystem will vary depending on their role. For example, a data scientist requires a different set of skills compared to an individual performing an ethical review of an AI use case or someone using the output of an AI system to take decisions. However, there is a common knowledgebase that is required regardless of skillset to ensure the bank can safely innovate in the use of AI.

1. The bank has created training sessions on RAI to increase the awareness of individuals and these are available in the form of videos and slides, as well as periodically scheduled live sessions.
2. Walk-throughs, deep dives and overview sessions have been held, with these tailored to focus groups including boards, risk committees, data science teams, operational teams and compliance teams.
3. Mandatory e-learning is being developed to certify individuals who are stakeholders of AI projects.
4. With growing awareness, there is increasing acceptance that the governance associated with the Responsible use of AI is not an overhead but necessary to enable safe innovation.

6.6.5 Concluding Remarks

The bank's RAI Standards and Principles are based on MAS's FEAT Principles. The Veritas methodology provides strong guidance and recommendations for implementing these principles, and at an overall level the bank has implemented a substantial portion of the methodology across in-house solutions as well as in onboarded third-party solutions. The methodology has enabled the bank to validate the current approach to RAI Governance and identify opportunities to enhance the existing practice, which have been detailed in the above sections.

The bank selected Signature Verification for the use-case integration specifically due to the use of unstructured data. The generation and use of unstructured data is growing far faster than structured data, and there is an opportunity for financial institutions to improve client services, derive insights for business decisions, and solve business problems using AI models that use unstructured data.

The bank's observation is that the Veritas methodology does not yet fully address implementation on AI models with unstructured data, as described in use case study. The bank's recommendation is that the industry continues to collaborate as part of Veritas and provides additional guidance specifically to governing the use of deep learning and unstructured data, especially in the context of fairness and transparency.

6.6.6 Standard Chartered – Appendix

Veritas Methodology Checklists

Table 2 – Generic Checklist

Generic Checklist	Summary of Outcome
(G1) Is AIDA defined in the FSI?	These steps have been performed historically and are covered in Section 6.6.2.
(G2) Is there a framework in place to define roles and responsibilities for AIDA projects/use cases?	
(G3) Is there a materiality framework in place? If so, do you have a clear definition to determine the different level of materiality?	
(G4) Does the FSI have an up-to-date inventory that tracks all material AIDA use cases in use?	
(G5) Are the business objectives of the system defined? Are there quantitative performance estimates against those business objectives?	The primary objective of the model is to perform signature comparison to achieve operational efficiency. The expected effort saves have been quantified and can be assessed over a period of time. The below quantitative performance estimates have been defined for the use case: • False Acceptance Rate – Number of signatures that the model has passed incorrectly. • False Rejection Rate – Number of signatures that the model has failed incorrectly. A Lower Control Limit (LCL) and Upper Control Limit (UCL) have been defined for the above metrics.

Generic Checklist	Summary of Outcome
(G6) Are there scope boundaries of the proposed system? For example, specific geographies/segments?	The scope of solution is limited to a single country and will be expanded/scaled in the future.
(G7) Is AIDA use justified for this particular objective/use case (based on limitations of alternatives, proof of AIDA being used successfully elsewhere, availability of adequate data to train the AIDA system, valid legal basis, etc.)?	The use of AIDA is justified for this use case as the solution is required to handle unstructured data and possess image-recognition capabilities. Other non-AI automation options are not capable of solving this problem. The AIDA system can also adapt through progressive learning algorithms which are essential for Signature Verification and not feasible with a non-AI system.
(G8) Has the materiality of the use case been determined using the framework defined in Step 0?	The materiality of the use case has been assessed as per the RAI Impact Assessment defined in Section 6.6.2. Inherent materiality was assessed to be “medium” due to the use of personal data and potential impact on customers. The net impact was reduced to “low” materiality due to 100% human oversight, and the presence of a proactive and periodic use-case monitoring via metrics such as False Acceptance Rate and False Rejection Rate. These metrics give early warning alerts on performance variations.
(G9) Have you documented the data used by the system including types of data available, the provenance (source), attributes, dictionary (meaning), vintage (time period), data (quality) profile, representativeness and informed consent? For any personal data used by the system, has a DPIA been conducted?	Artefacts/technical documents have been completed in line with existing data quality and related technology policies. A Data Protection Impact Assessment has also been undertaken as personal data is utilised.
(G10) Have you documented the data pre-processing and engineering done by the system?	Data pre-processing steps such as resizing, noise removal, skew correction, etc. are performed on the images. As part of feature engineering, the in-house AI tool extracts relevant dimensional vectors for the cheque signature and specimen signature. All data pre-processing steps are documented appropriately.
(G11) Is the composition of the AIDA system defined? Is it clear how each component of the system AIDA models, as well as business rules and human judgement, if relevant, are used to achieve its business objectives?	The AIDA system consists of: 1) The model component predicts a similarity score for the customer transaction signature. Based on the similarity score, the signature is classified as pass/fail by the algorithm. 2) Human-in-the-loop component where the signatures are validated by a checker before making the final decision on classifying the transaction as pass/fail. All passed signatures proceed to the checker queue for final pass/fail confirmation. All signatures failed by the algorithm get routed to the maker queue for the normal manual verification process. The business function is instructed on the model working and composition. The checkers are trained on the AIDA system and overall process that needs to be followed to achieve the business objectives. This is illustrated in Section 6.6.4.

Generic Checklist	Summary of Outcome
(G12) Is the system's monitoring and review regime designed to detect abnormal operation (e.g., significant model and/or data drift)?	Periodic monitoring is in place to monitor performance and KPIs such as False Acceptance Rate and False Rejection Rate. These are evaluated regularly and assessed along with other elements such as data drift where out of time data is compared with training and testing data to identify differences between data sets. To identify data drift, patterns in the properties of data are checked to identify any outliers or extreme values. Data samples falling consistently outside the control limit range indicate data drift and warrant deeper analysis. For structured data, there are a few open-source packages and statistical methods for analysing data drift such as the Covariate shift, the Kolmogorov-Smirnov test, etc. However, for unstructured data it is difficult to use common divergence metrics on raw data. The image metadata baselined at the time of training is compared with data on an ongoing basis to monitor drift.
(G13) Are there fallback and/or mitigation plans in place in case of triggers from the system's monitoring and review regime?	The model will be retrained in the event of a decline in model performance. The model is regularly monitored, and any performance-related issues are identified with the help of metric thresholds. Adequate safeguards are also built-in in the form of human-in-the-loop.

Table 3 – Fairness Checklist

Fairness Checklist	Summary of Outcome
(F0) Has the FSI defined a standard, robust process for: (a) Identifying at-risk groups? (b) Identifying personal attributes and potential proxies? (c) Identifying explicit fairness objectives and associated measures and thresholds? (d) And algorithmic methods to identify and reduce unfairness?	These steps have been performed historically and are covered in Section 6.6.2. Explicit fairness measures are not specified at a standards/policy level but at a use-case level where feasible.
(F1) Have you identified and documented who are the individuals and groups that are considered to be at risk of being systematically disadvantaged by the system?	Language, age and health as possible grounds for systematic disadvantage have been considered. However, this cannot be assessed, as the relevant Ground Truth cannot be obtained. The model will be periodically monitored and revalidated to identify possible future deviations with respect to language, age or health with the help of False Acceptance Rate and False Rejection Rate metrics.

Fairness Checklist	Summary of Outcome
(F2) Have you identified and documented the potential harms and benefits created by the system's operation that are relevant to the risk of systematically disadvantaging the individuals and groups in F1?	There are four possible scenarios: 1) True positives where the algorithm correctly classifies the signatures as pass. 2) True negatives where the algorithm correctly classifies the signatures as fail. These are scenarios which do not negatively impact clients and further help to reduce fraudulent transactions. 3) False positives which may lead to fraudulent signatures getting passed by the algorithm. 4) False negatives which may lead to correct signatures being failed by the algorithm. To address the above scenarios, the robust human-in-the-loop element mitigates the possibility of any reputational or financial impact to the client.
(F3) Have you identified and documented the fairness objectives of the system and associated fairness metrics with respect to the individuals and groups in F1 and the harms and benefits in F2?	There is no mechanism to assess for biases because the relevant indicators of bias such as personal health, age-related motor degradation, etcetera are not available. Adequate safeguards are also available in the form of human-in-the-loop.
(F4) Have you documented key errors, biases or properties present in the data used by the system that may impact the system's fairness?	The data has been evaluated for multiple biases including language bias, measurement bias and pre-processing bias. (Details given in above section.)
(F5) Have you documented how these impacts are being mitigated?	Measurement bias: Stratified sampling has been done to ensure appropriate representation of signatures so that matching and non-matching samples are taken. This will help to train the algorithm to identify signatures to pass/fail. The model also has an extended parallel run where model prediction is compared against checker prediction. This will help to authenticate reproducibility and repeatability of the model and avoid bias. Pre-processing bias: Addressed at each step of pre-processing by assessing the impact of loss in the images. Language-, age-, health-related bias: Model is periodically monitored with the help of metrics such as False Acceptance Rate and False Rejection Rate. Data Drift: Any data drift-related biases will be managed as part of the model monitoring and improvement cycles where out-of-time data will be compared with training and testing data.

Fairness Checklist	Summary of Outcome
(F6) Have you assessed and documented the quantitative estimates of the system's performance against its fairness objectives and the uncertainties in those estimates assessed over the individuals and groups in F1 and the potential harms and benefits in F2?	Assessing fairness at a group level is not feasible for this specific use-case after taking into account the potential harms as the Ground Truth related to the drivers of such harm (personal health, age-related motor degradation) are not available. Additionally, quantification of fairness-related biases is a challenge for unstructured data.
(F7) Have you assessed and documented the achievable trade-offs between the system's fairness objectives and its other objectives?	Addressing pre-processing bias can add artefacts that impact the model. This is managed through increased human oversight.
(F8) Have you justified and documented why the fairness outcomes observed in the system are preferable to these alternative trade-offs?	100% human oversight was selected to prioritise operational robustness and mitigate any bias. This was not considered a trade-off but rather an overall risk management/safety decision.
(F9) Have you determined and documented whether personal attributes are used as part of the operation or fairness assessment of the system?	The customer signature and account number are the only personal attributes that are present in the final cropped image used as input for model training.
(F10) Does the process of identifying personal attributes take into account the ethical objectives of the system, and the people identified as being at risk of disadvantage?	This was analysed as part of the fairness assessment stage of the model. Only the minimum information absolutely necessary for the model to function is included.
(F11) Have you assessed and documented every personal attribute and potential proxy for a personal attribute and why its inclusion is justified given the system objectives, the data and the quantified performance and fairness measures?	The only personal attributes used within the algorithm are customer signatures and account numbers. The inclusion is justified as these are the minimum attributes for the model to work.
(F12) Does the system's monitoring and review regime ensure that the system's impacts are aligned with its business and fairness objectives?	Periodic monitoring is in place for performance and KPIs such as accuracy and precision. These are tested along with other elements such as model stability, data drift, data suitability, etc. If the model continuously performs below the agreed metric thresholds, appropriate action will be taken. Notwithstanding the monitoring, models will be revalidated depending on their materiality.

Table 4 – Ethics and Accountability Checklist

Ethics and Accountability Checklist	Summary of Outcome
(EA1) Are organisational values defined and described?	The bank's organisational values have been in place for many years.
(EA2) Are organisational principles for the ethical use of AI defined and described?	The Group RAI Standard defines the Principles for Responsible Use of AI.
(EA3) Is the portfolio of core concepts identified and described?	The bank has defined internal values that drive our behaviour. These are supported by the bank's Code of Conduct as well as policies and standards that implement various regulations, while aligning with our values.
(EA4) Is the workforce trained on values-based decision-making or is there a pilot planned?	The bank's Group Code of Conduct aligns with values and behaviours, and annual training/recertification is conducted across the bank.
(EA5) Is there a separate statement of relevant values, core concepts, principles, commitments and specifications, along with a written description for each use case?	Values, standards, principles and fairness assessments are defined at a bank level. Every use case is assessed against these and documented as part of a checklist that covers all the principles in the RAI Standard.
(EA6) Have relative materiality scores been recorded for values and core concepts, and has materiality of the commitments and/or specifications been considered? Part of this should include the identification of stakeholders who will be able to see the specification results and be able to hold the team accountable to their own commitments.	The bank has detailed guidance which allows us to assess materiality for stakeholder impact including fairness. These are reviewed by the RAI Review Forum and RAI Council before go-live and also monitored post-go-live.
(EA7) Are recourse mechanisms available to data subjects being used? Is there an explanation if the use of recourse is too high or too low?	Additional recourse mechanisms were not required for this use case as the overall process has not changed and existing recourse mechanisms remain in place.
(EA8) Are controls in place to revisit the specifications to ensure they are measuring progress toward the commitments and recalibrated as needed?	There is a tiering mechanism that assigns periodicity of model review and revalidation for the use case depending on the materiality assigned to the use case. Signature Verification has an annual model revalidation and quarterly model review in place. Based on this, the model will be reviewed to measure progress and recalibrate as needed.

Table 5 – Transparency Checklist

Transparency Checklist	Summary of Outcome
(T1) Has the FSI defined the factors it will use to determine whether external transparency is essential for a particular AIDA use case?	Covered in Section 6.6.4.
(T2) (Where an FSI has chosen to provide external transparency.) At each stage of the FSI's customer lifecycle, has the FSI determined what proactive or reactive communication may be needed, and standard templates/interfaces for the same?	
(T3) Has the FSI defined the factors it will use to determine the extent of, and audience for, internal transparency for individual AIDA use cases?	
(T4) Has the FSI defined an acceptable set of AIDA ML explanation method(s) for use within the firm?	
(T5) Has the FSI set minimum accuracy standards for such explanation methods?	
(T6) Has the AIDA use-case team determined whether there is a need for external (customer-facing) transparency? Apply standards from T1 to help answer the question.	The use case is not client-facing and there is a human-in-the-loop element that makes the final decision and communicates with the client as appropriate. Hence additional external transparency is not required.
(T7) If yes, has the team identified the proactive and reactive communication needed at each stage of the customer lifecycle, and the form of such customer-facing communication? Apply standards from T2 to help answer the question.	
(T8) Has the team determined the level of internal transparency needed and the audiences for the same? Apply standards from T3 to help answer the question.	The AIDA system owner has been clearly oriented on how the AI model works, its implications on the outcomes, including potential risks and limitations of the algorithm. This understanding is essential for the AIDA system owner to acknowledge, take ownership and demonstrate accountability of the AI model.
(T9) Has the team selected a suitable explanation method for this specific use case from the approved list in T4?	Acceptable AIDA ML explanation methods for unstructured data such as signatures are not yet available in the wider industry and hence yet to be defined at the bank level. In-house methods specific to the use case were developed to inspect a layer-level output which required the data scientists to interpret the model behaviour. In addition, Gradient Class Activation Mapping has been selected as most suited to the neural network algorithm for comparing two signatures. Thresholds are defined by the process owner based on the risk appetite of the process to handle exceptions.
(T10) Has the team ascertained that the chosen explanation method/implementation meets the minimum accuracy requirement for this specific use case (based on T5)?	

Transparency Checklist	Summary of Outcome
(T11) Have internal transparency dashboards/ reports been implemented in line with the requirements agreed in T8-T10?	Internal transparency is demonstrated through the AI Validation report that will be presented to the AI system owner.
(T12) Have relevant first- and second-line control teams – including model validation where relevant – reviewed and approved these outputs (e.g., local, and global explanations, conceptual soundness, etc.)?	Signature Verification was validated by one of the AI Validation teams in the bank. The outcome of the validation is the AI Validation report which ensures conformance of the model to the AI principles and identifies the current strengths/weaknesses of the model for further resolution.
(T13) Where the explanations have not met first-/ second-line expectations, have appropriate mitigation actions been taken? (e.g., switching to a simpler model despite a reduction in predictive accuracy, dropping difficult-to-explain features, introducing more human oversight, etc.)	There is 100% human oversight for Signature Verification. There are no other mitigating actions that have been recommended by the AI Validation team.
(T14) In line with the external transparency requirements agreed in T6-T7, has appropriate system functionality been developed and tested as part of the AIDA system's implementation plan?	(T14) and (T15) The AI model does not have an outcome that directly impacts the customer, and the final decision is always made manually by the human-in-the-loop element. The relevant internal stakeholders are appropriately informed on the model risks/limitations. The operations team has been trained on operating the output of the model. (T16) The use case is not client-facing and there is a human-in-the-loop element that makes the final decision and communicates with the client as appropriate. Hence additional external transparency is not required.
(T15) Have operational processes such as customer service and complaint handling been modified appropriately to incorporate AIDA customer transparency? Have relevant staff been provided with appropriate training to address customer queries?	
(T16) Have customer/website Terms and Conditions been appropriately updated (i.e., to explain how data is used, how benefits and risks to individuals are associated with the processing, and how individuals may participate and object where appropriate)?	
(T17) Does the AIDA system implementation support the agreed internal and external explanations even after “go live” (i.e., not just as a one-off before approval but throughout the lifetime of the AIDA system)?	Yes. Explainability will be revalidated along with fairness and other principles whenever the model goes through retraining or if there is any change to the algorithm due to reasons such as data drift, concept drift, etc. Additionally, explanation is validated as part of the periodic model monitoring.

Model Training and Testing – Representative Sample (Open-source Data Sets)



Figure 6.19.: Training snapshot – sample of a cheque containing client signature that serves as input to the use case.

image	average_red	average_green	average_blue	image_size	width	height	Edge_Mean	blurriness	area
00404013.png	0.90652776	0.90652776	0.90652776	10592	1285	465	6.110781976	6209.274005	0.90652776
00404092.png	0.90652776	0.90652776	0.90652776	10592	1285	465	6.110781976	6209.274005	0.90652776
07004070.png	0.910352969	0.910352969	0.910352969	4842	591	430	4.439145319	4610.831071	0.910352969
07702077.png	0.897086971	0.897086971	0.897086971	11236	931	665	5.650145773	6100.501684	0.897086971
07704077.png	0.887933409	0.887933409	0.887933409	10096	797	656	5.742513848	6246.644524	0.887933409
07705077.png	0.896613539	0.896613539	0.896613539	11416	919	658	5.437761079	6117.082712	0.896613539
08604086.png	0.908725641	0.908725641	0.908725641	7818	1096	436	5.199191807	5213.365114	0.908725641
08804004.png	0.934659847	0.934659847	0.934659847	4942	782	375	4.068695652	4217.921739	0.934659847

Figure 6.20.: EDA dataset of images.

The below charts and tables represent a synopsis of a section of the Exploratory Data Analysis (EDA) output. EDA refers to the critical process of performing initial investigations on data to discover patterns, spot anomalies, test hypotheses and check assumptions with the help of summary statistics and graphical representations. This has influenced some key decisions in the model training.

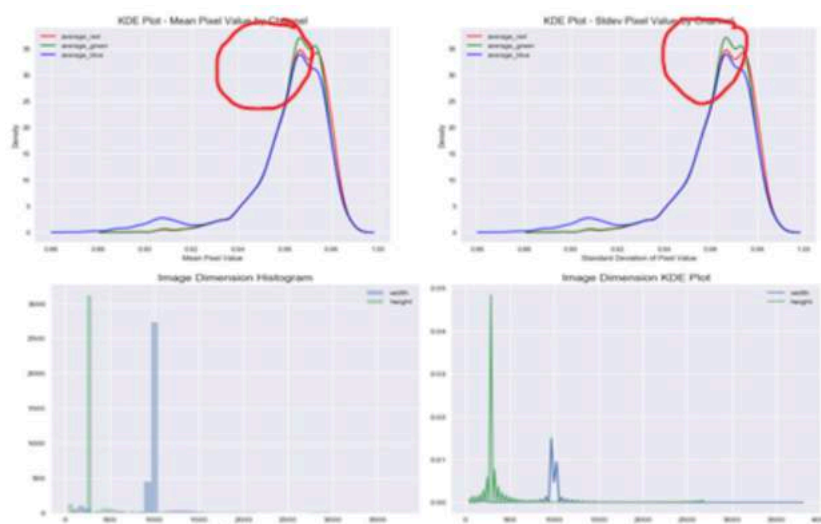


Figure 6.21.: Distribution plots.

Distribution plots visually assess the distribution of sample data by comparing the empirical distribution of the data with the theoretical values expected from a specified distribution



Image : 00501022.png (height : 2222)



Image : 02205022.png (height : 2691)

Figure 6.22.: Outliers samples of signatures (Kaggle.com)



Image : 00703023.png (Blurriness : 41.510599)

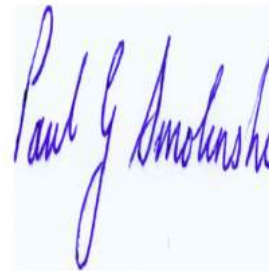


Image : 00702023.png (Blurriness : 50.543436)

Figure 6.23.: Blurriness in signatures.

```
In [12]: (EDA.EDA(Source)).T
```

Total Missing Values = 3479.0 Data Points, Contributing to 94.0 %
Columns with values as 0

Series([], dtype: object)

Out[12]:

	whiteness	average_red	average_green	average_blue	image_size	width	height	Edge_Mean	blurriness	area
count	208.000000	3687.000000	3687.000000	3687.000000	3.687000e+03	3687.000000	3687.000000	3687.000000	3687.000000	3687.000000
mean	74.350072	0.965031	0.964514	0.961089	2.650664e+05	981.769731	302.779496	3.923215	2885.466063	0.966710
std	15.258185	0.013100	0.013095	0.018814	3.950902e+05	265.016559	145.944252	2.960839	924.137994	0.015776
min	26.145000	0.887933	0.887933	0.867869	3.771000e+03	72.000000	36.000000	0.984226	41.510600	0.887933
25%	64.280000	0.958536	0.958354	0.955914	2.656280e+05	977.000000	281.000000	2.704392	2340.747669	0.958536
50%	74.150000	0.966907	0.966591	0.965622	2.785480e+05	978.000000	283.000000	3.258771	2845.314767	0.966920
75%	85.365000	0.974490	0.973662	0.973316	2.807860e+05	1027.000000	284.000000	3.967499	3446.555011	0.974928
max	100.000000	0.990357	0.990357	0.990357	1.383418e+07	3786.000000	2691.000000	25.414443	6246.644524	1.000000
Kurtosis	0.000726	2.507750	2.907489	3.630214	5.721159e+02	19.159754	79.597900	17.338478	1.314351	1.489690
Skewness	-0.342111	-1.246619	-1.334964	-1.814102	2.145840e+01	1.345246	6.812720	4.039468	-0.239187	-0.243331
Range	73.855000	0.102423	0.102423	0.122487	1.383041e+07	3714.000000	2655.000000	24.430217	6205.133924	0.112067
IQR	21.085000	0.015955	0.015308	0.017402	1.515800e+04	50.000000	3.000000	1.263107	1105.807341	0.016392
Missing Values	3479.000000	0.000000	0.000000	0.000000	0.000000e+00	0.000000	0.000000	0.000000	0.000000	0.000000

Figure 6.24.: Summary statistics.

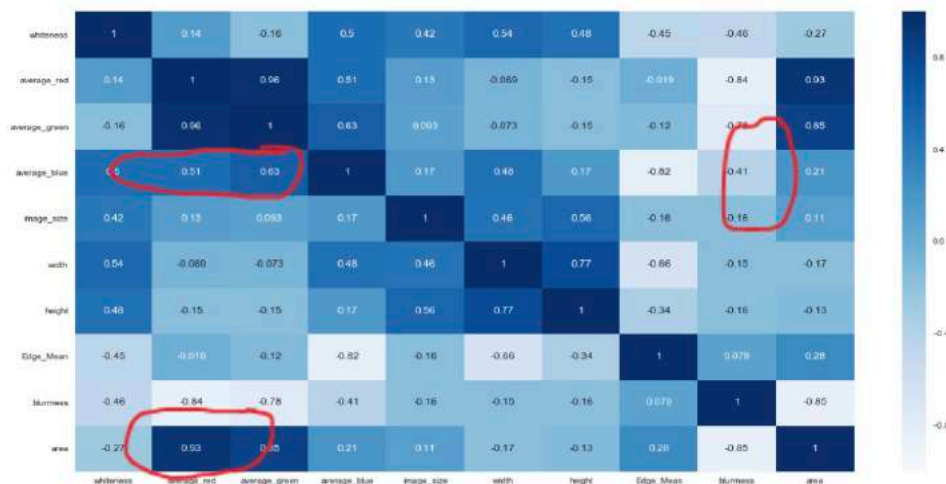


Figure 6.25.: Correlation plot.

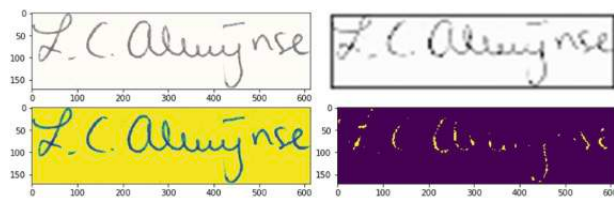


Figure 6.26.: Image clean-up

Based on the EDA, noise removals were done as part of pre-processing which includes grey scaling of data, thinning of the image, thresholding, removing seals from signature, etc.



6.7 UOB

6.7.1 Overview

At UOB, we believe in responsible innovation. As we reap the benefits from using Artificial Intelligence and Data Analytics (AIDA), it is increasingly important that we are guided by our vision and values to do right by our customers and stakeholders. By using AIDA fairly, ethically, with accountability and transparency, we strengthen customer trust and create significant business value and competitive advantage, achieving the vision of having the most trusted data and insights for the world's most trusted bank.

With the release of the MAS FEAT Principles, UOB, as one of the lead members of the Veritas consortium, collaborated with MAS and its partners to develop the Assessment Methodology for Fairness on a credit risk scoring use case and subsequently the Assessment Methodology for Ethics and Accountability on a customer marketing use case in Veritas Phase 1 and Phase 2 respectively.

The Veritas methodologies have since been applied progressively in UOB to ensure that our processes, technology and people align and adhere to the MAS FEAT Principles. Additionally, to establish a broad-based comprehensive implementation approach and to enable a diversity of thoughts, a multi-disciplinary Data Ethics Taskforce which comprises senior executive representation from various business segments and control functions such as the Data Management Office, Customer Experience and Advocacy, Risk, Legal, and Compliance was set up to help drive the strategy and operationalisation of the Data Ethics Governance Model across the bank.

This paper provides a summary of how UOB has adopted the MAS FEAT Principles and integrated the Veritas methodologies into our existing data ethics governance processes on the Advanced Scam Analytics use case.

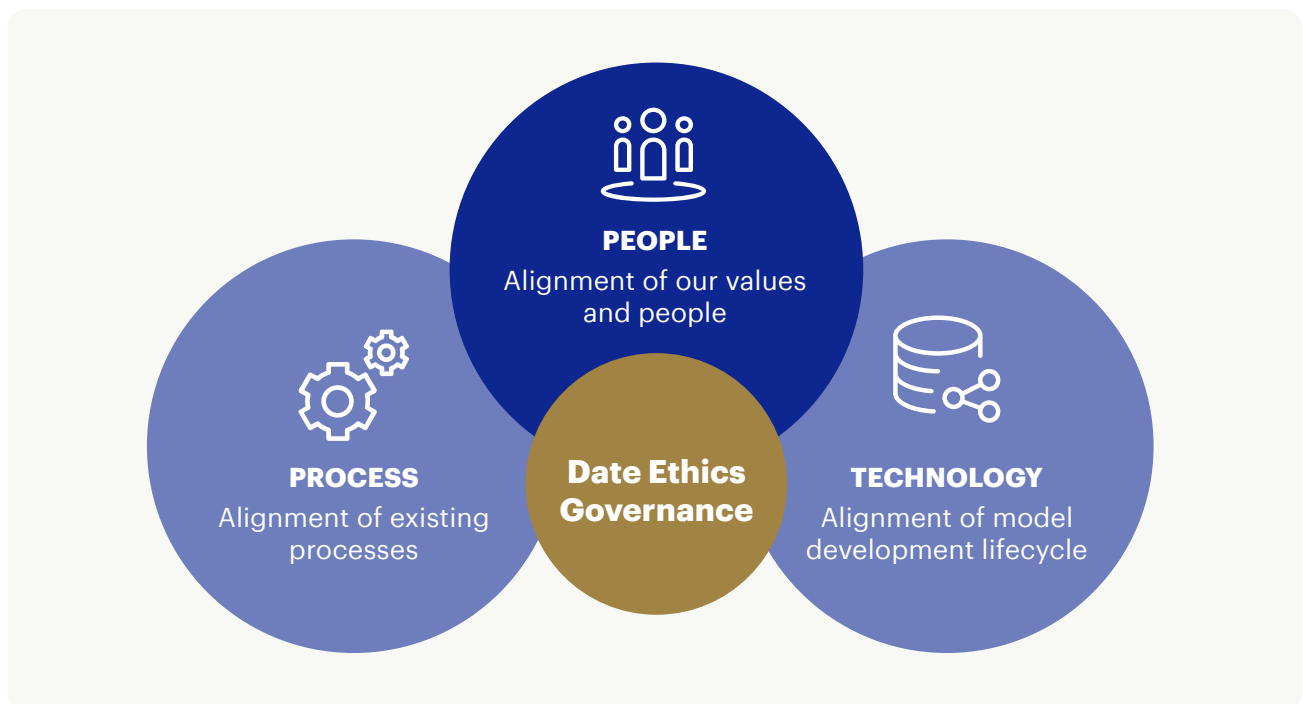


Figure 6.27.: People, process and technology.

A Multi-disciplinary Taskforce to look into the alignment of AIDA solutions with MAS FEAT Principles

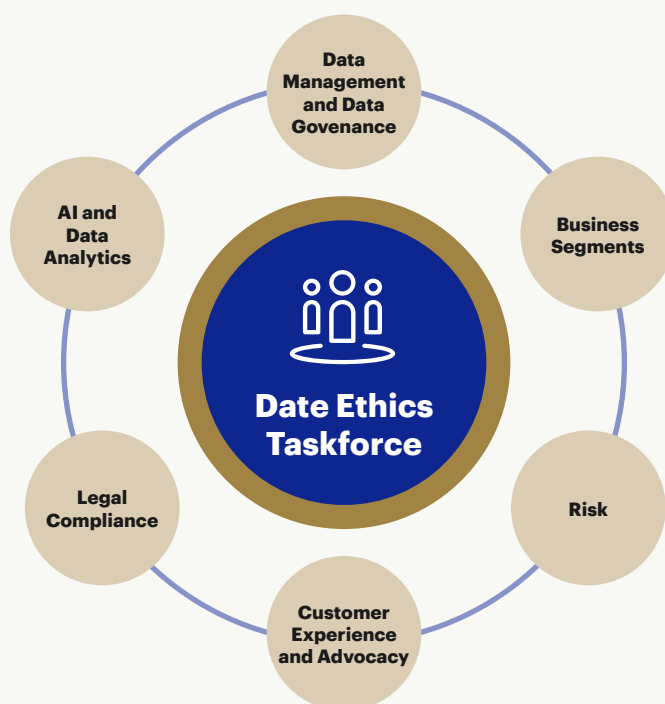


Figure 6.28.: Data Ethics Taskforce.

6.7.2 AI/ML Governance

Managing risk is integral for UOB to create long-term value for our customers and stakeholders. Our risk culture demonstrates behaviours that are consistent with our core values – Honourable, Enterprising, United and Committed. Being a well-established bank, UOB has frameworks and policies in place to manage model risks effectively and comprehensively.

Integrating the MAS FEAT Principles and Veritas methodologies into the existing governance framework has nevertheless presented new challenges. Traditionally, our Model Risk Governance Framework focuses on the management, validation and monitoring of risks for compliance and regulatory models. With the integration of the MAS FEAT Principles, the scope of assessment has been expanded to cover all AIDA models with results that directly impact customers or employees. In addition, the Veritas Fairness Toolkit is now used for the assessment of fairness in AIDA models.

The Data Ethics Governance Model was also introduced to ensure fair, ethical, accountable and transparent use of AIDA solutions in the bank.

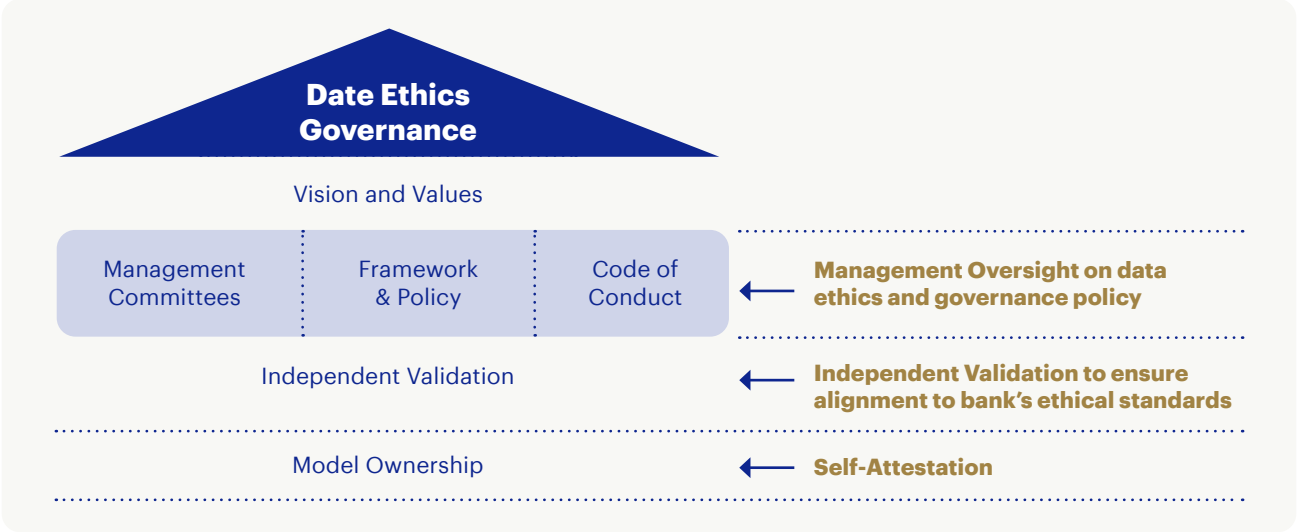


Figure 6.29: Data Ethics Governance model.

In the first tier of the governance model, the business and support functions will perform an annual self-attestation to ensure that they comply with the MAS FEAT Principles in their use of AIDA solutions. At the same time, the model owners are required to perform a materiality assessment of their models to assess the impact to the bank, regulatory compliance, as well as customers, employees and wider society. Records of the self-attestation and model details including the materiality assessment will be updated in the central registry and reviewed by the Data Ethics Validation Team.

High-materiality models which impact customers or employees will be subjected to independent data ethics validation in the second tier of the governance model. Under this tier, the Data Ethics Validation Team plays the role of an independent party to review the use of the AIDA models to ensure that they are aligned with the bank's business directions and ethical standards.

Under the last tier, corporate governance oversight has been put in place to review and oversee the policies and processes on AIDA risks.



Materiality Assessment

To ensure the risk profiles are commensurate with the materiality levels of the AIDA models, a risk-based approach has been adopted to subject only high-materiality models which impact customers or employees to a more extensive analysis and validation by the independent Data Ethics Validation Team.

The materiality assessment consists of both quantitative and qualitative considerations to ensure that the assessment is comprehensive and holistic. This includes assessing the severity and probability of occurrence of harms on the bank from financial, reputational and legal perspectives, on customers from a socio-economic perspective, and on regulatory compliance.

An illustration of the Materiality Assessment Matrix is shown in Figure 6.30. For example, models that perform credit scoring are generally assessed to be high materiality as they pose a major risk to the bank, and any inherent bias in the models could result in a systematic disadvantage to our customers.

		Severity of Harm		
Probability of Harm happening	Model Materiality	Severity of Harm	Severity of Harm	Severity of Harm
	Likely	High	High	Low
	Possible	High	High	Low
	Rare	High	Low	Low

Figure 6.30.: Materiality Assessment Matrix.



6.7.3 AI/ML Process

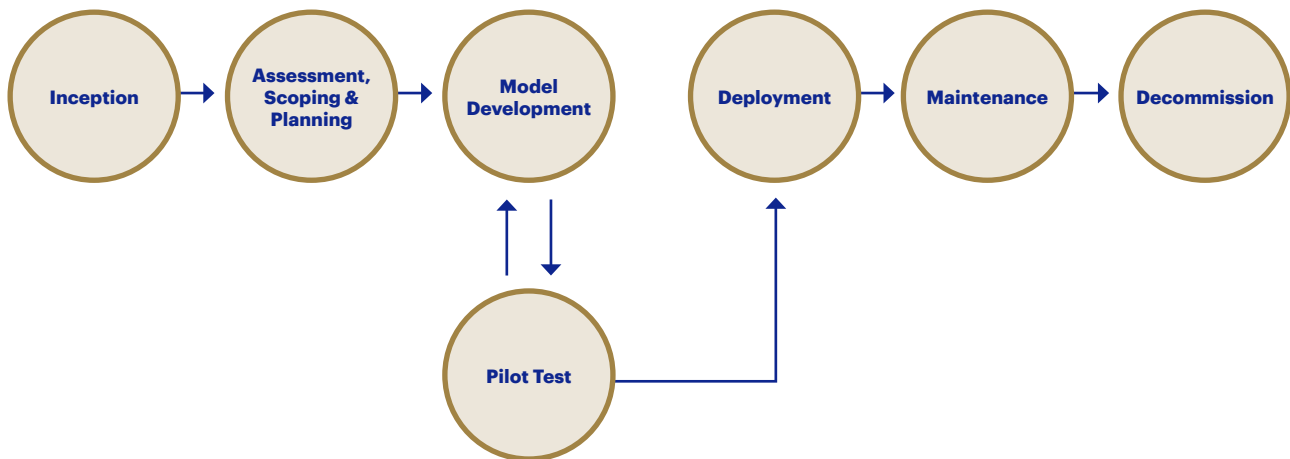


Figure 6.31.: AIDA model development lifecycle.

New requirements have been introduced as part of the integration process of incorporating the MAS FEAT Principles and Veritas methodologies into our AIDA model development lifecycle.

For example, during the model design and development phase, requirements such as a materiality and fairness assessment have been added to ensure that the model complies with data ethics governance processes. The steps involved in model development are generally summarised as follows:

- Assess the feasibility of the business case.
- Assess the materiality of the model.
- Develop the model and document the development process in the model documentation (which includes the business objective, fairness objective, use of personal attributes and their justification, and model explainability, among others).
- Perform the fairness assessment.
- Seek approval from the appropriate authority for the use of the model.
- Deploy the model.

All model owners and developers are also informed of their responsibility to ensure their AIDA models comply with the FEAT Principles.

For high-materiality models, an independent data ethics validation will be performed by the Data Ethics Validation Team before deployment. To support the validation team and model development team in the assessment, the Veritas Fairness Toolkit was deployed in our model development platform and made accessible to the teams for detailed and structured fairness assessment.

In addition, as part of the regular review process, the high-materiality models will be validated on a periodic basis for any unintended harms, and mechanisms will have to be deployed to mitigate these harms.

The Journey to Increase RAI Culture and Training

As the concept of data ethics is relatively new, a broad-based approach has been adopted to drive RAI awareness in the organisation across all levels. At the strategic level, an engagement session in partnership with an external industry leader was set up to raise awareness to senior management on the global AIDA landscape and data ethics principles. At the operational level, sharing sessions and workshops were organised to raise awareness on the FEAT Principles. On an annual basis, model owners and developers are briefed on the FEAT Principles as part of the annual FEAT self-attestation exercise.

In addition, UOB is committed to the professional development of our staff on RAI. In 2022, UOB emerged as one of the top 10 champions in Artificial Intelligence Ethics and Governance certification at the Tech3 Forum organised by the Singapore Computer Society (SCS) and supported by the Info-comm Media Development Authority (IMDA). Our Data Ethics Validation Team has also attained the SCS-NTU certification on AI Ethics and Governance by the AI Ethics and Governance Certification Committee, Singapore Computer Society.

6.7.4 Use Case Study

Description of the identified use case

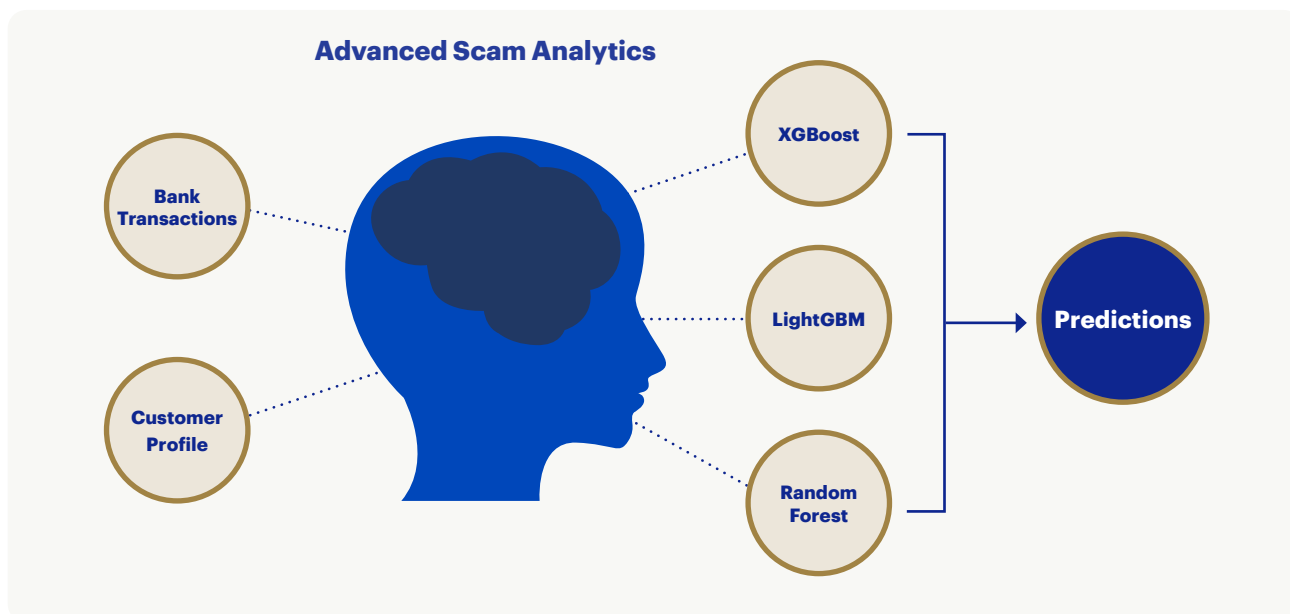


Figure 6.32.: UOB's Advanced Scam Analytics solution.

Over the past few years, Singapore has seen an increase in the number of scam cases. The situation has exacerbated recently with the higher volume of online activities due to COVID-19. Given the rising number of scam cases, UOB has developed an Advanced Scam Analytics solution to predict potential scam transactions using customers' banking transactions and demographic data. This model leverages an advanced machine learning algorithm to predict scam transactions with the highest accuracy and lowest false-positive rate possible.

Lessons Learned from the Application of the Assessment Methodologies

In applying the Veritas methodologies and the Veritas Open-Source Fairness Toolkit on the Advanced Scam Analytics solution we have made several observations with illustrative examples provided in the following Veritas checklist. The AIDA system in this use case is an illustrative example, and the scope of this case study focuses on assessing adherence to fairness, ethics and accountability of the MAS FEAT Principles.

• General

- G1** Is AIDA defined in the FSI?
- G2** Is there a framework in place to define roles and responsibilities for AIDA projects/use cases?
- G3** Is there a materiality framework in place? If so, do you have a clear definition to determine the different level of materiality?
- G4** Does the FSI have an up-to-date inventory that tracks all material AIDA use cases in use?

These foundational questions have been addressed in UOB's internal processes described in the earlier sections.

- G5** Have you documented the commercial objectives of the system and the quantitative measures to meet those commercial objectives? Is it documented how is AIDA used to achieve them?

The objective of this AIDA solution is to improve the accuracy of scam detection by learning from customers' past transactional behaviour and demographics.

- G6** Are there scope boundaries of the proposed system? For example, specific geographies/segments?

The scope of this solution focuses on retail customers in Singapore only.

- G7** Is AIDA use justified for this particular objective/use case (based on limitations of alternatives, proof of AIDA being used successfully elsewhere, availability of adequate data to train the AIDA system, valid legal basis, etc.)?

As machine learning models thrive in use cases where data is abundant and are better in identifying patterns and trends, the use of AIDA is justified for this solution.

G8 Has the materiality of the use case been determined using the framework defined in Step 0?

The materiality of the use case has been assessed with respect to the impact on the bank, the socio-economic well-being of the customers and regulatory compliance. The probability of harm happening has also been assessed and the materiality has been mapped out accordingly.

G9 Have you documented the data used by the system including types of data available, the provenance (source), attributes, dictionary (meaning), vintage (time period), data (quality) profile, representativeness and informed consent? For any personal data used by the system, has a DPIA been conducted?

All data used in the model development are assessed for quality, recency and traceability. DPIA will be conducted where applicable.

For the purposes of this illustration, we have used a publicly available dataset. It comprises transactional data and customer demographics such as gender, income level, education level and age.

G10 Have you documented the data pre-processing and engineering done by the system?

Different techniques such as binning, one-hot encoding and log transformation have been applied to generate the features.

G11 Is the composition of the AIDA system defined? Is it clear how each component of the system AIDA models, as well as business rules and human judgement, if relevant, are used to achieve its commercial objectives? Are the performance estimates and the uncertainties of those estimates documented?

The AIDA system comprises the AIDA model, business rules and operational processes. The results of the model will be reviewed by the investigators for further follow-up with our customers, if necessary.

G12 Is the system's monitoring and review regime designed to detect abnormal operation (e.g., significant model and/or data drift)?

Model performance metrics are monitored and assessed for model drift over time. The review process is guided and governed by our model risk governance and data ethics governance, detailing key modelling methodologies and steps that adhere to regulatory standards, guidelines and requirements.

G13 Are there fallback and/or mitigation plans in place in case of triggers from the system's monitoring and review regime?

Model retraining will kick in if model deterioration is detected during the continuous monitoring process.

• Fairness

F1 Have you identified and documented who are the individuals and groups that are considered to be at risk of being systematically disadvantaged by the system?

For the purposes of this illustration, we have analysed the fairness metric based on gender and age group. The intent is to determine whether there is any systematic disadvantage to any of these groups.

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

Outcome/ Decisions	Ground Truth	
	Scam Transaction	Non-Scam Transaction
Suspicious	True Positive Benefit to bank: Improved risk management. Benefit to society: Prevent scams, etc.	False Positive Harm to Bank: Unnecessary work and cost. Harm to Customer: Unnecessary disruption and requests for verification. Decline in customers' satisfaction.
Non-Suspicious	False Negative Harm to Bank: Potential penalty by authorities; potential negative reputational impact. Harm to Society: Societal impact due to failure to prevent scam.	True Negative Benefit: Improved risk management. Benefit to customer: Smooth banking experience with no unnecessary disruption.

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

The fairness objectives are as follows:

- To operate at the efficient frontier of trade-offs so that there are no harms (or lost benefits) that could have been avoided without incurring some other harms or lost benefits – e.g., if the model can be made more accurate for all transactions, there will be fewer harms without requiring trade-offs.
- To operate at the level where no group and individual is systematically disadvantaged by the distribution of benefits and harms i.e., low-risk transactions are smoothly executed regardless of their personal attributes.

In selecting the appropriate fairness metric for the assessment, we have taken into consideration the potential unintended bias against customer groups because of their personal attributes i.e., falsely classifying their transactions as scam transactions because of their personal attributes. To best fit our fairness objective, we have chosen "False Positive Rate" as the fairness metric.

- F4** Have you documented key errors, biases or properties present in the data used by the system that may impact the system's fairness?
- F5** Have you documented how are these impacts being mitigated?

Data quality checks have been performed during model development. Such checks include handling for missing value and data outliers. We have also applied an over-sampling technique on the unbalanced dataset where only 0.6 percent of the transactions are related to scams.

- F6** Have you determined and documented personal attributes that are used as part of the operation or fairness assessment of the system?
- F7** Does the process of identifying personal attributes take into account ethical objectives of the system, and the people identified as being at risk of disadvantage?
- F8** Have you assessed and documented every personal attribute and potential proxy for a personal attribute and why its inclusion justified given the system objectives, the data, and the quantified performance and fairness measures?

For the purposes of this use case, the personal attributes used as part of the fairness assessment are gender and age.

- F9** Have you assessed and documented the quantitative estimates of the system's performance against its fairness objectives and the uncertainties in those estimates, assessed over the individuals and groups in F1 and the potential harms and benefits in F2?

True positives (successful identification of scam transactions) improve risk management, while true negatives improve operational aspects such as manpower cost and customer satisfaction.

Performance estimates of the model i.e., precision, recall and F1 Score are computed for the out-of-time analysis base. Using the Veritas Open Source Fairness Toolkit for the fairness assessment, we are able to ascertain that the fairness result is acceptable within 20 percent threshold for gender and age groups.

	Performance Metrics
Recall	0.836 +/- 0.019
Precision	0.104 +/- 0.005
F1 Score	0.185 +/- 0.008

	Fairness Metrics (FPR)	Results based on 20% threshold
Gender (Male, Female)	-0.007 +/- 0.001	Fair
Age (<60 years old, >= 60 years old)	-0.004 +/- 0.001	Fair

F10 Have you assessed and documented the achievable trade-offs between the system's fairness objectives and its commercial objectives?

F11 Have you justified (vs alternative trade-offs) and documented the final fairness outcome for the AIDA system?

Trade-off analysis is not performed for this use case study as the fairness result is acceptable within the 20 percent threshold for gender and age groups.

F12 Does the system's monitoring and review regime ensure that the system's impacts are aligned with its commercial and fairness objectives?

All high-materiality models are subjected to periodic review and independent model validation to ensure models are aligned with commercial and fairness objectives on a periodic basis.

• **Ethics and Accountability**

EA1 Are organisational values defined and described?

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

EA3 Are context-relevant core concepts identified and described?

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

These foundational questions have been addressed in the earlier sections. More details can be found in "Veritas Phase 2 White paper – UOB Case study on E&A Assessment in Retail Marketing."

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

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

The following provides a simple illustration of the details that we have established to assess ethics and accountability of the use case, showing the lines from values and concepts to principles, commitment and specifications.

Values	Core Concepts	Principles	Commitments	Specifications	Priority
Honourable, Enterprising, United, Committed	Fair dealing, Trust, Accountability, Human Involvement	Non-discrimination	Non-discrimination	% of personal attributes having a fairness value within the 20% threshold.	H
		AIDA decisions are held to at least the same ethical standard as human led decisions	Maintain the same ethical standard as human driven decisions for high materiality models	% of review and investigation successfully completed.	H
		Use of AIDA in decision making is approved by an appropriate internal authority	Ensure proper management oversight and approval for the use of internally and externally sourced AIDA solutions	Number of issues raised during management review	H

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

UOB complies with the ABS Code of Consumer Banking Practice in the handling of customers' queries and disputes. There are several channels where customers are able to submit their requests for a review of the bank's decisions that affect them.

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

The model owner will review the commitments and specifications on a periodic basis to ensure that they are properly calibrated.

In addition, all high materiality models are subjected to independent data ethics validation periodically to ensure the commitments and specifications are aligned with our ethical standards and core values.

6.7.5 Concluding Remarks

Since the release of the MAS FEAT Principles, UOB has embarked on our journey to develop and implement an effective data ethics governance model by adopting the best practices in data ethics. A key learning point in our journey is that data ethics is constantly evolving. Therefore, to ensure that our decisions are grounded in our values and guided by our governance framework, it is important for us to continuously educate and create awareness on data ethics across all levels of the bank. Another key learning point is that a risk-based approach is essential for the bank to allocate resources effectively to prioritise and address AIDA risks based on the materiality of the models. Moving forward, the roadmap is to enhance the existing integration with best industry practices such as the Veritas Toolkit 2.0.

With the increasing adoption of AI, UOB will continue to support the AI ethics movement and advocate for the need to use AIDA responsibly. Through the Veritas projects, we are able to contribute to the development of the assessment methodologies for the MAS FEAT Principles. We will continue to strengthen this foundation of trust by building upon the best practices in Responsible AI and data ethics.

07 Conclusion

AI and Data Analytics are currently high priorities for FIs. While these technologies offer the promise of expanded access to financial products and services as well as improved customer experience, they also bring with them new risks on data security and unethical data usage that result in bias and discrimination on different grounds. For these reasons, regulatory agencies globally are in the process of laying down frameworks and methodologies to address the challenges involved in the responsible use of AI within the financial services industry.

This whitepaper has captured the landscape of Veritas methodology Integration by FIs in Singapore, and has illustrated this via the case studies from the seven FIs listed.

The whitepaper highlights common challenges that FIs faced during integration such as the need for additional resources to cover the scope of the work required, as well as the need to refer to multiple policy statements across geographies to draft their own governance frameworks.

In the future, industry will need to address these common challenges to help pave the way for the widespread adoption of AI for the benefit of the industry and for society more broadly.

08 Acknowledgements



James Gan
Project Lead

Bhavna Rawlley
Project Lead

Ray Eitel-Porter
Responsible AI Subject Matter Expert

Dr. Deepa Rajagopalan
Data Science Consultant

Dr. Thanh-Hai LE
Data Scientist



Prakash Arunachalam
Enterprise Data Science Director

Suryanto Ang
Enterprise Innovation Software Engineer

Manish Jain
Enterprise Data Science Group Manager

Vikram Lalit
Head of Data and Analytics –
Clearance and Collateral Management,
PMG and Credit Services

Fengzi Li
Enterprise Innovation Specialist Data Scientist

Hays McCormick III
Enterprise Data Science Fellow

Vikram Singh
Enterprise Data Science Platform Lead

Samprita Sasihithlu
Enterprise Data Science Specialist Data Scientist

Zakie Twainy
Digital Partnerships & Strategic Investing

Jitesh Ravindra Vaidya
Enterprise Data Science Data Engineering Lead

Michal Wojciechowski
Business and Legal Entity Advisory, Model Risk



Sameer GUPTA
Chief Analytics Officer

Simon CHATER
Data Governance

Sandhya JILSON
Data Governance

Jasmine NEO
Data Governance

John Jianan LU
Analytics Centre of Excellence

Wei WU
Analytics Centre of Excellence

Mei LENG
Analytics Centre of Excellence

Rajanikanth ANNAM
Consumer Banking Analytics

Durga Naga Venkata Ramarao ALAMURI
Consumer Banking Analytics

Lakshmi Kiran KANCHI
Consumer Banking Analytics

Juon Qiang Yin
Legal & Compliance



Gunjan Bhatt

Subject Matter Expert

Michael P Hutchinson

Subject Matter Expert

Dena A Beatty

Subject Matter Expert

Lian Y Wang

Subject Matter Expert

Diana L Y Chu

Model Risk Management Expert

David Tan

Project Manager



Guillermo Arbeiza

Chief Customer Officer

Samuel Chu

Head of Data Science and Analytics

Balaraghuvver Alampalli

Data Science and AI Lead

Nithya Krishnamoorthy

Data Scientist

Phanyrath Samnang

Data Scientist

Pauline Tan

Senior Legal Counsel



Robert Cheah

Project Lead

Tan Li Choo

Subject Matter Expert

Chia Bock Sim

Subject Matter Expert

Joyce Chua

Data Governance Expert

Sreeparna Majumder

Data Scientist

Sean Tan

Data Scientist



Donald MacDonald

Head, Group Data Office

Andrea Pisoni

Head, Data Science

Kevin Moe

Data Science Lead

Shou Ruyan

Data Scientist

Seah Tian En, Timothy

Data Scientist

Tan Chee Peng, Andrew

Group Data Governance Office

Tan Chong Lim

Group Data Governance Office



Siddhartha T

Lead Data Scientist

Pradeep Madhavankutty

Data Scientist

Thiripura Hithayakanthan

Data Scientist

Vinaya Menon

Data Scientist

Farooq Husain

Project Co-ordinator

Azra Razak

Project Lead

Emma Johnson

Project Manager

Adhinarayan Nammalvar

Subject Matter Expert

Vijay Jairaj

Subject Matter Advisor



Qiang Zhang

Deputy Director AI Development Office
(FinTech & Innovation Group)

Xuchun Li

Head of AI Development Office
(FinTech & Innovation Group)

09 Bibliography

References

1. <https://www.mas.gov.sg/-/media/MAS/News/Media-Releases/2021/Veritas-Documents-1-FEAT-Fairness-Principles-Assessment-Methodology.pdf>
2. <https://www.mas.gov.sg/news/media-releases/2022/mas-led-industry-consortium-publishes-assessment-methodologies-for-responsible-use-of-ai-by-financial-institutions>
3. Ibid.
4. <https://www.mas.gov.sg/-/media/MAS-Media-Library/news/media-releases/2022/Veritas-Documents-3A---FEAT-Fairness-Principles-Assessment-Methodology.pdf>
5. <https://consumerfed.org/reports/cfa-report-title-forthcoming/>
6. Privileged and Unprivileged Groups: An Empirical Study on the Impact of the Age Attribute on Fairness, Exploring privileged vs. unprivileged population groups in Artificial Intelligence and Machine Learning, available at <https://solar.cs.ucl.ac.uk/pdf/hort2022fairware.pdf>

LEGAL NOTICE

This report is prepared and issued by the MAS, Accenture, BNY Mellon, DBS, HSBC, OCBC, Singlife, Standard Chartered, and UOB.

All intellectual property rights in or associated with this report remain vested in the MMAS, Accenture, BNY Mellon, DBS, HSBC, OCBC, Singlife, Standard Chartered, and UOB and/or their licensors. This report and its contents are not intended as legal, regulatory, financial, investment, business, or tax advice, and should not be acted on as such.

Whilst care and attention has been exercised in the preparation of this report, MAS, Accenture, BNY Mellon, DBS, HSBC, OCBC, Singlife, Standard Chartered, and UOB do not accept responsibility for any inaccuracy or error in, or any inaction or action taken in reliance on, the information contained or referenced in this report.

This report is provided as is without representation or warranty of any kind. All representations or warranties whether express or implied by statute, law or otherwise are hereby disclaimed.