



# Veritas Document 4

---

## FEAT Principles Assessment Case Studies

3

3A

3B

3C

4



# Table of Contents

<b>1</b>	<b>Introduction</b>	<b><u>4</u></b>
1.1	Purpose of the document	<u>4</u>
<b>2</b>	<b>Fairness Assessment in Predictive Underwriting</b>	<b><u>5</u></b>
2.1	Background	<u>5</u>
2.2	Scope	<u>5</u>
2.3	Key Highlights of the Fairness Assessment Methodology	<u>6</u>
2.4	Applying the Fairness Assessment Methodology to a Use Case	<u>8</u>
2.5	AIDA Applications in Life Insurance Underwriting	<u>9</u>
2.6	Use Case Illustration – Predictive Underwriting in Life Insurance for a Cross Sell Campaign	<u>13</u>
2.7	Fairness Assessment Using the Methodology	<u>17</u>
2.8	FS Reflections of Fairness Assessment Methodology	<u>55</u>
2.9	Conclusion	<u>57</u>
2.10	Disclaimer	<u>58</u>
<b>3</b>	<b>E&amp;A Assessment in Fraud Detection</b>	<b><u>59</u></b>
3.1	Preface	<u>59</u>
3.2	Introduction	<u>59</u>
3.3	Learning from Application of the Assessment Methodology	<u>60</u>
3.4	Challenges	<u>64</u>
3.5	Conclusion	<u>64</u>
3.6	E&A Worksheet – AXA Sherlock Fraud Detection	<u>65</u>

# Table of Contents

<b>4</b>	<b>E&amp;A Assessment in Retail Marketing</b>	<b><u>81</u></b>
4.1	Preface	<u>81</u>
4.2	Introduction	<u>81</u>
4.3	Learning from Application of the Assessment Methodology	<u>83</u>
4.4	Challenges	<u>85</u>
4.5	Conclusion	<u>85</u>
4.6	E&A Worksheet – UOB Retail Marketing	<u>86</u>
<b>5</b>	<b>Transparency Assessment in Credit Decisioning</b>	<b><u>103</u></b>
5.1	Executive Summary	<u>103</u>
5.2	Introduction, Purpose and Scope	<u>104</u>
5.3	Use of AIDA in Credit Decisioning	<u>104</u>
5.4	Overview of Transparency Assessment Methodology	<u>106</u>
5.5	Approach to Reviewing the Assessment Methodology	<u>107</u>
5.6	Reflections	<u>123</u>
<b>6</b>	<b>Transparency Assessment in Customer Marketing</b>	<b><u>125</u></b>
6.1	Use Case	<u>125</u>
6.2	Context	<u>125</u>
6.3	Key Components for Transparency Evaluation	<u>126</u>
6.4	Transparency Assessment Approach	<u>126</u>
6.5	Transparency Assessment and Explanations	<u>127</u>
<b>7</b>	<b>Acknowledgement</b>	<b><u>135</u></b>
<b>8</b>	<b>Bibliography</b>	<b><u>138</u></b>

# 01 Introduction

## 1.1 Purpose of the Document

This document is one of a suite of documents published as an output of the Monetary Authority of Singapore (MAS) Veritas Phase 2 project. Its purpose is to illustrate implementation of the Fairness, Ethics, Accountability and Transparency (FEAT) Principles Assessment Methodology for Financial Institutions on selected use cases and it fits alongside the published documents as highlighted in the diagram below.

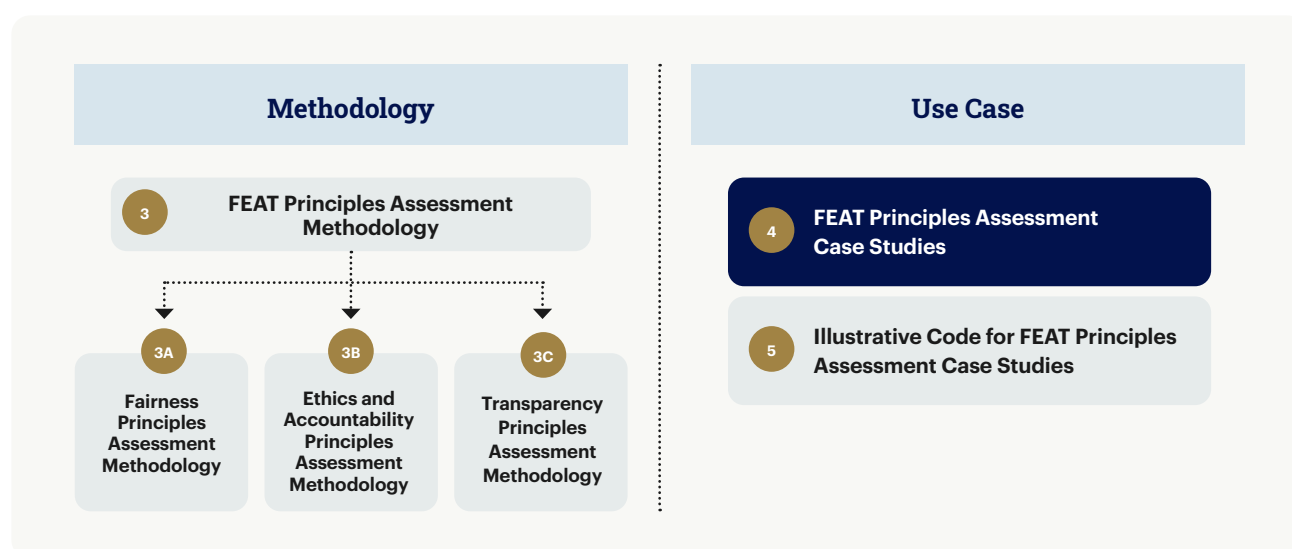


Figure 1.1 Illustrative Case Studies document in the suite of Veritas outcomes

# 02 Fairness Assessment in Predictive Underwriting

## 2.1 Background

Within the foundational FEAT framework on the use of Artificial Intelligence and Data Analytics (AIDA), which was published by the MAS in 2018, the principles that relate explicitly to fairness (“the Principles”) aim to ensure that AIDA driven decisions do not disadvantage any individual or groups of individuals, without appropriate justification of inputs, processes and outcomes. MAS defines AIDA as “technologies that assist or replace human decision making.”

To move the Principles to practical application, Veritas **Phase 1** (published in January 2021) formulated the FEAT Fairness Assessment Methodology (“the Methodology”). In this phase, Veritas provided use case examples from the banking sector. These use cases applied the Methodology to AIDA systems used in credit scoring and customer marketing (please refer to the Case Study document from Phase 1 to review the use cases in full).<sup>1</sup> Veritas **Phase 2** extends the Methodology to the **insurance sector**, with a use case around predictive underwriting in the Singaporean life insurance market for existing insurance customers.

## 2.2 Scope

In Phase 2 of the Veritas project, Swiss Re (**Swiss Re Asia Pte. Ltd.**), Great Eastern (**Great Eastern Life Assurance Singapore**) and Accenture have partnered to apply the FEAT Methodology and validate the outcomes produced in the use case document from a synthetic dataset.

Life Insurance underwriting is a crucial business process for Swiss Re and Great Eastern and requires a robust and effective assessment of associated applications and systems. In the section of the document where Financial Service Institutions (FSI) reflect on the use cases, we share their experiences and findings of the project.

The AIDA system in this use case is a simple and illustrative example. In some areas, we have identified alternative approaches that may be more suitable for complex or high risk AIDA systems. These examples are by no means exhaustive and there are AIDA systems where approaches other than those set out in this document would be more suitable if assessing for fairness. It should be recognised that the field of algorithmic fairness is relatively new, and for some complex areas there is ongoing active research to determine how to address current challenges.

The case study makes no claim on the alignment of the systems presented with the Principles: this is a value judgement to be made by the AIDA system assessor based on the answers to the assessment questions (see section 2.3 for an example assessment process). It is important to acknowledge that examples are illustrative and specific to the respective use cases in the Singapore jurisdiction and only for the specific attributes assessed (illustrations are not generalisable), recognising that the appropriate demographic attributes examined for fairness assessments may be different in other jurisdictions.

As covered in the methodology document, FSIs must continue to comply with all applicable laws and requirements. FSIs are encouraged to calibrate their internal governance frameworks for the FEAT assessment based on their own discretion and taking into consideration their existing frameworks, the materiality of AIDA systems and the cost of FEAT assessments and potential mitigation. The document acknowledges that an FSI's specific obligations around AIDA fairness will depend on the regulatory requirements and applicable laws of the jurisdiction in which they operate, as well as the organisation's values and existing governance standards, which are also likely to change over time. Therefore, the Methodology is not prescriptive, but instead is aspirational, providing guidance and recommendations for relevant audiences. ***It is important that the level of FEAT fairness assessment is proportional to the fairness risk of any use case as these assessments will need to be implemented with additional costs, which ultimately get passed on to consumers.***

The Methodology described in this whitepaper is only applicable if personal attributes, including personal and sensitive personal data, are collected and processed in accordance with all applicable regulatory requirements. Personal attributes are defined as features about individuals that should not be used as the basis for decisions without reasonable justification. Personal data is defined as data, whether true or not, about an individual who can be identified either from that data or from that data and other information to which the organisation has, or is likely to have, access.



## 2.3 Key Highlights of the Fairness Assessment Methodology

The Methodology presented in the Fairness Methodology paper for assessing alignment with the Principles indicates a series of guiding questions to be **answered by FSIs and presented to AIDA system assessors** for Fairness evaluation. The Methodology aims to propose a process that would enable fairness assessments to consider the **end-to-end AIDA systems development lifecycle** rather than simply focusing on the algorithmic models, thereby encompassing all the aspects of the FSIs' operations that contribute to the AIDA driven decisions including business rules, manual overrides and monitoring aspects. The diagram below shows how the Methodology aligns to a typical AIDA System development lifecycle:

## Embedding fairness checkpoints in a typical AIDA system development lifecycle

Use Case or AIDA System Specific

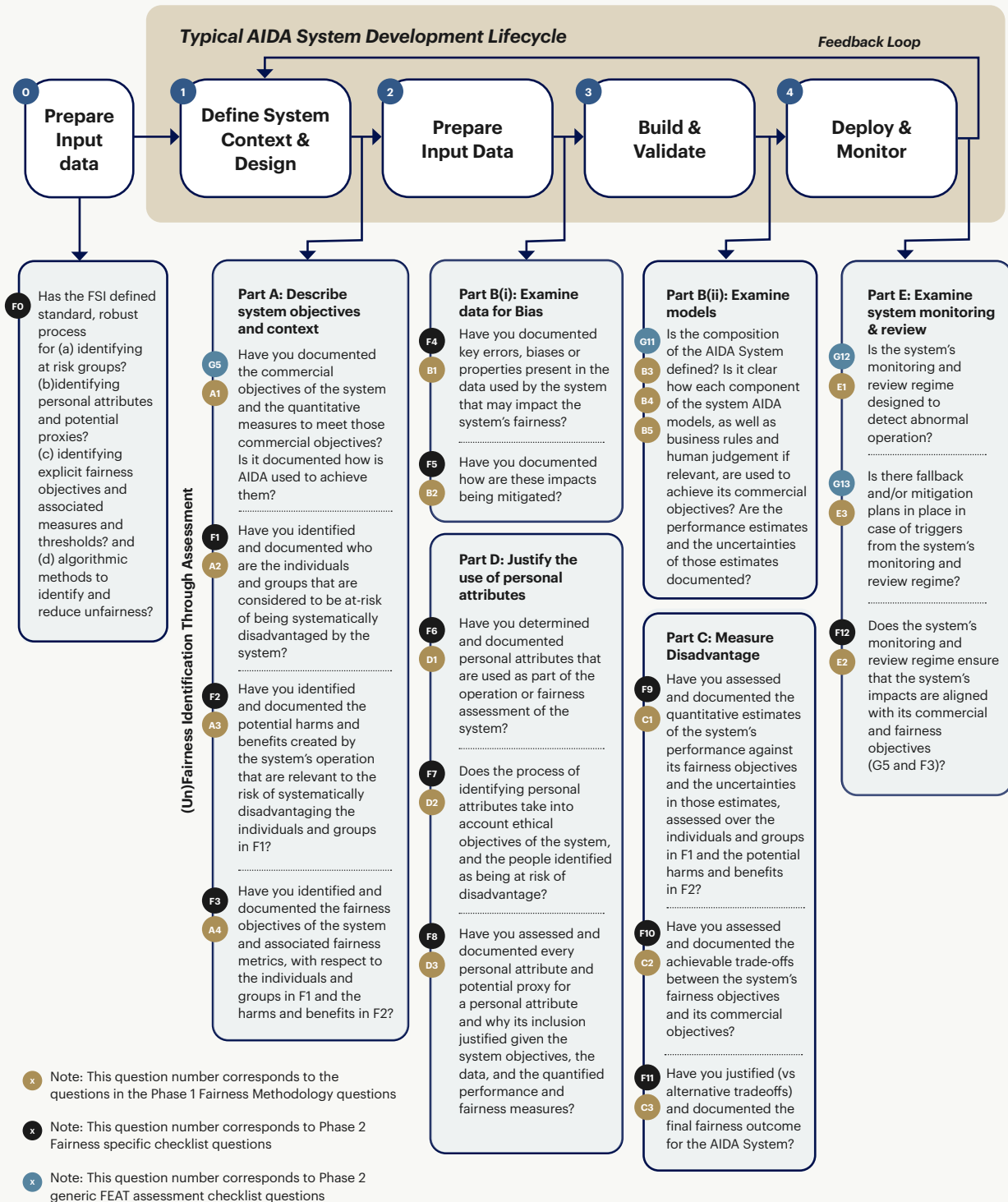


Figure 2.1 Mapping the FEAT Fairness Assessment Methodology (defined in Veritas Document 1 - FEAT Fairness Principles Assessment Methodology, Veritas Consortium, published in December 2020 by MAS) to a typical AIDA system development lifecycle

## 2.4 Applying the Fairness Assessment Methodology to a Use Case

Once an FSI's risk management processes have determined that an AIDA system is in scope, then the system can be assessed using the Methodology. The answers to the assessment questions in the Methodology are provided by AIDA system owners, by seeking details from the AI system developer (which can be a team within the FSI or a third party). Please refer to the FEAT Principles Assessment Methodology whitepaper and its glossary for full definitions of AIDA system roles. The following guide is an indicative example of how this assessment could be carried out:

1. AIDA system owner completes a risk assessment to determine an appropriate customisation of the Methodology for the system's risk (see Section 2.6.3). This could include completing "Part A: System Objectives and Context" of the Methodology, or a similar summary of the system.
2. AIDA system owner provides the summary to AIDA system assessor (who is chosen based on risk level), to further refine the scope of the assessment if necessary, such as agreeing on the boundaries of the system, and which elements of the Methodology are relevant to the system.
3. The AIDA system owner works with the AIDA system developer to gather the relevant information and perform the relevant analysis, producing answers for some or all of the questions of the Methodology, as appropriate to the risk level of the system.
4. The AIDA system owner presents the results of the analysis to the AIDA system assessor, who, based on this analysis, judges the system's alignment with the FEAT Fairness Principles.
5. Based on the feedback of the AIDA system assessor, the AIDA system owner works with the AIDA system developer to make changes to the system as appropriate.
6. After internal feedback has been addressed, the assessment results and resulting actions can be shared with external stakeholders, such as supervisory authorities, as appropriate.

The assessment process can be consistent, whether the AI system developer is a team within the FSI or a third party. When a third party is responsible for the development, FSIs may demand the development and testing process to adhere to their internal requirements and request conclusive documentation.

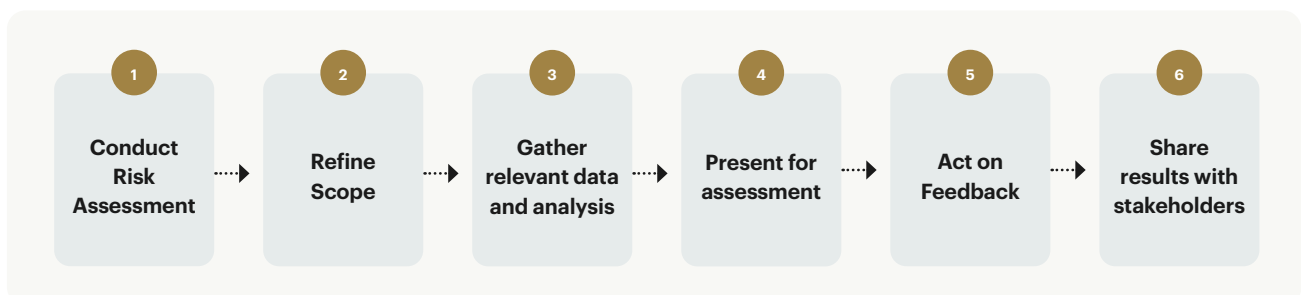


Figure 2.2 – A potential workflow for applying the FEAT Fairness Assessment Methodology

## 2.5 AIDA Applications in Life Insurance Underwriting

### 2.5.1 Introduction to Insurance Underwriting

The insurance industry operates around the premise of risk (where risk can be defined as an uncontrollable, potential loss of something of value) and risk sharing, where the insurer accepts a portion of an insured's financial risk in exchange for an agreed amount of money, or premium. An **insurance risk**, quite simply, is the threat that a potential loss covered by an insurance policy will occur. While insurers agree to assume certain risks in exchange for receiving premium payments from their policyholders, they also determine which risks to assume. To stay profitable, insurance companies need to be selective about the risks they assume or they are in danger of paying more in claims and operating expenses than they receive as premiums and indirect interests. Hence, an accurate insurance risk assessment of customers is needed in order to decide whether to issue a policy and at what price.

Underwriting is the process of evaluating and quantifying the financial risk associated with providing insurance coverage for an individual or group of individuals for an item of value – for example their assets, liabilities, life or health. **The Life & Health underwriting process involves assessing the likelihood of risk to be insured for an individual. The likelihood of risk can be based on personal attributes, which in Singapore can currently include age, health, occupation and medical history amongst others.** The decision process and data to be used is regulated based on the product line and depending on the jurisdiction. Having a view of the degree of insurance risk for a person or investment, underwriters are better informed to, among other things, set indicative/variable premiums to cover the cost of insuring with fair rates and coverage in exchange for taking on the identified risk.

The historical considerations and context of underwriting and pricing, such as information asymmetry, risk pooling and market competitiveness, must be considered when integrating FEAT principles for underwriting use cases to achieve an efficient and equitable outcome.

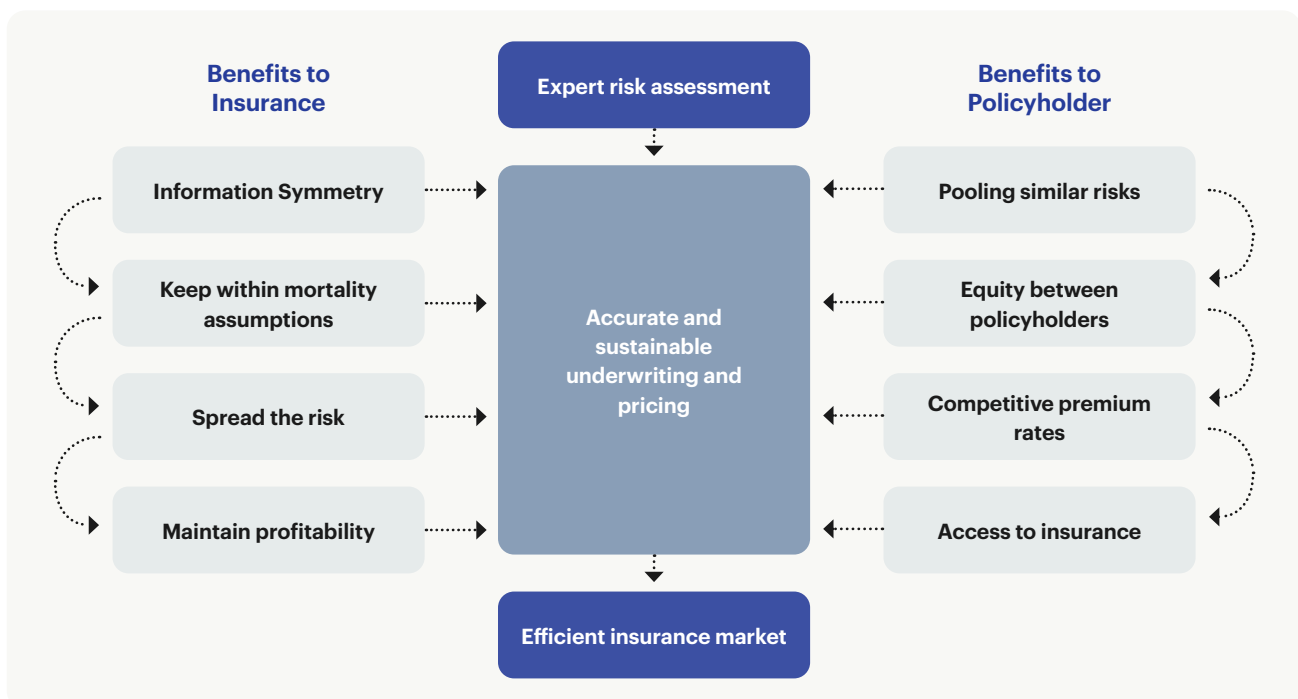


Figure 2.3 - Providing fair underwriting/premiums to all by assessing and pooling equal risks (for more details, see: Christoph E Nabholz. (2011). Fair risk assessment in life and health insurance)<sup>2</sup>

## 2.5.2 Increasing the Use of AIDA for Predictive Underwriting in Life Insurance and its Benefits

The life insurance underwriting journey involves numerous manual tasks ranging from information gathering to the assessment of the risk and issuing a policy. Underwriters traditionally rely on underwriting guides in the risk selection process, which are developed and built using various actuarial analysis, medicals journals, underwriters' experience and data collected over the years.

The traditional underwriting process starts with a customer applying for insurance, at which time the application is referred to the underwriter for review. Subsequently, the underwriter reviews the application and may request additional information or documents from the insured, or suggest medical evaluations to better understand the risk. Based on the provided information, the underwriter can accept or reject to enter into a policy depending on its anticipated potential losses and the company's underwriting guidelines. Depending on the risk classification for the individual and underwriting decisions, policies can have personalised premiums, differential benefits, exclusions, or terms and conditions.<sup>3</sup>

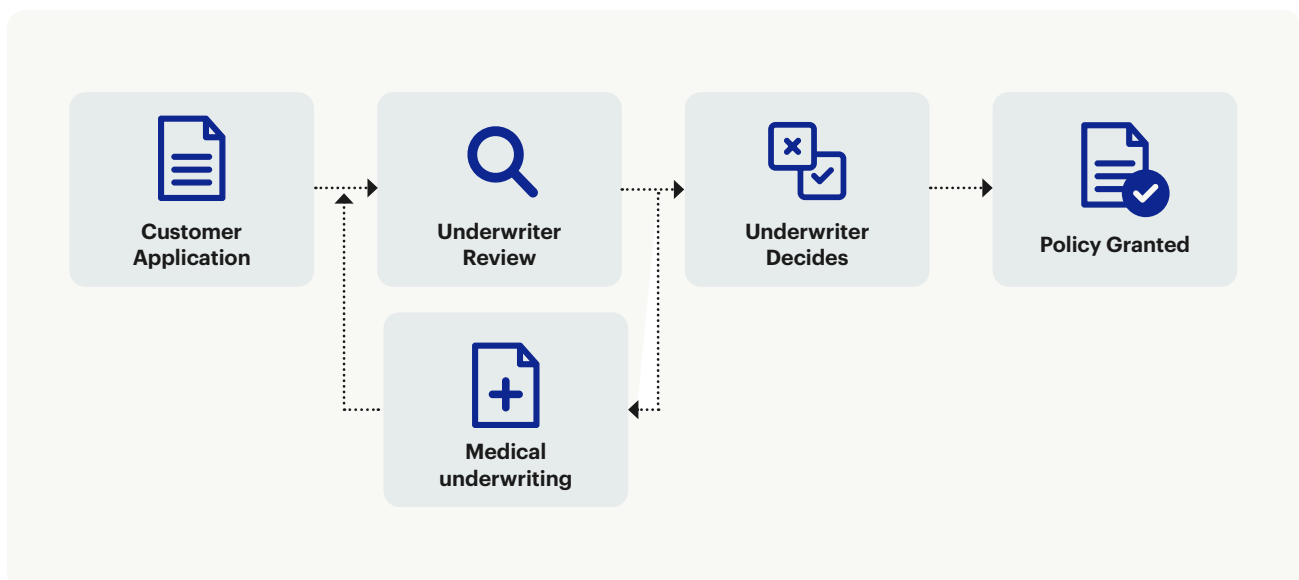


Figure 2.4 - Traditional underwriting process

It is quite common in complex cases for an insurance company to take weeks to gather information and evaluate an application before a decision on eligibility can be made. To accelerate the decision, automated underwriting engines can be used to orchestrate the underwriting journey in which the underwriting rules and forms are used under the guidance of underwriters and in accordance with all applicable laws and regulations. The use of predictive analytics encodes customer information collected by the insurer and underwriting rules as objective inputs, which are then used by AIDA systems to classify applications by their degree of risk. The applications with a low level of risk (Eligible Cases) may be processed directly (Accelerated Underwriting), which gives an opportunity to provide a seamless experience for the customer, while applications with higher or uncertain levels of risk (Low Confidence Cases) are referred to underwriters to follow the standard full underwriting process.

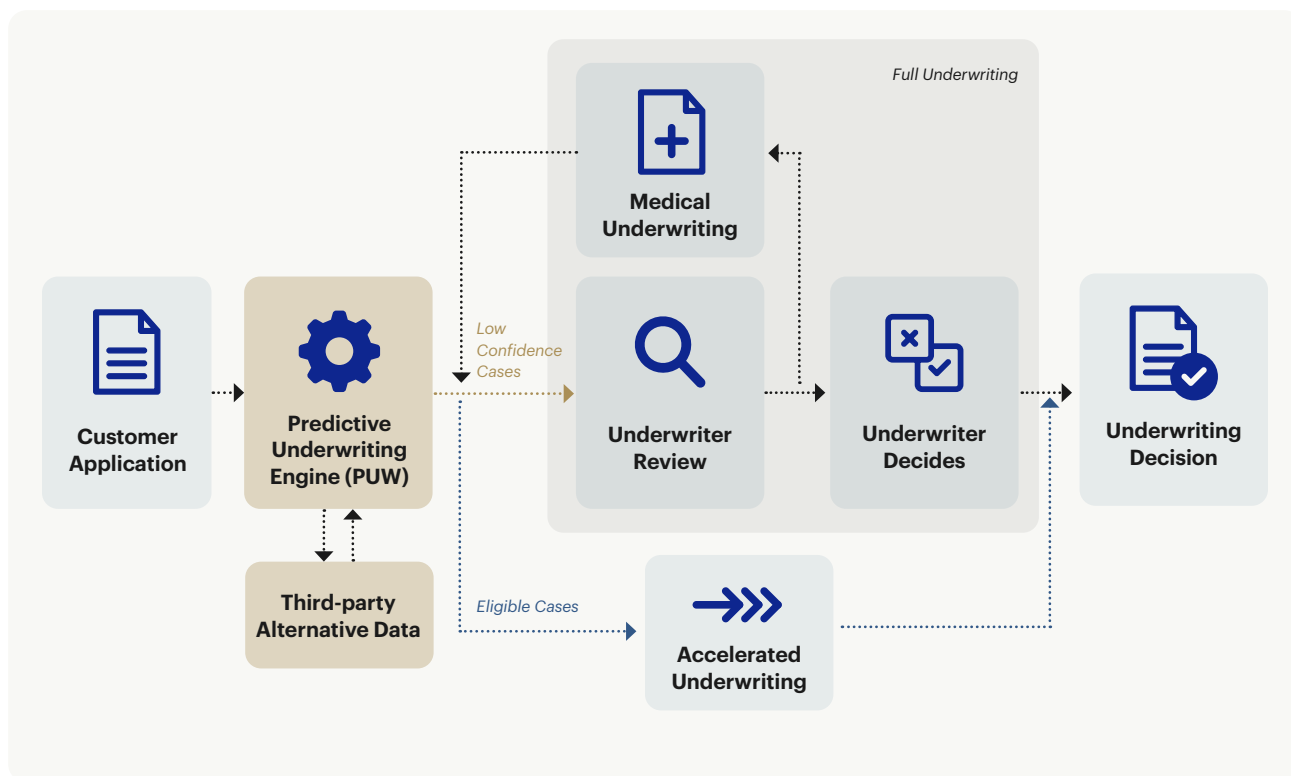


Figure 2.5 - Predictive underwriting process

Beyond the information provided by the applicant, insurers are also considering the use of additional third party data sources in accordance with all applicable laws and regulations. Where usage is approved by customers and regulators, third party information, such as from wellness apps, wearable devices or electronic health records, provide good risk insights that can help simplify the customer journey at the time of underwriting and benefits both the insurer and the insured. Predictive Underwriting (PUW) helps life insurance companies segment and underwrite risks using data and domain expertise from underwriters. This approach reduces costs and saves valuable time when processing an insurance application. In turn, these benefits result in an improved customer journey, increased customer satisfaction and enhanced affordability. For the insurer, the approach yields a better understanding of the risk they are onboarding.

PUW is an extension to traditional underwriting and has the potential to increase access to insurance by making coverage more available, cost effective and consumer friendly. With access to a greater variety/volumes of consented/approved data at the time of underwriting, PUW promises to accelerate decision making at scale for underwriters when managed properly.

In the future, data driven PUW solutions could also be used by insurers to encourage customers adopt healthier lifestyle by providing them with insights into their health risk exposure or around how to better manage existing conditions. Any development in this area would have different implications on the fairness assessment and would need to be considered accordingly.

For the use case covered in this document, we examine the application of a PUW solution for life insurance products in the Singaporean market as part of a simplified underwriting campaign with no price loading and only for existing customers (i.e., cross sell).

## 2.5.3 Fairness Considerations for Predictive Underwriting in Life Insurance

There is an increasing attention to the fairness – and, more generally, the ethical – aspects of AIDA driven decisions. This is for a number of reasons, including:

- The ability of automated AIDA decisions to scale and have an impact on a wider group of customers compared to alternatives that are solely based on human decision making.
- The ability of AIDA to differentiate risks to a finer degree than traditional guidelines and rules.
- Privacy concerns arising from the increasing use of personal and alternative data sources and the involvement of third party data and developers.

Erroneous or biased decisions from automated systems may or may not be frequent, but if not detected early, can lead to wider societal impacts. Having an appropriate governance process in place to detect erroneous or biased decisions is therefore important.

For example, an insurance company developing an AIDA system to determine who in an insured population should have cover for a procedure based on their health status, may assume that the amount an individual has historically spent on healthcare is a good proxy for their health status. Such an assumption could introduce measurement bias if certain subgroups in a population have spent less in the past because they have less money, not because they are healthier. In this case, if the naive AIDA system was subsequently deployed over the whole population it would scale to impact the whole of this subgroup.

Due to such (and many other) potential scenarios of unfairness in AIDA systems, it has become important to incorporate fairness assessments. **Incorporating fairness assessments depends on the business purpose of the use case at hand and the risk level of the AIDA system.** There may be scenarios where the same AIDA model could be used for different purposes. For example, a model used to predict the risk level of individuals for life insurance may be used for any of the following purposes:

- A marketing campaign.** Providing an added product/service to eligible customers with a guaranteed or simplified offer (i.e., cross selling alternate insurance products, other than the ones they already possess and have been underwritten for). This is the scenario used in this use case whitepaper.
- Auto approving.** Simplify the underwriting process and hence outcomes for eligible customers on customer initiated applications.
- Auto declining.** Ruling out certain products/services for ineligible customers on customer initiated applications.

The first two scenarios relate to providing additional benefits to customers. The third concerns denying opportunities to some customers that are available to others. Incorrect decisions made for this third scenario would be more severe than for the first two. For this reason, the fairness assessment risk levels of the AIDA systems change based on the purpose for which its outputs are to be used. **The Assessment Methodology section 4.1 acknowledges that the scaling of a fairness assessment depends on the fairness risk level of the AIDA system.** High risk AIDA systems would usually require a detailed and sophisticated fairness assessment while low risk AIDA systems would likely require less detail.

## 2.6 Use Case illustration – Predictive Underwriting in Life Insurance for a Cross Sell Campaign

### 2.6.1 Use Case Description

To demonstrate the application of the Fairness Assessment Methodology, we considered the scope of PUW AIDA systems to **cross sell life insurance** products to existing customers in the Singapore market. The system aims to predict if an individual existing customer of an insurer is an **eligible risk** for undergoing a simplified underwriting process to obtain a life insurance product. This concerns an active effort by the insurer to reach out to customers pre-identified as eligible, rather than a passive one where the insurer waits for existing customers to request life insurance before checking their eligibility. If the eligible customer accepts the simplified underwriting offer, they receive an expedited approval (simplified underwriting process with no price loading) and do not have to go through the full underwriting process. The ineligible customer may be able to receive the same insurance product if they contact the insurer and go through the full underwriting journey.

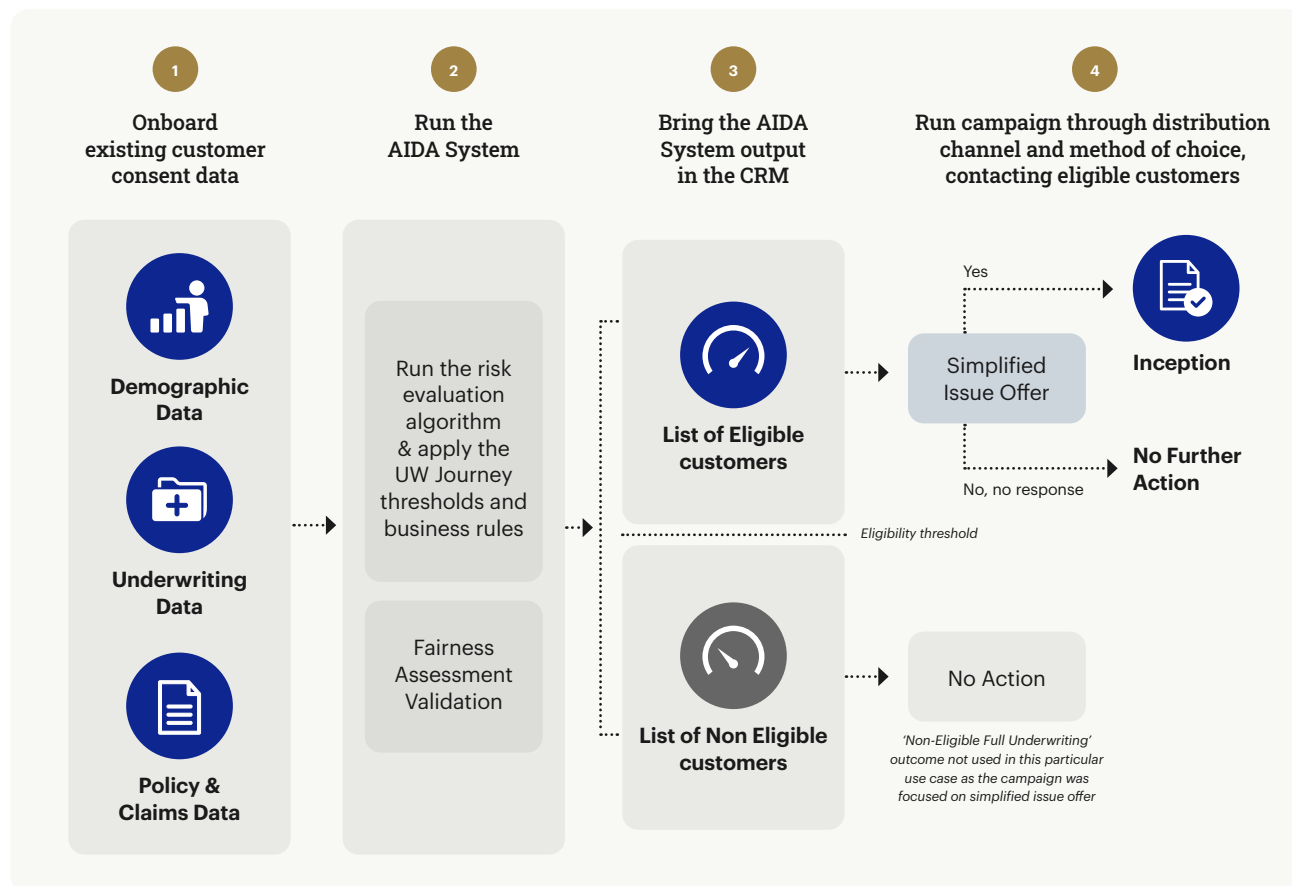


Figure 2.6 - PUW AIDA System Live and embedded in business process

## 2.6.2 Insurer Description

We have completed this use case for a hypothetical insurer using a synthetically generated dataset to preserve data privacy (see section 2.6.4 for more detail) the intention of which is to give FSIs, and particularly insurers, a worked example of a populated fairness assessment document.

**The company “Insurer A”** is a hypothetical Singapore insurer that is conducting a fairness assessment on its PUW AIDA system. Insurer A has defined fairness standards for its organisation, taking into account its organisational principles and values as well as all relevant regulation, and these standards have been approved through its internal governance processes. The fairness standards cover the areas of identifying at risk groups and harms, personal attributes, proxies, explicit fairness objectives and associated measures and thresholds. They are used consistently for AIDA systems across the organisation.

**Note:** It is up to the FSI to determine which standards are best for its organisation considering market practices, sociocultural views, organisational values, regulatory guidance and the legal requirements in its jurisdiction on related concepts that could impact fairness – e.g., personal data, privacy, security, anti-trust.

When it comes to the acceptable deviation threshold from fairness criteria, there is currently little regulation around what these thresholds should be. Insurer A selected a 20% deviation threshold from the ideal fairness metric value. If breached, this threshold would trigger investigative action. The threshold is based on the Four-Fifths Rule, an approach that originates in the US Equal Employment Opportunity Commission’s Uniform Guidelines for Employee Selection Procedures (see Appendix 8.3 in the Methodology whitepaper for more information).

## 2.6.3 AIDA System Fairness Risk Tier and Assessment

FSIs should customise the depth of the fairness assessment to be proportionate to the fairness risk level they determine for the AIDA system under consideration. Such customisation could apply to the fairness assessment process itself, the level of detail of the assessment, or the parties involved.

### 2.6.3.1 Determining the Fairness Risk Tier of the AIDA System

It is up to the FSI to determine the process to assess the risk level of an AIDA system for FEAT, and fairness specifically. Insurer A used its existing standard AIDA risk tiering process, updated to make it more relevant to apply to the Fairness Principles (see section 4.1.1 in the Fairness whitepaper).

Insurer A **assessed the PUW Insurance AIDA System for the Singapore market to be at the medium low risk tier for Fairness using a two-step process:**

Step 1: The PUW Insurance AIDA System was assessed as falling in the **medium risk** tier by Insurer A’s existing standard AIDA Risk Tiering Framework that considers the following:

- a. Extent of automation in the AIDA driven decision making process.
- b. Complexity of the AIDA model.
- c. Severity and probability of impact on different stakeholders, including the individuals affected.
- d. Monetary and financial impact.
- e. Regulatory impact.
- f. Reputational risk.
- g. Use of personal data.

Step 2: The Fairness Specific Risk was then assessed in more detail. This was done by answering questions relating to factor C in the above with a focus on individuals. This was assessed to be **Low Risk**.

**C1: Impact of AIDA system decisions on affected individuals - low**

- Low probability and severity of potential harms to these individuals – the product is offered only to a selected group of people who have been identified for this campaign. If the individual is not part of the campaign, he or she would still have access to the product but would need to go through another available underwriting journey, where the price of the same product may or may not be the same.

In Insurer A's overall assessment process, when C1 is scored as a low risk, this is accepted as the outcome of the Fairness Specific Risk. However, when the answer to C1 is "medium" or "high", the following questions C2 and C3 are also considered as they can mitigate or increase the risk [conversely, if the answer to C1 is "low" this remains the case irrespective of the answers to questions C2 and C3]. The replied to C2 and C3 are given here for illustrative purposes:

**C2: The number and type of individuals who could be affected by system outcomes - medium.**

- All are current insurance customers of the insurance company. The system is planned to be used on an ongoing basis (i.e., it is not just a one-off marketing campaign). This AIDA system is being developed inhouse, so there is a low risk it would become industry standard and used on individuals outside of Insurer A's portfolio.

**C3: Options for recourse – high.**

- Because the PUW is being used only for a specific list of targeted individuals as part of a marketing campaign there is no option for recourse.

Insurer A then combines the outcomes of Steps 1 and 2 to get the Fairness Risk Tier. If they assessed the same risk (e.g., medium-medium) then that is the risk tier level. If they are different by one level, then Insurer A combines the results, leading with the higher risk (in this case "medium-low"). If different by two levels, then the standard AIDA risk tier assessment (outcome of Step 1) is brought one level closer to the 'Fairness Specific Risk' (outcome of step 2) and they are combined in the same way as above. So if Standard AIDA Risk Tier assessment is Low, and Fairness Specific Risk is 'High', then the Fairness Risk Tier is 'High-Medium'.

**Note:** Different insurers could classify the same AIDA system use case at a different fairness risk level, as there is no standard way to assess. If a similar AIDA system is used for other purposes like the risk assessment of new customers or to auto-decline customers (as discussed in section 2.5.3 above), then the associated fairness risk level needs to be assessed and may result in being higher (as per the FSI's model governance) and requiring a more detailed fairness assessment (as outlined in section 4.1 of the Methodology document).

### 2.6.3.2 Determining the Application of the Fairness Assessment Methodology, Based on the Fairness Risk Tier of the AIDA System

Based on the Fairness Risk Tier of medium-low, Insurer A's fairness standards recommend

- Answering all assessment questions with short/summary answers and standard level depth of analysis.
- Applying its standard trade-offs between applicable commercial and fairness objectives.
- Applying its standard required monitoring regime for fairness.
- Applying its standard review and escalation process.

Examples are highlighted during the below assessment where a different approach would be required for AIDA systems with higher Fairness Risk Tier going by Insurer A's Fairness Standards.

**Note:** The above is an example of how an FSI might determine the application of the Fairness Assessment Methodology – it is up to the FSI to determine the method that is best for them. Based on materiality - in the case being the level of Fairness Risk of the AIDA System, the FSI can decide to answer all or a subset of the questions.

### 2.6.4 Data Sources and AIDA System Description

For the purpose of showcasing how the process outlined in the Methodology paper can be followed in practice, a real insurance model development dataset was used to produce synthetic data – data that is close enough to real data to meaningfully illustrate the application of the methodology, but different enough to preserve privacy. The synthetic data generation was conducted using generative network methods including generative adversarial networks (GANs) and variational auto-encoders (VAEs), optimising to meet KPIs for similarity, utility and privacy.

This Synthetic dataset has been used throughout the use case described below. More than 20,000 records were used to develop a simplified predictive underwriting model, on which the fairness Methodology was applied. As per common practices, a manual risk labelling process was used to label records as 1 and 0 representing an eligible and non-eligible risk case respectively. The ratio of eligible to non-eligible risk cases in the dataset c. 5:1.

Key specifications of the modelling process and outcomes are:

- **Labels.** Risk labels were created using a set of rules that consider the severity and frequency of past claims and the most recent underwriting decisions. Only the most confident cases (records) were labelled by the underwriters (several rounds and several reviewers were used to limit the impact of a single underwriter's bias).
- **Key features.** Demographics, claim history and underwriting data were among the most important features.

- **Model(s) approach.** We built a logistic regression binary classification model to predict the eligible/ineligible risk labels from the training data with a confidence score. We then fine-tuned the model and underwriters decide on a threshold to make the final classification decision (simplified offer or full underwriting). The threshold(s) depended on model performance metrics on training and testing data and on the risk appetite of both reinsurance and insurance parties.
- **Business rules overlay.** Certain business rules, agreed with underwriters and actuaries, were applied (see section 2.6.5.3) and where appropriate overruled the model's predictions to further control the risk and business context. For example, depending on the product, there are age or claims restricted offers and these individuals would need to go through the full underwriting process.

Manual overrides: there were no manual overrides as the business rules in place are sufficient to not offer simplified underwriting to the non-target group.

## 2.7 Fairness Assessment Using the Methodology

This section demonstrates how the fairness assessment is conducted using the Fairness Assessment Methodology mapped to a typical AIDA System Development Lifecycle Steps 1-4, and applying it to the specific use case of PUW for cross sell in Singapore. The 18 questions from Veritas Phase 1 have been mapped to four generic FEAT checklist questions (relevant to all of FEAT) and 12 FEAT fairness-specific checklist questions. A reference to the equivalent question label in Phase 1 can be seen under each question in brackets as shown in figure 2.1.

The answer to Step 0 Checklist question F0 is a yes - Insurer A has defined standard, robust processes for (a) identifying at risk groups (b) identifying personal attributes and potential proxies (c) identifying explicit fairness objectives and associated measures and thresholds and (d) algorithmic methods to identify and reduce unfairness. They are referred to in Steps 1-4 below.

The code to run some of this analysis can be found in the following GitHub repo (<https://github.com/veritas-project/phase2>). All key analysis documented in this use case document, other than Feature Importance and Phik Correlation, can also be performed using the Veritas Toolkit (<https://github.com/veritas-toolkit/diagnosis-tool>).

### 2.7.1 Step 1: Define System Context and Design

#### 2.7.1.1 Part A: Define System Objectives and Context



**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?**

[This question refers to question A1 in the Phase 1 methodology]

#### **Primary commercial objective:**

**Reduce cost of underwriting an existing customer's eligible risk for a life insurance product while maintaining portfolio risk levels.**

### **Quantitative targets:**

- 1. Reduce cost of underwriting = increase the simplified underwriting** of new policies per annum from 0 to 50% of eligible risk new life applications.
  - i. This would require the campaign offers approximately 50% of the current health insurance customer base life insurance (i.e., this 50% would need to be flagged as eligible risk).
  - ii. Given the level of exclusions from the model development dataset e.g.:
    - a. Customers that labelling classified as 'uncertain,' (i.e., not eligible or ineligible risk) and
    - b. The business rule exclusions that will be applied as part of the AIDA system.This would require the model development dataset before these exclusions to have approximately 80% flagged as eligible risk. Insurer A assessed that this would require a balanced accuracy of over 82%.
- 3. Maintaining the portfolio profit levels constraint = Insurer A assessed it could afford a maximum of 1 false positive for 24 true positives, which is a minimum precision of 96% for the new business underwritten by the campaign.** The calculation of precision could be post business rules, but as so few of the business rule exclusions had an "eligible risk" flag by the model, and the precision calculation is done on the model outcomes.

**Note:** the requirement for balanced accuracy of 82% to flag 80% eligible is arrived at following Insurer A's internally defined standards. Each FSI will have their own standards for this, based on their risk appetite, etc.

### **The two main secondary commercial objectives are to:**

- Improve and speed up customer journey for life insurance onboarding with the objective of increasing existing customer satisfaction and retention.
- To increase the size of the life insurance portfolio without increasing risk levels, thus increasing profit.

Meeting the primary commercial objective also enables meeting the secondary objectives.

**AIDA is used to achieve these commercial objectives by using predictive underwriting** to predict eligibility risk of individuals for a new life insurance product, which is derived from the data Insurer A has on customers who already hold a life and health insurance policy with them.

**Note:** Insurer A had secured consent from existing customers to use relevant data for the purpose of cross selling before including in the model development dataset, in accordance with Singapore laws and regulations.

**The main risk for Insurer A is a financial loss due to higher than expected claims if there is a poor estimation of the risk profile of eligible individuals.** A key external constraint that could lead to a poor estimation of risk is the potential possibility of an incomplete picture for medical claims for example due to individuals holding medical policies with other insurers or using other ways to cover medical expenses.

Insurer A extracted the above from the development document it created as part of standard (non-FEAT) AIDA development process.

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?**

[This question refers to question A2 in the Phase 1 methodology]

Reminder: this is synthetic data from a hypothetical FSI and the bias described below is created for illustrative purposes only. Other potential bias could have been created; our scenarios are not exhaustive.

**Insurer A identified individuals in ethnicities with low representation** (i.e., non-Chinese) to be at risk of systematic disadvantage. The main reason is that the relatively small size of the ethnic minority population in Singapore increases the risk that the AIDA system does not learn the behaviour of this population as accurately as for the majority population.

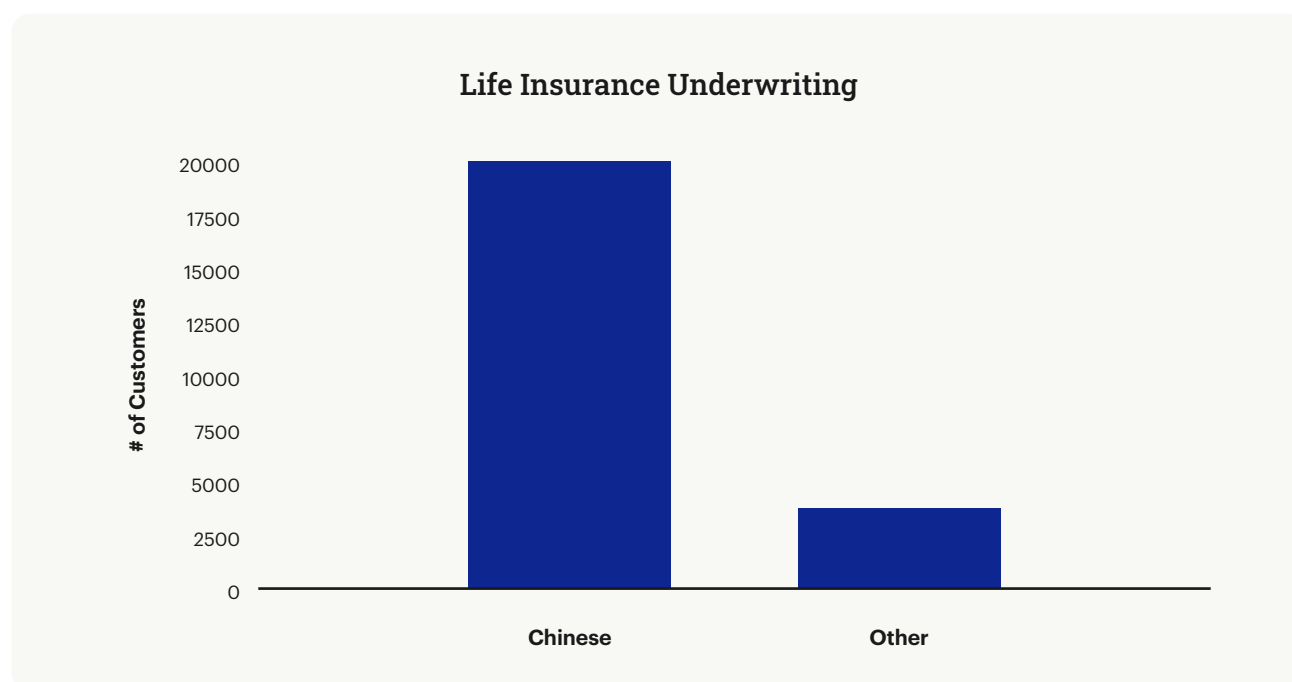


Figure 2.7: Ethnicity representation for Insurer A data

**Insurer A also chose to assess gender for fairness.** While not a protected variable in Singapore, using gender as a factor in decision making is prohibited in some jurisdictions, such as the European Union. Although Insurer A didn't consider any specific gender to be at risk of systematic disadvantage by the AIDA system, it wanted to validate this assumption. This is because the insurer plans to use gender as a factor in the predictions of its AIDA model, justified as it is material to prediction of the model target (see Part D Justify Use of Personal Variables for details).

Assessment was also done on the intersection between the two variables to check for compound disadvantage.

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?

[This question refers to question A3 in the Phase 1 methodology]

In line with Insurer A's fairness standards, a multidisciplinary group of data scientists, compliance and risk experts, business experts, lawyers, actuaries, underwriters and responsible AI experts participated in a workshop to determine the potential harms and benefits created by the system's operations (to meet commercial objectives) for the individuals it applies to. It did this by populating the confusion matrix below:

Confusion Matrix		
	Eligible Risk (Actual)	Non-Eligible Risk (Actual)
Insurance cover through PUW (predicted)	<p><b>Correct classification of eligible risk:</b> insurer offers a good-risk customer auto approved life insurance</p> <p><b>Benefits:</b> Convenient/cheaper insurance</p>	<p><b>Mis-classification of non-eligible risk:</b> insurer offers a bad-risk customer auto approved life insurance</p> <p><b>Harms:</b> Covered by surcharge; it impacts others in the group - LT harm</p> <p><b>Benefits:</b> Insurance at price lower than their risk warrants</p>
Insurance cover through full underwriting (predicted)	<p><b>Mis-classification of eligible risk:</b> insurer does not offer good-risk customer auto approved life insurance</p> <p><b>Harms:</b> Higher potential cost if obtain insurance from another campaign</p>	<p><b>Correct classification of non-eligible risk:</b> insurer does not offer bad-risk customer auto approved life insurance</p>

Figure 2.8.1: Confusion matrix for harms and benefits for individuals and groups.

**The potential harm that is most relevant to the at risk groups identified in F1 is the potentially higher cost of insurance for them if they are incorrectly assessed as ineligible by the system.** Another harm identified was of a surcharge on all insured to cover the cost of those that were incorrectly assessed as eligible by the system. However, this is a longer term potential harm and applicable to the whole population, not a systematic disadvantage to one particular group, and therefore was not the focus of this assessment.

**The potential benefits that are particularly relevant to groups identified in F1 are cheaper insurance and the convenience of simplified underwriting to reduce mortality protection gap.**

**Note:** As this is a medium-low risk use case with the probability of severity of potential harms to individuals being low, Insurer A's fairness standards do not require focus groups or customer surveys to obtain customers' views on the proposed offering and its benefits and risks. However, for higher risk use cases, Insurer A's standards recommend this to be done. A different insurer or Insurer A in a different jurisdiction may have different outcome for a similar scenario.

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?

[This question refers to question A4 in the Phase 1 methodology]

In line with Insurer A's fairness standards, the same multidisciplinary group as referred to in F2 used a fairness decision tree to determine for groups identified as most at risk in F1, and the associated harms from F2, the most important fairness objective and, aligned to that, the relevant measurable fairness metric.

**Insurer A determined the fairness objective to be:** For the eligible population (those that should get simplified underwriting) the distribution of errors (those that aren't offered simplified underwriting) does not differ by over 20% among subgroups.

**And the relevant measurable fairness metric for this objective is:** false negative rate (FNR) ratio.

Refer to the diagram below and the documented steps under it for the route down the fairness decision tree which led Insurer A to its system fairness objective and metric. This tree was adapted from Aequitas Toolkit for AI Bias by Insurer A.<sup>4</sup>

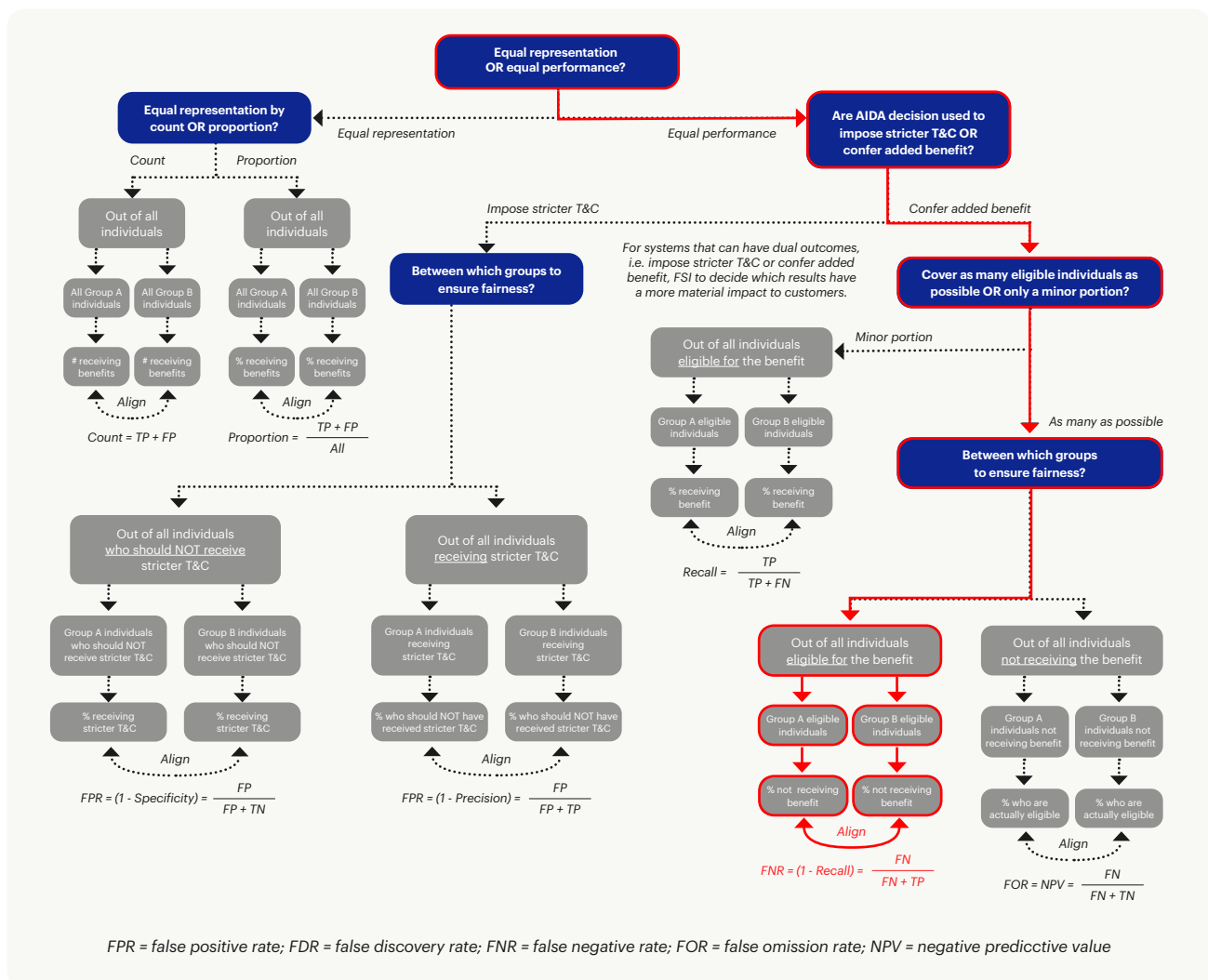
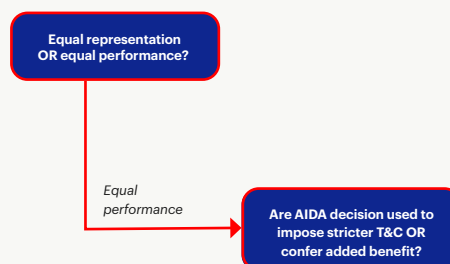


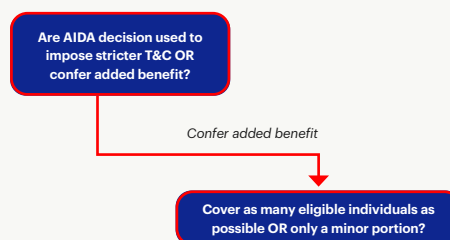
Figure 2.8.2: Insurer A's route down the fairness decision tree

For each decision point in the tree, there is a bullet below to explain why the decision was made:

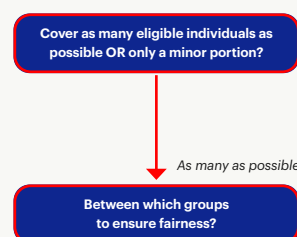
- Decision 1:** Equal representation or equal performance: Chose “equal performance”: For this use case, private insurers typically target equality in risk based performance across groups, rather than the alternative “equal representation,” which targets equality in count or proportion across groups and is not risk focused.



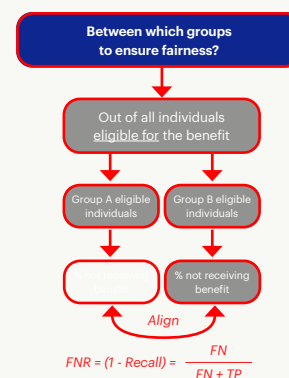
- Decision 2:** Are AIDA decisions used to impose stricter terms and conditions (“T&Cs”) OR confer added benefit: Chose “confer added benefit”: This use case confers the added benefit of simplified underwriting to customers, versus the regular full underwriting process. It does not impose stricter T&Cs, the other alternative.



- Decision 3:** Cover as many eligible individuals as possible OR only a minor portion. Chose “as many as possible”: For this use case, there is no constraint in terms of effort or cost which would restrict the number of eligible individuals that can be covered. Those covered actually reduce effort and cost because they do not have to be manually underwritten.



- Decision 4:** Between which groups to ensure fairness? Chose eligible individuals: For this use case, the fairness objective for Insurer A is to ensure that, for the eligible population (those that should get simplified underwriting) the distribution of errors (those that don’t get offered simplified underwriting) is balanced among subgroups of the personal attribute being assessed.



**The relevant measurable fairness metric in this case, as can be seen in the decision tree above, is false negative rate** – the objective being that Insurer A keeps outcomes similar (i.e., within the thresholds mandated by its fairness standards document for this type of decision and use case), for ethnicity and gender subgroups.

**Note:** The above is an example method to identify the relevant fairness objective and metric – it is up to the FSI to determine the method that is best for them.

## 2.7.2 Step 2: Prepare Input Data

### 2.7.2.1 Examine Data for Unintended Bias

F4

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

[This question refers to question B1 in the Phase 1 methodology]

#### Dataset summary:

The synthetic dataset Insurer A used for model development consists of 23,124 records/observation and 21 attributes. The synthetic dataset does not include personal data (i.e., information that would allow individuals to be identified). The synthetic dataset includes the following attributes about the customer:

Attribute	Description
BMI	BMI
Age	Current age
Tenure	Years as a customer
Gender	Gender
Race	Ethnicity of customer
Marital Status	Marital status of policy holder
Nationality	Primary nationality of the policyholder
Postcode*	Only first 2 digits which describes which district in Singapore the policyholder lives
Smoking	Does the policy holder smoke - Yes/No
Annual Premium	Annual premium
Previous pay-out amount	Pay-out amount over previous period
Number of new policies past period	Number of new policies past period
Number of life policies	Number of life policies
Number of personal accident policies	Number of personal accident policies
Number of single premium policies	Number of single premium policies
Number of exclusions	Number of exclusions for which the insurer does not provide coverage
Purchase recency	Latest recent purchase
Latest purchase distribution channel	Distribution channel of the policyholder
Latest purchase product category	Product category of the policyholder
Policy duration	Duration of the Policy

\*Postcode (first 2 digits) was dropped from the dataset used to develop the model after proxy analysis was done to meet data minimisation guidelines on the basis that it is not predictive and not a proxy for a personal attribute. See Section D on personal attributes.

Table 2.1: Customer attributes

The data attributes used to analyse the groups identified as at risk of being systematically disadvantaged in F1 are:

- Race attribute for ethnicity at risk group.
- Gender attribute for gender at risk group.

See answer to question F1 above for the reasoning behind choosing these two at risk groups in this hypothetical illustration.

Target variable is tagged as 1 (eligible risk) and 0 (ineligible risk). Distribution of target variable: 19,124 are eligible (82.7%) and the remaining 4,000 are not eligible (17.2%).

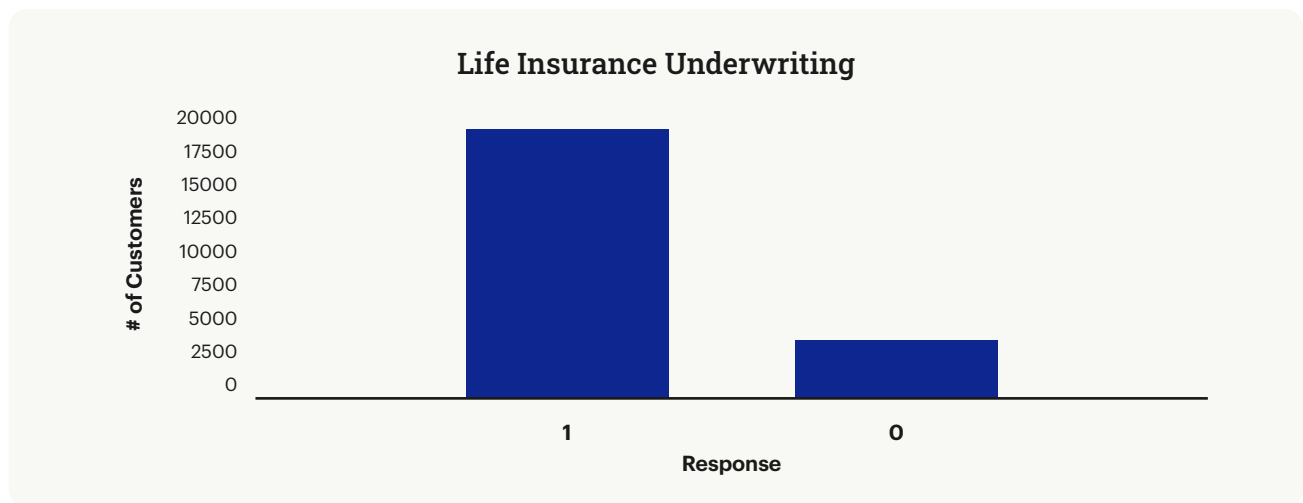


Figure 2.9: Response/target distribution for Insurer A data

### Bias checks:

Insurer A conducted the data bias checks required by its fairness standards: it checked for representation bias, measurement bias and data pre-processing bias and measurable proxy bias on its development dataset. See below for summary outcomes.

#### 1. Checks for representation bias:

*For Gender:*

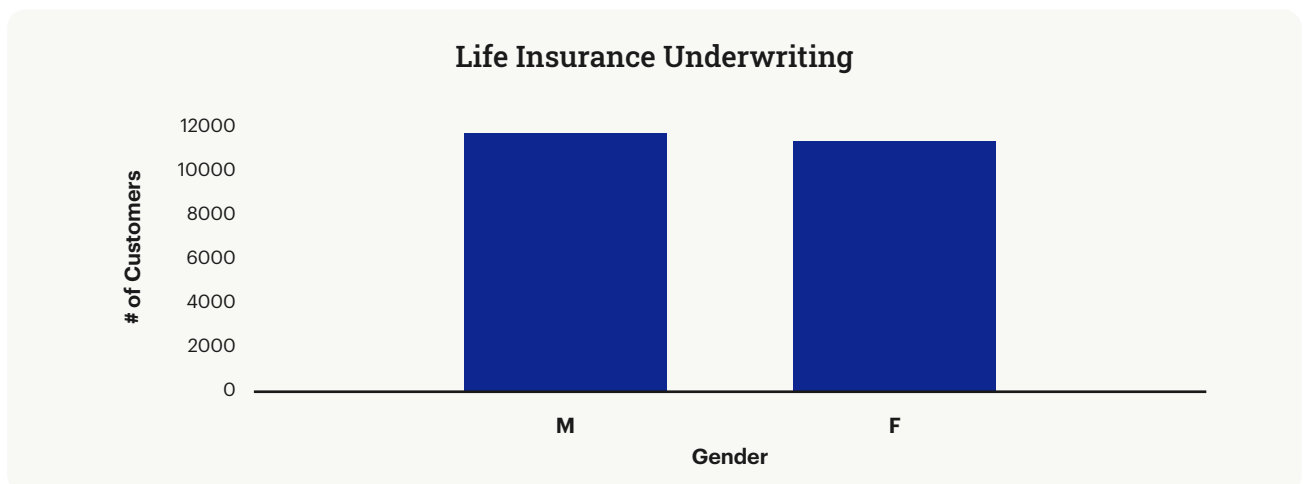


Figure 2.10: Gender distribution for Insurer A data

The risk of representation bias depends on both absolute and relative amounts of training data.

### Representation of classes within the model dataset:

For this dataset, **representation is similar for females compared to males (48.91% of customers are Male, 51.08% are female)** and in relative terms, there is a very small imbalance between the two groups.

In absolute terms, there is a large number of applications for each gender (11,813 and 11,311 female and male customers respectively).

### Representation of classes in model dataset vs. Singapore population:

The representation observed is also in line with the representation for gender in Singapore (gender in 2020 is 957 males per 1000 females)<sup>5</sup>, which is desirable from the point of view that the training/development dataset reflects the model use population with respect to gender.

**Outcome: there are no concerns with respect to representation bias for gender.**

### For Ethnicity:

### Representation of classes in model dataset vs. Singapore population:

For this dataset, representation is considerably larger for the Chinese population compared to the other ethnicities in the dataset: 83.2% of customers are Chinese, and the rest are non-Chinese. This representation is close to the ethnicity representation in Singapore (ethnic groups in Singapore in 2017: 74.3% Chinese, 13.4% Malay, 9% Indian, 3.2% Others)<sup>6</sup> and therefore positive from the point of view stated above which is that the training/development dataset approximately reflects the model use population.

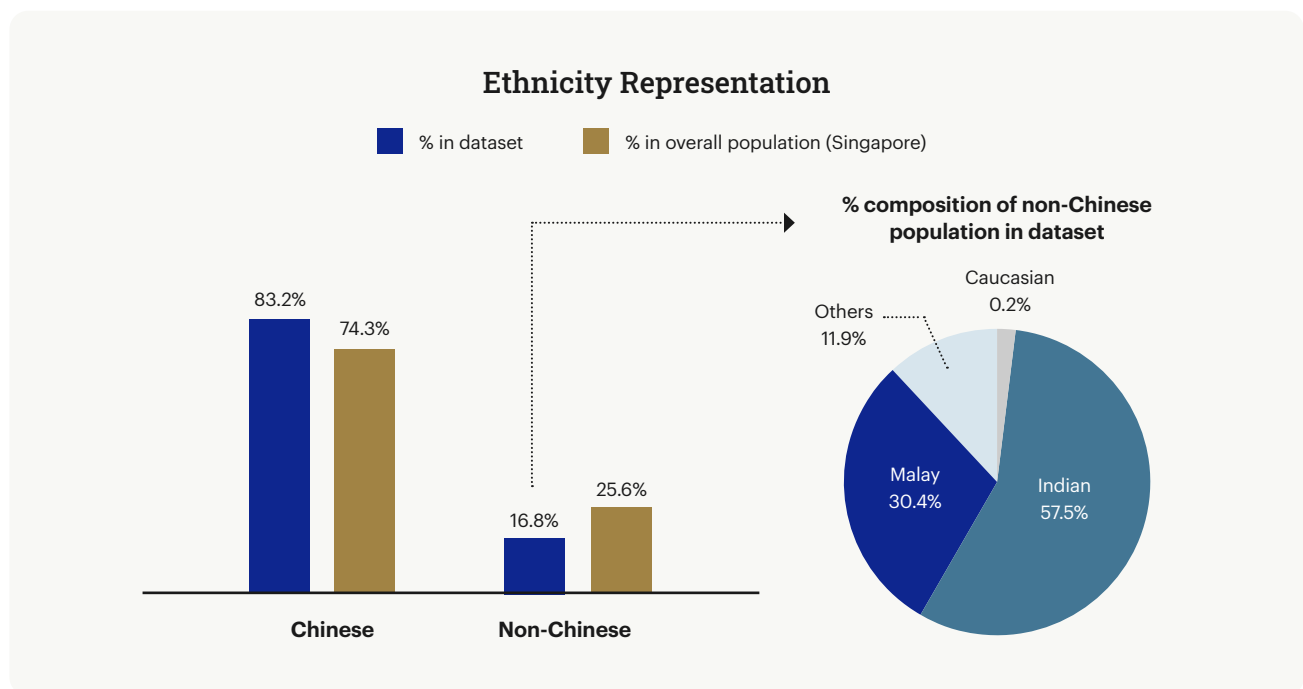


Figure 2.11: Ethnicity distribution for Insurer A data vs. population

## Representation of classes within the model dataset:

When combining non-Chinese groups together, the representation obtained is 83.10% for Chinese and 16.89% for non-Chinese or “other”. In Insurer A’s fairness standards, it states that on a relative basis, less than 50% of imbalance between the groups is considered as relatively low, whereas an imbalance above this threshold is considered high and should be looked at more closely. **The difference in representation between “Chinese” and “other” groups for this dataset is 66%, well above the 50% threshold and therefore pointing to a relative imbalance between Chinese and other ethnicities** (this was expected and was the main reason Insurer A classified this group as potentially at risk in A2). As a result, Insurer A examined the case more closely and in absolute amounts, it determined that there is sufficient training data for both groups as the “other” ethnicity group contains 3,884 customers.

**Outcome: More detailed analysis was conducted on model outcomes (in measurement bias and also question F9) to understand if the ethnicity representation observed might lead to some bias.** If the behaviour of “other” and “Chinese” with respect to the target under study is different, the model will predominantly be trained on the behaviour of the larger group (i.e., Chinese) hence possibly leading to a lower performance of the model for the minority group.

**An item to note:** In Singapore there is no single standard classification for ethnicity and nationality across FSIs at the time of data collection, and the one used by an FSI may also be different to the one used for national statistics. Depending on how the question is asked when the data is collected, the same individual could give different answers. Therefore, comparing the distribution of ethnicity within an insurer to the distribution in the overall population may not be an accurate analysis and should be treated with caution.

## 2. Checks for measurement bias (on human labelling):

Next, Insurer A examined the prevalence rates of eligible risk for both gender and ethnicity. This information can be useful in connection to a known source of bias: measurement bias with respect to the label.

**The prevalence rate** of a given historic outcome in a population is defined as the fraction of that population that experienced the outcome (this is based on true outcome labels rather than predicted values). The prevalence ratio is as the name suggests the ratio of the prevalence in one group of the population over the prevalence of another group. An unbalanced prevalence ratio could indicate bias in the human labelling process, which could subsequently be learned by the model trained on this data (this imbalance would be seen in the disparate impact ratio).

Insurer A’s fairness standards indicate a prevalence ratio of less than 0.8 or more than 1.2 (considering a deviation of +20% from parity in line with the Four-Fifths Rule) should be examined further.

### For Gender:

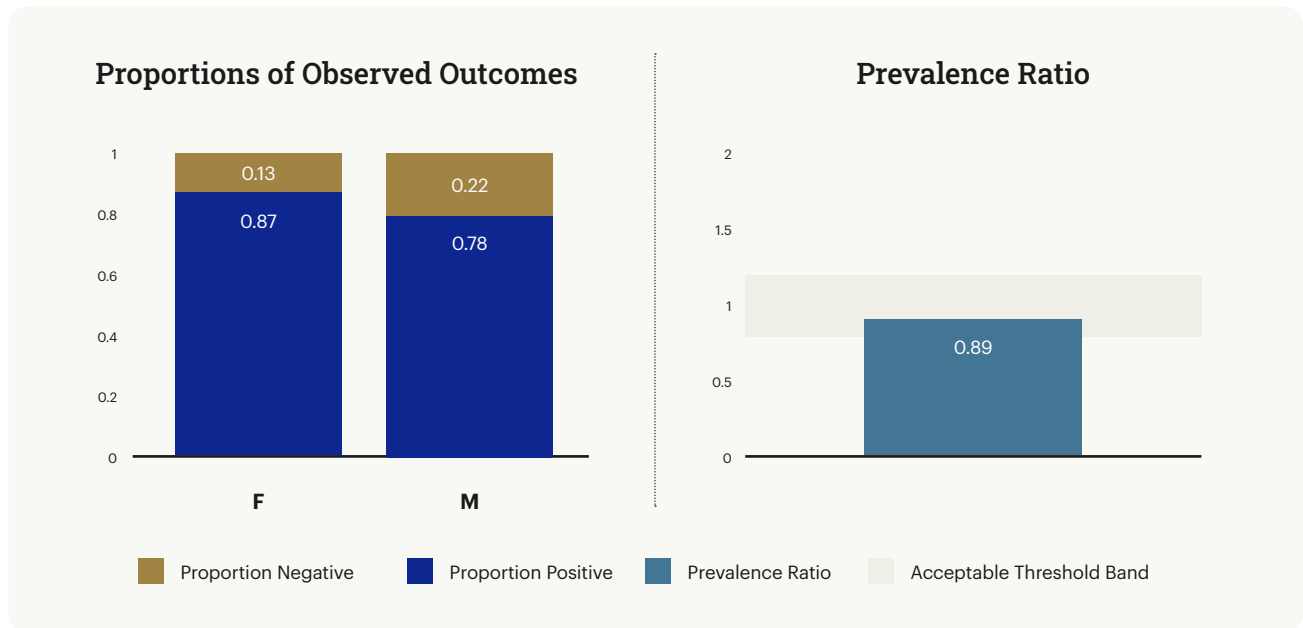


Figure 2.12: Prevalence ratio of eligible risk for gender for Insurer A data

As can be seen in figure 2.12 above, the prevalence rate of eligible risk for female customers is larger than that for male customers. When compared in the form of a ratio (male eligible rate/female eligible rate), the resulting value is 0.89. This value is above the fairness standards threshold of 0.8 and points to no significant difference in the rates of eligible risk outcomes for male customers compared to female customers. The fact that there is such a high level of eligible risk outcomes (which is the positive outcome the model will be aiming to predict) in the development data (i.e., approximately 80%) means that the absolute difference between the male and female rates would need to be large for the ratio to go below the 0.8 threshold. This should be checked again on the live population once the model is deployed and there is ground truth data as lower levels of positive outcomes may impact the ratio.

The level of difference could be explained by the manual labelling process where underwriters only label risks that are clear cut (i.e., either clearly ineligible or clearly eligible). The middle risk customers are not labelled and as a result are not included in the development dataset. It is possible that the middle risk population has more eligible males than females, which would explain the discrepancy, and if included bring the prevalence ratio closer to 1. This should be checked on the overall live population (which will include middle risk customers) once the AIDA system is deployed and there is sufficient ground truth post deployment (see Section E).

**Outcome: As the prevalence ratio is within the acceptable range no immediate concerns exist** with respect to measurement bias on the label for gender in the development dataset.

As mentioned, the difference in prevalence rates will be monitored when deployed and sufficient ground truth has accrued. When the model first goes live and before ground truth is available, the disparate impact ratio should be used as a proxy for prevalence, and should also be checked in Part C of the Fairness Assessment Methodology. The disparate impact is similar to prevalence but is based on a predicted model outcome rather than on true outcome labels.

### For Ethnicity:

Next, Insurer A examined the prevalence rates of eligible risk for ethnicity.

When combining all non-Chinese groups together, the prevalence ratio for eligible risk is 1.01, meaning the two rates are very similar and within the threshold band as defined by Insurer A's fairness standards.

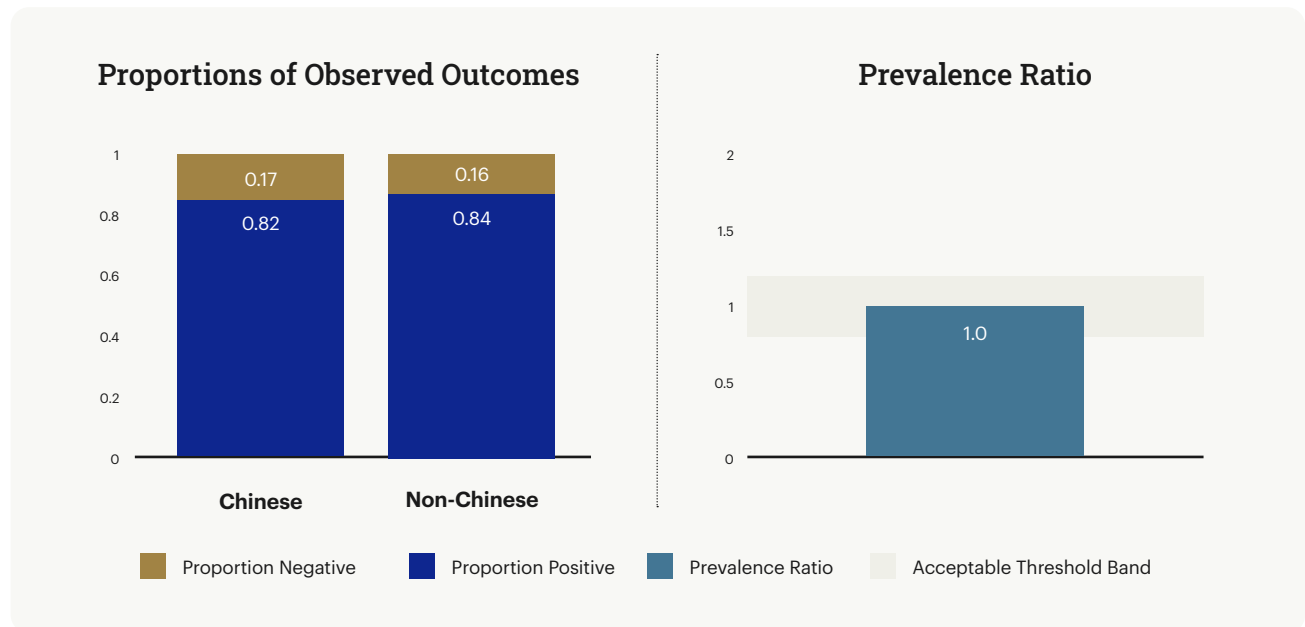


Figure 2.13: Prevalence ratio for ethnicity for Insurer A data

**Outcome: there are no concerns with respect to measurement bias in the label for this dataset with respect to ethnicity.**

Checking for measurement bias in the those labelled ineligible risks by human labellers in monitoring.

As the full population in scope for this AIDA system are existing customers, it is possible to check for bias in those labelled as ineligible risks by human labellers. As part of monitoring post deployment, Insurer A will undertake an exercise sometime around the first full year of deployment to reassess customers' risk status and model outcomes to check again for bias across gender and ethnicity.

### 3. Checks for data pre-processing bias:

Data pre-processing was not performed on either gender or ethnicity attributes (i.e., no data cleansing, missing data imputation or other similar processes were carried out). Therefore, there was no change in the distribution of gender and ethnicity so no need to check for pre-processing bias on the distribution of these attributes.

#### 4. Checks for measurable proxy bias:

There were no measurable proxies for unavailable data used in the model development dataset, and so no reason to check for this type of bias.

**Note: other potential sources of bias that should be evaluated where the data allows.**

- Bias due to data collection options. Some people do not identify as male or female but instead as a non-binary gender. When Insurer A collects information on gender, it forces a binary choice, only giving the options “male” or “female.” This in itself might create additional forms of bias and has not yet been addressed due to a lack of data. A recommended mitigation is to start collecting this data according to the applicable data protection and privacy laws.
- Historic decision biases. Risk evaluation is changing with medical advancements. An example would be customers with diabetes seeking mortality cover in life insurance. Insurer A is now willing to consider such customers eligible risks compared to in the past when diabetes was generally a more serious condition. However, the data used to train the model may still contain outdated bias against people with diabetes. The data should be assessed for this type of bias where that level of granularity exists, both at an overall population level and for at risk subgroups. This level of granularity was not available in the synthetic dataset generated for this use case.

F5

#### Have you documented how are these impacts being mitigated?

[This question refers to question B2 in the Phase 1 methodology]

Below summarises the mitigation actions taken on the data due to identified bias/errors in the data.

##### Representation bias:

- **Gender. No action was taken with respect to gender** as the representation observed for female and male customers can be considered as balanced.
- **Ethnicity.** A relatively large imbalance was observed in the dataset. However, representation bias requires being addressed only if, in addition to the representation bias, there is indication of different behaviours for different groups with respect to the outcome of interest – which is measured by seeing if the accuracy rates, as well as the error rates, are similar across the groups. If the behaviour is the same, there is no need to adjust for representation bias. A more detailed analysis was completed on this (see question F9) and the behaviour was assessed to be sufficiently similar, and **therefore no mitigating action was taken to change the distribution of ethnicity in the model development data.**

#### Measurement bias (label):

- **Gender: No mitigation action was taken with respect to gender as the differences observed were within the threshold.** Insurer A will be monitoring post deployment for evidence of potential measurement bias in the labelling. In the case that the results obtained would point to differences that cannot be justified or explained, a qualitative analysis of the labels might be recommended to randomly audit a sample of customers and review the accuracy of the label/risk for each of the groups (i.e., female and male).
- **Ethnicity: No mitigation action.** The results obtained are near parity (1) and therefore do not raise concerns with respect to measurement bias for the label, therefore no mitigation action was required.

**Summary outcome for F5: it was assessed that no mitigation was required for the personal attributes of gender or ethnicity.**

For use cases where data bias mitigation is required, see section 3.4 in the fairness methodology document for potential approaches.

**Note:** FSIs should look to minimise unintentional bias by incorporating reasonable steps in the AIDA system development lifecycle to identify and address potential sources of both data and human (cognitive) bias. See Fairness methodology whitepaper section 2.2.5 for more detail.



## 2.7.2.2 Part D: Justify the Use of Personal Attributes

F6

**Have you determined and documented personal attributes are used as part of the operation or fairness assessment of the system?**

[This question refers to question D1 in the Phase 1 methodology]

Insurer A is following data minimisation and proportionality principles to validate the data to be used in order to protect the privacy of individuals. There was no personal data in the development dataset, but if there had been, the synthetic data generation method we used would also have allowed the development to have been conducted without using real personal data.

Insurer A followed the standard process, as defined in its internal fairness standards, for determining personal attributes – this is outlined in answering F7. Personal attributes are defined as features about individuals that should not be used as the basis for decisions without reasonable justification. Personal attributes are defined by FSIs in the context of each specific use case, and at a minimum covers any personal data that may be included in the AIDA system as defined in relevant data protection and anti-discrimination laws, and can also include non-personal data.. The personal attributes identified are:

Data Type	Nos.	Field	Description	Personal Attribute Group 1	Personal Attribute Group 2
Demographic	1	Gender	Gender		Y
	2	Age	Current age		Y
	3	Marital status	Last recorded marital status of customer	Y	
	5	BMI	BMI based on last recorded height/weight		Y
	6	Ethnicity	Ethnicity	Y	
	7	Nationality	Last recorded nationality of customer	Y	
	8	Previous Payout Amount	# new policies last 3 years		Y

Table 2.2: Personal Attributes for Insurer A's model development dataset

Group 1 attributes are not to be used in decision making and will be checked and monitored for fairness. Group 2 attributes are justified for use in decision making and can be used in the model. See answer to question F8 for the justification.

There are many additional personal attributes not available in the data that are likely to be beneficial for use in decision making (historic serious medical conditions, for example). The model developed from the existing data met the accuracy and precision requirements for commercial objectives, so there was no immediate need for additional personal attributes.

F7

## 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 question refers to question D2 in the Phase 1 methodology]

Insurer A followed the standard process for identifying personal attributes as defined in its internal fairness standards. The same multidisciplinary group as referred to in F2 attended a workshop to review each of the data attributes included in the development dataset as directed in the standard process to determine if they are personal attributes or proxies and if so, if they are Group 1 or Group 2. A high level diagram and decision tree of this process is shown below.

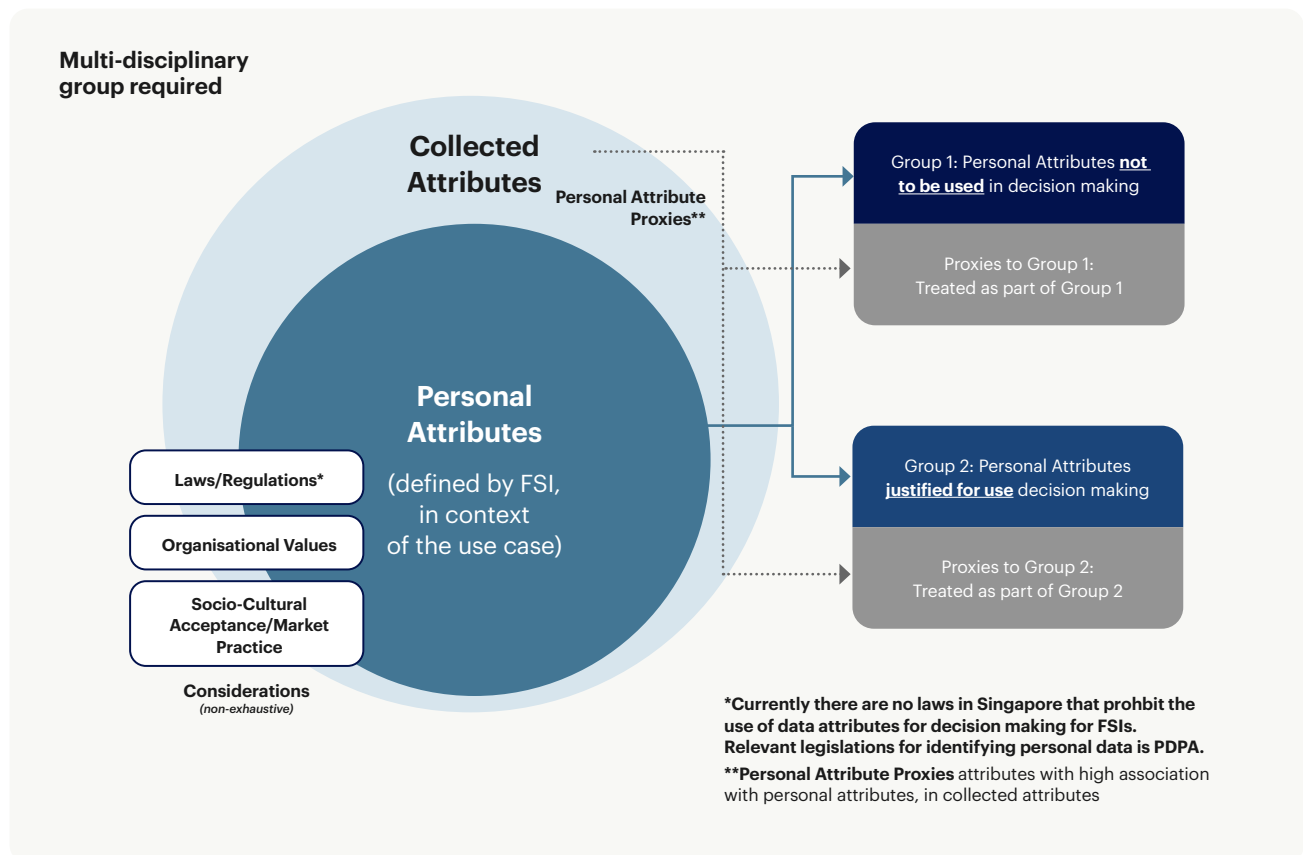


Figure 2.14: Identifying Personal Attributes

## Personal Attribute Identification and Classification Decision Tree (Jurisdiction Use Case Specific)

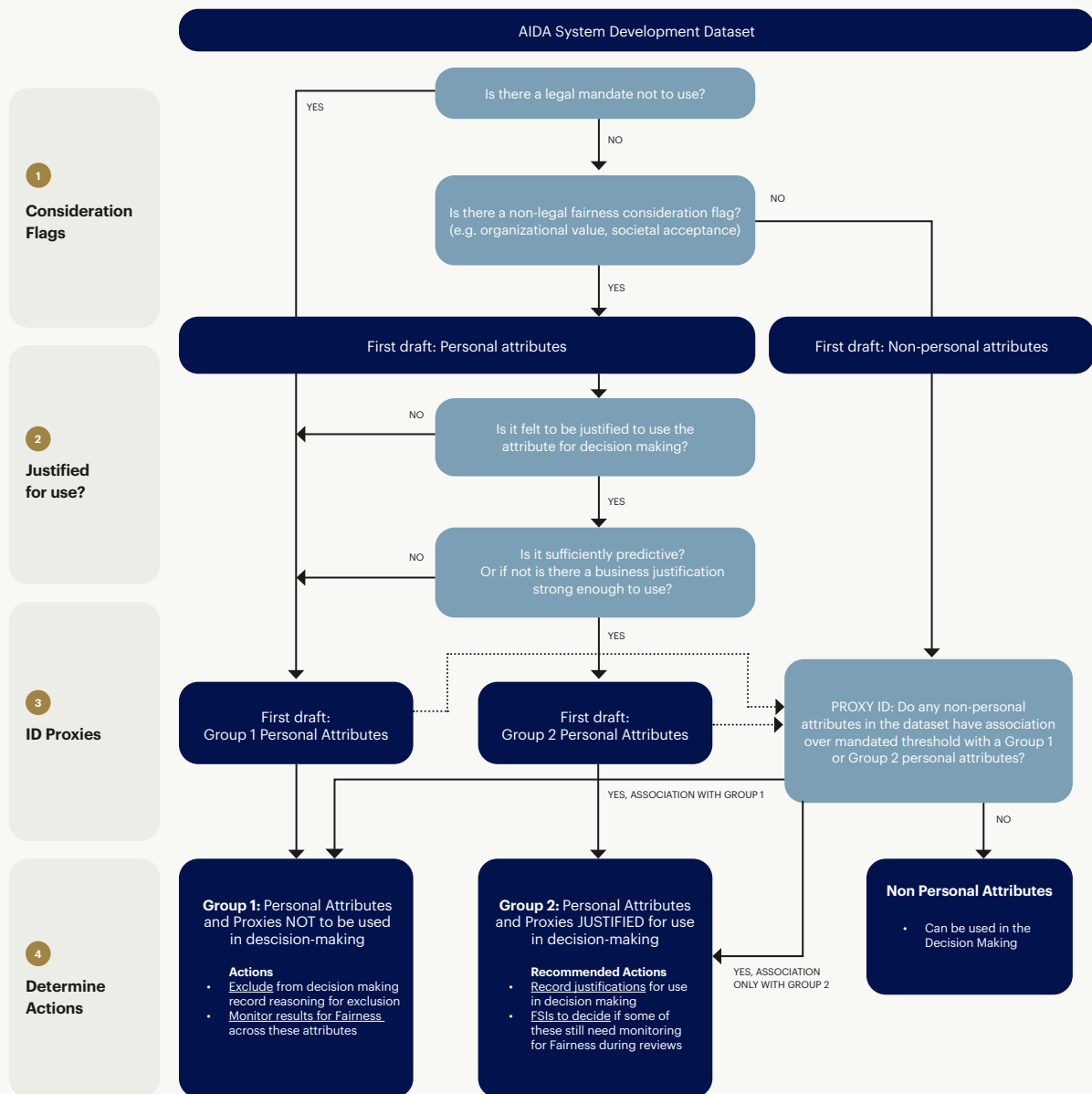


Figure 2.15: Personal attributes identification and classification decision tree

The full list of data fields in the model development dataset, and the information used to make the classification decisions, along with the final outcome, is in the table below.

The data-fields have no personally identifiable information about an individual (no names, IDs, etc.). Therefore, no personal data is involved.

## Mapping of Data Attributes to Personal Attributes/Proxies

Data Type	No.	Data Attribute	Description	1 Consideration Flags			Personal Attribute?	2 Justified for use in decisioning		3 Proxies	Personal Attribute Group		4 Actions	
				Laws/Regulations	Societal (Un) Acceptance/Market Practice	Organisational Values		Justification is Known Risk Factor for Target	Predictive of target in development data?		NOT TO BE USED IN DECISION MAKING	JUSTIFIED FOR USE IN DECISION MAKING	Assess for Fairness?	Data Attributes that can be used in the model
demographics	1.	gender	gender	N	Y	Y	Y	Y	Y	-	-	Y	Y	Y
	2.	age	current age	N	Y	N	Y	Y	Y	-	-	Y	N	Y
	3.	marital status	last recorded marital status of customer	N	Y	Y	Y	N	N	-	Y	-	Y	N
	4.	postcode	first 2 digits of last recorded postcode of customer	N	N	N	N	N	N	N	-	-	N	Y
	5.	BMI	BMI based on last recorded height/weight	N	Y	N	Y	Y	Y	N	-	Y	N	Y
	6.	smoking	last recorded smoking status of customer	N	N	N	N	Y	Y	N	-	-	N	Y
	7.	race	ethnicity	N	Y	Y	Y	N	N	-	Y	-	Y	N
	8.	nationality	last recorded nationality of customer	N	Y	Y	Y	N	N	-	Y	-	Y	N
existing plan(s) info	9.	tenure	years as customer	N	N	N	N	N	Y	N	-	-	N	Y
	10.	number of exclusions	no. of exclusions in product that insurer doesn't provide cover on	N	N	N	N	Y	Y	N	-	-	N	Y
	11.	purchase recency	latest purchase	N	N	N	N	Y	Y	N	-	-	N	Y
	12.	annual premium	last recorded annual premium	N	N	N	N	Y	Y	N	-	-	N	Y
	13.	latest premium dist channel	latest purchase dist channel	N	N	N	N	N	Y	N	-	-	N	Y
	14.	latest purchase product category	latest purchase product category	N	N	N	N	N	Y	N	-	-	N	Y
	15.	previous payout amount	previous payout amount	N	Y	N	Y	Y	Y	-	-	Y	N	Y
	16.	# new policies	number new policies	N	N	N	N	N	Y	N	-	-	N	Y
	17.	# of life policies	number of life policies	N	N	N	N	N	Y	N	-	-	N	Y
	18.	# of single premium policies	number of single premium policies	N	N	N	N	N	Y	N	-	-	N	Y
	19.	# of personal accident policies	number of personal accident policies	N	N	N	N	N	Y	N	-	-	N	Y
	20.	policy duration	duration of policy	N	N	N	N	N	Y	N	-	-	N	Y

■ Personal Attribute Group 1
 ■ Personal Attribute Group 2

Table 2.3: Full list of Insurer A's data attributes with classification to Personal Attribute Group

Gender and ethnicity were selected as the personal attributes for the fairness assessment in this illustrative example. The rationale behind choosing these attributes is outlined in the answer to Question F1.

**Note:** The above is an example method to identify and classify the relevant personal attributes – it is up to the FSI to determine the method that is best for them.

F8

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

[This question refers to question D3 in the Phase 1 methodology]

The data attributes in the development dataset that were determined to be personal attributes are analysed here and an initial analysis conducted on all data attributes to check for association with a personal attribute over a threshold that would classify it as a material proxy for that attribute. This process creates a full list of attributes that Insurer A classifies as a personal attribute or a proxy. For personal attributes in Group 2 and the material proxies for that group, Insurer A then justifies their use in decision making.

### Analysis to identify any material proxies

To determine the level of association of non-personal attributes with personal attributes Insurer A used the Phi\_K correlation matrix, which is shown below. Phi\_K works consistently between categorical, ordinal and interval variables and captures non-linear dependencies. Insurer A uses the threshold of 70% correlation with a personal attribute to signify a material proxy, as defined in Insurer A's fairness standards. Note that the association measure and threshold to identify material proxies should be defined by each FSI and should be approved by relevant internal governance.

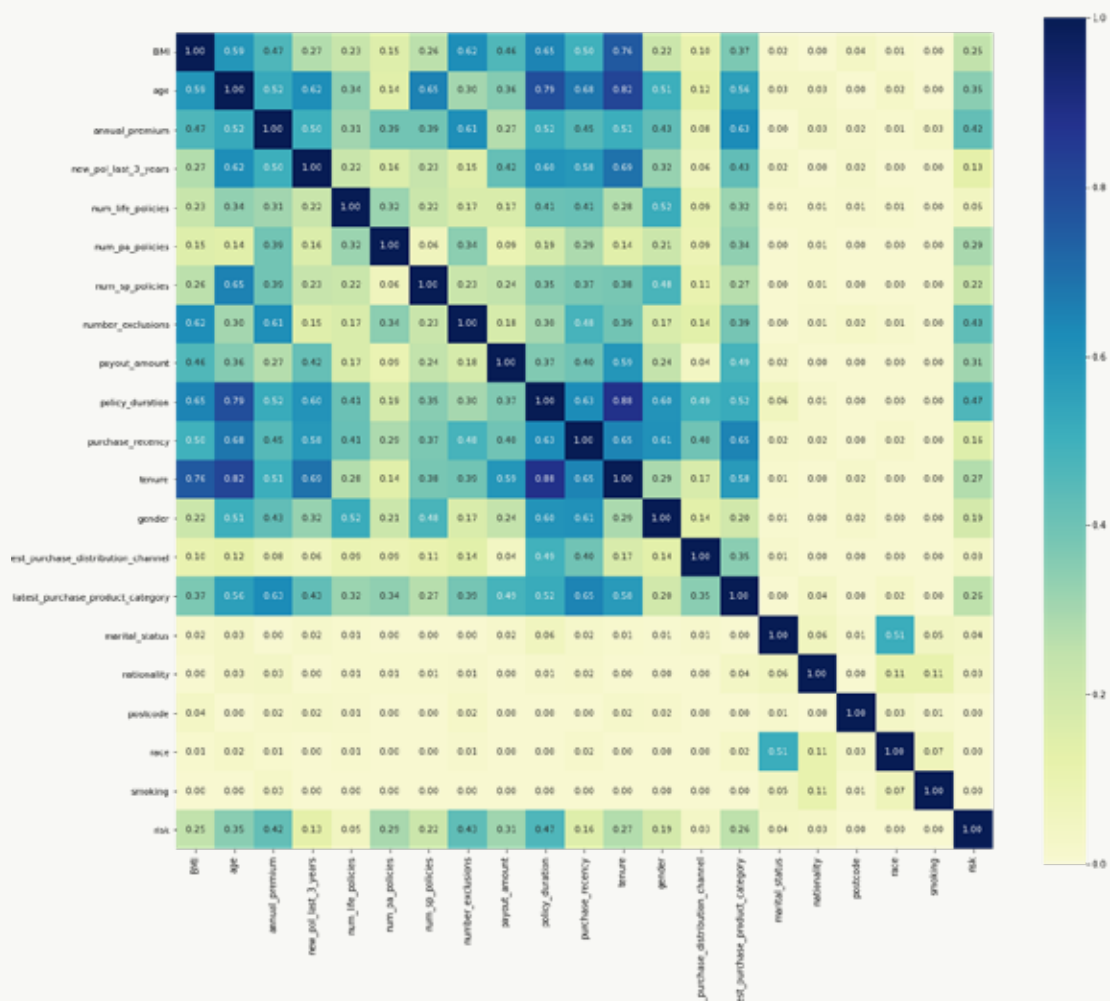


Figure 2.16: Heat map association of non-personal attributes with personal attributes

As can be seen in the matrix, none of the personal attributes classified as Group 1 (i.e., attributes that should not be used in decision making) - ethnicity, nationality and marital status - have a correlation of over 70% with any other variable. Therefore, as per Insurer A's fairness standards, there are no material proxies in the data.

For the personal attributes classified as Group 2 (i.e., attributes that are justified to be used in decision making) - age, gender, Body Mass Index (BMI), previous pay-out amount - there are a number of non-personal attributes that have a correlation of over 70% (e.g., tenure has an over 70% correlation with both BMI and age), but as these personal attributes are justified to be used in decision making, proxies for them are also justified to be used. For higher risk use cases, Insurer A's fairness standards require more in-depth proxy analysis, which includes analysing potential associations between groups of attributes.

Partial postcode data (capturing only the first two digits to ensure against it being used to identify individuals), was included in the initial data set for development. However, it is not expected to have predicting power and is therefore not an attribute included in the model. Nevertheless, Insurer A retained the dataset in its correlation matrix to confirm it was not a material proxy for Group 1 personal attributes, specifically ethnicity, as if it were, the insurer would need to include it in its fairness assessment. As can be seen in the matrix above, the dataset's correlation with other attributes is low across the board and so it is not a material proxy for any personal attribute. As a result, Insurer A removed the postcode attribute from the development data on which it built its model.

**Note:** This document is suggested guidance. It is up to each FSI that decides to apply the FEAT Framework, and specifically the Fairness Principles, to decide the best way for their organisation to determine personal attributes and proxies. For example, if the feature candidate list is large for a use case, an FSI may determine visualisation is not practical and instead of computing association for all pairs using full data, may only compute interesting pairs using sampled data. Also, a business driven framework could be used to identify more complex and potentially dangerous associations (e.g., between subsets within an attribute and those of another, such as between income and gender only in the 30-45 age range).



## Analysis to Justify Inclusion of personal attributes in decision making

The impact of each personal attribute on systematic disadvantage as measured by the fairness metric of false negative rate ratio, and as based on the personal attributes of gender and ethnicity, was calculated. Insurer A computed the impact using a Leave-One-Covariate-Out (LOCO) approach on the logistic regression model. In this approach, a new model is trained by dropping each feature, one at a time, to ascertain the impact on the fairness metric - False Negative Ratio, and the commercial performance metric – balanced accuracy.

Grp	Personal Attribute	FNR Ratio based on gender (acceptable range threshold 0.8 - 1.2)			FNR Ratio based on ethnicity (acceptable range threshold 0.8 - 1.2)			Balanced accuracy (minimum threshold to meet business objectives = 82%)		
		Baseline	LOCO	Impact (LOCO - Baseline)	Baseline	LOCO	Impact (LOCO - Baseline)	Baseline	LOCO	Impact (LOCO - Baseline)
G2	Gender	1.35	1.22	-0.32	0.98	0.98	-0.00	0.831	0.818	-0.013
G1	Ethnicity	1.53	1.56	0.02	0.98	0.95	-0.03	0.831	0.832	0.001
G1	Nationality	1.53	1.58	0.04	0.98	1.01	0.03	0.831	0.828	-0.003
G1	Marital Status	1.53	1.54	0.01	0.98	0.98	0.00	0.831	0.828	-0.003
G2	BMI	1.53	1.59	0.06	0.98	0.96	-0.02	0.831	0.812	-0.019
G2	Age	1.53	1.52	-0.01	0.98	1.02	0.04	0.831	0.827	-0.004
G2	Pay-out Amount	1.53	2.66	1.12	0.98	1.11	0.13	0.831	0.770	-0.061

■ Within Threshold Range ■ Outside Threshold Range

Table 2.4: Leave-One-Covariate-Out Analysis for all Personal Attributes with FNR impact for Gender and Ethnicity

## Personal attributes classified as Group 2 – justification to include in decision making

### Gender:

Exclusion of gender from the model reduces FNR Ratio from 1.53 to 1.22, which is very close to the maximum threshold, but it also decreases balanced accuracy by 1.3% from 83.1% to 81.8% (top line in yellow). As the target minimum balanced accuracy to meet Insurer A's commercial objective is 82%, Insurer A's fairness standards state that this can be used as justification to include gender in the model, as removal of this data brings the model below the threshold.

However, it is just below the minimum balanced accuracy threshold and given the reduction in FNR Ratio achieved by removing gender, if a mitigation solution had not been found to reduce the gender FNR Ratio, a consideration would have been made to remove gender on the basis of the above.

**In terms of the other personal attributes classified as Group 2, BMI, age and previous pay-out amount** are commonly used attributes to estimate the risk eligibility for life insurance with a causal relationship with the target (risk of claim). This is Insurer A's main justification for using them in the model. This justification is reinforced by the quantitative outcomes:

- Insurer A also checked the impact on balanced accuracy (see the table above), and the exclusions of each of these attributes reduced the balanced accuracy.
- In addition, in the feature importance diagram below, BMI and previous pay-out data are both in top 10 features. Age is lower down, but that is partially due to the fact that ages <18 and >60 were excluded from the model development dataset, as the AIDA System will have a pre-processing business rule to exclude these age groups from the campaign when deployed, which would reduce the feature importance of the attribute in the model.

### Personal attributes classified as Group 1 (should not be used in decision making)

The attributes classified as Group 1 personal attributes, were classified as such mainly due to Insurer A's organisational values and societal acceptance. The quantitative analysis below was not required as Insurer A had already decided not to use these attributes, but was conducted nonetheless to reinforce its decision.

#### Ethnicity:

As can be seen in the table above, the exclusion of ethnicity from the model reduces the FNR Ratio for Ethnicity a small amount – from 0.98 to 0.95 – but it is still well within the 0.8-1.2 acceptable band. Additionally, removing ethnicity actually increases balanced accuracy a small amount. This reinforces the decision to keep ethnicity out of the model.

**In terms of the other personal attributes classified as Group 1 – nationality, and marital status** – it is observed that balanced accuracy drops, but only by a small amount (it is still over the 82% minimum).

Furthermore, none of the Group 1 attributes feature in the top 25 features in terms of importance. Again, this reinforces the decision to exclude these attributes from the model.

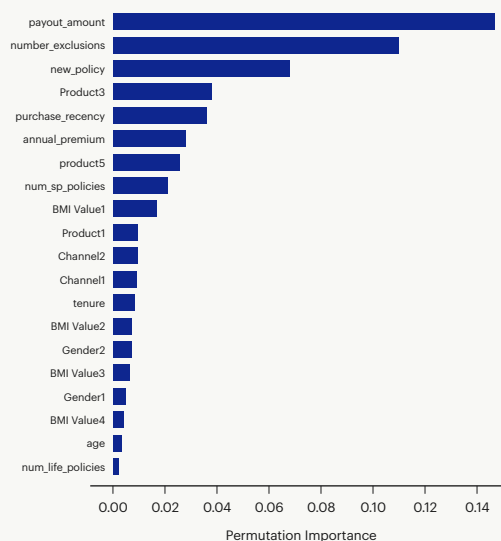


Figure 2.17: Permutation importance

**Note:** Because of the dynamic nature of data driven approaches and the use of AIDA systems, there remains an ongoing discussion around the definition and use of certain personal attributes with differences appearing across regions. FSI need to monitor the latest regulatory developments in this field. FSIs also need to be attentive to the appropriateness of demographic attributes in the different jurisdictions that they operate in to avoid unintended harm.

Another point to note is that data that can be used by AIDA systems may not be classified as personal data at the time it is designed/developed, but if combined with other information could become personally identifiable information later in the process. That means data which initially falls outside the scope of applicable data protection and privacy laws may fall subject to these laws when it is combined with other information. Additionally, in some jurisdictions it is prohibited to differentiate along certain dimensions (e.g., gender and ethnicity) due to anti-discrimination laws. As per the fairness methodology document, FSI should continuously monitor for these issues, and respond accordingly throughout the AIDA system development lifecycle.

## 2.7.3 Step 3: Build and Validate

### 2.7.3.1 Part B(ii): Examine Models

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?**

[This question refers to question B3, B4 and B5 in the Phase 1 Fairness Methodology questions]

#### Components of the AIDA system:

The AIDA system is made up of one pre-processing business rule, one logistic regression model and two post-processing business rules.

#### 1. Logistic regression model

The main component of the AIDA system is the logistic regression model – a predictive underwriting model that was developed to predict the likelihood or probability of an individual (an existing customer) being eligible for a life insurance product offered by a simplified underwriting campaign with no price loading. An individual is defined as eligible for life insurance if their risk fits the appetite of the insurer to include them into the risk pool.

With respect to the model type, a logistic regression model was trained to predict the likelihood of an individual eligible for life insurance based on a mixture of underwriting, demographic and policy information. Logistic regression was chosen as the prediction model in the context of life insurance underwriting for the following reasons:

- Relative Interpretability.
- Robustness.
- Readily and easily extensible by bias mitigation methods

The model outputs the likelihood that an individual will be eligible for life insurance as a probability value between 0 and 1. This is interpreted as an eligible risk score. Applicants with an eligible risk score above a certain threshold can be offered life insurance without the need for full underwriting. This represents a shorter and simpler process between applying and obtaining life insurance when compared to those individuals that must undergo full underwriting. In this simplified use case, the threshold was chosen to maximise balanced accuracy. The business looks into the pooling characteristics of each range of thresholds and decides the best threshold for the eligible/ineligible decision based on product nature (medical, critical illness or mortality), insurer's risk appetite (balance in the book) and market context (the specific market experience).

The logistic regression model was trained on a synthetic dataset. For details of the dataset see section F1.

The features on which the final logistic regression model was trained on (a subset of the features in the model development dataset) are as follows:

Attribute	Description
<b>BMI</b>	BMI
<b>Age</b>	Current age
<b>Tenure</b>	Years as a customer
<b>Gender</b>	Gender
<b>Smoking</b>	Yes/No
<b>Annual Premium</b>	Premium for a year
<b>Previous pay-out amount</b>	Pay-out amount over previous period
<b>Number of new policies past period</b>	Number of new policies past period
<b>Number of life policies</b>	Number of life policies
<b>Number of personal accident policies</b>	Number of personal accident policies
<b>Number of single premium policies</b>	Number of single premium policies
<b>Number of exclusions</b>	Number of exclusions for which the insurer does not provide coverage
<b>Purchase recency</b>	Latest recent purchase
<b>Latest purchase distribution channel</b>	Distribution channel of the policyholder
<b>Latest purchase product category</b>	Product category of the policyholder
<b>Policy duration</b>	Duration of the Policy

Table 2.5: Feature used in model development

Attributes excluded were:

- **Marital status, nationality and ethnicity** were not used to train the model, but were used to assess fairness, as they are assessed by Insurer A to be personal attributes in Group 1. See F8 for more information on how this was assessed.
- **Postcode** data was dropped from the model development dataset, in line with data minimisation principles, once assessment had concluded that it was not a material proxy for any of the Group 1 personal attributes.

List of pre-processing steps:

- One-hot-encoding of categorical variables (a pre-processing step to convert categorical variables into numerical variables).
- Standardisation of numerical variables.

## 2. Business rules working with the predictive model

Certain business rules are applied and overrule the model predictions to further control the risk and business context. For this AIDA system the business rules used are:

1. Pre-processing rule (before the model is applied):
  - a. Restrict the potential offer to those aged between 18 and 60 years. The reasons are:
    - i. people under 18 years are most likely not decision makers for an insurance policy. Furthermore, there is limited information (such as income, lifestyle) on this young demographic, which is likely to impact the accuracy of the system for this group.
    - ii. For age 60 and above full underwriting will be applicable to this age group.This excludes 5% of the development population and these customers will not get the offer by default.
2. Post-processing rules (on the model outcome) exclude the below even if the model let through:
  - a. Exclude offer from those with less than three years with Insurer A – the reason for this is that the model works less well (i.e., it has lower accuracy and higher errors) on this population as a number of the model factor attributes are using historic data going back beyond three years. This rule excludes approximately 1% of the population and these customers will not get the offer.
  - b. Exclude offer from those with a very serious claim history. This is because these customers should always be manually assessed by an underwriter, due to the high level of risk.

These rules excludes 1% of the population and these customers will not get the offer.

## 3. Human manual overrides

In some cases, humans override the decisions of AIDA systems, such as when a model's outcomes are particularly uncertain. The AIDA system described in this use case includes no such manual overrides. Instead, strict business rules are used to exclude cases that would typically be sent for manual review (e.g., high cost claims). In this simplified example, only three business rules are used, but a real PUW AIDA system would typically have many more.

## How is each component of the system - including AIDA models, business rules and human judgement if relevant - used to achieve the insurer's commercial objectives?

### 1. Models

There is only one model in the example AIDA system. It is a logistic regression model trained to predict the likelihood of a customer being an eligible risk for life insurance.

All performance measures were calculated on a test set, which were held out from the 'train' dataset used for model training and hyperparameter tuning. Insurer A applied an 80/20 split between train and test, and the test set contained 4,600 data points. Modelling choices such as feature selection, tuning of L2 regularisation parameters, and the tuning of the prediction threshold were made using k-fold cross validation (k=10) on the training set.

Performance measures were calculated after the pre-processing business rule on age, but before post-processing business rules were applied.

The performance metrics used when training the model were:

- Area under the receiver operating characteristic curve (AUC) for threshold-free metrics to check the model performance.
- Balanced accuracy tackles the imbalanced labels. This is the quantitative measure for the primary business objective as highlighted in answer to question G5.
- Precision captures the portfolio risk level and Recall captures the business opportunities.

The uncertainty for each of these measures is calculated using the empirical bootstrap method with 50 replications and 5-95% confidence intervals used, the plus-minus intervals representing two standard deviations.

Performance measure (rate) measure	Value	Formula	Meaning
Balanced accuracy	0.827+/- 0.014	$\text{Balanced accuracy} = \frac{\text{TPR} + \text{TNR}}{2}$	The arithmetic mean of the true positive rate and true negative rate.
Precision	0.963+/-0.07	$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}$	The proportion of customers offered convenient/simplified life insurance policies who did not (or hypothetically would not) claim life insurance.
True positive rate (Recall)	0.805 +/- 0.012	$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}$	The proportion of customers did not (or hypothetically would not) claim life insurance who are correctly predicted.
True negative rate	0.855 +/- 0.019	$\text{TNR} = \frac{\text{TN}}{\text{TN} + \text{FP}}$	The proportion of customers did (or hypothetically would not) claim life insurance who are correctly predicted.
AUC	0.898+/-0.01	-	AUC is an important evaluation metric for calculating the performance of any classification model's performance. AUC gives the rate of successful classification.  The higher the AUC, the better the model is at distinguishing between eligible and ineligible customers.

Table 2.6: Performance metrics

The test set confusion matrix for the model is displayed below:

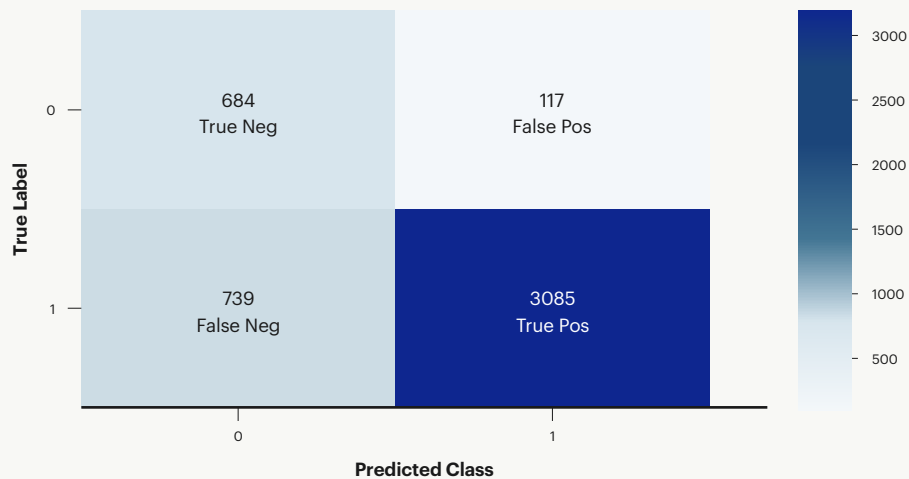


Figure 2.18: Confusion matrix of test data

**As can be seen from the table above, the model meets the target of over 82% balanced accuracy and the constraint of over 96% precision – therefore meeting the quantitative targets for insurer a’s commercial objectives as outlined in answer to question G5.**

It is important to note that the confidence interval for the balanced accuracy is relatively small – less than 2% of the balanced accuracy value and fully above the 80% minimum threshold. On the other hand, precision has a wider confidence interval, which does go below the minimum threshold. This metric will be monitored closely post deployment for early signs that the precision of the model is below the minimum threshold, as this will impact the loss rate on the portfolio.

## 2. Business rules

As can be seen from the table below, the model alone contributes to identifying customers that are eligible risk to fulfil the main business outcome of simplified underwriting for existing customers. The business rules only exclude customers (i.e., those flagged as ineligible).

AIDA system outcomes – development population				
	Pre-processing business rules: age	Model	Post-processing business rule: claims, new customer	Totals
Total %	5%	93%	2%	100%
<b>Model Outcomes</b>				
Model - Eligible	-	74%	-	74%
Model - Ineligible	-	19%	-	19%
<b>Non-Model Outcomes</b>				
Non-Model - Eligible	-	-	-	-
Non-Model - Ineligible	5%	-	2%	7%
<b>AIDA System Outcomes:</b>				
Total - Eligible	-	-	-	74%
Total - Ineligible	-	-	-	26%

Table 2.7: Business rules

## 2.7.4 Part C – Measuring Disadvantage

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?

[This question refers to question F9 in the Phase 1 methodology]

The quantitative estimates of the system's fairness objectives below are made through the fairness metric selected in F3 above using the fairness metrics decision tree based on Aequitas' model. This is the process to use, as defined in Insurer A's fairness standards.

**The key metric determined to optimise for fairness is equal opportunity – or false negative error rate (FNR) = FN/(FN+TP).**

That is, "unfairness" for this use case is considered to mean that one group has a higher proportion of customers not offered simplified underwriting among those who are eligible for this. The selection of this metric best fits the use case's fairness objectives (see question F3).

The unit of measure used is the number of unfair events, rather than estimating a monetary impact on customers, as this is a medium-low risk use case and further estimates to get to this level of detail are not required for medium-risk use cases, as per Insurer A's fairness standards.

Insurer A's fairness standards state that a difference of over 20% in fairness measure rates (i.e., the rate of occurrence of harms or benefits) between personal attribute subgroups is "significant" and should be investigated.

Insurer A also ran the 'standard set' of fairness metrics, as defined in its standards, to check that there were not large imbalances elsewhere that it should be aware of. Uncertainty is calculated for the fairness metrics in the same way that it was for model and system performance measures, using the empirical bootstrap method with 50 replications and 5-95% confidence intervals used and the plus-minus intervals representing two standard deviations.

### Standard fairness measure

### Calculation formula

### Interpretation in personalised underwriting context

False positive rate ratio

$$\text{FPR Ratio} = \frac{\text{FPR}_1}{\text{FPR}_2}$$
$$\text{FPR} = \frac{\text{FP}}{\text{FP} + \text{TN}}$$

An equal fraction of convenient/simplified life insurance policies is marketed (across groups) to the individuals who are bad risk and not eligible for life insurance.

False omission rate ratio

$$\text{FOR Ratio} = \frac{\text{FOR}_1}{\text{FOR}_2}$$
$$\text{FOR} = \frac{\text{FN}}{\text{TN} + \text{FN}}$$

An equal fraction of individuals who are eligible for life insurance among those customers who are not marketed/approached for convenient/simplified underwriting life insurance.

False negative rate ratio

$$\text{FNR Ratio} = \frac{\text{FNR}_1}{\text{FNR}_2}$$
$$\text{FNR} = \frac{\text{FN}}{\text{TP} + \text{FN}}$$

An equal fraction of convenient/simplified life insurance policies is not marketed (across groups) to customers who are good risk and are eligible for life insurance.

### Standard fairness measure

### Calculation formula

### Interpretation in personalised underwriting context

False discovery rate ratio

$$\text{FDR Ratio} = \frac{\text{FDR}_1}{\text{FDR}_2}$$

Among individuals who are marketed (across groups) for convenient/simplified life insurance, an equal fraction of individuals is bad risk and not eligible for life insurance.

$$\text{FDR} = \frac{\text{FP}}{\text{TP} + \text{FP}}$$

Disparate impact

$$\text{DI Ratio} = \frac{\text{DI}_1}{\text{DI}_2}$$

An equal fraction of convenient/simplified life insurance policies is marketed to the customers across groups.

$$\text{Disparate Impact} = \frac{\text{FP} + \text{TP}}{\text{FP} + \text{TP} + \text{TN} + \text{FN}}$$

Prevalence

$$\text{Prevalence Ratio} = \frac{\text{Prevalence}_1}{\text{Prevalence}_2}$$

An equal fraction of individuals who are eligible for life insurance among all the individuals (across groups).

$$\text{Prevalence} = \frac{\text{TP} + \text{FN}}{\text{FP} + \text{TP} + \text{TN} + \text{FN}}$$

Table 2.8: Fairness metrics formula

## Fairness metrics for gender

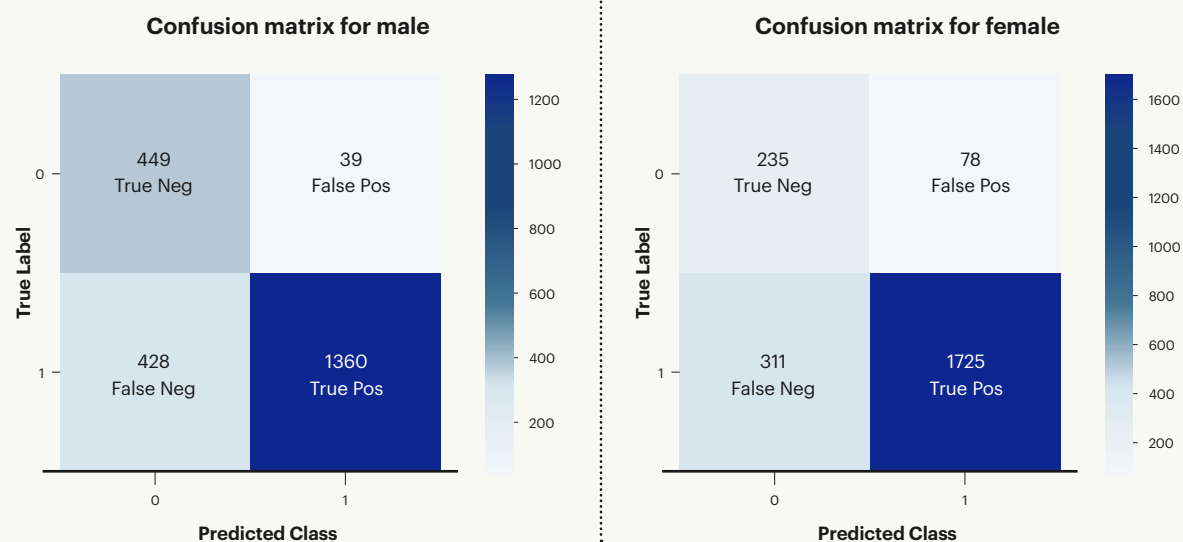


Figure 2.19: Confusion matrix for gender

The fairness metrics along with confidence scores are below:

Fairness metrics - Gender - male/female

	Prevalence ratio	Disparate impact	False negative rate ratio	False positive rate ratio	False discovery ratio	False omission ratio
Value	0.906	0.799	1.593	0.339	0.685	0.864
Acceptable Range	0.8 < value < 1.2					
Confidence Intervals @95%						
Range	± 0.022	± 0.033	± 0.240	± 0.128	± 0.284	± 0.076
Lower	0.884	0.766	1.353	0.211	0.401	0.788
Upper	0.928	0.832	1.833	0.467	0.969	0.940

Within Threshold Range

Outside Threshold Range

Confidence Range stays within threshold

Confidence Range goes outside threshold

Table 2.9: Fairness metrics with confidence score for gender

As can be seen in the table above, the false negative rate ratio for gender is 1.593 which means that male customers who are eligible for life insurance are 1.593 as likely to not have been offered convenient/simplified life insurance than female customers. This is our key fairness measure, and the value breaches the threshold of 1.2 set in Insurer A's fairness standards. The confidence intervals tell us that there is a 95% likelihood that the range 1.353 – 1.833 covers the true False Negative ratio. From the point of view of measuring disadvantage, and the 4/5th rule threshold, all the values in this range are over the threshold of 1.2.

For the standard set of fairness metrics, the results were run to check there were not large imbalances elsewhere that Insurer A should be aware of. The results were as follows:

In the case of false positive rate ratio, a value of 0.339 means that male customers who are not eligible for life insurance are as 0.339 as likely as female customers to be approached/ marketed for convenient/simplified life insurance, which is again below the 0.8 threshold and a disadvantage for males.

False discovery rate ratio is also below the 0.8 threshold. Disparate impact: the ratio of predicted positive outcomes for male customers over that of female customers is 0.685 which is slightly below the 0.8 threshold, again to the disadvantage of males. This value is lower compared to the same calculation (ratio of positive outcomes) on the observed or historical data (prevalence rate ratio) which for the test dataset is 0.906. It would be expected that the prevalence ratio and disparate impact ratio would be fairly similar. This will be checked again on the full population for the first campaign (it is the only fairness metric that doesn't require a performance period post deployment), to see if it improves/ comes closer to the more balanced prevalence rate for the live population.

Confidence intervals are quite wide on most of the metrics, so if the value had been within band, there would have been low confidence that this would actually be the case in a live environment (the width of the bands is likely to have been impacted by size of sample).

Insurer A also checked the outcome of the accuracy measures for subgroups male and female (precision, recall, accuracy and balanced accuracy) to see if there were large imbalances there, but all ratios were within the 0.8-1.2 threshold band.

## Outcome for gender:

Overall, the metrics above suggest a potential disadvantage for male customers over female customers with the key Fairness Metric, False Negative ratio, falling outside the 1.2 threshold, along with a number of other the standard fairness metrics. In the next section (F10), mitigation methods are assessed to rectify this imbalance.

**Note:** For higher risk use cases, Insurer A's fairness standards would have required further investigation into the reasons for the high level of imbalance in the False Negative ratio. A potential reason is the higher prevalence rate in females, and therefore a different risk distribution,<sup>7</sup> or the lower performance of the model for male customers.

## Fairness metrics for ethnicity

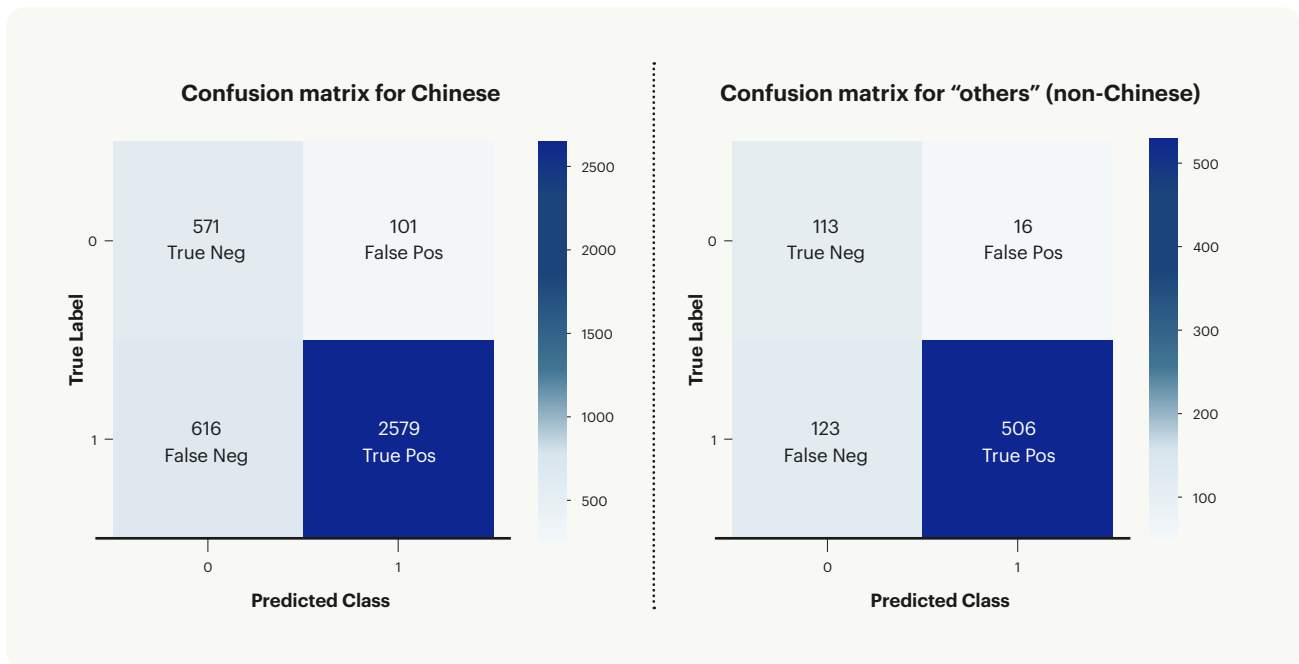


Figure 2.20: Confusion matrix for ethnicity

The fairness metrics along with confidence scores are below:

Fairness metrics - Ethnicity - Chinese/non-Chinese						
	Prevalence ratio	Disparate impact	False negative rate ratio	False positive rate ratio	False discovery ratio	False omission ratio
Value	0.989	1.001	0.987	1.190	1.255	0.974
Acceptable Range	0.8 < value < 1.2					
Confidence Intervals @95%						
Range	± 0.031	± 0.054	± 0.172	± 0.567	± 0.542	± 0.128
Lower	0.95869	0.947	0.815	0.623	0.713	0.846
Upper	1.0231	1.0557	1.159	1.757	1.797	1.102

■ Within Threshold Range
■ Outside Threshold Range
■ Confidence Range stays within threshold
■ Confidence Range goes outside threshold

Table 2.10: Fairness metrics with confidence score for ethnicity

The false negative rate ratio for ethnicity is 0.987. This is well within the acceptable 0.8 to 1.2 range as defined in Insurer A's fairness standards. Similarly, the other standard fairness metrics all fall within 0.8 and 1.2.

Insurer A also checked the outcome of the accuracy measures for each subgroup (precision, recall, accuracy and balanced accuracy) to check for large imbalances. However, all ratios were within the 0.8-1.2 threshold band. This finding is especially relevant in the context of ethnicity, as the non-Chinese population is a small proportion of the overall population and differences in model outcomes would likely be indicated by diverging accuracy metrics. Conversely, if the model behaviours are the same for both groups, there is no need to adjust for representation bias. As there were no large imbalances in the accuracy and fairness metrics for the gender subgroups, the behaviour was assessed to be sufficiently similar, and **as a result no mitigating action was taken on to change the distribution of ethnicity in model development data.**

Once again, the confidence intervals are quite wide for a number of the fairness metrics examined, and specifically for the key metric of false negative rate. However, as the value sits right in the middle of the band, the top and bottom of the band are just within the allowable thresholds.

### Outcome for ethnicity:

Overall, the outcomes of the fairness metrics above suggest that outcomes for ethnicity subgroups are balanced. This completes the analysis that was referred to in section F4, where it was stated that further analysis is needed to determine if representation bias is an issue. The fact that the key Fairness Metric outcome is within threshold, combined with the fact that outcomes are also balanced for accuracy measures, indicates that representation bias for ethnicity does not seem to be occurring.

**Note:** As this is a medium-low risk use case for fairness, no further action was taken regarding the fact that some of the confidence intervals are wide for both gender and ethnicity.



## Individual fairness

In addition to these group fairness considerations, Insurer A aims to be fair to individual customers. Final eligibility decisions for simplified underwriting are made based on the model score following the application of business rules. Insurer A considers the risk score output to be the measure of individual similarity. That is, it interprets individual fairness to mean that customers with the same risk scores receive the same eligible/ineligible decisions. This is measured by identifying the volume of applicants with identical risk scores that receive different decisions due to post-processing business rules. Below, the final outcome is calculated for customers in the test dataset that trigger one of the post-processing business rules:

Total test dataset		4625	
Post-processing business rule	# observations	Model outcome	# outcome changed due to business rule
1. Customers with very serious claims history	46	2 eligible, 44 ineligible	2
2. Customers with <3 years with Insurer A	45	13 eligible, 32 ineligible	13

15 customers in total received a different outcome than other customers with the same score.  
This is approximately 0.3% of the population.

Table 2.11: Summary of customers with same model score and different final outcome

**The outcome for the individual fairness assessment** is that the % of the population that receive a different outcome than the rest of the population “similar” to them is **low at 0.3% and Insurer A's fairness standards direct that no further action is required at this level.**

**Note:** Quantifying individual fairness and determining an acceptable level is currently a nascent area of research with no agreed standard methodology.

F10

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

[This question refers to question C2 in the Phase 1 methodology]

### Mitigation explored to meet fairness objectives for gender:

**Insurer A looked to explore the trade-offs between the systems' fairness objectives and its other objectives by applying post-processing algorithmic interventions for bias mitigation to reduce the false negative rate ratio for gender to 1.2, to be within the 0.8-1.2 band.**

Algorithmic interventions are available at the pre-processing, in-processing and post-processing stages to improve fairness of the AIDA system. Bias mitigation algorithms attempt to maximise AIDA system model performance yet also conform with respect to user-provided fairness constraints.

Insurer A tried a number of in-processing mitigation methods, including AIF360: adversarial debiasing and fairlearn reduction gridsearch (see the appendix to the Veritas Phase 2 Fairness Methodology document for more detail on technical bias mitigation methods). However, the best results in terms of meeting commercial and fairness objectives was post-processing mitigation in the form of constrained balanced accuracy for gender, which is documented below.

In constrained balanced accuracy, the threshold for the classification model (eligible/ineligible risk) is a key operating parameter that affects the fairness metrics. The mitigation approach selects separate classification thresholds for groups for which it aims to optimise fairness, in this case males and females.

To minimise imbalance with respect to the chosen fairness metric amongst subgroups while maximising model performance, a grid search for thresholds was conducted to bring the FNR ratio to within  $\pm 0.2$  of neutrality while maximising the balanced accuracy.

Once the analysis had been run, the fairness-performance trade-offs of operating the model at various eligible risk thresholds was visualised – see figure 2.8 below. The **heatmap indicates the model's expected performance (balanced accuracy) when operated at each pair of male/female risk thresholds. The white contour lines indicate the false negative rate ratio group fairness metric with respect to gender.** It is optimal when equal to one (1), but within the threshold of 0.8 to 1.2 is acceptable in terms of Insurer A's fairness standards.

Insurer A plotted three points of interest:

- **The blue diamond** allows different eligible risk thresholds for men and women and maximises the unconstrained model performance.
- **The red X** maximises model performance while keeping the same eligible risk threshold for both men and women.
- **The purple star** allows different thresholds for men and women and maximises the model performance while constraining for gender fairness, as measured via the false negative rate ratio. The constraint of ensuring gender fairness stays within the 0.8-1.2 threshold band for false negative rate ratio.

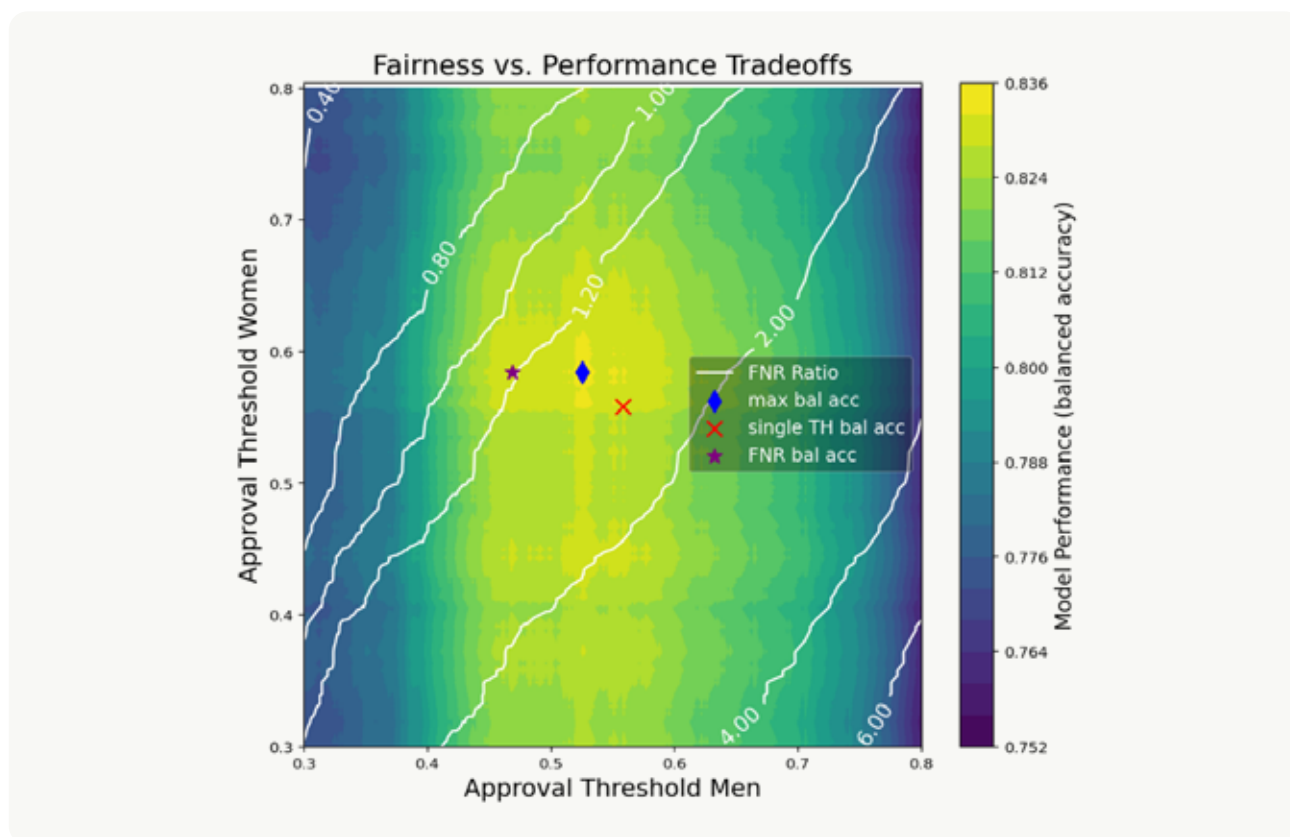


Figure 2.21: Fairness vs. performance trade-offs

When optimising gender fairness (the purple star), the model's balanced accuracy drops slightly, but is still above the 80% minimum required. Precision drops from 96.51% to 96.01%, which is still above the minimum constraint of 96.00% to maintain portfolio risk levels.

Post mitigation the FNR Ratio is still in band for ethnicity.




	Single Threshold	Threshold Male	Threshold Female	Commercial Objective: Balanced Accuracy	Commercial Objective: Precision	Fairness Objective: FNR Ratio
<b>Required Constraints</b>	NA	NA	NA	>82%	>96.00%	0.8 < 1.2
 <b>Single threshold for max balanced accuracy</b>	0.54	NA	NA	83.04%	96.34%	1.56
 <b>Split threshold for max balanced accuracy</b>	NA	0.54	0.59	83.27%	96.45%	1.42
 <b>Split threshold for max balanced accuracy constraining for fairness</b>	NA	0.48	0.59	83.14%	96.00%	1.18
<div> <div></div> Within Threshold Range           <div></div> Outside Threshold Range         </div>						

Table 2.12: Summary of impact on commercial and fairness objectives of mitigation strategies

A caveat is that when you optimise for one aspect of fairness, it can have trade-offs and subsequent knock-on effects in other aspects. These trade-offs can include other fairness metrics, impacts on subgroups of individuals, or the overall model performance.

The table below demonstrates that when Insurer A applies post-processing mitigation for gender, bringing the FNR down from 1.56 to 1.18, the ethnicity ratio goes from 0.998 to 0.949, which is slightly lower but still within the acceptable threshold.

The false positive rate ratio is included to check for significant knock-on imbalances, and shows that while ethnicity increases to slightly above the 1.2 upper bound at 1.29, for gender the FPR improves, increasing to 0.528 although still falling well below the 0.8 lower bound. As this is not the chosen fairness metric to optimise, this is acceptable by Insurer A's standards.

#### Fairness metrics before gender fairness mitigation

	Disparate Impact	False Negative Rate Ratio	False Positive Rate Ratio
<b>Ethnicity</b>	1.004	0.998	1.21
<b>Gender</b>	0.801	1.563	0.318

#### Fairness metrics after gender fairness mitigation (post-processing)

<b>Ethnicity</b>	1.017	0.949	1.29
<b>Gender</b>	0.870	1.18	0.528

 Within Threshold Range
  Outside Threshold Range

Table 2.13: Summary of impact on commercial and fairness objectives of mitigation strategies

### Impact on individual fairness of post-processing mitigation of split gender thresholds:

Having different eligibility thresholds for males and females has an impact on individual fairness. As stated in the previous section, Insurer A considers the risk score output to be the measure of individual similarity. That is, the firm interprets individual fairness to mean that customers with the same risk scores receive the same eligible/ineligible decisions. Below, the model is calculated for customers in the test dataset that are between the split thresholds, separated by males and female, as they receive different outcomes for the same scores between thresholds:

Total test dataset		4625
Individual with prob between (.48 to. 589)	# observations	Model outcome
1. Males	160	Eligible
2. Females	85	Ineligible

Table 2.14: Summary of customers with same model score and different final outcome

With the split thresholds, 85 females receive a non-favourable outcome compared to males with the same score. This is approximately 2% of the total population. Individual fairness is therefore negatively impacted by the post-processing mitigation to reduce group unfairness. Insurer A's fairness standards state that when a trade-off has to be made between group fairness and individual fairness, that group fairness is prioritised.

### Summary of exploring the trade-off of reducing the false negative ratio for gender:

1. Impact on commercial performance.

As can be seen above, when classification thresholds are set to constrain for fairness on gender (i.e., reducing the FNR to 1.18 to fall within the acceptable threshold), then the balanced accuracy reduces slightly to 83.14%, which is it still above the 82% minimum threshold.

2. Impact on risk of portfolio.

When classification thresholds are set to constrain for fairness on gender (i.e., reducing the FNR to 1.18), the resulting precision is 96.00% which is same as the 96% threshold required to maintain the portfolio risk levels for the new business underwritten by the campaign.

3. Impact on ethnicity.

When classification thresholds are set to constrain for fairness on gender (i.e., reduce the FNR to 1.18), the FNR for ethnicity stays within fairness thresholds, reducing from 0.998 to 0.949.

4. Impact on individual fairness.

When classification thresholds are set to constrain for fairness on gender (i.e., reducing the FNR to 1.18), individual unfairness goes from impacting approximately 0.3% of the population to approximately 2% of the population.

The decision on mitigation is provided in next question.

**F11**

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

[This question refers to question C3 in the Phase 1 methodology]

#### Decision on mitigation:

The AIDA system can be operated with gender specific thresholds to meet the fairness objective for gender, and to keep the chosen fairness metric of false negative rate (FNR) ratio in the acceptable deviation range of 20% from parity (between 0.8-1.2).

These gender-specific thresholds also kept accuracy and precision above the minimum thresholds required to meet the firm's primary commercial objectives, as well as kept ethnicity FNR ratio within the acceptable fairness threshold. Individual fairness was negatively impacted but this trade-off is acceptable by Insurer A's fairness standards.

**On this basis, Insurer A has decided to implement the post-processing mitigation of split gender thresholds, thereby bringing the false negative rate ratio within the acceptable threshold band for the personal attribute of gender.**

As noted, this is a hypothetical insurer with hypothetical standards – each FSI will define their own standards in line with their individual circumstances.

## 2.7.5 Step 4: Deploy and Monitor

### Part E: Examine System Monitoring and Review

**G12**

### Is the system's monitoring and review regime designed to detect abnormal operation?

[This question refers to question E1 in the Phase 1 Fairness Methodology questions]

Insurer A has set up monitoring in line with its model development lifecycle standards, with standard regular reporting on key KPI metrics as well as continuous monitoring to quickly identify large changes in inputs or outputs that could signal an issue/malfunction in the system. This monitoring will stay in place as long as the AIDA system is used for campaigns.

The reports produced are assessed by a member of the AIDA system assessor team, which is also the AIDA system validation and monitoring team, independent of the AIDA system development team. They will set out the relevant action to take if there is a breach of KPI thresholds and present to the AIDA system owner who is responsible to put required actions in place.

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

[This question refers to question E3 in the Phase 1 Fairness Methodology questions]

In general, the mitigation plans for Insurer A in the case of triggers from monitoring for either commercial or fairness metric breach triggers is an investigation, with no further campaigns run until the issue has been addressed to the satisfaction of the AIDA system owner.

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

[This question refers to question E2 in the Phase 1 methodology]

Insurer A has added to standard regular monitoring of this AIDA System:

1. Accuracy metrics that are being reported for the overall population will also be reported for each subgroup in gender and ethnicity, along with the ratio metric and a flag will be raised for investigation if the ratio goes outside the 0.8-1.2 threshold band.
2. The key fairness metric – false negative rate ratio – will be monitored for gender and ethnicity, with a flag raised for investigation if the ratio for either goes outside the 0.8-1.2 threshold.
3. The other standard fairness metrics documented in F9 for ethnicity and gender will be run and included for informational purposes (no flags assigned).

For the first campaign, the outcomes of the disparate impact ratio for gender on the full campaign population will be monitored to see if it goes above the minimum threshold of 0.8 on the live population. This is the only fairness metric that can be measured without a performance period post deployment, as it only uses model outcomes and does not require ground truth. If the number moves further outside the 0.8-1.2 band, this will be discussed with the AIDA system owner, who will decide the appropriate action.

Insurer A will add to the annual manual review of this model:

1. Relabelling a sample of those customers labelled as “ineligible” in the development population, to see if the label changes and if so, whether there been a change in the underlying risk over the past year, or if it is a sign of measurement bias.
2. Assuming there is not a large number of customers that bought life insurance as part of the campaign and have since claimed, a manual labelling exercise will be undertaken to understand both the model accuracy and the fairness of the AIDA system in the live environment. The exercise involves labelling a sample of the portfolio, including both eligible and ineligible decisions, and calculating the relevant accuracy and fairness measures of the sample, taking the relevant trigger action if the measures fall below the mandated thresholds.

## 2.8 FS Reflections of Fairness Assessment Methodology

### a. Swiss Re Context And Considerations

Digitalisation across the re/insurance value chain is accelerating. It is expected that this will enhance the value provided to customers and help close protection gaps. With their strong expertise and experience implementing data driven solutions, global reinsurers are particularly well placed to support the insurance industry, regulators and partners in building robust insights with AIDA enabled solutions. At Swiss Re more than 200 data scientists are working together with our technical and business experts. The Swiss Re Global Advanced Analytics Centre of Expertise, part of the Group Data Services unit, has already delivered more than 1200+ advanced analytics projects across the world as of January 2021.

The use of artificial intelligence and digital personal data raises ethical concerns regarding fairness, inclusion, hardship, and solidarity. In this context, Swiss Re has started to develop its own ethical guidance to enable swift digitalisation, while at the same time ensuring that we maintain customers' trust, differentiate our services and safeguard our reputation as a leading re/insurer.

Various regulators around the world have started to evaluate the need for regulations on the topic of big data and digital ethics, or have issued initial guidelines. Swiss Re actively engages in discussions with regulators by participating in regulatory expert groups, contributing to studies and reports, such as the project Veritas led by MAS, as well as by giving feedback to consultations and responding to regulatory questionnaires.

Swiss Re's internal governance also recognises that while technology offers many business opportunities, it also creates new risks. Digital governance requirements are increasing in both number and complexity. However, fragmentation and, as a result, partially uncoordinated governance approaches make it difficult for owners of digital services to navigate these requirements.

Swiss Re has developed a Digital Governance Framework (DGF) that aims to balance the often competing needs for fast business innovation and effective risk management. Designed to be comprehensive, risk-based and user-friendly, it makes requirements transparent and positions governance as a fundamental pillar of digitalisation.

Swiss Re was honoured to be the only Reinsurer organisation within the Veritas consortium working with the Monetary Authority of Singapore (MAS) and other financial industry partners to create the Veritas framework for financial institutions to promote the responsible adoption of artificial intelligence and data analytics.

We are thankful to MAS for having selected Swiss Re to lead the Veritas Phase 2 fairness assessment workstream, and to Accenture for the strong collaboration on delivering this initiative. We hope that the fairness methodology whitepaper and this predictive underwriting use case fairness assessment will be valuable guidance for the industry.

Since fairness is a key consideration in underwriting for re/insurers, Swiss Re is continuously working enhancing the fairness assessment methodology applicable to predictive underwriting, such as the scenario selected for this use case white paper. We will continue to support our clients and partners to implement AIDA solutions in accordance with all applicable laws and regulations. Our team of experts across the world will continue to further develop our DGF and digital responsibility frameworks to align with the development and adoption of AIDA systems, technical advancement, and regulatory changes in order to create value for individuals, society and the industry.

For more information, please refer to Swiss Re Group Governance Sustainability Report 2020.<sup>8</sup>

## b. Swiss Re & Great Eastern Reflections and Main Learnings from the Application of the Methodology

While applying the FEAT Fairness Assessment Methodology on our real portfolio of insurance customers we have made several key findings and observations. We will share the details of our findings in this section of the document.

### **Part 1: Define system context and design**

The answers to the questions on describing the system's commercial objectives, potential groups at risk, and potential harms, are reasonable to us as is the reasoning to get to the fairness objective and chosen fairness metric for the Singapore market.

### **Part 2: Prepare input data**

The attributes in the synthetic dataset are similar to a subset of attributes that exist at Great Eastern for current insurance customers. The answers to the questions on checking for data bias and potential mitigations again are reasonable. In addition, the answers to the questions on identifying personal attributes and potential proxies and justifying use in the AIDA system where relevant are sensible and logical for the Singapore market.

### **Part 3: Build and validate**

Great Eastern performed the same 'build and validate' steps that were applied to the synthetic dataset on an internal dataset. In this way, this helps to

1. validate that synthetic dataset was 'valid' (i.e., it produced realistic outcomes when used as a development dataset for predictive underwriting),
2. understand the likely outcomes that would be obtained when developing and deploying an AIDA system in this space.

The same steps were applied to retrain the logistic regression model built on the synthetic dataset, checked if the model met the commercial quantitative metrics, tested the outcomes for fairness and applied the mitigation techniques.

The model retrained on Great Eastern data met the key commercial quantitative metrics, but further refinement would have been required if it was going to be deployed.

When we ran the standard set of fairness metrics, our outcomes were as close or closer to parity for nearly every metric than outcomes with the synthetic data. When we applied the post-processing bias mitigation on gender, similar to the use case on the synthetic data, both personal attributes of interest, gender and ethnicity, had the key fairness metric of FNR ratio within the acceptable band of 0.8-1.2.

### **Part 4: Deploy and monitor**

The answers to the question on how monitoring will be extended to ensure systems impacts are aligned with its fairness objectives make sense.

## 2.9 Conclusion

At Great Eastern and Swiss Re, we are committed to putting our customers and their insurance needs first, and the commitment includes managing our customers' data ethically. Through the Veritas project, we have obtained a clear understanding of an approach for a fairness assessment of an AIDA system, which will help us verify the fairness of AIDA systems we may develop and deploy in the future.

The initiative helped to continue assessing the impact and start calibrating our internal governance frameworks for the FEAT assessment taking into consideration our existing frameworks, the materiality of AIDA systems and the cost of the FEAT assessment and potential mitigation, to ultimately achieve our business objectives, bringing more value to our customers and society.

**It has also become clear to us that finding the right balance for the level of FEAT fairness assessment is to stay proportional to the fairness risk of any use case, as these assessments could be implemented with additional costs which ultimately get passed on to consumers.**

Gaining an understanding of how the privacy enabling technology of synthetic data generation can be leveraged in this space was also a valuable experience. The synthetic dataset could also enable third parties to develop codes and models, reducing the time, cost and effort needed to develop and test AIDA systems for fairness.

For the purpose of demonstrating the application of the Fairness Assessment Methodology, we considered the scope of predictive underwriting AIDA system to cross sell life insurance products to existing customers for the Singapore market.

Our hypothetical scenario applied the fairness assessment methodology to gender and ethnicity subgroups. The model fairness assessment on ethnicity was satisfactory and no further action was required, however some areas of improvement were identified on the quality of the data collection for that personal attribute. The model fairness assessment on gender was just outside the limit of our acceptance threshold, therefore the suggestion for this use case was to make the model fairer while reducing slightly its performance. Our scenario was relatively simple, for more complex cases involving a big number of protected personal attributes, finding the right balance between model fairness for all subgroups and model performance is a much more challenging tasks and might require difficult trade-off choices.

One challenge is the identification, collection, management, and usage of personal attributes for fairness assessment, even if those attributes are not necessarily used by the AIDA system. In order to test and make models fairer to subgroups of a population, FSIs should first be able to identify such subgroups.

Therefore, one of the opportunities is to create the right data protection and privacy regulatory environment providing adequate data protection and privacy for individuals, while allowing AIDA system owners to lawfully access the necessary data for assessing fairness for relevant population subgroups, with the objective to limit discrimination of the designed AIDA systems. Close coordination among data privacy policy makers and regulators across different jurisdictions will be required to achieve optimal outcome in this area for individuals and society. We believe industry has a role to play to support these efforts, such as the input provided by the Veritas consortium initiative.

## 2.10 Disclaimer

While the information in this whitepaper is from reliable sources, the authors do not accept any responsibility for the accuracy or comprehensiveness of the information given or forward looking statements made. The information and forward looking statements provided in this document are for informational purposes only and in no way constitute or should be taken to reflect the authors' position, particularly in relation to any ongoing or future dispute. In no event shall the authors be liable for any loss or damage arising in connection with the use of this information and readers are cautioned not to place undue reliance on the forward looking statements provided in this document. The authors undertake no obligation to publicly revise or update any forward looking statements, whether as a result of new information, future events or otherwise.

# 03 E&A Assessment in Fraud Detection

## 3.1 Preface

With the purpose to act for human progress by protecting what matters, AXA's goal is to move from "payer to partner" for its customers. AXA has always been a leader in innovation, fostering progress in all its dimensions. Technology and data are often seen as enablers to accelerate our progress on this journey. Claims being the primary costs for insurance companies, effective claims management is key.

While adopting AIDA solutions, it is crucial to act with care and responsibility towards customers. A lack of caution could easily result in the violation of company values and principles and deteriorating trust relationships between AXA and our customers. In this case study, we look into a fraud detection AIDA solution in the claims value chain and explore the ethics and accountability principles needed to ensure the interests of stakeholders are safeguarded.

## 3.2 Introduction

Influenced by our values and stance as a responsible insurer, we have created an agile internal governance body with the aim of managing in the most effective way our mixed risk and principles based approach to AI development and adoption: "AXA Responsible AI Circle". Under the sponsorship of AXA leadership committee, this structure gathers key professional teams and internal responsible AI experts to give directions towards trustworthy, performant and value delivering AI.

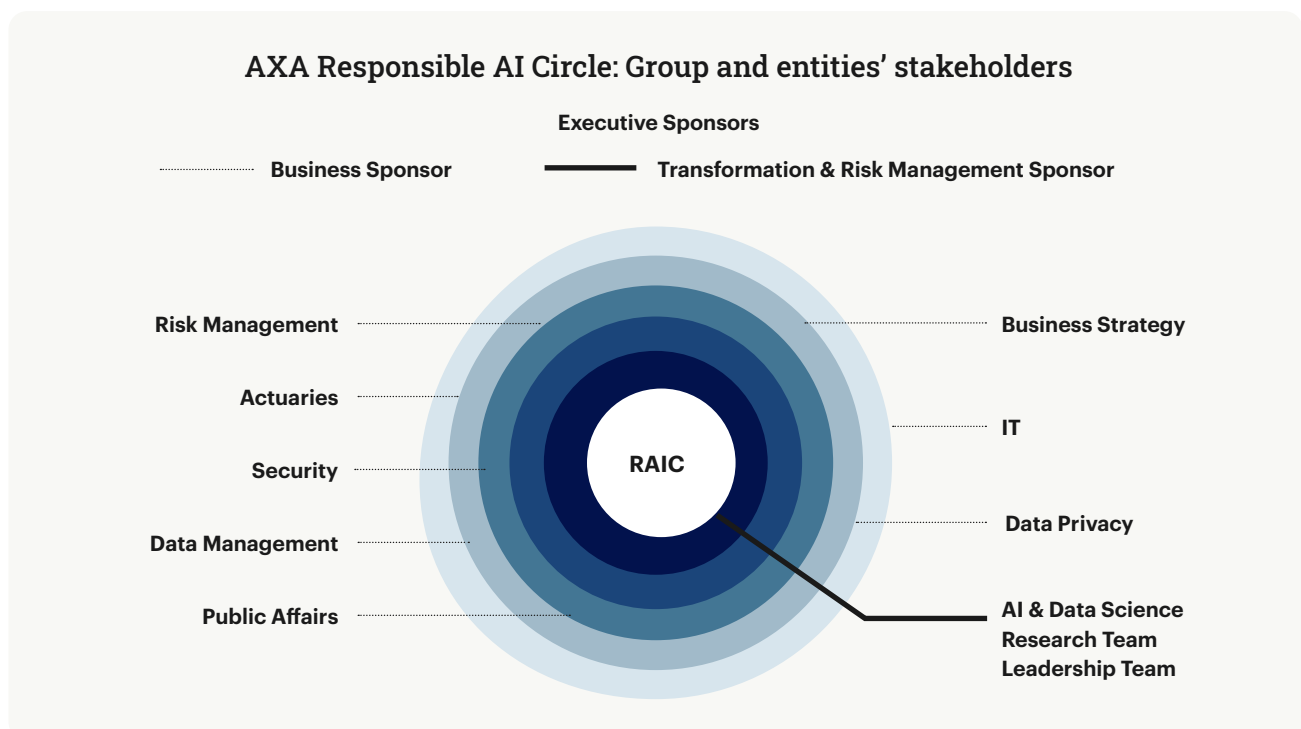


Figure 3.1: AXA Responsible AI Circle

### 3.3 Learning from Application of the Assessment Methodology

This section of the report gives a more detailed view of our fraud detection AIDA solution and our approach to applying the Ethics and Accountability Framework (“the Framework”) to this use case.

#### 3.3.1 System Objectives and Context

Sherlock is an advanced fraud detection AIDA solution that was built inhouse. It leverages a mix of batch and real time technology and aims to enhance savings by eliminating undue fraudulent claims. This constantly evolving product includes developments spanning underwriting fraud to document fraud. It uses a wide variety of machine learning techniques such as image recognition and fuzzy matching.

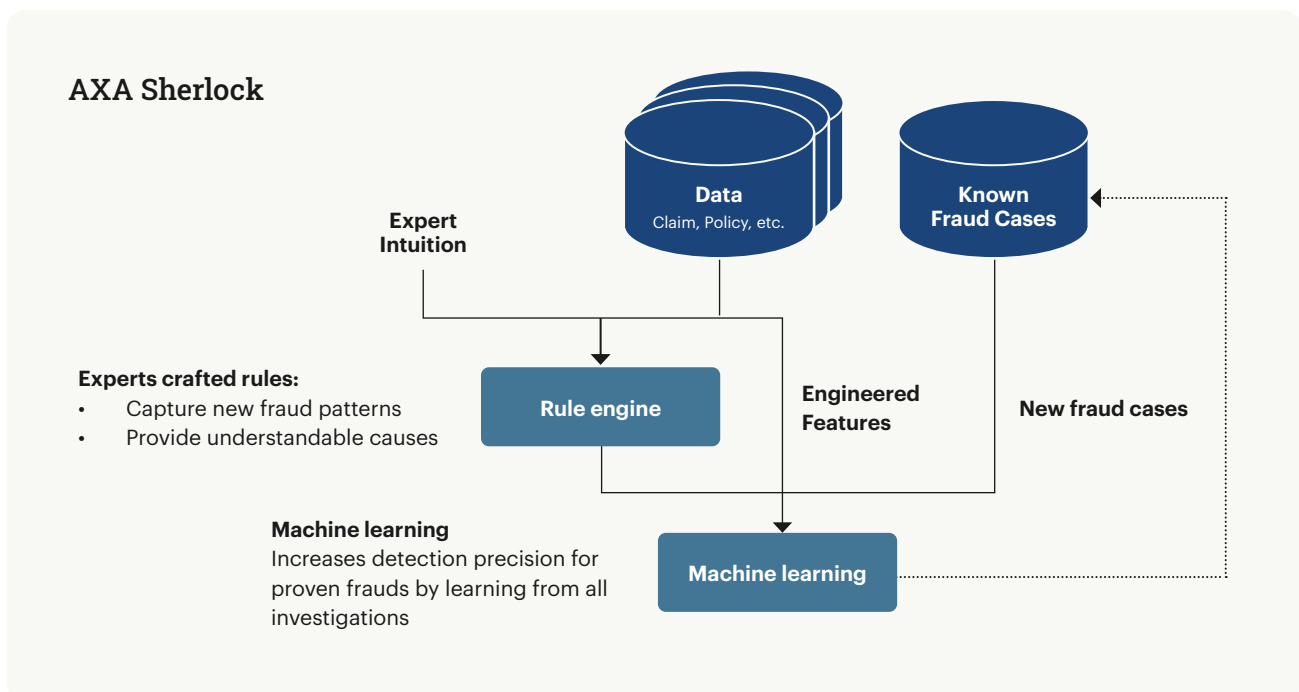


Figure 3.2: AXA Sherlock model

The philosophy of the product is to allow contributions both from global and local teams, to achieve the highest collaboration as well as to leverage a fruitful cocreation mode. Local experience flanked by global technical expertise are the cornerstones of the platform, which is currently helping numerous fraud teams in several countries across the AXA Group in Europe and Asia.

The strengths of Sherlock are its capability to connect all data points (i.e., people, claims, policies, providers, vehicles, phone numbers, e-mails, addresses, etc.) into one sophisticated network visualisation that allows for a pervasive investigation.

### Interactive Claim Network

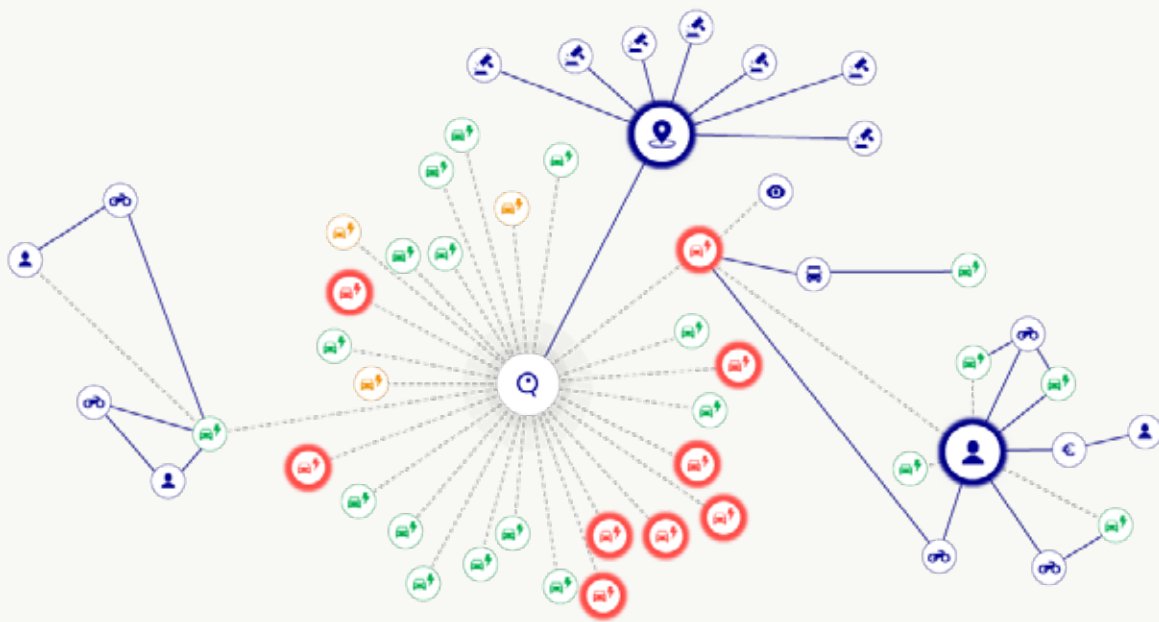


Figure 3.3: Interactive claim network

Leveraging AIDA models not only allows us to tackle fraud in an innovative and comprehensive way, but it is also clear evidence of the pervasive willingness across the AXA Group to heavily rely on new technologies. The continuous enhancement and development of the platform around its different pillars (data, machine learning and web) is pushing the boundaries of how fraud detection is conceived and addressed at AXA. Instead of outsourcing fraud activities, the Group has opted for an internal investment that has started to pay off in terms of market recognition, industry awards and, ultimately, increased benefits.

The short term objective of Sherlock is to protect against fraud schemes and enhance savings from illegitimate claims. In the long run the goal is to create a fairer, more transparent and trustworthy relationship with AXA customers. It is part of our strategic goal of being customer first: eliminating undue fraudulent claims will enable decreases in premium payments, rewarding high valued clients. Sherlock is the proof that technology innovation can increase customer satisfaction in new and differentiating ways.

### 3.3.2 Application of the Assessment Methodology

**Note:** This study is performed in “dry run” mode, as part of the Veritas Phase 2 initiative. Further work is required to adopt and integrate the Ethics and Accountability Framework into AXA’s internal governance and business processes.

In this section, we go through each step of the framework and show how it is applied to a fraud detection use case and the underlying thought process. We start with defining our core values, which in turn are used to define core concepts, principles and commitments required by the Framework.

Our values reflect the culture that the AXA Group's teams around the world live and express each day. **Customer first, courage, integrity and one AXA** are our core values since 2016. While One AXA is oriented towards internal alignment and collaboration, the three others are fostering a corporate culture that encourage innovation and focus on our customers and partners. All AXA employees, including leaders, are encouraged to follow these values and define the company's strategy and actions around them.

- **Customer first.** All our thinking starts with our clients. Their aspirations and challenges. Their triumphs and setbacks. And how we can continue to be relevant and impactful.
- **Integrity.** Strong ethical principles are fundamental. We trust our judgment to do the right thing for our customers, employees, stakeholders, and partners. We do not tolerate dishonest behaviours.
- **Courage.** We speak plainly and act to make things happen. We push boundaries and are emboldened to take decisive actions that add value. We rely on cutting edge technology to make our job more relevant and effective.
- **One AXA.** Diversity of profiles (in terms of background, experience, and qualifications) working together in the same direction, with the same ultimate objective. Common and shared culture.

Using our core values as foundation, we define core normative concepts, which are defined in the Ethics and Accountability Framework as individual, societal, and system level concerns that can help guide the priorities of the organisation. Core concepts help us to decompose values into more specific, closer to use case statements. We identify five core normative concepts: **fair market, efficiency, customer satisfaction, benefit sharing** and **same treatment** (please refer to the case study workbook for detailed description of core normative concepts).

The review of fraud detection use case is done through the lens of core concepts. These concepts help to design principles, directly reflecting the use case context in a descriptive manner. This is a good moment to gather inputs from AIDA solution owner and business users. Our internal governance structure – “AXA Responsible AI Circle” – supports to define roles and responsibilities ensuring that our product is designed, developed and deployed according to values and core concepts.



Below is a set of principles that covers both pillars of this exercise: ethics and accountability:

- **To fulfil the fair market concept, we need to participate in regulatory initiatives and committees.** Contribute to emerging challenges discussions, policies development and maintaining a fair insurance market. Leverage and contribute to industry wide knowledge.
- **Maintain high efficiency fraud detection.** Continuous improvement of the tool to maintain high efficiency and enhance benefits/savings. Retraining of AIDA models to address novel fraud patterns.
- **Use gained efficiency to improve customer satisfaction.** Improve customer satisfaction by simplifying the claim process and speeding it up. Create a fairer, more transparent, and trustworthy relationship with AXA customers.
- **Ensure benefit sharing to further enhance customer satisfaction.** Return fair share of savings back to customers by readjusting premiums and investing into the continuous improvement of claims settlement processes and customer satisfaction.
- **For ensuring equal treatment we act for non-discrimination in technology.** Strive to create data and technology solutions that do not discriminate customers based on race, gender, and social-demographic attributes.
- **Attend to the downstream uses of datasets.** Strive to use data in ways that are consistent with the intentions and understanding of the disclosing party.
- **Our internal initiatives (based, among other, on the Responsible AI Circle mission) and current review process lead us to assume that products will be subjected to further internal and external ethical reviews and audits.** Prioritise establishing consistent, efficient, and actionable ethics review practices for new products, services, and research programmes. Consistent review practices can mitigate risk while building institutional capacity. Independent and external reviews can contribute significantly to public trust.

Last but not least, we translated the above principles to commitments. It is especially important to define commitments in an actionable and measurable way, so we can not only act on them, but track our progress on a way to fulfil them. Each commitment is prioritised with one of three priority levels: high, medium or low. Priority represents the importance of the commitment, considering its relevance to the current stage of the project.

While we defined over 10 different commitments during this exercise, we highlight only some of them in this section (you can find the full list in the workbook for our use-case).

For example, aligned to the principle of **participating in regulatory initiatives and committees**, our commitments are: **actively participate in regulators initiatives, contribute to Responsible AI research and promote Responsible AI principles within insurance industry**, which we can measure by number of relevant initiatives where AXA participates, number of talks on Responsible AI and number of papers published.

Another example is for the principle of **using gained efficiency to improve customer satisfaction**. In this case we focused on parameters that directly affect customers and identified the following commitments: **speed up claim process time, simplify claims processes and improve perceived customer satisfaction**. These can be measured by average claim process time and average number of interactions customers had with claim handler.

## 3.4 Challenges

There are multiple challenges faced in the journey to achieve responsible and fair solution. Some are resolved but still many more remain to be tackled. Here we list key ones where relevant action has already been taken or shall be addressed in the near future.

- **Responsible AI awareness.** Education on key topics related to responsible AI principles (fairness, transparency and explainability, robustness) is a priority so that workforce, clients and partners are aware of what is at stake and how to manage these issues. *For this, curating content from AI research on the topic has been developed: this led to the creation of a responsible machine learning crash course available for AXA employees, and to be launched publicly in the first quarter of 2022.*
- **Gathering feedback.** Challenges in gathering real feedback from business end users on use of the AIDA solution with respect to ethics and accountability. The users as well as team leaders tend to position themselves (and the team) and processes in a good light.
- **Tools and measures.** Providing appropriate content, tools and mitigating solutions in a large, decentralised and diverse group is another challenge. Scalability of AI risk management tools and appropriate measures for insurance use cases is yet to be achieved.
- **Evolving regulation.** Regulation is evolving at different pace in various regions and countries. This both requires agility and simplicity in tackling rules, recommendations and incentives from local regulators regarding product development and technology innovation. However, consistency with core values should always be taken into prior consideration when deploying and AI product in new environment.

## 3.5 Conclusion

The Ethics and Accountability Framework, created by Veritas consortium and MAS, provides clear guidelines that pave the way for responsible AI adoption. By identifying core normative concepts, principles, commitments, and their evaluation metrics, we align our actions to our ethical principles in a structured way, in contrast to the more intuitive and subjective way used previously. Leveraging AI Circle – AXA's governance framework developed by the Group – we were able to reach out to and gain attention of key stakeholders around the ethics and accountability topic. The Framework proved to be a complementary tool for the AI Circles. Further work is required to adopt and integrate the Framework into AXA's internal governance and business processes. We shall also explore application of the Framework for AIDA models in different verticals.

AXA strongly believes in the importance of actively contributing to the responsible AI research community. We are doing so not only by participating in the Veritas and other financial industry initiatives, but also conducting independent research in the fairness, interpretability, safety and robustness fields. This work allows us to shape the next steps for unleashing the power of AI in the most effective and responsible way.

## 3.6 E&A Worksheet – AXA Sherlock Fraud Detection

### A Overall Guidance, Workbook Instructions, and Use Case Definition

#### Operationalising Ethics and Accountability: Workbook

##### Please bring this information to the workshop:

- Existing **core values** for the organisation
- Existing **AI principles** or other published commitments
- A **use case** related to an AIDA implementation
- (Optional) Any preexisting risk evaluation rubric/scale/process, whether technology focused (e.g., model risk) or not

##### Instructions:

Proceed through this workbook in sequential order. The outputs from each page will become inputs to following pages.

**This framework can be used to hold organisations accountable and drive consistent ethical decision making across geographies. It starts from organisational values and is best applied to specific use cases.**

An early goal is to gain familiarity and comfort with the process and concepts. This will be a highly iterative and stakeholder intense process

##### Outcomes from workshop and workbook:

These materials are based on a framework for ethics and designed to take a set of organisational values and get to commitments and specifications for measuring those commitments.

At the end of the activities, each participant will be able to:

- Establish a line from values and concepts to principles, commitments, and specifications for a particular use case
- Be able to arrive at consistent decisions when values, concepts, or principles are in conflict
- Have ways to measure, communicate, and report on progress toward commitments

##### Describe the use case:

Sherlock is an advanced AIDA fraud detection model based on historical data, leveraging a mix of batch and real time technology, aimed at enhancing savings by eliminating undue fraudulent claims. The objective in the long run is to create a fairer, more transparent and trustworthy relationship with AXA customers by rewarding them with reduced premiums.

## B MAS FEAT Concepts and Principles:

Principles to Promote Fairness, Ethics, Accountability and Transparency (FEAT) in the Use of AIDA in Singapore's Financial Sector

### Normative Concept: Principle:

#### **Fairness: Justifiability**

1. Individuals or groups of individuals are not systematically disadvantaged through AIDA driven decisions unless these decisions can be justified.
2. Use of personal attributes as input factors for AIDA driven decisions is justified.

#### **Fairness: Accuracy and Bias**

3. Data and models used for AIDA driven decisions are regularly reviewed and validated for accuracy and relevance, and to minimise unintentional bias.
4. AIDA driven decisions are regularly reviewed so that models behave as designed and intended.

#### **Ethics**

5. Use of AIDA is aligned with the firm's ethical standards, values and codes of conduct.
6. AIDA driven decisions are held to at least the same ethical standards as human driven decisions.

#### **Internal Accountability**

7. Use of AIDA in AIDA driven decision making is approved by an appropriate internal authority.
8. Firms using AIDA are accountable for both internally developed and externally sourced AIDA models.
9. Firms using AIDA proactively raise management and board awareness of their use of AIDA.

#### **External Accountability**

10. Data subjects are provided with channels to enquire about, submit appeals for and request reviews of AIDA driven decisions that affect them.
11. Verified and relevant supplementary data provided by data subjects are taken into account when performing a review of AIDA driven decisions.

#### **Transparency**

12. To increase public confidence, use of AIDA is proactively disclosed to data subjects as part of general communication.
13. 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.
14. Data subjects are provided, upon request, clear explanations on the consequences that AIDA driven decisions may have on them.

## C Operationalising Ethics and Accountability: The Framework



FOUNDATIONAL → OPERATIONAL

	FOUNDATIONAL				OPERATIONAL
DEFINITION	...indicate what is desirable to promote and protect	...embody or represent foundational and organisational values	...express general norms or guidance on how to honour the concepts and realise the values	...describe specific ways to abide by the principles that will be used or implemented	...provide standards to measure the extent to which commitments have been met
EXAMPLES	Respect for the individual, integrity, customer first	Privacy, fairness, justice, autonomy, democracy, transparency, accountability	Nondiscrimination, protection from harm, inclusion, equality of access, responsibility to act with integrity	No differential consideration, presumption of eligibility, savings returned to customer	Can AI decisions be explained? How long does it take different groups to reach customer service?
GUIDANCE	The most common starting point for operationalising ethics, often grounded in the equal worth and political standing of all people	Notions, ideas, or concepts rooted in justice or fairness. They add context to the values and establish a bridge to define principles	Guardrails for what ought and ought-not be done to implement the values and concepts, sometimes in specific use cases	Specific policies or outcomes the organisation is willing to be held accountable to and are concrete representations of values, concepts, and principles	Metrics to measure accountability to commitments over time; each commitment might have numerous specifications

### PLEASE NOTE:

'Normative' is a term-of-art for philosophers. Strictly defined, normative statements make claims about how organisations should or ought to be designed, how to value them, which things are good or bad, and which actions are right or wrong.<sup>1</sup> More colloquially, it is a **commonly understood moral behaviour for a community that is generally accepted and socially enforced by most in that community**. The term is used herein for all of these reasons, but where it is used is significant in that it underscores the highly contextual nature of ethical decision-making.

It is possible that groups within an organisation could have conflicting norms – for instance, legal may have a norm of protecting the organisation from legal risk, whereas marketing's norm is to protect the organisation from reputational risk. These different norms may yield different commitments and specifications that could conflict with each other – legal might be comfortable talking about a sensitive topic publicly, whereas marketing may prefer to not enter public debate.

Being able to navigate competing normative values to arrive at a solution that is most aligned with the entire organisation – in the local context (with many different definitions of 'local') – is what this work is all about, this is "normative" work.

## D Potential Harms from Automated Decision Making

### Articulating potential harms

Harms can happen at the scale of individuals or societies. It's often helpful to name individuals or groups that could be harmed. Below is a map of potential harms described by Future of Privacy Forum. Use these to describe potential harms on the following pages.

Individual Harms		Collective/ Societal Harms
Illegal	Unfair	
Loss of Opportunity		
Employment Discrimination		Differential Access to Job Opportunities
e.g., filtering job candidates by race or genetic/health information	e.g., filtering candidates by work proximity leads to excluding minorities	
Insurance & Social Benefit Discrimination		Differential Access to Insurance & Benefits
e.g., Higher termination rate for benefit eligibility by religious group	e.g., Increasing auto insurance prices for night-shift workers	
Housing Discrimination		Differential Access to Housing
e.g., Landlords relies on search results suggesting criminal history by race	e.g., Matching algorithm less likely to provide suitable housing for minorities	
Education Discrimination		Differential Access to Education
e.g., Denial of opportunity for a student in a certain ability category	e.g., Presenting only ads on for-profit colleagues to low-income individuals	
Economic Loss		
Credit Discrimination		Differential Access to Credit
e.g., Denying credit to all residents in specified neighbourhoods ("redlining")	e.g., Not presenting certain credit offers to members of certain groups	
Differential Pricing of Goods and Services		Differential Access to Goods and Services
e.g., Raising online prices based on membership in a protected class	e.g., Presenting product discounts based on "ethnic affinity"	
Narrowing of Choice		Narrowing of Choice for Groups
e.g., Presenting ads based solely on past "clicks"		

Individual Harms		Collective/ Societal Harms
Illegal	Unfair	

### Social Detriment

#### Network Bubbles

e.g., Varied exposure to opportunity or evaluation based on “who you know”

#### Filter Bubbles

e.g., Algorithms that promote only familiar news and information

#### Dignitary Harms

e.g., Emotional distress due to bias or a decision based on incorrect data

#### Stereotype Reinforcement

e.g., Assumption that computed decisions are inherently unbiased

#### Constraints of Bias

e.g., Constrained conceptions of career prospects based on search results

#### Confirmation Bias

e.g., All-male image search results for “CEO”, all-female results for “teacher”

### Loss of Liberty

#### Constraints of Suspicion

e.g., Emotional, dignitary, and social impacts of increased surveillance

#### Increased Surveillance

e.g., Use of “predictive policing” to police minority neighbourhoods more

#### Individual incarceration

e.g., Use of “recidivism scores” to determine prison sentence length (legal status uncertain)

#### Disproportionate Incarceration

e.g., Incarceration of groups at higher rates based on historic policing data

## E Describing Values and Core Concepts

### E.1 Defining Organisational Values

Values indicate what stakeholders care about and want to protect and promote.

1. Fill in existing core values, describing them and any potential for harm within the defined use case. When in doubt, use the example provided as a model.

Organisational Value:	Description/ Notes:	Potential Harm(s):
<b>Customer first</b>	All our thinking starts with our clients. Their aspirations and challenges. Their triumphs and setbacks. And how we can continue to be relevant and impactful.	Not staying relevant in long run. Price ineffectiveness because of high level of fraudulent claims and losing our customers to competitors.
<b>Integrity</b>	Strong ethical principles are fundamental. We trust our judgment to do the right thing for our customers, employees, stakeholders and partners. We do not tolerate dishonest behaviours.	Lose our way by pursuing short term priorities.
<b>Courage</b>	We speak plainly and act to make things happen. We push boundaries and are emboldened to take decisive actions that add value. We rely on cutting edge technology to make our job more relevant and effective.	Not being able to implement required innovations / changes and thus failing to enhance savings (and pass it to customers).
<b>One AXA</b>	Diversity of profiles (in terms of background, experience and qualifications) working together in the same direction, with the same ultimate objective. Common and shared culture.	Loss of opportunity owing to inefficient collaboration. Duplication of efforts, waste of resources leading to economic loss.

## E Describing Values and Core Concepts

### E.2 (Optional) Choosing Normative Concepts

**Normative concepts are individual-, societal-, and system-level concerns that can help to guide the priorities of the organisation.**

Below is a set of example normative concepts [of justice]. They are illustrative and may or may not be relevant for the given context. There is space on the next worksheet to define a custom set of core concepts for an organisation.

1. Select any of the core normative concepts that might be relevant or would be good for discussion.
2. Copy any core concepts that are relevant to the use case to the list on the next page.

#### Procedural:

<b>Non-discrimination</b>	A system or decision-making process should not be biased against certain groups.
<b>Equality of Opportunity</b>	Everyone should have equal/similar chances at success (e.g., educational, social, or economic).
<b>Equality of Participation</b>	People should be similarly empowered in social and political decision-making contexts and processes.
<b>Just Desserts</b>	People's social and economic outcomes should reflect their efforts and contributions.

#### Distributive:

<b>Equality of Access</b>	Everyone should be provided the same/similar access to benefits and services.
<b>Benefit Sharing</b>	Everyone who contributes to a collective endeavour should share in its benefits.
<b>Decent Outcome</b>	Everyone should have good enough (or minimally decent) social and economic outcomes.
<b>Prioritising Worst-Off</b>	Practices and policies should prioritise those who are most vulnerable, dependent, or in need.

#### Recognition

<b>Same Treatment</b>	Everyone should be treated the same regardless of the groups to which they belong.
<b>Representational Accuracy</b>	People or groups of people (and their interests) should not be mischaracterised.
<b>Inclusion</b>	People or groups of people (and their interests) should not be marginalised or excluded.
<b>Reparative Justice</b>	Past wrongful harms should be made up for and addressed so as not to create further harms or future disadvantages.

## E Describing Values and Core Concepts

### E.3 Defining Core Normative Concepts

**Core normative concepts are individual-, societal-, and system-level concerns that can help guide the priorities of the organisation.**

1. Fill in any organisational core normative concepts. When in doubt, use the example provided as a model.

Core Concept:	Description/Notes:	Potential Harm(s):
<b>Fair market</b>	Participate in governance and legal recourse to improve overall functioning of insurance market.	Unhealthy insurance market which could lead to collapse of the industry, affecting people, companies and even the country's economy.
<b>Efficiency</b>	Maintaining a high level of fraud detection with high consistency will help to improve the expense ratio and rationalise resources.	Risk of losing to competitors, overspending and an inefficient claims process.
<b>Customer satisfaction</b>	More efficient claim processing will result in easier claiming process and faster claims settlement.	Increased churn.
<b>Benefit sharing</b>	With reduced fraud/waste/abuse cases company will be able to adjust premiums.	Loss of opportunity to be "customer first" and gain competitive advantage.
<b>Same treatment</b>	Claims from any group of people are treated in a same way, no discrimination based on socio-demographic attributes of claimant.	Violation of organisational value of "integrity."

## F Reference List of Common Principles

### Choosing the Principles Set

The following is a list of most **frequently cited themes** in principles from public and private sector organisations and academia. This is not comprehensive, merely a set of common principles observed globally. It's also important to remember that many of the items organisations publish as "AI principles" might fall under "core concepts" in this worksheet.

### MAS FEAT Principles

Please reference the MAS FEAT Principles on page 66 for more inspiration. This is a great example of localisation – below, is a general set of principles that appear most frequently in large-scale surveys of existing principles. The MAS FEAT Principles are industry-, and in some contexts, geography-specific. Principles are highly context dependent and having a diversity of guiding principles is normal. Organisations will need to satisfy their own principles, those of their customers, suppliers, regulators and other stakeholders.

#### Principle:

#### Description:

**Autonomy and respect of persons**

When technologies and/or practices could impact the human condition, the potential harm to individuals and communities should be the paramount consideration. Avoid unfairly limiting an individual's possibilities.

**Attend to the downstream uses of datasets**

Strive to use data in ways that are consistent with the intentions and understanding of the disclosing party.

**Provenance of the data and analytics tools shapes the consequences of their use**

All datasets and accompanying analytics tools carry a history of human decision-making. As much as possible, that history should be auditable, including mechanisms for tracking the context of collection, methods of consent, the chain of responsibility, and assessments of quality and accuracy of the data.

**Meet and exceed privacy and security expectations**

Data subjects hold a range of expectations about the privacy and security of their data and those expectations are often context dependent. Strive to match privacy and security safeguards with privacy and security expectations.

**Be wary of collecting data just for the sake of more data**

Give due consideration to the possibility that less data may result in both better analysis and less risk

**Explain methods for analysis and marketing to data disclosers**

Maximising transparency at the point of data collection can minimise more significant risks as data travels through the data supply chain.

**Incorporate privacy, transparency, configurability, and auditability into design**

Not all ethical dilemmas have design solutions but being aware that design choices can have outsized downstream implications is profoundly important. Ethics and accountability is a strategy, product, and engineering challenge that requires widespread stakeholder engagement as early as possible.

[CONTINUED ON THE NEXT PAGE]

## Principle:

## Description:

**Assume products will be subjected to internal and external ethical reviews and audits**

Prioritise establishing consistent, efficient, and actionable ethics review practices for new products, services, and research programs. Consistent review practices can mitigate risk while building institutional capacity. Independent and external reviews can contribute significantly to public trust.

**Accountability through commitments, specifications, and governance**

Organisations signal how their values show up in their products/services by making commitments and specifying how those commitments will be measured. This is how in/external governance mechanisms – governance assesses whether norms are satisfied in a particular case – can hold organisations accountable.

**Inclusivity, solidarity, and non-discrimination in Technology**

While everyone deserves the social and economic benefits of data, not everyone is equally impacted by the processes of data collection, correlation, and prediction. Data professionals should strive to mitigate the disparate impacts of their products and listen to the concerns of affected communities.

**Internal Diversity and Non-Discrimination**

Strive to create an internal culture and set of hiring practices where people with different backgrounds and experiences can thrive professionally and personally.

**Sustainability and the Environment**

As much as possible, protect the basic preconditions for life on our planet, continued prospering for mankind, and the preservation and restoration of a thriving environment for future generations.

**Stewardship, awareness and education**

Contribute to the knowledge and furtherance of ethical leadership

**Protection of whistleblowers**

Conscientious objectors, employee organising, and ethical whistleblowers should be protected as a force for accountability and ethical decision making.

**Hidden costs and externalities**

Take potential hidden costs seriously. Including underpaid and unrecognised workers and potential misuse of work in downstream applications.

## G Designing the Principles

### Connecting Organisational Values with Core Concepts and Principles

#### Bringing Values, Concepts, and Principles Together

1. List the values and core concepts that are relevant to the Use Case.

##### Organisational Values

Customer first

Integrity

Courage

One AXA

##### Core Normative Concepts

Fair market

Efficiency

Customer satisfaction

Benefit sharing

Same treatment

#### Principles are guardrails that describe how to honour the concepts and implement the values.

1. Use the lines below to list existing principles and/or ideas for new ones. Use the principles on pages 66 and 73-74 for inspiration.

##### Principle Statement:

##### Description/Notes:

**Participate in regulatory initiatives and committees**

Contribute to emerging challenges discussions, policies development and maintaining fair insurance market. Leverage and contribute to industry-wide knowledge.

**Maintain high efficiency of fraud detection**

Continuous improvement of the tool to maintain high efficiency and enhance benefits/savings. Retraining of AIDA models to address novel fraud patterns.

**Use gained efficiency to improve customer satisfaction**

Improve customer satisfaction by simplifying claim process and speeding it up. Create a fairer, more transparent, and trustworthy relationship with AXA customers.

**Give back to customers their fair share of benefits**

Return a fair share of savings back to customers by readjusting premiums and investing in the continuous improvement of claims settlement processes and customer satisfaction.

## G Designing the Principles (continued)

### Guardrails that Describe how to Honour the Core Concepts and Implement the Values

Principle Statement:	Description/Notes:
<b>Non-discrimination in technology</b>	Strive to create data and technology solutions that do not discriminate customers based on race, gender, and social-demographic attributes.
<b>Attend to the downstream uses of datasets</b>	Strive to use data in ways that are consistent with the intentions and understanding of the disclosing party.
<b>Assume products will be subjected to internal and external ethical reviews and audits</b>	Prioritise establishing consistent, efficient, and actionable ethics review practices for new products, services, and research programmes. Consistent review practices can mitigate risk while building institutional capacity. Independent and external reviews can contribute significantly to public trust.

## H Identifying Commitments and Measuring Accountability

### Commitments, with Meaningful Specifications, are the Foundation of Accountability

**Principles:**

Norms or guard-rails that describe how to honour the concepts and implement the values

**Commitment:**

A professed obligation that explains how a principle will be implemented in a specific context

**Specification:**

A set of quantifiable assessments, or metrics, that can account for a commitment being met

**Commitments are individual-, societal-, and system-level promises an organisation makes to its customers and other stakeholders that are informed by values, core concepts, and principles.**

These pages are to define and/or brainstorm a set of possible commitments and specifications.

1. Copy a single principle to each box.
2. Brainstorm possible commitments for each principle in the context of the use case. Commitments should be in the context and furtherance of fulfilling product requirements while living the principles. Commitments should be phrased in a positive manner as something that will happen, as opposed to a negative statement – we can neither measure nor prove a negative.
3. Brainstorm possible specifications (ways to measure) for each commitment. Specifications should be measurable over time to determine progress toward a commitment.
4. Add a priority – low, medium, or high – to each commitment. Prioritisation scores serve to prioritise resources and optimise for higher priority commitments when outcomes might be in conflict.



## H Identifying Commitments and Measuring Accountability

Commitments, with Meaningful Specifications, are the Foundation of Accountability

Priority	Commitment	How to Measure (Specifications)
	Participate in regulatory initiatives and committees	Number of initiatives AXA is part of
H	Actively participate in regulatory initiatives	% of caught fraudulent cases with a help of fraud bureaus
H	Contribute to Responsible AI research	Number of published papers in the field of Responsible AI
M	Promote Responsible AI principles within insurance industry`	Number of talks in industry events Number of industry collaborations
	Maintain high efficiency of fraud detection	Loss ratio
M	Improve loss ratio	Saved money
H	Maintain fraud detection efficiency	% of fraudulent claims Average time spent by claim handler per claim % of appealed fraud findings and proportion of them highlighted by AIDA system
H	Reveal new fraud schemas	Number of new fraud schemas revealed Total number of fraud schemas being tracked

Priority	Commitment	How to Measure (Specifications)
----------	------------	---------------------------------

Use gained efficiency to improve customer satisfaction

**M**

Speed up claim process time

Average claim process time

**L**

Simplify claims process

Average number of client's interactions with claim handler / customer service

Average time spent for processing and payment of a claim

Average number of supporting documents for each claim

**M**

Improve perceived customer satisfaction

Average satisfaction score

Give back to customers fair share of benefits

**M**

Readjust premiums

Premium / loss ratio

**L**

Invest saved money in customer satisfaction and further tool improvement

% of saved money invested in customer satisfaction projects

% of saved money invested in tool improvement

Non-discrimination

**H**

Do not discriminate customers based on social-demographic attributes

Relevant fairness metric

% of appeals that claim unfair treatment

## J Identifying Commitments and Measuring Accountability

### Stakeholders and Accountability to Commitments

**Individuals, committees, organisations, and governments can be stakeholders in accountability**

These pages help to define a basic stakeholder map.

1. Copy the commitments and specifications from the previous page(s).
2. Define an individual owner for each specification. *This should be a person or a committee of known individuals.*
3. Define the stakeholders for each specification. This is who the owners will need to be accountable to. *Stakeholders can be in/external individuals, committees, public or private organisations, government entities, or any party that stands to experience benefit or harm from a commitment.*

**Please note: Due to privacy, confidentiality, and trade secrets concerns, the content offered in this section is intentionally anonymized and generalized. Organizations will want to name individuals, roles, and committees or other governance bodies in internal documents.**

Commitment	Specification(s)	Owner(s)	Stakeholders
Simplify claims process	Average number of client's interactions with claim handler / customer service	Head of Claims	Claim handlers, Customer service staff
	Average time spent for processing and payment of a claim	Head of Claims	Claim handlers, Customer service staff, Payments
	Average number of supporting documents for each claim	Head of Claims	Claim handlers, Customer service staff
Actively participate in regulatory initiatives	Number of regulatory initiatives AXA is part of	Group Public Affairs and Chief Data Scientist	AI Governance Lead and Research Data Scientists

*Note: Illustrations are provided for selected commitments*

# 04 E&A Assessment in Retail Marketing

## 4.1 Preface

Many banks are investing in the key area of AIDA to help their customers understand and reach their financial goals. Using AIDA, it is possible for banks to gain insights to the financial needs and preferences of their customers by learning from their transaction behaviour and to thereby anticipate who might be interested in or benefit from specific financial products. This case study explores the journey of a financial institution in assessing ethics and establishing accountability when using AIDA in customer marketing.

## 4.2 Introduction

At UOB, we are committed to providing our customers with progressive solutions that help them to achieve their financial goals across their different life stages and changing priorities. To do this, we focus on deepening our understanding of their changing needs and preferences, addressing their concerns and always doing what is right for them.

We believe in responsible innovation. It is important that we uphold the highest standards in safeguarding and using data appropriately for our customers. At UOB, we have established a data ethics governance model and set up a multidisciplinary data ethics task force to develop policies and processes to ensure the responsible and ethical use of data.

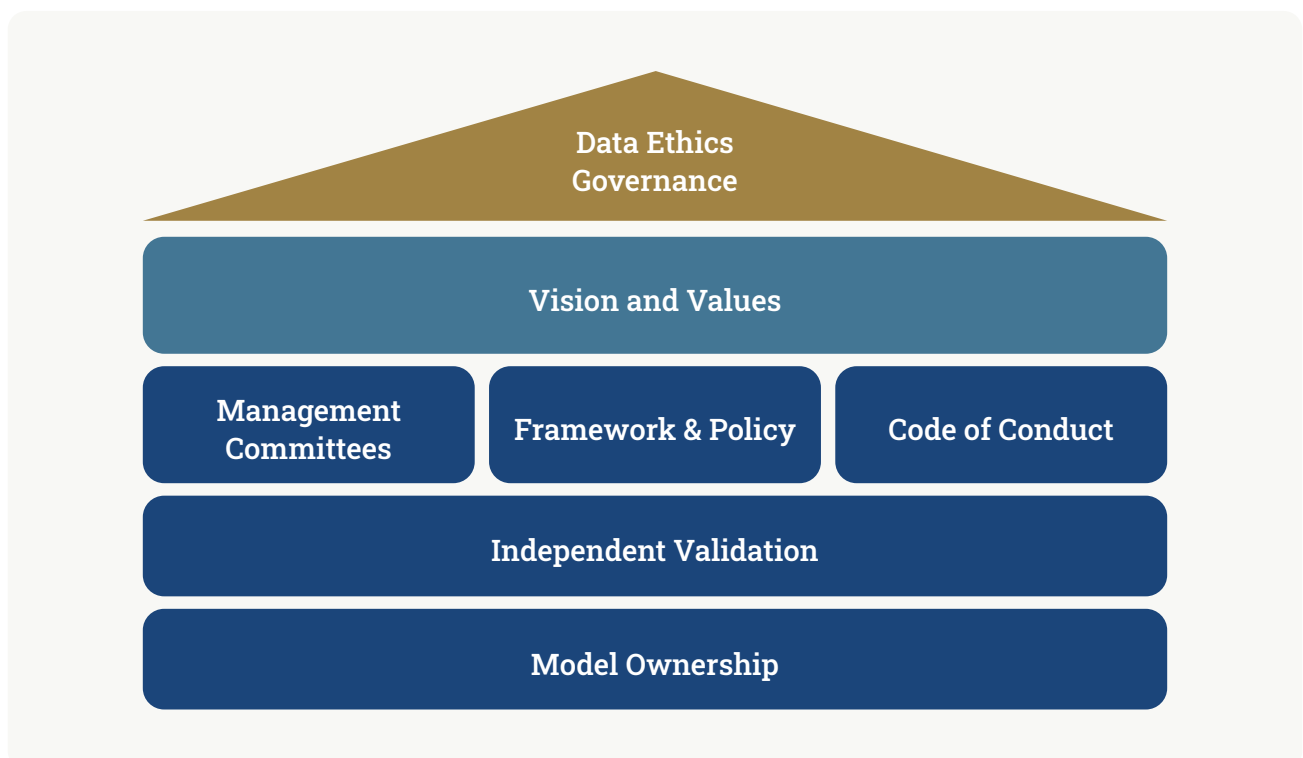


Figure 4.1: Data Ethics Governance Model

## UOB Data Ethics Taskforce

A Multi-disciplinary Task Force to look into the alignment of AIDA solutions with the FEAT principles



Figure 4.2: Data Ethics Taskforce

UOB is also one of the lead members of the Veritas Consortium which aims to create a framework for financial institutions to promote the responsible adoption of AIDA. In Phase 1 of the Veritas initiative, UOB partnered with Element AI to develop the assessment methodology for fairness on a credit risk scoring use case.

In the second phase, UOB partnered with Accenture to develop an assessment methodology for ethics and accountability and to apply it to a customer marketing model and its business processes. The customer marketing model was developed by UOB's Retail Business Analytics team to ensure that our marketing campaigns are robust and effective in helping us understand customers' needs and recommending suitable products and services to serve these needs.

## 4.3 Learning from Application of the Assessment Methodology

In this section of the document, we share the findings and observations from applying the Ethics and Accountability assessment methodology on our customer marketing model.

### 4.3.1 System Objectives and Context

The use case for the Veritas Project is a customer marketing model developed by UOB's Retail Business Analytics team to identify customers' financial needs and to match their needs with relevant banking products, such as insurance, deposit or investment solutions in a timely manner. The model is used to generate deeper insights on the varying financial priorities and needs of the bank's retail banking customers across different demographics and life stages.

The use of AIDA in customer marketing is guided by the bank's values, our code of conduct and our commitment to fair dealing.

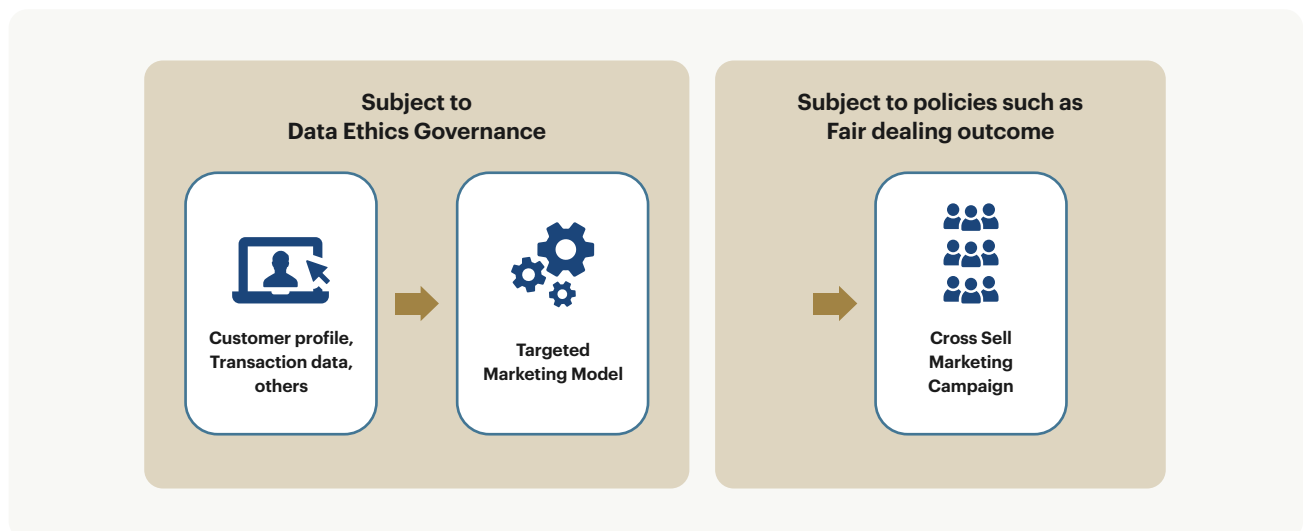


Figure 4.3: Customer marketing model

### 4.3.2 Application of the Assessment Methodology

UOB's core values – **Honourable, Enterprising, United and Committed** – guide our day-to-day decisions and actions and shape our interactions with our colleagues and our customers. They are at the core of the UOB code of conduct, which sets out the principles of personal and professional conduct expected of all employees. These core values also serve as a starting point for us to define the concepts, principles, commitments and specifications for this use case.

For example, the core value of Honourable underpins our commitment to ensure fair dealing at work, regardless of our individual roles, by putting our customers and their financial goals first. By doing so, we aim to achieve the five fair dealing outcomes established by the Monetary Authority of Singapore:

- **Outcome one.** Customers have confidence that they deal with financial institutions where fair dealing is central to the corporate culture.
- **Outcome two.** Financial institutions offer products and services that are suitable for their target customer segments.
- **Outcome three.** Financial institutions have competent representatives who provide customers with quality advice and appropriate recommendations.
- **Outcome four.** Customers receive clear, relevant and timely information to make informed financial decisions.
- **Outcome five.** Financial institutions handle customer complaints in an independent, effective and prompt manner.

These principles are integral to the design and development of the customer marketing model and campaigns, ensuring that the use of AIDA is aligned with the bank's ethical standards and core values. Guided by these principles, we also adopt the "human-in-the-loop" approach, which ensures that human input always forms a part of our AIDA driven decisions. With the understanding of these principles, we proceeded further to identify the commitments specific to the use case and to define the specifications where quantitative measurement can be made.

**Please refer to the workbook for more details.**

In the application of the assessment methodology on the customer marketing model, we are well aligned with the guiding principles of accountability and ethics. As an established bank with 86 years of experience operating in the region, we have governance and processes in place to ensure that all AIDA driven solutions in use comply with the highest ethical standards. For example, it is mandatory for all employees to complete an annual training to attest their knowledge and understanding of the fair dealing guidelines. This ensures that our employees will always do what is right for the customer and that AIDA driven decisions are made in the customers' best interest. The bank also has processes in place to manage all non-compliance instances of fair dealing in an independent, effective and prompt manner and to ensure that complaints are addressed within the established procedures and timelines.

## 4.4 Challenges

In the use case analysis, the assessment methodology led us to examine the core values to establish a comprehensive list of commitments and quantifiable metrics that account for these commitments being met. As core values describe the fundamental beliefs and practices by which a company abides and are usually qualitative in nature, it may pose a challenge to define a measurable specification. For example, our core value of United shapes the way we create an inclusive culture in which people's voices and opinions are heard and considered. While we have determined a quantifiable way to measure how best we meet our commitment of creating an inclusive culture, it will likely need to be refined over time and as it is applied to new use cases.

## 4.5 Conclusion

Using data in an ethical manner to serve our customers better is the responsible and sustainable way to do business. UOB has put in place robust standards, processes and policies to ensure the ethical governance and use of data. Through the Veritas Project, we are able to contribute to the development of the assessment methodology for ethics and accountability of 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.

## 4.6 E&A Worksheet – UOB Retail Marketing

### A Overall Guidance, Workbook Instructions, and Use Case Definition

#### Operationalising Ethics and Accountability: Workbook

##### Please bring this information to the workshop:

- Existing **core values** for the organisation
- Existing **AI principles** or other published commitments
- A **use case** related to an AIDA implementation
- (Optional) Any preexisting risk evaluation rubric/scale/process, whether technology focused (e.g., model risk) or not

##### Instructions:

Proceed through this workbook in sequential order. The outputs from each page will become inputs to following pages.

**This framework can be used to hold organisations accountable and drive consistent ethical decision-making across geographies. It starts from organisational values and is best applied to specific use cases.**

An early goal is to gain familiarity and comfort with the process and concepts. This will be a highly iterative and stakeholder intense process.

##### Outcomes from workshop and workbook:

These materials are based on a framework for ethics and designed to take a set of organisational values and get to commitments and specifications for measuring those commitments.

At the end of the activities, each participant will be able to:

- Establish a line from values and concepts to principles, commitments, and specifications for a particular use case
- Be able to arrive at consistent decisions when values, concepts, or principles are in conflict
- Have ways to measure, communicate, and report on progress toward commitments

##### Describe the use case:

An AIDA system in Retail Marketing with an objective to identify customers' financial needs and to match their needs with relevant banking products, such as insurance, deposit or investment solutions in a timely manner.

## B MAS FEAT Concepts and Principles:

Principles to Promote Fairness, Ethics, Accountability and Transparency (FEAT) in the Use of AIDA in Singapore's Financial Sector

### Normative Concept: Principle:

#### **Fairness: Justifiability**

1. Individuals or groups of individuals are not systematically disadvantaged through AIDA driven decisions unless these decisions can be justified.
2. Use of personal attributes as input factors for AIDA driven decisions is justified.

#### **Fairness: Accuracy and Bias**

3. Data and models used for AIDA driven decisions are regularly reviewed and validated for accuracy and relevance, and to minimise unintentional bias.
4. AIDA driven decisions are regularly reviewed so that models behave as designed and intended.

#### **Ethics**

5. Use of AIDA is aligned with the firm's ethical standards, values and codes of conduct.
6. AIDA driven decisions are held to at least the same ethical standards as human driven decisions.

#### **Internal Accountability**

7. Use of AIDA in AIDA driven decision making is approved by an appropriate internal authority.
8. Firms using AIDA are accountable for both internally developed and externally sourced AIDA models.
9. Firms using AIDA proactively raise management and board awareness of their use of AIDA.

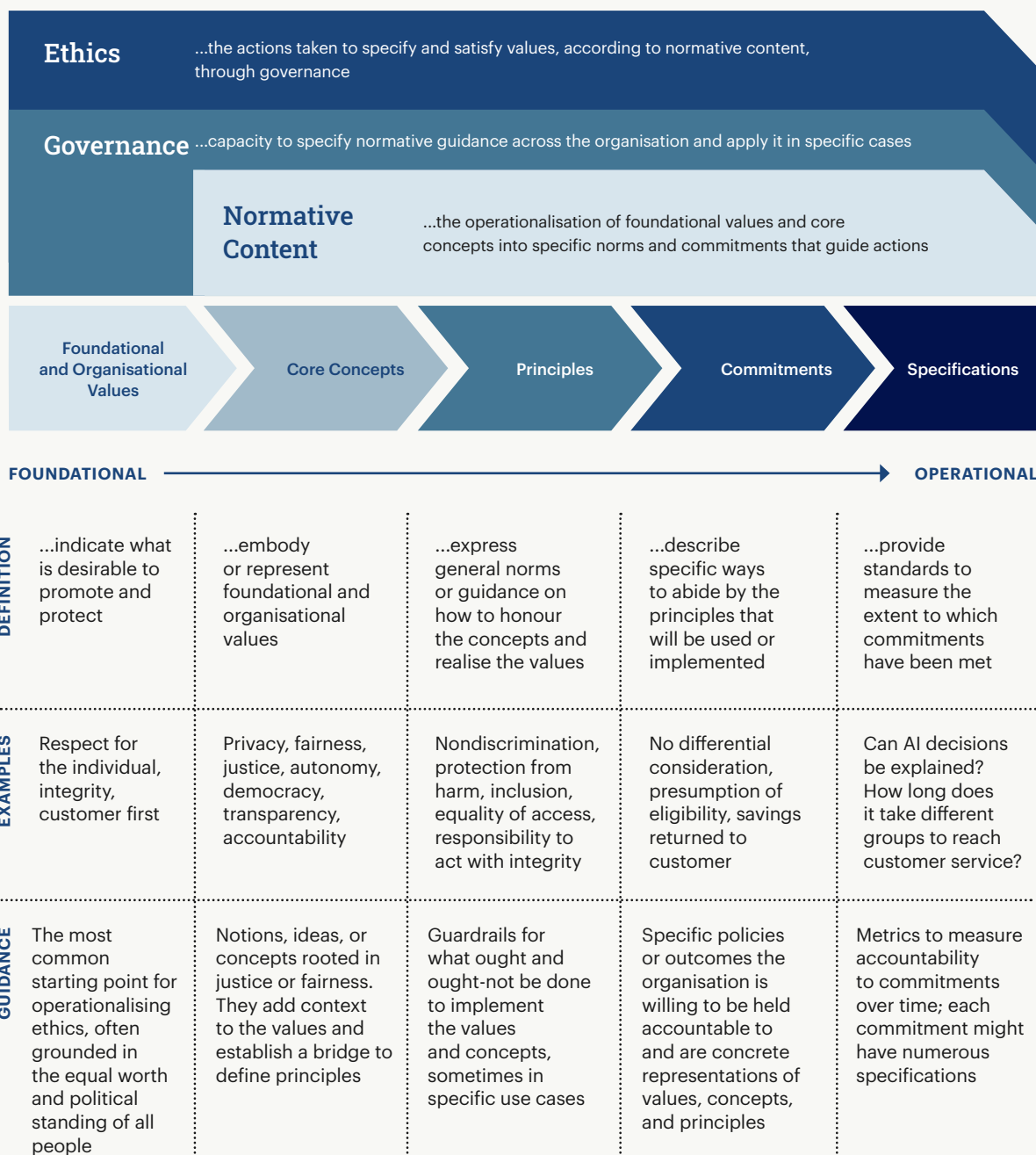
#### **External Accountability**

10. Data subjects are provided with channels to enquire about, submit appeals for and request reviews of AIDA driven decisions that affect them.
11. Verified and relevant supplementary data provided by data subjects are taken into account when performing a review of AIDA driven decisions.

#### **Transparency**

12. To increase public confidence, use of AIDA is proactively disclosed to data subjects as part of general communication.
13. 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.
14. Data subjects are provided, upon request, clear explanations on the consequences that AIDA driven decisions may have on them.

## C Operationalising Ethics and Accountability: The Framework



### PLEASE NOTE:

'Normative' is a term-of-art for philosophers. Strictly defined, normative statements make claims about how organisations should or ought to be designed, how to value them, which things are good or bad, and which actions are right or wrong.<sup>1</sup> More colloquially, it is a **commonly understood moral behaviour for a community that is generally accepted and socially enforced by most in that community**. The term is used herein for all of these reasons, but where it is used is significant in that it underscores the highly contextual nature of ethical decision-making.

It is possible that groups within an organisation could have conflicting norms – for instance, legal may have a norm of protecting the organisation from legal risk, whereas marketing's norm is to protect the organisation from reputational risk. These different norms may yield different commitments and specifications that could conflict with each other – legal might be comfortable talking about a sensitive topic publicly, whereas marketing may prefer to not enter public debate.

Being able to navigate competing normative values to arrive at a solution that is most aligned with the entire organisation – in the local context (with many different definitions of 'local') – is what this work is all about, this is "normative" work.

## D Potential Harms from Automated Decision Making

### Articulating potential harms

Harms can happen at the scale of individuals or societies. It's often helpful to name individuals or groups that could be harmed. Below is a map of potential harms described by Future of Privacy Forum. Use these to describe potential harms on the following pages.

Individual Harms		Collective/ Societal Harms
Illegal	Unfair	
Loss of Opportunity		
Employment Discrimination		Differential Access to Job Opportunities
e.g., filtering job candidates by race or genetic/health information	e.g., filtering candidates by work proximity leads to excluding minorities	
Insurance & Social Benefit Discrimination		Differential Access to Insurance & Benefits
e.g., Higher termination rate for benefit eligibility by religious group	e.g., Increasing auto insurance prices for night-shift workers	
Housing Discrimination		Differential Access to Housing
e.g., Landlords relies on search results suggesting criminal history by race	e.g., Matching algorithm less likely to provide suitable housing for minorities	
Education Discrimination		Differential Access to Education
e.g., Denial of opportunity for a student in a certain ability category	e.g., Presenting only ads on for-profit colleagues to low-income individuals	
Economic Loss		
Credit Discrimination		Differential Access to Credit
e.g., Denying credit to all residents in specified neighbourhoods ("redlining")	e.g., Not presenting certain credit offers to members of certain groups	
Differential Pricing of Goods and Services		Differential Access to Goods and Services
e.g., Raising online prices based on membership in a protected class	e.g., Presenting product discounts based on "ethnic affinity"	
Narrowing of Choice		Narrowing of Choice for Groups
e.g., Presenting ads based solely on past "clicks"		

Individual Harms		Collective/ Societal Harms
Illegal	Unfair	

### Social Detriment

#### Network Bubbles

e.g., Varied exposure to opportunity or evaluation based on “who you know”

#### Filter Bubbles

e.g., Algorithms that promote only familiar news and information

#### Dignitary Harms

e.g., Emotional distress due to bias or a decision based on incorrect data

#### Stereotype Reinforcement

e.g., Assumption that computed decisions are inherently unbiased

#### Constraints of Bias

e.g., Constrained conceptions of career prospects based on search results

#### Confirmation Bias

e.g., All-male image search results for “CEO”, all-female results for “teacher”

### Loss of Liberty

#### Constraints of Suspicion

e.g., Emotional, dignitary, and social impacts of increased surveillance

#### Increased Surveillance

e.g., Use of “predictive policing” to police minority neighbourhoods more

#### Individual incarceration

e.g., Use of “recidivism scores” to determine prison sentence length (legal status uncertain)

#### Disproportionate Incarceration

e.g., Incarceration of groups at higher rates based on historic policing data

## E Describing Values and Core Concepts

### E.1 Defining Organisational Values

Values indicate what stakeholders care about and want to protect and promote.

1. Fill in existing core values, describing them and any potential for harm within the defined use case. When in doubt, use the example provided as a model.

Organisational Value:	Description/ Notes:	Potential Harm(s):
<b>Honourable</b>	We act prudently to fuel our customers' success. We maintain the highest professional and moral standards in all our dealings – with our customers and with each other.	<p>Discrimination and offering of unsuitable product and services may occur.</p> <p>It may result in reputational damage and loss of opportunity if we do not treat our customers fairly.</p>
<b>Enterprising</b>	We were built with an enterprising spirit. We demonstrate this today through thought leadership, keen insight and a forward looking mindset.	<p>Offering of unsuitable product and services may occur.</p> <p>It may result in low customer engagement and loss of opportunity if we are not able to provide the customers with the right solutions.</p>
<b>United</b>	We work as a team. Every one of us is united to reach individual and corporate goals through cooperation, mutual respect and loyalty.	<p>Poor customer service and lack of accountability may occur.</p> <p>It may result in loss of customer trust and loss of opportunity because of the disconnection with our customers.</p>
<b>Committed</b>	We are committed to performance. We are accountable for ensuring that UOB is a trusted source of stability, security and strength.	<p>Unethical practices and lack of accountability may occur.</p> <p>It may result in loss of customer trust if there is a lack of commitment and accountability.</p>

## E Describing Values and Core Concepts

### E.2 (Optional) Choosing Normative Concepts

**Normative concepts are individual-, societal-, and system-level concerns that can help to guide the priorities of the organisation.**

Below is a set of example normative concepts [of justice]. They are illustrative and may or may not be relevant for the given context. There is space on the next worksheet to define a custom set of core concepts for an organisation.

1. Select any of the core normative concepts that might be relevant or would be good for discussion.
2. Copy any core concepts that are relevant to the use case to the list on the next page.

#### Procedural:

<b>Non-discrimination</b>	A system or decision-making process should not be biased against certain groups.
<b>Equality of Opportunity</b>	Everyone should have equal/similar chances at success (e.g., educational, social, or economic).
<b>Equality of Participation</b>	People should be similarly empowered in social and political decision-making contexts and processes.
<b>Just Desserts</b>	People's social and economic outcomes should reflect their efforts and contributions.

#### Distributive:

<b>Equality of Access</b>	Everyone should be provided the same/similar access to benefits and services.
<b>Benefit Sharing</b>	Everyone who contributes to a collective endeavour should share in its benefits.
<b>Decent Outcome</b>	Everyone should have good enough (or minimally decent) social and economic outcomes.
<b>Prioritising Worst-Off</b>	Practices and policies should prioritise those who are most vulnerable, dependent, or in need.

#### Recognition

<b>Same Treatment</b>	Everyone should be treated the same regardless of the groups to which they belong.
<b>Representational Accuracy</b>	People or groups of people (and their interests) should not be mischaracterised.
<b>Inclusion</b>	People or groups of people (and their interests) should not be marginalised or excluded.
<b>Reparative Justice</b>	Past wrongful harms should be made up for and addressed so as not to create further harms or future disadvantages.

## E Describing Values and Core Concepts

### E.3 Defining Core Normative Concepts

**Core normative concepts are individual-, societal-, and system-level concerns that can help guide the priorities of the organisation.**

1. Fill in any organisational core normative concepts. When in doubt, use the example provided as a model.

Core Concept:	Description/Notes:	Potential Harm(s):
<b>Fair dealing</b>	Culture is built on integrity, trust and respect. Treating customers fairly.	Discrimination, offering of unsuitable product and services, reputational and legal risk, biased model results, etc.
<b>Accountability</b>	Being responsible for a particular set of outcomes, defined by commitments. In combination with recourse, it is a central component of any system of governance.	Lack of customer trust, low customer engagement, unclear ownership, etc.
<b>Trust</b>	Belief that the other party's goals are aligned with yours, there is competence to achieve the goals, and one or both parties will work toward those goals.	Lack of customer trust, low customer engagement, loss of opportunity, reputational and legal risk, etc.
<b>Human involvement</b>	AIDA driven decisions are held to at least the same ethical standard as human driven decisions.	Unethical practices, discrimination, reputational and legal risk, etc.

## F Reference List of Common Principles

### Choosing the Principles Set

The following is a list of most **frequently cited themes** in principles from public and private sector organisations and academia. This is not comprehensive, merely a set of common principles observed globally. It's also important to remember that many of the items organisations publish as "AI principles" might fall under "core concepts" in this worksheet.

### MAS FEAT Principles

Please reference the MAS FEAT Principles on page 86 for more inspiration. This is a great example of localisation – below, is a general set of principles that appear most frequently in large-scale surveys of existing principles. The MAS FEAT Principles are industry-, and in some contexts, geography-specific. Principles are highly context dependent and having a diversity of guiding principles is normal. Organisations will need to satisfy their own principles, those of their customers, suppliers, regulators and other stakeholders.

#### Principle:

#### Description:

**Autonomy and respect of persons**

When technologies and/or practices could impact the human condition, the potential harm to individuals and communities should be the paramount consideration. Avoid unfairly limiting an individual's possibilities.

**Attend to the downstream uses of datasets**

Strive to use data in ways that are consistent with the intentions and understanding of the disclosing party.

**Provenance of the data and analytics tools shapes the consequences of their use**

All datasets and accompanying analytics tools carry a history of human decision-making. As much as possible, that history should be auditable, including mechanisms for tracking the context of collection, methods of consent, the chain of responsibility, and assessments of quality and accuracy of the data.

**Meet and exceed privacy and security expectations**

Data subjects hold a range of expectations about the privacy and security of their data and those expectations are often context dependent. Strive to match privacy and security safeguards with privacy and security expectations.

**Be wary of collecting data just for the sake of more data**

Give due consideration to the possibility that less data may result in both better analysis and less risk

**Explain methods for analysis and marketing to data disclosers**

Maximising transparency at the point of data collection can minimise more significant risks as data travels through the data supply chain.

**Incorporate privacy, transparency, configurability, and auditability into design**

Not all ethical dilemmas have design solutions but being aware that design choices can have outsized downstream implications is profoundly important. Ethics and accountability is a strategy, product, and engineering challenge that requires widespread stakeholder engagement as early as possible.

[CONTINUED ON THE NEXT PAGE]

## Principle:

## Description:

**Assume products will be subjected to internal and external ethical reviews and audits**

Prioritise establishing consistent, efficient, and actionable ethics review practices for new products, services, and research programs. Consistent review practices can mitigate risk while building institutional capacity. Independent and external reviews can contribute significantly to public trust.

**Accountability through commitments, specifications, and governance**

Organisations signal how their values show up in their products/services by making commitments and specifying how those commitments will be measured. This is how in/external governance mechanisms – governance assesses whether norms are satisfied in a particular case – can hold organisations accountable.

**Inclusivity, solidarity, and non-discrimination in Technology**

While everyone deserves the social and economic benefits of data, not everyone is equally impacted by the processes of data collection, correlation, and prediction. Data professionals should strive to mitigate the disparate impacts of their products and listen to the concerns of affected communities.

**Internal Diversity and Non-Discrimination**

Strive to create an internal culture and set of hiring practices where people with different backgrounds and experiences can thrive professionally and personally.

**Sustainability and the Environment**

As much as possible, protect the basic preconditions for life on our planet, continued prospering for mankind, and the preservation and restoration of a thriving environment for future generations.

**Stewardship, awareness and education**

Contribute to the knowledge and furtherance of ethical leadership

**Protection of whistleblowers**

Conscientious objectors, employee organising, and ethical whistleblowers should be protected as a force for accountability and ethical decision making.

**Hidden costs and externalities**

Take potential hidden costs seriously. Including underpaid and unrecognised workers and potential misuse of work in downstream applications.

## G Designing the Principles

### Connecting Organisational Values with Core Concepts and Principles

#### Bringing Values, Concepts, and Principles Together

1. List the values and core concepts that are relevant to the Use Case.

##### Organisational Values

Honourable

Enterprising

United

Committed

##### Core Normative Concepts

Fair dealing

Accountability

Trust

Human involvement

#### Principles are guardrails that describe how to honour the concepts and implement the values.

1. Use the lines below to list existing principles and/or ideas for new ones. Use the principles on pages 86 and 93-94 for inspiration.

##### Principle Statement:

##### Description/Notes:

##### Inclusive culture

Strive to create an internal culture where people or groups of people have their voices heard and opinions considered.

##### Non-discrimination

While everyone deserves the social and economic benefits of data, not everyone is equally impacted by the processes of data collection, correlation, and prediction. Careful considerations should be taken to mitigate the disparate impacts of products and listen to the concerns of customers and affected communities.

##### Adhere to data privacy

Ensure personal data is used responsibly in accordance with the legislation and our ethical standards. Access to and disclosure of data are strictly on a need-to-know basis.

##### Providing confidence to customers that fair dealing is central to corporate culture

Maintain high professional and moral standards in all our dealings. Uncompromising discipline, clarity and bravery to do what is right for the customer and make decisions in the customers' best interest.

## G Designing the Principles (continued)

### Guardrails that Describe how to Honour the Core Concepts and Implement the Values

Principle Statement:	Description/Notes:
<b>Offering products and services that are suitable for target customer segments</b>	Offer customers with the solutions that work best for them.
<b>Providing customers with quality advice and appropriate recommendations</b>	Provide suggestions, appropriate recommendations and knowledgeable guidance to help customers manage their day-to-day and future financial needs.
<b>Handling customer complaints in an independent, effective and prompt manner</b>	Provide effective and prompt solving of customer complaints in a fair, effective and independent manner
<b>Providing customers with clear, relevant and timely information to make informed financial decisions</b>	Provide customers with clear and timely information for making informed decisions.
<b>AIDA decisions are held to at least the same ethical standard as human led decisions</b>	Ensure that the AIDA driven solutions must be able to maintain the highest ethical standard.
<b>Use of AIDA in AIDA driven decision making is approved by an appropriate internal authority</b>	Ensure proper management oversight and approval for the use of internally developed and externally sourced AIDA solutions.

## G Designing the Principles (continued)

### Guardrails that Describe how to Honour the Core Concepts and Implement the Values

Principle Statement:	Description/Notes:
<b>Data subjects are provided with channels to enquire about, submit appeals for and request reviews of AIDA driven decisions that affect them.</b>	<p>Comply with the ABS Code of Consumer Banking Practice in the handling of customers' queries and disputes.</p> <p>Provide channels for customers to submit their requests for a review of the bank's AIDA driven decisions that affect them.</p>
<b>Verified and relevant supplementary data provided by data subjects are taken into account when performing a review of AIDA driven decisions.</b>	<p>Provide channels for customers to update the bank formally of the change in their information in accordance with the bank's established guidelines and regulations.</p>

## H Identifying Commitments and Measuring Accountability

### Commitments, with Meaningful Specifications, are the Foundation of Accountability

**Principles:**

Norms or guard-rails that describe how to honour the concepts and implement the values

**Commitment:**

A professed obligation that explains how a principle will be implemented in a specific context

**Specification:**

A set of quantifiable assessments, or metrics, that can account for a commitment being met

**Commitments are individual-, societal-, and system-level promises an organisation makes to its customers and other stakeholders that are informed by values, core concepts, and principles.**

These pages are to define and/or brainstorm a set of possible commitments and specifications.

1. Copy a single principle to each box.
2. Brainstorm possible commitments for each principle in the context of the use case. Commitments should be in the context and furtherance of fulfilling product requirements while living the principles. Commitments should be phrased in a positive manner as something that will happen, as opposed to a negative statement – we can neither measure nor prove a negative.
3. Brainstorm possible specifications (ways to measure) for each commitment. Specifications should be measurable over time to determine progress toward a commitment.
4. Add a priority – low, medium, or high – to each commitment. Prioritisation scores serve to prioritise resources and optimise for higher priority commitments when outcomes might be in conflict.



## H Identifying Commitments and Measuring Accountability

Commitments, with Meaningful Specifications, are the Foundation of Accountability

Priority	Commitment	How to Measure (Specifications)
	Inclusive culture	Percentage of customer feedback to which the bank responded
M	People or groups of people should have their voice heard and opinion considered	Number of implemented initiatives arising from the customers' feedback
	Non-discrimination	
H	Minimal or no unintended bias for different groups	Number of documented incidents of unintended bias
H	No discrimination based on personal attributes e.g., race, colour, creed, religion, ethnicity, gender, gender identity or expression, national origin, nationality, citizenship, age, disability, marital status, culture, sexual orientation, ancestry, veteran status, socioeconomic status or any other legally protected characteristic	Percentage of personal attributes used in the model that are justified
		Percentage of personal attributes that have a fairness value above the acceptable threshold
	Adhere to data privacy (PDPA)	Number of times data breaches occurred
H	Protect the confidentiality of data/ ensure confidentiality of data in all forms	Number of data breaches affecting more than 500 customers
		Number of stakeholders affected by data breaches
	Providing confidence to customers that a fair dealing is central to corporate culture	Percentage of employees who have completed training in fair dealing
H	Strengthen customer trust	Number of substantiated customer complaints / sales transaction volume
	Offering products and services that are suitable for their target customer segments	Cross sell conversion rate
		Customer satisfaction score
H	Offer customers with the solutions that work best for them	Net promoter score
	Providing customers with quality advice and appropriate recommendations	Cross sell conversion rate
		Customer satisfaction score
H	Provide suggestions, appropriate recommendations and knowledgeable guidance	Net promoter score

# H Identifying Commitments and Measuring Accountability

Commitments, with Meaningful Specifications, are the Foundation of Accountability

Priority	Commitment	How to Measure (Specifications)
	Handling customer complaints in an independent, effective and prompt manner	Percentage of complaints that have been responded to within the SLA
H	Provide effective and prompt solving of customer complaints in a fair, effective and independent manner	Percentage of high risk customer complaints resolved within the SLA
	Providing customers with clear, relevant and timely information to make informed financial decisions	Customer satisfaction score
H	Enable customers to make timely informed decisions	Net promoter score
	AIDA decisions are held to a high ethical standard	Percentage of customer requests for review of AI decision resolved within SLA
H	Maintain the same ethical standard as human driven decisions for high materiality models	Explainability of AIDA model
	Use of AIDA in AIDA driven decision making is approved by an appropriate internal authority	Number of issues raised during management review
H	Ensure proper management oversight and approval for the use of internally developed and externally sourced AIDA solutions	Percentage of issues resolved
	Data subjects are provided with channels to enquire about, submit appeals for and request reviews of AIDA driven decisions that affect them	Percentage of request for a review of the AIDA decision resolved within SLA
M	Provide channels for customers to submit their requests for a review of bank's decisions that affect them	
	Verified and relevant supplementary data provided by data subjects are taken into account when performing a review of AIDA driven decisions	Percentage of request for a change of data resolved within SLA
M	Provide channels for customer to update the bank formally of the change in the customer's particulars	

## J Identifying Commitments and Measuring Accountability

### Stakeholders and Accountability to Commitments

**Individuals, committees, organisations, and governments can be stakeholders in accountability**

These pages help to define a basic stakeholder map.

1. Copy the commitments and specifications from the previous page(s).
2. Define an individual owner for each specification. This should be a person or a committee of known individuals.
3. Define the stakeholders for each specification. This is who the owners will need to be accountable to. *Stakeholders can be in/external individuals, committees, public or private organisations, government entities, or any party that stands to experience benefit or harm from a commitment.*

**Please note: These pages were added in response to Veritas consortium feedback and have not been reviewed by consortium members nor were they part of the industry partner use cases.**

Commitment	Specification(s)	Owner(s)	Stakeholders
Minimal or no unintended bias for different groups	Number of documented incidents of unintended bias	Retail Business Analytics	Internal stakeholders are our management committees and business teams.  External stakeholders are our retail customers.
Use of AIDA in AIDA driven decision making is approved by an appropriate internal authority	Number of issues raised during management review	Retail Business Analytics	Stakeholders are our management committees and business teams.
	Percentage of issues resolved	Retail Business Analytics	Stakeholders are our management committees and business teams.

*Note: Illustrations are provided for selected commitments*

# 05 Transparency Assessment in Credit Decisioning

## 5.1 Executive Summary

Standard Chartered (“the bank”), in partnership with HSBC and TruEra, developed a methodology for adoption of the transparency principles (“the Methodology”) from the Monetary Authority of Singapore’s FEAT Guidelines. The Methodology was then tested on an AIDA use case being implemented at the bank.

Current credit decisions are based on models that do not employ AIDA techniques. The use case chosen for the deep dive involves implementation of AIDA in a challenger mode, where a limited number of credit decisions will be made using the AIDA model.

The deep dive of the credit decision use case provided an opportunity to assess and identify areas for enhancement in the bank’s existing governance framework for the responsible use of AI and testing the Methodology, apart from assessing the use case itself. This also helped identify areas where existing business practices intending to use AIDA techniques may require further assessment to enable and support adoption of the transparency principles.

### Considerations for AIDA policies/standards

The bank has a principles based internal standard for governing the use of AIDA. The deep dive involved mapping the practices proposed in Methodology to the standard, followed by assessing the use case against the Methodology.

The exercise revealed that the following capabilities defined by the Methodology are already in place (when adopting AIDA techniques):

- Factors to determine whether customer facing transparency is essential for a use case.
- Mechanisms to establish whether proactive or reactive communication is required over the customer lifecycle as well as the artefacts and channels for the same.
- Factors to determine the extent of internal and external transparency and audiences.

In addition, the bank identified the following areas for future consideration that were defined in the Methodology but not adopted by the Group’s responsible AI standard:

- Prescribing acceptable explanation methods in line with the materiality of the use case and the nature of the underlying algorithms deployed.
- Specifying minimum accuracy standards for such explanation methods (for the use case).

These are currently not feasible due to the evolving nature and understanding of the explainable AI domain. The bank will assess these areas in the future when proven techniques are more widely adopted, and also establish where in the governance structure they may be included, taking into account trade-offs of the current principles based approach against a more prescriptive and rule-based standard.

### Considerations for current transparency related business practices

The level of transparency required for AIDA driven decisions (defined in the methodology paper) is higher than when AIDA techniques are not used. In cases where a hybrid approach combining AIDA driven systems with traditional systems is adopted, this could result in a two tier process and pose operational challenges. The frontline teams will assess these considerations, also accounting for impact to intellectual property, competitive advantage, and risks associated with “gaming” critical algorithms before implementing suitable transparency mechanisms.

The steps involved in reviewing the standard and assessing the credit decisioning use case and the how they lead to the stated considerations are discussed in this document.

## 5.2 Introduction, Purpose and Scope

This document is intended for developers, owners and assessors of AIDA systems, business users of such systems and other internal audiences including those in the governance, risk, and control functions within Financial Services Institutions. It is expected that the readers have familiarised themselves with the Methodology for implementing the transparency principle, which is the foundation for this document.

This document describes how the current practices related to the transparency principles were assessed against the Methodology for a credit decisioning use-case prior to its implementation.

The document does not comment on alignment of the AIDA use case’s systems and processes to the FEAT principles, which is left to the AIDA assessors evaluating the implementation.

While the assessment of an AIDA use case is performed against all the FEAT principles (through their mapping to the standard), this document covers only the transparency principle.

## 5.3 Use of AIDA in Credit Decisioning

The typical workflow involved in providing credit is depicted below.

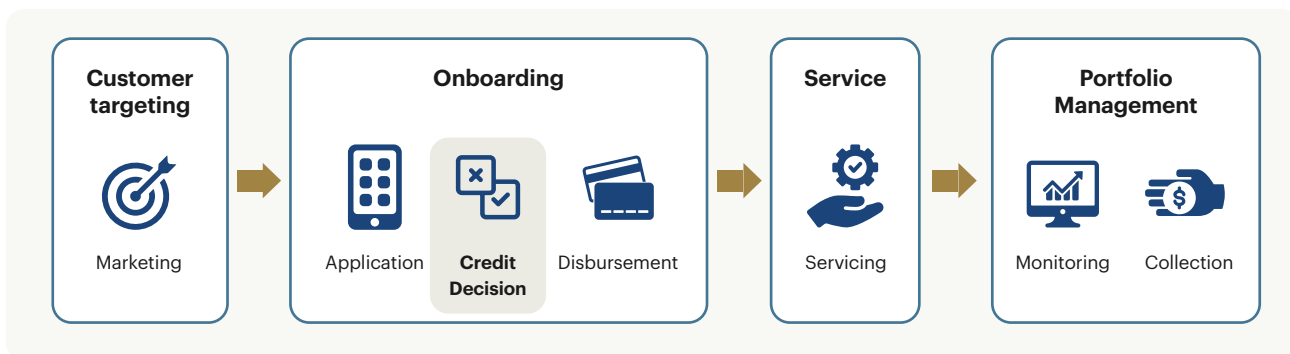


Figure 5.1: Typical workflow in credit decisioning

Credit decisioning is an integral part of the customer onboarding process and involves assessment of the customer's (or prospect's) suitability for credit. Typically, this analyses the customer's credit worthiness based on size and length of the credit along with an assessment of their ability to repay. This assessment is usually based on a credit scorecard which consists of a group of characteristics, statistically determined to be predictive in separating good and bad loans (or customers). Examples of scorecard characteristics include demographic data, credit account performance, bank transaction data and real estate data.

A credit score is the quantification of the customer's likelihood to repay a loan. Credit scores are combined with business rules (such as eligibility criteria and risk management strategies) to arrive at credit decisions. There is now an increased drive among FSIs towards automation of such quantification and decisions to scale operations for handling greater volumes with improved consistency across decisions.

Credit scoring has the potential to impact large number of "underbanked" consumers. For this reason, the use of credit scoring has increased significantly in recent years, owing to access to additional sources of data, the rise of computational power, regulatory requirements, and demand for efficiency and economic growth. Furthermore, the application of credit scoring has evolved from the traditional decision making of accepting or rejecting an application for credit to include other facets of the credit process such as the pricing of financial services to reflect the risk profile of the consumer and setting of credit limits.

Credit scoring methods are evolving from traditional statistical techniques to innovative methods using AIDA, including machine learning algorithms such as random forests, gradient boosting and deep neural networks.

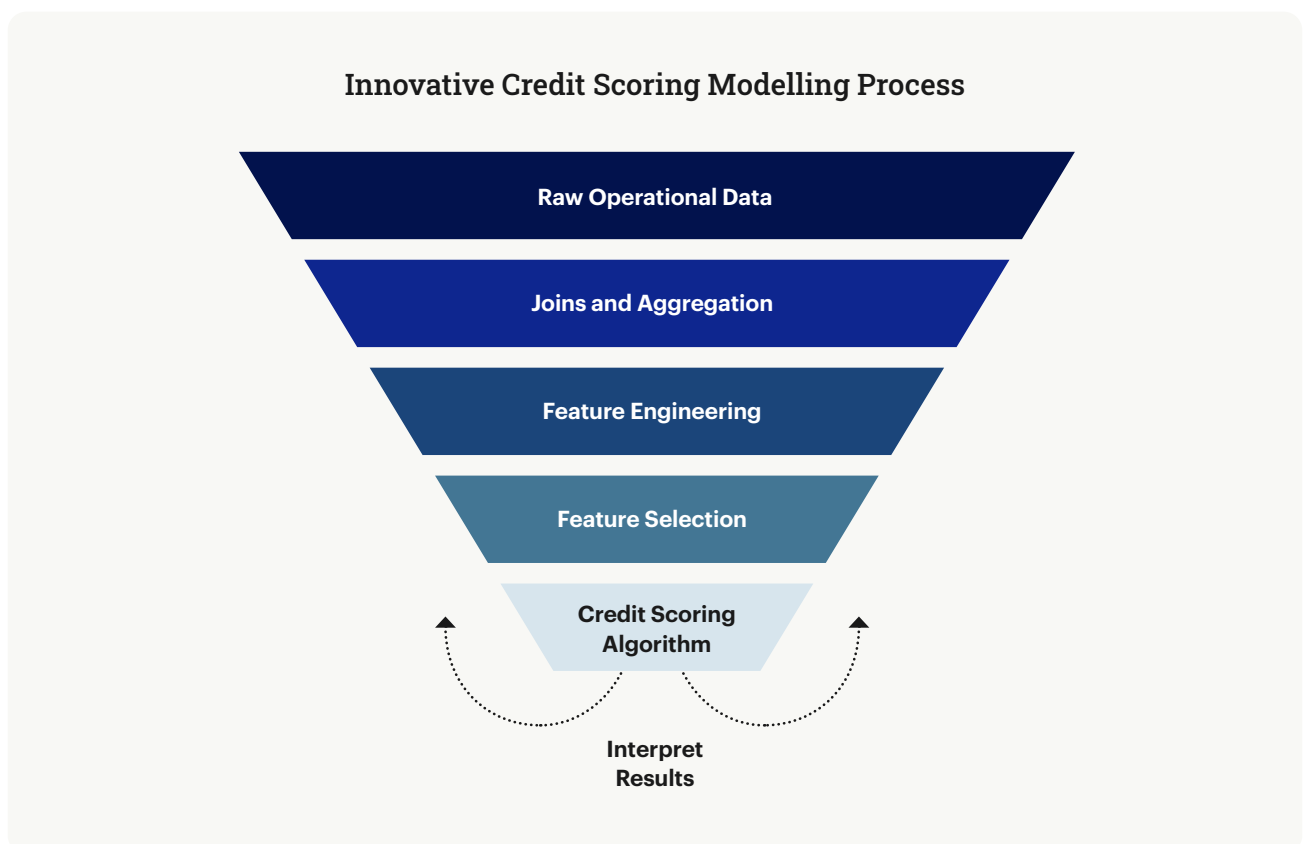


Figure 5.2: Innovative credit scoring modelling process

The adoption of alternative modelling techniques has also been necessary to handle the broadening range of data that could be considered relevant for credit scoring models and decisions. Digitalisation and the digital footprints left by consumers and businesses have caused a rapid growth in the data sources available for credit scoring, broadening the possibilities to generate insights beyond traditional data sources. Data has become a vital resource for organisations, entities, and governments. Financial institutions are now leveraging these non-traditional data sources to score consumers and businesses that have limited credit bureau information (referred to as “thin-bureau” or “thin-file” customers). Use of such alternative data may enhance the ability of financial institutions to serve customers that have difficulty accessing affordable credit within developing and developed economies due to lack of traditional data inputs from credit bureaux. Today, the data used for credit scoring come from diverse and multidimensional sources.

Increases in the diversity of data sources coupled with larger volumes necessitates the use of innovative and sophisticated methods that are relatively “opaque” when compared to traditional statistical methods. This is because unlike traditional credit scoring models, innovative methods are often viewed as challenging to interpret and explain. In addition, such innovative methods may be prone to overfitting and raise concerns about fairness.

The potential benefits and risks of using alternative data sources in credit decisions is highlighted in the DFS report on Apple Card (released March 2021)<sup>9</sup>. The report highlights that the use of alternative data can bridge gaps in traditional datasets as well extend credit to a wider section of customers, while cautioning that any such use should ensure appropriate understanding and explainability of the outcome.

Standard Chartered uses traditional models in the current credit decision making process and is evaluating and implementing an AIDA driven model using alternative data sources to work with the traditional model in a challenger mode. In the initial phase, a limited number of credit decisions will be processed using the AIDA model. This is expected to provide insights into the AIDA model's performance in comparison to the traditional model and will eventually provide the incentive to fully operationalise the AIDA driven credit decisions. At the time of writing, the AIDA model is under implementation and yet to be operationalised.

## 5.4 Overview of Transparency Assessment Methodology

This section is based on Section 3 from the Methodology whitepaper. The transparency assessment consists of a set of questions mapped to the steps in a typical AIDA lifecycle, outlined in the diagram below.

The original FEAT document outlines the Principles to be used to guide FSI's work on Transparency (Principles 12, 13 and 14). FSI's should first translate these Principles into appropriate internal standards that are not specific to individual use cases but applicable to the FSI as a whole.

Teams accountable and responsible for individual AIDA systems should then adopt and operationalise the appropriate standards during development, validation, deployment, and ongoing monitoring of their AIDA systems.

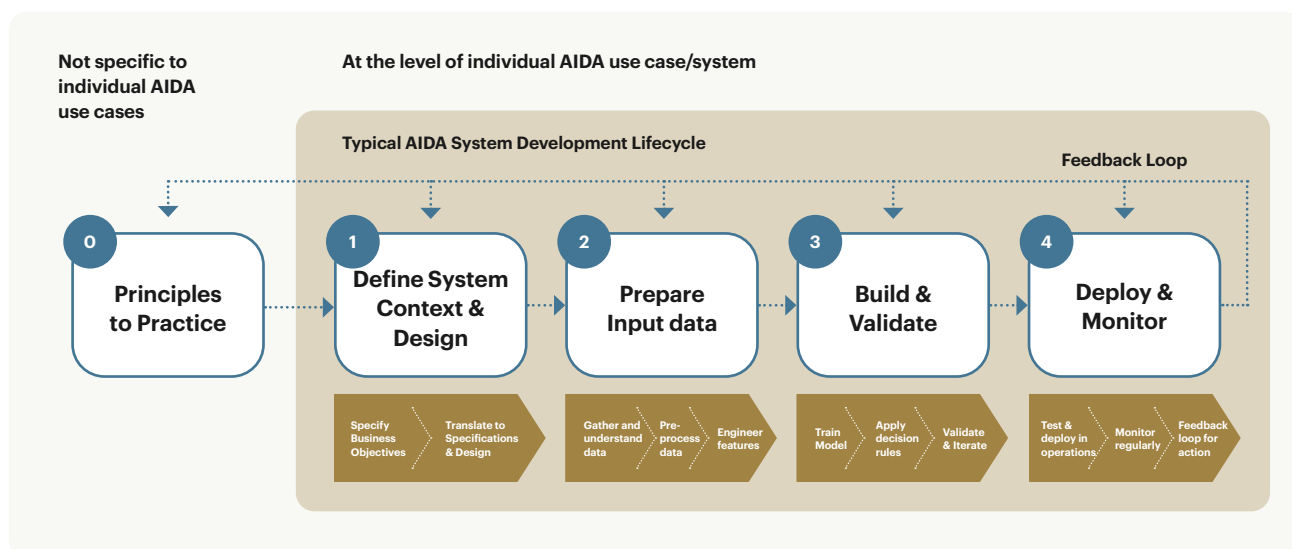


Figure 5.3: AIDA system development lifecycle

Note that for both external and internal transparency, explanations should cover the final outcome of the AIDA system after any planned human interventions have been made. In addition, for internal transparency, stakeholders may also be interested in explanations of individual ML models that form part of the AIDA system, prior to planned human interventions.

The steps involved in the typical AIDA lifecycle are listed below, and the Methodology whitepaper has the questions applicable at each step.

- Step 0: Define Principles and Standards Internally (within the bank)
- Step 1: Define System Context and Design
- Step 2: Prepare Input Data
- Step 3: Build and Validate the AIDA System
- Step 4: Deploy and Monitor AIDA System

## 5.5 Approach to Reviewing the Assessment Methodology

The Methodology was the basis for assessing the actions required to align with the FEAT Transparency principles prior to implementing the AIDA driven credit decisioning system.

The bank's standard for the responsible use of AI is mapped to the FEAT Principles and formalised under the enterprise risk management framework. The standard is principles based and supported by a governance mechanism that includes a materiality assessment and a detailed checklist which covers all the requirements in the standard.

The approach adopted for the deep dive was to assess how much of the bank's existing practices already cover the proposed methodology and what additional changes could be considered based on practicality to the standard, as well as the use case, before it was operationalised.

For each step in the Methodology, the relevant section of the standard and the supporting governance mechanisms were reviewed. This was followed by an analysis of the credit decisioning

use case to cover the standard and steps in the Methodology. The results helped identify the parts of the Methodology we may adopt to further enhance our alignment with the FEAT principles, and establish whether the enhancements are to be considered in the standard, the supporting governance process, or embedded in the AIDA development lifecycle itself.

This exercise was not a maturity assessment, considering the guidelines are not mandatory regulations currently. The following sections describe the analyses performed.

### 5.5.1 Step 0 – Internal Standards (Transparency)

The bank’s AIDA governance framework consists of the following components:

- a. AI Regulations.** AIDA standards are guided by existing regulatory expectations and the bank’s set of values. The standard incorporates MAS’ FEAT Guidelines as well as Hong Kong Monetary Authority (HKMA) requirements holistically, and is continuously reviewed against upcoming regulatory pronouncements
- b. AI Standard.** The standard establishes common requirements for all implementations of AI in the bank. A set of key principles underpin the standard, and include data suitability, fairness, ongoing monitoring, auditability, and transparency. These are mapped to the Fairness, Ethics, Accountability and Transparency principles in the FEAT guidelines.
- c. Institutional Values.** The bank’s values provide overarching guidance on how regulations and standards are to be adhered to. For example, certain practices may be undesirable according to the values even when not explicitly prohibited by regulations.
- d. AIDA Controls.** The standard is supported by a set of controls, which provide detailed requirements on how to identify, assess, build, and validate AIDA systems.
- e. AIDA Risk Monitoring.** This involves monitoring the risk across the entire lifecycle and describes how risks introduced on account of AIDA usage are identified, recorded, monitored, and remediated within the bank.

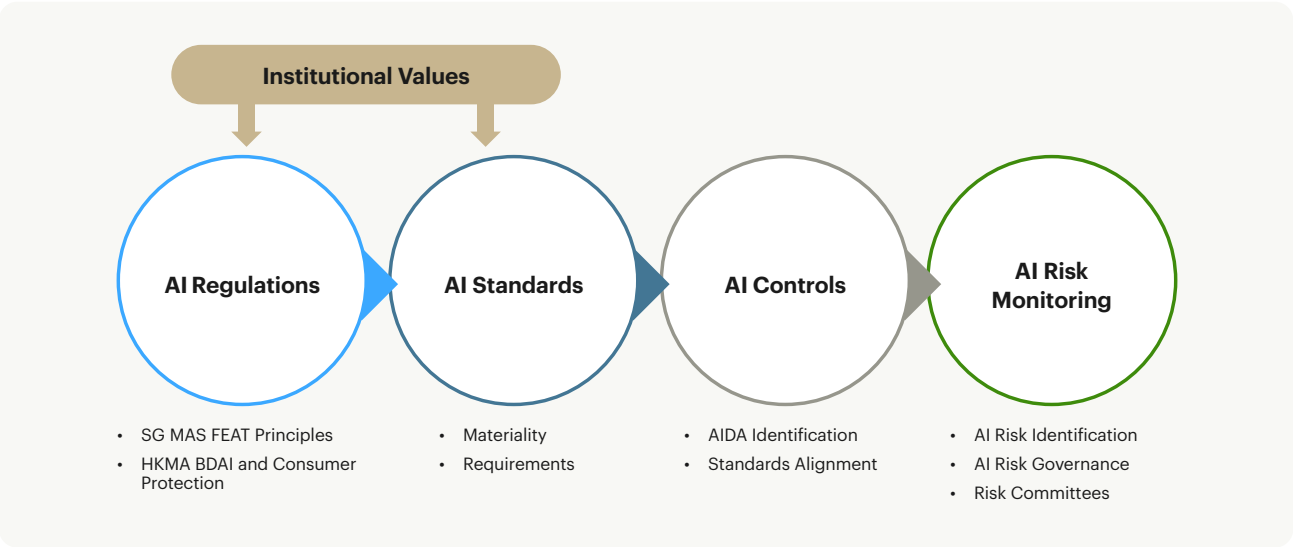


Figure 5.4: AIDA governance framework

The use of large amounts of data from a variety of sources in AIDA and the potential impact on customers makes it important to build transparency (and fairness) into the AIDA solution.

As part of Step 0, the Methodology prescribes the following five questions as a means to scope the key activities for achieving transparency:

- **(T1)** Has the FSI defined the factors it will use to determine whether external (customer facing) transparency is essential for a particular AIDA use case?
- **(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 the 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 FSI?
- **(T5)** Has the FSI set minimum accuracy standards for such explanation methods?

The questions are answered below. Each question is followed by an explanation of the objective and requirements implemented in the standard or controls. These references will be used in Step 1 to finalise the credit decision related transparency requirements.

### 1. Determine use case materiality

The materiality of the use case is determined by assessing the use case against a set of criteria and rating the overall outcome on a three point scale (low/medium/high).

**Requirement #1 in the standard**

All use cases must be assessed for the level of impact attributed to use of AIDA and classified as material or otherwise (final outcome). The following table lists the key criteria as described in the Veritas Methodology whitepaper. These are included in the AIDA controls that are in place.

Assessment Criteria	Outcome
Extent of Influence of AIDA on the final outcome	Low/Medium/High
Impact to Customers	Low/Medium/High
Reputational Impact	Low/Medium/High
Financial Impact	Low/Medium/High
Regulatory Impact	Low/Medium/High
Final Outcome	Low (Not Material)/Medium, High (Material)

Table 5.1: Materiality assessment components

## **2. (T1) Has the FSI defined the factors it will use to determine whether external (customer facing) transparency is essential for a particular AIDA use case?**

This helps establish whether external transparency is required for an AIDA use case, based on factors specified in the standard. These include assessment for unjust bias, fairness and other principles enshrined in the standard, in addition to transparency.

### **Requirement #2 in the standard**

- Transparency (both internal and external) requirements should be assessed and aligned in line with the materiality of the use case.
- Input data should be assessed for use of protected variables and their proxies. Such use is not allowed and should be justified in any exceptional cases.
- The extent of external transparency is determined based on the influence of the use case's outcome on one or more of the following factors:
  - a. Deny access to products/services.
  - b. Limit access to products/services.
  - c. Provide access to products/services that may be unsuitable for the customer.
- External transparency requirements can be reduced where the use case relates to any of the following scenarios:
  - a. Fraud detection.
  - b. Financial crime detection and compliance.
  - c. Susceptible to gaming the bank's systems or processes.
  - d. Loss of bank's intellectual property, resulting in loss of competitive advantage.

## **3. (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 the standard templates/interfaces for the same?**

This question addresses the appropriate nature and form of communication required across every stage of customer lifecycle to meet the transparency requirements.

### **Requirement #3 in the standard**

- Use of AIDA in decision making should be proactively communicated to the data subject.
- The data subject should be provided (upon their request) with clear explanations:
  - a. On the data used in the AIDA decision and how the data affects the decision.
  - b. On the consequences of the AIDA decision for them.

This process also helps identify the channels that should be used at each stage where external transparency is essential, and the content of those transparency related communications.

### **Requirement #4 in the standard**

- Appropriate channels should be made available to the data subject to receive, request for, and submit information in relation to their case. The data subject should be sufficiently made aware of these channels.

**Note:** Channels may include (but are not limited to) websites/internet banking, call centres/phone banking, sales representatives and the methods may include (but are not limited to) terms and conditions, telephonic communications, emails and face to face meetings.

#### 4. (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?

The audiences for internal transparency include AIDA users who are frontline staff engaged in customer facing interactions, AIDA validators who are involved in testing and validation of the AIDA use case, internal risk and controls users involved in the monitoring of the AIDA solution.

The following transparency requirements apply across the following audiences in line with materiality considerations.

Transparency audiences	Category	Requirements
<b>Data scientists, developers, technology teams</b>	AIDA system developer	Fairness/explainability metrics/ transparency reports including various metrics for fairness/ explainability/ etc.
<b>Frontline staff</b>	AIDA users	Transparency reports/dashboards in simple, clear language
<b>BAU owners</b>	AIDA users	Transparency reports/ dashboards including various metrics for fairness / explainability /etc.
<b>Model validation group</b>	AIDA validators	Fairness/explainability/ performance metrics
<b>Second line and governance</b>	AIDA reviewers/approvers	AI review checklist AI controls implementation

Table 5.2: Transparency Considerations



## 5. (T4) Has the FSI defined an acceptable set of AIDA ML explanation method(s) for use within the FSI?

The AIDA use case can implement one or more of the available explainability techniques based on the use case and the audience. The standard does not prescribe the technique(s) to be used for each use case due to the evolving nature (and understanding) of the explainable AI domain, and will accommodate emerging techniques once proven. Below are some of the methodologies available/under research in industry and widely discussed in the context of internal transparency.

<b>Model agnostic methods</b>	<b>Global explanation methods</b>	<ul style="list-style-type: none"> <li>Partial dependence plots - PDPs</li> <li>Accumulated local effects - ALE</li> <li>Permutation Importance</li> <li>Global surrogate models</li> </ul>
	<b>Local explanation methods</b>	<ul style="list-style-type: none"> <li>Local interpretable model-agnostic explanations (LIME)</li> <li>Local surrogate techniques like breakdown and high precision anchors</li> <li>Shapley values (popular approximation techniques include quantitative input influence (QII) and Shapley additive explanations (SHAP))</li> <li>Global surrogate models</li> </ul>
<b>Model Specific methods</b>	<b>Linear model and decision tree methods</b>	<ul style="list-style-type: none"> <li>Examining coefficients in linear models</li> <li>Generating global feature importance in decision trees by examining Gini impurities of feature splits within a tree</li> </ul>
	<b>Gradient based deep neural networks (DNN) methods</b>	<ul style="list-style-type: none"> <li>SmoothGrad saliency maps</li> <li>Guided backpropagation</li> <li>Layer wise relevance propagation</li> <li>Grad-CAM</li> <li>Integrated gradients</li> </ul>
<b>Conceptual soundness method</b>		<ul style="list-style-type: none"> <li>Influence Sensitivity plots</li> </ul>

Table 5.3: Non-exhaustive list of explanation methods

## 5. (T5) Has the FSI set minimum accuracy standards for such explanation methods?

The standard is principles based and does not prescribe minimum accuracy expectations for the model explanations. These are determined for each use case based on the type of underlying algorithm and data. Model explainability is an evolving area and it is too early to prescribe techniques or accuracy measures. This will be reconsidered in the future as the subject area evolves, and there is greater experience and feedback from using the various techniques.

The outcomes from Step 0 are summarised below

#	Checklist question	Yes/No
<b>T1</b>	Has the FSI defined the factors it will use to determine whether external (customer-facing) transparency is essential for a particular AIDA use case?	Yes
<b>T2</b>	At each stage of the FSI's customer lifecycle, has the FSI determined what proactive or reactive communication may be needed, and the standard templates/interfaces for the same?	Yes
<b>T3</b>	Has the FSI defined the factors it will use to determine the extent of, and audience for internal transparency for individual AIDA use cases?	Yes
<b>T4</b>	Has the FSI defined an acceptable set of AIDA ML explanation method(s) for use within the FSI?	For future consideration
<b>T5</b>	Has the FSI set minimum accuracy standards for such explanation methods?	For future consideration

## 5.5.2 Step 1 – Define System Context and Design

This section defines the interpretation of the Methodology and the standard for the AIDA use case. This involves evaluating materiality, and the questions in Step 0 (T1 to T5) in the context of the use case to establish coverage in current practise as well as identify areas for future consideration.

- **(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.
- **(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.
- **(T9)** Has the team selected a suitable explanation method for this specific use case from the approved list in T4?
- **(T10)** Has the team ascertained that the chosen explanation method/implementation meets the minimum accuracy requirement for this specific use case (based on T5)?

### 1. Use case materiality

Using the assessment described in Step 0 as well as the bank's internal criteria, the credit decision use case was assessed as "medium" materiality. This was in the context of the current implementation being limited to a challenger model. The materiality will be reassessed when the scope of the current implementation is expanded.

## 2. (T6) Has the AIDA use case team determined whether there is a need for external (customer facing) transparency?

Considering the medium materiality rating driven by impact to customers, both external and internal transparency were required to be established. The use case does not relate to fraud or financial Crime, and there are no grounds to consider any reduction in the requirements for external transparency requirements.

## 3. (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?

The drivers for the type of communication and information required will vary according to the stage and status of the customer journey with the bank. In a credit decisioning context, both proactive and reactive communication are relevant to establish external transparency.

Proactive communication is required at the beginning of customer engagement; this could be at the prospecting stage for a new customer or at the time of cross selling or upselling a new product or service to an existing customer. At this stage, it is important that the customer understands the use of AIDA in the decision making process related to the products and services offered and the information used in the process.

These requirements for proactive transparency are achieved through information in the product details, application forms, terms and conditions, and during face to face or telephonic interactions as part of the customer's application.

Once the customer applies for the credit product, the need for transparency around decision making arises both in case of AIDA driven and non-AIDA driven decisions. It is not practical or feasible to have different transparency standards and expectations for AIDA and non-AIDA driven decisions (for external transparency), especially in the context of a "challenger" model, as this would impose a significant overhead on frontline staff. For this reason, external transparency requirements for AIDA driven decisions will be achieved by using existing practices where applicable or by establishing a suitably enhanced common practice across both (AIDA and non-AIDA driven outcomes) where required. This decision will be taken prior to operationalising the model.

The following table provides current state view (prior to AIDA implementation) of the various stages of engagement with a customer/prospect, overlaid with the information shared as part of external transparency.

Stage	Decision Making (Pro-active)	Reject Reasons (Reactive)	Impact to Customer (Reactive)
<b>Prospect Engagement</b>	Not Applicable/Not practiced	-	-
<b>Cross Sell/Up-sell</b>	Not Applicable/Not practiced	-	-
<b>Apply for Products and Services</b>	Not Applicable/Not practiced	-	-
<b>Approval/Rejection</b>	Not Applicable/Not practiced	Practiced (Made Available on Demand)	
<b>Account Servicing</b>	Not Applicable/Not practiced	Practiced (Made Available on Demand)	
<b>Hardships</b>	-	Practiced (Made Available on Demand)	

Table 5.4: Stages of engagement with a customer / prospect

Current decision making does not use AIDA, hence this requirement on proactive communication in the context of AIDA driven decisions does not apply today. It should be also noted that as is common practice, proactive transparency has not been established in current (non-AIDA) practice with respect to the decision making process.

With regard to reactive communication, the customer is provided with all information related to their application, (except information that may be susceptible to gaming or part of the bank's intellectual property) when they request for it. Additionally, the customer is informed that a "free of charge" credit report is available from the credit bureau once their application is processed, and this practice will be continued in the future.

The bank will consider establishing common practices for proactive and reactive communications across non-AIDA and AIDA driven decisions as part of use case operationalisation.

Customer facing communication decisions involves both the channels and the forms of communication. In current practice, there are several channels already involved in the customer engagement, and these channels are not expected to change with the introduction of AIDA to the decision making process. The following channels of communication are currently available:

- 1. Application Forms.** These have extensive terms and conditions that address consent requirements. Consent is a prerequisite for application submission, and consent to the use of personal data for processing is already in place.
- 2. Websites.** Information available on the bank's websites related to the products is extensive but does not explain the current (non-AIDA) decision making process. This is to ensure the current credit policies are not public and is common practice in the industry.
- 3. Telesales.** When the products are offered and applied over the phone, the terms and conditions are read out to the applicant. As current terms and conditions do not explain the decision making process, telesales channels do not include them in scope.

Details of the decision making process are currently not communicated proactively, and a similar approach is expected to be continued. The bank may consider enhancements to existing channels in future as this has an impact on all credit decisions and is not limited to AIDA driven decisions.

Stage	Websites	Telesales	Face-to-Face
<b>Prospect Engagement</b>		Pro-active (Terms and Conditions on Application Forms)	
<b>Cross Sell/Up-sell</b>		Pro-active (Terms and Conditions on Application Forms)	
<b>Apply for Products and Services</b>		Pro-active (Terms and Conditions on Application Forms)	
<b>Approval/Rejection</b>	-	Reactive (Information provided on demand)	
<b>Account Servicing</b>	-	Reactive (Information provided on demand)	
<b>Hardships</b>	-	Reactive (Information provided on demand)	

Table 5.5: Existing channels of customer facing communications

#### 4. (T8) Has the team determined the level of internal transparency needed, and the audiences for the same?

To facilitate the external transparency requirements, it is important to understand and meet the requirements of internal stakeholders. This is important to those in customer facing roles and situations that require reactive communication (e.g., explaining reject decisions).

Internal audiences include:

- Frontline staff, telesales/support executives who interact with customers and prospects.
- Teams that build, assess, test, and validate the AI during design and periodically in BAU.
- Second line staff that provide assurance on the AI.

Audience/Stakeholder	Transparency artefacts
Frontline staff	<ul style="list-style-type: none"><li>• Transparency reports</li></ul>
Call centre staff	<ul style="list-style-type: none"><li>• Standard operating instructions to handle customer queries on transparency.</li></ul>
Aida validators	<ul style="list-style-type: none"><li>• Explainability reports/metrics</li><li>• Fairness reports/metrics</li></ul>
Second line/governance	<ul style="list-style-type: none"><li>• AIDA validation reports</li><li>• Unjust bias assessment</li><li>• Ongoing monitoring plan</li></ul>

Table 5.6: Internal stakeholders

Reports related to explainability and fairness are available for AIDA validators and reviewers through the tools implemented in the bank.

Transparency reports for the frontline staff will be developed based on operationalisation considerations. The requirements could include but are not limited to:

- Non-technical, simple language reports containing information related to the top reasons driving the decision in the specific instance. This should be based on the output of explainability techniques applied to validate the model and will help staff handle requests/enquiries from customers, and also ensure that explanations are consistent with the techniques applied.
- Counterfactual reports that will help staff explain the factors that can help the customer get a favourable decision.
- Training on how to handle requests for explanations for a customer's specific case, as well as operating instructions on handling different scenarios, for staff who handle customer calls.

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

To ensure the requirements for both internal and external transparency are met, it is important to establish explainability for the outcome of the AIDA use case.

As part of this implementation, the following techniques are applied to the use case from the available techniques listed in Step 0. Details of the techniques along with examples are available as part of step 3.

<b>Model agnostic methods</b>	<b>Global explanation methods</b>	<ul style="list-style-type: none"> <li>Partial dependence plots – PDPs</li> <li>Permutation importance</li> </ul>
	<b>Local explanation methods</b>	<ul style="list-style-type: none"> <li>Shapley values (popular approximation techniques include quantitative input influence (QII) and Shapley additive explanations (SHAP))</li> </ul>
<b>Conceptual soundness method</b>		<ul style="list-style-type: none"> <li>Influence Sensitivity plots</li> </ul>

Table 5.7: List of the explanation methods

## 6. (T10) Has the team ascertained that the chosen explanation method/implementation meets the minimum accuracy requirement for this specific use case (based on T5)?

The standard does not prescribe minimum accuracy expectations for explanation methods. As part of the testing and validation exercise, AIDA validators assess based on their understanding of the methods used.

The outcomes from Step 1 are summarised below

#	Checklist question	Yes/No
<b>T6</b>	Has the AIDA use case team determined whether there is a need for external (customer facing) transparency?	Yes
<b>T7</b>	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?	Yes
<b>T8</b>	Has the team determined the level of internal transparency needed, and the audiences for the same?	Yes
<b>T9</b>	Has the team selected a suitable explanation method for this specific use case from the approved list in T4?	PDP, permutation importance, feature importance based on Shapley values
<b>T10</b>	Has the team ascertained that the chosen explanation method/implementation meets the minimum accuracy requirement for this specific use case (based on T5)?	Not applicable

## 5.5.3 Step 2 – Prepare Input Data

The standard has a set of requirements specific to the data suitability principle which check for representativeness in the data, imbalanced datasets, as well as aspects related to completeness, fairness, and associated principles. These were applied to the use case but are not elaborated here. There are no special considerations in this step for the transparency principle.

## 5.5.4 Step 3 – Build and Validate AIDA System

Having considered questions T6 to T10 in the Methodology, this section describes how these considerations will be applied in the specific credit decisioning use case implementation. While many of the transparency related considerations are in place due to the existing standard and governance mechanisms, additional factors in the build and validate stage may be considered as part of ensuring operational readiness of the use case.

### 1. (T11) Have internal transparency dashboards/reports been implemented in line with the requirements agreed in T8-T10?

Internal transparency requirements related to explainability, bias checks, fairness assessment, materiality assessment for validators and reviewers have been implemented, and were used during assessment and validation of the AIDA use case.

The current nature of the credit decisioning AIDA implementation (challenger mode) constrains the utility of the transparency dashboards/reports to the frontline teams considering that not all customers would be processed through the challenger model, and equivalent dashboards/reports were not a requirement for the incumbent non-AIDA “champion” model. This has potential for creating an inconsistent user (and client) experience. The bank will consider this factor while creating transparency dashboards/reports for frontline staff in the future.

### 2. (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)?

The AIDA use case was reviewed for alignment with all principles in the standard, including fairness/bias, explainability, testing and validation, stability, data privacy, cyber security and third party risks, as well as the Methodology for transparency.

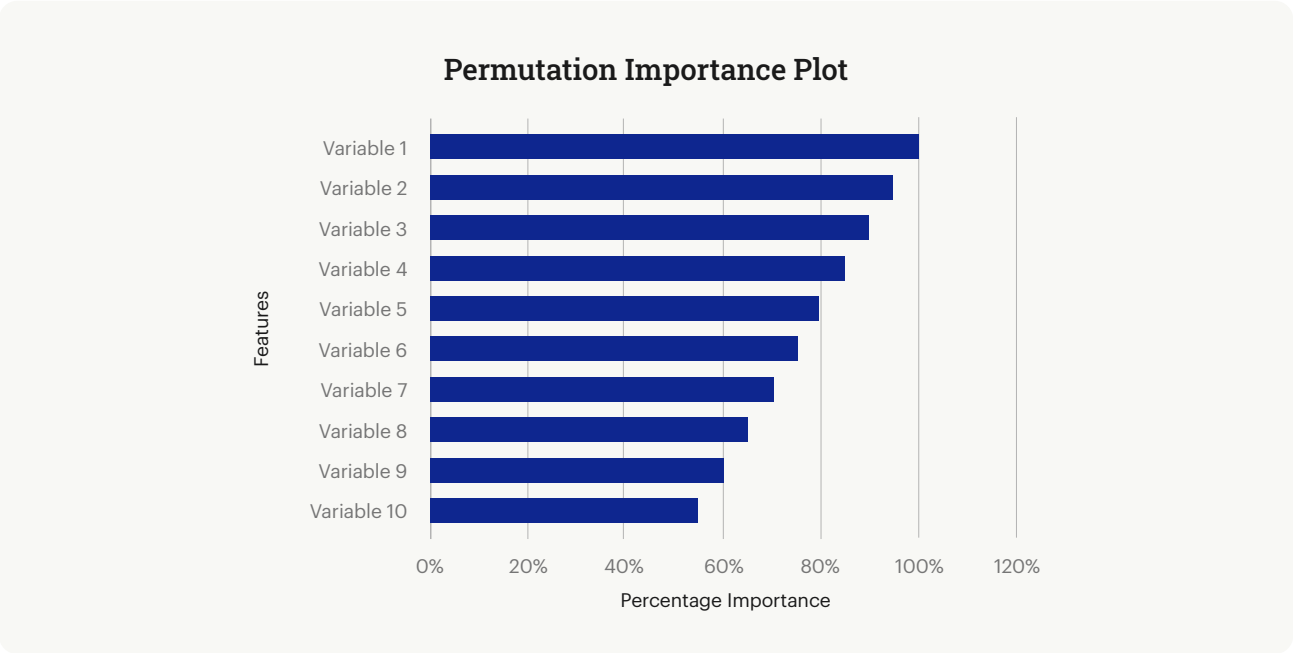
These validations included global and local explanations, which were also used as inputs to assess conceptual soundness. The outcome against each method is described below.

#### **Overall feature importance**

The feature importance plot (plot 1) shows the top 10 highest contributing features and their relative percentage importance in descending order, with the actual feature names redacted. The most important feature i.e., variable 1 is assigned the highest importance (100%) and all other variables are measured relative to variable 1, for e.g., variable 10 has the lowest importance (55%) in relation to variable 1 (among the top 10 features).

Plot 1 depicts that the feature importance distribution is smoothly decaying, which means that none of the features is dominating the performance of the model or is highly correlated with the target variable.

The top features for credit decision use case could include (for e.g.) number of unsecured loan enquiries by external clients in the last six months, the number of months since card issuance, maximum ever delinquency status seen, as in the below permutation importance plot.

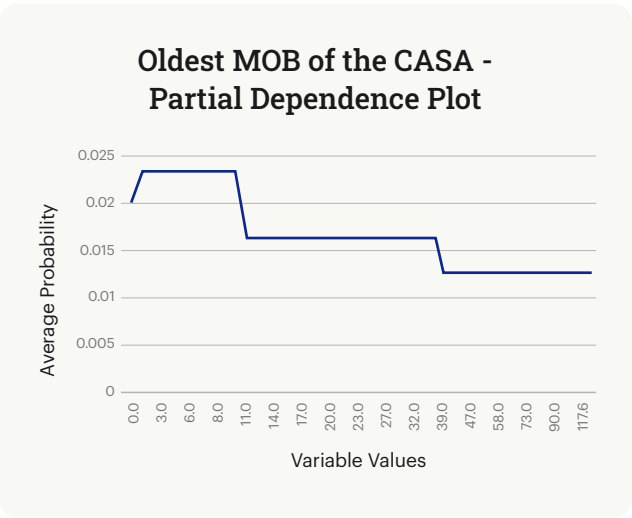


Plot 1: Overall feature importance

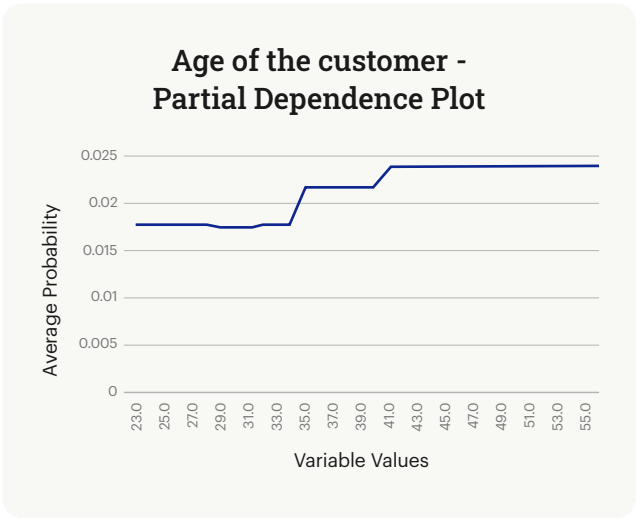
**Partial dependence plot**

Plot 2 shows decreasing bad rate with increasing value of the “months on book” feature. The stepwise change indicates the variable is grouped into ranges by the model and is in line with expectations and observed behaviour. Beyond a point, an increase in the value of the variable does not affect the outcome.

Plot 3 shows an increasing bad rate with increasing value of the “age of customer.” Again, the step like behaviour of the plot shows grouping by range with no impact from the increase of values beyond a particular point. The steps on the plot align and reflect major life events of customers.



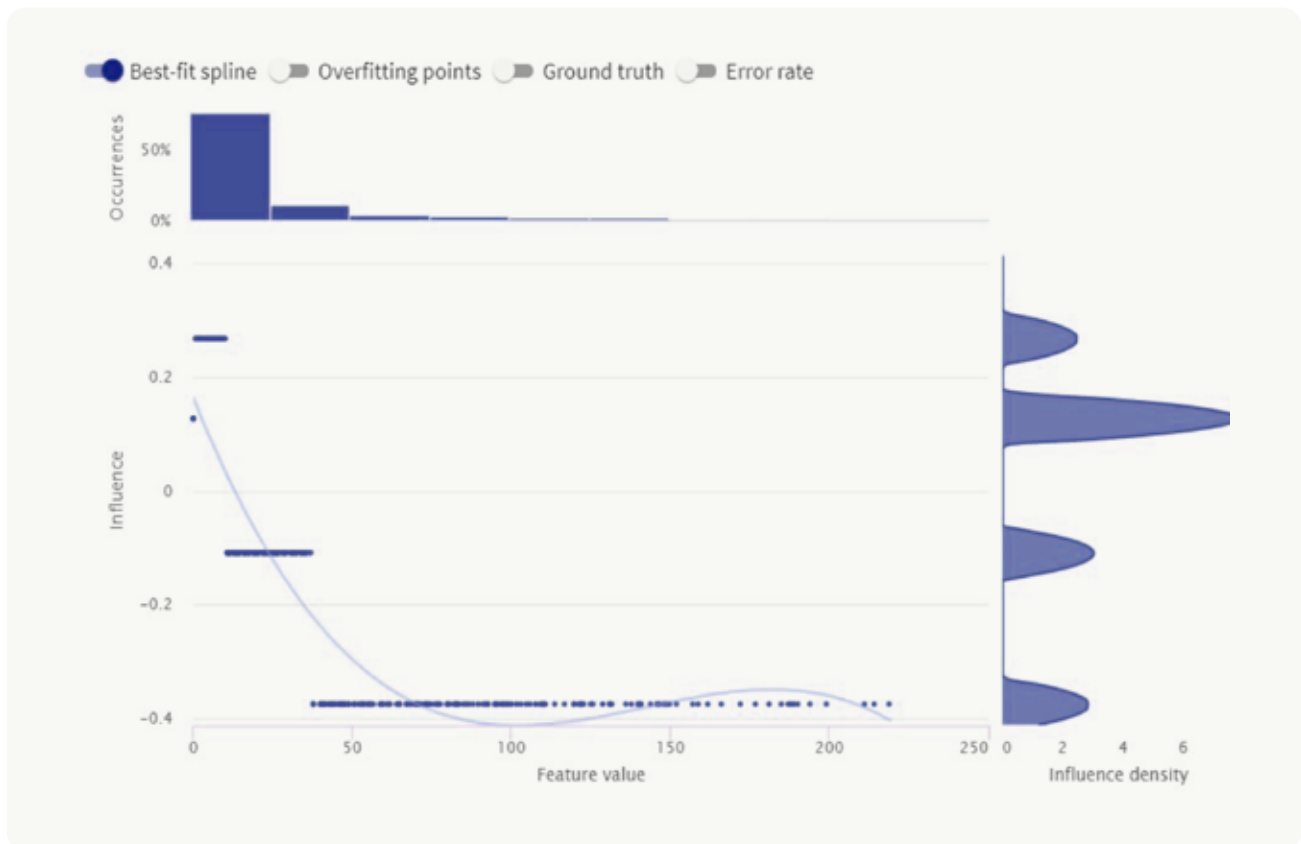
Plot 2: Partial dependence plot for variable 1



Plot 3: Partial dependence plot for variable 2

## **Influence sensitivity plot**

The influence sensitivity plot below shows the relationship between a feature's value and its contribution to the output. One can also view it as the transformation internally performed by a black box model as it plots a feature's influence against its raw values.



Plot 4: This view contains (i) a distribution of the feature values, (ii) the distribution of influence and (iii) the relationship between the two as an influence sensitivity plot.

Plot 4 (iii) depicts the influence of the “months on book” feature values towards risk. As the feature value increases the influence decreases, which explains the negative correlation of this feature with the risk. Feature values around 0 have the highest influence ( $>0$ ) towards risk whereas the risk decreases as the value increases.

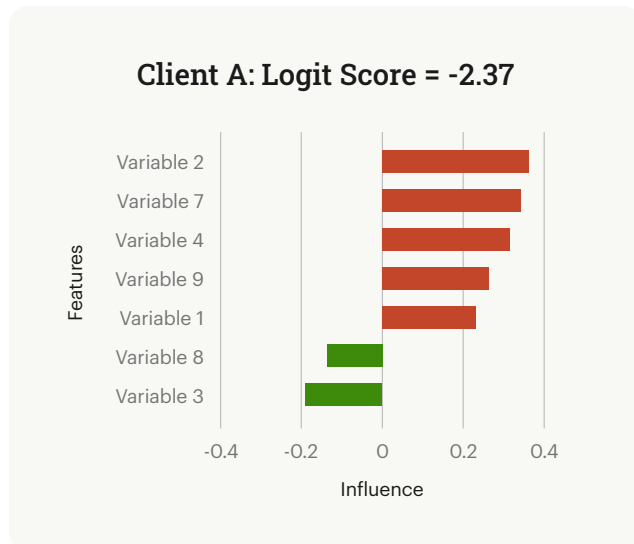
## **QII plot**

The QII plot shows the contribution (QII value) of different features towards the prediction of a data point, compared to the average prediction for the dataset. Given a set of feature values for an individual, the x-axis (influence) represents the contribution of a feature value to the difference between the actual prediction and the mean prediction.

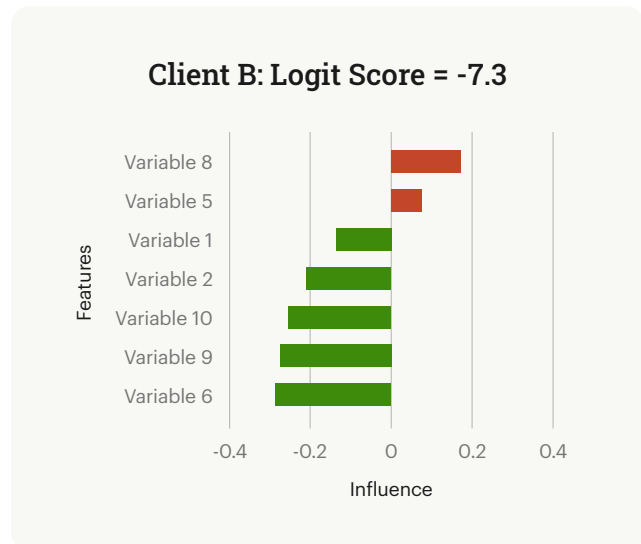
Plots 5 & 6 below show the QII values of the top influencing features for 2 different data points (customers). Red bars are contributing towards high risk (higher value of logit score/predicted probability) and greens are contributing towards low risk. The x-axis represents the degree of influence for each feature.

For Customer A, the top influencers are contributing towards higher risk, resulting in a higher logit score. For Customer B, the top influencers pushed the logit score to a lower value than average. As a result, Customer A's application will be more likely to get rejected than Customer B's.

Another observation would be, for a particular feature, the direction of influence changes with the value it takes for different customers. One example would be the customer's relationship with the bank. A longer relationship might help to decrease the risk than a newer client. In Plot 5 & 6, although variables average balance, age of customer and number of credit cards are some common influencers for customer A and B, based on their different values for A and B, in one case they are contributing towards higher risk and in the other case, the opposite.



Plot 5: QII Plot for Customer A with higher predicted risk



Plot 6: QII Plot for Customer B with lower predicted risk

### 3. (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)?

The bank's internal processes ensures that as part of the "validate" lifecycle stage, the AIDA use case has been reviewed by independent validators. This was performed iteratively and met the first and second line expectations. The independent validation included conceptual soundness and bias checks which were within tolerance thresholds, as well as an assessment of transparency related requirements as per the standard.

### 4. (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?

Reports required for assessment and validation are available as part of development and validation phase. Internal transparency reports for stakeholders required to enable external transparency will follow the outcome from T11.

### 5. (T15) Have operational processes such as customer service and complaint handling been modified appropriately to incorporate AIDA customer transparency? Have relevant staff been provided appropriate training to address customer queries?

## 6. (T16) Have customer/website Terms and Conditions been appropriately updated?

The AIDA use case is currently under implementation. The bank has taken additional external transparency requirements identified as future considerations to update business practices across both AIDA and non-AIDA driven decisions.

## 7. (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)?

All reports available as part of T11 and T14 during the testing and validation stage will be available after go-live.

The outcomes from Step 3 are summarised below

#	Checklist question	Yes/No
T11	Have internal transparency dashboards/reports been implemented in line with the requirements agreed in T8-T10?	Yes
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)?	Yes
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)?	Not applicable
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?	Not yet implemented
T15	Have operational processes such as customer service and complaint handling been modified appropriately to incorporate AIDA customer transparency? Have relevant staff been provided appropriate training to address customer queries?	Not yet implemented
T16	Have customer/website terms and conditions been appropriately updated?	Not yet implemented
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

## 5.5.5 Step 4 - Monitor AIDA System

Monitoring is in place for current non-AIDA use case on input data quality and model performance. This will be updated to include the AIDA use case as part of ongoing monitoring implementation to cover not just the transparency requirements, but also metrics related to data drift, unjust bias, and accuracy. As the use case is not yet operational, these are yet to be implemented.

## 5.6 Reflections

The Methodology is comprehensive and operates at the level of policies and standards to set up a structure that can be applied consistently to individual use cases. It covers the key components of transparency: data, model, outcomes, lifecycle of the AIDA system, and the stakeholders/consumers of transparency (internal and external).

This is accomplished using a set of 17 questions across the AIDA lifecycle, which ensures that all transparency considerations are considered and can be applied in a proportionate manner depending on the materiality/impact of the use case.

### **Considerations for AIDA policies/standards**

From the deep dive exercise, the bank established the following capabilities are already in place:

- Factors to determine whether customer facing transparency is essential for a use-case.
- Mechanisms to establish proactive or reactive communication required over the customer lifecycle as well as the artefacts/channels for the same.
- Factors to determine the extent of internal transparency, and associated audiences.

In addition, the bank has identified the following areas for future consideration:

- Prescribing acceptable explanation methods in line with the materiality of the use case and the nature of the underlying algorithms deployed.
- Specifying minimum accuracy standards for such explanation methods (for the use case).

These are currently not practicable due to the evolving nature and understanding of the explainable AI domain. The bank will assess these areas in the future when proven techniques are more widely adopted, and establish where in the governance structure they may be included, taking into account trade-offs of the current principles based approach against a more prescriptive and rule based standard.

### **Considerations for current transparency related business practices**

The level of transparency required for AIDA driven decisions is higher than when AIDA techniques are not used.

An AIDA driven challenger model will process only a subset of all customers as opposed to the non-AIDA incumbent model. The transparency requirements for the output from these models are different and could result in variations in the customer engagement process depending on which model processed the specific transaction. This could pose operational challenges to frontline teams and needs to be factored in while enabling transparency capabilities for AIDA models in a hybrid (AIDA and non-AIDA) environment.

### **Considerations for explanations related to internal and external transparency**

The explanations are expected to cover the final outcome of the AIDA system. This could pose a challenge for transparency to different stakeholders, internally and externally.

For internal stakeholders, especially AIDA developers and assessors, it is necessary to evaluate the accuracy and other parameters of the AIDA component itself before it is processed by downstream components. The influence of any non-AIDA post-processing overlays or human actions on the AIDA output could introduce “noise” in the explanations, especially for troubleshooting and validation activities. This will need to be addressed by validating the AIDA-driven model and the final outcome in separate steps.

For external stakeholders who do not have visibility into the workflow (e.g., credit approval), it may not be feasible to provide clear and detailed explanations on the final outcome in all cases, as these could be driven by factors beyond the AIDA-driven model. For example, a customer whose credit application was rejected sees the rejection result regardless of where in the lifecycle it was rejected. The rejection could be unrelated to the AIDA-driven model: it may have occurred earlier in the lifecycle due to fraud checks, or later in the lifecycle because the annual portfolio allocation for the product was fully utilised for the year. However, there are precautions to be taken before sharing such information (especially in relation to fraud and anti-money laundering, to avoid the risk of tipping off). Therefore, the steps proposed in the Methodology must be assessed and applied appropriately for each use case.

# 06 Transparency Assessment in Customer Marketing

## 6.1 Use Case

AIDA models are regularly used for marketing banks' products to fulfil customers' borrowing needs. This use case demonstrates approaches that can create more transparency in such modelling processes and decisions.

## 6.2 Context

HSBC proactively contacts existing credit card customers to discuss solutions about customers' borrowing needs. A selection system (shown in Figure 1) consisting of event triggers, business rules and machine learning (ML) models are used to prioritise leads for proactive contact.

Based on past interaction data and other signals, the selection system tries to infer customers' need for credit and the propensity to subscribe the bank's lending solutions. Safeguards like credit and contact exclusions, regular model validation and ongoing performance reviews are integral to the system development and use.

Transparency evaluation toolkits, including open source code libraries, are utilised to create better explanations of the selection process and impacted customer outcomes.

### Customer marketing selection system

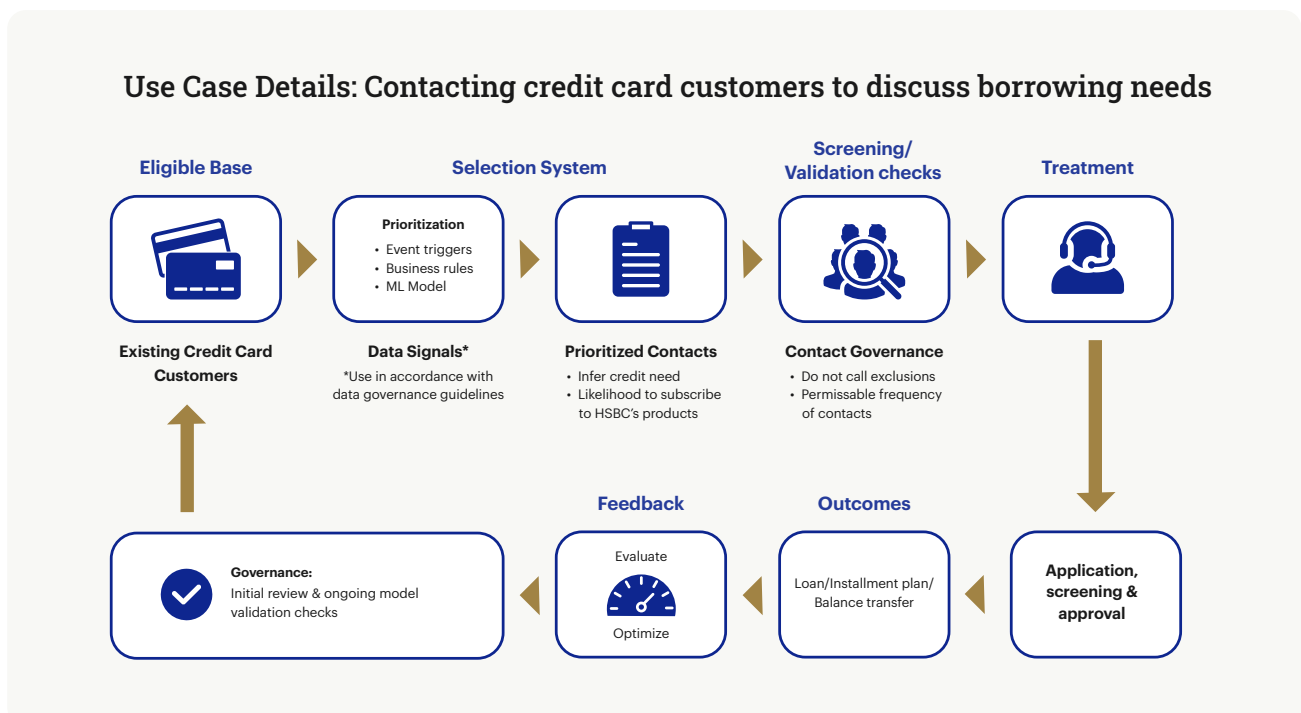


Figure 6.1: Customer marketing use case overview

## 6.3 Key Components for Transparency Evaluation

1. **AIDA selection models** which prioritise customers to be proactively contacted for a borrowing needs conversation. Specific to a borrowing needs conversation, existing credit card customers form the eligible base subject to prevalent regulatory and credit risk exclusion criteria. Three distinct types of AIDA models are used – event triggers, business rules and machine learning models. Based on recent customer behaviour and past engagement history, these models try and infer the need for credit and likelihood to subscribe to bank's solutions. The use of data signals to build these three AIDA models is strictly governed by internal guidelines. Data elements which do not meet the permissible guideline are discarded, even though they may bring good discriminatory power. Some defined criteria help prioritise and establish a hierarchy among AIDA model selections which are actioned subject to outbound direct marketing capacity (call, sms, email, etc.)
2. **System transparency** addresses understanding of the mechanism by which the selection system works (i.e., which input factors lead to specific system outputs). For the purpose of creating greater transparency, the selection system has been approximated to a single machine learning model. The approximation was built on the same input data features, and was trained to produce an output similar to the underlying system. The transparency assessment carried out here is focused on building better explanations for internal audiences (i.e., business sponsors, model owners, reviewers, validators, etc.) of the FSI and is consequently algorithm agnostic.
3. **Treatment** includes a proactive direct marketing outreach. In this use case, only outbound telephone conversations about borrowing needs and related personal loan solutions, are considered in scope for system evaluation.
4. **Outcomes** relate to “approve” or “decline” decisions on customer's loan application as per the bank's policies during the decision period. A customer conversation may result in take up of other credit products, such as an instalment plan or credit card. Such outcomes are, however, excluded to maintain comparison rigour.
5. **Exclusions:** The impact of product or credit solution design, features including pricing, treatment scripts/messaging and credit risk criteria, on customer outcomes are out of scope for transparency assessment of customer selection system.

## 6.4 Transparency Assessment Approach

- **Selection system** is assessed as a “single” AIDA model. Assessment is algorithm agnostic in the sense that we focus on general diagnostics that enables better explainability of the overall selection system as opposed to a subset/component of the modelling process.
- **Outcomes** (approve or decline decisions) are not fully attributable to customer selection system as other factors have an influence, such as credit risk policies, a customer's choice to not subscribe, etc. However, creating transparency with respect to the selection system helps provide a clear understanding of why specific a customer population was selected for proactive contact over others from the eligible base.

- **Proactive customer contact** (“treatment”) may lead to borrowing needs being fulfilled by appropriate credit products such as instalment plans, personal loans or credit cards. For the purpose of this study, however, only specific outcomes – successful personal loan applications – are considered to maintain comparison rigour. Business impact is measured in terms of customer needs fulfilled and the modelling process aims to maximise such outcomes within given constraints.
- **Transparency assessment** aims to explain the impact of input parameters on selection of customers for contact prioritisation to address borrowing needs. Global (most influential features generally) and local (particular outcome) interpretations are featured here.

## 6.5 Transparency Assessment and Explanations

### 6.5.1 Step 0 - Set (Transparency) Standards Internally Within the FSI

#### **(T1) Has the FSI defined the factors it will use to determine whether external (customer facing) transparency is essential for a particular AIDA use case?**

Yes. The decision is guided by prevailing regulatory disclosure requirements in the market, internal model development standards and HSBC’s “Principles for the Ethical Use of Big Data and AI”. For example, one principle that directly addresses this topic is: “we aim to be transparent with our customers and other stakeholders about how we use their data, unless there is an overriding public interest (e.g., the prevention of financial crime).”

#### **(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 the standard templates/interfaces for the same?**

Yes. Internal guidelines cover some aspects of the proactive and reactive communication. However, these are evolving constantly. Given these are proprietary, we are unable to share here.

#### **(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?**

Yes. By our internal model development standards, the internal transparency artefacts are prepared by the model developer, and the audiences are the business stakeholder, the model owner, model monitoring manager, and the business analyst (model user). The models which are of low-risk materiality will not go through independent risk validation process, while the rest will pass independent model review.

#### **(T4) Has the FSI defined an acceptable set of AIDA ML explanation method(s) for use within the FSI?**

Yes. LIME (local interpretable model-agnostic explanation, and SHAP (Shapley additive explanation) are recommended by internal model development standards to interpret the model.

#### **(T5) Has the FSI set minimum accuracy standards for such explanation methods?**

No. While the explanation requirements are clearly defined, benchmarks are yet to be calibrated as the practice is not fully matured.

### **6.5.2 Step 1 - Define system context and design**

The gradient boosting model was used as the selection system approximation model. Twenty-one original factors (predictors) were used, including business rules, triggers and scoring model features. All parameters were masked (as business sensitive). The model's accuracy is 86%, which is reasonable for the customer targeting model.

#### **(T6) Has the AIDA use case team determined whether there is a need for external (customer facing) transparency?**

Yes. For this use case, limited external transparency is required as the activity relates to proactive marketing effort to address potential customer needs via a communication channel already approved/consented by the customer.

#### **(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?**

Yes. Proactive communication allows the customer to know about product eligibility criteria, the product offer and various terms and conditions via the website or during telemarketing outreach. Additionally, customer consent for receiving marketing offers via specific channels is recorded prior to contact. Since the choice to apply or not is made by the customer, reactive communication (customer facing), other than confirmation of the chosen offer, is not required. Please note, the outcome of credit decision is out of scope for this use case.

However, there are several requirements with regards to internal transparency or explainability to aid an organisation's internal review monitoring and to aid audits. These are covered later in the document.

#### **(T8) Has the team determined the level of internal transparency needed, and the audiences for the same?**

Yes. The internal transparency evaluation is conducted as part of HSBC's model development standards requirements. It is not a standalone exercise but rather a part of the internal regulatory framework. Along with transparency, the underlying system's fairness will be evaluated (which is covered by Veritas Phase 1), as well as its quality, risks and list of possible actions during the model's lifecycle. All these requirements are defined at the model planning stage and will be used by business stakeholders, the model review team, and for internal audits.

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

Yes. There are two standard algorithms for complex models interpretability in the industry – LIME (local interpretable model-agnostic explanation)<sup>10</sup> and SHAP (Shapley additive explanation).<sup>11</sup> LIME helps to illuminate the AIDA model and to make its predictions individually comprehensible. The method explains the classifier for a specific single instance and is therefore suitable for local consideration. The advantage of SHAP lies in the fact that it unifies all available frameworks for interpreting predictions. Another popular methodology to support the model's transparency evaluation is ELI5,<sup>12</sup> which helps to debug the machine learning classifier and explain its top prediction, but which is mostly limited to tree based and other parametric/linear models. Although the number of the proposed techniques continues to grow, there has been little evaluation of whether they can help business stakeholders to achieve their desired goals.

We used SHAP values to interpret the approximation model and show how the evaluation approach works. The benefits of using SHAP are both at an overall and at a local level, as follows:

- At a global level, the collective SHAP values helps stakeholders to interpret and understand the model. They show how much each predictor contributes, either positively or negatively, to the target variable. It allows for very intuitive interpretation of the model structure and is generalisable across a number of different modelling methodologies.
- At a local level, each observation gets its own set of SHAP values (one for each predictor). This greatly increases transparency, by showing contributions to predictions on a case by case basis, which traditional variable importance algorithms are not able to do. In addition, local interpretability can aid in segmentation and outlier detection.

### **(T10) Has the team ascertained that the chosen explanation method/implementation meets the minimum accuracy requirement for this specific use case (based on T5)?**

No. As there are no common standards on this requirement, its evaluated based on the particular use case, in agreement with the business stakeholders and model owner.

## **6.5.3 Step 2 - Prepare Input Data**

We use the cohort of customers who responded to a previous campaign as the customer targeting system input. Every customer is represented with 21 masked features.

## 6.5.4 Step 3 - Build and Validate the AIDA System

### Internal transparency:

**(T11) Have internal transparency dashboards/reports been implemented in line with the requirements agreed in T8-T10?**

Yes, outlined below.

The advantage of SHAP is that it offers the same level of interpretation regardless of the model type. This is especially important as we approximate the initial selection system by the model.

### Global interpretability:

At a global level, the below graph at Figure 2 summarises the effects of all the explanatory variables on the model output, colour coded to show the direction of the impact (red means an increase, while blue shows a decrease). SHAP value that are further away from zero reflect a bigger impact. It is also visually easy to see which variables have the strongest relationship with the target variable. In this way, SHAP is also useful as a tool for variable selection.

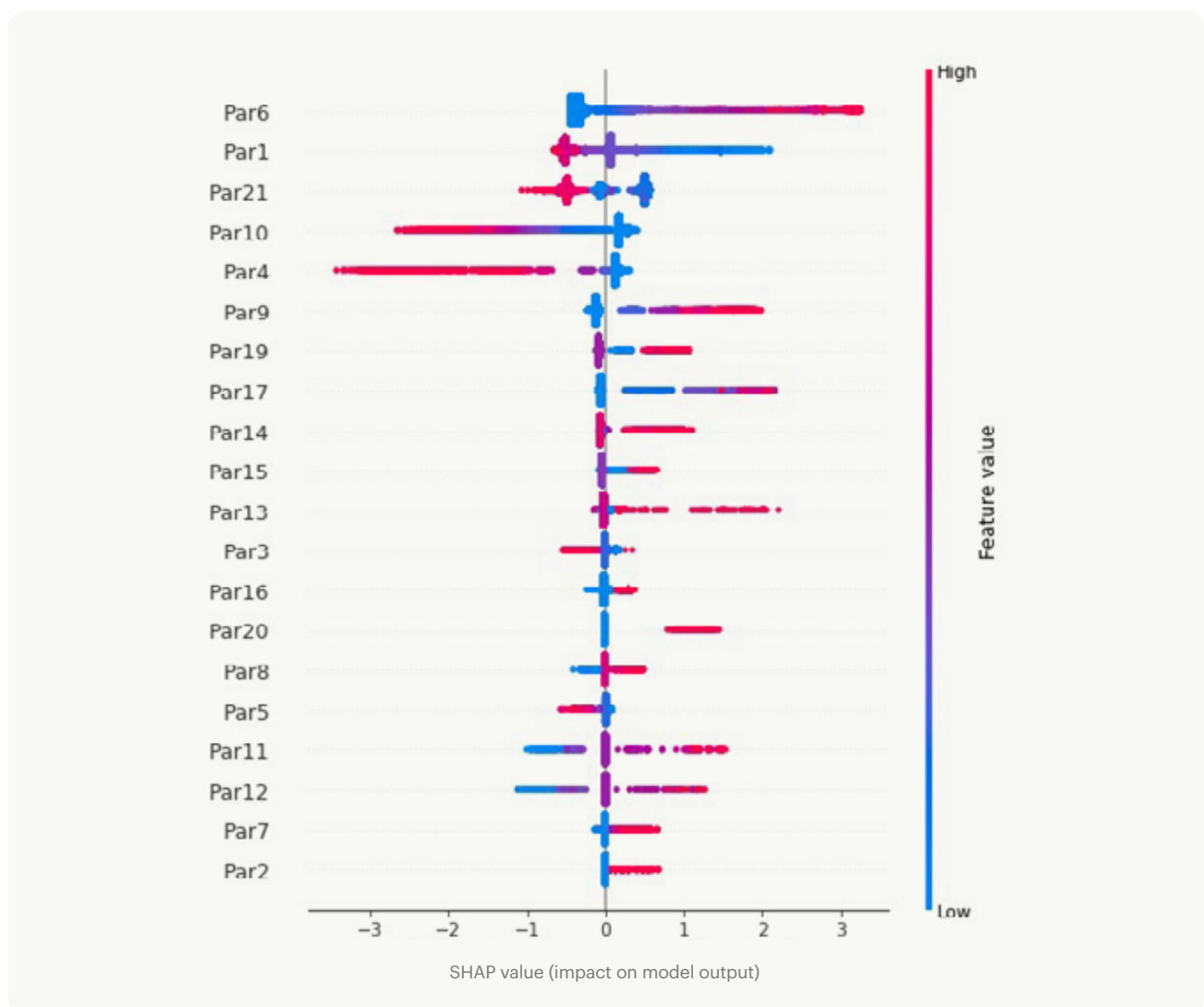


Figure 6.2: Global model interpretability with Shapley values

To summarise, the above plot demonstrates the following information:

- **Feature importance.** Variables are ranked in descending order.
- **Impact.** The x-axis location shows whether the effect of that value is associated with a higher or lower prediction.
- **Original value.** Colour shows whether that parameter is high (in red) or low (in blue) for the particular customer. Every customer would have his own weights (or Shapley/SHAP values) for every parameter.
- **Correlation.** Positive and negative relationships of the predictors with the outcome (customer get called or not).

Example:

Parameter	Contribution	Business Impact
Par 6	Top 1	The higher Par6 value, the more likely the customer will be called
Par 1	Top 2	Par1, is working in opposite direction: the higher its value, the lower probability for that customer to be called.

Table 6.1: Parameters global interpretability example

Figure 3 below represents the complex relationships between the two most influential parameters, Par6 and Par1, and the possibility to get selected. It explains that probability is increasing with higher Par6 and decreasing with Par1.

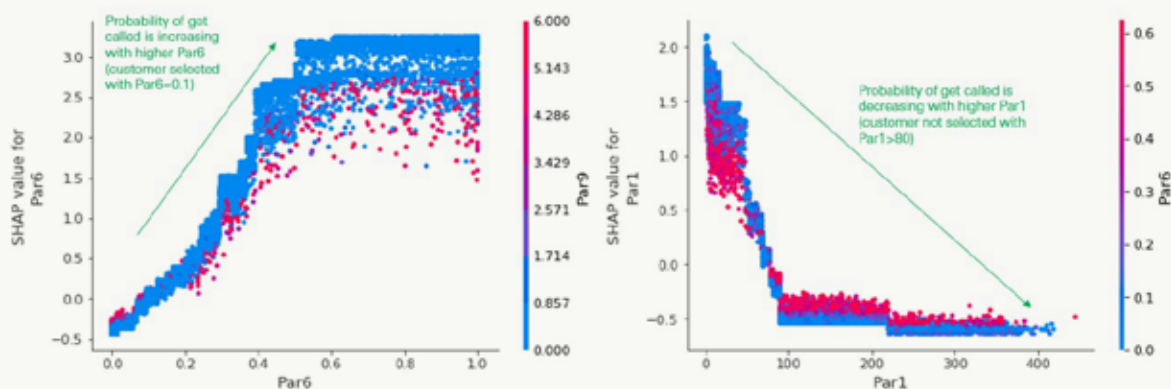


Figure 6.3: Par6 and Par1 impact on the model output

## Local interpretability:

We use local interpretability to explain the model on a more granular level. This information is used by the model developers (to meet business requirements and optimise the model), business stakeholders (to monitor the selection performance), model review team (for internal audit).

For a local view which makes it clearer which way each variable is “pushing” the model output towards, the following plot in Figure 4 can be used, selecting the row/observation we want to show:

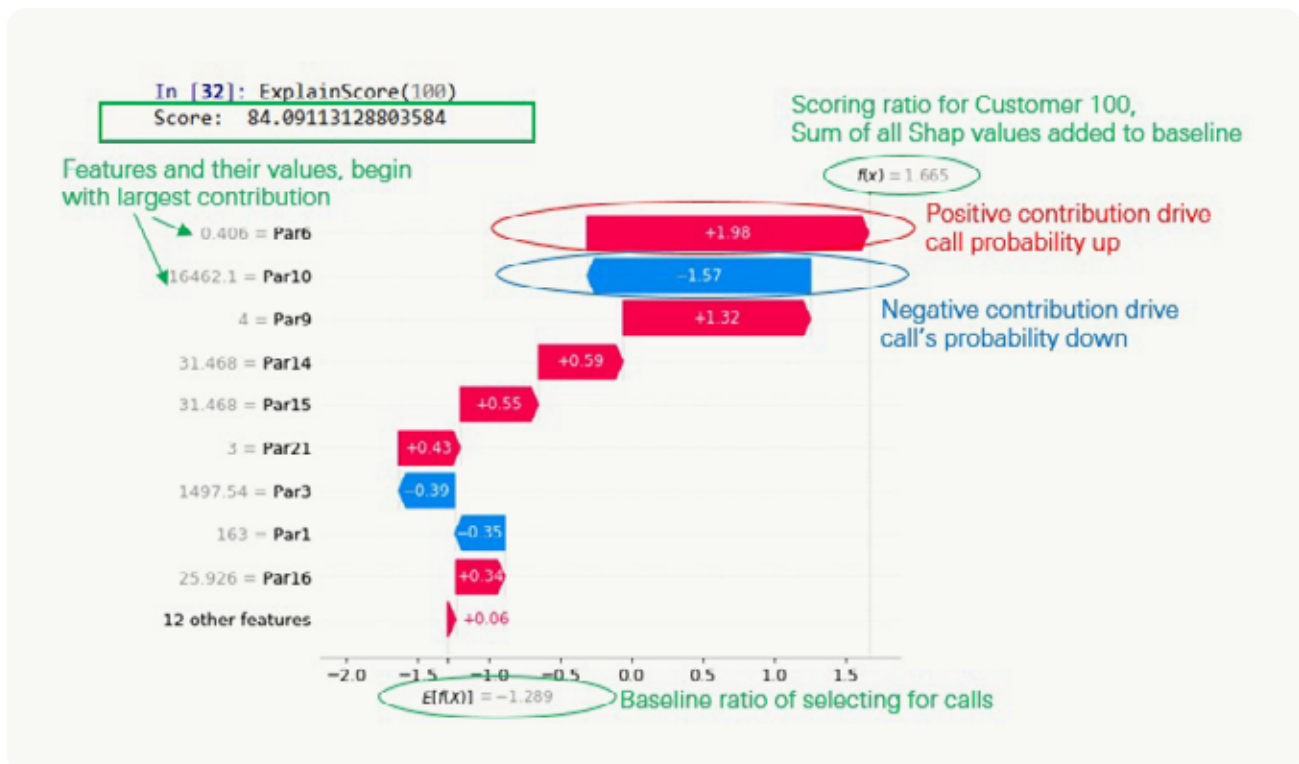


Figure 6.4: Explaining model output for a particular customer

- The x-axis is the Shap value,  $E[f(x)] = -1.289$  indicates the baseline ratio of customer gets selected (average Shap value for all customers). All cases with  $f(x) > -1.289$  have scores  $>50\%$  (high probability for the customer to get selected for the call), and  $f(x) < -1.289$  have scores  $<50\%$  (low probability).
- Shap values of all the input features will always sum up to the difference between baseline (expected) model output and the current model output for the prediction being explained.

The above graph shows the base value (the average model output over the training dataset), compared to the model output. Shown in red are the variables pushing the prediction higher, while the opposite holds true for the variables in blue. This highlights both the direction and the measure of impact for the variables. By putting them side by side, we can also compare the predictions of multiple models (variable impact, direction, and base value vs. model output), on observations or segments.

Example:

Parameter	Value	Contribution	Business Impact
Par 6	0.406	1.98	Par 6 add to the baseline ratio 1.98, increasing probability of getting called for customer 100
Par 10	15462.1	-1.57	Par 10 remove 1.57 from baseline ratio, increasing probability of not selected for call, for customer 100

Table 6.2: Parameters global interpretability example

**(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)?**

Yes. The model passed several approval stages. The risk materiality rating was assigned at the planning stage, which defines the necessary audience and content. At the development and testing stage, the transparency assessment was conducted, documented, and requirements for future monitoring created. All artefacts were reviewed and approved (by model owner and business stakeholders), and the model went to the implementation stage, where it was assessed again to ensure consistency and again reviewed and approved before the production stage.

**(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)?**

Yes. As an example, to improve communications with the internal model's users (part of internal audience), we can skip a particular feature's explanation – the one which has high weight to the outcome but which is not converted to clear action.

## External transparency

**(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?**

Not applicable.

**(T15) Have operational processes such as customer service and complaint handling been modified appropriately to incorporate AIDA customer transparency? Have relevant staff been provided appropriate training to address customer queries?**

Not applicable.

**(T16) Have customer/ website terms and conditions been appropriately updated?**

Yes

**(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, as described in Step 4

## 6.5.5 Step 4 - Monitor AIDA System

The selection model is monitored in accordance with HSBC's internal policies and model development standards. Additionally, specific to transparency assessment, the following aspects may be covered for certain customer selection models, depending on complexity and materiality.

- **Transparency metrics (local/global).** We calculate SHAP values and monitor their distribution and discrepancy from their benchmarks (SHAP values for the model training data).
- **Transparency outcomes** observed in the system and compared to their benchmarks.
- **Transparency service provision**, i.e., any relevant party can have access to the data.

Once established, the transparency evaluation may be automated with minimal manual intervention. Results can be automatically updated on a regular basis and the assessment system can be programmed to trigger manual review subject to significant deviation from baseline. Establishing thresholds and defining "significant" changes is an ongoing effort and still under development.

# 07 Acknowledgements

Workstream	Role	Name	Organisation
<b>Fairness</b>	Lead Subject Matter Expert	Medb Corcoran	Accenture
<b>Fairness</b>	Subject Matter Expert	Laura Alvarez	Accenture
<b>Fairness</b>	Data Scientist	Smitha HD	Accenture
<b>Fairness</b>	Data Analytics	Sneha Ramteke	Accenture
<b>Fairness</b>	Synthetic Data Generation Subject Matter Expert	Jer Hayes	Accenture
<b>Fairness</b>	Lead Academic	Adrian Weller	Turing
<b>Fairness</b>	Lead Subject Matter Expert	Yannick Even	Swiss Re
<b>Fairness</b>	Project Coordinator & Data Scientist	Guan Wang	Swiss Re
<b>Fairness</b>	Project Coordinator & Data Scientist	Yuxuan Zhang	Swiss Re
<b>Fairness</b>	Lead Data Scientist	Ming Yang	Swiss Re
<b>Fairness</b>	Lead Data Scientist	Luca Baldassarre	Swiss Re
<b>Fairness</b>	L&H Underwriting Expert	Farooque Ahmed	Swiss Re
<b>Fairness</b>	L&H Underwriting Expert	Jon Lambert	Swiss Re
<b>Fairness</b>	Risk Management Expert	Lutz Wilhelmy	Swiss Re
<b>Fairness</b>	Legal & Compliance Expert	Tomas Abelovsky	Swiss Re
<b>Ethics &amp; Accountability</b>	Lead Subject Matter Expert	Steven Tiell	Accenture
<b>Ethics &amp; Accountability</b>	Subject Matter Expert	Lara Pesce Ares	Accenture
<b>Ethics &amp; Accountability</b>	Lead Subject Matter Expert	Chaouki Boutharouite	AXA

Workstream	Role	Name	Organisation
<b>Ethics &amp; Accountability</b>	Lead Subject Matter Expert	Lorenzo Morganti	AXA
<b>Ethics &amp; Accountability</b>	Project Manager	Mohit Gupta	AXA
<b>Ethics &amp; Accountability</b>	Lead Data Scientist	Valerii Ishchenko	AXA
<b>Ethics &amp; Accountability</b>	Subject Matter Expert	Tan Li Choo	UOB
<b>Ethics &amp; Accountability</b>	Subject Matter Expert	Chim Chau Seng	UOB
<b>Ethics &amp; Accountability</b>	Subject Matter Expert	Andy Ang Jun Long	UOB
<b>Ethics &amp; Accountability</b>	Data Governance Expert	Chow Wai Pun	UOB
<b>Ethics &amp; Accountability</b>	Project Manager	Robert Cheah Tong Ngee	UOB
<b>Ethics &amp; Accountability</b>	Data Scientist	Sreeparna Majumder	UOB
<b>Ethics &amp; Accountability</b>	Data Scientist	Murari Mohan	UOB
<b>Transparency</b>	Subject Matter Advisor	Professor Anupam Datta	TruEra
<b>Transparency</b>	Research Engineer	Divya Gopinath	TruEra
<b>Transparency</b>	Lead Subject Matter Expert	Shameek Kundu	TruEra
<b>Transparency</b>	Lead Subject Matter Expert	Gunjan Bhatt	HSBC
<b>Transparency</b>	Data Scientist	Oxana Samko	HSBC
<b>Transparency</b>	Project Manager	Amos Ong Seng Keong	HSBC
<b>Transparency</b>	Subject Matter Advisors	Pauline Tee	Standard Chartered
<b>Transparency</b>	Subject Matter Advisors	Vijay Jairaj	Standard Chartered
<b>Transparency</b>	Lead Subject Matter Experts	Adhinarayan Nammalvar	Standard Chartered
<b>Transparency</b>	Lead Subject Matter Experts	Natalia Goh	Standard Chartered
<b>Transparency</b>	Lead Subject Matter Experts	Sanjeev Satija	Standard Chartered
<b>Transparency</b>	Data Scientist	Kaushik Das	Standard Chartered
<b>Transparency</b>	Data Scientist	Niladri Sekhar Samanta	Standard Chartered

We would also like to express our appreciation to Karen Tan (Swiss Re), Daisy Ning (Swiss Re), Marcin Detyniecki (AXA), Richard Lowe (UOB), Johnson Poh Wei Li (UOB), Wong Kee Joo (HSBC), Anurag Mathur (HSBC), Yusuf Demiral (HSBC), Manohar Chadavalavada (Standard Chartered), Vishu Ramachandran (Standard Chartered) and Manoj Piplani (Standard Chartered) who have supported and contributed to this project.

# 08 Bibliography

<sup>1</sup> <https://www.mas.gov.sg/-/media/MAS/News/Media-Releases/2021/Veritas-Documents-2-FEAT-Fairness-Principles-Assessment-Case-Studies.pdf>

<sup>2</sup> [https://www.researchgate.net/publication/239682035\\_Fair\\_risk\\_assessment\\_in\\_life\\_and\\_health\\_insurance](https://www.researchgate.net/publication/239682035_Fair_risk_assessment_in_life_and_health_insurance)

<sup>3</sup> <https://www.globallogic.com/wp-content/uploads/2021/04/Automating-Underwriting-to-Increase-Direct-Sales-1.pdf>

<sup>4</sup> <http://www.datasciencepublicpolicy.org/projects/aequitas/>

<sup>5</sup> [https://en.wikipedia.org/wiki/Demographics\\_of\\_Singapore](https://en.wikipedia.org/wiki/Demographics_of_Singapore)

<sup>6</sup> [https://en.wikipedia.org/wiki/Demographics\\_of\\_Singapore](https://en.wikipedia.org/wiki/Demographics_of_Singapore)

<sup>7</sup> <https://arxiv.org/pdf/1808.00023.pdf>

<sup>8</sup> <https://reports.swissre.com/sustainability-report/2020/governance.html>

<sup>9</sup> [https://www.dfs.ny.gov/system/files/documents/2021/03/rpt\\_202103\\_apple\\_card\\_investigation.pdf](https://www.dfs.ny.gov/system/files/documents/2021/03/rpt_202103_apple_card_investigation.pdf)

<sup>10</sup> <https://lime.readthedocs.io/en/latest/>

<sup>11</sup> <https://shap-lrjball.readthedocs.io/en/latest/index.html>

<sup>12</sup> <https://eli5.readthedocs.io/en/latest/overview.html>

## **LEGAL NOTICE**

This report is prepared and issued by the MAS, Accenture, AXA, HSBC, Standard Chartered, Swiss Re, UOB, and TruEra.

All intellectual property rights in or associated with this report remain vested in the MAS, Accenture, AXA, HSBC, Standard Chartered, Swiss Re, UOB, and TruEra 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, AXA, HSBC, Standard Chartered, Swiss Re, UOB, and TruEra 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.