IMPLEMENTATION OF FAIRNESS PRINCIPLES IN FINANCIAL INSTITUTIONS' USE OF ARTIFICIAL INTELLIGENCE / MACHINE LEARNING

- OBSERVATIONS FROM A THEMATIC REVIEW

INFORMATION PAPER

June 2022

MAS
Monetary Authority of Singapore

TABLE OF CONTENTS

1 Introduction 1

2 Background 1

3 Observations from Thematic Review 4

A. Scope 4

B. Materiality 5

C. Fairness 6

D. Others 9

4 Conclusion 10

1 Introduction

1.1 Financial institutions (FIs) are increasingly using Artificial Intelligence and Data Analytics (AIDA) to bolster operational efficiency, mitigate risks, and improve business outcomes. However, these benefits are coupled with concerns that the extensive use of AIDA could amplify existing risks or create new ones, if not well managed.

1.2 Innovation in the financial sector should be underpinned by good governance and risk management, as well as driven by sustainable strategies. While MAS encourages FIs to experiment and innovate with AIDA, we also expect FIs to do so responsibly.

2 Background

FEAT Principles

1.3 In late 2018, MAS co-created with the financial industry, principles of Fairness, Ethics, Accountability and Transparency (the "FEAT Principles") to promote the deployment of AIDA in a responsible manner[1].

1.4 To provide guidance to FIs in implementing FEAT, MAS worked with the Veritas consortium[2] to create a framework known as Veritas in November 2019. Veritas aims to provide FIs with a verifiable way to incorporate the FEAT Principles into their AIDA solutions. It comprises open-source tools that can be applied to different business lines in various markets. For Phase 1[3], the consortium published two whitepapers on the FEAT Fairness Assessment Methodology and case studies on two banking use cases in January 2021. Phase 2[4] was concluded in February 2022 with the issuance of five whitepapers that included:

- A FEAT checklist for FIs to adopt during their AIDA software development lifecycles.
- An Enhanced Fairness Assessment Methodology to enable FIs to define the fairness objectives of their AIDA solutions, and identify personal attributes of individuals as well as any unintentional bias.
- A new Ethics and Accountability Assessment Methodology, which provides a framework for FIs to carry out quantifiable measurement of ethical practices, in addition to the qualitative practices currently adopted.
- New Transparency Assessment Methodology, which helps FIs determine whether and how much internal/external transparency is needed to explain and interpret the predictions of machine learning models.

[1] https://www.mas.gov.sg/~/media/MAS/News%20and%20Publications/Monographs%20and%20Information%20Papers/FEAT%20Principles%20Final.pdf

[2] The Veritas consortium is made up of various FIs, technology companies and MAS. Refer to the link for the full list of consortium members in Phase 1 and 2 of Veritas: https://www.mas.gov.sg/schemes-and-initiatives/veritas

[3] Press release on Veritas Phase 1 (dated 6 Jan 2021): https://www.mas.gov.sg/news/media-releases/2021/veritas-initiative-addresses-implementation-challenges

[4] Press release on Veritas Phase 2 (dated 4 Feb 2022): https://www.mas.gov.sg/news/media-releases/2022/mas-led-industry-consortium-publishes-assessment-methodologies-for-responsible-use-of-ai-by-financial-institutions

1.5 Looking ahead, the Veritas consortium will be applying the Transparency Assessment Methodology to the insurance underwriting use case and the full FEAT assessments for payments use cases. Pilots will be run with selected FIs to integrate the methodologies with their existing governance frameworks. The Veritas Toolkit[5] will also be incorporating a newer version of the diagnostic tool for Fairness and Transparency, and an enhanced assessment tool for FEAT. Integration with APIX[6] will also allow FIs and FinTechs to have their own Veritas Toolkit environment on the platform. In developing the Veritas ecosystem, the consortium will invest in talent development, providing bite-size training and roadshows to raise awareness and develop expertise in FEAT. It will also look into defining norms and collating acceptable practices for areas relevant to the industry, such as reject inference methods, fairness metrices thresholds and protected attributes. These initiatives will take into consideration common issues and challenges faced by FIs as noted by the thematic review, to support FIs in adhering to the FEAT Principles.

Thematic Review on FIs' Use of Artificial Intelligence / Machine Learning (AI/ML)

1.6 In late 2021, MAS conducted a thematic review of selected banks and insurers to understand the extent of their AI/ML adoption, as well as the maturity of governance frameworks and controls in place to meet the Fairness component[7] of the FEAT Principles:

> Justifiability
> F1. Individuals or groups of individuals are not systematically disadvantaged through AIDA-driven decisions unless these decisions can be justified.
> F2. Use of personal attributes as input factors for AIDA-driven decisions is justified.
>
> Accuracy and Bias
> F3. Data and models used for AIDA-driven decisions are regularly reviewed and validated for accuracy and relevance, and to minimize unintentional bias.
> F4. AIDA-driven decisions are regularly reviewed so that models behave as designed and intended.

1.7 These principles provide broad guidance on how fairness can be achieved when AIDA models are designed and assessed, without which, the AIDA models could cause unintended harms and reinforce existing disadvantages in our society. For example, a credit model without fairness considerations may unfairly grant fewer loans to females compared to males even though the pool of female applicants may have credit scores that are better

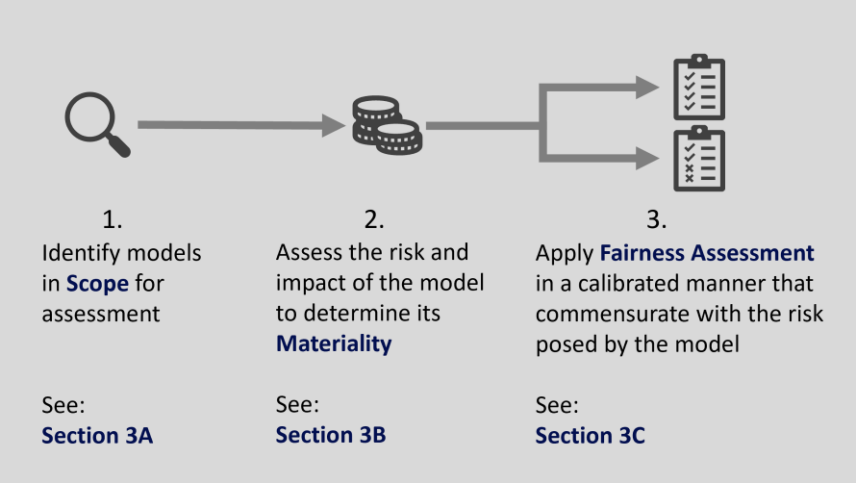
[5] Veritas Toolkit version one is available at https://github.com/veritas-toolkit/

[6] APIX is purpose-built to transform and radically simplify the FI - FinTech collaboration process end-to-end. The Veritas Toolkit will be deployed as a docker image on APIX so that FIs and FinTechs can have their own Veritas Toolkit environment by simply launching the Toolkit. The Veritas Toolkit environment will be owned by the FIs or FinTechs and they will have full control over it for their POC projects. For more info on the APIX Platform please refer to: https://apixplatform.com/

[7] The thematic review was conducted prior to release of the enhanced Fairness Assessment Methodology and insurance use cases on fairness in February 2022.

than or equivalent to males[8]. Besides causing harm to customers, there would also be a risk of reputational damage to FIs.

1.8 To meet the objectives of these principles, FIs may calibrate actions and requirements under their internal governance framework based on the materiality of the AIDA-driven decisions. In line with this risk-based approach, the typical process an FI uses to manage its AIDA models is outlined below:



1.9 The following areas were covered in this review:

(i) FIs' policies, procedures, frameworks and governance structures relating to the three themes above; and

(ii) Implementation effectiveness, i.e. whether the policies and frameworks translated into practical and effective implementation in AI/ML use cases.

1.10 Despite the huge potential of AI/ML, FIs are still at a nascent stage of using AI/ML models for decision-making in a manner that has direct impact on consumers. Many FIs are experimenting with AI/ML use cases and proofs-of-concept, but are not using AI/ML extensively in actual decision-making. For this thematic review, MAS reviewed sample AI/ML use cases that are utilised by FIs to make decisions that have the potential to cause harm to consumers or other individuals[9].

1.11 MAS expects FIs to have risk management and internal controls that are commensurate with their risk profile and complexity of their operations. In the context of AIDA, FIs that use or have plans to use AIDA models extensively in decision-making should set up robust AIDA governance structures and demonstrate how their controls meet the FEAT Principles. This is especially if the models could have an impact on the prudential soundness of the FI or cause consumer detriment. FIs that do not use such models extensively may calibrate their risk management processes accordingly.

[8] This could be due to a variety of reasons depending on how the model was built. For example, females may be less likely to apply for loans and hence the model may not be able to predict as accurately if the training data for females is small, or the model may have been trained on human-adjudicated data where the gender biases already existed, or it may have been trained on past data which indicated that females had lower wages and/or education level.

[9] FEAT applies to AIDA solutions. The thematic review focused on AI/ML use cases, which is a narrower scope.

1.12 This information paper shares observations and good practices noted during the thematic review, taking into account the current maturity of AI/ML deployment by FIs. FIs may take reference from these as they build their foundation for more extensive AI/ML adoption in future.

1.13 Sections 3A, 3B and 3C of the paper reflect observations from the three themes of scope, materiality and fairness respectively. Section 3D highlights broader observations around governance, resourcing and other challenges. Recommendations are summarised in boxes in the paper.

3 Observations from the Thematic Review

3.1 FIs are in the early stages of implementing AIDA-specific risk management processes, given that AIDA models are not widely used for decision-making. Some foreign FIs have principles and controls on AIDA that have been defined at their Head Offices. These are guided by the group's organisational values, as well as AIDA principles from relevant jurisdictions, and generally encompass aspects of the FEAT Principles. Other FIs have incorporated the FEAT Principles within their existing governance frameworks, or are in the process of setting up frameworks to address them. The full FEAT assessment methodologies were only released in February 2022[4].

3.2 With regard to the governance of AIDA models, some FIs leverage existing governance structures (e.g. risk committees), while others set up dedicated committees (e.g. AI councils). Regardless of approach, the relevant structures or committees should have the expertise and appropriate level of authority to govern the AIDA models. Other good practices observed include having the policies and frameworks approved at a sufficiently senior level (e.g. board committees), and having a representative from the Singapore office at relevant group committees to provide local perspectives.

A Scope

FIs that implement frameworks to determine whether models are in scope for materiality and fairness assessments should ensure that these frameworks are sufficiently comprehensive to take into consideration all relevant AIDA models.

FIs with separate governance frameworks for regulatory models[10] and AIDA models should consider whether models under the former have AIDA features and should also be subject to materiality and fairness assessments.

3.3 The typical process that FIs use to manage their AIDA models starts with identifying the models that are in scope for assessment. For FIs that deploy a wide range of models, determining whether a model is in scope helps them to consistently select and manage models that should be subject to the subsequent materiality and fairness assessments. FIs

[10] Regulatory models are models subject to specific MAS regulatory requirements such as MAS Notice 637.

may establish an internal definition of models that are in scope, and they have generally included a combination of the following factors:

(i) Whether the model uses AIDA
(ii) Whether the model is in production
(iii) Whether the model impacts customers, employees or the FI

3.4 In incorporating the FEAT Principles, FIs either overlay their existing frameworks with the principles, or set up separate frameworks or governance structures for them. However, for one of the FIs, regulatory models are not in scope and hence not subjected to the materiality and FEAT assessments.

3.5 A key consideration in determining whether models are in scope should be whether models use AIDA, regardless of the types of models. As regulatory models could also be using AIDA, FIs should consider whether these models should be in scope for the subsequent materiality and FEAT assessments, and build this consideration into their governance structures and processes.

3.6 Not all FIs have a framework to determine whether a model is in scope. At FIs where the use of AIDA models is not extensive, some have chosen to simply subject all models to the materiality assessments. This approach may not be sustainable if the number of AIDA models used by the FI increases beyond a certain point. An appropriate scope framework can facilitate a more effective risk-focused approach.

B Materiality

FIs should adopt frameworks for materiality that take into account a comprehensive range of factors.

3.7 The FEAT Principles specify non-exhaustive examples of factors that may be used when determining materiality:

(i) Extent to which AIDA is used in decision-making;
(ii) Complexity of AIDA model;
(iii) Extent of automation of process of AIDA-driven decision-making;
(iv) Severity and probability of impact on different stakeholders, including individuals;
(v) Monetary and financial impact;
(vi) Regulatory impact; and
(vii) Options for recourse available

3.8 The Veritas methodology further recommends the consideration of:

(i) Reputational risk; and
(ii) Use of personal data

3.9 FIs vary in the number of factors considered in their materiality frameworks, with one FI using only two factors – the extent of automation and the severity of impact. For that FI, only the models that are of high automation *and* impact would be subject to a

higher degree of governance. This could run the risk of applying a lower degree of governance on models once there is a human-in-the-loop, even if the impact could be significant. Adding a human-in-the-loop does not always eliminate FEAT or fairness risks.

3.10 FIs should consider a comprehensive set of factors and use more flexible criteria for classifying a model's materiality. For example, some FIs have applied a scoring system based on a broad set of factors, and classified a model to be of high materiality when the total score exceeds a certain threshold.

Where possible, FIs should adopt frameworks for materiality with quantitative parameters to supplement qualitative ones, to facilitate more consistent and holistic assessments.

3.11 In determining materiality, the FIs generally have established some qualitative guidance on how their models should be classified. For example, in determining the severity of impact, one FI has classified models that "can result in the withholding of a product or service" or "can result in systematic disadvantage to individuals or groups" to be of high impact, while models that "have no direct impact on customers" or "target customers but the content and purpose is benign and customers can opt in or out" to be of low impact.

3.12 In addition to such qualitative criteria, some FIs have used quantitative parameters, such as the dollar impact in the case of financial loss, duration of negative news coverage in the case of reputational impact, and percentage likelihood to determine the probability of harm. While some estimates are needed in quantifying these parameters, having the quantitative criteria helps to align the understanding across different stakeholders who perform and validate the assessments, hence promoting greater consistency. These quantitative criteria also complement qualitative ones by providing additional perspectives that contribute to a more holistic materiality assessment.

C Fairness

FIs may apply a calibrated approach when performing FEAT or fairness assessments based on the materiality of their models.

3.13 The FIs with materiality frameworks have used the materiality assessments of their models to guide them on the extent of the fairness assessments necessary for those models. Some FIs have a binary framework, where full fairness assessments are performed for high materiality models, while no assessments are done for the low materiality ones.

3.14 The purpose of risk-tiering is to facilitate meaningful resource allocation commensurate with the materiality of the models, and ensure that models with higher risks are accorded more resources and attention. Based on the materiality of the models, FIs may calibrate the extensiveness of their fairness assessments according to their levels of risk.

3.15 The customisation of the fairness assessments may include changing the assessment process itself and/or its associated governance requirements. Examples of how FIs can customise the assessment process include:

- Requiring more extensive analysis and more thorough responses to the higher risk models, and short summary answers to some or all the assessment questions for lower risk models; and

- Having more frequent and in-depth monitoring of the models' performance and FEAT metrics for higher risk models.

3.16 In terms of customising the governance requirements, FIs could implement more stringent review and approval processes which may be reflected in the number and levels of parties required to sign off the models for deployment; require greater levels of escalation and intervention for deviations from pre-agreed model performance or FEAT metric thresholds; and involve independent validation through third party audits, for the higher materiality models[11].

FIs' fairness assessments should be substantiated and supported by adequate justifications. This will facilitate subsequent validation.

3.17 Most FIs have developed checklists or scorecards to facilitate the fairness assessments of their models. These checklists and scorecards typically include questions to review the business and fairness objectives of the models; examine the data and models for biases; calculate fairness metrics to determine whether there is any disadvantage between different groups or individuals; justify the use of personal attributes; and examine the monitoring and review of models so that they behave as designed and intended.

3.18 Some FIs have developed checklists or scorecards with predominantly Yes/No questions that do not require explanations or justifications. While simple Yes/No answers reduces the burden of documenting the fairness assessments, the lack of explanations or justifications makes it hard for a reviewer to understand the considerations and context of the assessment. Hence, in designing questions for their checklists or scorecards, FIs should allow room for justifications of the Yes/No answers, especially for high materiality models. This will allow an independent second line of defence function to better validate the assessments. For low materiality models, simple justifications may suffice.

FIs should consider specifying a list of protected attributes and their proxies, the use of which is subject to justifications and approvals, as a safeguard against discrimination.

3.19 Protected attributes are features that are typically prohibited to limit the risks of unfair discrimination, but could nevertheless have legitimate use cases. Common protected

[11] Veritas Documents 1 Section 3.3 and Document 3 Section 3.2.3 provide some suggestions on how this calibration can be performed.

attributes include gender, race and age. Proxies for protected attributes may also pose similar risks. For example, zip codes are proxies for race in some countries as racial groups may be clustered in certain neighbourhoods, and years of driving experience can be a proxy for driver's age. In line with Principle F2 of the Fairness Principles, the use of protected attributes and their proxies as input factors for AIDA-driven decisions should be justified.

3.20 FIs have used protected attributes and their proxies as input variables for some of their models. Common reasons cited for the use of protected attributes and their proxies include the general acceptance of the use of such attributes for certain use cases (e.g. using age as an input variable for a motor insurance pricing model) and the necessity of the use of such attributes (e.g. demographic data are needed to allocate customers to the right products). FIs need to differentiate protected attributes and their proxies from other features when they perform the fairness assessments so as to provide justifications and obtain approval for their use. Hence it will be beneficial for FIs to specify a list of protected attributes and their proxies at the organisation level, so that they can be consistently identified. The use of protected attributes and their proxies in models should also be justified by taking into consideration the business objectives and data, as well as the performance and fairness measures. Appropriate approvals should be sought on their use to safeguard against potential discrimination.

FIs should review models for fairness on a regular basis, especially those with high materiality.

3.21 An AIDA model that has been extensively tested for its performance and fairness prior to deployment, may degrade over time. There could also be unintended consequences that only arise after the model is deployed. This degradation in performance and fairness can be detected through monitoring, which is especially important for models involved in high-volume or high-consequence decisions. Monitoring and review are also key components of the Fairness Principles – Principle F3 requires data and models used for AIDA-driven decisions to be regularly reviewed and validated for accuracy and relevance, and to minimise unintentional bias, and Principle F4 requires that AIDA-driven decisions be regularly reviewed so that models behave as designed.

3.22 Some FIs review their models on a regular basis for performance, but did not do the same for fairness. Reasons for not assessing models for fairness on a regular basis included data availability (e.g. difficulty in obtaining reject inference data for credit models), resource constraints, and time needed to implement and operationalise the Fairness Principles.

3.23 FIs should implement a process to review models for fairness on a regular basis, especially for high materiality models. Models and data may evolve over time and result in shifts in the fairness considerations and measures.

FIs should base their fairness objectives and measures by comparing across individuals and groups.

3.24 The FEAT Fairness Principle F1 states that "individuals or groups of individuals are not systematically disadvantaged through AIDA-driven decisions, unless these decisions can be justified". Demonstrating alignment with this principle requires FIs to identify the individuals and groups that may potentially be subject to systematic disadvantage, determine the harms and benefits created by the system, and define how the distributions of these harms and benefits should be measured across individuals and groups to assess systematic disadvantage.

3.25 One FI assessed the fairness of their model by comparing the model's output against a human assessor's decision. This measures the extent of adherence of a model's decision to that of a human, but does not measure whether the model is fair across individuals and groups. Instead, FIs should measure fairness by assessing the differences in harms and benefits across individuals and groups, such as by comparing the models' predictions for males versus females, across different age groups, and other protected attributes that may have potential fairness considerations. This method of determining fairness is in line with standard fairness measures available in literature[12].

D Other Considerations

FIs should incorporate FEAT considerations during the model development life cycle and develop a roadmap to put existing models through the FEAT assessments.

3.26 Many FIs are still in the process of developing frameworks and governance around the Fairness Principles, and operationalising them through the process of classifying models based on their materiality and performing the fairness assessments. For FIs with large numbers of existing models, the effort to perform the materiality and fairness assessments is not trivial.

3.27 To manage this effort, FIs should consider the following:

- For new models, incorporate FEAT considerations during the model development life cycle, i.e. when new models are developed, or when a decision is made to put model into production.

- For existing models, develop a roadmap or plan to put models through the materiality and FEAT assessments. The roadmap should take into consideration the extent of use of AIDA in the FI and how it plans to expand its use.

FIs may also consider the use of the Veritas toolkit[7], which is an open-source software that can be used to automate the fairness assessment.

[12]Refer to Veritas Documents 1, 2 and 3A.

FIs should set up appropriate structures and processes for an independent, or second line of defence function, to perform validations on the materiality and fairness assessments.

3.28 The three lines of defence model and the segregation of duties between development, validation and independent review also apply to AIDA models. There should be independence between the AIDA model developers and the AIDA model validators who review the fairness assessments presented to them.

3.29 The FIs generally have governance structures or committees to oversee their AIDA models, validate them for performance and fairness, and approve their deployment for use. However, some FIs perform the model validation within the business units themselves, meaning that there was no second line of defence to perform independent assessments.

3.30 While FIs may still be adjusting their governance structures to incorporate the Fairness Principles, they should set up appropriate structures and processes that encompass second line of defence responsibilities when their use of AIDA models become more pervasive.

FIs should ensure that the units involved in assessing and validating models against the FEAT Principles are adequately resourced and have the necessary skillsets as they scale up their use of AIDA.

3.31 Related to the point above, some FIs may not have adequate resourcing in terms of headcount and experience to perform the assessments and validations of their AIDA models given that this is a rather new area. This may result in incomplete, erroneous or delayed implementation of the Fairness Principles. FIs should develop a plan to build up the needed resources and skillsets for FEAT if they do intend to scale up their use of AIDA.

4 Conclusion

4.1 FIs' use of AIDA must be accompanied by good governance and risk management, as well as sustainable strategies. The FEAT Principles have been developed to assist FIs in deploying AIDA in a responsible manner. In line with a risk-based approach, FIs that use or have plans to use AIDA models extensively in decision-making should take into account the recommendations in this paper as they incorporate the FEAT Principles in their AIDA policies and workflows.

4.2 The work around the responsible use of AIDA in the financial sector is ongoing and will involve the active participation of various stakeholders including FIs, FinTechs and MAS. MAS will continue to work with the industry to promote the responsible use of AIDA through the effective implementation of the FEAT Principles.



Monetary Authority of Singapore