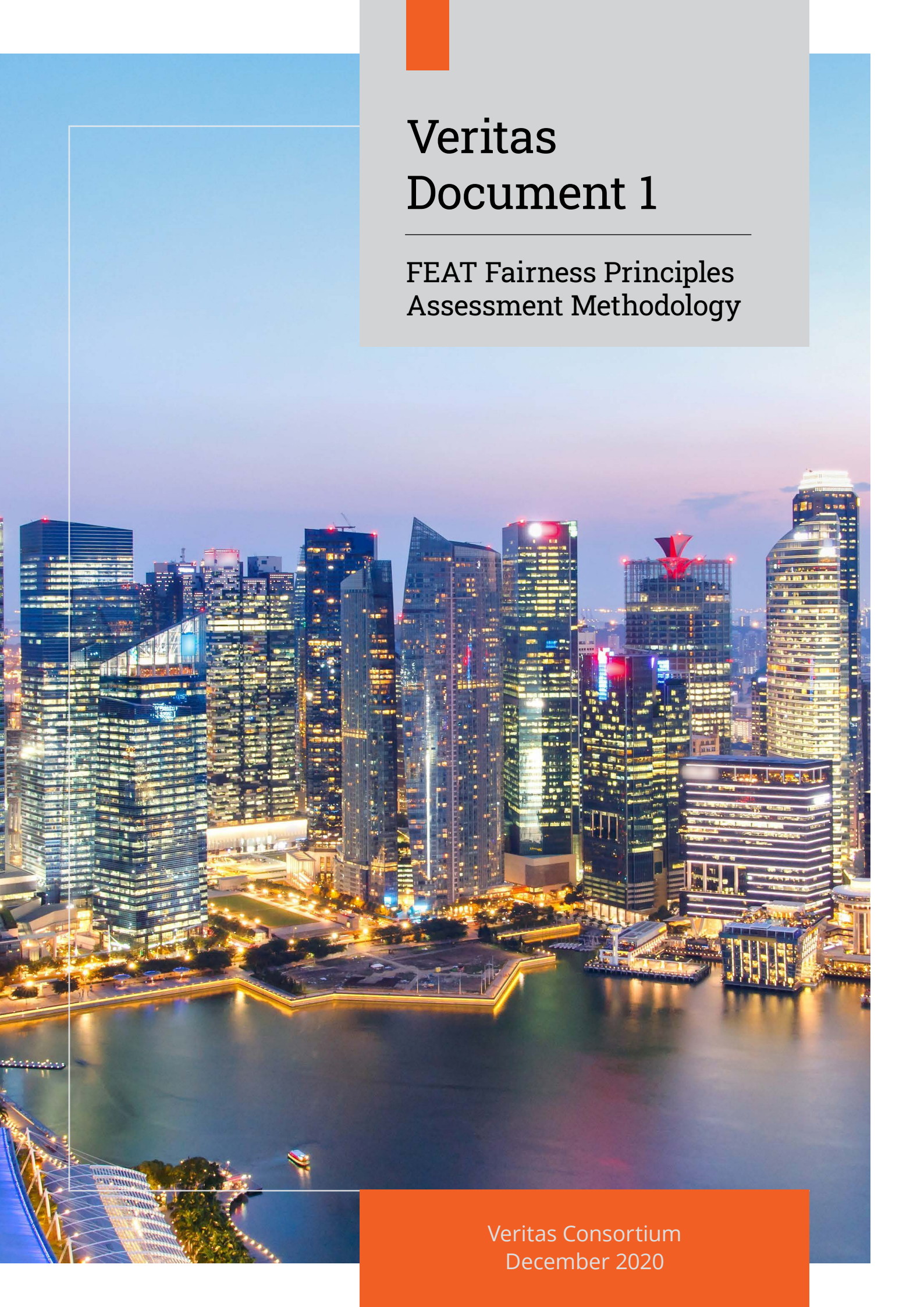


A nighttime photograph of a dense urban skyline, likely Singapore, featuring numerous illuminated skyscrapers and modern buildings. The city is reflected in the water in the foreground. The sky is a mix of blue and purple hues from the twilight.

Veritas Document 1

FEAT Fairness Principles Assessment Methodology

Foreword





Veritas consortium, comprising the Monetary Authority of Singapore (MAS) and industry partners, are pleased to present the Veritas* phase one white papers to assess the FEAT Fairness principles.

Veritas consortium, comprising the Monetary Authority of Singapore (MAS) and industry partners, are pleased to present the Veritas* phase one white papers to assess the FEAT Fairness principles. These white papers contain the collective experience of the Veritas consortium and share the methodology and open source code for validating the fairness principles against two selected use-cases of credit risk scoring and customer marketing.

The Veritas initiative was started/announced in 2019 to enable financial institutions to assess their Artificial Intelligence and Data Analytics (AIDA)-driven solutions against the principles of fairness, ethics, accountability and transparency (FEAT). The FEAT principles were co-created by MAS and the financial industry to provide guidance to firms offering financial products and services on the responsible use of AI and data analytics, and to strengthen internal governance around data management and use.

The Veritas initiative is the first step towards a framework to operationalise the FEAT principles in guiding the responsible use of AIDA in financial services and translate these guidances into practical implementation of the FEAT principles. The methodology described in these white papers serve to bring about perspectives the financial industry must keep in mind when putting AIDA concepts and systems into practice, and brings the philosophical discussions about fairness to reality.

Beyond this phase, the consortium will extend this effort and expound the Veritas methodology for other FEAT principles to additional use-cases in financial services, such as insurance. MAS remains committed to supporting the industry in developing practical implementation of the FEAT principles and supporting financial institutions in validating their AIDA solutions against the FEAT principles.

We would like to express our thanks to all members of the consortium for their active participation and generous support in the development of this report. We would also like to acknowledge the contributions of Element AI, EY, HSBC, IAG Firemark Labs & Gradient Institute and UOB in this ground-breaking endeavour.

As we continue into the next phases of Veritas, we encourage the financial industry and technology firms to read about the accomplishments and learnings of this project, and share their experience and feedback with the consortium to further advance the trustworthiness of AIDA solutions in our financial landscape. We also hope that these whitepapers will help you better implement the FEAT fairness principles in your organization.

MAS CFTO

*Veritas was announced as part of Singapore's National AI Strategy by Mr Heng Swee Keat, Deputy Prime Minister of Singapore, at the Singapore FinTech Festival and Singapore Week of Innovation and TeCHnology (SFF x SWITCH) 2019.

Executive Summary





Artificial intelligence and data analytics (AIDA) are technologies that assist or replace human decision-making. AIDA systems can give FSIs the ability to learn complex relationships from large volumes of data and perform quantitative risk-aware decision-making at a vast scale.

The use of AIDA systems by Financial Services Institutions (FSIs) can create new benefits for stakeholders but also introduces new risks from data-driven automation.

Artificial intelligence and data analytics (AIDA) are technologies that assist or replace human decision-making. AIDA systems can give FSIs the ability to learn complex relationships from large volumes of data and perform quantitative risk-aware decision-making at a vast scale. These technologies can improve business processes, allowing FSIs to serve customers better and more efficiently, and potentially avoid the variability from individual humans making decisions. However, AIDA systems operate according only to the data and objectives explicitly provided by their designers and do not understand the broader context or moral constraints that humans take for granted, and adding a “human in the loop” does not automatically eliminate all fairness risks. As a result, AIDA systems can cause unintended harms and perpetuate or reinforce existing disadvantages that exist in society, which in turn can create reputational, operational, and legal risks for FSIs. The heightened risks for businesses and customers arise whenever consequential decisions are being made at high speed or volume, whether the underlying decision-making algorithms are relatively simple or more complex (such as neural networks trained with machine learning).

MAS released a set of principles for the use of AIDA and convened the Veritas consortium to support FSIs in implementing the principles. This report focuses on fairness principles.

In 2018, the Monetary Authority of Singapore (MAS) released a set of principles (co-created with the financial industry and other relevant stakeholders) to promote Fairness, Ethics, Accountability and

Transparency (FEAT) in the use of AIDA by the financial sector. The FEAT Principles are not prescriptive, and recognise that FSIs will need to contextualise and operationalise the governance of AIDA in their own business models and structures. To help FSIs do so, MAS convened the Veritas consortium, which includes FSIs and technology partners, to create guides and tools that support the implementation of the FEAT Principles. This report, the first publication of the consortium, focuses on the four FEAT Fairness Principles:

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.

F3

Data and models used for AIDA-driven decisions are regularly reviewed and validated for accuracy and relevance, and to minimise unintentional bias.

F4

AIDA-driven decisions are regularly reviewed so that models behave as designed and intended.



The report was developed with expertise from a variety of areas including: AIDA practitioners and academics, industry perspectives from FSIs and domain experts, and policy and legal expertise from compliance experts. This document was also designed so that FSIs following Personal Data Protection Commission (PDPC) guidance are not acting in contradiction to any guidance found here.

The FEAT Fairness Principles ask FSIs to consider both whether the decision-making process and the outcomes from AIDA-driven decisions are justified.

AIDA systems encompass more than mathematical models and data. An AIDA system often includes downstream or upstream business rules (such as eligibility criteria), employees who may review or change automated decisions, and processes and mechanisms to monitor and review decisions and outcomes.

AIDA systems produce decisions, such as approving a customer's loan application, which in turn create outcomes, such as a customer having access to credit or in some cases defaulting. The four FEAT Fairness Principles are concerned with both the ways in which a system makes those decisions (for instance by asking whether the use of personal attributes is justified) and the outcomes that follow (for example by asking that any systematic disadvantage produced by the system be justified). What types of decision-making processes and outcomes are acceptable or desirable will be context-specific and FSIs may reasonably come to different conclusions based on their company values, business strategies, and technical maturity. For example, organisations may set goals to achieve higher levels of fairness as they develop more expertise with AIDA and FEAT. Nonetheless, in beginning to align their AIDA systems with the FEAT Fairness Principles any FSI will need to consider a specific set of measures and tradeoffs that go beyond traditional risk management practices. For instance, an AIDA system can achieve high overall accuracy while applying the majority of its erroneous decisions to a disadvantaged minority population. As a result, members of this population might have worse outcomes (as compared to the average) simply

because their relative rarity in the data would make their outcomes more difficult for the system to predict. The FEAT Fairness Principles are concerned with whether and how this type of outcome distribution and data usage are justified.

There is no single answer about what is “fair” so FSIs will need to separately (a) define their fairness objectives and (b) measure how well AIDA systems achieve these objectives.

There is no single, agreed upon definition of fairness or formula for calculating disadvantage or bias, even for a particular AIDA use case. In translating the FEAT Fairness principles into practice, FSIs are faced with a series of questions such as: How should groups be defined? How should differences between groups be measured? How big of a difference is acceptable? How much should be invested in reducing differences? In almost all cases it will be impossible to build a “perfectly fair system” that avoids any differences between any groups, especially because the many mathematical measures of fairness usually cannot be satisfied simultaneously. Tradeoffs are mathematically inescapable, and cannot be avoided in either human or AIDA decision-making. As a result, the goal of implementing AIDA fairness principles is not “making AIDA systems fair”, but about improving the operation of an AIDA system with respect to the fairness objectives chosen by its operators, and ensuring those objectives align with the values of the organisation and the society in which the system operates.

A comprehensive five-part methodology to assess alignment with the FEAT Fairness Principles has been developed that can be applied to AIDA systems.

Many AIDA systems deployed in FSIs present significant and novel fairness risks. Credit scoring systems, especially those using non-traditional data and modern machine learning techniques, are one such example. Credit scoring fairness risks arise from the systems: involving consequential decisions; having evidence in other jurisdictions of historical discrimination; and showing patterns of systemic disadvantage in existing deployments. For instance, one

study from the found that lenders charge otherwise-equivalent borrowers from some demographic groups higher rates for mortgages, and that while algorithms discriminated less than face-to-face lenders there were still significant disparities [72]

To help assess the fairness risks of consequential AIDA systems, this document presents a comprehensive 18-question assessment methodology (the “FEAT Fairness Assessment Methodology”). Questions in the five parts (A-E) of the Methodology are designed to help FSI make their objectives for fairness explicit and provide evidence that the system’s operation is aligned with those objectives:



A. Describe system objectives and context

- A1. What are the business objectives of the system and how is AIDA used to achieve these objectives?
- A2. Who are the individuals and groups that are considered to be at-risk of being systematically disadvantaged by the system?
- A3. What are potential harms and benefits created by the system's operation that are relevant to the risk of systematically disadvantaging the individuals and groups in A2?
- A4. What are the fairness objectives of the system, with respect to the individuals and groups in A2 and the harms and benefits in A3?

B. Examine data and models

This section references AIDA models which are algorithms (such as a linear regression or set of business rules) used within AIDA systems (such as a credit scoring system). AIDA systems may have, as components, one or more AIDA models.

- B1. What errors, biases or properties are present in the data used by the system that may impact the system's fairness?
- B2. How are these impacts being mitigated?
- B3. How does the system use AIDA models (with, or separately from, business rules and human judgement) to achieve its objectives?
- B4. What are the performance estimates of the AIDA models in the system and the uncertainties in those estimates?
- B5. What are the quantitative estimates of the system's performance against its business objectives and the uncertainties in those estimates?

C. Measure disadvantage

- C1. What are 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 A2 and the potential harms and benefits in A3?
- C2. What are the achievable tradeoffs between the system's fairness objectives and its other objectives?
- C3. Why are the fairness outcomes observed in the system preferable to these alternative tradeoffs?

D. Justify the use of personal attributes

- D1. What personal attributes are used as part of the operation or assessment of the system?
- D2. How did the process of identifying personal attributes take into account ethical objectives of the system, and the people identified as being at risk of disadvantage?
- D3. For 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?

E. Examine system monitoring and review

- E1. How is the system's monitoring and review regime designed to detect abnormal operation and unintended harms to individuals or groups?
- E2. How does the system's monitoring and review regime ensure that the system's impacts are aligned with its fairness and other objectives (A1 and A4)?
- E3. What are the mechanisms for mitigating unintended harms to individuals or groups arising from the system's operation?

To illustrate the application and relevance of the Methodology, it was applied to two simulated data sets and models, and two real, in-production AIDA systems in participating FSIs from the Veritas consortium. These case studies are available in a second, accompanying document (Document 2: Assessment Case Studies).

The risk associated with the AIDA systems should be taken into account when using the Methodology to conduct an assessment.

There is a broad set of use cases for AIDA systems. The risk associated with individual AIDA systems differs on a case-by-case basis and may be impacted by several factors including:

- **materiality** of the system for the FSI, which may be assessed based on the social, reputational, economic, or regulatory consequences of a potential system error or the number of users that apply the system
- **impact** of the AIDA system based on the number and type of internal and external stakeholders that can be affected as well as the probability and severity of potential harms
- **complexity** of the system related to the type and amount of ingested data sources, the mathematical computation, or the nature of human intervention in decision-making

FSIs may define a set of standardised factors based on their risk management principles to categorise AIDA systems by their level of risk. In line with a risk-based approach, FSIs can reasonably apply less onerous customisation of the Methodology to AIDA systems in the lower risk buckets or exclude systems from assessment, if deemed adequate.

In line with general good governance and risk management practices, independent assessors should review and approve the results of an assessment.

The three lines of defense model, as well as the associated segregation of duties between development, review/validation and independent

evaluation, also apply to AIDA systems. Therefore, a 2nd Line function of the FSI should be responsible for defining the applicable risk and control standards for fairness based on the FSIs principles and relevant external requirements. An assessment of how well an AIDA system aligns with the FEAT Fairness Principles should have some degree of independence between the AIDA System Developer (whether this is a team within the FSI or a vendor) and AIDA System Assessors who review the fairness assessment presented to them by the AIDA System Owner (the FSI or business unit using the AIDA system to help make decisions to fulfill their business goals).

The Assessor role could be played by an external audit firm, or internally by a model risk management, internal audit group or other independent function with the according competency. An internal assessor might be purely focused on whether an AIDA system meets the FSI's self-defined fairness goals, whereas an external assessor (if used) might also be interested in whether the goals themselves have been well-defined and align with industry standards (if and where these exist). Assessors should provide their opinions to the appropriate level of management of the AIDA System Owner. For systems with significant risks to customers or the FSI, this might be at the senior management or Board level, whereas for other cases with lower risk levels, the model risk management function may be sufficient.

Implementation of the FEAT Fairness Assessment Methodology and review process could be led by an AIDA center of excellence, the model risk management group, or an ethics body.

One or more groups within an FSI will need to lead the process of requesting assessments from AIDA System Developers and AIDA System Owners, and of supporting any resulting interventions or changes requested by the AIDA System Assessor. FSIs are taking different approaches. Three potential strategies are:

- **Embed fairness considerations within a centralised AIDA governance function.** Some FSIs have established AI centres of excellence which are responsible for developing, sourcing, and managing AIDA systems across their lifecycle.

These technologists can extend their scope of responsibility from technical requirements to also cover fairness goals.

- **Extend the technical Model Risk Management (MRM) framework to include fairness considerations.** The financial sector has well-established MRM practices, and most FSIs have established committees (e.g., model governance committee, model adequacy committee or model oversight committee) that are experienced in receiving reports from model developers, performing technical validations and making recommendations. This function could be extended to include fairness assessments.
- **Extend the ethics and conduct framework for employees to include AIDA systems.** Most FSIs have existing frameworks dedicated to ethics and employee conduct, and another FEAT Principle recommends that “AIDA-driven decisions are held to at least the same ethical standards as human driven decisions” [57]. An ethics body will often already have expertise in social and ethical considerations and so could also lead assessments of AIDA fairness, with support from more technical groups.

AIDA System Owners may ask AIDA System Developers to make design choices and interventions based on the results of the fairness assessment; these choices should consider long term effects.

Once fairness objectives have been selected and measured, the challenge remains to design or update AIDA systems to try and more fully achieve them. One design approach to creating fairer AIDA systems that is common but not generally effective is “fairness through unawareness”. This refers to the removal of personal attributes from an AIDA model to prevent the model using them to discriminate. This is ineffective because a model may infer the personal attribute from other variables even if it is omitted, and when sensitive personal attributes are predictive of good outcomes for some groups, removing them can lead to more discriminatory outcomes. Similarly, simply adding human intervention or oversight to AIDA decision-making is not guaranteed to improve fairness. This strategy may be ineffective if, for example, the intervening humans have their own unconscious biases.

The choice of effective intervention strategies will vary depending on the use case, AIDA system, findings of fairness assessments and other factors, but three categories should be considered:

- **data collection and preparation:** collecting more data, especially for underrepresented groups, is one way to mitigate selection and evaluation biases that can create systematic disadvantage. If additional data collection is impractical, there are statistical techniques such as data reweighting that can be helpful. Practitioners can also consider changes to how the data is preprocessed. For instance, personal attributes may be changed from being binary (e.g., “white” and “not white”) to codings that can handle multiple categorisations that better reflect people’s identities.
- **model design and development:** the behaviour of an AIDA system will be driven by what it is optimised to do within the constraints it is given. If a practitioner is looking to mitigate some aspect of unfairness, one strategy is to add regularisation terms or constraints that enforce a particular fairness objective. It may also be possible to constrain how the model learns in order to “mask” out a personal attribute.
- **post-processing:** in many AIDA systems, there are steps between the output of a model and a real-world outcome. For instance, credit scores output default probabilities, which are combined with other business rules to make actual approval decisions. These intermediate steps offer additional opportunities for practitioners to intervene. For instance, different decision thresholds can be set for different groups (i.e. an individual in one group might need to be in the top 80% of applicants to be approved while individuals in a less well represented group might only need to be in the top 70% of applicants in their group)

All strategies (pre-, in-, and post-processing) may change outcome patterns but create their own issues. For instance, if historical bias has meant that members of one group on average have less wealth or income than another group, initiatives to lower approval thresholds may increase the disadvantaged group’s access to credit in the short term, but also raise their default rate over time. These types of long term effects and feedback loops should be considered when intervening in AIDA systems.

Future work will focus on ensuring the generalisability and practicality of the FEAT Fairness Assessment Methodology, and providing integrated guidance on the other FEAT Principles.

This first phase focused on generating practical guidance for implementing the FEAT Fairness Principles. The Methodology was developed and tested for two use cases (credit scoring and customer marketing) with two FSI Veritas consortium members. The consortium intends to extend and refine the

Methodology for new use cases and with other types of FSIs (such as Insurers). Additional future work will focus on FEAT Principles related to Ethics, Accountability, and Transparency, as well as the relationships between these Principles.

This document provides guidance to support FSIs' adoption of the FEAT principles. It does not preclude MAS, in its capacity as a financial regulator, from assessing the adequacy of FSIs' implementation of the FEAT principles and recommending areas for improvement where appropriate.



Contents

01	Introduction	14
1.1	AIDA systems in context	16
1.2	This Document	19
02	A Risk-Based Approach to FEAT Fairness Assessment	22
2.1	Overall process for applying the FEAT Fairness Assessment Methodology	23
2.2	Existing AIDA risk management practices	27
2.3	Integrating the Methodology with existing risk management	30
03	A Methodology to Assess AIDA Systems	32
3.1	Key concepts related to FEAT Fairness Principles	34
3.2	The FEAT Fairness Assessment Methodology	42
3.3	Guidance on scaling the Methodology to system risk	56
	Acknowledgements	58
	Bibliography	62
	Appendices	70
	Annexure	90



01 Introduction





In 2018, the Monetary Authority of Singapore (MAS) released a set of principles for the responsible use of Artificial Intelligence and Data Analytics (AIDA) in the financial sector.



In 2018, the Monetary Authority of Singapore (MAS) released a set of principles for the responsible use of Artificial Intelligence and Data Analytics (AIDA) in the financial sector. These principles promote Fairness, Ethics, Accountability and Transparency (FEAT) in AIDA systems and provide high-level guidance to Financial Services Institutions (FSIs) on how to design, operate and review them.

This document aims to help FSIs assess how their AIDA systems align with the four FEAT Principles that address fairness. The document presents a step-by-step methodology that assists FSIs to collate the information and analysis required for an assessment, and provides guidance for internal or external assessors or governance bodies on how to conduct an assessment. Case studies of credit scoring and customer marketing use cases are also provided (in Document 2) to illustrate the application of the Methodology.

1.1 AIDA systems in context

Increasing use of AIDA

The increasing use of AIDA is enabling FSIs to deliver better experiences and more resilient services to their customers. It is also improving the efficiency of back-office support functions to ensure security, manage risks, streamline operational processes and allocate resources more efficiently. A global AI survey in financial services jointly conducted by the Cambridge Centre for Alternative Finance at the University of Cambridge Judge Business School and the World Economic Forum found 77% of FSIs are expecting AI to become an essential driver with a high level of importance to their business in two years' time [77].

Adopting more sophisticated predictions, personalisations and insights with AIDA is becoming a necessity for FSIs, rather than an advantage. These institutions are facing transformation and disruption from an array of traditional competitors and new entrants like FinTechs, InsurTechs, and digital banks. External factors have also hastened the adoption of AIDA systems: the recent (and ongoing) COVID-19 pandemic has disrupted our economy and way of working, forcing the rapid adoption of new technologies to compensate for much of the workforce working remotely [81].

AIDA's novel risks

AIDA systems give FSIs the ability to automate decision-making at a vast scale. Modern AIDA techniques such as machine learning allow those decisions to be based on complex relationships learned from large volumes of data. However, without careful design and control, AIDA systems can also introduce new risks for businesses and customers. Poorly designed AIDA systems can cause unintended harms and perpetuate or reinforce existing disadvantages in our society [64, 30] (though when carefully deployed, they also offer the opportunity to reduce such harms). Some widely publicised examples of these unintended harms include systems for patient triage [63], for screening job applicants for interviews [21], for performing facial recognition [13], and for conducting web searches [62].

These harms arise from the fact that, unlike humans, AIDA systems operate according only to the data and objectives explicitly provided by their designers. AIDA systems cannot, even in principle, understand the broader context or moral constraints that humans take for granted. For example, when an AIDA system is simply designed to "be accurate", the system can still achieve that aim while applying the majority of its erroneous decisions to a disadvantaged minority population. Members of this population would have worse outcomes as a result (such as mistakenly being denied a credit product) simply because their relative rarity in the data would make their outcomes more difficult for the system to predict. In addition to harming customers, AIDA systems that unfairly discriminate risk harming the reputation of trustworthiness of an FSI [2] and the organisation and business itself [86]. Eroding the trust of customers may be particularly damaging for the financial sector, which already suffers from a trust deficit compared to other industries [28].

Addressing AIDA fairness

Fairness objectives

Addressing unfair discrimination and harms caused by AIDA systems requires designing the systems with explicit fairness objectives. These objectives must be made mathematically precise to be encoded into an AIDA system. However, for any given AIDA system, there is no one “right answer” to the question of what its fairness objectives should be.

Fairness is an “essentially contested” concept [38, 48]: whilst the general nature and importance of fairness are widely understood, its precise definition, or the outcomes that are fair or unfair in a particular situation, continue to be debated in both moral philosophy and wider society. Even for a particular person, the concept of fairness is highly context-sensitive and difficult to quantify.

In almost all cases it will be impossible to build a system that is universally agreed to be fair. Different notions of fairness often conflict so that applying one notion violates another, and there will rarely be universal agreement on which notion is preferred. Such fairness tradeoffs are made explicit by the attempts to quantify measures of fairness: of the many fairness measures in the literature, few can be satisfied simultaneously in all-but-trivial situations [74]. These tradeoffs are mathematically inescapable, and cannot be avoided in either human or AIDA decision-making.

Hence, addressing fairness in AIDA systems is not about “making AIDA systems fair”, but about improving the operation of an AIDA system with respect to the fairness objectives chosen by its operators, and ensuring those objectives align with the values of the society in which the system operates. The choice of fairness objectives for an AIDA system depends on the system’s exact purpose, the consequences of its operation, the people affected, and the values of the people responsible for its design. That choice will also likely involve making tradeoffs between different kinds of fairness for different people affected by the system.

Designing AIDA systems to meet fairness objectives

Once fairness objectives have been selected for an AIDA system, a challenge remains to design the system to try and achieve them. One design approach to creating fairer AIDA systems (that is unfortunately common but demonstrably not effective) is “fairness through unawareness”. This refers to the removal of personal attributes from an AIDA model to prevent the model using them to discriminate. The existence of “redundant encodings” in the data means that the model may infer the personal attribute from other variables even if it is omitted [15]. For example, given a history of segregation and redlining in the United States which led to racial groups being clustered in neighbourhoods, it is often the case that zip codes are proxies for race, or that they encode both location and race data. When the personal attributes are predictive of good outcomes for some groups, removing them can lead to more discriminatory outcomes [18].

Instead of simply removing information from the AIDA system that would enable it to discriminate, approaches to designing fairer AIDA systems rely on a variety of data processing and modelling approaches being developed within the emerging field of “algorithmic fairness” [71]. These approaches may, for example, remove statistical dependencies from data, modify the prediction objective of a model itself, or even modify the outputs of a model once created to alter the resulting decisions.



Developing the FEAT Principles

Designing and validating fairer and more trustworthy AIDA systems is a challenge that organisations are only beginning to address. FSIs, as owners of AIDA systems making consequential decisions about customers' lives, have an opportunity to lead the effort.

MAS recognised this opportunity and collaborated with leading FSIs to create a set of principles guiding the responsible use of AIDA in financial services. The principles consist of four key pillars – Fairness, Ethics, Accountability, and Transparency (FEAT) [\[57\]](#) and were released in 2018 as part of Singapore's National AI Strategy to build a progressive and trusted environment for AI adoption. The principles have established a standard across the financial sector in Singapore, with various jurisdictions following suit and releasing similar principles. They act as voluntary guidelines to encourage innovation, being conscious of the balance that governments and regulators have to maintain between governing the ethical use of advanced technologies like AIDA and stifling innovation.

Context behind the Veritas initiative

The first step towards building AIDA systems that align with the FEAT Principles is to establish ways to measure that alignment. To facilitate this work and accelerate the adoption of responsible AIDA in the financial services industry, MAS commissioned the Veritas initiative. This initiative assembled a consortium of FSIs, AI experts, and global technology vendors, to study and cocreate a set of practical methodologies that provide insights into how an AIDA system's operation aligns with the FEAT Principles.

1.2 This Document

This document is the Veritas consortium's first effort to provide a practical methodology for assessing the alignment of an AIDA system with the FEAT Principles. In particular, this initial work focuses on the FEAT Fairness Principles and on applications in the customer marketing and credit scoring domains. The Veritas consortium anticipates future work encompassing the rest of the FEAT Principles and targeting additional AIDA use cases.

The FEAT Fairness Assessment Methodology presented in this document is designed to account for elements of ambiguity, subjectivity and contestability in judging a system's fairness, acknowledging the impossibility of removing all forms of unfair treatment and outcome, and that fairness is but one of many competing objectives relevant to a system's operation. Rather than provide a set of rules or criteria, the Methodology asks FSIs to explicitly present the objectives and constraints governing the fairness of the system, then suggests analyses that enable humans to judge the system's performance against those objectives.

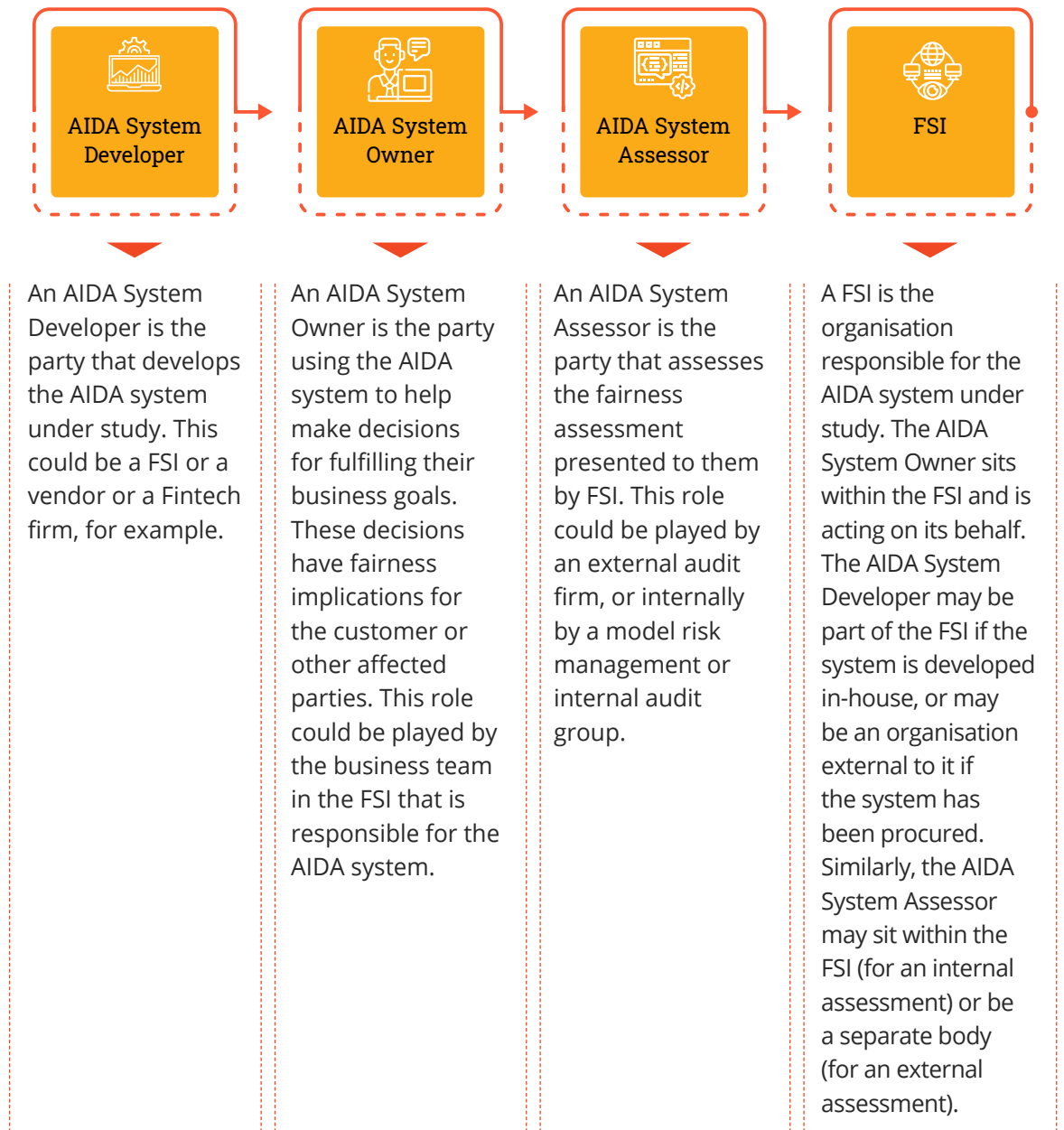
The Methodology is presented as questions and considerations that step FSIs through the assessment process. The considerations provide detailed guidance to assist practitioners in developing responses to the questions. The answers to the questions in the Methodology should provide internal or external assessors with the information and analysis they need to judge the system's alignment with the FEAT Fairness Principles. Case studies of example systems in customer marketing and credit scoring are also presented to illustrate the application of the Methodology, along with additional use case specific considerations.

The scope of work applying the Methodology is specific to the system under study and is designed to be calibrated to the level of risk associated with the system's operation. While AIDA as described by the FEAT Principles encompasses a broad range of use cases, each FSI would apply the Methodology based on how sophisticated their models are, the impact on customers and on the financial performance or reputation, and an FSI's individual standards and governance principles (some guidance on integrating the Methodology into existing risk management structures is provided in Section 2.3).

This document provides guidance to support FSIs' adoption of the FEAT principles. It does not preclude MAS, in its capacity as a financial regulator, from assessing the adequacy of FSIs' implementation of the FEAT principles and recommending areas for improvement where appropriate.

Audience of this document

This document distinguishes between four groups involved in AIDA system assessments in different ways:



These groups may already be present in an FSI's existing risk management framework, or may need to be assigned. In either case, they have roles consistent with the standard model development life cycle of system development; verification, validation, and deployment; and ongoing operation and monitoring (note that for assessments on pre-existing systems, the first phase of system development is assumed).

Navigating this document

The structure of the document is as follows: after this introduction, Section 2 expands on the risk-based approach to FEAT fairness assessment taken by the Methodology. The section includes background on the current market practice for AIDA risk management in FSIs, guidance for developing scoping and scaling processes to apply the Methodology to systems with different risk levels, and guidance on integrating fairness assessment into existing risk management practices.

Section 3 introduces the FEAT Fairness Principles and presents the FEAT Fairness Assessment Methodology. The section begins by defining key terms in the principles, then moves to motivate and present the five-part Methodology. Each part of the Methodology also contains a set of considerations to assist AIDA System Owners, Developers, and Assessors in performing and reviewing an assessment.

Finally, a series of appendices provide additional technical reference material to complement the Methodology ([Appendix 1](#)), advice on mitigation strategies and actions to improve AIDA system fairness ([Appendix 2](#)), resources for software implementations, fairness measurement and mitigation ([Appendix 3](#)), and references for further reading ([Appendix 4](#)).

Accompanying this document are a set of “FEAT Fairness Assessment Case Studies” (Document 2). Document 2 presents complete, detailed, case studies of applying the Methodology to hypothetical but realistic systems: one for customer marketing and another for credit scoring. It also provides reflections from two FSIs (HSBC for customer marketing and UOB for credit scoring) that applied the Methodology to their real systems, as well as series of use case specific considerations to assist FSIs in completing assessments of customer marketing and credit scoring systems.



02 A Risk-Based Approach to FEAT Fairness Assessment





2.1

Overall process for applying the FEAT Fairness Assessment Methodology

Identifying AIDA models in scope for assessment

The ultimate scope of the FEAT Fairness Assessment Methodology is defined by the FEAT Principles [57], which state that:

“AIDA” refers to artificial intelligence or data analytics, which are defined as technologies that assist or replace human decision-making. While this definition is broad, the Principles in this document can apply to both complex and straightforward AIDA techniques.’

Since AIDA systems refer to a broad set of techniques and have no standard definition across the industry, FSIs may establish an internal definition of applications that are in scope of their AIDA risk management framework.

However, some systems identified by an FSI as AIDA may have low fairness risks, and hence not meet the FSI’s threshold for conducting a fairness assessment. This is especially likely to be true of systems that are not in production, or for which robust risk management processes that encompass fairness are already in place. FSIs should determine which systems are in scope based on the identified level of risk (discussed below) associated with each unique AIDA system.

Assessing the risk level of the system

Once AIDA systems have been identified and inventoried, FSIs can perform an assessment of the risks associated with a system based on a set of pre-defined criteria. The FEAT Principles acknowledge that their application should be scaled according to the risk of a particular AIDA system, stating that “firms can calibrate actions and requirements under their internal governance framework based on the materiality of the AIDA-driven decisions” [57]. The FEAT Principles present the following non-exhaustive and indicative list of example considerations that may be used in assessing the risk associated with AIDA systems (the commentary is not part of the original document).

The Personal Data Protection Commission (PDPC) has adopted a similar approach in their Model Artificial Intelligence Governance Framework, which elaborates that: “The extent to which organisations’ adopt the recommendations in this Model Framework depends on several factors, including the nature and complexity of the AI used by organisations’, the extent to which AI is employed in the organisations’ decision-making, and the severity and probability of the impact of the autonomous decision on individuals” [67]. Accordingly, FSIs following PDPC guidance are not acting in contradiction to any guidance in this report.

The result of the risk assessment may be to place the AIDA system into a specific risk tier or level such as “low”, “moderate”, or “high”. The goal of risk tiering is to ensure that models with similar levels of risk are treated similarly and to allocate more assessment resources to higher risk systems through an appropriate customisation of the Methodology (as described in Section 3.3) rather than to directly quantify risk.



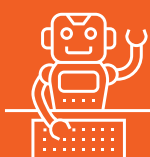
Extent to which AIDA is used in decision-making

more complete substitution of technology for human decision-making may increase the risk of accidental systematic disadvantage, especially if the AIDA models are more complex and opaque.



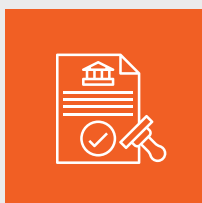
Complexity of AIDA model

more complex models may obscure, for example, the ways in which personal data is used which might increase the risk that the use of particular attributes is not fully justified.



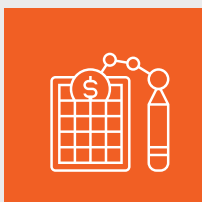
Extent of automation of process of AIDA-driven decision-making

all else being equal, more automation is likely to mean a higher volume and/or speed of AIDA-driven decision-making which may increase the scope of impact and decrease opportunities for timely human intervention to prevent issues.



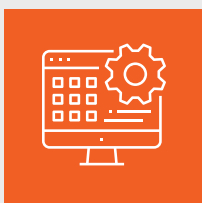
Severity and probability of impact on different stakeholders, including individuals

for example, the use of AIDA systems to make credit approval or pricing decisions is more consequential than adjusting marketing messages. Impact could also include the volume and scope of usage, and the type and number of external and internal stakeholders that can be affected.



Monetary and financial impact

AIDA models may be higher risk because of the potential impact to the FSI, such as its exposure and balance sheet, or to customers.



Regulatory impact

some AIDA models may be higher risk because of the regulatory exposure of their decision-making for applications such as anti-money laundering.



Options for recourse available

all else being equal, if there are efficient and effective ways for customers or other stakeholders to challenge or change AIDA system's decisions then the associated risks (such as systematic disadvantage) may be lower.

Applying the Methodology

Once the risk management processes in place within the FSI have determined that an AIDA system is in scope, then the system can be assessed using the Methodology (introduced in Section 3). The following guide is an indicative example of how this assessment might proceed:

1. AIDA System Owner completes a risk assessment to determine an appropriate customisation of the Methodology for the system's risk (see Section 3.3). 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 what 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, perform the relevant analysis, and produces 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 might 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 the internal requirements and request conclusive documentation.



Figure 1: A potential workflow for applying the FEAT Fairness Assessment Methodology.

2.2 Existing AIDA risk management practices

FSIs leverage various elements of existing risk management frameworks to govern the risks associated with AIDA systems. A common standard for applying these elements has not yet been established across the industry. FSIs are in different maturity stages in their approaches to adopting and specifying existing risk management processes to AIDA systems based on their specific risk management frameworks, the varying scale of AIDA deployment, and the availability of relevant in-house expertise.

Yet, FSIs acknowledge the need to enhance existing frameworks in order to address the increased complexity and specific risks associated with AIDA systems such as Fairness risks. According to the EY/IIF 2019 Global Risk Survey, firms plan to enhance their Model Risk Management (MRM) processes for AI/ML over the next three years across a range of areas, including model risk assessment, ongoing monitoring, change management, policies and procedures, and model inventory framework [31].

The following subsections highlight market observations across the elements of risk management for AIDA systems.

Initiation stage – identification

FSIs commonly establish a consistent definition to facilitate a **cohesive identification** of applications that are in scope of their AIDA risk management framework. In order to ensure a comprehensive view of relevant systems across the FSI and their consideration within risk management frameworks, AIDA systems are typically captured in existing or newly established (model) inventory management processes. Some FSIs amend the minimum documentation standards that apply to the capturing of AIDA systems in such inventories to include additional elements like the frequency of retraining or the scope and type of data input.

Initiation stage – risk assessment and risk tiering

Once identified and inventoried, FSIs typically perform an assessment of the risks associated with the identified systems based on a set of pre-defined criteria. These criteria cover three main dimensions: complexity, impact, and materiality.

- **Complexity** could refer to the mathematical computation, the number and type of data sources used as inputs, or the level of human intervention opposed to autonomous decision-making.
- **Materiality** could encompass the social, regulatory, reputational, economic or end-user consequences (i.e. exposure, balance sheet, reputation etc.) associated with a potential wrongdoing or error of the AIDA system, the possibilities or options for recourse as well as the number of users that apply the AIDA system.
- **Impact** could refer to the consequentiality of the decision: the use of AIDA systems to make approval or pricing decisions is more consequential than adjusting marketing messages. The impact dimension could also include the volume and scope of usage, the type and number of external and internal stakeholders that can be affected as well as the probability and severity of any potential harms that can be caused.

As a result of the risk assessment, AIDA systems are categorised into risk tiers that are associated with specific requirements across the model lifecycle and risk management operating models. These requirements could include:

- The level of required documentation regarding utilised data sources, the development methodology, the requirements for calibration, the test plan, the user's manual, or the technology environment
- The required approval level for use of the model, deployment of the model or model changes
- The monitoring frequency and depth
- The defined escalation procedures
- The controls framework including the extent of controls that apply in various lifecycle stages (i.e. development tests)
- The validation activities including frequency (i.e. at development, annually, bi-annually) or breadth and depth (i.e. review of documentation, developmental data, performance testing, backtesting, benchmarking, sensitivity testing, stability testing etc.)

Development stage

Amongst many factors, the quality of AIDA systems depends strongly on their underlying data. Since AIDA systems operate based on relationships learned from data, that data plays a fundamental role in determining their output. As a result, data acquisition and processing take on greater importance for AIDA systems. In consequence, FSIs are enhancing their control frameworks and considerations of relevant constraints, like personal data usage or transfer bounded by personal data privacy rules.

FSIs are also emphasising the explainability of AIDA systems to ensure they operate in a manner consistent with their business purpose. This includes justifications and explanations of how the business problem statement is addressed with a designated system, and the reasoning behind the choices made in the design of the system, including tradeoffs that have been considered such as those between predictive accuracy and the interpretability of the system's outputs.

Review stage

Model review remains a key component of the model lifecycle and is reflected in established governance frameworks for initial, ad hoc, and regular review cycles. For complex systems that are associated with an increased level of risk and thus a higher risk tiering, FSIs have established sophisticated model validation practices. The model's conceptual soundness (its fitness for purpose) plays a key role in development and review, especially due to the increased level of complexity of non-linear relationships that can be captured by modern algorithms.

As traceability between model inputs and outputs decreases with the complexity of the AIDA system, explainability becomes increasingly important for effective review. FSIs address these challenges with various enhancements in the review practice. These can include the establishment of centralised model validation teams or the use of technological support to standardise validation tasks and concentrate human efforts on judgement and interpretation.

Ongoing monitoring

Ongoing monitoring aims to detect whether the AIDA system is performing as intended and evaluate any relevant changes. An annual review is typically applied as a minimum standard to determine whether systems continue to be fit for purpose given the availability of new data or potential changes to the business, economic, or regulatory environment. However, FSIs may consider this frequency to be potentially insufficient for complex AIDA systems given their dynamic nature. Whilst the specific frequency of ongoing monitoring is regularly linked to the risk assessment and tiering of the AIDA system it is also differentiated by model type. For example, a supervised learning model with infrequent batch updates to training data might require testing semi-annually, whereas a reinforcement learning model receiving continuous feedback might require testing on a quarterly or even monthly basis. FSIs may also consider a dynamic monitoring frequency to AIDA systems with shorter intervals at the onset of the system and a slightly extended frequency once the system performance is deemed robust. In general, FSIs emphasise model performance as a monitoring target because of the difficulty of assessing their contextual soundness by simple introspection. The scope of performance definitions is therefore often extended to encompass benchmarking, local interpretability, and ethical or societal considerations.

Active changes to an AIDA system's components such as models, inputs, outputs, or production infrastructure are also typically monitored for potential impacts to the system's operation. In many cases, technological and societal dynamics cause these changes to occur continually, requiring increased scrutiny. Monitoring rapid changes to the AIDA system's components can be particularly challenging if the model is retrained dynamically as a result, as there is little time or data to assess a particular model configuration before another retraining occurs. The system should be robust enough and not impacted by retraining or incorrect training data.

2.3 Integrating the Methodology with existing risk management

Existing risk management in FSIs follows the industry standard three lines-of-defense model: the first line (typically the business) owns and manages risks; the second line (risk and compliance) sets standards and controls, monitors whether the first line is adhering to them, supports and challenges the first line, and measures, aggregates, and reports on enterprise-wide risk; and the third line (internal audit) provides independent assurance of the first and second line with direct accountability to the senior board.

From a governance perspective, there is no uniform approach to integrating fairness into the overall risk management frameworks of FSIs. Below are three approaches being used by some FSIs that focus on different risk governance elements to manage fairness risks. FSIs might choose one or a combination of these strategies.

- Embed fairness considerations within a centralised AIDA governance function

Some FSIs have established AI centres of excellence which are responsible for managing AIDA systems across their lifecycle. These centralised functions can extend their scope of responsibility to cover fairness risks and the Methodology. The main advantage of placing the responsibility with such a centre is that fairness outcomes depend on many design choices, and a centralised team of AIDA specialists will be well-placed to understand the choices that are being made and integrate fairness mitigation strategies into their system lifecycle. For instance, this group may have the best expertise to define a set of appropriate fairness definitions, and a set of appropriate fairness metrics to measure disadvantage. The main downside of this centralised approach is the increased necessity to involve business units to ensure business needs and requirements are fully understood and considered, as these may be relevant to identify and assess fairness risks.

- Extend the technical Model Risk Management (MRM) framework to include fairness considerations

The industry has largely developed mature MRM practices. This development was accelerated through local regulatory initiatives of varying scope and detail that have had a direct or indirect impact on MRM functions of FSIs. Examples for such regulatory work are the standards and guidelines defined by the Federal Reserve [32] or by the European Central Bank [26]. Some FSIs have extended the responsibility of the second line of defense (the MRM function) beyond the technical model development, review, and monitoring processes to include fairness. The benefit of integrating the Methodology into the MRM Framework is that MRM groups are experienced in receiving reports from model developers and assessing their safety and soundness, and are staffed with experts who can perform the technical analyses suggested in the Methodology. For instance, this group would be well-placed to decide whether fairness should be an additional component to the model risk assessment/ tiering to reflect the potential ethical and societal impacts, how the potential for bias and lack of fairness should be measured, and what minimum standards apply to monitoring. The downside of this approach is that these experts may be further removed from business operations and have less knowledge of issues (such as community concerns about particular products) that are unique to fairness assessments.

- Extend the ethics and conduct framework for employees to include AIDA systems

Most FSIs have existing frameworks dedicated to ethics and employee conduct. These can be extended to include AIDA fairness, and the bodies in charge of ensuring these standards are upheld can be given the responsibility to define relevant standards and/or review fairness assessments. The benefit of this approach is that this type of body will often already have more expertise in social and ethical considerations than highly technical areas of the FSI and may be involved with other FEAT Principles, such as those relating to ethics. They will also have experience with the need to make tradeoffs between different stakeholders and FSIs commercial objectives. This type of group would be well-placed to review the selected fairness definition specific to the underlying use case of the system, as well as the associated risk. The downside of this strategy is that ethics teams may lack technical knowledge, which would in turn require the involvement of and alignment with other functions that assess the technical soundness to derive a holistic review of such systems.

No matter which overall strategy is chosen, it is important that the choices made by the first line of defense may be challenged by an independent function to safeguard compliance with the established standards and policies. Similarly, the standards as defined by the second line of defense may be challenged to ensure that risk and control frameworks for fairness are consistent with overall principles and guidelines of the FSI. This includes assessing whether the justifications given for choices are actually supported by the evidence. The scope and extent of requirements that apply to managing the risk of such systems can be based on the system's risk level as discussed in Sections 2.1 and 2.2.

Involving IT and security

Many AIDA systems or data pools are open source or developed by or with third parties, and overarching considerations around aspects such as data privacy, cyber risks or information security, equally apply to such systems. AIDA System Developers and Assessors often work closely with Information Technology (IT) teams to ensure the consistency of different tools and compliance with corporate policy. For instance, FSIs may verify key functionalities of the libraries and for public open source libraries may assess the credibility of the underlying ML libraries and benchmark those to alternative libraries.

Training

No matter where the assessment Methodology is integrated in AIDA risk management, the function with this new responsibility is likely to benefit from enhanced training to understand the unique characteristics of AIDA systems, their associated risks, and the relevance of the FEAT Principles as well as how they concur with the code of conduct and overarching purpose of the FSI. The scope of such training may include techniques associated with AIDA systems to strengthen the technical understanding across the different Lines of Business of an FSI. Such learnings may aim to include parties that have previously not been involved in system developments, but are likely to be included as AIDA System Owners in the future. Beyond this scope of training, FSIs may choose to offer dedicated learning for fairness considerations associated with AIDA systems. Such initiatives may not only create awareness for the relevant risks across the workforce, but can strengthen the skill set that is required for the specific roles of AIDA System Developers, Owners, and Assessors. Both streams may be necessary to ensure an understanding of system development standards as well as the potential ethical and societal impacts of AIDA systems. The specific training program needs to be customised for various stakeholders. Beyond a dedicated training curriculum, targeted hiring or a relocation of resources may be required to ensure adequate staffing across all functions.

03 A Methodology to Assess AIDA Systems





This section presents the FEAT Fairness Assessment Methodology to help FSIs assess an AIDA system's alignment with the FEAT Fairness Principles. The Methodology guides FSIs through collating the information and analysis required for an internal or external AIDA System Assessor to judge alignment of the system under study. The Methodology should be applicable to a broad range of AIDA use cases in the financial sector, and can serve as a starting point for assessments of specific use cases such as the two presented in Document 2 (Assessment Case Studies).

Since the very notion of fairness is context-dependent and essentially contested, there is no universal formula or algorithm to measure an AIDA system's fairness. The broadness of fairness as a concept requires that every system be assessed individually in a way that accounts for contextual information such as the particular impacts of a system, the cohort with which it engages, and the communities in which it operates.

Even with contextual information, judging whether the system is sufficiently fair can only be assessed with respect to a particular set of ethical norms, and such norms are also different for different use cases, organisations, and contexts: they contain questions of moral judgement for which there are no agreed upon or universal answers.

After historical baselines and norms are established by analysing AIDA systems in use today, it may be possible for an AIDA System Assessor to be more prescriptive about normative fairness requirements. This might include, for example, requiring certain fairness measures to be computed, or even constraining the acceptable values of these measures.

At this early stage in the development of fairer AIDA systems, however, the only practical approach for assessing the fairness of an AIDA system is to compare its measured impact to the ethical objectives and constraints with which it was designed, or to a set of broader ethical governance policies that exist within the responsible organisation. Of course, these fairness objectives and constraints themselves may also be subject to scrutiny and review.

The effectiveness of the Methodology is enhanced by AIDA systems having their fairness objectives and constraints declared in the initiation phase of the model lifecycle (see Section 2.2). For example, by completing Part A (“Describe system objectives and context”) of the Methodology before the system is developed, AIDA System Owners have provided the fairness goals against which different possible system designs can be assessed. Upfront declarations of the system’s fairness goals also prevent AIDA System Owners from retroactively declaring “fair operation” of a system to be whatever it is currently doing. Of course, as an AIDA system is developed and deployed, additional fairness concerns may arise and the system’s goals would be updated accordingly (addressed Part E of the Methodology, see Section 3.2.5).

Subsequent to applying the Methodology and receiving feedback from the AIDA System Assessor, the AIDA System owner may seek to improve the system’s alignment with the FEAT Fairness Principles. [Appendix 2](#) provides details of some potential mitigation strategies and actions to improve an AIDA system’s performance with respect to fairness objectives.

3.1

Key concepts related to FEAT Fairness Principles

The FEAT Principles are a set of 14 principles addressing AIDA fairness, ethics, accountability and transparency. Though the principles are interrelated, the scope of this document is the four fairness principles, which are grouped into “justifiability” and “accuracy and bias” (note: emphasis added):

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 minimise *unintentional bias*.

F4. AIDA-driven decisions are regularly reviewed so that *models behave* as *designed and intended*.

Each principle is followed by some elaboration and examples, though the FEAT Principles, by design, are not prescriptive. To get from principles to practical applications, however, requires understanding how to select or derive appropriate measures of fairness, interpret these measures, and to build systems that account for them. Considerations to inform this process and complete the Methodology are provided, that draw from well-established risk management approaches for statistical models implemented in FSIs, as well as the newer domain of algorithmic fairness. The questions and considerations presented in the Methodology rely on key concepts (emphasised above) from the FEAT Fairness Principles. These concepts are introduced in the following subsections.



AIDA-driven decisions and AIDA systems

For this document, an AIDA system is a system (i.e., a set of components working together) containing one or more AIDA models that is used to make decisions to achieve a set of objectives. The decisions made by an AIDA system are “AIDA-driven decisions”. AIDA systems encompass not just the AIDA model (or models), but also other parts of a decision-making process such as the humans involved in those decisions (who might be giving input or even using the model output to inform their final decision), other governance or rules that affect the decisions, the monitoring and review mechanisms, and the actions taken on the basis of the models’ output.

Focusing on AIDA systems rather than models acknowledges that fairness is not confined to being a mathematical property of a predictive model. Rather, in the context of the FEAT Fairness Principles it concerns the treatment and outcomes for people impacted by the (AIDA-driven) decisions of an FSI: the AIDA system encompasses the aspects of the FSI’s operation relevant to those impacts.

Models and model behaviour

AIDA encompasses diverse algorithmic tools including traditional statistical modelling, business rule automation, and machine learning. For this document, a “model” can be any AIDA algorithm, implemented in software, that has well-defined inputs and outputs, and for which there exists a set of performance measures (for example, accuracy). Typically, models are used to make predictions about unknown or future states, though they may also directly make choices or decisions. Examples of some different types of models that might be used in an FSI include:

- a logistic regression model trained to estimate the probability of a customer refinancing a loan in the next 6 months
- a natural language translation algorithm that translates customer queries from other languages into English
- a set of automated business rules that examine the customer database to determine when a customer will be sent a marketing email

The behaviour of a model in the context of the FEAT Fairness Principle F4 can be thought of as referring primarily to its operation of mapping inputs (data) into outputs (predictions or decisions). However, other operational considerations may be relevant, such as the reliability or maintainability of the model’s implementation or its computational cost.

Accuracy and relevance

This document defines “accuracy” as used in the FEAT Principles to mean how well a model’s predictions match the true state of the world. The most informative measures for this match will vary by application, and may not always include accuracy as defined strictly by the ratio of correctly predicted observations to the total observations. The document interprets

the intent of the FEAT Fairness Principle F3 as requiring FSIs to verify that their models are sufficiently performant to achieve their stated objectives, and to ensure that the performance of their models is regularly reviewed. Similarly, “relevant” data are taken to be data that are useful in achieving the system’s objectives. The data could, for example, encode a target for prediction, or be statistically or causally dependent on a prediction target, or could inform membership of a group for fairness analysis. These definitions imply that it may be necessary for the AIDA System Owner to provide its measures for relevance and performance. They could justify the measures by explaining how they capture risks in the system’s business and fairness objectives.

System design and intention

This document defines the “design and intention” of an AIDA system to mean its objectives and the chosen constraints on how it achieves those objectives. These objectives and constraints may or may not be the same as the objectives of any models the system contains. For example, an AIDA system that approves or denies loan applications may have multiple objectives that balance the risk and profit for the bank, comply with external governance requirements, attempt to minimise harmful defaults, and impose certain fairness constraints. Within this system there is an AIDA model whose objective is simply to predict the probability an individual will default.

An AIDA system’s objectives and constraints encompass all the things that matter about the system’s design, use, and impact. Typical business objectives might include generating revenue or maximising profit, whilst ethical objectives might include minimising disparities in outcomes between certain groups of customers. Constraints could also be business oriented (such as limits on value at risk) or ethically oriented (such as setting a maximum threshold for the rate a particular harm occurs). **Note that hereafter, “objectives” are assumed to encompass objectives and constraints.** The Methodology asks an AIDA System Owner to present its system objectives qualitatively, then, with guidance, to create measures that quantify them, and finally, to present the processes by which system objectives and their measures are reviewed (and can be modified if necessary).

Clearly setting out both the business and ethical objectives of a system allows for its operation to be assessed against those stated objectives. Doing so has a number of benefits to both the AIDA System Owner and the AIDA System Assessor:

- If these objectives and their review and modification processes are decided and stated before the system is designed, it becomes difficult for organisations to retroactively choose the objectives to fit the system’s performance (for example, by declaring that “fair operation” is whatever the system is currently doing).
- Making the objectives, especially those relating to fairness, transparent to other parties allows them to be examined, challenged, and defended, ultimately improving the system’s alignment with societal norms.

- The approach separates intent from implementation, helping uncover sources of unfair outcomes (for example, if a system is intended to satisfy a gender fairness constraint, but outcomes are unfair, the assessment of a system would likely focus on the implementation, whereas if the system objectives did not explicitly address gender, the assessment would start with addressing intent).
- As the fairness measures of the system can be tracked over time, it presents an opportunity to the AIDA System Owner to demonstrate iterative improvement of their systems against its objectives.

Systematic disadvantage

The operation of an AIDA system creates harms and benefits, directly and indirectly, for people who interact with it. Inevitably, these impacts will be distributed unevenly: some individuals or groups will be disadvantaged relative to others or relative to other possible outcomes. However, asking when these individuals or groups are being “systematically disadvantaged” is equivalent to asking when they are being treated unfairly: both are context dependent and essentially contested concepts. The Methodology therefore asks AIDA System Owners how they choose to measure when individuals or groups have been systematically disadvantaged by an AIDA-driven decision, and to defend that choice.

A distinct but related consideration for understanding AIDA fairness is systemic disadvantage. This is a widespread, persistent disadvantage that individuals or groups experience in many aspects of their lives. An example of systemic disadvantage is the well-documented history of discrimination experienced by women in many parts of the world in the labour market, where women are sometimes paid less for equal work, or not hired for certain types of jobs despite being qualified [83]. One reason for working to improve the fairness of AIDA systems is to prevent those systems from contributing to systemic disadvantage, and even to actively reduce it. On the other hand, if AIDA systems are poorly designed, they can perpetuate and reinforce existing systemic disadvantage, for instance, advertising systems trained on data from economies where women were historically disadvantaged in the labour market may learn to show women less highly paid jobs [40].



Personal attributes

Like the notion of fairness itself, the question of what constitutes a “personal attribute” is an essentially contested and context-sensitive value judgement. For example, a person’s race or ethnicity is generally considered a personal attribute but their personality type is not, even though personality is a highly individual and influential characteristic that is used in AIDA systems for use cases such as psychographic marketing. This document therefore regards personal attributes as those defined by the AIDA System Owner, but requires them to justify their selection.

FSIs may have data governance rules that define personal attributes for all systems or particular use cases. In some countries, there is anti-discrimination legislation that protects a minimum set of personal attributes and can inform FSIs’ data and model governance rules. For example, in the United States, the Equal Credit Opportunity Act prohibits discrimination on the basis of “race, color, religion, national origin, sex, marital status, age, receipt of public assistance, or good faith exercise of any rights under the Consumer Credit Protection Act” [27], and in the United Kingdom, the Equality Act protects “age; disability; gender reassignment; marriage and civil partnership; pregnancy and maternity; race; religion or belief; sex; sexual orientation.” [29]. Additionally, customer consultation can be undertaken to understand how customers view the collection or use of attributes about them in particular AIDA systems, or to determine which attributes they may not wish to reveal about themselves.

A separate, but related notion that may inform the choice of personal attributes is personal data, as defined by the Singaporean Personal Data Protection Act (PDPA) [66]:

“personal data” means data, whether true or not, about an individual who can be identified —

- from that data; or
- from that data and other information to which the organisation has or is likely to have access;

This document assumes that there are data that may be considered “personal attributes” even if they might not meet the criteria laid out in the PDPA, such as proxies for personal attributes.

Given that proxies pose similar risks as the personal attributes themselves [3], FEAT Fairness Principle F2 suggests their use should also be justified (see Section 3.2.4).

Groups of individuals

The FEAT Fairness Principles refer to systematic disadvantage for “groups of individuals”. In theory, any attribute can be used to define groups by grouping individuals with the same value of that attribute. In practice, fairness principles and more binding policies in other jurisdictions (such as anti-discrimination law) are motivated by concerns for groups as defined mainly by demographic or socio-economic attributes such as gender, race, and age, rather than groups defined by more operational attributes (such as “house owners” or “bird watching enthusiast”). Group-defining attributes are likely to (but do not necessarily) overlap with personal attributes as defined above.

It is important to note that individuals will be members of more than one group. The intersection of groups (such as gender and race) creates additional groups (such as individuals that identify as Black women or women of colour). Systematic disadvantage may be heightened due

to this intersectionality. For example, commercial AIDA facial recognition software was found to perform better on male faces than female faces, and better on lighter faces than darker faces, and especially poorly on females with dark skin [13]. In other cases, an AIDA system may show relatively little disadvantage to single-attribute groups but still show strong disadvantage for individuals at the intersection, perhaps because there are especially few training examples of that subset of the population [13, 34]. This document uses groups to mean subsets of the population defined by both single and compound attributes, and urges practitioners to keep issues of intersectionality in mind.

Justifications

The FEAT Fairness Principles require that both an AIDA system's systematically disadvantageous decisions and its use of personal attributes, should be "justified". This document adopts the common perspective of the literature on justification that a justification is the presentation of reasons for a decision or action [35]. The reasons given by AIDA System Owners in the framework for their design choices should be understandable, accessible and persuasive to the AIDA System Assessor [35], aiming to convince them that these design choices were reasonable and consistent with the FEAT Fairness Principles and the AIDA System Owner's stated ethical stance.

Some relevant considerations for justifying the inclusion of personal attributes in an AIDA model include:

- whether the attribute is required to meet a fairness objective (for example, if gender parity is a fairness objective it may be necessary to incorporate gender information into the model to meet that objective)
- whether the attribute is required to comply with legal or regulatory requirements (for example, if an FSI may only serve Singaporean citizens, then citizenship information is required to comply with the requirement)
- the privacy implications of collecting, storing and using the attribute
- the extent to which the attribute is causal factor in a prediction target
- the extent to which the inclusion of the attribute affects the system's performance, for both business and fairness objectives
- whether inclusion of the attribute is consistent with the FSI's ethical and governance principles
- the reliability of the attribute as it is measured

Part D of the Methodology (Section 3.2.4) provides some additional considerations for FSIs developing justifications for the use of personal attributes.

As part of justifying AIDA decision-making that systematically disadvantages individuals or groups, the Methodology asks AIDA System Owners to examine the set of possible tradeoffs that exist in their system between the measured disadvantages and other objectives (such as business

objectives). Typically, enforcing one type of fair decision-making will create unfairness elsewhere and may also reduce other performance measures [60]. Justifying observed outcomes of the system therefore requires explaining why the system was designed to produce those outcomes rather than the possible alternatives.

Unintentional bias

There are multiple notions of “bias” in the statistical and algorithmic fairness literature [12], some of which may be artefacts of measurement or modelling [59], while others may reflect current or historic discrimination present in society [39]. Each of these types of bias may or may not cause systems to systematically disadvantage individuals or groups (depending on how this disadvantage is defined). Reducing some types of bias may also create other types [74]. Therefore, this document interprets the FEAT requirement to minimise unintentional bias as a requirement to take reasonable and justified steps to discover and remove imperfections in models and data that may lead to avoidable systematically disadvantageous decision-making.

Consider, for example, data that systematically under-reports the income of women (a systematic error or “measurement bias”). AIDA credit scoring systems that rely on that data may underestimate female credit scores as a result. This could lead to the system systematically disadvantaging women by unduly denying them loans at a greater rate.



3.2 The FEAT Fairness Assessment Methodology

The FEAT Principles describe the properties of an AIDA system aiming to be fair, ethical, accountable and transparent. However, understanding an AIDA system's alignment with the FEAT Principles is an ongoing process. For example, FEAT Fairness Principle F1 states: "Individuals or groups of individuals are not systematically disadvantaged through AIDA-driven decisions, unless these decisions can be justified." Demonstrating alignment with F1 would require institutions to:

- define the individuals and groups in scope for analysis of systemic disadvantage
- determine the harms and benefits created by their system that may cause systematic disadvantage (i.e. define "systematic disadvantage")
- define how the distributions of these harms and benefits should be measured across individuals and groups to assess systematic disadvantage
- estimate the possible tradeoffs between these systematic disadvantages and the system's other objectives and constraints
- justify any measured systematic disadvantages created by the design choices they have made including the chosen tradeoff between different system objectives.

This section presents a general methodology for assessing alignment with the FEAT Fairness Principles that forms the basis of the case studies on customer marketing and credit scoring (found in Document 2). As well as common terminology defined above, the Methodology includes an ordered list of questions in five parts. The questions are intended to be answered by, and from the perspective of, the AIDA System Owner. Some questions likely require consulting the AIDA System Developer. Once the answers and analyses from the questions have been collated, this information goes to the AIDA System Assessor who uses it to judge the system's alignment with the FEAT Fairness Principles.

The following subsections introduce the five parts of the Methodology, the questions in each part, and considerations to assist in developing responses to the assessment questions. A technical reference has been developed in [Appendix 1](#) for further detail and discussion of the considerations. The two case studies presented in Document 2 follow this five-part structure, and also include additional use case specific considerations. As discussed in Section 2, it is expected that the Methodology will be adapted to each use case, as part of the risk-based and context-sensitive approach that this document applies.

Part A: describe system objectives and context

The purpose of this first part of the Methodology is to make explicit the objectives and constraints of the AIDA system to be assessed and to better understand the level of risk posed by the system. These objectives include both those relating to fairness, as well as the business and other purposes of the system. Material developed through application of Part A allows AIDA System Assessors to gain a qualitative understanding of the system before analysing its operation in more detail in subsequent parts. Answers to Part A may also assist FSIs in

performing initial risk assessments of an AIDA system, in order to determine how best to assess the system given its level of risk (See Section 2 for details of the risk-based approach the Methodology applies and Section 3.3 for guidance on customising the Methodology for risk).

Explicitly stating the business objectives of the system is a vital step in analysing its fairness. These objectives serve as the considerations against which design decisions and the system's operation can be justified in subsequent parts of the Methodology, particularly tradeoff-type decisions such as the degree to which systematic disadvantage has been mitigated, or the choice to include or exclude particular person attributes.

A key part of understanding the fairness risks associated with an AIDA system is understanding who is at-risk from the system's operation, and what the potential impact of the system will be on them. These properties also motivate mathematical measures of fairness required in later parts of the Methodology.

Based on the at-risk individuals and groups, and the potential harms and benefits created by the system's operation, FSIs are presumed to have determined objectives for the system's fairness. Ideally, these objectives were determined and declared before the AIDA system was designed and implemented, serving as an accountability mechanism to ensure that fair operation is not defined in retrospect to be in accordance with the current operation of the system.

Once a system is in operation, new considerations and unexpected consequences of its operation may come to light, and the system objectives would typically be updated through a well-defined process to reflect this new information. That process is examined in the system monitoring and review step in Part E. Note also that data used to determine individuals and groups for fairness assessment (in relation to Question A2) is considered and discussed in detail in the justifying the use of personal attributes step in Part D.

Questions for Part A

A1

What are the business objectives of the system and how is AIDA used to achieve these objectives?

A2

Who are the individuals and groups that are considered to be at-risk of being systematically disadvantaged by the system?

A3

What are potential harms and benefits created by the system's operation that are relevant to the risk of systematically disadvantaging the individuals and groups in A2?

A4

What are the fairness objectives of the system, with respect to the individuals and groups in A2 and the harms and benefits in A3?

Considerations for Part A

These reflections may help practitioners develop their responses to the questions in this section:

System objectives and use of AIDA

- The business objectives should encompass not just the primary objective of the system, but to the extent practical, everything that matters about its operation. This would include other objectives and also internal and external (regulatory and other) constraints. Often systems have multiple objectives and constraints, and understanding fairness tradeoffs made by the AIDA System Owner when deploying the system requires assessors to have as complete a view of these objectives as possible.
- The structure and design of the AIDA system, even at a high level, can provide important clues to potential causes of unfair operation. Examples include human design choices encoding discriminatory attitudes into the AIDA system [7], or the deployment of an AIDA system outside the context for which it was designed creating unintended harms [64].
- Does the organisation have in place principles, values or guidelines relevant to the fairness objectives of the system? These may be especially relevant if the system was deployed previously without specific fairness objectives.

At-risk individuals and groups

- Are there people that are already disadvantaged, whose interaction with the system could perpetuate or reinforce that disadvantage?
- Are some individuals more vulnerable than others to the harms caused by the system?
- Which of the at-risk groups, defined across other organisations or in anti-discrimination law and regulation in other countries, may have fairness risks in this case? (See Section 3.1.6 and 3.1.7)
- Are there combinations of group memberships that may compound disadvantage? For example, intersectional groups, such as Black women, combine both race and gender groups that are, on their own, disadvantaged. This disadvantage can be compounded when combined.
- Does the product or service have eligibility criteria that prevent it being used for some people, and is there a risk these exclude certain individuals from opportunities provided by the system?
- To explore potential individuals and groups that may be subject to greater risk of harm, it may be helpful to examine data from the system's deployment, for example, attributes used to demarcate a decision boundary, or attributes associated with higher rates of model errors.

Harms and benefits

- What are the customer's views about the benefits or risks of the system, and are there any potential harms to vulnerable or at-risk customers that may have been missed by any consultation processes undertaken?
- Are there different types of impacts the system could have, such as financial, reputational, social or emotional?
- Does the product or service have potential harms or benefits that occur over longer time frames, such as the possibility of the customer defaulting on a loan offered by a marketing intervention?
- Are there opportunity costs associated with particular outcomes that should be taken into account when assessing harms and benefits?
- Why are these the most important harms and benefits to consider? What evidence is there that these harms and benefits occur, and what steps were taken to ensure important harms and benefits do not go unnoticed?

Part B: examine data and models for unintended bias

How the data were collected and prepared for modelling can have a significant impact on the fairness of a system: there are various types of bias that data can contain, reflecting real-world disparities or introduced through sampling, measurement or processing. If not addressed, these biases can cause systems to exhibit unfair behaviour (and in some cases, this behaviour is not revealed by standard fairness measures. See [Appendix 1.1](#) for details). The performance of models used as part of an AIDA system also has a substantial impact on its fairness. Having accurate models reduces the number of mis-predictions, and having well-calibrated models ensures that the model is not over- or under-confident. In Part B, AIDA System Owners are asked to examine their data and models, since the FEAT Fairness Principle F3 requires that FSIs regularly validate that the data and models used for data-driven decisions are accurate, relevant and without unintentional bias.

In some circumstances, an AIDA system may involve a number of models, each of which contributes to achieving the overall objectives of the system. Therefore, in addition to understanding the performance measures of individual models (Question B4), AIDA System Owners are asked to provide the quantified measures for the overall system performance (Question B5). These measures will also be used in the Methodology Parts C, D and E to understand tradeoffs between performance and fairness objectives, justifications for the inclusion of personal attributes, and the monitoring and review of the system.

Questions for Part B

This section references AIDA models which are algorithms (such as a linear regression or set of business rules) used within AIDA systems (such as a credit scoring system). AIDA systems may have, as components, one or more AIDA models.

- B1** What errors, biases or properties are present in the data used by the system that may impact the system's fairness?
- B2** How are these impacts being mitigated?
- B3** How does the system use AIDA models (with, or separately from, business rules and human judgement) to achieve its objectives?
- B4** What are the performance estimates of the AIDA models in the system and the uncertainties in those estimates?
- B5** What are the quantitative estimates of the system's performance



Considerations for Part B

These reflections may help practitioners develop their responses to the questions in this section (with more details provided in [Appendix 1.1](#)):

Data collection

- One way in which data may impact the fairness of an AIDA system is through its sampling procedure, whereby different parts of the relevant population are under- or over-represented in a dataset. Examples of mechanisms that may create unrepresentative data samples include:
 - data are drawn from the results of previous systems, which themselves are targeted/ non-representative
 - there is attrition within the sampled cohorts so that certain longer term outcomes are under-represented. For example, people refinancing or moving banks in a way that is correlated with an outcome of interest (like successfully paying off a loan)
- Systematic errors in measurement can also impact system fairness and can be difficult to detect. Such errors are particularly damaging in target data, because ML algorithms will tend to learn to reproduce the error. Measurement errors can be mitigated by obtaining another, more reliable source of data, then using that source to estimate and remove the systematic error from the original data.
- Are predictions from other models used as data for the system? These other models will themselves be biased and uncertain in different ways, which may need to be accounted for explicitly.
- Many properties are not directly measurable, but require defining measurable proxies. These proxies are imperfect, and may introduce fairness risks when they are less suitable for some people compared to others. Consider for example “low socio-economic status” as codified by “earns less than X per annum” or “lives in X neighbourhood” or “own X brand of a car” or “shops at X online portals”. For such variables (especially those used as targets for modelling), consider:
 - Could variables be differently accurate for different people, or fail to capture important effects for some people? Consider especially the individuals and groups at-risk (Question A2)
 - AIDA models that measure performance through a proxy which is less effective for some people can create unintended systematic disadvantage. For example: an AIDA system that measures the utility of an internet banking app by “time spent in app” may come to a false conclusion that the app is very useful for older people, when in fact this group may, on average, simply use an unfamiliar app more slowly. Detecting this form of proxy bias involves using different proxies to measure the same property, providing only some of these measurements the AIDA model, and using the others to verify that the underlying process being modelled is indeed being captured.

Data processing and models

- How the data are pre-processed before being used in an AIDA model can impact the fairness of the overall system. One example of this is the handling of missing data. If an at-risk group has more missing data, for example, then imputation of that data may induce the system to disadvantage that group. Similarly, different imputation methods for missing data, even with the same overall performance, may introduce different advantages or disadvantages for at-risk groups.
- Different models may have the same aggregate performance across a validation set (e.g. as measured by accuracy or mean-squared error) but have very different predictions at the individual level. This even applies to the same model initialised with different random seeds, for example.

AIDA system performance

- Could the loss functions of any machine learning models be inconsistent with the system's fairness objectives (for example, by considering performance averaged over the whole population and so de-emphasising minorities)?
- What aspects of the system's objectives do its performance measures fail to capture, and could this discrepancy hide fairness risks?

Part C: measure disadvantage

The focus of this section is the extent to which the AIDA system's data-driven decisions systematically disadvantage individuals or groups. AIDA System Owners are asked to examine the potential harms and benefits to customers that they have identified as relevant to disadvantage. The purpose of the examination is to evaluate how these harms are distributed, and to justify any disparities, with reference to the tradeoffs in the system objectives that reducing the disparities would require.

The FEAT Fairness Principle F1 requires that "Individuals or groups of individuals are not systematically disadvantaged through AIDA-driven decisions, unless these decisions can be justified". The existence of disadvantage for an individual or group implies a different distribution of harms and benefits (i.e. more harm and/or less benefit than a point of comparison). Therefore an analysis of disadvantage first requires AIDA System Owners to describe the harms and benefits they are considering (Question A3) and how they and the disparities between them are being measured (Question C1). Because there are no objective or uncontested answers to these questions, AIDA System Owners are asked to consider how the system could have operated differently with regard to systematic disadvantage (Question C2) and then to provide justifications for why the system's current operating conditions are preferable to the alternatives (Question C3).

Questions for Part C

C1

What are 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 A2 and the potential harms and benefits in A3?

C2

What are the achievable tradeoffs between the system's fairness objectives and its other objectives?

C3

Why are the fairness outcomes observed in the system preferable to these alternative tradeoffs?

Considerations for Part C

Attempting to quantify the harms and benefits to the customers should be part of an FSI's long term approach to aligning with the FEAT Fairness Principles. This is the most complete way to measure the degree to which individuals or groups are being systematically disadvantaged by the operation of the system. However, quantifying the precise size or nature of a harm or benefit is difficult even for a simple system. Standard fairness measures in the literature instead typically quantify rates of incidence of particular outcomes, implicitly associating them with a harm or benefit of fixed magnitude (see [Appendix 1.2](#)).

These reflections may help practitioners develop their responses to the questions in this section (with more details provided in [Appendix 1.2](#)):

Measuring fairness objectives

- Which aspects of the system's harms and benefits are captured by the chosen fairness measures, and which are not? Though it may be possible to quantify, for example, money lost due to defaulting on a loan, such a measure will fail to capture qualitative notions that may be important such as associated changes in physical and emotional wellbeing, access to credit, and opportunities that arise from credit access. These impacts may be immediate, or may occur (and compound) over longer time frames.
- As a precursor to quantifying harms and benefits, it may be simpler to instead quantify the circumstances under which harms and benefits occur. These circumstances can be measured by a particular combination of intervention and outcome values for the individuals in the query cohort. Measuring incidence rates of benefits and harms is reasonable as long as the magnitude of a harm or benefit in a given situation is similar between customers across the cohort. It is then possible to define fairness measures by comparing these rates at which benefit- or harm-causing circumstances occur in the system's operation within cohorts for individual fairness, and across cohorts for group fairness.
- Many fairness measures have been developed in an effort to address the problem of mathematically specifying a fairness objective for an AIDA system (See [Appendix 1.2](#)). Computing and analysing a selection of standard fairness measures can mitigate the risk that a harm or benefit was missed in an initial analysis.

- Standard fairness measures in the literature such as equality of opportunity and precision parity are often implicitly or explicitly measuring absolute counts or rates of occurrences of an outcome associated with a harm or benefit [78]. Understanding which of the standard measures may be relevant to a given system therefore begins with an analysis of that system's harms and benefits, and the circumstances under which they occur.
- How might the same outcomes caused by the system affect different people differently? FSI's should always be mindful of the constant magnitude assumption underlying incidence rate approaches to quantifying fairness, and seek wherever possible to validate or extend their analysis to include estimates of magnitudes of benefits and harms, particularly for people already disadvantaged or vulnerable.

Justifying systematic disadvantage

- As AIDA System Owners are asked to provide justifications for instances of systematic disadvantage, note that:
 - fairness objectives will be different for different systems, and different people may select different objectives and tradeoffs even for the same system
 - reducing one type of systematic disadvantage may increase another type (for example, reducing a particular notion of age discrimination may increase a measure of gender discrimination)
 - inescapable tradeoffs may also exist between measures of systematic disadvantage and other performance considerations.
- What operating parameters and design choices expose the key tradeoffs between fairness and other objectives (for example, per-group selection thresholds in a logistic classifier)?
- What operating parameters and design choices were not analysed, but could affect the tradeoffs between fairness and other objectives?
- There will exist a Pareto-optimal set of tradeoffs for any particular system: systems that operate outside this optimal set can improve one or more objective measures without hurting any others [60].
- Tradeoffs can often be improved (in the sense of having to make fewer compromises between objectives) with more accurate predictions, so any approaches to improving predictive accuracy will be beneficial (such as collecting more data).
- Justifying the choice of operating position on a Pareto front is a question of priorities, and the relative importance of different objectives to the AIDA System Owner.

Part D: justify the use of personal attributes

The use of personal attributes presents privacy and fairness risks to individuals that have disclosed these data to an FSI. Some individuals may be concerned that this personal information will be used in a way that unfairly disadvantages them or adversely affects them in other ways. This is especially true if that personal information is used as features in AIDA modelling. However, excluding personal attributes from machine learning models (so-called “fairness through unawareness”) is generally unsuitable as a method for preventing discriminatory treatment or outcomes. In some cases, removing a personal attribute from a model may actually increase harm to the cohort it describes (see [Appendix 1.3](#)).

Part D of the Methodology derives from the FEAT Fairness Principle F2, which requires that the “use of personal attributes as input factors for AIDA-driven decisions is justified”. Because whether an attribute is considered “personal” varies by context, Part D asks AIDA System Owners to list the attributes they consider personal (Question D1), to explain how their choices consider the objectives and constraints discussed in Part A (Question D2).

As many use cases may not have any specific attributes that are prohibited from use (for example, based on Singaporean legislation or internal policy), whether the inclusion of an attribute is justified can alternatively be assessed with reference to its effect on the system. For this reason, Part D asks AIDA System Owners to provide justifications based on the effect of including an attribute on the system, as measured through the quantified performance and fairness measures (Question D3). Finally, AIDA models, especially ones that use non-traditional data sources and machine learning, can learn close statistical proxies for personal attributes. Question D3 also asks AIDA System Owners to examine proxies for personal attributes, preventing the requirement to justify the use of personal attributes being circumvented by substituting personal attributes with these proxies.

Questions for Part D

D1

What personal attributes are used as part of the operation or assessment of the system?

D2

How did the process of identifying personal attributes take into account ethical objectives of the system, and the people identified as being at risk of disadvantage?

D3

For 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?

Considerations for Part D

Introductory guidance about personal attributes and justifications is found in Section 3.1. Information about attributes that define groups for fairness analysis can be found in Section 3.1 (“Groups of individuals”) and Part C above. Additionally, these reflections may help in developing your responses (with more details provided in [Appendix 1.3](#)):

- The process for determining personal attributes benefits from active consultation with individuals affected by the system and the broader community. Determining personal attributes may consider, in addition to customer and community views, anti-discrimination law (if applicable), the ethical objectives and constraints of the system, the people identified as being at risk of disadvantage, and local context and history. See Section 3.1.6 for more details.
- For non-personal attributes, are there statistical dependencies with personal attributes that may cause them to act as proxies for personal attributes? In this case, some AIDA models are able to use a combination of non-personal attributes to infer a personal attribute (including using them to create systematic disadvantage). This can occur even if, individually, each proxy attribute has only weak dependence on the personal attribute. Proxy attributes can be identified using correlation or statistical dependence tests, or by training a model that uses the system’s data to predict the personal attribute, then examining the feature importance of that model.



Justifying inclusion of personal attributes

- In justifying the inclusion of personal attributes, it is important to assess the associated tradeoffs. For example, a justification may compare the degree to which including or excluding a personal attribute contributes to unequal or unfair outcomes against the attribute's contribution to the legitimate business purposes of a credit process (such as maximising profit, limiting the risk of default, or limiting loss exposure in the event of a default). Feature importance techniques may be useful in understanding the impact of inclusion of personal attributes on model performance and system objectives (see [Appendix 1.3](#)).
- The use of some personal attributes may be justified as being required for the basic operation of the system. For example, for products that are available only to local citizens or individuals under or above a certain age, personal attributes such as age, citizenship or residency status are essential to deliver those products to the appropriate individuals and groups.
- One justification for inclusion of a personal attribute is that the system needs that attribute to measure or achieve an objective. For a fairness objective to be measured or acted upon, the personal attributes that objective relies on must be used: for example, it is difficult to estimate and control the fairness of a system's operation with respect to gender if the gender of the cohort interacting with the system is not known (see also "fairness through unawareness" in [Appendix 1.3](#)).
- If some of the enumerated attributes are included and excluded, are these choices justified with specific facts about the local context and history? For example, the composition of ethnicities in Singapore is different from that in other countries, with a unique historical context, and therefore may engender different policies around governance of public and private services.
- Is the personal attribute serving as a proxy for a more relevant non-personal attribute that is not being collected? In this case, justifying the use of this personal attribute may require explaining why it is impractical to access the more relevant data.

Part E: examine system monitoring and review

Even if an AIDA system is performing well when it is deployed, its performance may degrade over time or unintended consequences of its operation may arise that were not apparent when the system was designed. This type of degradation can be detected through monitoring and review of both the technical characteristics of the AIDA model, as well as the decision rules and human interventions that are also part of the larger AIDA system. Such monitoring and review is especially important for systems involved in high-volume or high-consequence decisions.

Monitoring and review are key components of the FEAT Fairness Principles. Principle F4 requires that AIDA-driven decisions are regularly reviewed so that models behave as designed

and intended. Principle F3 also requires data and models used for AIDA-driven decisions to be regularly reviewed and validated for accuracy and relevance, and to minimise unintentional bias. Finally, questions on monitoring and review are also supported by Principle F1 — if the performance of an AIDA system degrades over time, then appropriate monitoring and review would be required to detect (and then prevent or mitigate) systematic disadvantage that occurs as a result.

To address the monitoring and review requirements of the FEAT Fairness Principles, Part E of the Methodology examines both technical failures and decision rules and human interventions that affect AIDA systems. Because the particular mechanisms for review and monitoring will depend on the context within the AIDA System Owner, its priorities and its risk appetite, the AIDA System Assessor can only judge whether the mechanisms described are appropriate by referencing the AIDA System Owner's justification for their approach. The investigation and mitigation of these risks should be commensurate with the risk level of the system, but nonetheless should be explained and justified so that the AIDA System Assessor can evaluate whether the monitoring and review regime is appropriate.

Questions for Part E

E1

How is the system's monitoring and review regime designed to detect abnormal operation and unintended harms to individuals or groups?

E2

How does the system's monitoring and review regime ensure that the system's impacts are aligned with its fairness and other objectives (A1 and A4)?

E3

What are the mechanisms for mitigating unintended harms to individuals or groups arising from the system's operation?

Considerations for Part E

These reflections may help in developing your responses (with more detail provided in [Appendix 1.4](#)):

- It is common for all performance objectives including fairness measures to be continually updated and monitored as outcomes from the system's decisions are observed (such as when a customer responds), with automatic triggers to alert people responsible when a measure decreases significantly.
- Some outcome measurements are impossible to obtain in the scope of a system's operation: for example, a marketing campaign may last three months but is marketing loans that have much longer terms. In this case, assessing harms or benefits relating to successful repayments or defaults would require predictive models. See [23, 14, 82] for details of incorporating the uncertainty from such a model into fairness measure calculations.
- Monitoring and review remain important, but tend to be de-prioritised as the system ages. Are there mechanisms in place to ensure monitoring and review through the lifetime of the system, particularly with regard to funding?

- To review the systems objectives as new information or new unintended consequences of the systems operation arise, it may be important to involve diverse stakeholders, including customers, management, domain experts, and data scientists. This is particularly important for addressing or anticipating concept shift (for example, are there concepts that may be changing in meaning?)
- “Dataset shift” occurs when query data for an AIDA model drawn from a different statistical distribution than the data used to train the model. Dataset shift can degrade model performance and increase systematic disadvantage, and can be difficult to detect. This creates a risk that an AIDA system is performing worse with respect to its fairness or business objectives than standard performance measures indicate. For details on detecting and addressing dataset shift, see [Appendix 1.4](#).
- Because performance measures defined for AIDA systems may be created with reference to a control group, it is necessary to also ensure the continuation of the experimental control so as not to introduce problem shift into the system’s performance measures.
- How can particular decisions made by the system be reproduced or examined after the fact? Reproducible AIDA systems can be difficult to implement, requiring careful versioning of data, models, associated code and infrastructure, and even random seeds. However, without reproducibility it is difficult to understand the mechanisms causing an undesirable outcome during a review.
- How, by whom, and how often are the system’s overall design, performance and impact reviewed?
- How are customers or other affected parties able to contribute to reviews of the system’s performance, challenge a decision that affects them, or seek redress for an error? These avenues may be important sources of information about unintended consequences of the AIDA system’s operation that can be used to improve the fairness of its operation.



3.3

Guidance on scaling the Methodology to system risk

Based on a risk assessment, FSIs are encouraged to customise the extensiveness of the FEAT Fairness Assessment Methodology to be commensurate with the level of risk of the AIDA system under study as well as other factors, such as the system's objectives. See Section 2 for a discussion of AIDA risk assessments and risk management in the context of the FEAT Fairness Principles.

The customisation of the Methodology may include changing both the assessment process itself, and its associated governance requirements. FSIs should determine how this customisation occurs to best fit their existing practices. For example, to scale the assessment to lower risk AIDA systems, FSIs could:

- Use Part A of the Methodology for triage: answers to Part A of the Methodology contain an overview of the business and fairness objectives of the system, its use of AIDA, the potential harms and benefits the system may cause and the individuals and groups at-risk from systematic disadvantage. As such, it forms a summary of the system and an indication of its fairness risks, which could be completed and used to determine whether further analysis of the system (and completion of the subsequent parts of the Methodology) is warranted.
- Conduct a more limited tradeoff analysis: answering Question C2 by computing the Pareto-optimal tradeoffs between a system's different objectives as a function of parameters like decision thresholds may be a time-consuming endeavour for many systems. For lower risk systems, it may be possible to justify the current operating point with only qualitative references to the alternative tradeoffs.
- Complete the assessment questions at a lower level of detail: for lower risk systems, it may be sufficient to provide short summary answers to some or all of the assessment questions. Document 2 Section 2.4 is an example of a short assessment conducted on a lower risk AIDA marketing system.
- Complete a subset of the assessment questions: lower risk systems might only require a subset of the Methodology's assessment questions to be answered. For example, some lower risk systems may not have extensive monitoring and review regimes, enabling FSIs to omit some of the questions in Part E. Similarly, for lower risk systems not using complex predictive models, it may be appropriate to omit Questions B3 and B4.

Some examples of governance procedures for the FEAT Fairness Assessment Methodology that could be scaled to account for a system's risk include:

- The level of required documentation as evidence that the assessment was performed: higher risk systems might need more extensive documentation of, for example, the justification for the choice of metrics, to support additional review activities. Systems that include more human intervention might need different documentation, such as gathering training materials.

- The review and approval levels that apply: lower risk systems might be reviewed and approved by the model risk management team while higher risk systems might require approval from senior management. Review and approvals may also depend on how AIDA fairness is integrated with existing model governance. See Section 2.3 below for a discussion of some options.
- The escalation procedures that apply in case of deviations from pre-defined standards: ongoing monitoring is a key part of the Methodology, and higher risk systems might require more frequent monitoring, and more rapid or different interventions if standards (such as fairness metric thresholds) are breached.
- The frequency of assessment: systems that are mostly static and lower risk might be subject to a one-time fairness assessment before deployment with ongoing monitoring of some key metrics, while higher risk systems might be subject to fuller reviews even after they have been in production for some time.

The case studies in Document 2 illustrate applications of the Methodology for AIDA systems with different levels of risk. The customer marketing case study in Document 2 Section 2.3 is an example of a detailed assessment suitable for a higher risk AIDA system, while the shorter case study in Document 2 Section 2.4 is intended to illustrate application to a lower risk system. For the credit scoring case study, Document 2 Section 3.3 illustrates a higher risk application of the Methodology. Answers to questions in Part A of the Methodology in these case studies also serve as examples of possible “triage” assessments suggested above.



Acknowledgements





We acknowledge the following people for their contribution on this project.

Role	Name	Organization
Project Sponsor	Sopnendu Mohanty	Monetary Authority of Singapore
Project Director	Li Xuchun	Monetary Authority of Singapore
Project Lead	Zhang Qiang	Monetary Authority of Singapore
Project Manager	Hardeep Arora	Element AI
Project Manager	Heok Kee Oon	EY
Project Manager	Oliver Cheung	IAG Firemark Labs
Project Manager	Dave Vijay	HSBC
Project Manager	Amos Ong	HSBC
Project Manager	Ng Kok Keong	UOB
Subject Matter Advisor	Prof. Robert Williamson	Fellow of the Australian Academy of Science
Subject Matter Advisor	Prof. Richard Zemel	University of Toronto
Lead Subject Matter Expert	Marc Etienne Brunet	Element AI
Lead Subject Matter Expert	Grace Abuhamad	Element AI
Lead Subject Matter Expert	Robin Schlarb	EY
Lead Subject Matter Expert	Lachlan McCalman	Gradient Institute
Lead Subject Matter Expert	Daniel Steinberg	Gradient Institute

Role	Name	Organization
Subject Matter Expert	Hardeep Arora	Element AI
Subject Matter Expert	Richard Zuroff	Element AI
Subject Matter Expert	Marie-Claude Ferland	Element AI
Subject Matter Expert	Heok Kee Oon	EY
Subject Matter Expert	Vikas Deep Sharma	EY
Subject Matter Expert	Neena Antal	EY
Subject Matter Expert	Jason Tuo	EY
Subject Matter Expert	Tiberio Caetano	Gradient Institute
Subject Matter Expert	Gunjan Bhatt	HSBC
Subject Matter Expert	Dexter Ang	HSBC
Subject Matter Expert	Marjorie Chan Bona	HSBC
Subject Matter Expert	Huah Cheng Jiann	UOB
Data Scientist	Summit Bhalla	Element AI
Data Scientist	Eniola Alese	Element AI
Data Scientist	Angus Leigh	Element AI
Data Scientist	Andrey Kostenko	IAG Firemark Labs
Data Scientist	Oxana Samko	HSBC
Data Scientist	Jun Xu	HSBC
Data Scientist	Robert Cheah Tong Ngee	UOB

We would also like to express our appreciation to Amy Shi-Nash (HSBC), Bill Simpson-Young (Gradient Institute), Chris Lim (EY), Johnson Poh (UOB), Nicolas Chapados (Element AI) and Richard Lowe (UOB) who have supported and contributed to this project.



Bibliography



01 - 10

- [1] Angwin, J., Larson, J., Mattu, S., & Kirchner, L. (2016). Machine Bias — There's software used across the country to predict future criminals. and it's biased against blacks. Retrieved from <https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing>
- [2] Bagalini, A. (2020). 5 ways racism is bad for business – and what we can do about it. Retrieved from <https://www.weforum.org/agenda/2020/07/racism-bad-for-business-equality-diversity/>
- [3] Barocas, S., & Selbst, A. D. (2016). Big Data's Disparate Impact. *California Law Review*, 104, 671–732. <https://doi.org/10.15779/Z38BG31>
- [4] Barocas, S., Hardt, M., & Narayanan, A. (2019). Fairness and machine learning. Retrieved from <https://fairmlbook.org/>
- [5] Bellamy, R. K. E., Dey, K., Hind, M., Hoffman, S. C., Houde, S., Kannan, K., ... Zhang, Y. (2018). AI Fairness 360: An Extensible Toolkit for Detecting, Understanding, and Mitigating Unwanted Algorithmic Bias. *Advances in Neural Information Processing Systems*, 2017-Decem(Nips), 5681–5690. Retrieved from <http://arxiv.org/abs/1810.01943>
- [6] Berk, R., Heidari, H., Jabbari, S., Joseph, M., Kearns, M., Morgenstern, J., ... Roth, A. (2017). A Convex Framework for Fair Regression. Retrieved from <https://arxiv.org/pdf/1706.02409.pdf>
- [7] Bertrand, M., & Mullainathan, S. (2003). Are Emily and Greg More Employable than Lakisha and Jamal? A Field Experiment on Labor Market Discrimination. <https://doi.org/10.3386/w9873>
- [8] Bickel, S., Brückner, M., Brückner, B., & Scheffer, T. (2009). Discriminative Learning Under Covariate Shift. In *Journal of Machine Learning Research* (Vol. 10).
- [9] Binns, R. (2019). On the Apparent Conflict Between Individual and Group Fairness. <https://doi.org/10.1145/3351095.3372864>
- [10] Breiman, L. (2001). Random forests. In *Machine Learning* (Vol. 45). <https://doi.org/10.1023/A:1010933404324>

11 - 20

- [11] Buja, A., Stuetzle, W., & Shen, Y. (2005). Loss Functions for Binary Class Probability Estimation and Classification: Structure and Applications. Retrieved from www.wharton.upenn.edu/~shenyi
- [12] Blodgett, S. L., Barocas, S., Daumé, H., & Wallach, H. (2020). Language (Technology) is Power: A Critical Survey of “Bias” in NLP. *ArXiv*. Retrieved from <http://arxiv.org/abs/2005.14050>
- [13] Buolamwini, J., & Gebru, T. (2018). Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification. In *Proceedings of Machine Learning Research* (Vol. 81).
- [14] Caelen, O. (2017). A Bayesian interpretation of the confusion matrix. *Annals of Mathematics and Artificial Intelligence*, 81(3–4), 429–450. <https://doi.org/10.1007/s10472-017-9564-8>

- [15] Calders, T., & Zliobaite, I. (2013). Why Unbiased Computational Processes Can Lead to Discriminative Decision Procedures. In L. Magnani, A. Aliseda, G. Longo, C. Sinha, & P. Thagard (Eds.), *Studies in Applied Philosophy, Epistemology and Rational Ethics* (3rd ed., pp. 43–60). Retrieved from <http://www.springer.com/series/10087>
- [16] Chen, I., Johansson, F. D., & Sontag, D. (2018). Why Is My Classifier Discriminatory? Retrieved from <https://arxiv.org/pdf/1805.12002.pdf>
- [17] Corbett-Davies, S., Pierson, E., Feller, A., & Goel, S. (2016, October). A computer program used for bail and sentencing decisions was labeled biased against blacks. It's actually not that clear. The Washington Post. Retrieved from https://www.washingtonpost.com/news/monkey-cage/wp/2016/10/17/can-an-algorithm-be-racist-our-analysis-is-more-cautious-than-propublicas/?utm_term=.0b6e663330a4
- [18] Corbett-Davies, S., Goel, S., Chohlas-Wood, A., Chouldechova, A., Feller, A., Huq, A., ... Shroff, R. (2018). The Measure and Mismeasure of Fairness: A Critical Review of Fair Machine Learning. Retrieved from <https://arxiv.org/pdf/1808.00023.pdf>
- [19] D'Amour, A., Srinivasan, H., Atwood, J., Baljekar, P., Sculley, D., & Halpern, Y. (2020). Fairness is not static: Deeper understanding of long term fairness via simulation studies. FAT* 2020 - Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency, 525–534. <https://doi.org/10.1145/3351095.3372878>
- [20] Dankers, C., Kronseder, V., & Wagner, M. (2019). Chapter 8 Introduction to Feature Importance |. In *Limitations of Interpretable Machine Learning Methods*. Retrieved from https://compstat-lmu.github.io/iml_methods_limitations/pfi.html
- [21] Dastin, J. (2018). Amazon scraps secret AI recruiting tool that showed bias against women. Reuters. Retrieved from <https://www.reuters.com/article/us-amazon-com-jobs-automation-insight-idUSKCN1MK08G>
- [22] Datta, A., Tschantz, M. C., & Datta, A. (2015). Automated Experiments on Ad Privacy Settings. *Proceedings on Privacy Enhancing Technologies*, 2015(1), 92–112. <https://doi.org/10.1515/popets-2015-0007>
- [23] DiCiccio, T. J., & Efron, B. (1996). Bootstrap confidence intervals. *Statistical Science*, 11(3), 189–212. <https://doi.org/10.1214/ss/1032280214>
- [24] Dudík, M., Bird, S., Wallach, H., & Walker, K. (2020). Fairlearn: A toolkit for assessing and improving fairness in AI. Retrieved from <https://www.microsoft.com/en-us/research/publication/fairlearn-a-toolkit-for-assessing-and-improving-fairness-in-ai/>
- [25] Dwork, C., Hardt, M., Pitassi, T., Reingold, O., & Zemel, R. (2011). Fairness Through Awareness. Retrieved from <https://arxiv.org/pdf/1104.3913.pdf>
- [26] European Central Bank. (2019). ECB Guide for the Targeted Review of Internal Models (TRIM). <https://doi.org/10.2866/849746>
- [27] 12 CFR 1002 - EQUAL CREDIT OPPORTUNITY ACT (REGULATION B) <https://www.ftc.gov/enforcement/statutes/equal-credit-opportunity-act>
- [28] Edelman. (2020). 2020 Edelman Trust Barometer. Retrieved from <https://www.edelman.com/trustbarometer>

31 - 40

[29] Equality Act 2010. UK Public General Acts (2010) c. 15 Part 2 Chapter 1 Section 4 <https://www.legislation.gov.uk/ukpga/2010/15/section/4>

[30] Eubanks, V. (2018). Automating Inequality: how high-tech tools profile, police, and punish the poor. New York, NY: St. Martin's Press.

[31] Ernst & Young. (2019). An endurance course: surviving and thriving through 10 major risks over the next decade Tenth annual EY/IIF global bank risk management survey

[32] Board of Governors of the Federal Reserve System. (2011). SR 11-7: Guidance on Model Risk Management.

[33] Fogliato, R., G'sell, M., & Chouldechova, A. (2020). Fairness Evaluation in Presence of Biased Noisy Labels. Retrieved from PMLR website: <http://proceedings.mlr.press/v108/fogliato20a.html>

[34] Foulds, J. R., Islam, R., Keya, K. N., & Pan, S. (2020). An intersectional definition of fairness. Proceedings - International Conference on Data Engineering, 2020-April, 1918–1921. <https://doi.org/10.1109/ICDE48307.2020.00203>

[35] Forst, R. (2011). The Right to Justification: Elements of a constructivist theory of justice. Retrieved from <https://cup.columbia.edu/book/the-right-to-justification/9780231147095>

[36] Frye, C., de Mijolla, D., Begley, T., Cowton, L., Stanley, M., & Feige, I. (2020). Shapley explainability on the data manifold. Retrieved from <http://arxiv.org/abs/2006.01272>

[37] Friedler, S. A., Hamilton, E. P., College, H., & Roth, D. (2018). A comparative study of fairness-enhancing interventions in machine learning. Retrieved from <https://arxiv.org/pdf/1802.04422.pdf>

[38] Gallie, W.B. (1955). *Essentially contested concepts. Proceedings of the Aristotelian society. Vol. 56. Aristotelian Society, Wiley.*

[39] Gebru, T., Morgenstern, J., Vecchione, B., Wortman Vaughan, J., Wallach, H., Daumé III, H., & Crawford, K. (2018). Datasheets for Datasets. Retrieved from http://www.fatml.org/media/documents/datasheets_for_datasets.pdf

[40] Gibbs, S. (2015). Women less likely to be shown ads for high-paid jobs on Google, study shows. The Guardian. Retrieved from <https://www.theguardian.com/technology/2015/jul/08/women-less-likely-ads-high-paid-jobs-google-study>

41 - 50

[41] Gneiting, T., & Raftery, A. E. (2007). Strictly proper scoring rules, prediction, and estimation. Journal of the American Statistical Association, 102(477), 359–378. <https://doi.org/10.1198/016214506000001437>

[42] Gretton, A., Borgwardt, K. M., Rasch, M. J., Smola, A., Schölkopf, B., & Smola, A. (2012). A Kernel Two-Sample Test Bernhard Schölkopf. In Journal of Machine Learning Research (Vol. 13). Retrieved from www.gatsby.ucl.ac.uk/

[43] Hajian, S., & Domingo-Ferrer, J. (2013). Direct and Indirect Discrimination Prevention Methods. In B. Custers, T. Calders, B. Schermer, & T. Zarsky (Eds.), Discrimination and Privacy in the Information Society (Vol. 3, pp. 241–254). Retrieved from https://link-springer-com.libproxy.mit.edu/content/pdf/10.1007%2F978-3-642-30487-3_13.pdf

- [44] Hardt, M., Price, E., & Srebro, N. (2016). Equality of opportunity in supervised learning. *Advances in Neural Information Processing Systems*, 3323–3331. Retrieved from <https://arxiv.org/abs/1610.02413v1>
- [45] Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The Elements of Statistical Learning Data Mining, Inference, and Prediction*. Springer Series in Statistics.
- [46] Hébert-Johnson, U., Kim, M. P., Reingold, O., & Rothblum, G. N. (2018). Multicalibration: Calibration for the (Computationally-Identifiable) Masses. Retrieved from PMLR website: <http://proceedings.mlr.press/v80/hebert-johnson18a.html>
- [47] Ingold, D., & Soper, S. Amazon doesn't consider the race of its customers. should it? Bloomberg, April 2016. Retrieved from <https://www.bloomberg.com/graphics/2016-amazon-same-day/>
- [48] Jacobs, A. Z., & Wallach, H. (2019). Measurement and Fairness. Retrieved from <http://arxiv.org/abs/1912.05511>
- [49] Jan, T. (2018). Redlining was banned 50 years ago. It's still hurting minorities today. *The Washington Post*. Retrieved from <https://www.washingtonpost.com/news/wonk/wp/2018/03/28/redlining-was-banned-50-years-ago-its-still-hurting-minorities-today/>
- [50] Kamiran, F., Calders, T., & Pechenizkiy, M. (2010). Discrimination Aware Decision Tree Learning.
- [51] Kamiran, F., & Calders, T. (2012). Data preprocessing techniques for classification without discrimination. *Knowledge and Information Systems*, 33(1), 1–33. <https://doi.org/10.1007/s10115-011-0463-8>
- [52] Kamiran, F., Žliobaite, I., & Calders, T. (2013). Quantifying explainable discrimination and removing illegal discrimination in automated decision making. *Knowledge and Information Systems*, 35(3), 613–644. <https://doi.org/10.1007/s10115-012-0584-8>
- [53] Kleinberg, J., Lakkaraju, H., Leskovec, J., Ludwig, J., & Mullainathan, S. (2018). Human decisions and machine predictions. *Quarterly Journal of Economics*, 133(1), 237–293. <https://doi.org/10.1093/qje/qjx032>
- [54] Leben, D. (2020). *Normative principles for evaluating fairness in machine learning*. *AIES 2020 - Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, 86–92. <https://doi.org/10.1145/3375627.3375808>
- [55] Lei, J., G'Sell, M., Rinaldo, A., Tibshirani, R. J., & Wasserman, L. (2018). *Distribution-Free Predictive Inference for Regression*. *Journal of the American Statistical Association*, 113(523), 1094–1111. <https://doi.org/10.1080/01621459.2017.1307116>
- [56] Liu, L. T., Dean, S., Rolf, E., Simchowit, M., & Hardt, M. (2018). *Delayed Impact of Fair Machine Learning*. Retrieved from <https://arxiv.org/abs/1803.04383>
- [57] Monetary Authority of Singapore. (2018). Principles to Promote Fairness, Ethics, Accountability and Transparency (FEAT) in the Use of AI and Data Analytics in Singapore's Financial Sector. Retrieved from <https://www.mas.gov.sg/~media/MAS/News%20and%20Publications/Monographs%20and%20Information%20Papers/FEAT%20Principles%20Final.pdf>
- [58] De Maio, F. G. (2007). Income inequality measures. *Journal of Epidemiology & Community Health* 61.10: 849–852.
- [59] Mitchell, M., Wu, S., Zaldivar, A., Barnes, P., Vasserman, L., Hutchinson, B., ... Gebru, T. (2018). Model Cards for Model Reporting. *FAT* 2019 - Proceedings of the 2019 Conference on Fairness,*

Accountability, and Transparency, 220–229. <https://doi.org/10.1145/3287560.3287596>

[60] Menon, A. K., & Williamson, R. C. (2018). The Cost of Fairness in Binary Classification. In Proceedings of Machine Learning Research (Vol. 81). Retrieved from <http://proceedings.mlr.press/v81/menon18a/menon18a.pdf>

61 - 70

[61] Molnar, C. (2019). Interpretable machine learning. A Guide for Making Black Box Models Explainable. <https://christophm.github.io/interpretable-ml-book/>

[62] Noble, S. (2018). Algorithms of oppression: How search engines reinforce racism. New York University Press.

[63] Obermeyer, Z., Powers, B., Vogeli, C., & Mullainathan, S. (2019). Dissecting racial bias in an algorithm used to manage the health of populations. Retrieved from <http://science.sciencemag.org/>

[64] O'Neil, C. (2016). Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy. New York, NY: Crown.

[65] Ovadia, Y., Fertig, E., Ren, J., Nado, Z., Sculley, D., Nowozin, S., ... Snoek, J. (2019). Can You Trust Your Model's Uncertainty? Evaluating Predictive Uncertainty Under Dataset Shift. Advances in Neural Information Processing Systems, 13991–14002. Retrieved from <http://arxiv.org/abs/1906.02530>

[66] Singaporean Personal Data Protection Act. (2012). <https://sso.agc.gov.sg/Act/PDPA2012>

[67] Singapore Personal Data Protection Commission (PDPC). (2020). Model AI Governance Framework. Retrieved from <https://www.pdpc.gov.sg/-/media/Files/PDPC/PDF-Files/Resource-for-Organisation/AI/SGModelAIGovFramework2.pdf>

[68] Pedreschi, D., Ruggieri, S., & Turini, F. (2008). Discrimination-aware Data Mining. Retrieved from <http://pages.di.unipi.it/ruggieri/Papers/kdd2008.pdf>

[69] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... Duchesnay, É. (2011). Scikit-learn: Machine Learning in Python. In Journal of Machine Learning Research (Vol. 12). Retrieved from <http://scikit-learn.sourceforge.net>

[70] Perrone, V., Donini, M., Kenthapadi, K., & Archambeau, C. (2020). Fair Bayesian Optimization. ArXiv. Retrieved from <http://arxiv.org/abs/2006.05109>

71 - 80

[71] Pessach, D., & Shmueli, E. (2020). Algorithmic Fairness. AEA Papers and Proceedings, 108, 22–27. Retrieved from <http://arxiv.org/abs/2001.09784>

[72] Desilver, D. & Bialik, K. (2017). Blacks and Hispanics face extra challenges in getting home loans. Pew Research Center. Retrieved from <https://www.pewresearch.org/fact-tank/2017/01/10/blacks-and-hispanics-face-extra-challenges-in-getting-home-loans/>

[73] Platt, J. C., & Platt, J. C. (1999). Probabilistic Outputs for Support Vector Machines and Comparisons to Regularized Likelihood Methods. Advances in Large Margin Classifiers, 61–74. Retrieved from <http://citeseer.ist.psu.edu/viewdoc/summary?doi=10.1.1.41.1639>

[74] Pleiss, G., Raghavan, M., Wu, F., Kleinberg, J., & Weinberger, K. Q. (2017). On Fairness and Calibration. Advances in Neural Information Processing Systems, 2017-December, 5681–5690. Retrieved from <http://arxiv.org/abs/1709.02012>

- [75] PrevueHR. Adverse Impact Analysis / Four-Fifths Rule. (2009). Retrieved from <https://www.prevuehr.com/resources/insights/adverse-impact-analysis-four-fifths-rule/>
- [76] Quiñonero-Candela, J., Rasmussen, C. E., Sinz, F., Bousquet, O., & Schölkopf, B. (2005). Evaluating Predictive Uncertainty Challenge. Machine Learning Challenges Workshop, 1–27. Retrieved from <http://predict.kyb.tuebingen.mpg.de>
- [77] Ryll, L., Barton, M. E., Zhang, B. Z., & McWaters, J. (2020). Transforming Paradigms: A Global AI in Financial Services Survey. Cambridge.
- [78] Speicher, T., Heidari, H., Grgic-Hlaca, N., Gummadi, K. P., Singla, A., Weller, A., & Zafar, M. B. (2018). A Unified Approach to Quantifying Algorithmic Unfairness: Measuring Individual & Group Unfairness via Inequality Indices. Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. ACM, 10. <https://doi.org/10.1145/3219819.3220046>
- [79] Suresh, H., & Guttag, J. V. (2019). A Framework for Understanding Unintended Consequences of Machine Learning. arXiv preprint: 1901.10002v1.
- [80] Sweeney, L. (2013). Discrimination in online ad delivery. Commun. ACM, 56(5):44–54..
- [81] Adrade D., Nielsen T., Kerrigan, B. (2020). Accelerating digital transformation while grappling with credit and reputational risks. Bank Governance Leadership Network. Retrieved from https://assets.ey.com/content/dam/ey-sites/ey-com/en_gl/topics/banking-and-capital-markets/ey-bank-governance-leadership-network.pdf
- [82] Tötsch, N., & Hoffmann, D. (2020). Classifier uncertainty: evidence, potential impact, and probabilistic treatment. Retrieved from <http://arxiv.org/abs/2006.11105>
- [83] UN Women. (2020). Equal pay for work of equal value. Retrieved from <https://www.unwomen.org/en/news/in-focus/csw61/equal-pay>
- [84] Wachter, S., Mittelstadt, B., & Russell, C. (2020). Why Fairness Cannot Be Automated: Bridging the Gap Between EU Non-Discrimination Law and AI. SSRN Electronic Journal. <https://doi.org/10.2139/ssrn.3547922>
- [85] Williamson, R.C., Menon, A.K. (2019). Fairness risk measures. International Conference on Machine Learning. <https://arxiv.org/abs/1901.08665>
- [86] Yeoman, R. & Mueller Santos, M. (2016). Mutuality in Business. Fairness and Organization Performance: Insights for Supply Chain Management.
- [87] Zemel, R., Wu, Y. (Ledell), Swersky, K., Pitassi, T., & Dwork, C. (2013). Learning Fair Representations. Retrieved from <http://proceedings.mlr.press/v28/zemel13.pdf>
- [88] Zhang, Y., Bellamy, R. and Varshney, K. (2020). Joint optimization of AI fairness and utility: A human-centered approach. Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society.



Appendices





Methodology technical reference

Note: there are no additional technical references for Part A of the Methodology.

1.1

Part B: Examine data and models

Data used in an AIDA system

Different data may be used in the operation of the system in a number of ways, for example:

- historical customer outcomes used as targets in a supervised learning model
- features about customers used to make predictions of behaviour
- features about customers used to estimate membership to a group
- customer and bank outcomes used to assess the performance of the system

All these data are likely imperfect, in that they are measurements that fail to exactly capture the qualitative concept the AIDA System Developers are attempting to represent with them. Some types of errors in measurement, when present in data used in an AIDA system, can not only reduce the fairness of the system's operation, but do so in a way that is not observable by standard fairness measures [79, 39, 4]. Some of these types of imperfection also fall under the label of “bias” as referred to in the FEAT Fairness Principles. Other challenges, including ground truth labels being often noisy [33] or partially missing [53], and data and concept drifts occurring over time, have not yet been associated with common or practical solutions.

The following are some common types of bias — selection bias, measurement and proxy bias, and missing data bias —with reflections to help uncover their existence in the above data.

Selection bias

Selection bias occurs when the distribution of data sampled from a population (for example customers) fails to represent that population accurately. For example, when a sample of 100 people in the general population contains 80 males, then women are under-represented in the sample compared to the population (since women represent roughly 50% of the population). Algorithms trained with data containing representation bias may exhibit worse-than-expected performance on under-represented cohorts, increasing the risk that the cohort is systematically disadvantaged through incorrect predictions. This is because typical machine learning algorithms are trained to minimise loss functions averaged over the training cohort. The smaller the proportion of the training cohort a group makes up, the smaller that group's contribution to the average loss. In training, model parameters that induce poor performance on the group then have little penalty.

Measurement and proxy bias

Using AIDA systems for prediction requires that objectives and data be quantified. However, many qualitative concepts are not directly measurable, and therefore can only be represented by a choice of “proxy” variable [48]. One such example is the difference between the concept of a customer wanting to purchase a product, and the proxy used to measure it: say, the customer clicking on a displayed ad.

The use of proxy variables is a relevant consideration for fairness because sometimes a proxy may be systematically better at capturing information for some people compared to others. This can affect the performance of AIDA models for those people, in a way that may not be detectable by standard validation approaches. A high-profile example of this kind of AIDA failure was caused by using the cost of an individual's healthcare as a proxy for their overall health in a system for assigning healthcare resources, resulting in widespread, systematic under-resourcing for Black citizens in parts of the US [63]. A related form of bias that can affect even quantities that seem directly measurable is measurement bias: this is when systematic errors exist in the data measurement itself, making the data more accurate for some people compared to others. Again, if such data are used to train an AIDA model, there is a risk that this bias lowers the performance of the system of some people [33].

Missing data bias

Sometimes it may not be possible to obtain features consistently for every person in a dataset. One result of this is that the data used to train and query a system may have missing features for some individuals. This missing data may affect the way in which an AIDA system makes decisions, and so it is vital that AIDA System Developers understand the statistical properties of this missingness, and determine how it varies across the customer cohort.

Considerations on quantifying system objectives

Whether a system's models or data are accurate and without unintentional bias can only be examined in terms of whether they are able to satisfy the objectives and constraints of the system: measurements that are sufficiently accurate for one application may not be so for another. The first step toward making that judgement is to understand how the qualitative considerations in a system's objectives have been translated into quantitative measures.

Assessing the models

For all calculations relating to system performance and fairness, it is vital that the measurements from machine learning or hybrid systems occur on held-out data, that is, data that has not been used to inform the creation of the model. For details of schemes for selecting models with good generalisation performance and creating held-out data sets, see for example [45]. Using recent production data with outcomes to compute performance measures minimises the risk of covariate or problem shift rendering the measures inaccurate.

Quantifying model calibration and confidence

Many AIDA selection systems rely on a predicted score that signifies an individual's suitability for an action by the system. Most often this score is actually the probability of the individual having the desired outcome if an action is applied to them; for example if they are selected for a loan, it is the probability that they will not default on the repayment. Thus, for an AIDA System Owner to correctly reason about the risks of selecting individuals for an action, and consequently to properly measure the fairness of their AIDA system, the score must be calibrated to the probability of the desired outcome. More formally following the convention in [4], the label $Y = 1$ signifies the designed outcome (successfully repays a loan) for an individual, and $R = r$ is the score variable that is predicted by an AIDA system for the same individual, then calibration can be defined as $P(Y = 1 \mid R = r) = r$

That is, all individuals assigned a score of r will have a proportion of r desired outcomes amongst them. Many machine learning and statistical models will by default produce a calibrated score, such as those that minimise log loss, Brier loss, cross-entropy loss or some other proper loss [11, 41]. For those that do not such as many ensemble methods (e.g. random forests) or max-margin classifiers (e.g. support vector machines), calibration reliability curves [69] and methods such as Platt scaling [73] can be used to check and correct output scores for calibration. Calibration of a score amongst different groups in a population is also related to

the sufficiency or multi-calibration notions of fairness [4, 46]. If a certain population has a score that is uncalibrated with respect to the rest of the population, this can result in them being unfairly selected or unselected by an AIDA system.

Establishing confidence in the performance and fairness measures

AIDA systems, particularly those using machine learning or statistical models, require performance measures (including those relating to fairness) to be quantified mathematically so that they can be optimised or constrained. However, these measures are necessarily computed from finite samples of data, or may even be estimates from other models, and so have an inherent statistical sampling and estimation uncertainty. Furthermore, some of the sample sizes may be very small, particularly for measures of fairness where underrepresented demographic groups make up the sample. Since consequential decisions will be made based upon the values of these measures it is also important to convey the uncertainty inherent in their values.

For example, a decision made based on a system that is measured to be 1.2 times as accurate on men as it is on other genders may be different from a decision that is made on the same system if it is measured to be between 0.7 and 1.3 times as accurate with 95% confidence. In the first instance (1.2 times) a decision may be made to differentially select men and other genders in order to be fair. However with more information about the uncertainty of this measure, a different decision may be made to acquire more data in order to satisfactorily measure the system's performance. Furthermore, the first decision to differentially select men and other genders may actually lead to unintended consequences if the true differential accuracy was closer to 1.0 or parity.

How to compute uncertainty in the measures will depend on the measure used, and whether the measure is being evaluated on training data or deployment cohort data. There are a number of widely applicable and commonly used ways of quantifying uncertainty in estimates derived from finite samples. Some suggestions follow.

Training data (or training/validation splits) - When computing measures on the training data, these measures must be computed on leave-out validation data, ideally using cross validation with as many replications or folds as feasible (though some data, such as time series, will require different leave-out approaches). Model retraining should occur using the leave-in data on each replication or fold, and the measures computed on the leave-out replications or folds may be used as samples from which to create uncertainty estimates. Note that the uncertainty computed from these samples itself contains error:

- If the cross validation procedure is used for model selection, then overfitting is possible and the measured performance may be overstated in each fold/replication. If this data is to be used for model selection, then a separate test dataset should be kept separate that is not used for model selection, and purely used for uncertainty estimation.
- If a cross validation procedure such as *K*-Fold cross validation is used, then the samples produced from each validation fold are not independent, and so the resulting uncertainty estimates (such as variance) may not reflect the true uncertainty.

For a thorough discussion of these points and of appropriate validation procedures, see Chapter 7 of [45].

Test and Deployment data - It is also vital to estimate uncertainty of the performance measures when evaluating a model's performance and fairness on a test or deployment dataset, where model retraining does not occur (and the data is not used for model selection). There are many contextual and model specific approaches to this scenario, but two are broadly applicable to the use case examples in this document:

- Bootstrap confidence intervals: this is a suite of nonparametric resampling methods for estimating the percentiles of a distribution [23]. This can be applied to the calculation of many standard measures of performance and fairness.
- Bayesian credible intervals: this approach requires defining parametric models describing the statistical distributions of observations and unknown parameters. In particular [14, 82] describe Bayesian approaches for estimating credible intervals of measures derived from classification confusion matrices. For some choices of prior and likelihood function, this procedure can be computationally inexpensive and may also be simpler to apply to very small sample sizes.

Chaining predictive models and predictive uncertainty

When feasible it is useful to have models used in AIDA systems output an uncertainty associated with each prediction. This can help inform decision-making at an individual level, for example, to manually treat an individual's case where there is high predictive uncertainty. This predictive uncertainty must also be calibrated, see [76] for a detailed review, and [65] for guidance on situations when this predictive uncertainty may or may not be trusted.

Some data used as features or targets in modelling may itself be the result of another model's predictions. In this case, all the concerns about data bias in Part B of the Methodology still apply. Note that the practice of chaining predictions in this way risks amplifying errors even if the outputs of each predictor are well calibrated and contain uncertainty information [65].

1.2

Part C: Measure disadvantage

Existing fairness measures in the literature

Many "fairness measures" have been developed in an effort to address the problem of mathematically specifying a fairness objective for an AIDA system. Computing and analysing a selection of these existing fairness metrics can mitigate the risk that a harm or benefit was missed in an initial analysis. However, note that all fairness measures capture different notions of fairness, each with their own limitations. Which of these notions of fairness are important for a particular system, and whether any are sufficient, is a context-dependent ethical question with no objective or universally accepted answer.

Some taxonomy of fairness measures has developed in the literature, with two broad categorisations for fairness commonly in use: individual fairness, where the treatments of individuals are compared, and group fairness, where the treatments of groups of individuals are compared [87]. Even once determined to be relevant, both types of measure still rely on value judgements to apply: group fairness measures require defining the groups between which comparisons should be made, whilst individual fairness measures require defining a notion of similarity between individuals. The following subsections examine some of these well-known fairness measures. For a more complete review of group and individual fairness measures in the literature, see [9].

Measures based on independence, separation and sufficiency criteria

Standard measures of fairness for both classification and regression settings commonly derive from notions of independence, separation and sufficiency [4]. These classes of measures capture many aspects of treatment and error distributions for individuals and groups, and disparities in these measures can serve as warning signs to investigate the causes and consequences of these errors more closely. Each of the three fairness criteria has its own benefits and issues, and crucially, it is likely impossible to satisfy all of them simultaneously if there are underlying differences (such as, in the context of credit scoring, different average income or wealth levels across groups). See [5, 24] for overviews of these measures with code implementations.

Given a personal attribute A , the target variable Y , and the classifier or score R :

Independence means that $P\{R = 1 \mid A = a\} = P\{R = 1 \mid A = b\}$, and has been discussed as demographic or statistical parity [25] or disparate impact. In other contexts (for instance, hiring discrimination), bounds on independence have been used as the fairness criteria by supervisory agencies [75], expressed mathematically as:

$$\frac{P\{R = 1 \mid A = a\}}{P\{R = 1 \mid A = b\}} \geq 80\%$$

Conditional Demographic Parity (or Disparity) [52] is another independence measure that controls for a set of chosen properties of the cohorts, and is being considered as a judicial criterion in the EU [84].

The benefit of independence is that it expresses common ethical theories about human equality — the belief that people’s race, gender or other characteristics do not make them inherently more or less qualified to attend a school, perform a job, pay back a loan, or anything else. The downside of independence is that it can have unintentional negative effects over time if there are correlations — for example, one group might have a higher default rate on loans than another. In this case, if a credit model approves applications from groups a and b at similar rates, but is less accurate with its predictions on members of group b (perhaps because the FSI has less training data about their historically underserved group), then applicants from the group b pool will have higher default rates over time.

As a matter of business priorities, an AIDA System Owner may prefer to consider base rates of outcomes when developing fairness objectives, for example having different lending rates for groups with different default rates. Separation allows correlation between the score and

the personal attribute to the extent that it is justified by the target variable (such as default rates). This translates mathematically to:

$$\mathbb{P}\{R = 1 \mid Y = 1, A = a\} = \mathbb{P}\{R = 1 \mid Y = 1, A = b\}$$

$$\mathbb{P}\{R = 1 \mid Y = 0, A = a\} = \mathbb{P}\{R = 1 \mid Y = 0, A = b\}$$

which means that all groups experience the same true positive rate and the same false positive rate. The downside of separation is that the two populations will have different tradeoffs between false positive and false negatives, and with multiple groups it may be impossible to find a point that is achievable for all groups. In this case, some individuals will end up being treated randomly to fulfill the group level criteria. A well-known group fairness measure motivated by separation is “equality of opportunity” (which relaxes the condition to only consider positive labels) [44].

Sufficiency which is often discussed in terms of calibration implies that out of all individuals assigned a score r , there is a fraction r of positive instances, for both groups a and b $\mathbb{P}\{Y = 1 \mid R = r, A = a\} = \mathbb{P}\{Y = 1 \mid R = r, A = b\}$. In another context (recidivism risk prediction) the makers of software tools have claimed equal calibration is a good metric for the fairness of a system [17]. The downside of calibration as a metric for fairness is that it can simply reproduce existing inequalities. If past social conditions (such as redlining in the US [49]) have produced issues such as income or wealth gaps, calibration will not increase “fairness” as it would be understood in an egalitarian or desserts-based ethical context [54]. That is, the approach would not correct for an individual’s “luck” in being born to a certain set of parents, nor move toward a more equal distribution of benefits across the population. “Precision parity” is a commonly used group fairness measure motivated by the sufficiency condition [44].

Quantifying deviance from fairness criteria

When any of the above fairness measures deviate from their fairness criteria a choice must be made as to how that deviation should be quantified. For example, consider measuring equal opportunity between groups A and B . The fairness criteria require that the true positive rate (TPR) in group A match that in group B , i.e. $TPR_A = TPR_B$. If they do not exactly match, the AIDA System Owner (in concert with the AIDA System Developer) must choose how to express the deviance. One approach may be via a difference, i.e. $TPR_A - TPR_B$, another is via a ratio, i.e. TPR_A / TPR_B . These measures correspond to the English language statements such as “women are 10% more likely to receive the benefit than the average” for a difference or “women are twice as likely to receive the benefit than the average” for a ratio.

The interpretation of the resulting equal opportunity metric is quite different in each case. In the first case a value of 0.2 should be interpreted as 20% more of the “non-defaulters” in group A gain access to credit than the “non-defaulter” in group B . In the latter case it should be interpreted as 1/5 as many of the “non-defaulters” in group A receive loans as compared to those in group B . Care should be taken to communicate the meaning of these numerical values. Regardless of how the deviations are quantified, the uncertainty on their values should also be estimated. In circumstances where group membership is relatively small this is especially important. This may be approached using parametric (e.g. chi-squared), non-parametric (e.g. bootstrap resampling) or Bayesian techniques. The following reflections may be helpful in quantifying deviance:

- Can we know what measures (difference, ratio or other) different stakeholder groups (e.g., customers, risk managers) would prefer?
- How will the choice of difference measure be justified?
- Who, within the FSI, will validate that choice? Will they be able to interpret how this choice defines the tradeoffs that are being made?

Fairness as a distribution of harms or benefits

Many standard measures of fairness from the economics or mathematical finance literature take the form of inequality indices or risk measures that directly measure the distribution of harms or benefits amongst the population, see [58] for a review.

This literature has recently been revisited in the context of machine learning systems [78, 85]. The approach of this work is to first quantify the harms or benefits from an AIDA system in terms of an individual benefit or loss function directly. Once these functions have been quantified, the next step is to measure the distribution of these losses or benefits amongst the population using thoroughly analysed and well-understood economic indices such as generalised entropy [78] or risk measures such as conditional value at risk [85].

The work based on these approaches is able to encode many of the common machine learning fairness measures as special cases. But it has the potential advantage of being more conceptually simple to translate specific harms or benefits from an AIDA system into measures of fairness that have well-understood mathematical properties. Furthermore, some of these measures have properties, such as additive decomposability [78], that make them robust to “fairness gerrymandering” — the practice of making an algorithm seem fair by choosing personal attributes (or combinations thereof) that exhibit less disparate outcomes. These properties are also valuable when arguing or evaluating anti-discrimination cases that rely on fairness gerrymandering as a defense [84] or when assessing performance of AIDA systems as exemplified in the work of [13].

Measures accounting for long term and feedback effects

Group fairness measures such as those described above typically consider a set of decisions at a moment in time, neglecting the possibility that individuals or groups interacting with the system over time may induce feedback or other long term effects. In this case, even a system that satisfies a particular fairness measure at any given point in time, may not do so when examined over a longer period [56]. Some approaches to measuring and addressing long term fairness considerations have recently been developed that may be particularly relevant to AIDA applications in the financial sector such as credit scoring or direct marketing [19].

Individual Fairness

While issues relating to group fairness map over most closely to discussions in policy and law, individuals are not defined solely by their personal attributes, and two individuals sharing the same personal attributes are not interchangeable on that basis. Care must be taken, especially when making changes to a system in order to improve a measure of group fairness, that they are not treated as such. For example, suppose one wishes to increase the number of female applicants accepted for a credit product from 10,000 to 15,000, and changes to the system are made accordingly. There is nothing preventing these changes from rejecting 3,000 of the women that would have been accepted previously, so long as 8,000 additional women are now accepted. While outcomes may have been improved for women as a broad group, those 3,000 individuals are still harmed.

In the literature, individual fairness tends to mean that similar individuals are treated similarly. What constitutes similarity between individuals is an open question, and is usually left to the determination of a domain expert for the use case. What it means to be treated similarly is also more subtle when talking about individuals, since similar risk scores are more likely to lead to different outcomes when making comparisons at an individual level. When assessing similar treatment, it is unclear whether one should focus on similar risk scores or similar outcomes. For example, if person **A** is given a risk score 0.51 and person **B** (a similar individual) is given a risk score of 0.49, it is possible that **A** gains access to credit while **B** is rejected. This is further complicated by certain post-hoc group fairness mitigation strategies, for example those proposed for equalised odds [\[44\]](#). Under these strategies, fairness may be achieved by rejecting all applicants below a certain risk score, accepting all applicants above a different risk score, and randomly accepting those with scores in between. The following reflections may be helpful in considering individual fairness:

- Is some degree of individual unfairness an acceptable consequence of promoting group fairness?
- Are there post-hoc interventions (such as explanations) that can decrease customers' feelings of individual unfairness (even if they do not change the actual incidence of unfairness)?

1.3

Part D: Justify the use of personal attributes

Limitations of fairness through unawareness

Removing a personal attribute from the system may not prevent the system from inferring the attribute using the remaining features and potentially disadvantaging the individual or group as a result. It is therefore important to monitor the system's treatment of groups defined by personal attributes even when their identifying attribute(s) has been withheld from the system.

AIDA models are commonly said to be attempting “fairness through unawareness” if personal attributes are not explicitly used in the prediction process. Fairness through unawareness approaches are often ineffective at removing disparities in outcomes relating to the personal attribute, however, especially with large datasets. There are numerous real-world examples [68, 1, 22, 80, 47] where excluded personal attributes still influenced an AIDA model through proxies that have resulted in unwanted outcomes.

The ability for an attribute to influence outcomes when not present in an AIDA model is due to the existence of redundant encodings — where features that are used in the model are correlated with a personal attribute, allowing the model to operate as if it had access to the attribute. The personal attribute can even be reproduced from weak correlations, if these correlations occur across many features [15].

Understanding the impact of using attributes

To understand the consequences of including or excluding attributes, AIDA System Developers can utilise feature importance techniques. These are a collection of methods that can estimate the extent a feature contributes towards both the AIDA fairness and operating performance objectives. Feature importance techniques such as LIME or SHAP could be used to illustrate the impact of the feature on the performance objectives [61], though these and other explainability techniques may themselves introduce bias [36].

The following two feature importance approaches are model-agnostic and simple to implement [20]:

- **Permutation Feature Importance (PFI) [10]:** the idea behind this approach is that when permuting the values of a given feature, its explanatory power mitigates as it breaks the association to the outcome variable. Therefore if the model relied on the feature, the performance error of the model should increase when predicting with the “permuted feature” dataset instead of with the “initial feature” dataset. The importance of the given feature is then evaluated by the difference of the prediction error and fairness metric.
- **Leave-One-Covariate-Out (LOCO) [55]:** this approach follows the same objective as PFI, to understand the importance of a specific feature against the AIDA model performance and fairness objectives. This approach involves training a new model by dropping each feature, one at a time, and observing the impact to the fairness and performance objectives.

Note that PFI does not require model re-training for each feature importance calculation, whereas LOCO does. Thus PFI has computational advantages, especially in case of complex models and large feature spaces.

1.4

Part E: Examine system monitoring and review

Dataset shifts

A common assumption in machine learning and other AIDA models is that the input data in the training set, that is the targets, y and features or covariates, x , follow the same probability distribution as the input data, x , and true outcomes, y , when the system is in production. Dataset shift is the general term to describe this assumption being violated, a state that can considerably degrade model performance and is difficult to detect.

Dataset shifts can result from a number of different processes: selection bias in the training data itself, or changes in the distribution of the customer cohort over time, or even external events that alter the relationship between the model's features and its targets. For instance, overall economic conditions or labour market conditions may have deteriorated making it less likely that applicants (even those with relatively good credit scores) would repay their loan at the same rate as in the past. These processes can result in the distribution of features/covariates changing between the training set and the query set (covariate shift), or the relationship between the features and the targets changing between the training set and the query set (concept shift).

Dataset shift notation

Formally, dataset shift is when the probability distribution of training data, $P_{\text{train}}(x,y)$, used to train a model is not the same as the distribution of data used to query the model in production, $P_{\text{query}}(x,y)$. Two common manifestations of dataset shift can be understood from looking at different components of this joint distribution of features or covariates, x , and targets, y . Using the chain rule of probability, the joint distribution can be equivalently expressed as

$$P(x,y) = P(y|x) \times P(x)$$

When there is a shift in the distribution of covariates or features, this is known as covariate shift,

$$\text{Covariate shift: } P_{\text{train}}(x) \neq P_{\text{query}}(x)$$

When the relationship between covariates and the targets changes, this is known as concept shift,

$$\text{Concept shift: } P_{\text{train}}(y|x) \neq P_{\text{query}}(y|x)$$

In both of these cases, any cross validation used to check the model's performance on the training data is invalid since cross validation is built upon the assumption that the probability distribution of data remains the same between the training and query data.

Detecting and addressing problem shift

There are two common approaches in the literature to detect covariate shift. Both require having samples of features (x) from the production query set for comparison with the training set.

- The first approach is to train a binary classifier to differentiate between features from the training set and features from the query set [8]. High accuracy of this classifier indicates separation between the two distributions, i.e. that covariate shift is present. However, the threshold of accuracy that signifies significant covariate shift will be problem-dependent.
- The second approach to detecting covariate shift is to use two-sample statistical testing procedures that attempt to determine if two samples originate from different distributions (i.e. training or query). For example, the maximum mean discrepancy test [42] is a generally applicable non-parametric testing framework.

Unfortunately, no general method exists for detecting concept shift prior to the deployment of the system from the query data set. Concept shift often manifests as poorer prediction performance than that estimated from cross validation on training data, only after a system has been deployed and the true target outcomes, y , have been observed.

The risk of concept shift underscores the need, wherever possible, to obtain correct target labels for predictions made by an AIDA system as soon as possible after those predictions are made. Furthermore, before an AIDA system is deployed the AIDA System Owner (in concert with the AIDA System Developer) should consider carefully if there could be a concept shift between the training data used for the system, and the data the system will encounter in production. For example, has regulation surrounding a financial product central to the system changed between training and deployment?

Redress mechanisms

In some jurisdictions outside Singapore, there is either specific or general legislation that requires applicants who receive an adverse decision to also receive some sort of explanation for that decision. In the US, ECOA requires adverse action reasons to be provided to loan applicants [27]. In the EU, GDPR requires the decision-subject to receive an explanation (including the “logic involved”) of a fully automated decision and the opportunity to challenge the decision or request human intervention [84]. In Singapore, there is no regulation requiring explanations, or providing redress or challenge rights. Nonetheless, for some AIDA systems, providing an actionable explanation (such as that a loan decline was due to a high debt-to-income ratio) might encourage applicants to adapt to address the issue. Actionable explanations might also allow applicants to challenge a decision, avoiding lost opportunity for the FSI (for example, if something like a data entry error caused an issue with the algorithmic assessment). The downsides of having a manual redress mechanism include additional cost, and that it introduces the potential for more discretion or overrides which can themselves be a source of systematic disadvantage.

If the FSI does have a redress or challenge mechanism for its decisions, then this is a valuable signal on the reliability of the model. For example, if most requests for redress come from a specific group, this might be a sign that there is unintentional bias. Similarly, if most requests for explanations come from a specific group, this might be a sign that they perceive the accuracy to be low for them.

Potential mitigating actions

The potential mitigation actions related to the FEAT Fairness Principles will be as diverse as the use cases for AIDA systems and depend on the specific findings of an assessment. There are three main intervention points that FSI or their AIDA System Developers may consider. For a review, see [\[79\]](#).

2.1 Data collection and preparation

As discussed in the Methodology Part B (Section 3.2.2), selection bias (which occurs when the population used for model development under-represents, and thus fails to generalise well for some part of the population which the model is applied to) can be a source of systematic disadvantage. The same is true for the evaluation population: if the testing or benchmark populations do not represent various parts of the population of decision subjects equally well then the model's error profile may be skewed in various ways.

Collecting more data, especially for underrepresented groups, is one way to mitigate this type of risk. If additional data collection is impractical, there are statistical techniques such as data reweighting that can be helpful [\[16, 50, 51, 43\]](#). Practitioners can also consider changes to how the data is preprocessed. For instance, personal attributes may be changed from being binary (e.g. "white" and "not white") to codings that can handle multiple racial categorisations. Practitioners can also consider whether the proxies that are being used for variables of interest are the best available. For example, is "self-employed" used as a proxy for "likely to have uncertain future cash flows"? The downstream techniques that follow are sensitive to these types of choices (such as how data is split into train and test sets and what preprocessing is performed), so it is important to have a coherent mitigation strategy [\[37\]](#).

2.2 Model design and development

Many populations of interest are heterogeneous and a single model may be unlikely to perform well on all subgroups. This is more obvious in the medical domain because patients from different groups with similar conditions may have different symptoms and disease progressions. Nonetheless, there may be important group differences in the financial realm — perhaps risk models that are appropriate for most retail customers are less relevant for entrepreneurs. In these cases, practitioners can consider ensembles of models.

Whether a single or multiple models are being used, the behaviour of the AIDA system will be driven by what it is optimised to do within the constraints it is given. For example: a credit scoring and approval system might be optimised to generate the highest expected revenue while keeping total expected defaults below a certain threshold. If a practitioner is looking to mitigate some aspect of unfairness, one strategy is to add regularisation terms or constraints that enforce a particular fairness objective (see for example [\[6\]](#)). It may also be possible to constrain how the model learns in order to "mask" out a personal attribute [\[87\]](#), or to automatically tune the hyperparameters of a standard model to optimise performance given fairness constraints [\[70\]](#).

Whichever method for incorporating fairness objectives into an AIDA model is used, the challenge of balancing tradeoffs between these and other objectives remains. One potential approach for determining these tradeoffs is to elicit preferences about outcomes or objectives from relevant policy-makers, and adjust the model to adhere to these preferences (see for example [\[88\]](#)).

2.3 Post-processing

In many AIDA systems, there are steps between the output of a model and a real-world outcome. For instance, credit scores output default probabilities, which are combined with other business rules to make actual approval decisions. These intermediate steps offer additional opportunities for practitioners to intervene. For instance, different decision thresholds can be set for different groups (i.e., an individual in one group might need to be in the top 80% of applicants to be approved while individuals in a less well-represented group might only need to be in the top 70% of applicants in their group) [\[44, 18\]](#).

It should be noted that processing approaches (pre-, in-, post-) may change outcome patterns but create their own issues. For instance, if historical bias has meant that members of one group on average have less wealth or income than another group, initiatives to lower approval thresholds may increase the disadvantaged group's access to credit in the short term, but also raise their default rate over time if the post-processing intervention did not address the underlying issue of ability to pay. This may even exacerbate issues if the data used to train future models shows a heightened correlation between group membership and default, which might make it even harder for members of the group to access credit in the future. The issue of negative feedback loops due to interventions is an emerging field of scholarship [\[56, 19\]](#).

Software resources

The following list contains a set of commonly used open source tools for the measurement and improvement of fairness in AIDA systems.

AI Fairness 360 (IBM) is an open source library to help detect and mitigate bias in machine learning models. It includes a comprehensive set of metrics for datasets and models to test for biases, explanations for these metrics, and algorithms to mitigate bias in datasets and models. URL: <https://github.com/Trusted-AI/AIF360>

Fairlearn (Microsoft) is a Python package that empowers developers of artificial intelligence (AI) systems to assess their system's fairness and mitigate any observed unfairness issues. Fairlearn contains mitigation algorithms as well as a Jupyter widget for model assessment. URL: <https://github.com/fairlearn/fairlearn>

fairness-indicators (Google) is designed to support teams in evaluating, improving, and comparing models for fairness concerns in partnership with the broader Tensorflow toolkit. URL: <https://github.com/tensorflow/fairness-indicators>

tensorflow-constrained-optimization (Google) is designed to help address issues like fairness constraints and predictive parity and help machine learning practitioners better understand things like true positive rates on residents of certain countries, or recall illness diagnoses depending on age and gender. URL: https://github.com/google-research/tensorflow_constrained_optimization

ML fairness gym (Google) is a tool for exploring the potential long term impacts of deploying machine learning-based decision systems in social environments. It implements a generalised framework for studying and probing long term fairness effects in carefully constructed simulation scenarios where a learning agent interacts with an environment over time. URL: <https://github.com/google/ml-fairness-gym>

What-If Tool (Google) among other things contains tooling for investigating model performance and fairness over subsets of a dataset. URL: <https://github.com/PAIR-code/what-if-tool>

FAT Forensics is a Python toolbox, built on top of SciPy and NumPy, for evaluating algorithmic fairness, accountability and transparency of predictive systems. URL: <https://github.com/fat-forensics/fat-forensics>

Fairness is an R tool that provides tools to easily calculate algorithmic fairness metrics for given predicted probabilities or predicted classes between different groups. It also provides opportunities to visualise and compare other prediction metrics between the subgroups. URL: <https://github.com/kozodoi/fairness>

FairPut is a holistic approach to implement fair outputs at the individual and group level. FairPut is built on Alibi, AIF360, AIX360, SHAP, and PDPbox. URL: <https://github.com/firmai/ml-fairness-framework>

Aequitas (University of Chicago) is an open-source bias audit toolkit for data scientists, machine learning researchers, and policymakers to audit machine learning models for discrimination and bias, and to make informed and equitable decisions around developing and deploying predictive tools. URL: <https://github.com/dssg/aequitas>

WhyNot is a Python package that provides an experimental sandbox for decisions in dynamics, connecting tools from causal inference and reinforcement learning with challenging dynamic environments. It can be used to study the effects of a fairness mitigation strategy in a dynamic environment.

URL: <https://github.com/zykls/whynot>



Suggested reading list

The following is a compilation of selected academic readings, with brief summaries, for those seeking a foundation in fairness literature.

#	Reference and annotation	Category
1	Fairness and Machine Learning Barocas, S., Hardt, M., Narayanan, A. Online Publication, 2020. https://fairmlbook.org/ <i>Chapter 4 is a summary of many of the group fairness measures for classification algorithms. It also includes many links to relevant papers. This related simulation by some of the co-authors is also worth looking at http://research.google.com/bigpicture/attacking-discrimination-in-ml/</i>	Core reading on fairness
2	A Framework for Understanding Unintended Consequences of Machine Learning Suresh, H., Gutttag, J.V. arXiv preprint, 2019. https://arxiv.org/abs/1901.10002 <i>"In this paper, we provide a framework that partitions sources of downstream harm in machine learning into six distinct categories spanning the data generation and machine learning pipeline. We describe how these issues arise, how they are relevant to particular applications, and how they motivate different solutions."</i>	Core reading on fairness
3	Interpretable ML Book Molnar, C. Online Publication, 2020. https://christophm.github.io/interpretable-ml-book/ <i>A comprehensive presentation of algorithms for interpreting machine learning algorithms.</i>	Justifiability & explainability

#	Reference and annotation	Category
4	The Cost of Fairness in Binary Classification Menon, A.K., Williamson, R.C. Proceedings of the 1st Conference on Fairness, Accountability and Transparency, 2018. http://proceedings.mlr.press/v81/menon18a.html <i>A deep dive into the tradeoffs between accuracy for a given level of fairness in terms of costs, as well as establishing an efficient set (Pareto set) of classifier threshold settings.</i>	Fair classification
5	Why Is My Classifier Discriminatory? Chen, I., Johansson, F.D., Sontag, D. Advances in Neural Processing Systems, 2018. http://papers.nips.cc/paper/7613-why-is-my-classifier-discriminatory <i>Argues that classifiers should be made “fair” through additional data gathering rather than the usual approach of thresholding. This is a way of “getting around” the fairness-accuracy tradeoff analysed in the previous paper.</i>	Fair classification
6	Learning Fair Representations Zemel, R., Wu, Y., Swersky, K., Pitassi, T., & Dwork, C. International Conference on Machine Learning, 2013. https://www.cs.toronto.edu/~toni/Papers/icml-final.pdf <i>Trades off individual fairness (similar individuals should be treated similarly) with group fairness (statistical parity) with predictive utility (classification accuracy) by learning a fair representation of the data which aims to optimise all three.</i>	Group and individual fairness
7	Fairness through awareness Dwork, C., Hardt, M., Pitassi, T., Reingold, O., Zemel, R. Proceedings of the 3rd Innovations in Theoretical Computer Science Conference, 2012. https://dl.acm.org/doi/10.1145/2090236.2090255 <i>Original work on individual fairness, which basically enforces the notion of similar individuals having similar outcomes. This is an intuitive concept, but may be hard to implement in practice.</i>	Group and individual fairness

#	Reference and annotation	Category
8	Equality of Opportunity in Supervised Learning Hardt, M., Price, E., & Srebro, N. Advances in neural information processing system, 2016. https://arxiv.org/abs/1610.02413 <i>Introduces two fairness criteria: equalised odds and equal opportunity, and shows how they solve shortcomings of other group fairness metrics. Demonstrates how to enforce these criteria post-hoc on classifiers and regressors. Includes a case study on FICO credit scores.</i>	Group and individual fairness
9	Avoiding discrimination through causal reasoning Kilbertus, N., Carulla, M.R., Parascandolo, G., Hardt, M., Janzing, D., Schölkopf, B. Advances in Neural Information Processing Systems, 2017. https://papers.nips.cc/paper/6668-avoiding-discrimination-through-causal-reasoning <i>Presents an argument about when observational fairness criteria fail, and how a causal approach to tackling discrimination resolves these issues. The authors present two methods for including/excluding features in predictive models based on causal reasoning justifications.</i>	Causal fairness
10	Delayed impact of fair machine learning Liu, L.T., Dean, S., Rolf, E., Simchowitz, M., Hardt, M. International Conference on Machine Learning, 2018. https://arxiv.org/abs/1803.04383 <i>A first-order, theoretical exploration of how commonly used group fairness constraints specific to one-shot learning fail to enforce fairness in dynamical settings over time.</i>	Feedback and long term impacts on fairness

Annexure



List of Veritas Consortium Members for Phase One

1. Amazon Web Services
2. AXA
3. Bank of China
4. BAFT (Bankers Association for Finance & Trade)
5. BNP Paribas
6. BNY Mellon
7. CFA Society Singapore
8. Citi Singapore
9. DBS Bank
10. Element AI
11. EY
12. IAG Firemark Labs & Gradient Institute
13. Goldman Sachs
14. HSBC
15. Industrial and Commercial Bank of China Limited
16. Microsoft
17. Monetary Authority of Singapore
18. MUFG Bank, Ltd
19. Ping An of China
20. Prudential Assurance Company Singapore
21. SGInnovate
22. Standard Chartered Bank
23. Swiss Re
24. Unionbank of the Philippines
25. United Overseas Bank Limited

LEGAL NOTICE

This report is prepared and issued by the MAS, HSBC, UOB, EY, Element AI, Gradient Institute, and IAG Firemark Labs.

All intellectual property rights in or associated with this report remain vested in the MAS, HSBC, UOB, EY, Element AI, Gradient Institute, and IAG Firemark Labs 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, HSBC, UOB, EY, Element AI, Gradient Institute, and IAG Firemark Labs 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.