MAS Veritas FEAT原则评估方法论

新加坡金融管理局 (Monetary Authority of Singapore, MAS)

2022年

中文翻译版 · 仅供参考，以英文原文为准

原文来源：https://www.mas.gov.sg/-/media/mas-media-library/news/media-releases/2022/veritas-document-3---feat-principles-assessment-methodology.pdf

---

前言

由新加坡金融管理局（Monetary Authority of Singapore, MAS）和行业合作伙伴组成的Veritas联盟，欣然发布Veritas第二阶段白皮书，评估人工智能和数据分析（Artificial Intelligence and Data Analytics, AIDA）系统与公平、伦理、问责和透明（Fairness, Ethics, Accountability and Transparency, FEAT）原则的一致性。这些白皮书包含了Veritas联盟的集体经验，分享了通用的FEAT评估方法论。白皮书还包括了预测性承保、客户营销、欺诈检测和信用风险评分等选定用例的FEAT原则评估说明。

2021年1月，MAS作为Veritas第一阶段发布了两份文件——《FEAT公平原则评估方法论》和《FEAT公平原则评估案例研究》。第一阶段的重点是银行业信用风险评分和客户营销领域的公平评估方法论开发。在第二阶段，方法论扩展至涵盖所有FEAT原则，并适用于更多用例。白皮书还强调了现有的全球监管格局及其对金融机构负责任AI采用的指导。

随着第二阶段的结束，Veritas倡议在提供完整的FEAT评估方法论方面取得了重大进展。该评估方法论将使金融机构能够系统地评估其AIDA系统与FEAT原则的一致性。此外，Veritas联盟还开发了一个开源Veritas工具包（Veritas Toolkit），实施了Veritas FEAT评估方法论。

在Veritas的下一阶段，联盟将继续增强FEAT评估方法论，并将其应用于金融服务行业的更多用例。联盟还将开发Veritas工具包的第二版，提供完整的FEAT评估功能。

我们鼓励金融机构和科技公司阅读这些白皮书并在自身组织中实施该方法论，以及使用Veritas工具包。Veritas联盟欢迎反馈以便进一步改进。让我们携手合作，确保AIDA在新加坡金融领域以更负责任的方式使用。

我们向联盟的所有成员表示衷心感谢，感谢他们在白皮书开发中的积极参与和慷慨支持。我们也感谢行业合作伙伴——Accenture、AXA、HSBC、Standard Chartered、Swiss Re、TruEra和UOB在这项卓越工作中的贡献。

Sopnendu Mohanty（MAS首席金融科技官）

> Veritas由新加坡副总理王瑞杰在2019年新加坡金融科技节暨新加坡创新与科技周（SFF x SWITCH）上作为新加坡国家AI战略的一部分宣布。

目录

执行摘要

2018年，新加坡金融管理局（MAS）发布了促进新加坡金融领域使用人工智能和数据分析（AIDA）的公平、伦理、问责和透明（FEAT）原则。Veritas倡议随后于2019年作为新加坡国家AI战略的一部分启动。

Veritas是一个与金融行业合作的多阶段项目。2021年1月，MAS宣布Veritas第一阶段成功完成，重点是银行业信用风险评分和客户营销领域的公平评估方法论开发。Veritas联盟发布了公平评估方法论的白皮书和两个用例的开源代码。

Veritas第二阶段将关注范围扩展至涵盖所有FEAT原则，供金融服务机构（Financial Services Institutions, FSIs）采用，从银行扩展到保险业。第二阶段联盟利用第一阶段的方法论作为基础，演进并提供跨FEAT原则的综合指导。本FEAT原则评估方法论白皮书是Veritas第二阶段项目交付成果之一。

本白皮书是一份总括性文件，提供了FEAT原则在AIDA系统开发生命周期中应用的整体视图。作为基础步骤，本白皮书为金融机构提供了以下建议：

· 将FEAT评估扩展至规模化
· 定义重要性和风险层级

- 采用基于风险的治理方法将FEAT原则应用于用例
- 与现有风险管理实践整合

FEAT原则应在组织层面到用例或AIDA系统层面纳入（在相关的情况下）。FEAT原则高度相互依赖，必须映射到AIDA系统开发生命周期中才能运作化。提供了一份跨端到端AIDA生命周期的检查清单，以及帮助金融机构回答评估问题的考量因素。

建议首先从伦理和问责评估开始，因为它有助于在组织层面定义指导原则，然后在AIDA系统开发生命周期的各个步骤中同时进行公平性和透明度评估，因为它们是相互关联的且针对特定用例。

本总括性文件还提供了深入方法论白皮书的摘要以及关于负责任AI采用的全球监管格局回顾。

金融机构应努力在越来越多地部署AIDA进行决策时实现FEAT原则和成果。然而，金融机构有灵活性来探索如何最好地实现这些原则，包括建立适合其业务运营和风险状况的治理框架和控制流程。

第二阶段文件不规定金融机构必须做什么来遵守FEAT，而是提供关于金融机构如何对其AIDA模型和系统进行FEAT评估的实际和详细指导，金融机构可根据情况使用和调整。

---

1. 引言

1.1 Veritas第二阶段项目背景

2018年，MAS发布了促进新加坡金融领域使用AIDA的FEAT原则。Veritas倡议随后于2019年作为新加坡国家AI战略的一部分启动，旨在通过行业合作为金融服务机构提供将FEAT原则纳入AIDA解决方案的指导。

Veritas第一阶段于2021年1月完成，重点关注银行业信用风险评分和客户营销领域的公平原则评估方法论开发。

Veritas第二阶段于2021年3月启动，将关注范围扩展至涵盖所有FEAT原则，从银行扩展到保险业。

第二阶段启动时设定了以下目标：

- 翻译和运作化FEAT原则。
- 将第一阶段方法论（公平性）从银行扩展到保险业。

- 通过Veritas框架的解决方案原型（即指南、代码、可视化等）和用例说明，协助金融机构合作伙伴采用FEAT。
- 精简跨FEAT原则的协作和方法论。
- 通过学术界、行业研究人员和从业者的参与分享行业专业知识。
- 通过可行的原型推动FEAT原则评估方法论的行业采用。

第二阶段包含三个核心工作组，分别开发其重点领域的评估方法论：

- **公平性**工作组：保险承保、客户营销（银行）
- **透明度**工作组：欺诈检测（保险）
- **伦理与问责**工作组：信用风险评分（银行）、客户营销（银行）

Veritas第二阶段共有27个联盟成员，负责领导和支持FEAT原则评估方法论的开发工作。

1.2 第二阶段方法论文件结构

第二阶段涵盖四项FEAT原则和三个工作组，文件分为方法论文件和用例文件。本FEAT原则评估方法论白皮书是一份总括性文件，提供：

1. FEAT原则在AIDA系统开发生命周期中应用的整体视图。
2. 深入方法论白皮书的摘要以及关于负责任AI采用的全球监管格局回顾。

第二阶段发布的文件套件包括：

- **文件3**：FEAT原则评估方法论（本文件）
- **文件3A**：公平原则评估方法论
- **文件3B**：伦理与问责原则评估方法论
- **文件3C**：透明度原则评估方法论
- **文件4**：FEAT原则评估案例研究
- **文件5**：FEAT原则评估案例研究说明性代码

各原则方法论的详细文件不应被视为独立文件，而应被视为跨AIDA系统端到端方法论的组成部分。

1.3 第二阶段文件受众

与第一阶段方法论类似，本文件区分了以不同方式参与AIDA系统评估的四组人员：

- **AIDA系统开发者**：开发被研究的AIDA系统的一方，可以是金融机构、供应商或金融科技公司。
- **AIDA系统所有者**：使用AIDA系统帮助做出决策以实现业务目标的一方，通常是金融机构中负责AIDA系统的业务团队。
- **AIDA系统评估者**：评估金融机构提交的FEAT评估的一方，可以是外部审计公司或内部的模型风险管理或内部审计部门。
- **金融服务机构（FSI）**：负责被研究的AIDA系统的组织。

---

2. FEAT原则评估方法论

2.1 背景

近年来，AI已成为金融机构的首要技术优先事项。根据2021年Gartner CIO调查：

- 32%的银行受访者和36%的保险受访者预期AI将在2021年成为其企业的"颠覆性力量"。
- 约25%的银行和31%的保险公司已在2020年部署了AIDA，另有25%的银行和23%的保险公司计划在未来12个月内部署AIDA。

值得注意的是，根据另一项全球高管调查，虽然60%的受访者认为在数字经济中采用AI对获得竞争优势是必要的，但45%的受访者同意对AI的意外后果了解不足。

鉴于金融机构越来越多地采用AIDA，行业监管机构可能会确保技术的合规和道德使用，以防止FEAT原则中定义的潜在负面结果或伤害。

2.2 本文件导读

本文件结构如下:

- **第3节**概述了金融机构设计AIDA系统的总体FEAT考量、FEAT原则、它们如何映射到每个FEAT领域，以及原则间的相互依赖性。第3节还解释了如何将FEAT原则评估方法论嵌入AIDA系统开发生命周期，并提供基础步骤、风险方法和组织规模化方面的指导。
- **第4节**提供了三个工作组方法论框架的摘要。
- **第5节**包含关于金融领域负责任AIDA采用的全球监管格局信息。
- **附录**包含一份FEAT遵守检查清单、常用术语表和全球AI监管格局章节。

2.3 基于FEAT考量设计AIDA系统

人工智能和数据分析涵盖辅助或替代人类决策的技术。过去十年中，来自数字渠道和第三方来源的数据可用性增加，使金融机构对AIDA的使用得以广泛发展。

AIDA系统包含一个或多个协同工作以实现一组目标的AIDA模型。AIDA系统所做的决策是"AIDA驱动的决策"。AIDA系统不仅包括AIDA模型，还包括决策过程中的其他部分:

- 影响AIDA决策的人类贡献（人类可能提供输入或使用模型输出作为最终决策的参考因素）。
- 影响决策的其他治理或规则。
- 基于AIDA系统输出采取的行动。

监控和审查机制也是端到端AIDA系统的组成部分。

使用FEAT原则评估方法论开发、部署和监控的AIDA系统可以最大限度地减少人为偏见。为评估FEAT原则在AIDA系统中的应用，金融机构需要:

- 定义与其组织价值观和原则一致的总体FEAT原则。
- 评估、衡量和监控AIDA系统如何实现这些原则。
- 定期评估总体承诺并完善原则。

FEAT原则和考量既可以存在于AIDA系统/用例层面，也可以存在于整体组织层面，且可能因地理位置和社会规范而异。鼓励受监管实体在组织层面到用例/AIDA系统层面纳入FEAT原则。

2.4 FEAT原则回顾

MAS发布了一套涵盖公平、伦理、问责和透明的14项原则，用于金融机构负责任地采用AIDA。14项原则映射到各FEAT领域如下：

公平性（Fairness） - P1：个人或群体不会通过AIDA驱动的决策系统性地处于不利地位，除非这些决策有正当理由。 - P2：使用个人属性作为AIDA驱动决策的输入因素是有正当理由的。 - P3：用于AIDA驱动决策的数据和模型定期审查和验证其准确性和相关性，并最大限度地减少非故意偏见。 - P4：AIDA驱动的决策定期审查，以确保模型按设计和预期运行。

伦理（Ethics） - P5：AIDA的使用与公司的伦理标准、价值观和行为准则一致。 - P6：AIDA驱动的决策至少遵守与人类驱动决策相同的伦理标准。

问责（Accountability） - P7：在AIDA驱动的决策中使用AIDA需经适当的内部权威机构批准。 - P8：使用AIDA的公司对内部开发和外部来源的AIDA模型均承担责任。 - P9：使用AIDA的公司主动提高管理层和董事会对其使用AIDA的认识。 - P10：为数据主体提供查询、提交上诉和请求审查影响其的AIDA驱动决策的渠道。 - P11：在审查AIDA驱动的决策时，考虑数据主体提供的经验证的相关补充数据。

透明度（Transparency） - P12：为增强公众信心，作为一般沟通的一部分，主动向数据主体披露AIDA的使用。 - P13：应数据主体请求，提供关于用于做出AIDA驱动决策的数据以及数据如何影响决策的清晰解释。 - P14：应数据主体请求，提供关于AIDA驱动的决策可能对其产生的后果的清晰解释。

2.5 FEAT原则间的相互依赖性

金融机构需要从公平、伦理、问责和透明的整体角度评估AIDA系统，以满足负责任AI采用的义务。由于部分FEAT原则是相互依赖的，金融机构在评估原则时应考虑以下方面：

公平性原则对其他FEAT原则的依赖性： - 依赖伦理原则P5和P6（使用AIDA与组织伦理价值观一致）。 - 依赖内部问责原则P7和P8（AIDA驱动决策需要内部利益相关方批准）。 - 依赖外部问责原则P10（为客户提供申诉/救济渠道）。 - 依赖透明度原则P13和P14（定期审查模型并解释决策过程）。

伦理和问责原则对其他FEAT原则的依赖性： - 依赖透明度原则P13和P14（公司管理层/董事会只有在能展示其AI系统如何达到特定结果时才能履行其问责义务）。

透明度原则对其他FEAT原则的依赖性： - 依赖伦理原则P5（客户透明度水平的决策取决于公司的伦理标准）。 - 依赖外部问责原则P10（为客户提供提问和寻求救济的渠道）。

3. AIDA系统开发生命周期中FEAT的应用

3.1 将FEAT原则评估方法论嵌入AIDA系统开发生命周期

为进行FEAT原则的全面评估，金融机构应将方法论嵌入其AIDA系统开发生命周期。作为前提步骤，建议组织根据其企业价值观制定关于开发AIDA系统的指导原则。这些价值观按照伦理和问责方法论在组织层面（或对于拥有不同业务部门的大公司则在业务单元层面）定义。

公平性和透明度原则是相互关联的且针对特定用例，因此建议金融机构同时进行公平性和透明度评估。

第一阶段建立的FEAT公平原则评估方法论设置了五个部分（A-E），包括一组检查清单问题。第二阶段以此为基础，将五个部分映射到典型的AIDA系统开发生命周期，并增加了伦理、问责和透明度方面的新考量。

FEAT评估检查清单映射到AIDA系统开发生命周期的五个步骤：

- **步骤0——原则到实践**：基础性（跨用例）
- 通用：G1-G4
- 透明度：T1-T5
- 公平性：F0
- 伦理与问责：EA1-EA4（需首先回答）
- **步骤1——定义系统背景和设计**：
- 通用：G5-G8
- 透明度：T6-T10
- 公平性：F1-F3
- 伦理与问责：EA5-EA6
- **步骤2——准备输入数据**：
- 通用：G9-G10
- 公平性：F4-F8

- **步骤3——构建和验证**:
- 通用：G11
- 透明度：T11-T17
- 公平性：F9-F11
- 伦理与问责：EA7-EA8
- **步骤4——部署和监控**:
- 通用：G12-G13
- 公平性：F12

通用问题详细说明：

步骤0——原则到实践： - G1：金融机构是否定义了AIDA？ - G2：是否有框架来定义AIDA项目/用例的角色和职责？ - G3：是否有重要性框架？是否有明确定义来确定不同的重要性水平？ - G4：金融机构是否有追踪所有重要AIDA用例的最新清单？

步骤1——定义系统背景和设计： - G5：是否记录了系统的商业目标和量化衡量标准？AIDA如何用于实现这些目标？ - G6：提议系统是否有范围边界？ - G7：AIDA的使用是否对此特定目标/用例有正当理由？ - G8：是否使用步骤0中定义的框架确定了用例的重要性？

步骤2——准备和输入数据： - G9：是否记录了系统使用的数据，包括数据类型、来源、属性、字典、时间范围、数据质量概况、代表性和知情同意？对于系统使用的个人数据，是否进行了数据保护影响评估（DPIA）？ - G10：是否记录了系统进行的数据预处理和工程？

步骤3——构建和验证： - G11：是否定义了AIDA系统的组成？每个组件如何用于实现商业目标？是否记录了性能估计和不确定性？

步骤4——部署和监控： - G12：系统的监控和审查机制是否设计用于检测异常运行（如重大模型和/或数据漂移）？ - G13：是否有后备和/或缓解计划？

重要的是，金融机构在进行与个人数据的收集、使用、处理或披露相关的活动时，应始终遵守所有相关的法律和监管要求。可以通过在审查FEAT检查清单之前进行数据保护影响评估（DPIA）来实现。

3.2 基础设置

3.2.1 为组织规模化FEAT创建正确的基础

金融机构需要确保其FEAT评估能够以一致和稳健的方式扩展到整个组织的AIDA使用中。关键建议包括：

- **采用重要性评估框架**：使用本文件中概述的评估框架或使用现有风险评估框架对AIDA系统的风险进行分级，并定义满足不同业务需求所需的治理政策。
- **将FEAT相关考量嵌入现有风险框架和治理结构**：围绕模型风险、网络安全、数据隐私和客户公平待遇的框架可以适配FEAT考量。
- **定义标准和流程**：定义应涵盖所有关键领域以提高流程的一致性和清晰度。
- **投资于适当的培训**：培训计划应包括客户、员工、管理层和风险/控制职能。
- **不要孤立地处理任何FEAT原则**：建议在整个AIDA系统生命周期中嵌入FEAT原则。
- **在内部和（如可能）外部阐明关键原则**：金融机构需要就其数据和算法的使用及AIDA决策对客户的影响传达积极、一致的信息。
- **创建并鼓励赋权个人对AIDA系统提出疑虑的组织文化**。
- **补充用例级别的工具投资**：提供更系统化的支持以帮助团队满足AIDA义务。
- **在FEAT评估过程中纳入多学科团队的见解**：包括人力资源、行为科学、法律、以人为本的设计等领域的人才。

3.2.2 确定AIDA系统的重要性

定义AIDA治理模型的关键要素之一是评估每个AIDA用例的重要性。管理风险的第一步是确定AIDA系统所构成的风险水平及其对组织的影响。

第一阶段定义了以下识别风险的参数：

- **AIDA在决策中的使用程度**：技术对人类决策的替代程度越高，可能增加意外系统性不利影响的风险。
- **AIDA驱动决策过程的自动化程度**：更多自动化可能意味着更高的决策量和/或速度，可能增加影响范围并减少人类及时干预的机会。
- **AIDA模型的复杂性**：更复杂的模型可能导致可解释性有限的模型或增加特定属性使用未充分证明的风险。

- **对不同利益相关方（包括个人）影响的严重程度和概率**：例如信用审批或定价决策比新营销信息对个人的后果更为重大。
- **货币和金融影响**：对金融机构风险敞口和资产负债表的潜在影响。
- **监管影响**：某些AIDA模型因其在反洗钱等应用中的决策而可能具有更高的监管风险。
- **救济选项**：如果客户有高效方式质疑AIDA系统的决策，相关风险可能较低。
- **声誉风险**：某些AIDA模型的决策可能对公司声誉产生更高的负面影响风险。
- **个人数据的使用**：在某些用例中可证明合理的个人数据使用程度。

举例而言，带自动拒绝功能的信用评分将是高风险用例，因为模型复杂性高、自动化程度高且对个人有潜在影响。

3.2.3 纳入基于风险的治理方法以应用FEAT原则

鼓励金融机构根据所考虑的AIDA系统的风险水平和后续风险分级来定制FEAT原则评估方法论的深度。定制维度包括：

- **分析范围和文档水平**：较高风险用例需要更广泛的分析和更详尽的回答。
- **业务目标与FEAT目标之间的权衡**：较高风险用例可能需要更多努力来维护FEAT原则。
- **监控机制的频率和广度**：较高风险用例可能需要更频繁和深入的监控。
- **审查和批准**：较高风险用例需要更严格的审查和批准流程。
- **升级流程**：较高风险用例的偏差需要更高级别的升级和干预。
- **内部与外部第三方审计**：较高风险用例可能受益于独立的第三方审计验证。

3.2.4 将FEAT原则评估方法论与现有风险管理实践整合

银行传统上有着完善的模型风险管理（Model Risk Management，MRM）实践。然而，这些风险实践关注的是减少机构因重要模型造成的重大财务损失风险，并未纳入FEAT原则的要素。

以下说明了如何将上述建议和框架结合起来增强传统MRM组件：

MRM组件	传统视角	将FEAT纳入AIDA系统的MRM
MRM的范围和适用性	关注风险管理或监管/合规用例	扩展应用到组织所有使用AIDA系统的用例；纳入基于风险的方法
模型清单	关键模型信息的捕获；验证活动和模型有效期的跟踪	

MRM组件	传统视角	将FEAT纳入AIDA系统的MRM
		模型信息目录（包括模型使用、目标、所有者等）以确保可追溯性和透明度；基于一致定义的组织范围模型识别
模型开发、验证和实施	关注数据质量、概念健全性、结果分析和定期监控	充分覆盖模型目的和使用；确保负责任地使用可用数据；提供公平性和非预期结果的模型测试和监控
风险分级和评估	关注模型重要性和关键性考量	基于AIDA系统的增强风险开发新的模型风险分级视角
治理标准	董事会层面的企业范围模型风险视图；明确的角色和职责	在组织治理战略中纳入指导性FEAT原则；定义可衡量和可追踪的KPI和指标
人才和技术发展	业务单元层面的相关产品和技术技能	识别与FEAT原则应用相关的所需技能/新角色；构建和部署可信且公平设计的AIDA模型和系统

---

4. FEAT原则评估方法论摘要

4.1 公平性

作为概念，公平性对决策而言并不新颖，也不是AIDA驱动决策所独有的。一般意义上，公平涉及正确、公正和平等，没有偏袒或歧视。

AIDA系统由高质量数据、复杂算法和不断增长的处理能力驱动，为数据驱动的决策过程带来了规模和竞争价值。然而，如果没有仔细的设计和控制，这些系统可能带来新的风险和非预期的伤害，延续或强化社会中现有的不利因素，甚至引入新的不利因素。

越来越多的监管机构在指导金融机构防止伤害和提供利益方面发挥着关键作用。在新加坡，FEAT原则中明确与公平相关的原则（P1-P4）旨在确保AIDA驱动的决策不会在没有适当理由的情况下系统性地使个人或群体处于不利地位。

第二阶段对公平评估方法论的主要扩展包括：

- 更详细地定义了个人属性、偏见类型和缓解方法、公平目标和指标等关键概念，并附有说明性框架和应用指导。

- 将第一阶段定义的A-E五部分公平评估方法论对齐到典型的AIDA系统开发生命周期。
- 为第一阶段的18个公平评估方法论问题提供了详细考量和指导。
- 将18个问题整合到整体FEAT检查清单中以实现端到端评估。
- 提供了关于跨金融机构规模化和嵌入公平原则评估方法论的指导。

第二阶段的用例是人寿保险的预测性承保。

4.2 伦理与问责

随着AIDA系统在金融领域使用的增加，监管机构、同行、消费者和整个社会对问责和道德行为的期望也在增加。伦理与问责方法论文件及配套工作手册可以帮助确保组织内一致且关注背景的道德决策。

虽然企业通常有一套核心价值观，有时还有AI原则，但在将这些价值观应用于决策方面进展甚少。然而，金融机构在管理重大和系统性风险方面拥有强大的治理基础，因此比其他行业的公司更有条件将价值观和原则应用于决策。

运作化伦理与问责框架提供了一套术语分类体系来组织道德决策中的重要概念，并推动组织从抽象观念走向具体且可衡量的行动承诺。这些术语包括：

- **价值观和原则**：组织最熟悉的概念。
- **规范性概念**（如公平）：用于阐明、塑造和落实承诺。
- **承诺**（commitments）：使组织可以被追究责任的具体政策。
- **规格**（specifications）：使承诺可衡量的具体指标。

4.3 透明度

随着金融服务行业中越来越多的决策由算法塑造或做出，人们对关键问题的透明度需求更大（例如：算法在哪里使用？使用什么数据？如何做出决策？有什么影响？等）。

全球监管机构正鼓励金融机构提高AIDA决策对受影响者的透明度，包括MAS 2018年发布的FEAT指南中的三项原则（P12、P13、P14）。

金融机构可以仅从监管合规角度看待透明度。然而，在日益数字化的世界中，透明度是一种未满足的客户需求，提供更高的透明度可以帮助金融机构建立更强大的客户关系和竞争优势。

金融机构的AIDA透明度方法必须解决两个方面：

- **外部透明度**：在各种情况下应向客户/潜在客户和其他外部利益相关方提供什么水平和形式的透明度？
- **内部透明度**：金融机构应如何构建可靠的解释以支持外部信息共享，并符合内部政策和监管要求？

关于外部透明度，AIDA驱动决策对受影响客户的重要性（如是否导致被拒绝获得金融服务）必须与防止欺诈或恶意利用等考量进行权衡。

还需要考虑实际运作方面：人类解读复杂的算法决策解释的能力是有限的。有时，强调他们可以采取的改变结果的行动可能比说明决策原因更有价值。

关于内部透明度，近年来在设计AIDA系统解释方法方面活动频繁，特别是关于机器学习模型。金融机构应审查现有方法论并考虑实施内部批准的解释方法标准。

透明度方法论文件提出了五项建议：将AIDA透明度考量嵌入现有风险框架；投资于主动沟通；培训客户、员工、管理层和风险/控制职能；投资于AIDA透明度工具；在整个AIDA生命周期中整合透明度考量。

---

5. 负责任AI监管的全球监管格局

5.1 全球监管格局

全球各监管机构正在制定框架来应对AIDA系统带来的独特机遇和挑战：

新加坡： - 人工智能治理模型框架 - 组织实施和自我评估指南 - 用例汇编

香港： - 人工智能高级原则 - 授权机构使用大数据分析和AI的消费者保护

美国： - 五家监管机构就金融机构使用AI和ML发出信息和意见征询 - 美国保险监督官协会AI指导原则

英国： - 英格兰银行金融未来报告 - 人工智能公私论坛（AIPPF）

欧盟： - 欧盟委员会提出了首个AI监管框架：《关于欧洲人工智能方法的法规提案》

5.2 关键监管趋势

监管格局的总体趋势是关注澄清和修订现行法律和标准，使其能够更清楚地适用于涉及AIDA系统的情况，而不是建立新的监管机构和法律。

亚洲的监管机构发布了关于AI采用的非约束性指南和原则，支持基于原则的、技术中立的负责任AI使用方法。这一监管方法在新加坡和香港最为明显，与亚洲以外的英国等司法管辖区采取的方法类似。

5.3 评估金融服务中AIDA项目的监管影响

不同的法律和监管要求适用于金融机构使用AIDA的不同背景。关键考量因素包括：

1. 采用基于重要性的方法评估监管考量。
2. 利用现有的内部风险管理和治理框架。
3. 与技术合作伙伴合作以促进透明度和问责。
4. 引入多元化视角讨论负责任AI相关监管问题。
5. 尽可能分享用例和示例以增强行业理解。

5.4 采用基于重要性的方法评估监管考量

MAS在其FEAT原则中支持了基于重要性的方法，鼓励公司根据AIDA驱动决策的重要性校准其内部治理框架下的行动和要求。

5.5 利用现有风险管理和治理框架

现有的风险管理和治理框架是金融机构实施负责任AI措施的有用起点。金融机构已在高度监管的行业中运营，大多数司法管辖区已有涵盖外包技术使用、风险管理和治理的稳健监管要求。

以新加坡为例，关键的监管考量涉及三个阶段：

阶段	监管考量	管辖示例
阶段1：数据输入AI技术	数据保护法、保密法、外包指南	新加坡《个人数据保护法》的个人数据框架
阶段2：AI处理数据		MAS发布的多项风险管理实践指南

阶段	监管考量	管辖示例
	技术法、网络安全法、风险管理实践	
阶段3：AI用例生成的输出	消费者保护法、反歧视法	新加坡《数字顾问服务提供指南》

5.6 与技术合作伙伴合作

金融机构部署AIDA系统与科技公司提供AI解决方案之间存在互补角色。科技公司可以与客户分享技术和非技术信息，帮助其负责任地部署AI产品。

5.7 确保视角的多样性

在讨论监管问题时，多元化的视角至关重要。多样性可以来自与一系列利益相关方的合作，也可以来自组织内不同团队的参与。团队内部的多样性（性别、族群、残障、种族等）也同样重要。

5.8 与行业分享用例

鉴于触发监管影响的主要是AIDA系统使用的背景，分享用例和示例可以增强行业对适用监管问题的理解。

5.9 Veritas的前进方向

随着亚洲金融机构AIDA系统部署的快速增加，实施MAS　FEAT原则等监管指导提供了负责任AI标准的基线，帮助金融机构在部署AI解决方案时提出正确的问题。

指导原则应承认背景在将高层原则转化为实践中的重要性。每项原则在不同背景下都有不同的含义，重要性评估对于确定哪项原则相关以及如何缓解风险至关重要。

---

6. 致谢

姓名	组织	角色
Sopnendu Mohanty	新加坡金融管理局	项目赞助人
Li XuChun	新加坡金融管理局	项目总监
Zhang Qiang	新加坡金融管理局	项目负责人
Bhavna Rawlley	Accenture	项目经理
Rupini Pandian	Accenture	数据科学家
Shen Kai	Accenture	数据科学家
Marcus Bartley Johns	Microsoft	主题专家
Dennae Smith	Microsoft	主题专家
Nick Lewins	Microsoft	主题专家
Dave Dadoun	Microsoft	主题专家

同时感谢Accenture的James Hwa Jaan Gan和Joon Seong Lee对本项目的支持和贡献。

附录A：FEAT检查清单

FEAT原则评估方法论检查清单

如前所述，此检查清单是可选的，仅在金融机构选择采取步骤遵守FEAT时才相关。应认识到使AIDA系统公平、道德、可问责和透明不是一项简单的任务。因此，一定数量的检查项目是不可避免的。

但是，顶层检查清单问题下的子问题应被视为帮助回答这些问题的考量因素。金融机构可能希望通过顶层检查清单项目而非更详细的考量因素来追踪遵守情况。

基础设置检查清单

编号	问题
G1	金融机构是否定义了AIDA?
G2	是否有框架来定义AIDA项目/用例的角色和职责?
G3	是否有重要性框架？是否有明确定义来确定不同重要性水平?
G4	金融机构是否有追踪所有重要AIDA用例的最新清单?
T1	金融机构是否定义了用于确定特定AIDA用例是否需要外部（面向客户的）透明度的因素?
T2	金融机构是否定义了用于确定个别AIDA用例内部透明度范围和受众的因素?
T3	（如金融机构选择提供外部透明度）在客户生命周期的每个阶段，金融机构是否确定了所需的主动或被动沟通?
T4	金融机构是否为其内部使用定义了可接受的AIDA ML解释方法集?
T5	金融机构是否为此类解释方法设定了最低准确性标准?
F0	金融机构是否定义了标准、稳健的流程来：(a) 识别高风险群体？(b) 识别个人属性和潜在代理？(c) 识别明确的公平目标及相关措施和阈值？(d) 识别和减少不公平的算法方法?
EA1	组织价值观是否已定义和描述?
EA2	是否定义了AI道德使用的组织或特定群体原则?
EA3	是否识别和描述了与背景相关的核心概念?
EA4	员工是否接受了基于价值观的决策培训?

各AIDA系统的FEAT原则评估检查清单

步骤1——定义系统背景和设计：

编号	问题
G5	是否记录了系统的商业目标和量化措施?
G6	是否有提议系统的范围边界?

编号	问题
G7	AIDA的使用是否有正当理由?
G8	是否确定了用例的重要性?
EA5	每个用例是否有相关价值观、核心概念、原则、承诺和规格的声明?
EA6	是否记录了每项承诺的相对优先级?
F1	是否识别和记录了可能被系统系统性不利对待的个人和群体?
F2	是否识别和记录了系统运行可能产生的对F1中个人和群体的潜在伤害和利益?
F3	是否识别和记录了系统的公平目标和相关公平指标?
T6	用例团队是否确定了外部透明度的需求?
T7	（如是）团队是否确定了客户生命周期各阶段的主动和被动沟通需求?
T8	团队是否确定了所需的内部透明度水平和受众?
T9	团队是否从T4的批准列表中为此用例选择了合适的解释方法?
T10	团队是否确认所选解释方法满足最低准确性要求?

步骤2——准备和输入数据：

编号	问题
G9	是否记录了系统使用的数据？是否进行了DPIA?
G10	是否记录了数据预处理和工程?
F4	是否记录了数据中可能影响系统公平性的关键错误、偏见或属性?
F5	是否记录了如何缓解这些影响?
F6	是否评估和记录了系统针对公平目标的量化性能估计?
F7	是否评估和记录了系统公平目标与其他目标之间可实现的权衡?
F8	是否论证和记录了为何系统观察到的公平结果优于这些替代权衡?

步骤3——构建和验证：

编号	问题
G11	是否定义了AIDA系统的组成？是否记录了性能估计和不确定性?
T11	是否按照T8-T10的要求实施了内部透明度仪表板/报告?
T12	相关的一线和二线控制团队是否审查和批准了这些输出?
T13	解释未满足预期时是否采取了适当的缓解措施?
T14	是否按照T6-T7的外部透明度要求开发和测试了系统功能?
F9	是否评估和记录了系统针对公平目标的量化性能估计?
F10	是否评估和记录了公平目标与其他目标之间的权衡?
F11	是否论证了为何观察到的公平结果优于替代权衡?

步骤4——部署和监控：

编号	问题
G12	系统的监控和审查机制是否设计用于检测异常运行?
G13	是否有后备和/或缓解计划?
T15	客户服务和投诉处理等运营流程是否已适当修改以纳入AIDA客户透明度?
T16	客户/网站条款和条件是否已适当更新?
T17	AIDA系统实施是否支持在"上线"后持续提供内部和外部解释?
EA7	数据主体的救济机制是否可用？使用情况是否过高或过低?
EA8	是否有控制措施来重新审视和校准承诺及其规格?
F12	系统的监控和审查机制是否确保系统影响与业务和公平目标一致?

附录B：术语表

术语	定义
问责性（Accountability）	对一组特定结果（如行为、行动、产品、服务或决策）负责的状态。
内部问责	组织治理系统追究个人或群体对非预期结果责任的能力。
外部问责	政府、监管机构、个人和其他机构追究组织对其运营结果责任的能力。
准确性（Accuracy）	评估模型质量的指标，通常计算为正确或不正确预测占所有考虑案例的百分比。
AIDA	人工智能或数据分析，定义为辅助或替代人类决策的技术。
AIDA系统	包含一个或多个协同工作的AIDA模型以实现一组目标的系统。
精算公平（Actuarial fairness）	量化公平的一个示例。每份保险合同应贡献其全部生产成本和公平分配的共同成本。
行为公平（Behavioural fairness）	基于个人控制范围内的行为选择提供奖励或施加处罚。
偏见（Bias）	有利于或不利于某一想法或事物的不成比例权重。AIDA系统中的偏见可以是训练数据或系统本身的功能。
合规性（Compliance）	遵守适用法律、法规、行为准则或其他法律和道德规范的程度。
混淆矩阵（Confusion matrix）	比较真实结果与预测结果的表格，用于描述分类模型的性能。
数据主体（Data subject）	可识别的与特定数据项相关的活人。
漂移偏见（Drift bias）	因个人情况或一般人群特征随时间变化而发生的数据偏见。
伦理（Ethics）	为满足一组价值观、按照原则、支持治理而执行的工作。
可解释性（Explainability）	对AIDA模型和/或人类过程做出的决策有普通成年人可理解的解释的能力。
公平性（Fairness）	正确、公正和平等，没有偏袒或歧视。公平是一项旨在防止伤害不均匀分配的指导原则。

术语	定义
特征工程（Feature engineering）	基于领域/行业知识将原始数据转换为可用作预测模型输入的特征的过程。
治理（Governance）	设计用于评估特定情况是否满足规范的结构、过程和机制。
历史偏见（Historical bias）	当先前决策中的偏见反映在训练数据中时发生。
可解读性（Interpretability）	训练有素的专业人员能够解释模型如何得出结论的程度。
正义（Justice）	以人的平等价值和政治地位为基础。人们获得他们应得的东西。
测量偏见（Measurement bias）	数据收集中的系统性或非随机错误。
模型（Model）	在数据科学背景下，任何AIDA算法和算法过程。
模型验证（Model validation）	在测试/保留数据集上评估模型性能的过程。
规范性内容（Normative content）	关于"应该"或"应当"做什么的指导。
目标（Objective）	公司为实现预期结果将采取的具体步骤和约束。
个人属性（Personal attribute）	不应在没有合理理由的情况下用作决策基础的特征。
原则（Principles）	描述如何实施价值观并为运作化价值观设定护栏的规范。
预处理偏见（Preprocessing bias）	模型开发中数据预处理操作导致的偏见。
代理偏见（Proxy bias）	使用代理属性估计目标变量时发生的偏见。
救济（Recourse）	数据主体在AIDA系统产生异常或不当结果时寻求帮助的能力。
纠正（Redress）	纠正AIDA系统异常或不当结果的行为。
可复现性（Reproducibility）	从头开始复现模型及其结果的行为。
代表性偏见（Representation bias）	群体在数据中代表性不足导致的偏见。
信任（Trust）	相信另一方的目标与你一致、有能力实现目标、并将努力实现目标。

术语	定义
透明度（Transparency）	对系统或组织采取的行动及其决策过程的可见性。
价值观（Values）	描述组织关心和想要促进和保护的理想集合。

附录C：全球AI监管格局

关键监管发展

全球各司法管辖区——包括新加坡、香港、澳大利亚、英国和欧盟——在负责任使用AI方面都有监管发展。

新加坡：MAS于2018年11月发布了14项FEAT原则。2020年又发布了人工智能治理模型框架、组织实施和自我评估指南以及用例汇编。这些指南共同促进了公众理解和信任。

香港：2020年8月发布的报告旨在促进对AI在银行业采用的广泛影响的理解。

英国：2019年6月，金融行为监管局（FCA）与英格兰银行合作发布了《金融未来报告》。五家金融监管机构最近发出了信息征询。

美国：2020年8月，美国保险监督官协会发布了AI指导原则。五家监管机构就金融机构使用AI和ML发出了信息和意见征询。

欧盟：2021年4月，欧盟委员会提出了首个AI监管框架，采用比例和基于风险的方法，将AI使用按风险级别分类（最低风险、有限风险、高风险和不可接受的风险）。

关键监管趋势

亚洲最显著的趋势是发布非约束性指南和原则，支持基于原则的、技术中立的方法：

- **新加坡**：FEAT原则及AI治理模型框架提供了可随时采用的实际建议。
- **香港**：2019年11月发布了高级别AI原则通告，银行可根据其应用的性质和风险水平按比例适用原则。
- **英国**：2019年6月发布了《金融未来报告》，鼓励制定数据标准和协议以促进机器学习和AI的负责任使用。

评估金融服务中AI项目的监管影响

五项关键考量：

1. **采用基于重要性的方法**：帮助金融机构确定所需治理和控制的比例应用。MAS的FEAT原则和欧盟的AI法规提案均支持此方法。
2. **利用现有监管和治理框架**：现有的风险管理框架是有用的起点，大多数监管要求是技术中立的。
3. **与技术合作伙伴合作**：科技公司可以分享关于AI软件能力和局限性的信息，帮助金融机构负责任地部署AI产品。
4. **确保视角多样性**：包括与政府、监管机构、行业协会、技术开发商等一系列利益相关方的合作。
5. **与行业分享用例**：用例分享可以增强行业对适用监管问题的理解。

Veritas的前进方向

Veritas代表了一个支持增加对AI部署问题认识的重要平台。基于迄今为止的工作，在AI实践和评估AI部署重要性的标准方面有明显的更大一致性潜力。未来的Veritas工作可以致力于制定金融机构在评估AI部署重要性时可以借鉴的指示性标准。展望未来，关键是建立在现有工作之上，鼓励更多用例分享以改进对金融机构如何实施负责任AI的理解。

Veritas联盟第二阶段成员名单

（联盟共27个成员机构参与）

法律声明

本报告由MAS、AXA、HSBC、Microsoft、Standard Chartered、Swiss Re、UOB、Accenture和TruEra编制和发布。