人工智能治理模型框架（第二版）

新加坡个人数据保护委员会 (Personal Data Protection Commission, PDPC)

2020年1月

中文翻译版 · 仅供参考，以英文原文为准

原文来源：https://www.pdpc.gov.sg/-/media/files/pdpc/pdf-files/resource-for-organisation/ai/sgmodelaigovframework2.pdf

---

目录

更新摘要

版本	发布日期	摘要
第一版	2019年1月23日	在瑞士达沃斯举行的2019年世界经济论坛年会上发布了《人工智能治理模型框架》（第一版）。
第二版	2020年1月21日	在瑞士达沃斯举行的2020年世界经济论坛年会上发布了《人工智能治理模型框架》（第二版）。

第二版的主要变更包括：

· 在每个章节中增加了行业案例，说明各机构如何实施该章节所述的AI治理实践；
· 更新了两个章节的标题以准确反映其内容：
· "确定AI决策模型"更名为"确定人工智能增强决策中的人类参与程度"；
· "客户关系管理"更名为"利益相关方互动与沟通"。

确定人工智能增强决策中的人类参与程度章节的具体变更：

· 阐明了"人类在环上方"（human-over-the-loop）方法，解释了人类在AI增强决策中的监督角色。
· 阐明了组织在确定涉及AI的决策过程中人类参与程度时，可以考虑其他因素，如伤害的性质和可逆性以及操作可行性。

运营管理章节的具体变更：

· 为组织提供了采用基于风险的方法实施措施的指导：
· 识别对利益相关方影响最大的特性或功能；
· 考虑哪种措施最能有效地建立利益相关方的信任。

- 提供了关于各项措施的必要性和相关性的指导：
- 阐明用于构建AI模型的数据集可能包含个人数据和非个人数据；
- 纳入了稳健性（robustness）、可复现性（reproducibility）和可审计性（auditability）等新措施，并提供了有益实践示例。

利益相关方互动与沟通章节的具体变更：

- 强调了与各类内部和外部利益相关方沟通的重要性。
- 强调了在与各类利益相关方互动时需要考虑目的和背景。
- 提供了与各类利益相关方互动时应提供的信息水平建议。

附录A——现有AI伦理原则汇编：

- 阐明所提供的AI伦理原则列表是现有AI原则的汇编，仅供参考。并非所有列出的原则都在模型AI治理框架中得到了解决。组织可以考虑将附录A中的其他原则纳入自身的企业原则。

附录B——算法审计：

- 阐明只有在需要发现模型中算法的实际运作情况时，且仅应监管机构要求（作为法证调查的一部分），才应进行算法审计。

附录C——用例：

- 附录C已删除。取而代之的是单独发布了一份用例汇编（go.gov.sg/ai-gov-use-cases）。

前言

2019年，世界见证了人工智能（"AI"）在复杂性和普及性方面的重大进展。例如，我们目睹了下一代AI驱动的自然文本生成器如GPT-2的出现，它能够生成难以与人类写作区分的段落。我们还看到了Dactyl（一种机器人手）的开发，它使用强化学习（reinforcement learning）以类似人类的灵巧度抓取和操作常见家用物品。这些例子证明了AI进步的速度及其将如何在我们日常生活中变得无处不在。

AI伦理和治理方面的讨论也在推进。在过去两年中，各国政府和国际组织开始发布关于AI伦理和治理的原则、框架和建议。2019年1月，新加坡在达沃斯世界经济论坛上发布了《人工智能治理模型框架》（"模型框架"）。模型框架对全球AI伦理讨论的独特贡献在于将伦理原则转化为组织可

以随时采用的实际建议，以负责任地部署AI。我们对采用模型框架中所述实践的组织的多样性感到鼓舞，这印证了其易用性和相关性。

新加坡自豪地推出了模型框架的第二版。本版纳入了已采用AI的组织的经验，以及我们参与欧盟委员会高级别专家组和经合组织（OECD）AI专家组等领先国际平台所获得的反馈。这些意见使我们能够为组织负责任地实施AI提供更清晰有效的指导。

新加坡的资讯通信媒体发展局（Info-communications Media Development Authority, "IMDA"）和个人数据保护委员会（Personal Data Protection Commission, "PDPC"）还与世界经济论坛第四次工业革命中心合作，制定了《组织实施和自我评估指南》（Implementation and Self-Assessment Guide for Organisations, "ISAGO"）。ISAGO通过允许组织评估其AI治理实践与模型框架的一致性来补充模型框架，同时提供有用的行业案例和实践。

我们还发布了一份《用例汇编》，其中展示了组织如何实施或使其AI治理实践与模型框架保持一致的真实案例。这些举措共同使任何组织都能以具体和实际的方式建立和完善其AI治理实践。

这些举措在新加坡国家AI战略中发挥着关键作用。它们体现了我们发展以人为本的AI治理方法以建立和维持公众信任的计划。它们也反映了我们以协作和包容的方式共同创建AI生态系统的重点。模型框架和ISAGO将为未来的发展铺平道路，例如关于AI伦理部署的专业人员培训，以及为新加坡乃至全世界更好地应对AI对社会的影响奠定基础。

我们今天采取的步骤将在我们的集体未来上留下不可磨灭的印记。模型框架已被公认为负责任使用AI及其未来发展的坚实基础。我们将以此势头推进以人为本的AI方法——促进创新并保障公众信任——以确保AI对世界的积极影响延续至未来世代。

S Iswaran 新加坡通讯及新闻部部长 2020年1月

---

1. 序言

1.1 模型框架主要关注四个广泛领域：内部治理结构和措施、AI增强决策中的人类参与、运营管理以及利益相关方互动和沟通。

虽然模型框架的雄心确实不受限制，但其在形式、目的和范围的实际考量方面终究是有限的。鉴于此，有几点需要说明。本模型框架——

a. **与算法无关**（Algorithm-agnostic）：不关注特定的AI或数据分析方法。它适用于AI的设计、应用和使用的一般情况。

b. **与技术无关**（Technology-agnostic）：不关注特定的系统、软件或技术，无论开发语言和数据存储方法如何都适用。

c. **与行业无关**（Sector-agnostic）：作为任何行业组织可采用的基准考量和措施集。特定行业或组织可选择纳入额外的考量和措施或调整此基准集以满足其需求。PDPC鼓励并将与各公共机构合作，为其所在行业调整模型框架。

d. **与规模和商业模式无关**（Scale- and Business-model-agnostic）：不关注特定规模或大小的组织。它也可以被从事企业对企业或企业对消费者活动和运营的组织使用，或应用于任何其他商业模式。

1.2　公认有许多问题与AI的伦理使用和部署密切相关。本模型框架不关注这些特定问题，这些问题的范围通常足以值得单独研究和处理。这些问题的示例包括：

a. 为AI制定一套新的伦理原则。全球已有多项尝试建立一套通用原则。虽然一套一致的核心伦理原则正在形成，但在文化、司法管辖区和行业之间也存在差异。模型框架使用现有的、通用的AI伦理原则（汇编见附录A）并将其转化为可实施的实践。

b. 提供关于数据共享的模型框架和解决相关问题，无论是公共和私营部门之间、组织之间还是联盟内部。已有许多相关指南，例如IMDA的《可信数据共享框架》和《数据共享数据估值指南》。

c. 讨论与AI相关的法律责任、知识产权以及AI的社会影响（如就业、竞争、社会不同群体对AI产品和服务的不平等获取、AI技术落入不法之手等）等问题。这些问题仍然是重要的，可以通过新加坡管理大学法学院设立的AI和数据治理中心等平台另行探讨。

---

2. 引言

目标

2.1　数据和计算能力的指数级增长推动了AI等数据驱动技术的进步。AI可以被组织用于提供新的商品和服务、提高生产力、增强竞争力，最终实现经济增长和更高的生活质量。然而，与任何新技术一样，AI也带来了新的伦理、法律和治理挑战。这些挑战包括可能导致不公平结果的非预期歧视风险，以及消费者了解AI如何参与对其做出重大或敏感决策的相关问题。

2.2 PDPC在咨询委员会的建议下，提出了本模型框架的第二版，作为一个通用的、即用型工具，使大规模部署AI解决方案的组织能够以负责任的方式这样做。本模型框架不针对那些部署恰好在其功能集中纳入了AI的更新版商业现成软件包的组织。

> 注：根据新加坡《个人数据保护法》（2012年）第5条，IMDA被指定为PDPC。

2.3 本自愿性模型框架就需要考虑的关键问题和可以实施的措施提供指导。采用本模型框架需要根据实施组织识别的风险来调整措施。模型框架旨在帮助组织实现以下目标：

a. 通过组织负责任地使用AI来管理AI部署中的不同风险，从而建立利益相关方对AI的信心。

b. 展示合理的努力，使内部政策、结构和流程与数据管理和保护中的相关问责制实践保持一致（例如《个人数据保护法》（2012年）（"PDPA"）和经合组织隐私原则）。

2.4 为协助组织实施模型框架，PDPC还准备了配套的ISAGO。ISAGO帮助组织评估其AI治理实践和流程与模型框架的一致性。它还提供了额外有用的行业参考和案例，进一步阐明了本模型框架中提出的建议。

2.5 组织采纳本模型框架中建议的程度取决于多项因素，包括组织使用的AI的性质和复杂性、AI在组织决策中的应用程度，以及自主决策对个人影响的严重程度和可能性。

2.6 进一步说明：AI技术可以用于辅助人类决策者或自主做出决策。例如，医学诊断中自主决策的影响可以说比产品推荐中的影响更大。因此，AI部署的商业风险与对个人的影响成正比。一般而言，如果以合乎伦理的方式实施AI技术的成本超过预期收益，组织应考虑是否应采用替代的非AI解决方案。本框架中提出的考量和建议旨在指导已决定大规模部署AI技术的组织。

指导原则

2.7 模型框架基于两项高层指导原则，促进对AI的信任和对AI技术使用的理解：

a. **使用AI进行决策的组织应确保决策过程是可解释的、透明的和公平的**。

尽管完美的可解释性、透明度和公平性不可能实现，但组织应努力确保其对AI的使用或应用尽可能反映这些原则的目标。这有助于建立对AI的信任和信心。

b. **AI解决方案应以人为本**。

由于AI用于放大人类能力，人类利益的保护，包括其福祉和安全，应成为AI设计、开发和部署中的首要考量。

2.8 与其他技术一样，AI旨在提高人类生产力。然而，与早期技术不同的是，AI做出的某些自主预测或决策可能无法完全解释。由于AI技术可以做出影响个人或对社会、市场或经济产生重大影响的决策，组织应考虑使用本模型框架来指导其AI部署。

2.9 组织在其流程中或为增强其产品和/或服务而大规模部署AI时，应详细制定一套伦理原则。必要时，组织可参考附录A中的AI伦理原则汇编。组织还应尽可能审查其现有的企业价值观，并纳入其已阐明的伦理原则。某些伦理原则（如安全性）可以作为风险表述，纳入企业风险管理框架。模型框架旨在帮助组织将伦理原则纳入熟悉的、预先存在的企业治理结构，从而帮助指导组织中AI的采用。

假设

2.10 模型框架旨在讨论通用的良好数据管理实践。模型框架主要适用于机器学习模型（相对于纯决策树驱动的AI模型）。

2.11 模型框架不解决因对高度依赖AI的组织发动网络攻击而导致灾难性故障的风险。无论是否使用AI技术，组织仍有责任确保其产品和服务的可用性、可靠性、质量和安全性。

2.12 采用本自愿性模型框架不会免除组织遵守现行法律法规的义务。然而，由于这是一个基于问责制的框架，采用它将有助于组织证明其已在数据管理和保护方面实施了基于问责制的实践，例如PDPA和经合组织隐私原则。

2.13 此外，应注意某些行业（如金融、医疗保健和法律行业）可能受到与该行业相关的现有特定行业法律、法规或指南的监管。例如，新加坡金融管理局（Monetary Authority of Singapore）发布了《促进新加坡金融部门使用人工智能和数据分析的公平、伦理、问责和透明原则》（"FEAT原则"），为使用AI和数据分析提供金融产品和服务的公司提供指导。建议组织注意此类法律、法规和指南，因为采用模型框架并不意味着组织符合此类特定行业法律、法规或指南。

定义

2.14 以下简化图描述了模型框架中讨论的AI采用过程中的关键利益相关方。采用过程不区分企业对消费者（"B2C"）、企业对企业（"B2B"）和企业对企业对消费者（"B2B2C"）关系。

2.15 AI中使用的某些术语根据上下文和用途可能有不同的定义。本模型框架中使用的一些关键术语定义如下：

"AI"（人工智能）：指一组寻求模拟人类特征的技术，如知识、推理、问题解决、感知、学习和规划，并根据AI模型产生输出或决策（如预测、建议和/或分类）。AI技术依赖AI算法来生成模型。选择并部署最合适的模型到生产系统中。

"AI解决方案提供商"（AI Solution Providers）：开发AI解决方案或利用AI技术的应用系统的组织。这不仅包括消费者可以直接使用的商业现成产品、在线服务、移动应用程序和其他软件，还包括B2B2C应用程序（如出售给金融机构的AI驱动的欺诈检测软件）。它们还包括将AI驱动的功能集成到其产品中的设备和设备制造商，以及其解决方案不是独立产品而是要集成到最终产品中的组织。一些组织开发自己的AI解决方案，可以成为自己的解决方案提供商。

"组织"（Organisations）：在其运营中采用或部署AI解决方案的公司或其他实体，例如后台运营（如处理贷款申请）、前台服务（如电子商务门户或打车应用程序）或销售或分发提供AI驱动功能的设备（如智能家居设备）。

"个人"（Individuals）：根据上下文，可指组织打算向其提供AI产品和/或服务的人员，或已购买AI产品和/或服务的人员。这些人也可以被称为"消费者"或"客户"。

---

3. 人工智能治理模型框架

3.1 本模型框架包含组织应在以下关键领域采纳的促进AI负责任使用的措施指导：

a. **内部治理结构和措施**：调整现有或建立内部治理结构和措施，以纳入与算法决策相关的价值观、风险和责任。

b. **确定AI增强决策中的人类参与程度**：帮助组织设定AI使用风险偏好的方法论，即确定可接受的风险并识别AI增强决策中适当的人类参与水平。

c. **运营管理**：开发、选择和维护AI模型时需要考虑的问题，包括数据管理。

d. **利益相关方互动和沟通**：与组织利益相关方沟通的策略以及与其关系的管理。

3.2 采用本模型框架的组织可能会发现并非所有要素都适用。本模型框架旨在灵活，组织可以调整模型框架以适应其需求并采用那些相关的要素。

3.3 为帮助组织更好地理解模型框架，我们在每个章节中纳入了说明真实公司如何实施该特定章节所述某些实践的案例。此外，PDPC还发布了一份《用例汇编》，说明了各种本地和国际组织如何实施与模型框架所有章节一致的AI治理实践。

内部治理结构与措施

3.4　本节旨在指导组织制定适当的内部治理结构，使组织能够对AI技术如何引入其运营和/或产品和服务进行适当监督。

3.5　内部治理结构和措施有助于确保对组织使用AI进行稳健监督。组织现有的内部治理结构可以进行调整，必要时也可以实施新的结构。例如，与使用AI相关的风险可以在企业风险管理结构内进行管理，而伦理考量可以作为企业价值观引入，并通过伦理审查委员会或类似结构进行管理。

> 伦理考量可以作为企业价值观引入，并通过伦理审查委员会或类似结构进行管理。

3.6 组织还可以考虑确定其内部治理结构中的适当特征。例如，当完全依赖集中式治理机制不是最优选择时，可以考虑分散式机制，在必要时将伦理考量纳入运营层面的日常决策中。组织高层管理层和董事会在组织AI治理中的赞助、支持和参与至关重要。

3.7 组织可能希望考虑纳入与其内部治理结构发展相关的特征，例如：

1. 明确AI伦理部署的角色和职责

a. AI部署各阶段和活动的责任和监督应分配给适当的人员和/或部门。必要时且有可能的情况下，考虑建立一个协调机构，拥有相关专业知识和来自整个组织的适当代表。

b. 负有内部AI治理职能的人员和/或部门应充分了解其角色和职责，接受适当培训，并获得履行其职责所需的资源和指导。

c. 可分配的关键角色和职责包括：

i. 使用任何现有的风险管理框架并应用风险控制措施来：　- 评估和管理部署AI的风险，包括对个人的任何潜在不利影响（如谁最脆弱、如何受到影响、如何评估影响规模、如何从受影响者获取反馈等）。 - 决定AI增强决策中适当的人类参与水平。 - 管理AI模型训练和选择过程。

ii. 对已部署的AI模型进行维护、监控、文档记录和审查，以便在需要时采取补救措施。

iii. 审查与利益相关方的沟通渠道和互动，以提供披露和有效的反馈渠道。

iv. 确保处理AI系统的相关员工接受适当培训。在适用和必要的情况下，直接与AI模型工作和互动的员工可能需要接受培训，以解读AI模型的输出和决策并检测和管理数据中的偏见。其他工作涉及AI系统的员工（如回答客户关于AI系统查询的客户关系官员，或使用AI启用产品提出建议的销售人员）应至少接受培训，了解并对使用AI时的好处、风险和局限性保持敏感，以便他们知道何时向组织内的主题专家发出警报。

2. 风险管理和内部控制

a. 组织可以考虑实施一套健全的风险管理和内部控制体系，专门应对所选AI模型部署中涉及的风险。

b. 此类措施包括：

i. 使用合理的努力确保用于AI模型训练的数据集适合其预期目的，并评估和管理不准确或偏见的风险，以及审查模型训练期间识别的异常情况。几乎没有数据集是完全无偏见的。组织应努力了解数据集可能存在偏见的方式，并在其安全措施和部署策略中解决这一问题。

ii. 建立监控和报告系统及流程，确保适当级别的管理层了解已部署AI的性能和其他相关问题。在适当的情况下，监控可以包括自主监控，以有效扩展人类监督。AI系统可以设计为报告其预测的置信水平，可解释性功能可以关注为什么AI模型具有特定的置信水平。

iii. 确保涉及AI活动的关键人员变动时进行适当的知识转移。这将降低人员流动造成内部治理空白的风险。

iv. 当组织结构或涉及的关键人员发生重大变化时，审查内部治理结构和措施。

v. 定期审查内部治理结构和措施，确保其持续的相关性和有效性。

案例：CUJO AI的内部治理结构与措施

CUJO AI是一家在电信运营商市场运营的网络智能软件公司。总部位于美国，致力于开发和部署AI以改善家庭和企业联网设备的安全性、控制和隐私。

CUJO AI实施了明确的内部治理结构和措施，以确保对其AI使用的稳健监督。其多利益相关方治理结构促进在适当层级做出决策：

- **研究委员会**：由首席技术官、实验室负责人和首席数据科学家组成，负责批准AI开发和部署。四个技术团队角色明确：研究团队进行数据分析和开发ML模型；工程团队构建软件和云服务；运营团队部署AI模型；交付团队与运营商对接集成服务。
- **架构指导组（ASG）**：由首席技术官、首席架构官和首席工程师组成，确保AI/ML模型在部署前的稳健性。ASG每两周举行会议。
- **博士级员工**：监督AI开发和部署过程，并努力为每个新功能开发实施学术审查标准。

此外，CUJO AI为其员工制定了通用的道德准则。

案例：万事达卡（Mastercard）的内部治理结构与措施

万事达卡是全球支付行业的科技公司，其全球支付处理网络连接了210多个国家和地区的消费者、金融机构、商户、政府和企业。

为确保对AI使用的稳健监督，万事达卡成立了治理委员会（Governance Council），审查和批准被确定为高风险的AI应用实施。治理委员会由人工智能卓越中心执行副总裁主持，成员包括首席数据官、首席隐私官、首席信息安全官、数据科学家和业务团队代表。

- **首席数据官和首席隐私官**：审查AI实施方案，确保数据适合AI用途、AI用于道德目的、对个人的影响适当且潜在危害充分缓解。
- **首席信息安全官**：确保实施安全设计。
- **数据科学团队**：与数据办公室和隐私办公室保持持续对话，确保AI应用生命周期中的持续信息共享。

万事达卡还实施了风险管理和内部控制，包括初始风险评分、识别潜在缓解措施。高风险项目将提交治理委员会审查。

确定人工智能增强决策中的人类参与程度

3.8 本节旨在帮助组织确定AI增强决策中适当的人类监督程度。

3.9 明确使用AI的目标是确定人类监督程度的关键第一步。组织可以首先确定使用AI的商业目标（如确保决策一致性、提高运营效率和降低成本，或引入新产品功能以增加消费者选择）。然后将这些商业目标与在组织决策中使用AI的风险进行权衡。此评估应以组织的企业价值观为指导，而企业价值观又可以反映组织所在地区的社会规范或期望。

> 在部署AI解决方案之前，组织应确定其使用AI的商业目标，然后将其与在组织决策中使用AI的风险进行权衡。

3.10 在多个国家运营的组织在可能的情况下还应考虑社会规范、价值观和/或期望的差异。例如，游戏广告在一个国家可能是可接受的，但在另一个国家则不然。即使在一个国家内，风险也可能因AI部署地点的不同而有很大差异。

3.11 某些对个人的风险可能仅在群体层面显现。例如，股票推荐算法的广泛采用可能导致从众行为，如果足够多的个人同时做出类似决策，将增加整体市场波动性。

3.12 组织在权衡商业目标和使用AI风险时，理想的情况是以其企业价值观为指导。组织可以评估预期的AI部署和所选的算法决策模型是否与其自身的核心价值观一致。任何不一致和偏离都应是组织有意识的决策，并有明确定义和记录的理由。

3.13 由于识别商业目标、风险和确定AI增强决策中适当的人类参与程度是一个迭代和持续的过程，组织应持续识别和审查与其技术解决方案相关的风险、缓解这些风险，并维持一份应急计划以防缓解措施失败。通过定期审查的风险影响评估记录此过程有助于组织在使用AI解决方案时增强清晰度和信心。

AI增强决策中人类参与的三种广泛方法

3.14 基于上述风险管理方法，模型框架确定了在决策过程中对各种人类监督程度进行分类的三种广泛方法：

a. **人类在环中**（Human-in-the-loop）：人类监督是积极和参与式的，人类保持完全控制，AI仅提供建议或输入。如果没有人类的肯定行动（如人类命令继续执行给定决策），则无法执行决策。例如，医生可以使用AI来识别不熟悉的医学状况的可能诊断和治疗，但医生将做出最终的诊断和相应治疗决策。

b. **人类在环外**（Human-out-of-the-loop）：没有人类对决策执行的监督。AI系统拥有完全控制权，没有人类覆盖选项。例如，产品推荐解决方案可能基于预定的人口统计和行为画像自动向个人推荐产品和服务。

c. **人类在环上方**（Human-over-the-loop，或human-on-the-loop）：人类参与监督的程度在于人类处于监控或监督角色，能够在AI模型遇到意外或不希望的事件时（如模型故障）接管控制。这种方法允许人类在算法运行期间调整参数。例如，GPS导航系统规划从A点到B点的路线，提供多条可能路线供驾驶员选择，驾驶员可以在行程中更改参数而无需重新编程路线。

概率-严重程度矩阵

3.15 模型框架还提出了一个设计框架（以矩阵形式构建），帮助组织确定AI增强决策中所需的人类参与程度。此设计框架沿两个轴构建：(a) 概率；和 (b) 组织关于某个个人（或组织）做出的决策对该个人（或组织）造成伤害的严重程度。

3.16 "伤害"的定义以及概率和严重程度的计算将取决于上下文并因行业而异。

3.17 然而，该矩阵不应被理解为伤害概率和伤害严重程度是确定组织涉及AI的决策过程中人类监督程度时唯一需要考虑的因素。组织在各种情况下可能认为相关的其他因素还包括：(a) 伤害的性

质（即伤害是物理性质还是无形性质）；(b) 伤害的可逆性，以及作为其推论，个人获得救济的能力；以及 (c) 在决策过程中让人类参与在操作上是否可行或有意义。

3.18 对于安全关键系统，组织应审慎地确保允许人员接管控制，AI系统应提供充分的信息以使该人员做出有意义的决策，或在无法实现人类控制的情况下安全关闭系统。

概率-严重程度矩阵使用示例

一家在线零售商店希望使用AI全面自动化基于个人浏览行为和购买历史的食品推荐。

概率-严重程度评估：伤害的定义可以是做出不满足个人预期需求的产品推荐的影响。向个人做出错误产品推荐的伤害严重程度可能较低，因为个人最终决定是否购买。伤害的概率可能高或低，取决于AI解决方案的效率和效果。

决策过程中人类干预程度：鉴于伤害严重程度较低，评估指向不需要人类干预的方法（即人类在环外）。

定期审查：组织定期审查其方法以重新评估伤害的严重程度和概率，以及社会规范和价值观的演变。

案例：Suade Labs

Suade Labs（"Suade"）是一家全球运营的监管科技公司，也是世界经济论坛技术先锋。Suade提供AI驱动的解决方案，帮助金融机构处理大量细粒度数据并生成所需的监管数据、计算和报告。

在确定AI决策中的人类参与程度时，Suade考虑了：领域知识要求程度和监管不合规的成本。鉴于其解决方案需要人类专家的专业知识，且AI解决方案不正确建议导致的监管不合规成本显著，Suade采用了**人类在环中**的方法。在调优AI模型方面，Suade采用**人类在环上方**的方法，允许根据用户偏好调整模型。

案例：Grab

Grab是一家总部位于新加坡的公司，提供打车服务、食品配送和电子支付解决方案。Grab使用AI进行行程分配，考虑驾驶员偏好。

在确定人类参与程度时，Grab考虑了：每分钟需要完成超过5,000次行程分配的实时决策规模，以及AI模型在次优工作时对用户的严重程度和概率。Grab认为人类在如此短时间内完成大量行程分配在技术上不可行，且行程分配不够理想通常对生命几乎没有危害，因此采用了**人类在环外**的方法，同时持续审查AI模型。

运营管理

3.19 本节旨在帮助组织在其AI采用过程的运营方面采取负责任的措施。

3.20 模型框架使用以下通用的AI模型开发和部署流程来描述组织实施AI解决方案的各个阶段：

- **阶段1：数据准备** — 原始数据被格式化和清洗，以便准确得出结论。
- **阶段2：算法** — 在数据集上训练模型并应用算法，包括统计或机器学习模型。检验结果并迭代模型，直到产生最合适的模型。
- **阶段3：选定模型** — 选定的模型用于生成概率评分，可纳入应用程序以提供预测、做出决策、解决问题和触发行动。

模型开发数据

3.22 用于构建模型的数据集可能来自多个来源，可能包含个人数据和非个人数据。每个来源的数据质量和选择对AI解决方案的成功至关重要。如果模型使用有偏见、不准确或不具代表性的数据构建，模型产生非预期歧视性决策的风险将增加。

3.23 参与训练和选择部署模型的人员可能是内部员工或外部服务提供商。组织内负责数据质量、模型训练和模型选择的相关部门应共同制定良好的数据问责实践，包括：

a. **了解数据谱系**（Data lineage）：了解数据的原始来源、收集方式、在组织内的整理和移动方式，以及如何随时间保持其准确性。数据谱系可以可视化表示，追踪数据从来源到目的地的移动方式。三种类型的数据谱系：向后数据谱系、向前数据谱系和端到端数据谱系。保留数据来源记录可以让组织根据来源和后续转换确定数据质量、追踪潜在错误来源、更新数据并将数据归因于其来源。

b. **确保数据质量**：鼓励组织理解和解决可能影响数据质量的因素，如数据集的准确性、完整性、真实性、时效性、相关性、完整性、可用性和人工干预。

c. **最小化固有偏见**：组织应意识到其提供给AI系统的数据可能包含固有偏见，并鼓励采取措施减轻此类偏见。两种常见的数据偏见类型： - **选择偏见**（Selection bias）：当用于生成模型的数据不完全代表模型可能接收或运作的实际数据或环境时发生。包括遗漏偏见和刻板印象偏见。 - **测量偏见**（Measurement bias）：当数据收集设备导致数据系统性偏向某一特定方向时发生。

d. **不同的训练、测试和验证数据集**：模型使用训练数据训练，使用测试数据确定准确性，在适用的情况下通过在不同人口群体上测试检查系统性偏见。最后，使用验证数据集验证训练后的模型。

e. **定期审查和更新数据集**：数据集应定期审查以确保准确性、质量、时效性、相关性和可靠性。

3.24 即使仅使用非个人数据（包括已匿名化的个人数据）进行AI模型训练，上述良好数据问责实践仍然适用。

案例：Suade Labs的数据管理

Suade采用了多项良好数据问责实践：仅从相关监管机构获取和更新监管数据，为数据集添加元数据以实现可追溯性，使用大量标注人员降低标注偏见风险，开发标注系统以促进数据注释，使用验证模式检查确保数据模式准确表示源数据。

案例：pymetrics的偏见管理

pymetrics使用神经科学见解和经审计的AI模型帮助以更具预测性和更少偏见的方式评估求职者。为处理社会敏感特征并减轻固有偏见风险，pymetrics：使用基于已建立神经科学研究的客观数据；主动对所有AI模型进行去偏见处理，遵循"五分之四规则"等法律标准；在部署后持续测试AI模型决策的不利影响。

算法和模型

3.25 AI系统可能有多种通过AI模型中的算法启用的特性或功能。可解释性（explainability）、可重复性（repeatability）、稳健性（robustness）、定期调优、可复现性（reproducibility）、可追溯性（traceability）和可审计性（auditability）等措施可以增强AI模型中算法的透明度。

鼓励组织采用基于风险的方法进行双重评估。首先，识别对利益相关方影响最大的功能子集。其次，识别哪些措施最能有效建立利益相关方的信任。

可解释性（Explainability）

3.26 可解释性通过解释已部署的AI模型算法如何运作和/或决策过程如何纳入模型预测来实现。被解释的目的是建立理解和信任。

3.27 部署AI解决方案的组织建议采取以下做法：

a. 记录模型训练和选择过程的执行方式、决策理由和为应对已识别风险而采取的措施。 b. 将解决方案的设计和预期行为描述纳入产品或服务描述和系统技术规格文档。 c. 当AI系统从第三方获取或采购时，可考虑请求AI解决方案提供商的协助。 d. 辅助解释工具有助于解释AI模型，特别是不太可解释的模型（也称为"黑盒"系统）。

3.28 技术可解释性可能并不总是有启发性的。隐含的解释可能比明确的描述更有用。例如，向个人提供反事实（如"如果您的平均债务低15%，您将获得批准"）和/或比较（如"这些是与您有类似画像且收到类似决策的用户"）。

3.29 然而，在某些情况下，提供与算法相关的信息可能不切实际或不合理。例如，披露用于反洗钱检测、信息安全和欺诈防范的算法可能让不法分子逃避检测。

可重复性（Repeatability）

3.30 当可解释性在技术上无法实现时，组织可以考虑记录AI模型产生结果的可重复性。可重复性是指在相同场景下一致执行操作或做出决策的能力。有益实践包括：进行可重复性评估、执行反事实公平性测试、评估例外情况的处理方式、确保异常处理符合组织政策、识别和考虑随时间的变化。

稳健性（Robustness）

3.31 稳健性是指计算机系统应对执行过程中的错误和错误输入的能力。确保部署的模型足够稳健将有助于建立对AI系统的信任。

3.32 可以通过基于场景的前瞻性错误输入测试来测试稳健性。组织可以考虑与AI开发者合作对其模型进行对抗性测试（adversarial testing），以确保模型能处理更广泛的意外输入变量。

定期调优（Regular tuning）

3.34 建立内部政策和流程以执行定期模型调优对于确保部署的模型适应客户行为随时间的变化是有效的。

3.35 在可能的情况下，测试应反映计划的生产环境的动态性。一旦AI模型在真实环境中部署，建议进行积极的监控、审查和调优。

可追溯性（Traceability）

3.36 如果AI模型的 (a) 决策，以及 (b) 产生AI模型决策的数据集和过程以易于理解的方式记录，则认为该AI模型是可追溯的。

3.37 组织可以考虑的促进可追溯性的实践包括：构建审计跟踪、实施"黑匣子记录器"以捕获所有输入数据流、确保与可追溯性相关的数据得到适当存储。

可复现性（Reproducibility）

3.39 虽然可重复性是指组织内部结果的重复，但可复现性是指独立验证团队基于组织的文档使用相同的AI方法产生相同结果的能力。

3.40 有助于可复现性的实践包括：测试特定条件、实施验证方法、提供复制文件、向原始AI解决方案提供商确认结果可复现性。

可审计性（Auditability）

3.41 可审计性是指AI系统准备接受对其算法、数据和设计过程评估的程度。由内部或外部审计员对AI系统的评估可以增强AI系统的可信度。

3.42 组织可以采用基于风险的方法来确定哪些AI产品特性需要实施可审计性。

3.43-3.44 为促进可审计性，组织可以考虑保留数据来源、采购、预处理、谱系、存储和安全的全面记录，并将此类信息数字化集中在流程日志中。

案例：Symphony AyasdiAI的模型开发文档

Symphony AyasdiAI（"Ayasdi"）提供帮助其客户（主要在美国银行和金融领域）构建AI模型的解决方案。其模型加速器（AMA）首先识别相关变量，然后解释选择原因。整个模型创建过程自动记录，使客户能够确保初始特征和模型选择被记录和文档化。

利益相关方互动与沟通

3.45 本节旨在帮助组织在部署AI时采取适当步骤建立利益相关方关系策略中的信任。

一般性披露

3.46 鼓励组织提供关于是否在其产品和/或服务中使用AI的一般信息。这可以包括关于AI是什么、AI如何用于与消费者相关的决策、其好处、组织为什么决定使用AI、如何采取措施降低风险以及AI在决策过程中的角色和程度的信息。

3.47 组织可以考虑披露AI决策可能影响个人消费者的方式，以及决策是否可逆。

解释政策

3.48 鼓励组织制定关于向个人提供什么解释以及何时提供解释的政策。此类政策有助于确保沟通的一致性，并明确设定组织不同成员的角色和职责。等价原则可以提供一些指导，即人类驱动决策的同等披露标准适用于AI系统做出或增强的决策。

3.49 适当的互动和沟通能激发信任和信心。利益相关方关系策略不应保持静态。鼓励公司测试、评估和审查其策略的有效性。

3.50　不同的利益相关方有不同的信息需求。组织可以首先识别其受众（即外部和内部利益相关方）。外部利益相关方可能包括消费者、监管机构、合作企业和社会大众。内部利益相关方可能包括组织的董事会、管理层和员工。

与消费者互动

3.51 鼓励组织考虑消费者在与AI互动过程中的信息需求：

a. 确保消费者知道他们正在考虑的产品或服务是AI驱动的。 b. 提供信息以便消费者了解AI功能在正常使用中的预期行为。 c. 对于消费者定期互动的AI功能，提供信息使其了解AI功能为什么以某种方式运行，并在可能的情况下提供偏好设置。 d. 对于影响消费者的AI增强决策，考虑提供额外信息使其了解决策原因；对于某些类别的此类决策，提供适当的渠道对此类决策提出异议。

退出选项

3.52 组织在决定是否为个人提供退出AI产品或服务的选项时应仔细考虑。相关考量包括：对个人的风险/伤害程度、决策可逆性、替代决策机制的可用性、替代机制的成本或权衡、维护并行系统的复杂性和低效性以及技术可行性。

3.53 当组织权衡上述因素后决定不提供退出选项时，应考虑为消费者提供救济途径。

沟通渠道

3.54 鼓励组织为其客户设置以下沟通渠道：

a. **反馈渠道**：可用于客户提出反馈或查询，可由数据保护官（DPO）或质量服务经理（QSM）管理。 b. **决策审查渠道**：组织可以考虑为个人（如受影响的消费者）提供一个渠道，请求审查对其产生影响的重大AI决策。

用户界面测试

3.55 鼓励组织在部署前测试用户界面并解决可用性问题。如适用，还应告知个人其回复将用于训练AI系统。

易于理解的沟通

3.56 鼓励组织以易于理解的方式进行沟通。除文本沟通外，还可以考虑使用可视化工具、图形表示、摘要表格或其组合。

可接受使用政策

3.57 组织可以考虑制定可接受使用政策（AUP），确保用户不会恶意引入输入数据来操纵解决方案模型的性能和/或结果。

与其他组织互动

3.59 组织在与AI解决方案提供商或其他组织互动时，也可以应用上述部分方法和方法论。组织需要从AI解决方案提供商处获得足够的信息来帮助其实现业务目标。

3.60 组织可能需要考虑从AI解决方案提供商处获取的支持和详细信息的级别，包括：数据、模型训练和选择、人为因素、推断、算法存在以及缓解数据和算法偏见的措施和保障。

伦理评估

3.62 随着管理AI开发和使用的伦理标准的发展，鼓励组织评估其AI治理实践和流程是否与不断演进的AI标准一致，并向相关利益相关方提供此类评估的结果。

案例：Facebook的利益相关方互动与沟通

Facebook致力于对公众和用户在其运营和服务（包括使用AI）方面保持透明。具体做法包括：以易于理解的方式提供关于数据收集和使用的一般性披露；给予用户对其信息使用方式的易用且有意义的控制；发布解释政策的博客文章；推出AI教育计划。

Facebook为其News Feed功能实施了"为什么我看到这篇帖子？"功能，使用户了解其过去的互动如何影响帖子排名，并提供个性化快捷方式。

案例：MSD的利益相关方互动与沟通

MSD是一家跨国制药公司，部署了内部聊天机器人Jennie来回答IT相关问题。在部署前，用户体验团队测试了人机界面。三大原则指导开发和部署：理解用户心智模型、以人为本的方法、管理机器人到人类的交接（最多三次尝试后转交给客服人员）。MSD在登陆页面上披露Jennie是AI驱动的。

结论

3.63 本模型框架绝非完整或详尽，仍是一份开放接受反馈的文件。随着AI技术的演进，相关的伦理和治理问题也将演变。PDPC的目标是根据收到的反馈定期更新本模型框架，以确保它对部署AI解决方案的组织继续保持相关性和实用性。

> 关于AI使用的适当沟通能激发信任，因为它在组织和个人之间建立和维持开放的关系。

附录A：现有AI伦理原则汇编（仅供参考）

本附录包含从各种来源提炼的基础性AI伦理原则汇编。并非所有原则都在模型框架中纳入或涉及。组织可以考虑在相关和需要的情况下将这些原则纳入自身的企业原则。

1. **问责性（Accountability）**：确保AI参与者根据其角色、背景和技术水平，对AI系统的正常运作和对AI伦理与原则的尊重负责。
2. **准确性（Accuracy）**：识别、记录和阐明整个算法及其数据来源中的错误和不确定性来源，以便理解预期和最坏情况的影响并为缓解程序提供信息。
3. **可审计性（Auditability）**：使感兴趣的第三方能够通过信息披露来探究、理解和审查算法的行为。
4. **可解释性（Explainability）**：确保自动和算法决策以及驱动这些决策的任何相关数据可以用非技术术语向最终用户和其他利益相关方解释。
5. **公平性（Fairness）**： a. 确保算法决策不会在不同的人口统计线上产生歧视性或不公正的影响。 b. 在实施决策系统时开发和包括监控和核算机制，以避免非故意歧视。 c. 在开发系统、应用和算法时咨询多元化的声音和人群。
6. **以人为本和福祉（Human Centricity and Well-being）**： a. 追求数据实践利益的公平分配。 b. 追求从数据使用中创造尽可能大的利益。 c. 从事鼓励促进人类繁荣、尊严和自主的美德实践的数据实践。 d. 重视受数据实践影响的人员或社区的审慎判断。 e. 做出不会对个人造

成可预见伤害的决策。 f. 允许用户保持对正在使用的数据及其使用背景的控制。 g. 确保用户的整体福祉是AI系统功能的核心。

7. **人权对齐（Human rights alignment）**：确保技术的设计、开发和实施不侵犯国际公认的人权。
8. **包容性（Inclusivity）**：确保AI对所有人可及。
9. **进步性（Progressiveness）**：支持创造的价值远优于不参与该项目的实施。
10. **责任、问责和透明（Responsibility, accountability and transparency）**： a. 确保设计者和操作者对其系统负责。 b. 提供外部可见和公正的补救渠道。 c. 纳入下游措施和流程供用户或利益相关方验证。 d. 保留设计过程和决策的详细记录。
11. **稳健性和安全性（Robustness and Security）**：AI系统应安全可靠，不易被篡改或使训练数据受损。
12. **可持续性（Sustainability）**：支持能在合理时期内有效预测未来行为并产生有益见解的实施。

附录B：算法审计

1. 当需要发现模型中算法的实际运作情况时，进行算法审计。这需要应监管机构的要求（作为法证调查的一部分）进行。进行算法审计需要技术专业知识，可能需要聘请外部专家。审计报告可能超出大多数个人和组织的理解。进行算法审计的费用和时间应与审计报告的预期收益进行权衡。
2. 进行算法审计时可能相关的因素包括：

a. **进行算法审计的目的**：模型框架促进提供关于AI模型如何运作的信息作为可解释AI的一部分。在进行算法审计前，应考虑已向个人、其他组织或监管机构提供的信息是否充分和可信。

b. **审计结果的目标受众**：不同受众所需的信息不同。当受众是个人时，提供关于决策过程的信息将更有效地实现可解释AI的目标。当受众是监管机构时，应首先审查关于数据问责和算法运作的信息。

c. **一般数据问责**：组织可以提供关于如何在组织内实现一般数据问责的信息。

d. **商业价值**：AI模型中的算法可能是具有商业价值的信息。如果考虑进行技术审计，也应考虑相应的缓解措施（如保密协议）。

致谢

PDPC对以下个人和组织对模型框架的宝贵反馈表示诚挚感谢（按字母顺序）：

A*STAR、Accenture、AIG Asia Pacific Insurance、Apple、Asia Cloud Computing Association、AsiaDPO、BSA | The Software Alliance、Cambrian AI、CUJO.AI、Data Synergies、DBS、Element AI、Emerging Technologies Policy Forum、Facebook、Fountain Court Chambers、Google、Grab、Great Eastern、GSK、IBM Asia Pacific、LawTech.Asia、Mastercard、Microsoft Asia、MSD International GmBH（新加坡分公司）、Non-Profit Working Group on AI、OCBC Bank、PwC、pymetrics、Salesforce、Singtel、Standard Chartered Bank、Suade Labs、Symphony AyasdiAI、Telenor Group、Temasek International、Tookitaki、UCARE.AI、Untangle AI
