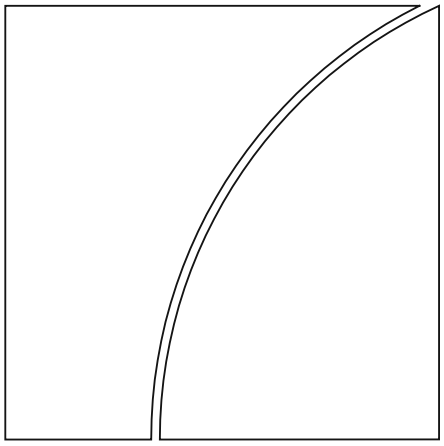
Financial Stability Institute | BIS



临时文件

24号

管理解释：监管机构如何应对人工智能可解释性问题

由Fernando Pérez-Cruz，Jermy Prenio，Fernando Restoy，Jeffery Yong

JEL分类：C60，G29，G38，O30

关键词：人工智能，机器学习，模型风险管理，风险治理

fsi临时文件旨在就与金融业及其监管和监督相关的广泛主题，为国际讨论做出贡献。其中表达的观点完全是作者的观点，并不一定反映国际清算银行或巴塞尔-based的标准制定机构的观点。

本出版物可在国际清算银行网站（www.bis.org）上获取。如需联系国际清算银行全球媒体与公共关系团队，请发送电子邮件至media@bis.org。您可以在www.bis.org/emailalerts.htm上注册邮件提醒。

ISSN 1020-9999 (在线)

摘要

金融机构越来越多地采用人工智能（AI）正在改变其运营、风险管理和客户互动方式。然而，复杂AI模型的有限可解释性，特别是在用于关键业务应用时，对金融机构和监管机构构成了重大挑战和问题。可解释性，即模型输出向人类解释的程度，对于透明度、问责制、监管合规性和消费者信任至关重要。然而，深度学习和大语言模型（LLM）等复杂AI模型通常难以解释。虽然现有的可解释性技术可以帮助揭示复杂AI模型的行为，但这些技术存在显著局限性，包括不准确性、不稳定性以及对误导性解释的易感性。

有限的模型可解释性使得管理模型风险具有挑战性。国际标准制定机构已发布——主要是高级别——模型风险管理（MRM）要求。然而，只有少数国家金融监管机构发布了具体的指导，而且它们往往侧重于用于监管目的的模型。其中许多现有指南可能并未针对先进的AI模型进行制定，并未明确提及模型可解释性的概念。相反，该概念体现在与治理、模型开发、文档记录、验证、部署、监控和独立审查相关的条款中。对于复杂的AI模型来说，遵守这些条款将具有挑战性。使用第三方AI模型将加剧这些挑战。

随着金融机构将人工智能模型应用于其关键业务领域，金融监管机构有必要寻求在人工智能背景下相关健全的模型风险管理与模型输出（MRM）实践。最终，可能需要在可解释性和模型性能之间做出权衡，只要风险得到适当评估和有效管理。允许使用可解释性有限但性能优越的复杂人工智能模型，或许能够使金融机构更好地管理风险并提升客户体验，前提是引入了充分的保护措施。对于监管资本应用场景，复杂人工智能模型可能被限制在特定的风险类别和敞口范围内，或受到输出下限的约束。监管机构还必须投入资源提升员工评估人工智能模型的能力，以确保金融机构能够发挥人工智能的潜力，同时不损害监管目标。

内容

管理解释：监管机构如何应对人工智能可解释性问题

第一部分 – 简介

人工智能（AI）模型正越来越多地应用于金融机构的所有业务活动，从内部运营到面向客户的业务。Crisanto等人（2024）和FSB（2024）强调了金融领域最近的人工智能应用案例，发现大多数应用是为了提高内部生产效率。金融机构在使用人工智能进行关键业务应用方面似乎比较谨慎，尤其是那些涉及客户互动的应用。尽管如此，随着公司寻求从时间和成本效率、改善客户服务以及加强监管合规和风险管理中获益，预计人工智能的使用将变得更加普遍，包括在关键业务领域。

一个关键的监管/监督关注点是人工智能模型的可解释性 [1] 特别是对于关键业务活动（例如面向客户的、核心活动，如承保或确定资本要求）。 [2] 虽然可解释性没有普遍公认的定义，但一些组织从各自的视角定义了这个概念。 [3] 美国监管机构将可解释性定义为“人工智能方法如何使用输入来生成输出”。 [4] 这是在FSB（2024）中引用的相同定义。另一方面，经合组织（OECD）的人工智能（AI）原则更侧重于以客户为中心，将可解释性定义为通过提供易于理解的信息，使受到人工智能系统结果影响的人能够了解其是如何得出的。IAIS（2025）将有意义解释定义为向人们提供关于人工智能系统如何做出决策或预测的理解性强、透明且相关的见解。PRA（2023）将可解释性描述为模型工作原理在非技术术语下可被理解的程度。

某些AI模型结果的缺乏可解释性可能引发审慎关切。FINMA（2024）指出，某些AI模型结果无法被理解、解释或重现，因此无法进行批判性评估。某些AI模型的缺乏可解释性也可能使监管机构难以确定金融机构在模型使用方面是否符合现有的监管要求，特别是在关键业务领域。IAIS（2025）强调了复杂AI模型缺乏可解释性所导致的模型风险，如何会导致不必要的或非法的趋势（例如低估风险）未被察觉，最终可能影响保险公司的盈利能力和资产负债表。

可解释性在使用 AI 模型计算监管资本方面也同样重要。当内部模型首次被巴塞尔框架允许时，巴塞尔委员会强调，“一个银行人员理解不佳的‘黑箱’并不能为评级过程或

[1] 例如，IIF-EY（2025）表明，在金融机构与监管机构/监督机构就人工智能进行交流时，人工智能可解释性是他们提出的主要问题。
[2] 例如，意意联合银行已经使用机器学习来计算信用风险的监管资本。
[3] 一个相关的——但通常被认为不同的——概念是可解释性。IBM 认为可解释性是专注于理解 AI 模型的内部工作机制，与旨在为模型输出提供理由的可解释性形成对比。
[4] 参见 OCC 等人 (2021)。OCC (2021) 将 AI 可解释性定义为 AI 决策过程和结果被银行人员合理理解的程度。

PDs的估计”。 [5] 巴塞尔iii通过移除某些风险类别中使用的高级选项并引入输入和输出地板来缓和此类模型的使用。 [6]

缺乏人工智能模型可解释性可能会潜在地导致系统性风险加剧。 [7] fsb (2024) 将模型风险、数据质量和治理确定为可能导致对金融稳定构成风险的ai相关漏洞来源。报告指出，某些ai模型的复杂性和有限的解释性，以及评估数据质量的困难，可能会增加缺乏稳健ai治理的公司面临的模型风险。

可解释的人工智能模型输出从消费者保护的角度也很重要，以避免歧视性决策。IAIS（2025）指出，可解释性和透明度是建立信任以及追究企业对消费者所面临风险（如非法歧视）责任的关键。某些人工智能模型的复杂性可能会使向消费者解释模型结果变得具有挑战性，并可能导致偏见被未检测到的风险增加。

从金融机构的角度来看，缺乏可解释性构成了采用和部署AI模型的障碍。如果金融机构不知道这些模型的工作原理，他们会谨慎使用AI模型。使用缺乏可解释性的AI模型会加剧金融机构的模型风险。因此，克服可解释性挑战对于避免在能够提升客户体验、监管合规、风险管理和运营效率的AI应用场景中错失机会至关重要。

因此，监管者通常期望企业能够解释用于关键活动或辅助决策的人工智能模型，以确定模型输出是否合适，并对企业进行问责。总而言之，如果人工智能模型的结果无法被理解，监管者不太可能信任其结果。

如前所述，人工智能方法论可以帮助金融机构提高其运营效率，改进风险管理，并为客户提供更多更好服务。因此，一些监管机构已经开始认识到，过于严格的可解释性要求可能会阻碍社会所期望的创新。例如，一些当局现在隐含地允许或明确地承认，一些金融机构可能正在使用人工智能模型用于监管资本目的。 [8] 这可能旨在鼓励创新以及更广泛的人工智能应用发展。至少，它表明金融当局需要批判性地审查其对受监管机构使用模型现行指南，并考虑到人工智能模型可能带来的可解释性挑战。

现存在关于模型风险管理（MRM）的国际标准和区域监管要求，其中一些已明确涵盖或隐含提及可解释性问题。 [9] 尽管如此，这些要求通常是高级别的，可能无法捕捉到人工智能模型的特殊性。因此，就人工智能模型而言，对可解释性概念进行更清晰的阐述将有助于补充现有的MRM要求。

本文旨在描述 MRM 的现行监管指南；讨论将其应用于 AI 模型时存在的挑战，特别是在可解释性有限的情况下；并提出解决其中一些挑战的考虑因素。为此，本文回顾了现有的

[5] BCBS (2001).
[6] 参见 BCBS (2017)。
[7] 参见 Daníelsson 等人 (2022)。
[8] 参见例如 EBA (2023) 和 Bank of England and Financial Conduct Authority (2024)。
[9] IAIS（2025）是国际保险监督官在如何监管保险公司使用人工智能方面的指导性文件的一个例子，同时涵盖了可解释性预期。

关于MRM的指导方针，识别了与模型可解释性相关的元素，分析了该指导方针在多大程度上可以被AI模型满足，并讨论了当前政策框架中可能的改进元素。 [10]

本文的结构如下：第二节讨论可能受缺乏人工智能可解释性影响的MRM需求。第三节概述了不同的AI可解释性方法论以及在更复杂的人工智能模型背景下实施它们的挑战。第四节描述了对现有MRM需求的潜在调整，以应对这些挑战。第五节总结。

第二节——混合模型推理和可解释性

全球标准制定机构（SSBs）已经对金融机构使用模型提出了一些高阶要求。巴塞尔核心原则（BCPs），特别是BCP 15（风险管理流程）基本标准6，规定使用风险模型的银行必须遵守模型使用的监管标准，包括对模型进行独立验证和测试。 [11] 保险核心原则（ICPs），特别是ICP 16（为偿付能力目的的企业风险管理），涉及风险测量的模型使用，包括用于测量技术准备金。 [12] 巴塞尔银行监管委员会（BCBS）也存在其他与模型使用相关的重要文件，例如有效风险数据聚合和风险报告的原则，这些原则要求银行准确可靠地生成风险数据，以及压力测试原则，这些原则规定模型应适合其目的并应接受挑战和定期审查。 [13]

ssbs还就监管资本目的下模型的使用发布了更详细的要求。iais通过其icp 17（资本充足率）提供了详细的要求，而bcbs则有风险特定的指导，特别是信用风险的内部评级基础（irb）方法和市场风险的内部模型方法（ima）。 [14] 所有要求有效治理和控制内部模型的使用，并将最终责任分配给董事会和高级管理层，以了解模型输出的后果和局限性。它们还要求金融机构确保模型方法和假设在概念上是合理的，适合预期用途，并具有良好的预测能力。另一个常见的要求是记录模型设计和假设。

最近，IAIS（2025）阐述了现有ICPs在保险公司使用人工智能的背景下的应用方式。它涵盖了治理和风险管理领域，这些领域在部署人工智能时已被确定为需要特别关注的方面。它还涵盖了技术方面（如数据治理和模型验证）以及其他旨在支持监管机构和保险公司管理人工智能引入或增强的风险的活动。框1概述了如何应用ICPs来解决人工智能模型的可解释性问题。

[10] iif-ey(2025)表明，在涉及人工智能方面，监管机构和公司在讨论中关注的最高问题是可解释性/某些人工智能算法的黑盒性质。
[11] 参见 BCBS (2024a)。
[12] 参见 IAIS (2024)。
[13] 参见BCBS（2013）和BCBS（2018）。
[14] 参见BCBS（2022）和BCBS（2024b）。

框 1

IAIS人工智能监管申请文件——与可解释性相关的指导

- ICP 7（公司治理）- 保险公司在制定方面应有明确的问责制
  对人工智能系统的期望，以便生成的输出可解释、公平且无偏见。
- icp 8（风险管理内部控制）- 保险公司应有有效的风险管理
  内部控制以尽量降低人工智能系统对其财务产生不利影响的风险
  健全性。
- icp 19（经营行为）-索赔决定和索赔的透明度和可解释性
  受人工智能系统影响的争议解决尤其重要。
- ICP 8（风险管理及内部控制）和ICP 19（业务行为）–
  - 监管机构应要求保险公司就人工智能系统的结果作出实质性说明
    它们所使用的，尤其是在对消费者或偿付能力有实质性影响的用例中
    或那些用于满足法律要求的。
  - 保险公司可以限制人工智能系统的部署，仅限于那些简单且可解释的系统或
    限制复杂 AI 系统的使用于挑战和微调更传统的
    数学模型。
  - 部署复杂的人工智能系统可能取决于伴随
    Shapley值或LIME等可解释性工具的部署，这些工具可用于
    说明不同变量对AI结果的影响，提高透明度
    信任，并认识到这些技术的局限性。
  - 在AI系统风险高且/或用于解释模型的工具
    它们自身有限制，保险公司可以考虑替代性的简单模型。
  - 当一个AI系统在新或意外的情况下无法提供足够的置信度时
    条件，保险人应确保其能够安全失效或升级到人工干预。
  - 对于无法实现预期水平的高度复杂人工智能系统
    可解释性，保险公司应考虑采用和记录补充
    治理措施，如使用防护栏（例如增强数据管理）和
    人工监督
  - 一个AI模型的解释水平和深度应该根据不同的
    利益相关者（例如审计员和监管者将需要更全面的技术）
    与保单持有人相比的信息)

来源：IAIS（2025）。

在国家层面，只有少数几家金融监管机构制定了模型风险（MRM）指导方针。基于单一监管机构（SSBs）的工作，这些监管机构——无论是制定了具体的模型风险管理（MRM）指导方针，还是制定了涵盖模型使用的通用风险管理指导方针——通常关注用于计算监管资本的模型。表1列出了本文所审查的MRM指导方针。[15]

[15] 由当局发布的仅仅将SSB模型使用标准移用于监管资本计算的规定（例如，*欧洲央行内部模型指南*）不包含在此处。

mrm指南 表1

	签发机构	文件标题	年/月发布
加拿大	校长办公室 金融机构 (OSFI)	草案指南 E-23 – 模型风险 管理*	2023年11月
日本	日本金融监管机构 (FSA)	模型风险原则 管理	2021年11月
阿拉伯联合 阿联酋	阿拉伯联合酋长国中央银行 阿联酋航空 (CBUAE)	模型管理标准	2022年11月
英国	审慎监管局 (PRA)	模型风险管理原则 对于银行	2023年5月
美国	联邦储备委员会/办公室 货币监理官 (FRB/OCC)	模型监管指南 风险管理	2011年4月
	OCC	模型风险管理 (总审计师手册)**	2021年8月

在参考文献部分提供了指向文档的链接。*原始指南于2017年发布。关于草案指南的咨询已于2024年3月结束，但尚未发布最终版本。**该手册与FRB/OCC 2011年指南保持一致，但更新了它以涵盖与人工智能相关的问题。

rmr指南具有共同要素。所有指南都涵盖治理和监督、模型开发与文档、模型验证与实施、监控与维护。这些指南还要求评估模型风险，以便采用基于风险的方法来应用rmr要求。此外，所有rmr指南都涵盖使用第三方模型时风险的管理。其中一些指南明确提到了模型可解释性的问题。

虽然模型可解释性的概念在许多现有的 MRM 指南中并未明确提及，但它隐含于这些指南中包含的许多条款之中。

• 治理：MRM指南强调董事会和高管在确保其公司MRM框架有效实施中的作用。此外，许多指南也期望他们充分理解模型并提供有效的挑战。

• 模型开发与文档：MRM指南要求模型文档应透明 [16] 以及包括模型的理论、假设、逻辑、规范和局限性等。一些指南还要求方法清晰、阐述详尽，以便所有利益相关者（CBUAE）了解，并要求理解并传达模型的优缺点给模型用户和其他利益相关者（PRA）。

• 模型验证：MRM指南要求模型验证以评估模型的适用性和概念上的合理性。此外，此类评估应由独立于模型开发过程的模型验证者进行。一项指南（OSFI）

[16] Prenio和Yong（2021）解释了透明度是如何成为实现人工智能模型监管评估的前提条件。人工智能模型的透明度取决于多种因素，包括是否为内部开发、开源或闭源，或由第三方服务提供商（包括大型科技公司）开发的专有应用程序。美国联邦信息安全办公室（2024年）强调，人工智能系统的透明度应涵盖人工智能系统及其生态系统的整个生命周期。缺乏对人工智能生态系统全面了解的监管机构可能无法恰当地评估人工智能模型的可解释性。

明确期望模型验证能够确保模型对相关利益相关者来说是可理解的。

- 部署和持续监控：一项指南（FRB/OCC）规定，使用该模型的业务领域如果模型输出似乎不合理，应当能够质疑模型的前提假设。另一项指南（CBUAE）提到，监控过程的目标是评估运营环境的变化是否对模型的表现、稳定性、关键假设和/或可靠性产生了影响。

- 独立审查或内部审计：通常，MRM指南仅期望内部审计评估MRM框架实施的有效性。然而，一些指南（BCBS（2022，2024b），PRA（2023））还规定了独立审查的必要性，以评估模型的适当性和稳健性。

评估模型风险性的要求，以便能够基于风险应用MRM要求，加剧了实施挑战。公司在评估模型风险性时需要考虑的因素包括使用的重要性、模型的复杂性。 [17] 对输入、方法、假设的不确定性，以及对潜在客户的影響。模型風險越高，MRM要求的应用就越是頻繁和強烈。因此，基於這些因素，缺乏可解釋性的AI模型，如果用於高風險領域，很可能會被評為高風險。因此，多少有些諷刺意味的是，MRM指南反而會使可解釋性對缺乏它的模型更具相關性。

使用第三方模型也加剧了缺乏可解释性所带来的挑战。英格兰银行和金融行为管理局（2024）发现，由于使用第三方模型，一半的受访者在调查中报告只对所使用的AI技术有部分了解。通常，期望公司遵守MRM要求，即使对于第三方提供的模型也是如此。一些指南认识到公司面临的挑战，这些公司通常无法完全了解第三方模型。FRB/OCC和FSA指南因此允许调整验证工作。前者规定银行可能不得不更多地依赖敏感性分析和基准测试。后者承认验证可能需要基于现有最佳信息进行。另一方面，CBUAE指南指出，如果公司未能完全理解第三方模型，则不得认为该模型适合用途。

MRM中一个现有指南未明确涵盖的方面与公司对受模型结果影响的客户的责任有关。 [18] 到目前为止的讨论主要关注可解释性问题如何影响大部分“面向内部”的MRM需求实施。然而，最近的AI政策发布强调了外部透明度、外部问责制和程序公平等概念。 [19] 这些概念涉及企业在与客户互动时要告知他们正在与人工智能互动，以及人工智能驱动决策的使用及其影响，同时提供关于人工智能模型如何运作、如何为决策做出贡献以及投诉和补救渠道的简单解释。这些是MRM指南应涵盖的重要要素，因为它们可能导致重大的声誉风险。同时，与

[17] 在PRA MRM指南中，缺乏可解释性被明确提及为导致模型复杂性的因素之一。OCC（2021）也指出，审查人员应评估模型评级是否考虑了可解释性。
[18] 马斯 FEAT 原则，例如，规定使用人工智能的金融机构应当为数据主体（如潜在金融客户）提供查询、提交申诉和请求审查人工智能驱动决策对其产生影响的渠道。欧盟人工智能法案有类似规定，而欧盟通用数据保护条例第22条规定：“数据主体有权不遭受仅基于自动化处理的决策（包括画像），该决策对其产生法律效力或类似重大影响。”
[19] 参见 Prenio 和 Yong (2021) 以及 Crisanto 等人 (2024)。

上述面向内部的需求数据，对外部实现这些面向外部的要求也会因缺乏可解释性而受到限制。

第三节 – 在人工智能背景下执行可解释性要求所面临的挑战

企业可能会发现，满足现有的关于人工智能模型可解释性的监管要求是一项挑战。[20] 深度神经网络等高级人工智能模型由于其众多参数和过度参数化（即参数比数据点多）而难以解释。与线性回归等具有可解释数学关系、可以展示数据输入如何导致模型输出的简单模型不同，高级人工智能模型之所以复杂，是因为它们涉及大量非线性推理。高级人工智能模型的这种黑盒特性可能会掩盖模型输出背后的计算过程。尽管高级人工智能模型可以提供卓越的预测性能，但它们往往以可解释性为代价，导致在问责、公平和伦理方面产生担忧。[21]

构建大型语言模型（LLM）使其功能比其他人工智能模型更为复杂。LLM在规模庞大（与互联网大小相当）的大型数据集上进行训练，其输出是基于涉及数十亿参数的复杂交互产生的。这些模型根据概率生成下一个标记的预测，这意味着随机数抽取决定了在众多最可能候选词中的后续单词。因此，即使输入保持不变，模型输出也可能发生变化。在某些极端情况下，一个不太可能的抽取可能导致模型产生所谓的“幻觉”输出。这种概率特性奠定了模型有效处理复杂查询的能力，但也引入了潜在的负面影响。此外，用于开发LLM的训练数据和过程往往是不可透明的。即使在具有开放访问权限（即公开可用）权重的模型中，评估其未来行为和界定其能力和局限性仍然具有挑战性。

3.1 现有解释人工智能模型的方法及潜在局限性

在大多数政策讨论中，使用“可解释性”一词，而在大多数学术文献中，则使用“可解释性”一词。虽然表面上看这些术语可以互换，但它们可能传达不同的含义。明确定义很重要，以便政策讨论能够清楚地了解预期的政策/监管目标。基于对一组学术论文的回顾，本文采用以下理解：[22] 为了正确地引出下一段文字：

[20] 荷兰央行和荷兰金融市场管理局（2024）强调，人工智能模型可能无法满足某些IRB要求，例如对预期结果的要求必须是“合理且直观的”——人工智能有时会产生不直观但结果却可能比传统模型提供更好的信用风险估计。
[21] EBA（2023）强调了可解释性和模型性能之间的权衡。虽然更复杂的模型可能会带来更好的性能，但它们更难解释或理解。
[22] 参见 Retzlaff 等人 (2024), Gilpin 等人 (2019) 和 Doshi-Velez 与 Kim (2017)。

- 可解释性是指模型的输出能在多大程度上被解释给人类（它回答了“模型为什么会生成这个输出？”或“模型为什么会推荐接受这个信贷申请？”的问题）。 [23]

- 可解释性是指人工智能模型的内部工作机制可以被人类理解的程度（它回答了“人工智能模型是如何得出这个输出的？”或“人工智能模型是如何确定这个属性不应该承保？”的问题）。 [24]

这些概念是相互关联的。 [25] 天生可解释的模型是可解释的，但反过来则不成立。可以通过描述模型在给定输入时会做出什么预测来“解释”一个模型。在实践中，这可以通过评估模型使用训练和测试数据集进行训练的方式，并追踪底层算法如何将新数据输入处理成模型输出来实现。对于一些复杂模型，这可能并不可行，但可以通过描述模型的行为来“解释”模型。这需要推理和论证模型做出某些决策的原因。 [26]

重要的是要认识到这些概念的非二元性，也就是说，很难断定或明确地说一个AI模型是或不是可解释的。当一位监督者评估一个AI模型是否可解释时，结论可能不总是“是”或“否”，而是结论可能是“在一定程度上是”。接受的标准取决于上下文和评估者的要求。

某些 AI 模型是固有的可解释的（也称为先验模型），例如：

- 决策树： [27] 通过遵循如下采用 if-then-else 规则方法的树状结构，对模型进行解释。

- 广义加性模型: [28] 一个解释输入变量和预测输出之间关系的模型。

然而，存在一些黑盒模型，由于其复杂性、非线性和大量参数的使用，本质上是不透明的，使得无法将模型输出与可理解的规则、模式或输入联系起来。 [29] 为了提高这些模型的可解释性，可以使用事后技术来分析黑盒模型在做出

[23] 见第1节，其中收集了金融领域可解释性的定义。其他有用参考包括欧盟数据保护专员（2023）。

[24] 在某些文献中，可解释性和可解释性被互换使用，但本文认为为了监管目的区分这些概念是有用的，因为它们可能对金融机构具有不同的含义。可解释性学术定义的例子有Gilpin等人（2019年）和Lipton（2017年）。Lipton（2017年）说明了可解释性概念并不简单，并确定了赋予人工智能模型可解释性的特征，例如透明性。人们承认可解释性和可解释性存在其他解释，例如NIST（2023年）持相反观点，认为可解释性回答了系统如何做出决策的问题，而可解释性回答了系统为什么做出决策以及其对用户的意义或背景的问题。

[25] 参见gilpin等人(2019)。thampi(2022)指出，可解释性需要以可解释性为基石。

[26] 本文关注不可解释的人工智能模型的可解释性，因为此类模型提出了监管挑战。根据上述定义，可解释的模型也应具有可解释性，因此，对监管机构来说，其引发的问题应更少。

[27] 参见 Molnar (2020)。

[28] 参见 Retzlaff 等人 (2024) 和 Saleem 等人 (2022)。

[29] 参见Retzlaff等人(2024)的研究。欧洲数据保护监督机构(2023)提出了相似的分类方法，但使用了不同的术语：“白盒”（可自解释的模型）与“黑盒”（事后解释）。可解释性技术也存在其他的分类方式，例如Thomas（2024）总结了针对GenAI模型的四种主要通用可解释性技术类别：基于特征的、基于样本的、基于机制的以及基于探测的。

预测/已交付输出。 [30] 后验技术可以根据全局和局部可解释性进一步细分。全局可解释性是指通过捕捉适用于其行为模式的总体规律、趋势和见解来解释模型整体功能的能力。而局部可解释性则是指识别驱动特定输出的具体输入。 [31] 前者更多是模型开发者、验证者和监管者的担忧，而后者更多是寻求在信用决策等用例中解释的个人客户的担忧。

事后技术包括：

- SHapley Additive exPlanations (SHAP)方法 [32] 归因于模型对单个输入特征/因素的预测； [33]
- 本地可解释模型无关解释（LIME）方法 [34] 这解释了最影响预测的显著特征，通过使用符合更简单模型的数据进行拟合与原始数据点略有改动；和
- 反事实解释 [35] 通过识别对特征的最小改变来改变预测/输出的模型输出解释。

后验技术有助于克服人类的认知限制，这些限制可能会限制人工智能模型的可理解程度。Candelon等人(2023)提到了认知负荷理论，该理论指出人类最多只能理解大约七条规则或节点，这使得人类几乎不可能完全理解复杂人工智能系统做出的决策。 [36] 换句话说，虽然公司可能会记录人工智能模型的工作原理，但人类监督者可能并不完全理解。IMDA和PDPC（2020）指出，技术可解释性可能并不总是足够的，特别是如果目标受众是普通消费者。在这种情况下，向普通人提供反事实信息（例如“如果你的平均债务低15%，你的抵押贷款就会被批准”）可能更合适。 [37]

[30] 开源可解释性算法可用于生成事后解释；参见Github。

[31] 参见MAS（2024）。

[32] 参见 Lundberg 和 Lee (2017)。例如，对于一个用于信贷审批的 AI 模型，SHAP 可以解释每个“特征”或风险因素（收入、信用记录等）对模型输出（审批决策）的贡献程度。Davis 等人 (2022) 提供了一个例子，说明了 SHAP 和 LIME 如何解释所选机器学习模型的信贷评估输出。英格兰银行和金融行为监管局 (2024) 发现，特征重要性和 SHAP 是英国调查的金融机构中最广泛使用的解释方法。可视化技术也可能有助于增强可解释性。Goldstein 等人 (2015) 提出了独立条件期望 (ICE) 图，它显示每个实例的预测-特征曲线，以揭示平均部分依赖性可能掩盖的异质性和交互作用，并附带一个简单的可加性可视化测试。Apley 和 Zhu (2020) 介绍了累积局部效应 (ALE) 图，这是一种通过整合预测中的局部变化来可视化特征效应的高效方法。

[33] IAIS（2025）通过保险公司，为选定的AI应用案例提供了可解释性技术的示例。例如，SHAP可用于向消费者解释：“您的保费主要受到您的驾驶记录（影响40%）、车辆类型（影响30%）和位置（影响20%）的影响。”

[34] 参见 Ribeiro 等人 (2016)。例如，为解释人工智能模型的信用贷款推荐，将一个或几个风险因素的价值稍作改变，并使用这些调整后的输入值重新运行该模型。对那些产生输出与原始数据输入更相似的 risk factor，分配更高的权重，从而说明模型是如何得出该推荐的。Ribeiro 等人 (2018) 提供了一种改进的 LIME 版本，称为 anchors。

[35] 参见 Dandl 等人 (2020) 的研究，其中提供了一个反事实陈述的例子：“由于您的年收入为 30,000 英镑，您被拒贷。如果您的收入为 45,000 英镑，您就会被提供贷款。” EBA (2023) 列出了受调查银行用于解释机器学习技术的几种措施；这些措施包括反事实解释。

[36] 讲解的方式可能会影响目标受众的理解。珍珠和麦克琴 (2018) 认为，人类更适应因果关系叙事，而不是抽象的统计概念。

[37] 见 Russell 等人 (2018)。

值得注意的是，这些可解释性技术并非相互排斥，每种方法都有其利弊。[38] 从监督角度来看，理解可解释性技术的局限性很重要（表2提供了可解释性技术局限性的非详尽列表）。

可解释性技术的局限性 表2

限制	描述
不准确	解释可能无法真实反映人工智能模型的实际决策。[1]
不稳定性 灵敏度	对数据输入的微小改动可能会导致截然不同的解释。[2]
无法 泛化	当推广到更广泛的人群/数据集时，解释可能不一定成立。[3]
真实标签	不存在性不存在用于评估正确性或完整性的普遍接受的指标 解释。[4]
误导性 解释	欺骗性的解释可能看起来似是而非。[5]

[1] 鲁丁（2019）解释了诸如LIME或SHAP之类的后验技术如何对底层人工智能模型不忠实。[2] 阿尔瓦雷斯-梅利斯并且 Jaakkola (2018) 解释了模型无关的扰动解释技术为何比...更易不稳定
基于梯度的技术。[3] molnar等（2020）解释说，即使是全局可解释性技术，欠拟合或过拟合也可能
导致模型泛化能力差。[4] 波尔特等人（2022）解释了事后技术在对抗性环境下为何无效
由于其固有的模糊性；易受操控；以及无法提供独特的、真实的原因用于算法
决策。[5] 拉克卡鲁朱和巴斯塔尼（2020）展示了如何操纵黑盒模型以提供误导性解释。

新的可解释性技术[39]正在进行开发，并改进现有方法，[40]所有这些将来都可能证明对协助金融部门监管机构更好地理解金融机构使用的AI模型有用。

3.2 应用现有需求面临的挑战

一个总体的MRM要求是，人工智能模型必须就其如何得出结果而言是可解释的。[41]如果严格解释，这项要求可能对某些人工智能模型（例如深度学习方法）具有挑战性。例如，深度神经网络由多个具有数千个参数的隐藏层组成，这些参数以非线性方式相互作用。无法明确地将模型输出归因于数据输入的组合。使用数十亿个参数的复杂模型（例如GPTs）[42]是另一个如何现有可解释性要求可能不适合用途的例子。

[38] 雷茨拉夫等（2024）概述了可用于比较可解释性方法的准则。阿隆索-罗比斯科和卡尔沃（2025）提供了一个评估SHAP和排列特征重要性的相对可靠性的例子，以解释所选AI模型对信用决策预测的解释。在银行监管的背景下，SHAP可以对“哪些因素总体上对申请人的信用评分贡献最大？”这个问题提供答案。巴克曼和约瑟夫（2023）介绍了一种三步工作流程，以使机器学习预测可解释，并揭示具有经济意义的非线性关系。
[39] 参见，例如，吴等人（2024）。
[40] 参见Dhurandhar等(2023)。Alonso-Robisco等(2025)提出在训练过程中对模型内部逻辑进行约束，以提高模型的信用预测性能及其可解释性。
[41] 参见OSFI（2023）。
[42] 天意（2023）报道OpenAI的GPT-3有1750亿个参数，这些参数是神经网路中在训练过程中进行调整的数值。

此外，可解释性要求可能需要根据目标受众进行调整，例如高级管理层、消费者或监管机构。[43] 然而，大多数现有要求并未做这样的区分。FINMA（2024）提到，当决策需要向投资者、客户、员工、监管机构或审计事务所进行解释时，FINMA对应用程序的可解释性进行了更深入的分析。该分析包括理解应用程序的驱动因素或在不同条件下的行为，以便能够评估结果的合理性和鲁棒性。

一些MRM要求规定了公司需要遵循的模型变更流程；然而，在人工智能模型方面，构成变更的内容尚不明确。有必要明确规定人工智能模型的变更构成。英格兰银行（2025）指出，复杂的、随着新数据可用而自动更新的人工智能模型具有动态性。

使用第三方提供的AI模型在遵守MRM要求方面带来了多重挑战。[44] 在某些司法管辖区，公司必须聘请独立第三方进行模型验证。对于从第三方供应商处获得许可的专有模型，这种独立的模型验证可能具有挑战性，因为供应商可能出于专有原因而不愿意解释其模型的工作原理或训练过程。这种情况在大型语言模型领域得到体现；公司通常既没有财力也没有计算资源来开发自己的模型。[45]

不同类型的AI模型在遵循MRM要求时可能会呈现不同级别的挑战。MAS（2024）强调，与银行通常用于专门针对AI模型已训练好的特定用例的常规AI相比，生成式AI（GenAI）在本质上更具通用性，并可用于银行更广泛的用例。在这些较新的用例中，为评估和测试生成式AI模型可能难以找到现成的真值。

缺乏既有的或全球公认的可解释性方法，特别是对于新型人工智能模型，是满足MRM指南的障碍。MAS（2024）观察到，普遍缺乏用于解释GenAI输出并评估其公平性的既定方法。

[43] osfi和全球风险研究院（2023）概述了在确定适当程度的可解释性时需要考虑的四个因素：需要解释的内容、目标受众、用例的重要性以及模型的复杂性。iais（2025）讨论了如何针对不同的利益相关者定制人工智能输出的解释。davis等人（2022）分析了消费信贷风险机器学习模型的可解释性，针对贷款公司、监管机构、贷款申请人和数据科学家等不同利益相关者。

[44] MAS (2024) 指出，外部模型提供者的透明度不足可能会导致在理解和解释 GenAI 的输出和行为方面存在挑战。

[45] 在欧盟，这通过模型开发者/提供者的实践指南得到一定程度的缓解。该指南的第一章涉及透明性，其中包括一个模型文档表格，概述了模型开发者/提供者需要共享的信息，以确保足够的透明性。

箱2

大语言模型的可解释性

大型语言模型（LLM）正越来越多地被金融机构应用于许多活动，从信息搜索和摘要到客户或管理报告的生成。LLM非常强大，因为它们的底层基础模型是在海量数据集上训练的，以至于人们通常在谈论人工智能时只会想到LLM。

尽管用户不知道模型具体如何运作，但LLMs能够回答任何问题的令人印象深刻的能力，吸引了越来越多的用户的钦佩和信心。

然而，解释和理解大型语言模型是一项极其复杂的任务，因为这涉及到大量的参数和非线性关系。Ameisen等人（2025）使用归因图，对Anthropic的大型语言模型Claude 3.5 Haiku进行了解读，这些图旨在部分追踪模型将特定输入提示转换为输出响应时所使用的中间步骤链。

理解 LLM 中正在发生什么的一种方法是使用思维链（CoT）提示，该术语最早由 Wei 等人（2023）提出。最初，CoT 指人类在提示中详细说明他们的推理过程，以帮助早期一代的模型获得更好的答案。人类会向模型提供推理步骤，然后模型会将这些步骤用作生成其他任务推理的示例。这些方法通过模仿结构化的人类推理来指导改进模型的性能。这些提示是可解释的；然而，LLM 如何使用它们则不是。

随着新一代推理模型的□世，焦点已□□移。□些新模型旨在更直接地复□□似人□的推理模式。□重要的是，□然□些模型生成的解释可能看似模仿了人类的推理方式，但它们并不一定反映模型实际得出结论的方式。换句话说，生成一个引人入胜的解释并不能保证它代表了模型内部的推理过程。

随着更多公司开发基于大型语言模型（LLM）的人工智能应用，根据具体应用场景，如果它们无法充分解释应用的工作原理，可能会成为一个监管问题。鉴于公司可能永远不会知道基础模型是如何训练的（包括使用的数据），这个问题在不久的将来不会消失。

□ OpenAI表示其推理模型是使用□化学□□行推理□□的□言模型，即它□在回复用□之前会□出一个内部思□□。模型会分解提示并考□多种生成回复的方法。

第四节 – mrm指南的潜在调整

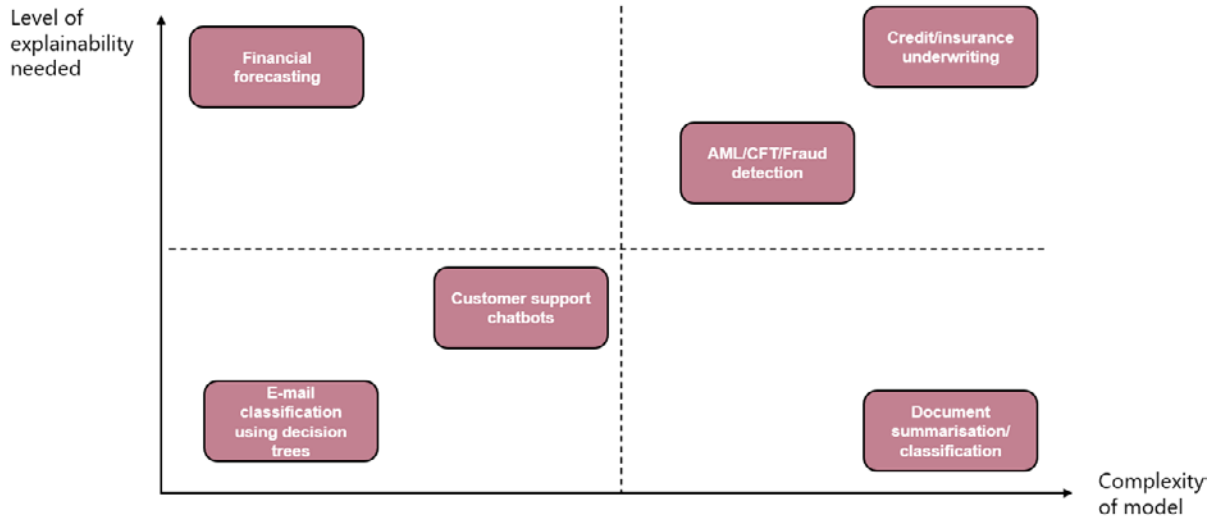
根据以上讨论，当局可能需要审查现有的MRM指南，并确定是否需要制定新指南或对现有指南进行调整，以在人工智能时代保持其相关性。随着金融机构在不同职能和业务领域扩大对人工智能模型的使用，金融当局可能需要就模型在监管资本目的之外的使用提供指导。清晰的AI指南提供监管确定性，并有助于以安全可靠的方式支持模型的使用，同时保护消费者利益。MRM指南不必针对特定技术，但确实需要定期更新，以反映与使用更新技术相关的问题。随着金融机构使用利用先进技术的模型变得更加普遍，这一点尤为重要。

原则上，若AI模型用于关键活动中进行决策，MRM指南可能要求金融机构使用本质上可解释的AI模型 [46] 或者至少采取足够

[46] 参见阿隆索-罗比斯科和卡博 (2025)。

针对黑盒模型的解释性技术。然而，对于复杂模型，仅使用一种现有的解释性方法可能无法完全提供信息。因此，可能需要一套方法来满足利益相关者的不同需求。私营部门和公共部门进行的各种盘点、调查或主题评审表明，金融机构正在使用多种解释性技术来更好地理解复杂的 AI 模型。 [47]

为选定用例中AI模型的相对复杂性和所需相应相对解释水平而进行风格化的插图



rmr指南可能需要要求金融机构为相关用例建立可接受的解释性标准。例如，occ（2021）要求，对于可能难以评估概念合理性的模型，应确定解释性水平是否适合特定模型的应用。公司的模型文件，包括验证报告，应包含有关如何解决高影响案例中使用的ai模型解释性的信息。提供的信息应使监管机构能够评估模型为何做出特定预测。图1说明了监管机构可能根据特定用例的关键性和模型的复杂性要求的不同解释性水平。

更普遍地，可以考虑根据人工智能用例的不同风险程度来定制监管可解释性要求。现有风险管理框架指南中已要求的根据风险程度对模型进行分层或评级，可用于识别可解释性更相关的用例。银行和其他金融机构及监管机构已经开始这样做。 [48] 通常，

[47] 参见 EBA (2023), Bank of England and Financial Conduct Authority (2024), MAS (2024) 和 IIF-EY (2025)。
[48] 参见 OCC（2021）、MAS（2024）、FINMA（2024）和 EIOPA（2025）。

需要更高可解释性标准的AI应用场景是那些可能导致需要向外部利益相关者（如监管机构和客户）解释决策的场景，或者那些模型输出不准确会带来重大（财务）影响场景。

应该明确认识到可解释性和模型性能之间可能的权衡。复杂模型可能会带来性能提升，但这应该与可解释性不足的后果相权衡。[49] 一个可能的初步步骤（EBA（2023））是要求金融机构通过避免不必要的复杂性，在模型性能和可解释性结果之间找到一个适当的平衡。这可以通过，例如，要求开发者证明选择复杂模型而不是简单模型的理由来完成。这可能涉及在各种涉及截然不同的模型输入的场景中，展示模型性能相对于挑战者模型得到改善的程度。

承认这种权衡的一个更具影响力的决定是允许使用那些不完全符合既定可解释性标准但性能明确且显著优于更传统和简单模型的复杂模型。可以说，禁止使用此类模型可能会限制银行和保险公司利用先进技术来有效管理风险和改善客户体验，从而支持诸如消费者满意度和金融稳定等具有社会价值的目标。

然而，可解释性豁免的引入应仅影响可解释性差距有限的人工智能模型，并考虑此类模型使用的风险程度。此外，它应意味着应用充分的保护措施。例如，频繁的稳定性/可重复性测试，[50] 同样也可能由第三方进行，并且可能需要对模型输出进行持续监控以确定模型输出在不同条件下是否保持一致。此外，可能需要替代的增强型风险管理措施、数据治理和人工监督来弥补可解释性不足。[51] 公司可以披露它们如何调整复杂的AI模型，以提高模型输出的可靠性。[52] 此外，断路器可以作为自动机制，在极端或意外情况下停止模型使用，提供一层防范不利结果的保护。对于可能对机构安全稳健或消费者利益产生重大潜在影响的用例，MRM指南可能要求银行随时准备在识别到相关性能缺陷时迅速停止使用这些模型。

解决用于监管目的的人工智能模型的低可解释性问题更加棘手。如上所述，复杂的人工智能模型可能无法满足现行的模型风险管理规定，包括用于监管目的的模型的相关规定。因此，用于此类目的的人工智能模型的使用可能会被禁止。然而，这会消除金融机构使用潜在表现良好的模型进行风险管理的激励。这可能会阻碍此类人工智能用例的发展，从而牺牲潜在的收益。一种折衷方案可能是，允许在一定限度内使用表现良好且复杂的人工智能模型来计算拨备、最低资本或其他监管义务。例如，此类模型可能仅被允许用于某些风险类别和风险敞口，或者使用此类模型计算的风险权重将受到比巴塞尔协议III中为更传统内部模型所设想的更严格的下限约束。

[49] osfi和全球风险研究院（2023）提供了一个不同的观点，即复杂模型并不总是比可解释模型产生更准确预测性能。
[50] 参见IMDA和PDPC（2020）。
[51] 参见 IAIS (2025)。
[52] 在 llms 的情况下，此类调整包括使用与所涉及用例更相关的数据对预训练模型进行微调，或使用检索增强生成向其提供相关的文档/信息，以便将输出限制在该特定信息集内。

第五节——结论

人工智能的应用预计将在金融机构的业务活动中更加普及，因为它们寻求优化新技术带来的收益。这意味着人工智能的应用将不仅限于内部提效的目的，还可能扩展到金融机构的关键业务领域。因此，金融监管机构需要关注人工智能应用对金融机构个体面临的以及整个金融体系面临的潜在影响。

某些人工智能模型的缺乏可解释性是金融监管机构的一个关键担忧。虽然金融机构有几种可解释性技术可供使用，但将这些技术应用于更复杂的AI模型（如LLMs）时存在局限性。随着金融机构在关键业务领域推出更复杂的AI模型，这将影响消费者、监管合规和系统性风险。本质上，这将增加金融机构的模型风险。需要特别关注闭源、专有模型，包括许多基础模型不透明的LLMs。使用LLMs的公司不知道基础模型是如何训练的，这种缺乏可解释性可能会限制其使用场景到低风险活动。

因此，金融监管机构寻求促进金融机构中考虑人工智能发展的稳健的MRM实践是至关重要的。监管机构可以通过发布以下MRM指南来实现这一目标：

(1) 解决金融机构更广泛地使用模型的问题，即不仅用于监管目的；（2） 认识到金融机构使用或将要使用的模型已经发展到包含使用人工智能的模型；以及 (3) 反映行业惯例并随着其发展而调整。

在人工智能可解释性的背景下，例如，MRM指南可以包括要求金融机构采用可解释性技术来解释黑盒模型，根据模型的潜在影响和风险性建立可解释性标准，并要求补充性保护措施，如加强数据治理和人类监督，以减轻在关键业务领域使用复杂人工智能模型相关的风险。最终，可能需要认识到可解释性与模型性能之间的权衡，只要风险得到适当评估和有效管理。可以考虑为具有经过充分验证的良好性能的模型在应用可解释性要求时提供一些有条件和受约束的灵活性。

当局也需要提升其员工技能，以便能够理解企业提交的可解释性提交。这并非一项微不足道的工作，因为即使是本身就可解释的模型，理解起来也可能具有挑战性，更不用说黑盒模型的可解释性技术了。

参考文献

阿隆索-罗比斯科，A 和 J 卡博 (2025): “我们应该信任机器学习模型提供的信用决策吗？”，*计算经济学* ，一月。

阿尔诺-罗比斯科，A，J卡博，G德哈罗和J吉伦加西亚（2025）：“用于信用违约预测的约束机器学习模型：谁胜出，谁失利”，佛罗伦萨银行与金融学院， *银行监管政策研究论文系列* ，无2025/05,六月。

阿尔韦雷斯-梅利斯 D 和 T 耶科拉 (2018): “关于可解释性方法的鲁棒性”,六月。

蚂蚁, E, J Lindsey, A Pearce, W Gurnee, N Turner, B Chen, C Citro, D Abrahams, S Carter, B Hosmer, J Marcus, M Sklar, A Templeton, T Bricken, C McDougall, H Cunningham, T Henighan, A Jermyn, A Jones, A Persic, Z Qi, T Thompson, S Zimmerman, K Rivoire, T Conerly, C Olah 和 J Batson (2025): “电路追踪：揭示语言模型中的计算图” *Transformer 电路线程* ，三月。

阿普利, D和朱J (2020): “可视化预测变量在黑盒监督学习模型中的影响” *英国皇家统计学会杂志B：统计方法* ，第82卷, 第4期, 9月, 第1059–86页。

英格兰银行（2025年）：聚焦金融稳定：金融系统中的人工智能，四月。

英格兰银行和金融行为监管局 (2024): *英国金融服务中的人工智能——2024* ，十一月。

巴塞尔银行监管委员会（BCBS）(2001): *内部评级法——咨询文件* ，一月。

——— (2013): 关于有效风险数据聚合和风险报告的原则，一月。

—— (2017): 巴塞尔iii改革高级摘要，12月。

——— (2018): 压力测试原则，十月。

——— (2022)：“CRE 36 – IRB方法：使用IRB方法的最低要求” *巴塞尔框架* ，十二月。

——— (2024a)：有效银行监督的核心原则，四月。

——— (2024b): “3月30日——内部模型方法：总则”, 巴塞尔框架, 七月。

联邦储备系统理事会和货币监理署 (2011): *模型风险管理监管指南* ，四月。

bordt, s, m finck, e raidl和u von luxburg（2022年）：\"事后解释在与对抗性相关的环境中无法实现其目的\"，五月。

巴克曼，M和A约瑟夫（2023年）：“一个可解释的机器学习工作流及其在经济发展预测中的应用” *国际中央银行期刊* ，第19卷,第4期,十月,第449-522页.

坎德隆，F，T Evgeniou和D Martens（2023）：“AI可以既准确又透明” *哈佛商业评论* ，五月。

阿拉伯联合酋长国中央银行 (CBUAE) (2022): 模型管理标准, 11月。

克里斯亚诺，J，C 莱特里奥，J 普雷尼奥和 J 尧（2024年）：“监管金融领域的AI：近期发展与主要挑战” *fsi关于政策实施的观点* 63号，十二月。

丹尼尔，S，C莫尔纳，M比纳尔和B比施尔（2020）：“多目标反事实解释”，在 *自然并行问题求解——PPSN XVI* ，第16届国际会议论文集, PPSN 2020, 荷兰莱顿, 2020年8月, 第448–69页。

丹尼尔森，J，R 麦克雷和 A 乌特曼（2022）：“人工智能与系统性风险” *银行与金融杂志* ，第140卷，七月。

Davis，R，A Lo，S Mishra，A Nourian，M Singh，N Wu和R Zhang（2022）：“可解释的机器学习模型消费者信用风险” *金融数据科学杂志* ，第5卷, 第4期, 1月, 第9–39页。

荷兰银行和荷兰金融市场监管局（2024）： *人工智能对金融行业和监管的影响* ,六月。

Dhurandhar, A, K Ramamurthy, K Ahuja和V Arya（2023）： *局部不变解释：通过局部不变学习实现稳定和单向解释* ，九月。

多希-维莱兹，F和金B（2017）：“迈向可解释机器学习的严谨科学”，3月。

欧洲银行管理局 (EBA) (2023): *用于IRB模型的机器学习：关于IRB模型讨论文件咨询的后续报告* ，八月。

欧洲数据保护专员 (2023): “可解释人工智能” *技术快讯* ,no 2/2023,十一月。

欧洲保险和职业养老金管理局（EIOPA）（2025）： *关于人工智能治理和风险管理意见的咨询文件* ，二月。

联邦信息安全局（2024年）：“人工智能系统的透明度”，白皮书，8月。

财务报告委员会（2024年）：技术精算指南——模型，10月。

日本金融厅 (FSA) (2021): 模型风险管理原则，十一月。

金融稳定委员会（FSB）（2024年）：人工智能对金融稳定的影响，十一月。

吉尔平，L，D Bau，B 元，A 巴贾瓦，M 斯佩克特和L 卡格尔（2019）：“解释解释：机器学习可解释性概述”，2月。

Goldstein, A, A Kapelner, J Bleich和E Pitkin（2015年）：“透视黑箱：使用个体条件期望图可视化统计学习”， *计算和图形统计杂志* ，第24卷, 第1期, 3月, 第44–65页。

天，W (2023)：“GPT-4 比ChatGPT更强大，但 OpenAI 不会说为什么” *麻省理工学院科技评论* 三月。

Infocomm媒体发展局（IMDA）和新加坡个人数据保护委员会（PDPC）（2020）： *人工智能治理框架模型——第二版* ,一月。

国际金融学院和艾德韦宣德 (IIF-EY) (2025): *IIF-EY年度调查报告：金融服务业中人工智能/机器学习使用情况——公共摘要* ,一月。

国际保险监督官协会（IAIS）（2024）： *保险核心原则和国际活跃保险集团监管通用框架* ，十二月。

（2025年）：人工智能监管应用文件，七月。

拉克卡尔朱，H和O巴斯塔尼（2020）：“‘我该怎么欺骗你？’：通过误导性黑盒解释操纵用户信任”，载于 *AIES '20* ，AAAI/ACM人工智能、伦理与社会会议论文集，纽约，二月，第79–85页。

Lipton，Z（2017）：“模型可解释性的神话”，三月。

隆德伯格，S和S-I李（2017年）：\"一种解释模型预测的统一方法\"，11月。

molnar, c (2020): *可解释机器学习：让黑盒模型可解释的指南。*

molnar, c, g konig, j herbinger, t freiesleben, s dandl, c scholbeck, g casalicchio, m grosse-wentrup 和 b bischl (2020): “解释机器学习模型时应避免的陷阱”，七月。

新加坡金融管理局 (MAS) (2024): *人工智能模型风险管理——主题回顾的观察*，十二月。

美国国家标准与技术研究院 (NIST) (2023): *人工智能风险管理框架（AI RMF 1.0）*，一月。

货币监理署 (OCC) (2021年)：\"模型风险管理\ *审计长手册——安全和稳健*，八月。

美国货币监理署、美联储理事会、联邦存款保险公司、消费者金融保护局和国家信用合作社管理局（2021年）：《关于金融机构使用人工智能（包括机器学习）的信息请求和意见征询》 *联邦公报*，卷 86, 第 60 期, 三月。

金融机构总监办公室（OSFI）（2023）： *起草指南 E-23 – 模型风险管理*，十一月。

OSFI和全球风险研究所（2023）： *金融行业人工智能论坛：对负责任人工智能的加拿大视角*，四月。

珍珠, J 和 D 麦克辛 (2018): 因果新书: 因果新科学, 五月.

普雷尼奥，J 和 J 杨（2021）：“人类控制人工智能——金融领域新兴的监管预期” *政策执行的FSI洞察*，第35期，八月。

保险监管局 (PRA) (2023): “银行模型风险管理原则” *监督声明*，没有 SS1/23, 五月。

雷茨拉夫，C，A 安格尔施米德，A 沙兰蒂，D 施奈贝格，R 罗特格尔，H 米勒和A 霍尔岑格（2024）：“事后解释与事前解释：面向数据科学家的可解释人工智能设计指南” *认知系统研究*，第86卷, 八月。

Ribeiro，M，S Singh和C Guestrin（2016年）：“‘我为什么要信任你？’解释任何分类器的预测”，8月。

——— (2018): “锚点：模型无关的高精度解释” *AAAI人工智能会议论文集*，第32卷, 第1期, 4月.

鲁丁，C (2019): “停止解释高风险决策中的黑盒机器学习模型，改用可解释模型” *自然机器智能*，第1卷，5月，第206–15页。

拉塞尔、C、S瓦赫特和B米特尔施塔特（2018年）：“无需打开黑盒的假设性解释：自动化决策和GDPR”，三月。

赛利姆、R、B元、F库鲁戈卢、A安祖姆和L刘（2022）：“解释深度神经网络：关于全局解释方法的综述” *神经计算*，第 513 卷，十一月，第 165–80 页。

瑞士金融市场监管机构 (FINMA) (2024): *FINMA 指南 08/2024 – 使用人工智能时的治理和风险管理*，十二月。

thampi, a (2022): 可解释人工智能：构建可解释的机器学习系统，七月。

托马斯，R（2024）：“揭露生成式人工智能：理解可解释性技术”，8月。

卫, J, X王, D Schuurmans, M Bosma, B Ichter, F Xia, E Chi, QLe和D Zhou（2023）：“思维链提示激发大型语言模型的推理”，一月。吴, Y, M Keoliya, K Chen, N Velingker, Z Li, E Getzen, Q Long, M Naik, R Parikh和E Wong（2024）：“DISCRET：为治疗效果估计合成忠实解释”，六月。