VOICE OF AMAC

（2026 年第 2 期，总第 205 期）

中国证券投资基金业协会　　2026 年 4 月 1 日

基金行业生成式人工智能可解释性治理探讨

【编者按】本研究立足行业实践与境外经验，揭示出大模型“算法黑箱”在赋能从业者与服务投资者场景中引发的信任危机与合规挑战，而破解这一困局需要构建兼顾技术特性、商业逻辑与监管需求的治理框架。基于境外经验比较，本文提出行业驱动与监管协同的治理路径：基金公司应前置合规风控要求，优化模型学习机制，落实信息披露与风险告知义务，筑牢投资者信任；监管部门可通过差异化监管、完善行业标准，共同推动可解释性建设规范发展。

一、引言

以大语言模型为核心的生成式人工智能（AI）正引发金融行业的效率革命。生成式 AI 指模拟输入数据的结构和特征以生成派生合成内容的 AI。基金行业中，生成式 AI 已赋能从业者使用、投资者服务两大场景。但在提升效率的同时，

可解释性问题也日渐凸显，即生成过程无法被清晰解释，“算法黑箱”难以溯源。但若详尽披露技术细节，或与基金公司的知识产权保护相冲突，甚至出现安全风险。

我国数字经济相关立法已初步形成生成式AI规制体系，但针对可解释性问题，还没有明确的规定，仅《国家人工智能产业综合标准化体系建设指南（2024版）》提出，要规范人工智能全生命周期的伦理治理要求，其中包括可解释性的技术要求与评测方法。以可解释性为抓手构建合规框架，是规范生成式AI在基金行业应用的重要路径。

二、基金业务智能化转型与可解释性挑战

（一）赋能从业者：从简单检索到智能助理

生成式AI在基金行业已形成多维度渗透，Morgan Stanley、Wells Fargo、J.P.Morgan等金融机构陆续部署生成式AI应用。在数据分析层，生成式AI可以高效解析新闻、财报、舆情等异构信息，打通非结构化数据价值链条。在投研决策层，可以通过深度学习构建市场趋势预测与资产定价模型，辅助投资组合优化。在风险管理层，可以持续监测交易合规边界，动态构建风控指标模型。在合规运营层，可以覆盖员工申报智能问答、专户合同条款审查、固收交易语料分析及双录材料审核等。生成式AI在“信息处理-策略生成-合规风控-流程执行”等核心环节，体现出系统提升基金前中后台业务效能与风控精度的优势。

然而，从业者使用生成式AI过程中，也存在可解释性挑战。大模型决策过程和输出结果难以被清晰解读，如在投资决策中，基金经理和投资者需要理解决策的逻辑和依据，

以判断决策可靠性。另外，金融市场本身具有高度复杂性，受政治、经济、心理等多种因素影响，而大模型基于历史数据进行训练，难以涵盖所有潜在因素。再者，若训练数据存在噪声、缺失或偏差，更加难以做出合理清晰的解释。

（二）服务投资者：从产品推荐到信任构建

生成式 AI 能通过对客户浏览历史、交易记录、咨询问题等多源数据的深度学习，精准洞察投资者的投资目标、风险承受能力和投资偏好等，再依据投资者的个性化特征，从海量基金产品中筛选并构建适配的投资组合，根据市场动态、持仓情况，生成个性化的市场解读、投资建议等内容，及时推送给投资者，该模式无疑更能得到投资者的信赖。目前市场上已有一些大模型或智能投顾助手在这方面做出积极实践。

但面向投资者时，生成式 AI 可解释性挑战更为突出。投资者难以知晓推理过程，所以当推荐的投资组合出现收益波动或未达预期时，投资者或对生成式 AI 的可靠性产生质疑。在市场突发“黑天鹅”事件时，大模型也无法准确解释市场异动。部分投资者认为，生成式 AI 的可解释性是其知情权的一部分[1]。若可解释性低，投资者会怀疑智能投顾给出的资产配置建议是推荐了高度关联公司利益的产品。但是，如果将基金公司的专有算法毫无保留地公开，又不利于知识产权保护，甚至使公司面临技术攻击。

基金行业投资者服务的核心在于构建投资者信任，生成式 AI 的深度应用拓展了信任构建的空间，而可解释性问题

[1] 徐凤.人工智能算法黑箱的法律规制——以智能投顾为例展开[J].东方法学,2019,(06):79.

成为了进一步建立投资者信任的瓶颈。如何构建有效的可解释性监管框架，已成为各司法辖区的共性核心议题。

三、境外 AI 监管实践的现状与反思

（一）欧盟：“基于风险”的分级监管

欧盟自 2016 年起就开始探索对 AI 的监管体系构建，最具代表性的是 2024 年通过的《人工智能法案》[2]。作为全球首部全面规范人工智能的法律，是可供借鉴的立法之一。（欧盟 AI 法规政策梳理见附表 1）

针对可解释性问题，2022 年《数字服务法案》就已有规范，尤其是涉及算法决策和内容推荐的部分，要求大型在线平台对 AI 生成的内容和决策提供清晰的解释。2024 年《人工智能法案》更进一步对于通用型人工智能（GPAI）提出了透明度要求。《人工智能法案》第 13 条第 1 款[3]概括规定了高风险 AI 的可解释要求、第 2 款[4]规定高风险 AI 应提交使用说明。值得借鉴的是，第 3 款[5]具体规定了使用说明应提供的信息。此外，《人工智能法案》第 86 条还规定了个体决策的解释权[6]，明晰了个体有权要求部署者进行解释的情况。

---

[2] 全称为《2024 年 6 月 13 日欧洲议会和欧盟委员会第（EU）2024/1689 号关于人工智能的统一规则条例，该条例修订了第（EC）300/2008 号、第（EU）167/2013 号、第（EU）168/2013 号、第（EU）2018/858 号、第（EU）2018/1139 号和第（EU）2019/2144 号条例以及第 2014/90/EU 号、第（EU）2016/797 号和第（EU）2020/1828 号指令（“人工智能法”）》。

[3] 《人工智能法案》第 13 条第 1 款规定：“高风险人工智能系统的设计和开发应确保其操作具有足够的透明度，使部署者能够解释系统的输出并加以适当使用。”

[4] 《人工智能法案》第 13 条第 2 款规定：“高风险人工智能系统应附有适当数字格式或其他形式的使用说明，其中包括简明、完整、正确和清晰的信息，这些信息对部署者来说是相关的、可获取的和可理解的。”

[5] 《人工智能法案》第 13 条第 3 款规定：“使用说明应提供（1）提供者及其授权代表的身份和详细联系信息；（2）高风险人工智能系统的特点能力和性能限制（详细规定在 i–vii 项）；（3）提供者在初始符合性评估时预先确定的对高风险人工智能系统及其性能的改变（如有）；（4）第 14 条所述的人类监督措施（包括为便于部署者解释高风险人工智能系统的输出而采取的技术措施）；（5）所需的计算和硬件资源、高风险人工智能系统的预期寿命以及任何必要的维护和保养更新措施与频率；（6）说明高风险人工智能系统所含机制以使部署者能够根据第 12 条适当收集储存和解释日志。”

[6] 《人工智能法案》第 86 条规定：“任何受到部署者根据附件三所列高风险人工智能系统（第 2 项所列系统除外）的输出结果所做的决定影响的人，如果认为该决定对其健康、安全和基本权利产生了不利影响，并产生了法律效力或类似的重大影响，应有权要求部署者就人工智能系统在决策程序中的作用和所做决定

（二）美国：多元监管模式

美国AI监管呈现出“以行政命令推进、以标准指南引导、联邦和州分级监管”的多元特征。2025年5月，众议院通过名为《一项宏大美好法案》的预算协调法案，其中包含一项禁止各州制定AI法规的10年禁令，旨在限制各州AI监管权。（美国AI法规政策梳理见附表2）

在此背景下，目前主要可参考的综合性监管依据是《关于安全、可靠和可信地开发和使用人工智能的行政命令》（第14110号行政命令）。行政命令第2节第（a）款提出了对AI进行评估测试的必要性[7]，以及对评估测试的要求[8]，该款还提出“本届政府将帮助开发有效的标签和内容来源机制，以便美国人能够确定内容何时是使用人工智能生成的”。要言之，该行政命令要求相关企业向联邦政府提交与模型训练、开发和生产相关的活动计划等信息。企业还需要在向用户发布新功能前进行测试，并将结果提交至联邦政府。另外，行政命令指出应建立内容认证和水印指南。

（三）中国香港：“应用为本”的务实策略

我国香港地区在2025年4月公布了《香港生成式人工智能技术及应用指引》（以下简称《指引》），该份《指引》提出“以应用为本”和“风险分级”的核心理念，其“四层次”风险系统与欧盟《人工智能法案》相似，分门别类设置了监管策略。（我国香港地区AI法规政策梳理见附表3）

的主要内容做出明确而有意义的解释。”

[7] 《第14110号行政命令》第2节第(a)款：“人工智能必须安全可靠。实现这一目标需要对人工智能系统以及政策、机构和其他适当的机制进行稳健、可靠、可重复和标准化的评估，以在这些系统投入使用之前测试、理解和减轻风险。”

[8] 《第14110号行政命令》第2节第(a)款：“测试和评估，包括部署后性能监控，将有助于确保人工智能系统按预期运行，能够抵御误用或危险修改，以合乎道德的方式以安全的方式开发和运行，并符合适用的联邦法律和政策。”

针对可解释性问题，《指引》在2.3.2节安全透明原则中提出了对模型层面和服务层面的要求[9]。接着，还以开源模型为例，指出了安全透明和商业秘密之间的权衡问题[10]。

《指引》区分了技术开发者和服务提供者。在3.1节对技术开发者提出了具体要求[11]，此外还提出“人工智能系统应内置来源核查、事实查证及验证机制”。在合规团队的工作内容上，也给出了详细指导[12]，要求开发者披露训练数据来源、模型架构及评估指标。对于服务提供者，《指引》也提出了高标准[13]，要求当人工智能系统输出的内容无法被验证时，应主动向用户发出提示。更进一步，在3.2.1节还要求服务提供者提供详细的服务说明和使用指南[14]。

不仅如此，我国香港地区作为国际金融中心，对生成式AI在金融行业的应用尤为关注，特别强调金融投资者和客户应对个人资料和偏好行使控制权[15]。在《指引》的附录2.2.2

[9] 《香港生成式人工智能技术及应用指引》2.3.2节：“在模型层面，应透过算法优化及数据治理，消除涉及违法、违规或违背伦理标准的有害内容。在服务层面，服务提供者必须充分向用户披露相关风险，并运用如加密技术及可解释人工智能等手段，提升模型的安全性与透明度，确保技术的可靠运行。”

[10] 《香港生成式人工智能技术及应用指引》2.3.2节：“开源模型通常具有更高的透明度，可审查其训练方法、数据来源及算法设计，从而有助于更广泛地验证安全措施及发现潜在偏见或漏洞。而专有模型虽然往往拥有先进的功能，但由于商业考虑，其透明度存在固有限制，使得外界独立验证变得更具挑战性。”

[11] 《香港生成式人工智能技术及应用指引》3.1节：“开发者应在可行情况下披露训练数据来源、模型架构及评估指标。机构应制定相关政策，规定何时可接纳人工智能生成的内容，例如要求用户在使用前仔细核对人工智能生成的资料、验证参考来源及确保内容准确。”

[12] 《香港生成式人工智能技术及应用指引》3.1节：“建立完善的文档制度，促使技术开发者遵循透明度原则，公开技术原理和使用规则，从而使服务使用者和监管机构能够理解和监督其技术应用，建立信任。”

[13] 《香港生成式人工智能技术及应用指引》3.1节：“人工智能模型必须以经过验证且可靠的来源进行训练，服务提供者亦应整合可信的参考资料，并公开披露其信息来源。当人工智能系统输出的内容无法被验证时，应主动向用户发出提示，以防止未经证实或虚假内容的流传。”

[14] 《香港生成式人工智能技术及应用指引》3.2.1节：“公开其服务背后的技术原理，说明服务使用者和监管机构理解和监督其提供的服务。提供详细的服务说明和使用指南，使得服务使用者能够正确理解和使用服务。”

[15] 中国香港《有关在金融市场负责任地应用人工智能的政策宣言》第12条(b)项：“金融机构在设计人工智能模型时，特别在制定商业决策时，应厘定模型演算法的关键原则，以缓解歧视及对待客户不公平或有偏差等风险。投资者及客户应充分了解人工智能在有关产品及服务的应用程度，掌握有关个人的数据是如何收集、处理及使用的，从而让他们就个人资料和偏好行使控制权。”

节中，还特别针对金融行业作出提示[16]，强调要给予用户选择权，确保用户在选择不使用时应能及时终止相关服务。

总体而言，我国香港地区在制定生成式AI监管规则时，秉持着一种务实的策略。针对具体问题进行了详细规制，但又与技术发展之间保持着一定平衡，具有重要借鉴价值。

（四）新加坡：灵活弹性的“软法”机制

新加坡主要是将监管职能分别纳入各个行业监管部门，通过发布非约束性的指南和建议实施AI治理。（新加坡AI法规政策梳理见附表4）

针对可解释性问题，新加坡《生成式人工智能治理模型框架》（以下简称《框架》）的执行摘要建议第（3）项提到：“尽管终端用户可能看不到开发过程，但围绕基本安全和卫生措施的有意义透明度是关键。”《框架》在正文“可信的开发和部署”一节中，具体提出业界应围绕最佳实践进行开发和安全评估[17]。并且，《框架》形象地将可解释性与透明度比作“食品/成分标签”，向下游用户提供信息，意义在于“将披露标准化有助于模型之间的可比性，并促进更安全地使用模型”。

对于应披露的内容，《框架》给出了示范性规定[18]，值

[16] 《香港生成式人工智能技术及应用指引》附录2.2.2节：“金融行业应注意加强使用生成式人工智能的公平性，如使用生成式人工智能提供推荐或辅助决策类服务时，应确保所有潜在的候选项均能公平地得到被推荐的机会，在可能的情况下，金融部门应考虑采取机制避免人为操纵，限制通过人工设置、干预模型训练或其他方式干预推荐权重。在适用的情况下，金融机构可能需要定制模型以满足特定的使用者需求。金融部门应尽可能考虑提供充分的资讯披露和可选择性，帮助使用者了解生成式人工智能的工作机制、效果及潜在的负面影响，应尽量确保使用者是出于主动意图使用生成式人工智能，同时在不选择使用时应能及时终止相关服务。”

[17] 新加坡《生成式人工智能治理模型框架》：“即使是开源模型，一些重要信息如方法和数据集也可能无法公开。业界应围绕最佳实践进行开发，进而进行安全评估，此后在所采取的基线安全和卫生措施方面实现有意义的透明。”

[18] 新加坡《生成式人工智能治理模型框架》：“（1）使用的数据。包括训练数据类型以及训练前如何处理数据的概述；（2）训练基础设施。包括对所使用的训练基础设施的概述，并在可能的情况下，估计对环境的影响；（3）评估结果。包括已完成的评估和主要结果概览；（4）缓解和安全措施。包括已实施的安全

得借鉴。对于披露详细程度，《框架》指出可以根据透明的需要和保护专有信息的需要进行调整，业界需要达成共识，政府和第三方也可以协助[19]。由此可见，新加坡虽然规定了具体披露内容，但围绕基线仍保留一定弹性空间。

四、针对可解释性构建生成式人工智能治理路径

立足中国特色金融文化的治理哲学，参考借鉴境外AI监管智慧的精髓，基金行业可构建以市场自律为核心、监管引导为支撑的双层治理机制，探索金融治理中国范式，又契合技术伦理全球共识的创新路径。对于可解释性，有学者提出两个关键维度：一是“面向用户”的可解释性，以用户可理解的方式呈现模型决策逻辑，满足透明度规范要求；二是“面向隐藏层”的可解释性，通过解析内部隐藏层机制，支撑透明度的技术实现。[20]基金公司需差异化推进两个维度的能力建设，监管部门可加以引导，推动行业生成式AI可解释性建设规范发展。

（一）基金行业生成式AI可解释性建设

1.面向用户的可解释性建设

（1）信息披露义务

基金公司面向用户时，即在赋能从业者和服务投资者场景中使用生成式AI时，应在前端交互界面或用户协议中披露注明：（a）该生成式AI应用使用的宏观技术原理；（b）参与模型设计、实施、管理、监督和审查的人员情况；（c）

措施，如偏差纠正技术；（5）风险和局限性。包括模型的已知风险；（6）预期用途。包括对模型预期用途的明确说明；（7）用户数据保护。包括对如何使用和保护用户数据的概述。”

[19] 新加坡《生成式人工智能治理模型框架》：“披露的详细程度可以根据透明的需要和保护专有信息的需要进行调整。进一步，业界应就基线透明度达成一致，将其作为向所有各方进行一般性披露的一部分。这既涉及模型开发者，也涉及应用部署者。另外，政府和第三方也可以协助制定这样的基线。”

[20] 郑飞,朱溯蓉.人工智能“可解释性”的两个维度及其适用[J].大连理工大学学报(社会科学版),2025,46(02):80-87.

建模训练和模型测试的数据来源、清洗情况等；（d）模型的稳健性、可靠性等性能。上述信息应是面向用户披露的最低要求。若有重大更新，还应以足以引起用户注意的方式进行提示，发布更新日志，或重新签订用户协议。

（2）风险告知义务

《互联网信息服务算法推荐管理规定》第16条明确了算法决策告知义务。[21]据此，基金公司若作为算法控制者，必须显著标识算法的部署情况，充分提示算法的潜在风险。该义务旨在保障用户知晓AI算法决策的存在，从而自由地选择AI算法决策或人工决策。同时，该义务不涉及复杂的技术细节，因此，境外监管机构也往往将其设置为刚性规范并要求严格执行。

2.面向监管的可解释性建设

（1）合规风控前置化介入

基金公司在生成式AI大模型部署的前期，就应该注重合规风控的前置化介入。即算法的搭建需要以法律法规、行业规章、内部规则为依据，先从法律构成要素中提取合规特征，再将该特征写入算法中，进行特征算法的模型构建。

如果在设计初期，算法就是以法律逻辑为基础进行构建的，那么之后基金业务行为的特征转化为数据行为特征后，对于监管者审查而言更易于理解。如此，后续的报告也能对算法进行法律化解释，有利于审查者进一步审查模型可靠性。

（2）优化模型学习机制

[21]《互联网信息服务算法推荐管理规定》第16条：“算法推荐服务提供者应当以显著方式告知用户其提供算法推荐服务的情况，并以适当方式公示算法推荐服务的基本原理、目的意图和主要运行机制等。”

多家中美研究机构联合发布的有关大型语言模型可解释性技术的综述中[22]，详细讨论了传统微调和基于提示的两种训练范式在超大规模模型中的应用及其解释技术，指出要使模型更具可解释性可能会最终降低其决策质量。可解释性与性能之间的关系推演可见下图：

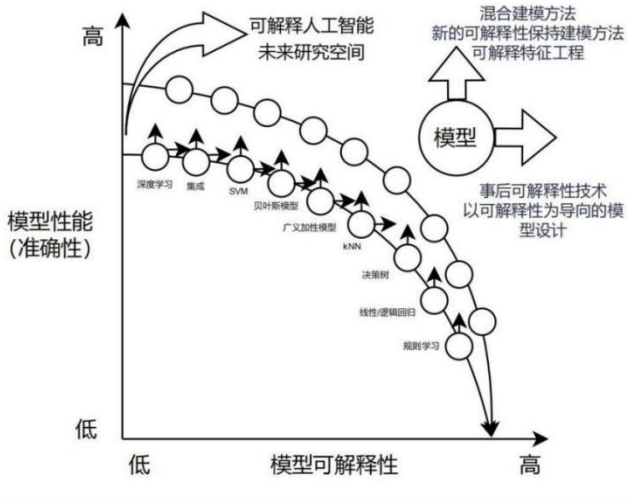

图 1 可解释性与性能之间的关系

研究表明，预训练语言模型常常表现出捷径学习的问题，其仅依赖数据中的表面特征（如个别关键词）来预测结果，只能反映局部统计偏差而缺乏因果逻辑支撑，导致模型在特定数据分布下表现良好，但在实际应用中（尤其是面对分布偏移或对抗性干扰时）性能显著下降，例如改动少数关键词则可能导致模型输出天差地别。所以，基金行业甚至是所有金融、科技公司共同面临的课题是，未来在技术上应重点研究解决模型捷径学习问题，以促进模型的泛化能力和可解释性提升。

（二）监管协同引导发展

1. 区分不同模式下平台方和部署方的责任

[22] ZHAO H Y,CHEN H J,YANG F,et al.Explainability for Large Language Models:A Survey[J].ACM Transactions on Intelligent Systems and Technology,2024,1,(2):1-38.

目前，众多基金公司并非自己开发AI大模型，而是选择接入头部AI公司开发的成熟产品，监管机构和行业组织应针对不同接入方式设置不同责任架构，总结如下表：

表1 生成式AI不同部署模式与责任架构

部署模式	技术特征	责任架构
官方API接入	最便捷的方式，简单直接	应对平台方提出较高披露要求，部署方责任可适当降低，但应协助提交说明
API-混合部署	部署方只需部署和维护当地的小模型，要进行数据脱敏再进入大模型	应相较于官方API接入模式提高部署方的披露责任
开源版本私有化部署	平台方中立，部署方拥有算法设计和再训练能力	应要求部署方承担更高程度的披露责任，平台方只承担协助说明义务
云服务商部署	云服务商已完成模型部署，部署方不可更改算法	应要求平台方和部署方各自承担所涉部分的披露责任，部署方责任应高于官方API模式，低于私有化部署模式
二次开发	部署方对大模型进行优化定制，储存并利用本地模型处理	应要求部署方承担更高程度的披露责任，与私有化部署类似

监管部门可充分考虑AI大模型平台方和基金公司等部署使用者的不同角色和特点，区分披露要求，制定差异化监管策略。例如，对于开源版本私有化部署接入模式，鉴于平台方的中立性，而部署方拥有算法设计和再训练能力，应要求部署方承担更高程度的披露责任，而平台方侧重于协助说明义务。

2.重点关注可信度评估

目前各国所提倡的AI算法透明与可解释性要求，以算法开源为主要方法。这一方法在算法透明的规范层次中属于“软法”[23]，受到商业秘密的制约，完全意义上的算法公开

[23] 安晋城.算法透明层次论[J].法学研究,2023,45(02):52-66.

很难实现。故目前可信性成为了可解释人工智能的检验标准。衡量可信性的指标主要是从系统中提取得到的可信证据。在评估可信性时，通常考虑训练数据的可信性、学习模型的可信性和预测结果的可信性[24]。（目前评估 AI 可信性方法的相关文献梳理见附表 5）

监管部门可引导行业构建一套既科学全面又合理可行的基金业人工智能系统可信度量评估模型，将度量评估转化为提升人工智能系统输出结果可信度的有效手段[25]。关于评估时点，可以要求基金公司在初次部署和每次重大更新生成式 AI 应用时，提交详细评估报告。还可参考欧盟《人工智能法案》第 86 条设置个体要求解释的权利。

五、结语

从理论逻辑看，可解释性监管的目的是平衡创新效率与金融安全。欧盟“基于风险”的分级监管、美国多元监管模式、中国香港“应用为本”的务实策略及新加坡灵活弹性的“软法”机制，虽路径各异，但均试图在商业秘密与监管透明之间寻找平衡点。这为构建差异化监管体系提供了镜鉴：避免“一刀切”式监管抑制技术创新，而是通过设计“面向用户”与“面向隐藏层”的二维框架，化解商业秘密保护与监管披露要求的内在冲突。

在实践层面，基金公司应主动在二维框架内加强可解释性能力建设：面向用户时，应恪守信息披露义务和算法风险告知义务；面向隐藏层时，应将合规风控要求嵌入算法设计

[24] 廖勇，韩小金，刘金林等．可解释性人工智能研究进展[J/OL].计算机工程,1-28[2025-06-16].https://doi.org/10.19678/j.issn.1000-3428.0069925.

[25] ZHOU Z H.Machine Learning[M].Beijing:Tsinghua University Press,2016.23-51.

全流程，在技术上优化模型捷径学习。监管部门可以进一步区分平台方与部署方的责任，重点关注可信度评估，通过政策引导与标准支持，推动形成具有行业特色的治理范式。

值得注意的是，生成式 AI 技术仍在快速迭代，系统复杂度持续提升，模型治理机制、跨机构数据协作与隐私保护框架等关键议题亟待深化研究。基金行业需立足中国特色金融文化的制度禀赋，在守正与创新中进一步构建可信、可靠、可解释的 AI 应用体系，以期持续夯实人民群众的金融信任基础，以自主技术革新不懈助力金融强国建设。

附件：AI 法规政策及评估 AI 可信性方法相关文献附表

【本文由兴证全球基金管理有限公司合规审计课题组供稿，供稿人声明本文为供稿人独立原创作品】

附件

AI 法规政策及评估 AI 可信性方法相关文献附表

附表 1 欧盟 AI 法规政策一览

时间	主体	文件
2018 年 4 月	欧盟委员会	《欧洲人工智能战略》
2018 年 5 月	欧盟	《通用数据保护条例》
2018 年 12 月	欧盟委员会	《人工智能协调计划》
2019 年 4 月	欧盟委员会人工智能高级专家组	《可信人工智能伦理指南》
2020 年 2 月	欧盟委员会	《人工智能白皮书》
2020 年 7 月	欧盟委员会人工智能高级专家组	《可信人工智能自评估清单》
2021 年 4 月	欧盟委员会	《关于制定人工智能底层监管规则的提案》
2022 年 7 月	欧盟委员会	《数字服务法案》
2022 年 9 月	欧盟委员会	《人工智能责任指令提案》
2024 年 5 月	欧盟	《人工智能法案》
2024 年 5 月	欧洲理事会	《人工智能与人权、民主和法治框架公约》
2025 年 2 月	欧盟委员会	《关于人工智能法案禁止的人工智能实践的指南》
2025 年 3 月	欧盟人工智能办公室	《通用目的人工智能业务守则》（三稿）

附表 2 美国 AI 法规政策一览

联邦层面	地方州层面
《关于安全、可靠和可信地开发和使用人工智能的行政命令》	加利福尼亚州《加州人工智能透明度法案》《生成式 AI 透明度法案》《年龄适当设计法》《自动决策系统问责法》
《删除法案》	纽约州《自动化就业决策工具法》《面部识别在学校禁令》
《反抗法案》	科罗拉多州《关于在人工智能系统交互中保护消费者权益的法案》
《算法责任法案》	华盛顿州《人脸识别服务法》
《无机器人老板法案》	伊利诺伊州《生物识别隐私法》《人工智能视频面试法》
《美国人工智能法的两党框架》	犹他州《人工智能政策法案》
《人工智能风险管理框架》	德克萨斯州《社交媒体内容审核法》《禁止

	深伪干预选举法》
《人工智能权利法案蓝图》	弗吉尼亚州《高风险人工智能开发者和部署者法案》
《国家人工智能研发战略计划》	阿肯色州《关于生成式人工智能工具生成的训练模型和内容所有权法案》

附表 3 中国香港 AI 法规政策一览

时间	主体	文件
2021 年 8 月	香港个人资料私隐公署	《开发及使用人工智能道德标准指引》
2024 年 6 月	香港个人资料私隐公署	《人工智能：个人资料保障模范框架》
2024 年 7 月	香港数字政策办公室	《人工智能道德框架》
2024 年 10 月	香港特区政府财库局	《有关在金融市场负责任地应用人工智能的政策宣言》
2025 年 3 月	香港个人资料私隐公署	《雇员使用生成式人工智能指引清单》
2025 年 4 月	香港数字政策办公室	《香港生成式人工智能技术及应用指引》

附表 4 新加坡 AI 法规政策一览

时间	主体	文件
2018 年 11 月	新加坡金融管理局	《促进新加坡金融业公平、道德、可问责和透明地使用人工智能和数据分析的原则》
2019 年 1 月	新加坡政府	《人工智能模型治理模式框架》
2019 年 11 月	新加坡政府	《国家人工智能战略》
2022 年 5 月	新加坡信息通信媒体发展局、个人数据保护委员会	《人工智能治理评估框架和工具包：AI Verify》
2023 年 12 月	新加坡政府	《国家人工智能战略 2.0》
2024 年 1 月	AI Verify 基金会、新加坡信息通信媒体发展局	《生成式人工智能治理的模型人工智能治理框架草案》
2024 年 5 月	新加坡政府	《生成式人工智能治理模型框架》

附表 5 评估 AI 可信性方法相关文献一览

可信证据类型	作者	验证方法
训练数据的可信性	GE L, GAO J, LI X Y, et al.	基于数据一致性和可靠性关联的方法
	TABIBIAN B, VALERA L, FARAJTABAR M, et al.	基于噪声评估数据的“时间痕迹”来建模的框架
	FOGLIARONI P,D'ANTONIO F,CLEMENTINI E.	基于版本更新的 VGI 质量指标量化模型
	ARDAGNA C A, ASAL R, DAMIANI E, et al.	一种创新的基于服务的可靠证据收集原子方法

	DISTERFANO S, Di GIACOMO A, MAZZARA M.	一个以车辆为核心的信息系统
学习模型的可信性	BAU D, ZHOU L. KHOSLA A, et al.	一种网络解剖的方法
	SLACK D, FRIEDLER S A, SCHEIDEGGER C, et al.	一项用户研究实验方法
	ROSENFELD A.	量化可解释人工智能模型
	SONG L W, MITAL P.	基于预测熵修正的推理攻击方法和新的细粒度隐私分析方法
	MA R, ZHAO Y Y, LIU Y W, et al.	SimHash 技术
预测结果的可信性	JHA S, Raj S, FERNANDES SL, et al.	一种适用于 DNN 的置信度度量方法
	T. H. Yang Zen, C. B. Hong, P. M. Mohan.	ABC-verify