DOI 10.15302/J-SSCAE-2025.08.012

数字金融场景中人工智能模型可解释性风险的特征与治理研究

张润驰[1]，岳中刚[1*]，张国法[2]，孙明明[3]，金磊[1]

（1. 南京邮电大学经济学院，南京 210023；2. 中国建设银行总行，北京 100033；
3. 中国证券监督管理委员会江苏监管局，南京 210019）

摘要：在数字金融快速演进的背景下，人工智能（AI）模型深度嵌入风险评估、资产定价与反欺诈等关键业务环节，其可解释性不足逐渐演化为制约金融安全与信任的重要风险源。本文旨在探索AI模型可解释性风险的成因、危害、识别与治理。研究发现，AI模型的可解释性风险主要源于算法结构的高复杂度、数据样本的隐性偏倚、建模目标与可解释监管目标的不一致以及模型迭代导致的解释失效。基于此，本文从金融稳定、社会包容、法律监管与技术安全4个维度系统揭示了AI模型可解释性风险的多层次危害，同时基于透明性量化、偏见识别、合规验证与安全检测为核心的识别思路，构建了AI模型可解释性风险的识别框架。最后，提出了一种涵盖模型工程优化、数据治理与特征管理、多方审计与监管协同、标准体系建设与责任界定的综合治理体系，以期在数字金融与AI的协同发展中实现技术效率、监管可控与社会信任的动态平衡。

关键词：数字金融；风险治理框架；可解释性风险；AI模型；可解释AI

中图分类号：F83；TP31　　**文献标识码：**A

Interpretability Risks of AI Models in Digital Finance Scenarios: Features and Governance

Zhang Runchi[1], Yue Zhonggang[1*], Zhang Guofa[2], Sun Mingming[3], Jin Lei[1]

(1. School of Economics, Nanjing University of Posts and Telecommunications, Nanjing 210023, China; 2. Head Office of China Construction Bank, Beijing 100033, China; 3. Jiangsu Office of China Securities Regulatory Commission, Nanjing 210019, China)

Abstract: In the context of the rapid evolution of digital finance, artificial intelligence (AI) models are deeply integrated into critical business processes such as risk assessment, asset pricing, and anti-fraud. The resultant lack of model interpretability has progressively become a significant source of risk, constraining financial stability and public trust. This study aims to comprehensively explore the causes, harms, identification, and governance of AI model interpretability risks. It finds that the interpretability risks of AI models primarily stem from the high complexity of algorithmic structures, implicit biases within data samples, inconsistency between modeling objectives and interpretability regulatory goals, and failure of explanations due to continuous model iteration. Building upon this, the study systematically reveals the multi-layered harms of AI model interpretability risks across four key dimensions: financial stability, social inclusion, legal compliance, and technical security. Concurrently, an identification framework for AI model interpretability risks is constructed, centered on the core methodology of transparency quantification, bias identification, compliance validation, and security detection. Finally, we propose a comprehensive governance system encompassing model engineering optimization, data governance and

收稿日期：2025-08-11；**修回日期：**2025-10-09

通讯作者：*岳中刚，南京邮电大学经济学院教授，研究方向为人工智能与数字金融；E-mail: yuezg@njupt.edu.cn

资助项目：国家自然科学基金项目(72401144)

本刊网址：sscae.engineering.org.cn

feature management, multi-party auditing and regulatory coordination, and construction of standards systems and responsibility delineation. This framework seeks to achieve a dynamic balance among technological efficiency, regulatory controllability, and social trust in the collaborative development of digital finance and AI.

Keywords: digital finance; risk governance framework; interpretability risk; artificial intelligence models; explainable artificial intelligence

一、前言

近年来，在诸如贷款信用风险评估、在线支付、反欺诈、智能投顾等代表性的数字金融场景中，人工智能（AI）模型的大规模应用持续提升了金融服务效率与智能化水平，中研普华集团的研究报告表明，截至2024年年底，我国数字金融用户规模达到9.6亿人，数字金融服务渗透率显著提升。但伴随而来的可解释性风险，即因AI模型存在“黑箱”特征（其内部决策过程复杂而不透明，难以直接被人类理解），进而导致金融决策失误、责任界定困难等一系列问题的不确定性[1]，引发了广泛的关注和担忧。较高的可解释性风险不仅使AI模型的输出结果缺乏可验证性，也掩盖了模型决策过程中的潜在偏差、误判与歧视性倾向[2]，最终削弱了数字金融机构对模型行为的可控能力，并影响公众对AI模型公平性、透明性及合法性的信任[3]。

针对上述问题，学术界与实务界近年来展开了诸多探索。在学术界，学者们重点探讨了AI模型可解释性风险的成因、危害、识别与治理，发现AI模型自身的复杂性与数据质量是导致可解释性风险的根源[4,5]，同时可解释性风险会带来用户信任缺失、系统安全风险、伦理冲突和法律不适配等问题[6,7]。可视化、反事实解释与特征重要性分析等技术，有助于识别可解释性风险[8,9]，此外可以通过开发具有内在可解释性的模型[10]、建立有效的监管框架[7]、使用一系列的可解释人工智能（XAI）方法，如局部可解释模型无关解释（LIME）与SHapley加性解释（SHAP）[11,12]进行事后解释等措施实现风险治理。但也有研究指出这些治理工具在稳定性、一致性与跨场景适用性上仍存在一定的不足[13,14]。随着大模型技术的崛起，一些研究也探索了针对性的解释技术以实现风险治理，代表性的成果包括利用大模型的自然语言生成能力，将复杂的AI模型决策转化为人类可理解的叙事性解释[15]。此外，面向金融时间序列分析场景，有研究探索以大模型的自动特征选择过程解释金融预测结果[16]；有学者提出了基于市场信息反馈的强化学习机制，使大模型驱动的金融交易策略能够适应不断变化的市场环境，同时通过内置解释机制保持透明度[17]；有研究构建了多智能体框架，通过大模型协调复杂的交易和投资组合管理任务，并强调了系统级可解释性的重要性[18]。

在实务界，不少数字金融机构探索在业务中使用XAI等方法提升模型的可解释性。例如，英格兰银行采用可解释梯度提升树明晰贷款违约模型的驱动因素[19]，招商银行的智能投顾结合自然语言处理模型解释投资组合的生成理由[20]，建设银行的智能客服模型对问答逻辑及业务依据进行解释[21]，微众银行的AI反洗钱模型输出特征贡献度以供监管机构复核[22]。此外，全球主要经济体的监管机构也逐渐关注模型的可解释性风险。美国在发布的《模型风险管理监督指南》（SR11-7）中要求金融机构必须深入理解模型的理论基础、设计思路和逻辑，此外还要进行持续监控和结果分析，以识别模型性能的偏差。欧盟在《通用数据保护条例》（GDPR）中明确了当AI决策对个人产生重大影响时，个人有权获得来自金融机构的关于该决策背后逻辑的解释。我国2022年印发的《关于银行业保险业数字化转型的指导意见》中明确指出要确保模型的可解释性和可审计性；2024年颁布的《银行保险机构数据安全管理办法》中第51条也强调模型算法投入使用前应当审查数据与模型的合理性、正当性与可解释性。

综上，尽管不少研究者近年来对AI模型的可解释性风险展开了一定探索，但大多面向一般性场景展开，对于数字金融场景中AI模型可解释性风险的异质性成因、危害、识别与治理的认识仍有待深化。同时尽管国内外数字金融机构与监管部门在实践中积累了不少有价值的经验，但对于如何将法律规范、行业标准与技术方法有效衔接，亦值得进一步探讨。

唯有打开AI模型的“黑箱”，构建更加公平、透明与可控的智能金融生态，方能推动数字金融的

高质量发展。在上述背景下，本文的研究内容主要包括3个方面：第一，厘清数字金融场景中AI模型可解释性风险的生成机制，揭示其成因的多维性与演化性；第二，分析其在金融稳定、社会公平、法律合规与技术安全方面的潜在危害，明确风险的表现形式；第三，提出包含模型研发与工程优化、数据治理与特征管理、多方审计与监管协同、完善法规与建设标准、加强沟通与明确责任的多维治理框架，以期为政策制定者、学术研究者与行业实践者提供理论参考和实践启示。

二、数字金融场景中AI模型可解释性风险的成因

AI模型在数字金融场景中的广泛应用极大提升了金融服务效率与智能化水平，但其可解释性风险却在复杂条件作用下积累与放大。深入剖析可解释性风险的成因，是后续识别与治理的前提。

（一）复杂且不透明的模型是工具性成因

AI模型的可解释性鸿沟从根本上说是模型技术问题[10]。数字金融背景下的大数据环境使得结构简单的传统模型如线性回归、逻辑回归等，难以捕捉复杂数据间的深层次关联。而结构复杂的AI模型如深度神经网络等，通过多层非线性变换，可以更好地拟合数据中的复杂模式，满足金融机构对模型精准度的需求，因而在实践中使用广泛[23,24]。然而，此类AI模型通常具有复杂的网络结构，涉及大量超参数设定、高阶特征交互与多层权重耦合，导致模型内部运行机制高度不透明。原始的数据在经过多轮映射与转换后，生成的新变量与原始金融语境中的概念联系被严重弱化甚至完全丧失，既难以对运算过程与变量的经济含义进行回溯解释，也无法识别决定输出内容的关键影响因素，导致“黑箱”效应愈发显著[25]。

（二）隐蔽的样本偏倚特征是数据性成因

AI模型的构建主要基于对海量历史样本的深度学习。然而，这些样本往往不可避免地携带着一些历史经济活动的偏倚性特征。此类特征庞杂且隐蔽，难以在训练AI模型前被全面人工识别并标记，AI模型会全盘进行学习并固化。例如，某一特定地域或职业群体由于历史原因信贷审批率较低，模型可能会在学习后将特定地域或职业作为负向预测因子，从而在未来决策中持续系统性地歧视该群体[26]。有研究发现样本选择偏差，如信用风险评估模型的训练数据集包含了更多的高收入群体样本[27]，会显著影响模型解释的稳定性。在偏倚样本环境下，SHAP值解释的一致性下降18%，且低收入群体的违约风险被系统性高估12.7%。该定量研究表明，数据偏倚直接提高了不同群体间的解释风险差异。同时在训练过程中，AI模型通过复杂的非线性函数关系将看似无关的特征与歧视性结果建立起关联，样本偏倚特征会转化为更为隐蔽的模型结构与参数组合特征[28]。因此，即使数字金融机构发现模型的输出结果具有歧视性，也难以快速定位并识别出歧视性结果的关键样本或特征组合，无法精准地调节复杂的模型结构与参数以快速消除歧视性问题。

（三）建模目标与解释需求的错位是制度性成因

AI模型的训练通常以预测精度或业务收益最大化作为建模目标，重视的是模型的经济效用，对可解释性特征的关注则被显著弱化，导致解释性设计在整体开发流程中处于边缘地位[29]。同时为了更好地适应大数据环境，模型设计者倾向于应用复杂的网络结构与多源高维特征融合策略，使得AI模型在训练过程中被动地牺牲了可解释性。此外，一些金融机构出于保护商业机密与保持竞争优势的考虑，对构建的AI模型细节信息保持高度封闭[30]，进一步削弱了外部评估者与监管者开展独立验证与解释的可能性。

（四）AI模型的高迭代特征是时效性成因

数字金融市场是一个高频变化的市场[31]，交易模式、用户行为与金融市场环境随时间变化频繁调整，这一特征在自动化交易、智能投顾与风险预警系统中尤为显著[32]。因此，AI模型的参数往往需要不断迭代更新。在每一次重新训练与迭代中，都可能引入新的决策路径，形成不同于原有逻辑的“黑箱”特征。这种动态演进过程使得模型的“黑箱”特性并非固定不变，而是在时间维度上不断累积和演化。尽管数字金融机构在某一固定时点可以通过特定的解释方法对AI模型进行剖析，但其有效性

会随着下一次迭代而迅速失效。新的模型版本可能因新的训练数据集或微调的建模目标，生成完全不同于前一版本的复杂特征权重，使旧有的解释框架无法复用。这种解释时滞导致可解释性风险在时间层面具有累积性。

三、数字金融场景中AI模型可解释性风险的危害

（一）金融稳定层面，AI模型的可解释性风险阻碍了数字金融体系的稳健运行

在数字金融时代，不透明的AI模型可能成为金融风险的放大器。在信用风险评估等核心数字金融业务场景中，当AI模型无法解释其决策逻辑时，若其内部存在功能缺陷或偏见，可能导致大规模的决策错误[33]，进而引发金融机构经营亏损，甚至诱发区域性或系统性金融危机。此外，AI模型的不可解释性也会加剧金融市场波动，在量化投资领域，多个机构的AI模型可能基于相似的、不可解释的逻辑同时做出一致性的交易决策，引发羊群效应[34]，导致市场价格在短时间内剧烈波动，甚至出现资产价格崩盘等极端事件[35]。最后，可解释性风险还会助长监管套利，金融机构可能利用AI模型的复杂性和不透明性，有意设计出难以被现有监管框架穿透和理解的产品或服务，从而规避监管审查，积累隐性风险，最终影响金融体系的整体稳定。

（二）社会包容层面，AI模型的可解释性风险削弱了数字金融的包容性

数字金融领域积累的历史样本，包含了社会与经济发展过程中形成的结构性偏见。AI模型在学习这些样本特征时，会无意识地将这些偏见内化，并将其视为预测违约风险的有效特征。研究发现，AI信用评分模型会对少数族裔或低收入群体产生偏见，即使其信用记录良好，也将面临更高的贷款利率或被拒绝贷款[36]。当AI模型做出决策时，其依据仅仅是基于学习到的数据相关性而非真正的因果关系，从而会掩盖其决策中隐含的歧视性[37]。这种偏见的内化与输出，使AI模型的决策看似客观，实则可能复制甚至加剧社会不公，且由于“黑箱”模型的不可解释性，偏见难以被高效率的识别、量化和纠正。长此以往，将引发公众对AI模型的信任危机，阻碍数字金融的包容性与健康发展。

（三）法律监管层面，AI模型的可解释性风险影响了经营合规性

“巴塞尔协议III”强调模型应具备透明性与可验证性，而AI模型的复杂结构与“黑箱”特征与此要求存在天然冲突。就当前全球范围内各主要经济体的监管法规实践来看，美国的《模型风险管理监督指南》（SR11-7）要求金融机构对模型假设、数据来源及结果进行可解释性验证；欧盟的《人工智能法案》和GDPR进一步规定，若算法决策对个人或机构权益产生实质影响，相关方有权获得“有意义的解释”；我国发布的《银行保险机构数据安全管理办法》《关于银行业保险业数字化转型的指导意见》亦要求模型可追溯、可审计。然而在实践中，AI模型在信贷审批、反欺诈监测等代表性数字金融场景中常出现解释性不足的问题。例如，先进的AI深度学习模型在信贷评分中可能因特征权重不透明，导致金融机构难以向监管部门说明具体决策依据；又如，AI反洗钱模型利用复杂网络结构识别可疑交易，但其风险标记逻辑难以复核。相关情形均可能被监管机构认定为模型治理不到位，从而引发风险[38]。

（四）技术安全层面，AI模型的可解释性风险使数字系统更加脆弱

可解释性不足会导致数字系统在面对恶意攻击时无法有效识别和防御。攻击者可能利用模型的复杂性进行对抗攻击，即通过微小、难以察觉的输入扰动，诱导AI模型做出错误决策[39]。例如，攻击者可以精心设计多个虚构的贷款人信息尝试申请贷款，直至发现能使欺诈识别模型失效、信用评分被恶意引导的样本模式，从而获得不合理的信贷批准。由于模型的决策过程无法被清晰解释，这种攻击难以被及时发现和防范。此外，AI模型在训练过程中可能过度拟合训练数据，导致其在面对新的输入样本时表现不佳。例如，在数字信贷审批场景中，若AI模型过度拟合了特定历史阶段的样本分布特征，可能对新经济周期中的贷款申请者的信用评估出现偏差[40]。由于模型缺乏可解释性，金融机构难以判断模型在新形势下是否依然可靠，加剧了

决策系统在面对未知状态时的脆弱性。

四、数字金融场景中AI模型可解释性风险的识别

由于数字金融场景中AI模型的可解释性风险是一种复合性和隐匿性并存的风险形态，因此需要对多层次的风险影响因素与表现因素分别展开分析。

（一）金融稳定层面，关注模型决策逻辑与输出一致性

梳理AI模型决策逻辑的可解释性，识别模型内部的“决策黑箱”区域。具体而言，可利用特征重要性分析、LIME或SHAP值分析等方法，对模型在各数字金融业务环节的关键特征贡献度进行检验。当模型的可解释特征占比过低或在不同样本间权重分布差异显著时，模型存在的可解释性风险往往也相对较大。

关注模型间的输出一致性，以防止系统性风险的积聚。在量化投资、智能投顾等代表性数字金融场景中，不同机构AI模型的输入特征、训练数据集及优化目标往往具有高度同质性。通过对模型输出行为的聚类分析与相关性检测，可以识别不同模型之间的同步性决策倾向。当模型在特定市场信号下表现出高度一致的交易反应，而其内部逻辑又缺乏可解释性支撑时，这种隐性一致性容易构成潜在的系统性金融风险。

建立跨模型的风险穿透识别机制。金融机构可通过可解释性评分、异常输出频率等监控指标，动态批量监测各场景下AI模型的异常行为模式。当特定模型在短时间内出现输出漂移、风险评估结果突变等情况时，应判定为可解释性风险暴露信号，并进行进一步的确认。

（二）社会包容层面，揭示模型输出中隐含的结构性偏见

数字金融机构应建立数据与特征层面的可解释性识别机制，利用包容性测度指标对AI模型输出进行分群对比分析，识别模型在性别、年龄、收入、地区等特征维度上的不平衡表现。当模型在不同群体间的预测结果出现显著差异，而其内部决策逻辑又难以说明这种差异的合理性时，即可认定其存在可解释性风险。

通过可解释的因果分析方法追踪偏见来源。针对信用评估、风险定价等高敏感的数字金融业务场景，通过剖析模型在特征选择、特征交互与输出映射过程中的因果路径，识别哪些变量可能在无意中承担了歧视性代理特征的角色。例如，当模型将职业类别、居住地址或社交数据隐含地作为信用风险指标时，应进一步判断这一做法的科学性，防止模型自动将相应指标与社会偏见形成关联。

强化模型输出的可验证性评价。通过设立人工对照组或基准规则模型，将AI模型输出与人工专家判断、传统评分模型结果进行比较，分析其偏差来源与合理性。当AI模型在某些社会群体中的预测偏差显著高于对照模型且缺乏可解释依据时，应将其列为包容性风险监测重点。此类识别方法有助于揭示算法歧视在模型“黑箱”结构中的隐蔽存在，为后续的包容性改进与透明性提升提供实证基础。

（三）法律监管层面，聚焦模型透明度、可追溯性与合规性

建立模型文档化与解释性披露识别机制。数字金融机构需对AI模型的假设前提、训练数据来源、算法架构、特征解释方法及验证结果形成系统记录，并定期进行可解释性披露。监管机构可通过审查模型说明文件与测试报告之间的一致性，识别出模型在信息披露中的不完整性或误导性，从而判断可解释性风险水平。

加强模型决策过程的可追溯性识别。对于涉及重大金融决策的AI模型（如信贷审批、反欺诈、反洗钱等），应通过日志留痕与可解释性报告体系，实现决策链条的溯源分析。若在追溯过程中无法明确界定模型决策的关键变量或逻辑路径，说明其在可解释性方面存在重大缺陷。监管部门可据此判定机构模型管理不到位，并形成合规性风险识别证据链。

依据前述的我国相关监管指引的要求，数字金融机构应定期开展AI模型可解释性的合规性审查，比较模型解释能力与法规标准之间的契合度。若模型未能提供可验证的决策依据、难以向客户说明拒贷原因或存在“自动化决策不透明”等情形，即可

判定为可解释性不足所引发的合规风险。

（四）技术安全层面，侧重模型稳健性、抗攻击性与泛化能力

采用对抗样本测试或敏感性分析方法，对AI模型的输入变量进行轻微干扰，观察输出结果的变化程度。当模型输出对微小扰动高度敏感，而模型内部无法提供稳定的解释性支撑时，说明其在安全性上存在可解释性风险点。

识别模型在迁移场景下的解释失效风险。金融机构可通过交叉时段与异质样本测试，验证模型在不同经济周期、地域市场或客户结构下的可解释稳定性。当模型解释结构在新数据集上显著变化，或原有特征贡献度失效时，应视为模型存在可解释性漂移，需纳入风险识别清单。

强化模型监测系统中的解释性预警机制。通过建立动态可解释性评分体系，对模型在运行过程中的解释一致性、特征稳定性和输出合理性进行量化监控。当模型可解释性指标下降或与历史基准偏离超过设定阈值时，即可触发风险预警信号。此类机制有助于在早期阶段识别模型异常行为，为防范安全漏洞与业务损失提供前置保障。

五、数字金融场景中AI模型可解释性风险的治理

AI模型在数字金融场景中的可解释性风险治理是一个复杂的系统性工程。以下从模型研发与工程优化、数据治理与特征管理、多方审计与监管协同、完善法规与建设标准、加强沟通与明确责任5个方面设计治理体系，治理框架如图1所示。

（一）模型研发与工程优化

在数字金融场景中，AI模型的研发过程往往直接决定了其可解释性水平，因此从源头环节介入是降低可解释性风险的关键。

研发阶段的治理应以不同场景下的可解释性需求为导向。一般而言，不同的数字金融场景对可解释性的要求差异显著。例如，信贷审批场景往往要求AI模型能够清晰追溯每一项风险评估指标的贡献，以确保决策的公平性与合理性；而反欺诈场景更强调AI模型对预警信号的触发机制说明；同时智能投顾场景则更加关注推荐的透明度与用户可理解性，需要将推荐逻辑直观呈现。因此，在研发阶段伊始需要对业务需求进行系统化梳理，并将其转化为模型开发的约束条件，避免在后续研发过程中因单纯追求预测精度而牺牲可解释性，造成“后置补救”的被动局面。

在数据清洗与特征工程处理阶段，治理的重点在于特征透明化与语义化。数字金融场景中的AI模型常依赖大规模异构数据，如交易流水、社交网络信息、历史行为数据等，这些数据中的特征一旦经过高维度非线性映射，往往难以解释其与预测结果之间的关系。因此，应在变量选择、转换与组合过程中，优先保留具有清晰业务语义的特征，并对高复杂度特征映射建立解释文档，使特征构建的逻辑链条可以被回溯[41]。

在算法设计阶段，应推动可解释性算法的优先应用。传统机器学习算法如逻辑回归、决策树等在可解释性方面具有天然优势，而深度神经网络、梯度提升树等复杂模型在预测精度上具备优势但透明性不足。在研发过程中，不应简单陷入性能与可解释性的二元对立，而应探索通过模型结构改进、约束正则化、特征可视化等手段实现兼顾。如可以通过模型蒸馏技术，将高复杂度模型的预测知识迁移至简化模型，从而在保留精度的同时获得较高的可解释性。

在模型训练与验证环节，需引入可解释性评估维度。传统AI模型训练往往以准确率、特异性、召回率等指标作为主要优化目标，无法全面反映模型在可解释性上的表现。因此，应建立一套覆盖全生命周期的可解释性评价体系，包括局部解释一致性、特征重要程度稳定性、跨群体公平性等维度，使训练过程不仅优化预测精度，还需满足可解释性指标的约束要求。同时，在验证环节应通过对比实验，系统评估不同解释方法的结果差异，避免因过度依赖单一解释指标而形成“解释假象”。

在模型开发与部署阶段，治理的重点在于增强工具化与自动化支撑。数字金融机构内部往往并行运行大量AI模型，人工逐一审查与解释的成本极高。因此，应在开发流程中嵌入自动化可解释工具，例如，建立统一的模型解释平台，实现对特征贡献度、局部决策路径的自动生成与可视化展示，不仅有利于提升解释模型的效率，还能在模型迭代

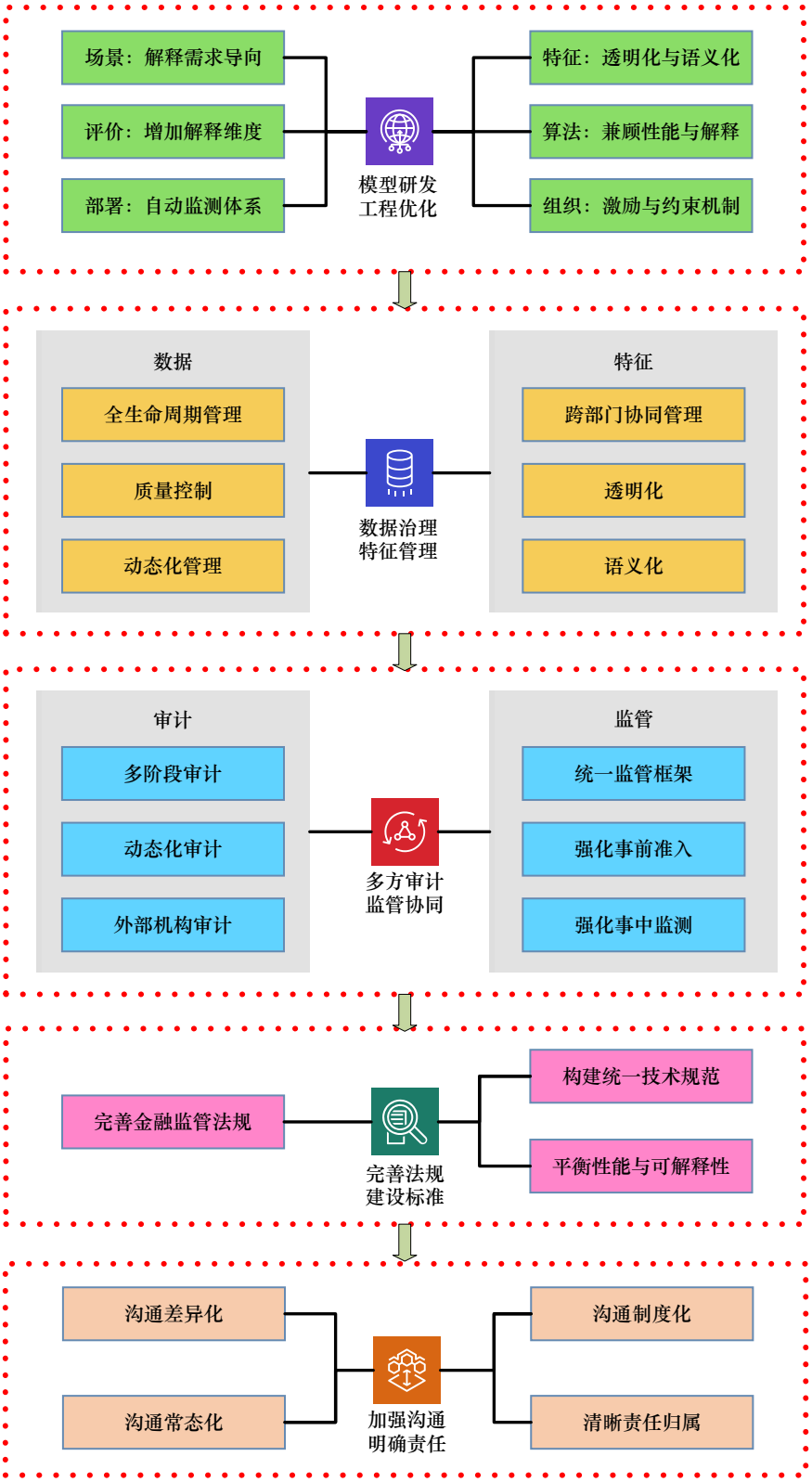

图1　数字金融场景中AI模型可解释性风险的治理框架

更新时提供动态对比。同时，自动化工具的引入还能标准化解释流程，避免因解释人员主观差异带来不一致性，提高可解释性评价的客观性。

最后，需要在组织层面建立激励与约束机制，

以确保可解释性目标的真正落地。若研发团队的考核仅以模型性能为导向，则可解释性往往会被边缘化。因此，应将可解释性纳入研发团队的绩效考核维度，在组织制度层面推动模型研发过程的可解释化。

（二）数据治理与特征管理

在数字金融场景中，AI模型的可解释性风险不仅源于算法结构的复杂性，也与其所依赖的数据基础密切相关。金融大数据的冗余与变量的语义模糊，都会使模型的决策逻辑难以被准确理解，从而放大可解释性风险。

数据治理的核心在于建立完整的数据全生命周期管理体系。数字金融机构若仅仅关注数据的规模与可用性而忽视其可解释性，就会导致在特征构建阶段出现信息语义断裂，使得后续模型难以溯源解释。对此，需要在数据采集时强化元数据标注，明确数据来源、采集时间、采集方法与质量等级[42]。此外，在存储环节应通过分级分类机制对数据进行结构化管理，清晰区分原始数据与衍生数据，从而在特征追溯时能够明确变量的属性。

在数据治理中需特别重视数据质量控制。低质量或缺乏代表性的数据会使模型在训练过程中出现数据迁就，导致结果难以解释。为此，数据治理流程必须引入多维度质量评估指标，包括完整性、一致性、准确性与时效性[43]，并通过自动化检测工具对数据集进行持续监测与清洗，确保AI模型所依赖的数据始终具备较好质量。

重视变量特征的透明化与语义化管理。金融模型往往涉及大规模高维特征，如账户行为、交易链路、社交关系、时空轨迹等，这些特征经过复杂的嵌套与组合后，虽然能够提升AI模型的预测性能，但也容易丧失与业务逻辑的直观关联，导致解释困难。对此，需要在变量构建阶段引入严格的语义定义流程，即每一个变量不仅要在数据层面明确其计算方式与生成逻辑，还要在业务层面建立与金融概念的对应关系。例如，对于交易频率变量应当明确其统计周期、覆盖范围及业务含义，以便在AI模型判别某用户的信用风险时，能够清楚说明高交易频率特征如何具体影响信用评价。通过语义化管理，特征的含义不再局限于技术表述，而能够以金融业务语言向监管者与用户传递解释，从而提升透明度。

此外，数据治理与特征管理需要在组织层面引入跨部门协同机制。数据科学团队虽然具备处理大数据与构建AI模型的能力，但在特征解释方面往往缺乏业务可解释性；金融业务部门能够理解数据的业务含义，却未必了解其在AI模型中的数学映射关系；监管部门则更关注数据使用是否合法合规以及解释结果是否符合标准要求，对业务与模型细节则知之甚少。若缺乏跨部门协作，数据与特征的管理极易陷入片面化，进而影响模型可解释性的全面性。因此，治理框架应推动建立跨学科特征管理委员会或工作小组，使数据治理、特征构建、模型解释与监管审查能够在同一平台上协同进行，确保可解释性信息在多部门间共享。

根据业务差异构建精细化的数据治理与特征管理机制。在贷款信用风险评估场景中，重点在于确保训练数据的完整性与公平性，应通过严格的样本平衡与特征去敏处理避免性别、地区等隐性偏倚扩散，保证信用因子解释的一致性。在支付场景中，重点在于通过多源数据交叉校验与异常值清理机制提升特征可信度，并对高频与低频特征进行分层管理。在反欺诈场景中，强调动态更新与多模态融合，应将设备地址、地理位置等非结构化数据纳入统一管理框架，通过特征重要性约束避免模型过度依赖某一类信号，提升可解释性与稳健性。在智能投顾场景中，需要通过特征可追溯机制保证客户风险偏好、投资经验等特征的来源与计算逻辑可验证，并建立模型输出与输入特征的映射规则，以便用户能够理解推荐背后的因果逻辑。

最后，还需考虑数据生命周期的动态性。金融业务环境变化频繁，数据分布随时间演化，特征的重要性与可解释性也并非一成不变[19,31]。例如，宏观经济波动可能导致某些特征的重要性显著上升，而另一些特征则失去预测力。此时，如果特征管理缺乏动态更新机制，模型解释可能会滞后于现实，形成“过时的解释逻辑”。因此，治理框架需要引入动态特征监控机制，对特征的有效性、相关性与重要性进行持续跟踪，并在发现漂移或失效迹象时及时调整或剔除特征，以确保解释逻辑与现实业务保持一致。

（三）多方审计与监管协同

金融行业的复杂性决定了模型应用涉及多重主

体，通过实施多方审计与监管协同，避免出现因监管缺位或责任断裂导致的系统性风险。

数字金融机构内部的审计流程应当涵盖AI模型研发、验证、部署与运行等多个阶段。在研发阶段，应审阅模型开发团队提供的可解释性评估报告，包括特征选择逻辑、解释方法选择及解释结果的稳定性检验等内容；在验证阶段，需对解释方法的有效性进行复核，确保解释结果与业务逻辑一致；在部署与运行阶段，应建立动态审计机制，定期抽样分析模型输出与解释的一致性，尤其在数据分布漂移或模型迭代更新后，必须重新评估其可解释性表现。

然而，单靠内部审计难以完全消解可解释性风险，尤其是在与业务利益冲突时，数字金融机构往往存在隐匿风险的动机。因此，需要引入独立的第三方审计机构作为外部监督力量。相应机构应依据统一的行业标准，对模型的特征重要性、局部解释一致性、跨群体公平性与解释稳定性进行评估，并出具可被监管机构采信的审计报告。

在行业监管层面，协同机制的构建至关重要。监管机构肩负着维护金融稳定、保护用户权益与推动金融科技健康发展的多重使命，而AI模型的可解释性风险恰恰涉及这些核心目标。对此，一方面，监管机构应推动形成可解释性评价的统一框架，包括指标体系、评估方法与合规要求，以避免各数字金融机构因解释标准混乱而各行其是[44]。另一方面，还需在事前准入与事中监测上加强监管力度。例如，对于涉及大规模信贷决策或高敏感性反欺诈场景的AI模型，监管部门可以要求数字金融机构在上线前提交详细的可解释性评估报告，并在运行过程中定期披露解释性指标，防止其因追求效率而牺牲解释性。

（四）完善法规与建设标准

当前，AI模型的可解释性缺乏统一的法律规制与技术标准体系，导致治理实践中存在制度空白与执行困境。因此，治理框架须对这两方面进行完善。

亟需完善针对AI模型可解释性的相关法规。现有的金融监管法规如《中华人民共和国商业银行法》《中华人民共和国证券法》《中华人民共和国数据安全法》等，更多强调数据安全、信息披露与审慎经营原则，但在模型可解释性方面缺乏明确规定。对此，立法机关与金融监管部门应探索增设针对AI可解释性的强制性条款。例如，应明确规定金融机构在采用AI模型时，必须建立可解释性评估机制并向监管机构定期提交解释性报告；要求金融机构在信贷、保险定价、投资顾问等代表性业务环节向用户提供简明而可理解的模型决策依据说明，并允许用户基于解释性结果行使异议权与申诉权。同时，结合不同金融业务场景的特征建立差异化的法规体系。在信贷场景中，法规应明确借款人有权获得基于核心特征的可解释说明，并禁止基于敏感属性（如性别、种族）的隐性差别化决策。在支付与反欺诈场景中，法规应要求金融机构保留操作日志记录，确保可追溯性，以满足司法审查与跨境合规的需求。在智能投顾场景中，则需强调投资建议的解释性与风险提示义务，规定金融机构必须提供可理解的推荐理由说明，而非仅给出最终推荐结果。通过分场景的法律条款设计，可有效实现因场景施策，在维护技术创新灵活性的同时，保障金融安全与用户权益。

治理必须依托于统一、科学且可操作的技术规范。目前，学界与业界在可解释性方法的研究中已提出局部可解释模型（如LIME、SHAP）、全局可解释框架（如可视化特征重要性分析）、因果推断等多种方法[11,12]，但这些方法的可靠性、可重复性与可对比性仍存在争议，尚缺乏权威的评价标准，使数字金融机构在实践中面临解释方法选择随意、解释结果不一致等问题，进而削弱解释效力与市场信任。例如，有研究发现在信用风险评估场景中[45]，SHAP和LIME在识别具有相似风险特征的观察值群体时，其对特征重要性的识别结果存在显著差异。对于同一笔拒绝的贷款，SHAP将“负债收入比”列为首要原因，而LIME更强调“企业年龄”的贡献。进一步比较两种解释器在特征权重空间上形成的聚类结果发现，它们的归类标签一致性通常在0.5~0.65之间波动，低于0.7，即一致性评分为中低度，表明模型给出的解释高度依赖所选的解释器。因此，应尽快建立覆盖模型开发、验证、部署与评估全过程的可解释性技术标准体系，明确不同类型AI模型的适用解释性指标、评估方法与合规要求。同时，应充分考虑业务类型与属性的差异，建立分场景的可解释性评估技术标准。在贷款信用

风险评估场景中，技术标准可围绕模型输入透明度、输出可追溯性及决策一致性展开，要求模型在提供评分结果的同时，明确列示核心驱动因素，并设定可量化阈值。在反欺诈场景中，技术标准应转向模型对异常交易的判别逻辑是否具备可审计性。在智能投顾场景中，技术标准需关注推荐逻辑的透明度及投资者的可理解度，要求AI模型解释输出以风险等级和关键参数呈现，而非单纯的“黑箱”化预测。通过分场景设定差异化的技术标准，可以在保证AI模型性能的同时，使解释性评估具备可操作性与针对性。

技术标准建设还需注重可解释性与性能之间的动态平衡。AI模型的复杂性提升了预测性能，但可解释性要求的提高可能导致模型性能下降。在标准制定中，应根据不同业务的风险容忍度，设定合理的解释性与性能的平衡阈值。例如，在消费者信贷审批场景中应优先保证解释性，即使牺牲部分预测精度；而在金融资产价格预测分析中，则可容许一定程度的“黑箱”特征，只要模型能够在关键节点提供必要的解释机制。通过标准化、定量化地设定平衡阈值，为数字金融机构提供明确的合规导向，避免过度追求性能或过度强调解释性。

（五）加强沟通与明确责任

充分沟通和提升不同主体对AI决策过程的理解与信任程度，当风险事件发生时能够清晰问责，是提升AI模型可解释性风险治理效果的重要举措。

加强沟通的核心在于消弭技术壁垒与信息不对称，使用户能够知悉AI模型的运行逻辑与潜在风险。数字金融机构常常将AI模型视作商业机密，导致外部主体难以获知模型的决策逻辑。对此，监管部门应督促数字金融机构建立差异化的沟通机制，在保证核心商业利益的前提下，向不同层级的受众提供异质性解释，实现信息传递的适配性与透明度的均衡。面向普通用户，应以通俗易懂的科普性语言阐释AI模型决策中最关键的逻辑链条，避免技术性描述过度复杂，降低信息壁垒；面向行业研究者与专业媒体，可开放更高层次的解释信息，如特征权重、决策路径与模型鲁棒性指标，以便社会监督力量进行独立评价；面向监管机构，则需提供完整的模型透明档案，包括训练数据特征、算法选择理由、解释性方法等，以便开展审慎监管。

在外部沟通过程中，还要注重开拓沟通渠道，形成沟通制度化与常态化。传统的金融信息披露多以年度报告或定期公告形式进行，而AI模型的快速迭代特性决定了其可解释性风险具有实时动态性。因而，沟通机制应当更加高频与灵活，可以通过在线平台、移动应用终端以及交互式解释工具实现即时反馈。例如，在信贷审批环节，消费者提交申请后应能够即时获得决策结果及简要解释，并在必要时进入更深入的解释层级；在投资顾问业务中，用户应能够实时查询推荐结果的依据与风险提示，而非仅依赖事后的总结报告。此外，金融机构可以在官方网站设立解释专区，定期发布模型更新信息与解释性描述，形成面向公众的长期沟通渠道。

然而，仅有外部沟通并不足以彻底化解可解释性风险，还必须清晰责任归属以保障治理框架的有效运作。AI模型的研发与运行往往涉及多方主体，包括模型开发者、数据提供商、金融机构运营部门等，当模型结果产生偏差或造成损害时，若责任界定模糊，极易出现无人担责的局面。为避免责任真空，应当明确不同主体在可解释性风险治理中的责任边界[46]。对于模型开发者而言，其责任应集中在提供符合可解释性标准的AI模型工具，确保可解释性方法具有科学性与可验证性。对于金融机构而言，其责任不仅包括在部署AI模型时验证解释性指标是否达标，还包括在业务操作中履行告知义务与纠错义务，即当AI模型输出结果存在明显偏差或引发用户争议时，必须及时提供解释并采取补救措施。对于数据提供方而言，其责任在于确保数据质量与合法性。通过清晰划分多主体责任，使每个环节都具备明确的责任承担者。

六、结语

本文面向数字金融场景，系统探讨了AI模型可解释性风险的生成机制、危害表现与识别方法，并构建了涵盖模型研发与工程优化、数据治理与特征管理、多方审计与监管协同、完善法规与建设标准、加强沟通与明确责任五大环节的系统性治理框架。相关成果旨在为学术研究与业务实践提供参考，为基于AI技术规划并构建更加稳健与可持续的数字金融生态贡献力量。

展望未来，可以在以下几个方面展开进一步的探索：一是探索在深度学习、强化学习等高复杂度模型框架下，在保持预测性能的同时实现更高水平的可解释性，并构建可度量的解释性评估指标体系；二是推动跨学科融合，引入因果推断、博弈论及人机交互等新兴方法，突破现有事后可解释工具的局限，发展兼具内生透明性与动态适应性的解释技术；三是关注可解释性与隐私保护、数据安全之间的潜在矛盾，探索联邦学习等技术与可解释性方法的融合路径。

利益冲突声明

本文作者在此声明不存在任何利益冲突或财务冲突。

Received date: August 11, 2025; **Revised date:** October 9, 2025

Corresponding author: Yue Zhonggang is a professor from the School of Economics, Nanjing University of Posts and Telecommunications. His major research fields include artificial intelligence and digital finance. E-mail: yuezg@njupt.edu.cn

Funding project: The National Natural Science Foundation of China Project (72401144)

参考文献

[1] Li X H, Xiong H Y, Li X J, et al. Interpretable deep learning: Interpretation, interpretability, trustworthiness, and beyond [J]. Knowledge and Information Systems, 2022, 64(12): 3197–3234.

[2] Mhlanga D. The role of big data in financial technology toward financial inclusion [J]. Frontiers in Big Data, 2024, 7: 1184444.

[3] Thekdi S, Aven T. Understanding explainability and interpretability for risk science applications [J]. Safety Science, 2024, 176: 106566.

[4] Ding W P, Abdel-Basset M, Hawash H, et al. Explainability of artificial intelligence methods, applications and challenges: A comprehensive survey [J]. Information Sciences, 2022, 615: 238–292.

[5] Arrieta A B, Díaz-Rodríguez N, Ser J D, et al. Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI [J]. Information Fusion, 2020, 58: 82–115.

[6] Leben D. Explainable AI as evidence of fair decisions [J]. Frontiers in Psychology, 2023, 14: 1069426.

[7] Eshete B. Making machine learning trustworthy [J]. Science, 2021, 373(6556): 743–744.

[8] Lundberg S M, Erion G, Chen H, et al. From local explanations to global understanding with explainable AI for trees [J]. Nature Machine Intelligence, 2020, 2(1): 56–67.

[9] Wachter S, Mittelstadt B, Russell C. Counterfactual explanations without opening the black box: Automated decisions and the GDPR [EB/OL]. (2017-11-01)[2025-07-10]. https://arxiv.org/abs/1711.00399.

[10] Rudin C, Chen C F, Chen Z, et al. Interpretable machine learning: Fundamental principles and 10 grand challenges [EB/OL]. (2021-03-20)[2025-07-10]. https://arxiv.org/abs/2103.11251.

[11] Ribeiro M T, Singh S, Guestrin C. "Why should I trust you?": Explaining the predictions of any classifier [R]. San Francisco: The 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2016.

[12] Lundberg S M, Lee S I. A unified approach to interpreting model predictions [R]. Long Beach: The 31st International Conference on Neural Information Processing Systems, 2017.

[13] Visani G, Bagli E, Chesani F, et al. Statistical stability indices for LIME: Obtaining reliable explanations for machine learning models [J]. Journal of the Operational Research Society, 2022, 73(1): 91–101.

[14] Zaeem R N, Barber K S. The effect of the GDPR on privacy policies: Recent progress and future promise [J]. ACM Transactions on Management Information Systems, 2020, 12(1): 1–20.

[15] Martens D, Hinns J, Dams C, et al. Tell me a story! narrative-driven XAI with large language models [J]. Decision Support Systems, 2025, 191: 114402.

[16] Feng M K, Gu J J, Qiu J, et al. From news to forecast: Integrating event analysis in LLM-based time series forecasting with reflection [R]. Vancouver: Advances in Neural Information Processing Systems 37, 2024.

[17] Cao Y P, Chen Z, Cui Z Y, et al. FinCon: A synthesized LLM multi-agent system with conceptual verbal reinforcement for enhanced financial decision making [R]. Vancouver: Advances in Neural Information Processing Systems 37, 2024.

[18] Wang H P, Yang Z J. A multi-agent approach to investor profiling using large language models [R]. Barcelona: 2025 International Conference on Control, Automation and Diagnosis (ICCAD), 2025.

[19] Bank of England. Machine learning explainability in finance: an application to default risk analysis [EB/OL]. (2019-08)[2025-07-10]. https://www.bankofengland.co.uk/-/media/boe/files/working-paper/2019/machine-learning-explainability-in-finance-an-application-to-default-risk-analysis.pdf.

[20] 招商银行. 智能投顾资产配置解释系统 [EB/OL]. (2025-08-27)[2025-08-28]. https://www.finebi.com/blog/article/68ae76dc28946ecca8878efc.
China Merchants Bank. Intelligent investment advisory asset allocation explanation system [EB/OL]. (2025-08-27)[2025-08-28]. https://www.finebi.com/blog/article/68ae76dc28946ecca8878efc.

[21] 中国建设银行. 智能客服系统解释框架[EB/OL]. (2025-08-27)[2025-08-28]. https://www.finebi.com/blog/article/68ae76dc28946ecca8878efc.
China Construction Bank. Interpretation framework for intelligent customer service system [EB/OL]. (2025-08-27)[2025-08-28]. https://www.finebi.com/blog/article/68ae76dc28946ecca8878efc.

[22] 微众银行. 联邦反洗钱可解释模型 [EB/OL]. (2024-08-14)[2025-08-28]. https://cloud.baidu.com/article/3324961.
WeBank. Federated anti-money laundering explainable model [EB/OL]. (2024-08-14)[2025-08-28]. https://cloud. baidu. com/article/3324961.

[23] Talaat F M, Aljadani A, Badawy M, et al. Toward interpretable credit scoring: Integrating explainable artificial intelligence with deep learning for credit card default prediction [J]. Neural Computing and Applications, 2024, 36(9): 4847–4865.

[24] 刘璇, 张蕾, 刘钟, 等. 基于数据挖掘技术的企业信用风险评估 [J]. 山西财经大学学报, 2023, 45(S2): 89–91.
Liu X, Zhang L, Liu Z, et al. Enterprise credit risk assessment based on data mining technology [J]. Journal of Shanxi University of Finance and Economics, 2023, 45(S2): 89–91.

[25] Hakkoum H, Idri A, Abnane I. Global and local interpretability techniques of supervised machine learning black box models for numerical medical data [J]. Engineering Applications of Artificial Intelligence, 2024, 131: 107829.

[26] 刘朝. 算法歧视的表现、成因与治理策略 [J]. 人民论坛, 2022 (2): 64–68.
Liu C . Manifestations, causes and governance strategies of algorithm discrimination [J]. People's Tribune, 2022 (2): 64–68.

[27] Ford J. Fairness in focus: quantitative insights into bias within machine learning risk evaluations and established credit models [J]. Management System Engineering, 2025, 4(1): 8.

[28] 罗熠琛, 周新衡. 算法自动化决策中就业性别歧视的类型与规范路径 [J]. 中国人力资源开发, 2025, 42(1): 92–103.
Luo Y C, Zhou X H. Types and normative pathways of employment gender discrimination in algorithmic automated decision-making [J]. Human Resources Development of China, 2025, 42(1): 92–103.

[29] Weber P, Carl K V, Hinz O. Applications of explainable artificial intelligence in finance—A systematic review of finance, information systems, and computer science literature [J]. Management Review Quarterly, 2024, 74(2): 867–907.

[30] Abbasi W, Mori P, Saracino A. Trading-off privacy, utility, and explainability in deep learning-based image data analysis [J]. IEEE Transactions on Dependable and Secure Computing, 2025, 22(1): 388–405.

[31] 陈贞竹, 李力, 余昌华. 我国货币政策传导效率及信号效应研究——基于金融市场高频识别的视角 [J]. 经济学(季刊), 2023, 23(1): 56–73.
Chen Z Z, Li L, Yu C H. Transmission of monetary policy shocks and signaling channels in China—Based on high frequency data in financial markets [J]. China Economic Quarterly, 2023, 23(1): 56–73.

[32] 许荣, 田文涛, 胡学峰. 量化交易的风险影响与监管体系构建研究 [J]. 金融监管研究, 2025 (2): 74–92.
Xu R, Tian W T, Hu X F. Research on the impact of quantitative trading risks and the construction of the regulatory system [J]. Financial Regulation Research, 2025 (2): 74–92.

[33] Zhang R X. Toward interpretable machine learning: Evaluating models of heterogeneous predictions [J]. Annals of Operations Research, 2025, 347(2): 867–887.

[34] 郑挺国, 葛厚逸. 中国股市羊群效应的区制转移时变性研究 [J]. 金融研究, 2021 (3): 170–187.
Zheng T G, Ge H Y. A study of the time-varying characteristics of herding effects in China's stock market based on a regime-switching model [J]. Journal of Financial Research, 2021 (3): 170–187.

[35] 荆思寒, 王振山, 隋聪, 等. 股票间的风险传染——基于对股价崩盘风险的预测 [J]. 系统工程理论与实践, 2022, 42(11): 3090–3104.
Jing S H, Wang Z S, Sui C, et al. The risk contagion among stocks—Based on the prediction of crash risk on stock price [J]. Systems Engineering—Theory & Practice, 2022, 42(11): 3090–3104.

[36] Garcia A C B, Garcia M G P, Rigobon R. Algorithmic discrimination in the credit domain: What do we know about it? [J]. AI & Society, 2024, 39(4): 2059–2098.

[37] Cui P, Athey S. Stable learning establishes some common ground between causal inference and machine learning [J]. Nature Machine Intelligence, 2022, 4(2): 110–115.

[38] 刘艳红. 人工智能的可解释性与AI的法律责任问题研究 [J]. 法制与社会发展, 2022, 28(1): 78–91.
Liu Y H. Research on the interpretability of artificial intelligence and the legal responsibility of AI [J]. Law and Social Development, 2022, 28(1): 78–91.

[39] 李青青, 张凯, 李晋国, 等. 基于集成学习的入侵检测系统对抗攻击检测 [J]. 计算机工程与设计, 2025, 46(3): 850–856.
Li Q Q, Zhang K, Li J G, et al. Adversarial attacks detection in intrusion detection systems with ensemble learning [J]. Computer Engineering and Design, 2025, 46(3): 850–856.

[40] Mo G L, Zhang G L, Tan C Z, et al. Reassessment of corporate credit risk identification: Novel discoveries from integrated machine learning models [J]. Computational Economics, 2025, 66(4): 2791–2841.

[41] Toy T. Transparency in AI [J]. AI & Society, 2024, 39(6): 2841–2851.

[42] 储节旺, 夏莉. 嵌入生命周期理论的科学数据管理体系构建研究——牛津大学为例 [J]. 现代情报, 2020, 40(10): 34–42.
Chu J W, Xia L. Research on scientific dta management construction based on life cycle theory [J]. Journal of Modern Information, 2020, 40(10): 34–42.

[43] Al-Okaily M, Al-Okaily A. Financial data modeling: An analysis of factors influencing big data analytics-driven financial decision quality [J]. Journal of Modelling in Management, 2025, 20(2): 301–321.

[44] 邓胜利, 丁威威, 汪璠, 等. "工具–结构"视角下国内外生成式人工智能监管政策比较研究 [J]. 信息资源管理学报, 2025, 15(1): 54–68.
Deng S L, Ding W W, Wang F, et al. A multinational comparative study of regulatory policies for generative AI from the perspective of "tools–structure" [J]. Journal of Information Resources Management, 2025, 15(1): 54–68.

[45] Gramegna A, Giudici P. SHAP and LIME: An evaluation of discriminative power in credit risk [J]. Frontiers in Artificial Intelligence, 2021, 4: 752558.

[46] Dastani M, Yazdanpanah V. Responsibility of AI systems [J]. AI & Society, 2023, 38(2): 843–852.