人工智能与金融安全：技术脆弱性、风险演化路径与防范机制

丁少斌　汪红驹

［摘要］随着人工智能在金融领域的深度嵌入，产生的新型金融风险问题日益凸显。当前AI的技术缺陷在金融实践中表现出五类关键技术脆弱性：解释性不足延误风险溯源、数据偏误产生算法歧视、对抗易感性放大入侵风险、模型漂移导致决策失效及算法共振诱发市场同步风险。这些技术脆弱性通过自动化决策偏误、算法歧视外溢、集体性误判和基础设施依赖等路径传导，呈现出由局部技术失效向系统性金融风险演化的链式机制。AI模型的判断偏差容易被自动化流程放大，形成信贷排斥和资产错配。集体误判和同步行为进一步加剧了市场价格的非理性波动，破坏金融定价与风险识别机制。关键金融基础设施中的深度AI嵌入，可能放大单点故障风险，对底层系统架构造成冲击，进而引发服务瘫痪与市场中断。据此，需从技术、制度与治理三个层面构建以技术安全为核心、算法透明与审计规范为基础、责任边界清晰为保障的多维防控体系，以实现人工智能与金融安全的可持续协同。

［关键词］人工智能；金融风险；技术脆弱性；风险演化；防范机制
［中图分类号］F830.91 ［文献标识码］A ［文章编号］1006—012X（2026）—02—0027（10）

［作者］丁少斌，博士研究生，中国社会科学院大学应用经济学院，北京，102488
汪红驹，研究员，中国社会科学院财经战略研究院，北京，100006

一、引言

近年来，随着人工智能（Artificial Intelligence，AI）技术的飞速发展，其在金融领域的应用场景不断拓展，从风险识别、智能投顾到信贷审核、欺诈检测，正逐步重塑金融市场的运行逻辑和机构行为。这一趋势在中国也表现得尤为显著。各级政府近年来多次强调要推动金融与科技深度融合，将人工智能等前沿技术视为提升金融效率、优化资源配置与强化风险管控的重要工具。国务院2025年8月印发的《国务院关于深入实施“人工智能+”行动的意见》（国发〔2025〕11号）明确提出“人工智能+”产业发展行动，创新服务业发展新模式，在金融领域推动新一代智能终端、智能体等广泛应用。与此同时，金融安全问题长期以来受到党中央的高度重视。习近平总书记多次强调“要把防范化解金融风险放到更加重要的位置”“牢牢守住不发生系统性金融风险底线”，有效防范化解金融风险特别是系统性金融风险是推动金融高质量发展的根本保障。

当前学术界与实务界高度关注人工智能在金融领域的应用价值与场景优化，认为AI具有提高信贷识别精度、增强投资组合回报、降低欺诈检测成本等方面的潜力。[1,2] 这些研究强化了AI“效率工具”的形

基金项目：研究阐释党的二十届三中全会精神国家社会科学基金重大专项项目“生成式人工智能发展规律和管理机制研究”（24ZDA085）；国家社会科学基金重点项目“促进实体经济和数字经济深度融合的理论机制与实践路径研究”（24AZD066）；中国社会科学院智库基础课题（ZKJC240903）。

中国知网 https://www.cnki.net

象，推动了“金融AI赋能”的正向叙事。然而，在金融科技迅速发展的背景下，人工智能技术也引入了一系列新的风险类型与监管挑战。人工智能技术在金融领域的快速渗透，不仅推动了金融服务模式的数字化转型，也深刻改变了风险形成与传导的基本逻辑。伴随其广泛部署，金融系统也日益暴露出由算法控制、数据驱动与模型决策所带来的新型技术风险，其复杂性、隐蔽性与系统性程度显著高于传统技术或操作风险。[3,4] 尽管国内部分研究开始注意到AI技术可能放大现有金融风险，如伦理风险、[5] 数据与操作风险、[6] 算法垄断风险等，[7] 但对于“技术缺陷—局部风险—系统性风险”之间的多级传导机制尚未形成结构化解释框架。在上述背景下，本文聚焦于AI金融系统的“技术脆弱性”与其“风险演化路径”，力图回答如下问题：人工智能技术缺陷导致其在金融实际运行中面临哪些技术脆弱性？这些技术脆弱性如何沿既定路径演化为系统性金融不稳定？如何从技术、制度与组织层级构建针对性防范机制？已有文献多关注人工智能技术在金融领域的应用价值，对于技术缺陷的关注不足。少部分关注技术缺陷的研究停留在技术自身层面，亦或只关注技术缺陷导致的单点金融风险，而对于技术缺陷与金融系统的关联，单点技术脆弱性向系统性风险演化的机制研究尚有待补充。本文将人工智能目前的技术缺陷延伸到金融领域的技术脆弱性，系统归纳AI金融系统中的五类核心技术脆弱性，并剖析其在金融嵌入中的具体表现与放大机制，从而在机制层面提出具有代表性的风险演化路径，系统揭示技术脆弱性如何沿链条传导并放大至系统性金融风险。在此基础上，本文在治理层面构建多层级治理框架回应现有治理体系碎片化与适应性不足的挑战。立足人工智能系统性嵌入金融领域的新现实，尝试建构“AI金融安全”研究框架，填补当前研究中过于聚焦应用而忽视风险机制的问题，为AI时代金融系统的稳健发展提供学理支撑与政策建议。

二、人工智能在金融领域中的技术脆弱性

人工智能的广泛应用正在重塑金融系统的运行逻辑，但与此同时，其底层技术的结构性缺陷也使金融系统面临前所未有的风险挑战。本文将“技术脆弱性”界定为：由于人工智能系统的内生技术限制与外部数据环境不完备，在高度复杂、动态反馈性强的金融应用中所引发的功能失效、误判放大与结构性风险积累现象。人工智能技术的核心能力源自以深度学习为代表的数据驱动型算法。这类算法依赖大规模训练数据与复杂模型结构来学习模式、生成预测，并在此基础上自动化决策。然而，这种“高度非线性+非透明性”的决策结构，虽然显著提升了处理复杂任务的能力，但也引入了多重内在技术缺陷。理论上，这些技术缺陷主要包括以下几个方面：一是算法黑箱性。深度学习模型如神经网络虽然具备强大的拟合能力，但其内部权重结构和决策路径高度复杂，缺乏可解释性，难以满足问责、合规与调试要求。二是数据依赖性。AI模型严重依赖数据分布的稳定性与多样性。一旦训练数据存在偏误、污染或样本不足，模型输出将不可避免地体现出结构性偏差。三是对抗易感性。AI系统对输入微扰极度敏感，对抗样本攻击可轻易诱导模型产生错误判断。四是外生参数依赖性。AI模型本质是对训练分布的映射器，训练过程和预测结果完全依赖于外生参数，而金融系统中的政策、市场与行为变量常发生突变，容易导致模型“脱离环境”、预测失灵。五是模型结构趋同性。AI技术标准化、工具链通用化导致模型架构趋同，易在集体应激状态下引发“同步性错误”，形成技术放大机制。上述缺陷原本可通过工程优化在一般商业应用中加以控制，但一旦嵌入金融系统，其风险结构将发生质变。金融系统具有高杠杆、高关联、高反馈的内在结构，AI系统若作为核心决策机制被深度嵌入，将不再只是“工具性风险”，而可能演化为具备系统性放大能力的结构性风险节点。人工智能的技术局限，在金融场景中因任务敏感性、操作自动化与行为耦合性而放大为多种“技术脆弱性”表现。这些脆弱性并非简单的“算法故障”，而是金融系统运行机制被“算法化”后的失衡结果，具有长期性、耦合性与演化性特征。本文将AI金融系统中的技术脆弱性归纳为五类典型形式，并围绕其生成机制、金融表现与风险逻辑展开分析。

1. 解释性不足延误风险溯源

人工智能技术在金融领域的广泛应用大幅提升了风险识别、信贷审核与市场交易的效率。然而，这

中国知网　https://www.cnki.net

表1 人工智能技术脆弱性的主要类别

技术脆弱性类别	主要表现	典型案例	风险机制
解释性不足延误风险溯源	模型运行机制不透明，难以解释决策结果，限制审计与合规	澳大利亚联邦政府解释性不足的自动化债务算法导致错误债务难以及时被纠正	降低金融透明度与可问责性，阻碍风险识别与纠偏
数据偏误产生算法歧视	训练数据存在结构性偏差、标签错误与样本选择性，产生歧视性决策	美国“种族歧视”衍生的历史数据在信用筛选模型中导致信贷歧视	将社会不公内化为模型规则，引发信任危机与服务排斥
对抗易感性放大入侵风险	通过微扰输入、模型逆向工程或数据投毒手段误导AI判断	对信用卡反欺诈检测系统进行对抗性攻击导致伪造信用卡识别失效	攻击传播性强，可能在系统间扩散
模型漂移导致决策失效	数据分布或变量关系随市场变化而改变，模型预测准确性降低	新冠疫情冲击导致Zillow的房地产定价算法失效带来大幅的资产减值	多个机构集体依赖失效模型，可能诱发同步性风险事件
算法共振引发交易同步化	算法对相似信号做出趋同反应，引发价格波动或市场脱锚	高频自动化卖单导致的英镑闪崩事件	模型结构同质化导致集体性误判，放大系统性联动风险

些基于复杂算法的系统往往具备高度的“黑箱性”，即其内部运行机制对使用者、监管者甚至开发者本身而言都难以完全解释和追溯。这一技术特性，在提升模型预测能力的同时，也带来了显著的法律、伦理和金融安全风险。

具体而言，模型不可解释性使金融决策过程丧失了透明性与可问责性。以信贷审批为例，若AI模型因某些高度非线性、不可观测变量而拒绝借款申请，用户将难以了解其拒贷原因，亦无法申请复核或申诉。进一步来看，模型不透明性限制了金融机构自身的风控能力。许多深度神经网络、梯度提升树等算法虽能实现较高的预测精度，但由于其参数维度较高，决策路径复杂，在模型失效时金融机构难以快速定位故障来源。如，澳大利亚联邦政府2015年推行的“自动化债务”计划，以收入平均自动化算法黑箱运行、透明度不足，致错误债务多年后才被系统性揭示与赔付，皇家委员会最终认定方案违法并要求强化可解释与问责。与此同时，AI模型的黑箱性阻碍了监管介入与合规审查，形成了所谓的“技术审计盲区”。目前多数监管技术工具本身尚未普遍具备处理复杂机器学习模型的能力。这使得AI模型在面对模型失效、风险转嫁或算法操纵时，缺乏外部可验证机制。

黑箱模型与不可解释性是人工智能在金融系统中所面临的首要技术脆弱性。其风险不仅表现在模型本身的行为不透明，也体现在由此导致的法律责任归属不清、市场信号误判与监管机制失效。

2. 数据偏误产生算法歧视

人工智能模型的性能高度依赖于数据的质量、广度与代表性。然而，金融领域的数据体系往往存在结构性偏误与污染问题，进一步放大了AI系统在应用过程中的脆弱性。数据风险不仅可能误导模型判断，更可能加剧金融服务中的歧视性结果与资源配置失衡，最终演化为系统性金融不公和信任危机。

从数据偏误的角度看，历史数据中的结构性偏差易导致“算法歧视”。如，美国金融机构使用内置机器学习的信用筛选器辅助于授信流程，但由于历史授信数据中隐含“种族歧视”产生的群体偏差，使得部分借款人在AI授信系统中面临贷款利率和授信比例的不公平。[8] 另一个关键问题是，数据污染与标签伪误在金融大数据中尤为常见，且难以察觉。由于金融行为记录高度依赖自申报、间接观察或第三方数据供应，存在信息滞后、标签伪误等问题。如，在反欺诈建模中，部分真实欺诈事件可能未被识别或未被标注为“欺诈”，从而使模型误将异常交易视为正常行为，降低其检测灵敏度。此外，外部数据源（如社交行为、地理位置信息等）若未经标准化清洗，极易引入干扰性特征，造成模型输出偏离业务逻辑。样本代表性缺失也可能使得AI模型在面对边缘群体或非主流金融需求时表现不佳，甚至完全失效。在传统金融体系覆盖不足的区域，相关行为数据本身即缺失或极端稀疏，AI模型因此在训练阶段缺乏多样性，难以适应新的信贷需求或交易模式。如，在农村金融、绿色金融或初创企业融资等领域，模型容易出现过度保守甚至拒绝提供服务的倾向，从而造成算法排斥。此外，数据偏误还可能引发模型不稳定性与不可转移性。金融环境的动态变化导致历史规律难以适用于未来数据。如，重大外部冲击期间消费行为、违约概率与市场预期发生显著转变，基于正常年份训练的AI模型难以捕捉到新环境下的非常态风险特征，从而产生预测崩溃。这不仅使模型效率骤降，也削弱了金融系统对突发事件的响应能力。

基于上述因素，数据偏误、污染与代表性缺失是AI金融应用中极为关键的脆弱性根源。这些问题看

中国知网 https://www.cnki.net

似属于建模之前的技术细节，实则深刻影响着金融系统资源配置、公平正义与系统稳定性，是构建可信金融AI系统过程中必须正视与优先解决的关键环节。

3. 对抗易感性放大入侵风险

不同于传统的网络安全问题，对抗攻击针对的是模型输入与参数结构本身，通过构造人类难以察觉的微扰样本，误导AI模型产生错误判断。在金融场景中，这类攻击方式一旦与交易行为、身份识别或信贷审批结合，可能引发风险外溢、服务欺诈乃至金融欺诈系统性事件。

对抗样本攻击是最典型的技术形式。如，攻击者通过使用对抗性攻击技术对信用卡业务的AI检测系统进行攻击以规避欺诈分类器，造成系统无法有效识别伪造信用卡给消费者、商家和银行带来经济损失。[9] 另一个重要威胁是，模型反向工程与参数盗用正成为金融AI安全的重要隐患。攻击者可通过有限的API访问接口，对金融机构训练出的风控模型进行“探测性试探”，推测其特征权重与决策边界，从而开发针对性的规避策略，或构建仿冒模型实施商业盗用与操纵交易。[10] 这类攻击尤其对开放式信贷平台、金融科技公司及依赖云平台部署AI服务的中小机构构成巨大威胁。此外，敌手数据“投毒”也是近年来受到关注的重要攻击方式。攻击者通过伪造或篡改模型训练数据（如上传虚假交易记录、构造异常客户画像），使得AI模型学习到错误的风险判断模式。[11] 如，在反洗钱系统中，大量虚假正常交易被注入训练集后，系统可能会降低对真实异常行为的识别灵敏度，最终造成监管失效。更严重的是，这类攻击具有“隐蔽性”与“累积性”双重特征，难以被金融机构在模型训练初期及时发现。

对抗攻击的系统性风险在于其可能同时破坏多个模型、多个机构乃至整个金融市场的信任基础。一方面，攻击者可能同时针对多个金融机构的模型，实施跨平台干扰；另一方面，若攻击行为被复制、共享或市场化出售，则可能诱发AI模型集体性错误，进而通过自动交易、信贷同步等机制引发系统性风险扩散。

4. 模型漂移导致决策失效

人工智能模型依赖于历史数据对未来状态进行推断，其根本前提是训练数据的分布与应用场景保持一致性。然而，金融系统本身是一个高度动态的非稳态环境，各类政策冲击、经济周期、市场情绪和制度演化常常导致数据分布发生变化，从而引发模型漂移问题。当模型所依赖的数据生成机制不再稳定，其预测准确性、风险评估能力与决策有效性都将受到系统性削弱，进而对金融稳定构成威胁。

模型漂移一般可分为两类：概念漂移和数据漂移。前者指目标变量与特征之间的关系发生变化，如金融危机期间违约行为的决定因素可能由收入水平转向现金流紧张或市场信心崩塌；后者则指输入变量的分布本身发生变化。[12] 这两类漂移往往同时存在，并相互作用，导致AI模型在实践中“脱离现实”。典型案例是Zillow的算法失效事件，Zillow以算法定价进行房地产的大规模收购与转售，但由于新冠疫情冲击导致其误判2021年房地产市场的涨势与成本导致项目被迫关停并大额减值，背后是样本外环境变更导致的模型失效。金融环境中的模型漂移具有以下典型表现：（1）违约风险模型失效：在利率快速上升、失业率飙升等冲击下，传统基于历史信用行为的违约预测模型可能因变量不稳定而预测崩溃。（2）异常检测能力下降：在反欺诈系统中，模型依赖的交易行为模式若因宏观扰动或市场行为变迁而发生转移，系统容易出现误报率飙升或漏报严重的双重问题，使欺诈者趁虚而入。（3）资本定价模型适应性下降：基于历史市场数据训练的AI投资策略，若未能及时感知宏观经济预期或市场结构变化，容易形成策略惯性，在新环境中导致投资失误。

更为严重的是，模型漂移可能诱发系统性风险，即大量金融机构在同一时间内因使用失效模型作出相似而错误的判断，进而通过信贷紧缩、资本回撤、资产再定价等路径引发市场波动。如，在高频交易系统中，一旦模型未能识别市场状态突变，可能导致大量同步撤单、流动性枯竭或定价断裂现象。金融AI系统中若缺乏对概念漂移与输入状态突变的应急识别机制，将难以在极端情况下维持市场稳定性。

5. 算法共振引发交易同步化

人工智能技术的标准化与可扩展性使得越来越多的金融机构在投资策略、交易执行与市场预警中部

中国知网　https://www.cnki.net

署相似甚至相同类型的AI模型。尽管这一趋势提高了信息处理效率，但也显著放大了模型间的相关性与行为同步性，为市场引入了新的系统性风险源。这种算法共振现象，是指多个AI系统在面对相似市场信号时做出趋同反应，从而形成“同向交易—价格波动—信号强化—行为复制”的自我增强机制。

从高频交易的角度看，AI交易系统在高频环境下对市场信号的反应具有高度灵敏性与低延迟。当多个金融机构部署的算法在类似市场指标（如波动率阈值、成交量突变、资产相关性）触发下同时做出买入或卖出决策，会导致价格瞬时剧烈变动。[13] 这种同步交易行为不仅可能触发后续算法的连锁反应，还可能激活基于相同或相似输入数据训练的风控机制，从而进一步压缩市场流动性，放大波动风险。以2016年10月7日的英镑“闪崩”为例，由于亚洲时段流动性稀薄，自动化卖单与止损集中触发，引发做市撤流与顺向成交，英镑数分钟内急贬后迅速回撤，呈现短时算法共振特征。尽管此事件并非直接由AI驱动，但也充分说明，当多个算法模型依赖相似信号源（如情绪热度、价格趋势）时，易发生“行为共振”，放大市场非理性波动并加剧系统性脆弱性。进一步扩展来看，算法共振不仅发生在交易层面，还可能出现在AI风控模型、定价模型乃至信贷评级模型之间。大量金融机构使用相似的训练数据和建模方法构建信用评分体系，若宏观冲击或数据系统性失真使模型集体“误读”风险，将导致信贷资源的大规模同向收紧，从而通过资产价格下跌与信用紧缩传导至整个金融系统。这种“结构性模型同质化”问题在金融科技平台高度集中化的发展趋势下尤为严重。此外，AI系统在训练过程中普遍缺乏对市场结构变化、制度背景和行为复杂性因素的深度嵌入，极易将短期有效的交易逻辑误判为长期可行的通用规律。这种所谓的过度拟合市场情绪，使得AI系统在面对未知冲击时极易陷入“模仿—误判—同步”的路径依赖逻辑。

面对算法共振风险，传统的金融风控与监管手段难以完全奏效。一方面，AI交易系统内部决策逻辑高度复杂且封闭，缺乏透明性和审计可行性；另一方面，不同机构部署的AI模型无法实现实时信息共享，难以及时识别集体行为的放大风险。这要求金融监管者引入系统层面的AI交互行为分析框架，重点识别模型间的输入相关性、响应路径重叠度与行为同步概率。算法共振与交易同步化风险是AI技术“普及性”与“趋同性”共同作用下的新型系统性脆弱点，其本质在于技术同质化与行为协同放大之间的系统性失衡。

三、风险演化路径：从技术脆弱性到系统性金融风险的逻辑链条

人工智能技术在金融领域的广泛应用，将技术风险从原本可控的系统边界逐步推向可能外溢的系统性金融风险前沿。尽管前述各类技术脆弱性在发生时多呈现为局部故障或单点误判，但在金融系统的高关联性、高杠杆性与高度自动化特征作用下，这些微观层面的AI技术缺陷极易沿特定路径向上传导，最终演化为结构性风险、信任危机甚至金融不稳定。本部分旨在梳理人工智能系统中典型技术脆弱性向系统性金融风险演化的路径机制，构建从个体模型失效到群体同步误判，再到市场行为错配与宏观系统扰动的风险链条。

1. 自动化决策偏误路径：从个体误判到系统性错配

自动化决策偏误路径主要体现了数据偏误、模型漂移和黑箱模型等脆弱性如何在信贷审批、资产配置等关键金融业务中积累误差，并通过机构内部的自动化流程结构固化，引发系统性错配与金融服务排斥，形成从个体误判到资源错配的风险演化通道。人工智能技术在金融机构日常运营中已广泛用于信贷审批、投资组合优化、风险评估与客户识别等关键环节。一旦AI系统出现误判，特别是在输入数据、模型逻辑或训练机制存在缺陷的情况下，原本局部性的判断偏差可能通过机构内部操作流程迅速放大，进而引发系统层面的资源配置错配、客户信用评估失真乃至金融服务排斥。

在信贷决策环节中，AI系统若因数据偏误或模型漂移而做出错误的违约概率预测，将直接影响信贷授信额度与定价水平。如，有研究发现，若历史数据中存在对低收入群体等的隐性歧视，AI模型在学习过程中将该偏见固化为风险信号，从而系统性地“低估”这些群体的信用水平。[14] 在无人工审核机制补

中国知网 https://www.cnki.net

充的全自动化审批模式下，这类判断误差将大规模推导为“信贷排除”，削弱了金融的普惠性与包容性目标。而在投资决策与资产配置中，模型漂移的不稳定性亦可能引发系统性错配。AI系统在组合优化中常采用风险因子模型、历史收益最大化逻辑或短期市场信号建模，一旦市场环境发生变化而未及时识别，便可能做出资产过度集中、风险低估或流动性误判的投资建议。如，某些AI量化模型在市场回撤初期可能自动增持此前高收益资产以维持风险平衡，反而在持续回调中进一步加剧亏损。这类“模型路径依赖”行为体现出AI系统在面对不确定性时容易固守原有策略，缺乏灵活响应能力。

更需警惕的是，由于黑箱模型的解释性不足，在信贷审批、投资组合优化等操作中的内部决策逻辑高度复杂、不透明。当模型出现误判时，相关业务部门往往无法明确识别是哪一类输入变量、哪个中间逻辑节点或哪种训练偏差导致了异常结果。如，一位信用状况良好的客户若被拒贷，黑箱模型无法提供可追踪的拒绝理由，从而使人工干预和纠错机制失效。这种不可解释性使得“误判—执行—再反馈”的闭环持续存在，误差被自动化流程不断放大。在金融机构内部流程高度自动化与模块化的背景下，AI系统的微观误判极易“嵌入”操作流程，形成结构化误判机制。以风控流程为例，当信用评分模型出现错误时，审批、额度管理、催收定价等下游模块仍按该评分进行操作，导致误差沿链条放大，形成全流程风险偏置。一旦此类偏差在多家金融机构中同时存在，便可能演化为行业性的资源错配与风险暴露。

从系统层面看，自动化决策偏误风险的集聚还可能带来更深层次的风险外溢。一方面，不合理的信贷结构可能导致资金流向效率低下，抑制实体经济潜在增长；另一方面，若错误地将大量中低风险客户识别为高风险客户，将迫使其转向影子银行、非正规借贷等灰色领域，从而引发监管盲区与系统性脆弱性。AI系统在金融机构微观层面的判断误差并非仅限于个别事件，而可能通过操作流程嵌套、行为惯性累积与机构间扩散路径，形成广泛而深远的资源配置扭曲。这种“由点到链、由链到面”的演化路径揭示了人工智能微观风险向系统性金融错配演化的潜在机制，是当前AI金融安全治理中亟需关注与优先防范的方向。

2. 数据歧视外溢路径：从算法偏误到社会不公

数据歧视外溢路径的起点源于数据偏误这一技术脆弱性，并在黑箱模型的不可解释性下通过算法模型持续传导出歧视性结果，并演化为金融服务可得性下降与群体信任削弱的问题，最终可能在社会层面触发对金融制度正当性与合法性的系统性挑战。人工智能系统依赖大规模历史数据进行学习与决策建模，而金融数据本身深嵌于既有制度环境、社会结构与行为选择的历史轨迹之中，难以保持客观中立。在训练数据存在歧视或代表性不足的情形下，AI模型极易在无意识中“学习”到结构性不公，并将其转化为决策依据。这种“算法偏误”不仅在技术层面造成局部的不准确与不公平，还可能在社会层面演化为金融服务歧视、群体排斥与信任机制弱化，形成“数据偏见—算法歧视—信任流失”的外溢性风险链条。

历史数据中的结构性偏差往往反映社会对特定群体的长期制度性歧视。以美国为例，因“红线划区”（Redlining）等历史实践造成的种族隔离与信贷歧视，使得某些地区、族群在信用记录中系统性缺乏“正面数据”。所谓“红线划区”，是指20世纪30年代起由美国住房贷款机构将少数族裔聚居区划为“高风险区域”，在城市地图上用红色标出，从而系统性限制这些地区居民获得住房抵押贷款与金融服务的资格。[15] 这一政策造成长期性的金融排斥，其历史遗产在现代金融数据中仍具显著痕迹。AI系统在此类数据上训练后往往会将“历史缺乏信贷参与”误判为“未来高风险”，进而在模型输出中反复强化这一标签，从而形成“算法再歧视”。同时，这种偏误由于黑箱模型的不可解释性在实际金融服务中表现为隐蔽性排斥。与传统“明示拒贷”不同，AI系统可能在无明确拒绝的情况下，通过设定高阈值、动态调整定价、精细化画像等方式将特定群体“筛除”出服务对象。由于这些过程高度技术化且不可解释，受影响用户难以识别是否遭遇歧视，更无从申诉，造成技术隐性权力的加剧。[16] 更为重要的是，这种偏误的积累不仅影响个体金融可得性，更可能引发社会层面的信任危机。一方面，长期被边缘化的群体会将技术工具视为延续不平等的制度安排，从而削弱其对金融体系的信任感；另一方面，一旦算法歧视问题广泛曝光，公众对金融机构采用AI技术的合法性与正当性将遭遇挑战。如，美国住房信贷市场曾因评分算法的种族偏误而引发广泛争议，迫使监管机构启动针对算法公平性的专项审查。这类事件表明，AI系统一

中国知网　https://www.cnki.net

旦被感知为“不公正的自动化”，将迅速从局部风险转化为系统性信任危机。

另一个深层的问题在于，数据歧视风险具有高度的外部性。个体用户难以改变自身数据轨迹，而金融机构又普遍缺乏纠偏动因，使得整个市场形成“技术路径依赖”：即便新数据进入系统，既有模型与流程机制仍可能维持旧有偏误，从而延续不公。更甚者，在多机构同时使用相似评分逻辑与数据源的背景下，歧视性行为可能在行业内部形成固化，难以通过市场竞争机制自动纠正。AI系统在数据偏误条件下形成的算法歧视，远非局部技术问题，而是一种可能破坏社会契约与金融信任基础的系统性风险路径。

3. 集体性误判路径：从集体偏误到市场错误定价

集体性误判路径重点展现了模型漂移、对抗攻击与算法共振导致的AI系统间响应趋同性如何在市场扰动中同步触发失效判断，形成市场错配与价格脱锚的链式机制。特别是在高频交易场景中，算法共振放大的同步行为被识别为新型系统风险源。不同机构使用相似模型、相似数据源与反应机制，从而在面对对抗攻击等外部扰动时出现集体性判断失误，进而扭曲市场定价与风险认知。这一风险路径可被概括为“模型同质性→响应同步性→市场错位性”的三阶段演化逻辑。

AI模型结构与数据来源的趋同性为同步性误判埋下隐患。随着机器学习算法标准化程度的提高，不同金融机构往往采用类似的建模技术、优化目标函数以及同质化的训练数据（如央行信用数据、公开财报、社交媒体情绪指标等）。这种结构性同质化使得各机构对市场信号变化的反应路径趋于一致，当外部环境改变导致模型漂移或遭遇外部对抗攻击时，集体性误判的可能性显著上升。如，在新兴市场资产定价中，若多个AI系统将美元流动性指标作为关键预测变量，则一旦美元出现短期紧缩信号，模型可能同步下调对新兴市场的风险评估，导致集体性资本撤出，最终形成与基本面不符的价格错杀。这种“个体误判—行为同步—资产偏离”链条，是AI模型同步性风险的典型表现形式。同时，模型在特定条件下的失败具有“协同触发”特征。一方面，面对罕见或极端事件，变量关系或变量分布特征发生变化，AI系统容易出现“样本外状态”识别失败的模型漂移，进而做出错误判断；另一方面，多机构系统集体进入此类盲区时，不仅无法纠偏，反而可能相互强化误判。在高频交易场景中，算法共振进一步放大同步性误判的市场影响。不同交易算法在面临类似市场信号（如价格跳跃、成交量异常）时同步执行买入或卖出操作，可能诱发价格断裂、流动性枯竭、误报触发等“微结构错乱”现象。这种技术结构上的同步响应机制，实质上构建了金融市场的新型联动风险渠道，即“算法相似性”替代传统“资产相关性”成为系统风险传播的纽带。

此外，AI模型的自动学习机制往往缺乏宏观调节功能。在集体误判导致市场定价错位之后，若后续模型继续以错位价格为输入进行训练，可能进一步加剧偏差，形成“行为惯性—信号伪造—认知循环”的反馈闭环。这种反馈机制使得市场失真状态可能持续存在，甚至形成新一轮风险错判的训练基础。AI模型在多机构部署、标准化建模与数据源同构背景下，极易形成集体性误判路径，对市场价格形成机制构成挑战。这种同步性风险不同于个体故障，更类似于系统耦合机制的“失调”，其识别与治理需要跨机构、跨模型的系统联动机制。

4. 基础设施脆弱性路径：从技术依赖到金融中断风险

基础设施脆弱性路径主要对应AI系统在关键金融基础设施中的高度嵌入受黑箱模型、攻击易感性与外包部署等因素影响形成新型单点故障机制，进而对整个金融系统运行稳定性构成挑战。随着人工智能技术深度嵌入金融系统的底层架构，越来越多的金融基础设施——包括支付清算系统、风控平台、账户认证机制、市场交易枢纽乃至监管接口——开始依赖智能化决策与自动化执行。虽然这提升了系统运行效率与响应速度，但也形成了“关键节点高度自动化 + 模型不可控”的新型脆弱结构，使得技术故障、对抗攻击或模型漂移可能在瞬间引发广泛的黑箱性导致模型不可解释、数据偏误引发算法歧视、对抗攻击破坏模型安全性、模型漂移削弱预测稳定性以及算法共振诱发市场同步风险，构成金融系统运行连续性的新威胁。

AI系统成为关键操作流程的“系统性依赖点”。当前大量金融机构将AI模型用于实时欺诈检测、资金交易审核、身份验证与清算风控等关键环节。一旦模型逻辑或数据源出现故障，将导致整个业务链条

中国知网 https://www.cnki.net

“阻断式中断”。这种因AI系统崩溃引发的“单点故障—业务中断—信任损失”路径，是金融基础设施新型风险传播机制的典型表现。同时，模型外包化与云端部署带来“黑箱式技术外依风险”。许多金融机构将AI模型开发、训练或运行外包给大型技术服务商，并部署于云计算平台，这虽然节约了成本并提高系统灵活性，但同时加剧了对不可控技术底座的依赖。由于服务方往往不提供模型源代码或决策规则，金融机构自身对AI系统的运行逻辑与潜在风险缺乏深度掌控，在面临系统错误、合规审查或攻击事件时难以及时排查与响应。此外，跨境数据传输与监管不一致还可能带来监管管辖模糊与数据泄露难追责等复合性风险。AI系统在关键场景下被攻击或误触将放大风险冲击的速度与范围。如，在支付清算系统中，若攻击者通过对抗样本误导模型判断，可绕过欺诈检测机制发起大规模虚假交易，造成链式清算失败或系统性挤兑。在资本市场中，若AI系统在重大经济事件发布前后误读市场信号并执行大规模交易，则可能引发算法踩踏或市场脱锚。这种风险一旦在系统性基础设施节点爆发，其影响将迅速波及全市场，难以通过局部隔离加以控制。

更深层次的问题在于，传统金融系统治理框架并未充分准备应对AI主导型基础设施风险。当前大多数监管规则仍停留在对“操作风险”“IT系统故障”的静态认定，而对AI系统在数据更新、模型漂移、输入操纵等方面引发的动态、演化性故障缺乏制度化识别机制。加之监管科技工具本身对AI模型的可解释性、可审计性能力尚不足，进一步加剧了治理盲区。AI系统在金融基础设施中的全面部署虽提升了效率与智能化水平，但也引入了前所未有的系统级运行风险。

四、构建AI金融风险防控机制的对策建议

在人工智能技术广泛嵌入金融体系的背景下，前述所揭示的模型不透明、数据偏误、行为共振与基础设施依赖等技术脆弱性，已具备演化为系统性金融风险的现实可能。如何有效识别、控制并防范这些新兴风险，不仅关系到金融机构的稳定运营与用户权益保障，更关乎国家金融安全的制度韧性与前瞻治理能力。本部分从技术层、制度层与治理层三个维度出发，系统构建AI金融风险的防控机制路径。

1. 在技术层面构建以系统内生韧性为核心的防控机制

将技术层面的风险防控作为整个AI治理体系的核心基石。首先，应将可解释性置于提升模型透明度的首要位置。在信贷审批、保险核保与反欺诈等高风险决策场景，实施双重策略：一方面，对现有复杂模型（如深度学习网络），运用事后解释方法进行局部归因分析，揭示关键输入对输出的边际影响；另一方面，在建模之初就应优先考虑采用具备内生可解释性的结构，如，决策树、广义线性/稀疏模型以及因果图等，这为后续的风险审计、决策追溯与错误验证提供坚实基础。其次，应建立覆盖数据全流程的质量控制闭环。在数据源管理上，实施严格的输入数据“白名单”制度，仅允许符合标准化、合规性与分布代表性要求的数据进入训练与推理流程。在数据的标签治理上，需结合半监督学习与专家经验复核，定期校验标签的时效性与准确度，以规避因标签漂移造成的结构性误导。[17] 在动态适应方面，则应通过持续学习等机制，[18] 为核心模型配置概念漂移的自适应能力，主动降低因市场环境切换导致的性能衰退。同时，必须配套设计故障安全修复机制，如，实时的异常输出监测与故障隔离“安全阀”，确保任何情况下模型的行为都处于“可控、可断、可恢复”的状态，防止单点故障引发系统性风险。最后，应构筑模型行为监控与人工介入的最后一道防线。为防止错误决策外溢，需在模型输出层建立实时的异常识别与响应机制。该机制以历史行为一致性、业务规则约束与输出置信区间为基线，对模型的每一次决策进行监测。一旦结果超越预设阈值，系统即自动执行降级（如转人工审批）或熔断（如暂停服务），并触发人工链路介入复核，形成“人机协同”的安全冗余。技术层的前置防御，使系统在设计之初即内嵌“可解释—可控—可恢复”的逻辑，从而为制度层与治理层的后续管控提供坚实基础。

2. 在制度层面构建算法透明和审计规范的防控机制

制度层面应构建以“可审计性、可披露性与协同性”为核心原则的监管框架，推动监管理念从被动静态合规到主动风险管理的转变。首先，在信息披露方面建立以可审计性为核心的算法登记与模型报备制度，要求机构在引入或更新AI模型时向监管机构提交完整文档，覆盖建模逻辑、变量来源、训练数据

中国知网　https://www.cnki.net

概况、敏感性与稳健性测试结果以及性能指标，并配套解释性报告与变更记录，以便实现穿透式监管与事后问责。其次，在审计机制方面推动独立的第三方算法审计，将验证、监控与再评估嵌入模型全生命周期，对公平性、稳定性、合法合规与歧视性输出开展定期与专项评估，避免“自审自证”。再次，在技术标准方面确立可解释的最低合规门槛与通用披露模板，对高风险场景中可解释性不足的模型施加风险加权或使用限制，并与数据治理、隐私保护等配套标准同频推进。同时，建立由央行、金融监管总局、数据安全与算法备案等部门组成的联动网络与预警平台，实现模型黑名单与重大异常的跨部门共享与协同处置，对存在重大合规风险的模型依法实施“暂停—整改—复评”的闭环管理。最后，从法律责任角度明确金融机构在自研或外采AI模型时的首要责任人地位，探索将算法信用档案与合规记录纳入监管评级与行政许可参考，形成“合规激励—违规惩戒”的制度闭环。

3. 在治理层面构建责任结构明晰的防控机制

在金融机构的内部治理架构中围绕AI系统的完整生命周期，从数据准备、开发、验证、部署到持续监控与最终停用，建立一个结构化、权责明晰的风险防控机制，将管控责任明确地分配至各相关职能单元。首先，严格划分各方职责。模型开发与业务使用部门作为执行单元，共同负责保障输入数据的质量、模型设计的业务逻辑对齐以及运行参数的设定。独立的模型风险管理与合规部门则构成关键的监督与制衡职能，负责对模型的概念完备性、算法稳健性和性能表现进行独立验证，并确保其应用符合机构整体的风险偏好与外部监管要求。所有关键模型的启用、重大参数变更或停用，均须经由专门的治理委员会审批，并形成可供审计与监管审查的完整决策记录。其次，实施覆盖全生命周期的动态风险管理流程。模型部署前，必须完成包括算法偏见评估、输入敏感性分析和对抗性测试在内的全面验证，以建立其性能基线。在运行阶段，通过对关键性能指标和数据分布的持续监控，执行预设的分级干预协议：从性能偏离时的自动预警，到触发人工复核的权限降级，直至在极端异常情况下暂停模型的自动化决策功能。任何由模型引发的风险事件都必须触发根本原因分析，其结论应被纳入知识库，用以指导未来治理框架的迭代优化。最后，该治理框架应无差别地覆盖所有外部采购的第三方模型。通过严格的尽职调查和合同约束，确保机构拥有对外部模型的审计权、获得必要解释的权利，并明确数据来源、使用范围、退出机制及事件响应协议，以消除潜在的治理盲区，实现对机构内所有算法系统的全面、一致的风险管控。

参考文献：

[1] 廖高可，李庭辉. 人工智能在金融领域的应用研究进展 [J]. 经济学动态，2023，(03)：141-158.

[2] 肖凯元. 人工智能赋能耐心资本：现实价值和路径探索 [J]. 经济学家，2025，(03)：56-66.

[3] Danielsson J，Macrae R，Uthemann A. Artificial Intelligence and Systemic Risk[J]. Journal of Banking & Finance，2022，140：106290.

[4] Gallagher M，Pitropakis N，Chrysoulas C，et al. Investigating Machine Learning Attacks on Financial Time Series Models[J]. Computers & Security，2022，123：102933.

[5] 马 挺，韩廷春. 人工智能在金融领域的伦理考量：风险与对策 [J]. 科学管理研究，2025，(02)：156-166.

[6] 程雪军. 生成式人工智能嵌入数字金融平台的算法权力风险及规制进路 [J]. 法治研究，2025，(02)：32-41.

[7] 刘春航. 人工智能、大数据与金融风险管理 [J]. 金融监管研究，2025，(05)：1-11.

[8] Fuster A，Goldsmith-Pinkham P，Ramadorai T，et al. Predictably Unequal? The Effects of Machine Learning Cn credit Markets[J]. The Journal of Finance，2022，77(01)：5-47.

[9] Tsai M Y，Cho H H，Yu C M，et al. Effective Adversarial Examples Identification of Credit Card Transactions [J]. IEEE Intelligent Systems，2024，39(04)：50-59.

[10] Oliynyk D，Mayer R，Rauber A. I Know What You Trained Last Summer：A Survey on Stealing Machine Learning Models and Defences[J]. ACM Computing Surveys，2023，55(14s)：1-41.

中国知网 https://www.cnki.net

[11] Chen J, Zhang X, Zhang R, et al. De-Pois: An Attack-Agnostic Defense Against Data Poisoning Attacks [J]. IEEE Transactions on Information Forensics and Security, 2021, 16: 3412-3425.

[12] Mannapur S B. Understanding Date Drift and Concept Drift in Machine Learning Systems[J]. Internationat Joumal of Scientific Research in Computer Science, Engineering and Information Technology, 2025, 11(01): 318-330.

[13] Brogaard J, Carrion A, Moyaert T, et al. High Frequency Trading and Extreme Price Movements[J]. Journal of Financial Economics, 2018, 128(02): 253-265.

[14] Bhutta N, Hizmo A, Ringo D. How Much Does Racial Bias Affect Mortgage Lending? Evidence from Human and Algorithmic Credit Decisions[J]. The Journal of Finance, 2025, 80(03): 1463-1496.

[15] Aaronson D, Hartley D, Mazumder B. The Effects of the 1930s HOLC "Redlining" Maps[J]. American Economic Journal: Economic Policy, 2021, 13(04): 355-392.

[16] Das S, Stanton R, Wallace N. Algorithmic Fairness[J]. Annual Review of Financial Economics, 2023, 15: 565-593.

[17] Borowicz M K. The Data Quality Problem (in the European Financial Data Space)[J]. International Journal of Law and Information Technology, 2024, 32: eaae015.

[18] Wang L, Zhang X, Su H, et al. A Comprehensive Survey of Continual Learning: Theory, Method and Application[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2024, 46(08): 5362-5383.

Artificial Intelligence and Financial Security: Technological Vulnerabilities, Risk Evolution Pathways and Mitigation Mechanisms

DING Shao-bin[1]　WANG Hong-ju[2]

(1. School of Applied Economics, University of Chinese Academy of Social Sciences, Beijing 102488, China;
2. Institute of Finance and Economics Strategy, Chinese Academy of Social Sciences, Beijing 100006, China)

Abstract: As artificial intelligence (AI) becomes deeply embedded in the financial sector, the accompanying new types of financial risk are becoming increasingly prominent. At present, AI's technical shortcomings in financial practice manifest as five key technical vulnerabilities: insufficient explainability that delays risk tracing, data bias that produces algorithmic discrimination, adversarial susceptibility that amplifies intrusion risk, model drift that leads to decision failure, and algorithmic resonance that induces market synchronization risk. These technical vulnerabilities are transmitted through channels such as automated decision bias, spillovers of algorithmic discrimination, collective misjudgment, and dependence on infrastructure, presenting a chain mechanism by which local technical failures evolve into systemic financial risks. Biases in AI model judgments are easily amplified by automated processes, resulting in credit exclusion and asset misallocation. Collective misjudgment and synchronous behavior further exacerbate irrational fluctuations in market prices, undermining pricing and risk identification mechanisms in finance. Deep AI embedding in critical financial infrastructure may amplify single-point-of-failure risk, impact underlying system architectures, and in turn trigger service paralysis and market interruptions. Therefore, it is necessary to build a multi-dimensional prevention and control system at the technical, institutional, and governance levels, centered on technical safety, based on algorithmic transparency and audit norms, and guaranteed by clear boundaries of responsibility, to achieve sustainable coordination between artificial intelligence and financial security.

Key Words: artificial intelligence; financial risk; technological vulnerability; risk evolution; mitigation mechanisms

责任编辑: 何　飞

中国知网 https://www.cnki.net