# 算法歧视视阈下金融监管的法治化思考——以信贷审批自动化为研究对象

沈亚亚

贵州财经大学法学院，贵州贵阳，550025；

**摘要：**随着大数据和人工智能学习技术在信贷审批领域的深度应用，自动化决策系统明显提升了金融服务效率，但数据偏差、模型缺陷等引发的算法歧视问题也愈发凸显，既破坏金融公平性，又带来了金融监管风险。研究分析后提出相关法治化回应路径，在立法层面构建“算法影响评估强制披露制度”，在执法层面创新“沙盒监管+算法审计”协同机制，在司法层面确立“举证责任倒置”的侵权认定规则。健全法律融合机制，监管层面从立法、监管科技、国际合作重构体系，责任层面确立过错推定责任并强化社会共治。

**关键词：**算法歧视；金融监管；法治化；信贷审批

**DOI：**10.69979/3029-2700.26.04.082

## 引言

数字经济时代，大数据与机器学习技术驱动信贷审批迈向自动化，显著提升了金融效率与风险管控能力。然而，这一技术跃进亦伴随着算法歧视的隐忧。由于训练数据的历史偏见、模型设计的缺陷或决策过程的不透明，自动化系统可能对特定群体（如少数族裔、低收入者、自由职业者）产生系统性、不公正的差别对待，侵蚀金融公平的基石。

此类歧视不仅引发公平性质疑，更对金融监管构成严峻挑战。算法的“黑箱”特性使得歧视难以被察觉与验证；法律规则的滞后导致监管依据不足；责任主体在金融机构、算法开发者与数据提供者之间的模糊，使得问责困难。因此，探究如何在法治框架下有效规制信贷审批中的算法歧视，平衡金融创新与公平、安全的价值，已成为亟待回应的重大课题。

算法歧视并非算法本身的问题，而是数据、模型设计和决策造成的不公平性，即对某类特定人群进行不同的歧视，造成金融市场不公平、金融服务不可及。

数据的偏见传导链条是造成算法歧视的重要原因。训练数据往往是算法学习和决策的基础，若数据中存在被歧视的数据，就会被算法放大。在发展差异比较大的地区，产生过违约率高的数据通过信贷审批模型的训练数据系统后，在后续的评估过程中会对该地区的借款人产生偏见。此时，即使该地区经济已好转、借款人信用良好，也可能会因为该地区存在地域标签提高贷款门槛或者拒绝贷款，就会形成算法歧视的“数据→模型→决策”传导链条。[1]

模型设计中的不公平性。部分算法模型没有考虑反歧视的规则，而是隐性关联性别、种族等特征，过度简化不合理的选取弱势群体，而后对自由职业者的贷款申请只使用固定收入证明作为评判收入是否稳定。[2]忽略了自由职业者通过项目收入、兼职收入等多种方式获得收入，导致自由职业者在信贷审批中处于劣势，通过率远低于传统职业者，存在结构性歧视。

## 1 算法歧视对金融监管法治化的挑战

### 1.1 现行监管框架的适应性困境

如今金融科技迅猛发展，法律规范的迭代速度往往滞后于技术革新，使得算法歧视问题的处理显现出明显的迟滞性。在面对新兴的算法歧视情形时，我国法律规范未清晰界定算法歧视的法律概念与判定基准，这导致监管机构在评估金融机构算法应用是否构成歧视时缺少具体法律支撑。歧视行为实施精准干预就成为一大难题，《个人信息保护法》对自动化决策的公平性审查提出要求，却缺乏配套的实施细则，使得金融机构在操作中难以把握公平性审查的界限。由于缺乏明确的法律标尺，监管机构在认定该行为是否构成算法歧视时陷入两难，无法及时采取有力的监管手段。

传统金融监管工具在很大程度上依赖于人工审核与合规性检验，当应对复杂且持续演进的算法模型时，效能显得捉襟见肘。人工审核难以彻底、全面地剖析算法的内在逻辑与运作机制，因而无法及时识别潜藏的歧视性因素，算法模型的迭代速度极快，常在短时间内完成多次调整与优化。而传统监管手段则难以同步适应这种变化，引发监管的滞后效应，实时监控与偏见识别技术尚未在金融监管领域得到普遍应用。监管机构因而缺少高效的技术工具来对算法实施即时监测与分析，难以迅速发现纠正算法中的歧视行为。金融机构的信贷审批算法在运行期间，输入数据的变化常常产生歧视性结果并且监管机构缺乏实时监控技术，未能及时察觉，最终

导致部分消费者遭受不公待遇，直到问题积累到相当程度将会对消费者权益和金融市场公平性造成重大损害。

### 1.2 金融公平与监管效率的价值冲突

自动化审批在提升金融服务效能、促进“普惠金融”进程方面扮演着关键角色，它高效处理海量信贷申请，让更多人能够轻松获取金融服务。在追求效率的过程中，算法歧视问题日益显现，对“实质公平”造成侵蚀，小微企业信贷绿色通道为小微企业提供快速融资渠道。但部分算法在简化审批环节时，过度依赖一些标准化信用评估指标，忽略初创企业信用评估的独特性，初创企业往往缺少充足的财务信息和持续的经营记录流水。按照传统算法的评判标准，难以获得理想的信用评级。

金融科技的迅猛发展离不开技术革新的驱动，而算法在金融行业的运用则是技术进步的关键展现。倘若监管过于严苛，对算法应用施加过多约束与规范，这将削弱金融机构的创新动力，进而阻碍金融科技的发展进程。监管若过于宽松，则可能导致歧视性算法的泛滥，破坏金融市场的公平性与稳定性，加剧金融风险。如何在推动技术创新与强化风险防范之间寻求协调已成为金融监管领域的一大难题。建立“原则导向监管＋技术合规性”的动态协调机制已然出现在众多立法者的脑海当中。[3]互联网银行采用“地域＋职业”的标签体系快速筛选借款人，将审批周期缩短在 10 分钟以内。但也被揭露存在对农村地区借款人的隐性排斥现象，这反映出金融机构在追求效率与创新时容易忽略公平性及风险管控的必要性。

## 2 算法歧视视阈下金融监管的法治化路径构建

### 2.1 数据治理的法治化规范

运用重采样技术解决样本不均衡问题是数据处理的核心步骤，在真实数据环境中，不同群体间的样本数量差异显著。特定职业群体的数据量明显多于他群体，这会造成算法对数据量丰富群体的过度拟合，而对数据量稀疏群体训练不足，进而产生不公平的输出结果。借助重采样技术，对数据量较少群体实施过采样，对数据量较多群体实施欠采样，使各群体数据样本数量达到相对平衡状态，增强算法对不同群体的兼容性与公正性。

确保算法歧视得以规避，数据应用阶段的公开性与适宜性扮演着核心角色，针对自动化决策所倚重的重要数据要素展开显著性评估。进而明晰各项数据要素对审批结论的作用范围，金融组织有必要公开对审批结论影响力超出 10%的要素及权重值，借此使借款方解审批决策的来龙去脉，提升决策过程的可理解性。借款方在提交贷款申请遭到拒绝之后，可借助查阅公开资料，掌握是哪些核心数据要素引致拒贷结果，例如负债收入比例过高、信用记录时间过短等情况，对决策流程形成更透彻的认知，监管机构可指令金融组织在信贷体系中整合“数据偏误侦测单元”。该单元是一种借助前沿技术手段对数据偏差实施实时监控的有效机制，能够自动辨识与性别、地域等敏感要素紧密关联的指标，当某一模型内“户籍所在地”要素同违约概率的关联强度超过 0.6 时，便会启动人工审核程序，借助人工审核，审视算法决策是否受到敏感要素的不当干扰，避免历史偏误演变为算法歧视，维护借款方的平等借贷权益。[4]

### 2.2 算法规制的技术——法律融合机制

#### 2.2.1 立法层面：披露责任与救济机制的强化

算法的公开性是维护金融消费者知情权与公平权的关键基础。建议金融机构为用户出具“决策解析文档”，帮助用户透彻解信贷审批的判定标准与操作流程，该文档需具体阐述贷款被拒或价格差异化的核心要素。例如负债收入比高于 70%或过去 12 个月内信用卡逾期 3 次等实际情形，用户能够明确自身申请遭拒或获得不同定价的缘由，加深对裁决结果的理解与认同，方便用户在察觉裁决存在偏颇时，能够有目的地提出质询或申诉。

引入第三方机构对算法实施周期性审查，是保障算法公正性的关键外部监管手段。第三方机构运用“disparate impact analysis”（差别影响分析）[5]来检验不同群体在审批结果上的统计性区别，以科学、客观的方式评判算法是否存有偏见。构建金融算法监管体系，要求相关机构提交算法类别、训练数据概要、风险控制指标等数据，。通过对模型参数及输出结果的实时监测，判断算法是否存在针对特定群体的偏差，一旦检测到偏差，可立即采取调整和修正措施。

制度设计应当清晰界定双重责任框架：针对未能履行评估职责的机构，需设定递进式的法律责任，涵盖整改指令、经济处罚（建议将罚金设定在非法所得的 1%至 5%区间）以及信誉惩戒手段。应构建消费者知情权益的维权路径，允许借款人针对算法决策提出异议，要求相关机构提供评估报告的核心内容，欧盟《通用数据保护条例》[6]（GDPR）第 22 条所确立的自动化决策异议权制度，为我国消费者权益救济机制的构建提供关键借鉴，根本要义在于借助程序正当性来确保实质公正。

#### 2.2.2 执法层面：“沙盒监管+算法审计"协同机制的创新

沙盒监管创新容错机制的实施平台，必须着重界定三大关键构成要素，在准入标准层面，需构建“创新程度—风险可管理性”双轨制评价框架。优先吸纳采用机器学习、自然语言处理等前沿技术且风险扩散性较弱的信贷决策模型，要求申请主体提供算法逻辑阐释文件、数据来源合法性凭证及风险应对预案。测试时段建议设定为 6 至 12 个月的灵活区间，针对涉及个人信用判断的核心算法可延长至 18 个月，此间监管机构应按月提取模型运行结果退出规则则区分三类状况，测试合格的

算法须依监管要求完成合规性调整后方可纳入常规管理，存在轻度偏见风险的需在规定期限内修正，若检测到系统性偏差则应即刻终止测试启动溯源审查。

算法合规性审计的关键着力点在于构建覆盖全生命周期的评估体系，参照《金融行业算法审计指南》。公平性测试指标可具体化为不同性别群体授信获批率差异阈值控制在 5%以内，不同收入水平群体违约预测精度差异不超过 8%，设定算法决策可解释性评分不低于 70 分（百分制），审计执行可采取"监管机构委派第三方与机构自查相结合"的双重模式，着重检查训练数据代表性、特征变量公平性以及模型迭代记录。[7]审计报告应当包含歧视风险的量化分析、整改建议的追溯问责机制，将沙盒测试作为评估的核心参考依据。

在具体操作层面，可构建“测试数据沙盒—审计问题清单—整改验收规范”协同体系。测试阶段收集实时数据为审计工作提供基础资料，审计过程中识别出的偏颇性指标将启动沙盒退出审核，算法的商业化应用需以二次审计后的整改成效为前提，协作机制在欧盟《AI 法案》监管沙盒的应用中已显现初步成效，关键优势体现在将事后惩戒转变为事前防范。

在规则推行阶段，应当注意两种潜在风险，首先是过度增加金融机构的举证责任可能阻碍算法技术的创新实践。其次是消费者初步举证门槛设定不合理或导致诉讼泛滥，有必要借助司法实践持续改进证明标准的各项细节。

2.2.3 司法层面:“举证责任倒置”的侵权认定规则确立

在消费者遭遇算法歧视所引发的权益受损争议中，实施“举证责任转移”原则，即要求金融机构证明算法决策过程符合公正性规范，这对维护消费者权益大有裨益。相较于常规法律诉讼中原告承担举证义务的做法，在算法歧视情境下，消费者往往难以获取金融机构的算法决策内幕信息及技术参数。从而导致举证过程异常艰难，一旦适用举证责任转移，金融机构就必须出示完备的证明材料，涵盖算法模型构建逻辑、数据采集渠道与处理流程、风险测度机制等内容。以证实算法决策遵循客观且中立的准则，未对消费者实施区别对待，倘若金融机构无法提供上述证明，则需承担相应的法律后果，赔偿消费者的实际损失。

倡导金融行业协会草拟《信贷算法公平性自律章程》，切实发挥行业自律功能，引导金融机构主动遵循算法公平性准则，章程需明晰行业内部的公平性规范及行为指南，对金融机构在算法开发、数据应用、决策流程等环节提出标准化要求。

建立“算法偏见申诉机制”，为用户创设高效反馈途径，允许用户匿名反映可疑的歧视性判定，维护用户隐私权与正当利益。管单位对投诉占比超出 0.05%的机构实施专项审查，及时识别处置潜在的算法偏见问题，网络借贷机构因遭投诉“对残障人员贷款额度擅自下调 30%”，监管单位展开调查后查明算法不当关联“残疾证明编号”字段与偿债能力评估模型，依据责任归属原则。该机构被要求补偿受影响方损失公开致歉，被列入金融科技创新企业信用不良名单，此举保障消费者的正当权益，对其他金融组织形成警示效果，也可推动金融产业规范发展。

## 3 结语

信贷自动化审批作为金融科技的关键成就，在显著提升服务效能与风控水平的同时，也如同双刃剑般衍生出隐秘而系统的算法歧视。破解此困局，法治化监管乃是破题关键。为此，须构筑一个覆盖“数据治理法治化—算法规制融合化—监管体系立体化—责任分配社会化”的协同框架。该框架旨在源头规范数据质量以涤清偏见，通过技术法律融合机制以增强算法透明度与公平性，进而依托科技赋能升级监管能力，并厘清责任以激励多元共治。如此，通过贯穿算法全生命周期的规制，方能有效驾驭技术风险，在捍卫金融公平与市场秩序的同时，引导金融科技回归其包容、普惠的本源价值，实现创新与安全的良性共生。

## 参考文献

[1]凌秋实、马万利、张海汝:《人工智能背景下算法歧视法治化路径研究 —— 典型场景、规制困境及对策》，《财经问题研究》2025 年第十期，第 42-55 页。
[2]沈艳：《数字金融发展中的数据治理挑战》，《清华金融评论》2021 年第 3 期，第 91—94 页。
[3]卜素:《人工智能中的“算法歧视”问题及其审查标准》，《山西大学学报》（哲学社会科学版）2019 年第 42 卷第 4 期，第 124-129 页。
[4]彭丽徽、张琼、李天一:《人工智能嵌入政府数据治理的算法歧视风险及其防控策略研究》，《农业图书情报学报》2024 年第 36 卷第 5 期，第 23-31 页。
[5]Margot E. Kaminski, “Binary Goveranee: Lessons from the GDPR's Approach to Algorithmic Accountability,”Southern CaliforniaLaw Review, Vol.92, No.6,2019,pp.1529-1616.
[6]李晓辉:《自动化决策拒绝权的属性、功能与限度》，《法学》2024 年第 7 期，第 46-53 页。
[7]中国社会科学院农村发展研究所:《中国县域数字普惠金融发展指数研究报告 2020》,2021 年 1 月 6 日,hp:lex.essn.cnglx/glx_zk/202101/t20210106 5242724.shtml,访问日期:2025 年 11 月 24 日。