ICS
CCS

T/GXDSL

团　　体　　标　　准

T/GXDSL　— 2026

金融领域人工智能算法应用伦理与安全评规范

Ethical and Security Evaluation Norms for the Application of Artificial Intelligence Algorithms in the Financial field

（工作组讨论稿）

（本草案完成时间：2026-01-22）

2026 - - 发布　　2026 - - 实施

广西电子商务企业联合会　发 布

目　　次

前　言

本文件依据GB/T 1.1-2020 《标准化工作导则 第1部分：标准化文件的结构和起草规则》的规定起草。

请注意本文件的某些内容可能涉及专利。本文件的发布机构不承担识别专利的责任。

本文件由广西产学研科学研究院提出并宣贯。

本文件由广西电子商务企业联合会归口。

本文件起草单位：

本文件主要起草人：

本文件为首次发布。

金融领域人工智能算法应用伦理与安全评规范

1 引言

随着人工智能技术在金融领域的深度融合与广泛应用，算法已成为驱动金融服务创新、提升业务效率、优化风险管理的核心引擎。然而，算法的复杂性与自主性也引发了一系列伦理挑战与安全隐患，如算法偏见与歧视、决策“黑箱”、数据隐私泄露、模型安全攻击以及责任界定模糊等。这些问题不仅可能损害金融消费者的合法权益，侵蚀市场公平与信任，还可能危及整个金融体系的稳健运行。为推动金融领域人工智能技术的健康、可信与可持续发展，贯彻落实国家关于发展数字金融、做好金融“五篇大文章”的决策部署，并在《网络安全法》等法律法规框架下，统筹好人工智能技术创新应用与安全合规的关系，亟需建立一套科学、系统、可操作的算法伦理与安全评价体系。本规范旨在为金融机构、科技公司及行业自律组织提供一个全面的评价框架，从伦理准则与安全基线两个维度，对人工智能算法在金融业务场景下的全生命周期应用进行引导与规范。本规范由广西产学研科学研究院，基于其在人工智能产学研合作与标准研制方面的经验，联合业界机构共同提出。

2 范围

本规范规定了金融领域人工智能算法在设计与应用过程中应遵循的伦理原则、安全要求以及相应的评价方法、流程与判定准则。本规范适用于银行业、证券业、保险业等各类金融机构，以及为金融业务提供人工智能算法技术服务的第三方科技公司，对其在信贷审批、风险管理、市场交易、保险定价、客户服务、反洗钱等业务场景中部署使用的机器学习、深度学习及其他先进算法模型进行伦理影响评估与安全性评价。算法伦理与安全评价应覆盖从需求分析、数据采集、模型设计、训练验证、部署上线到持续监控与退役的全生命周期。

3 规范性引用文件

下列文件对于本文件的应用是必不可少的。凡是注日期的引用文件，仅注日期的版本适用于本文件。

凡是不注日期的引用文件，其最新版本（包括所有的修改单）适用于本文件。

GB/T 35273-2020 信息安全技术 个人信息安全规范

GB/T 22239-2019 信息安全技术 网络安全等级保护基本要求

JR/T 0193-2020 金融数据安全 数据安全分级指南

JR/T 0171-2020 个人金融信息保护技术规范

《中华人民共和国网络安全法》（根据 2025 年 10 月 28 日第十四届全国人民代表大会常务委员会第十八次会议《关于修改〈中华人民共和国网络安全法〉的决定》修正）

《中华人民共和国数据安全法》（2021 年 9 月 1 日起施行）

《中华人民共和国个人信息保护法》（2021 年 11 月 1 日起施行）

《人工智能算法金融应用评价规范》（中国人民银行发布）

T/BFIA 034-2024 人工智能算法金融应用伦理影响评价规范

《银行业保险业数字金融高质量发展实施方案》（国家金融监督管理总局发布）

4 术语和定义

4.1 金融领域人工智能算法：指应用于金融业务场景，通过数据分析、模式识别、预测或决策，实现或辅助实现特定金融功能的数学模型、计算程序及系统。

4.2 算法偏见指：由于训练数据不具代表性、特征选择不当或模型设计缺陷等原因，导致算法决策结果对特定群体（如特定性别、年龄、地域、收入水平的用户）产生系统性、不公正的有利或不利影响。

4.3 可解释性：指能够以易于理解的方式，向相关方（如监管者、风控人员、受影响的客户）说明或呈现算法模型的输入、处理逻辑、决策因素及输出结果之间的关系。

4.4 算法安全性：指算法在面对对抗性攻击（如数据投毒、模型窃取、对抗样本攻击）、系统故障、异常输入等情况时，能够保持其功能完整性、决策稳健性、数据保密性与系统可用性的能力。

4.5 伦理影响评价：指对算法在金融场景中应用可能产生的伦理风险、社会影响及对用户权益的潜在影响进行系统性识别、分析、评估与管理的活动。

4.6 模型全生命周期：指算法模型从业务需求定义、数据准备、模型开发与训练、验证测试、部署上线、运营监控到最终下线退役的完整过程。

5 总则

金融领域人工智能算法的应用，必须坚持以人民为中心的发展思想，将保障金融消费者合法权益、维护金融市场公平秩序、促进社会公共利益置于首位。算法的研发与应用机构应树立负责任的创新理念，遵循“以人为本、公平公正、安全可控、透明可信、尽职问责”的基本原则。机构应将算法伦理与安全管理纳入公司治理和全面风险管理体系，建立覆盖董事会、高级管理层、风险管理部门和业务部门的协同治理架构。算法的伦理与安全要求应贯穿其全生命周期，实现事前预防、事中监控与事后处置的有效结合。评价工作应坚持客观性、科学性和系统性，采用定性与定量相结合的方法，确保评价结果真实、准确、有效，为算法应用的持续优化与风险管理提供可靠依据。

6 伦理原则与评价要求

算法的开发与应用机构必须将伦理考量内嵌于技术流程之中。首要原则是公平性与非歧视性。机构必须采取积极措施识别并缓解算法偏见。在数据层面，应确保用于训练和推理的数据集在关键人口统计学特征上具有充分的代表性和平衡性，对历史数据中可能存在的歧视性模式进行审查与修正。在模型层面，应选用或设计能够减少偏见放大的算法，并定期进行公平性测试。评价时，需针对不同业务场景设定具体的公平性度量指标（如 demographic parity， equal opportunity difference），并设定明确的阈值。例如，在信贷评分场景中，算法对不同性别或年龄组客户的通过率差异，经统计检验不应存在显著性（如 p 值大于 0.05），或差异率应控制在不超过 5%的范围内。

透明性与可解释性是构建算法信任的基石。机构应致力于提升算法的透明度，根据算法应用的风险等级和影响范围，提供不同层次的解释。对于高风险决策（如拒绝贷款、调整信用额度、实施反欺诈制裁），必须能够向受影响的客户提供清晰、易懂、个性化的主要决策原因说明。在内部，应向风险管理和审计部门提供足够的技术细节，使其能够理解和评估模型的决策逻辑。评价可解释性时，可考察是否采用了可解释性模型（如线性模型、决策树），或为“黑箱”模型配备了事后解释工具（如 LIME， SHAP）。机构需建立可解释性测试案例库，确保在典型和边界场景下，解释的合理性与一致性通过率不低于 90%。

人类的监督与最终决定权必须得到保障。人工智能算法应定位为辅助工具，而非完全自主的决策主体。机构须建立有效的人机协同机制，明确划分算法与人工的职责边界。对于涉及重大利益、复杂情形或高风险场景的决策，必须设置人工复核与干预环节。例如，对于超过特定金额的贷款申请、对模型不确定度极高的案例、或客户提出异议的决策，必须强制路由至人工处理。评价时应检查相关业务流程设计文档和系统日志，确认人工监督环节的硬性控制是否存在且有效。

此外，算法的应用应促进金融包容与社会福祉，避免助长过度负债、非理性投机等有害行为。机构应评估算法对客户金融健康、长期利益的潜在影响，确保技术应用符合普惠金融和负责任金融的理念。

7 安全要求与评价方法

算法安全是金融稳定的技术防线。数据安全与隐私保护是基础。算法的数据处理活动必须严格遵守《网络安全法》、《数据安全法》、《个人信息保护法》及相关金融标准的要求。在算法全生命周期中，应对训练数据、输入输出数据进行分类分级管理[JR/T 0193-2020]，实施访问控制、加密存储与传输、脱敏处理等措施。采用联邦学习、安全多方计算、差分隐私等隐私增强技术的算法，应在评价中验证其技术实现的有效性，确保在发挥数据价值的同时，满足个人信息保护的最小必要原则。评价需包括对数据供应链的审查，确保第三方数据来源的合法合规性。

算法模型的健壮性与抗攻击能力至关重要。机构须对算法模型进行安全性测试，评估其对常见对抗性攻击的抵御能力。这包括但不限于：针对预测模型的对抗样本攻击测试，确保在输入数据遭受轻微扰动时，模型预测结果不会发生颠覆性改变；针对推荐或风控系统的数据投毒攻击模拟，评估模型在训练数据被恶意污染情况下的稳定性。应设定模型健壮性指标，例如，对于图像识别类应用，对抗样本的成功攻击率应低于 10%;对于信贷模型,在模拟噪声数据下,核心风险指标的波动率不应超过基准值的 15%。系统层面的安全与可靠性不容忽视。承载算法运行的软件框架、依赖库及硬件环境应定期进行漏洞扫描与安全更新。算法服务接口应具备防滥用、防重放攻击的能力。需建立算法的版本管理、回滚机制和故障应急响应预案。在涉及高频交易、实时风控等对时效性要求极高的场景，算法系统的可用性应达到 99.99%，平均故障恢复时间（MTTR）应小于 10 分钟。评价应包含对相关技术文档、运维记录和应急演练报告的审查。

8 评价流程与组织实施

算法伦理与安全评价应作为一个结构化的管理流程嵌入机构治理。评价流程通常包括评价启动、影响分析、评价实施、报告编制、结果审议与持续监控六个阶段。在算法模型开发初期或重大变更前，必须启动评价程序。由跨部门的评价工作组（成员应涵盖合规、风险、技术、业务及独立伦理专家）负责实施。

评价工作首先进行全面的影响分析，识别算法应用的业务场景、涉及的数据类型、影响的用户群体及潜在的伦理与安全风险点。依据风险等级（可参照高、中、低三级划分），确定评价的深度与广度。高风险算法（如直接涉及信贷决策、大额资金交易、客户身份认证的算法）必须进行全面、深度的评价。评价实施阶段，应根据本规范第 5、6 章的要求，采用文档审查、技术测试、场景模拟、数据审计、利益相关方访谈等多种方法，逐项检查与验证。技术测试可引入内部或第三方专业测评工具。对于公平性、可解释性、健壮性等关键指标，应出具定量测试报告。

评价结束后，应编制详细的《算法伦理与安全评价报告》，清晰记录评价过程、方法、发现、结论及改进建议。报告应提交至机构高级管理层或指定的专业委员会进行审议。只有通过审议的算法方可获准部署或继续运行。

评价并非一次性活动。对已投入生产的算法，应建立持续监控机制，定期（至少每年一次）或在业务环境、法律法规、数据分布发生重大变化时进行再评价。监控内容包括算法性能漂移、公平性指标变化、用户投诉分析以及与算法相关的安全事件等。

9 评价结果与应用

评价结果应作为算法风险管理与业务决策的核心依据。根据评价结论，算法可被划分为“完全符合”、“有条件符合（需限期整改）”、“不符合”等不同等级。对于评价通过的算法，应明确其适用的业务范围和约束条件。对于“有条件符合”的算法，必须制定并落实具体的整改计划，在完成整改并经复核前，应限制其应用范围或加强人工监控。对于被判定为“不符合”且无法通过整改满足要求的算法，必须暂停使用直至下线。

算法伦理与安全评价的相关文档，包括评价报告、审议记录、整改跟踪及持续监控日志，应至少保存五年，以备内部审计与外部监管检查。在遵守商业秘密的前提下，机构应积极探索适度提高算法透明度的方式，例如通过用户协议、产品说明书、社会责任报告等渠道，向社会公众披露其算法治理的基本理念与主要措施。

10 监督与责任

金融机构对自身使用的人工智能算法承担伦理与安全管理的主体责任。机构董事会或高级管理层对建立和维护有效的算法治理体系负有最终责任。国家金融监督管理部门在法定职责范围内，依法对金融机构的算法应用及相关风险治理情况进行监督与管理。行业自律组织可依据本规范，开展行业最佳实践推广、标准符合性评估与认证等相关工作。

对于违反伦理准则、安全规定或造成实际损害的行为，机构应建立内部问责机制。同时，相关行为也可能构成对《网络安全法》等法律法规的违反，将依法承担警告、罚款（最高可达一千万元）、责令改正、暂停相关业务等行政责任；构成犯罪的，依法追究刑事责任。

11 附则

11.1 本规范自发布之日起实施。

11.2 金融领域相关机构可依据本规范，制定更具体的实施细则。

11.3 本规范所引用的国家法律、法规、规章及标准，其最新版本（包括所有的修改单）适用于本规范。

11.4 随着人工智能技术的快速演进与金融业态的持续创新，本规范将适时进行复审与修订。