ICS 35.240.01
CCS L 77

团 体 标 准

T/ISC 0134—2026

AI 算法可解释性与决策透明度通用技术规范

General technical specification for the interpretability of AI algorithms and decision-making transparency

（发布稿）

2026－09－08 发布　　2026－10－08 实施

中 国 互 联 网 协 会　发 布

目　　次

前　言

本文件按照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。

本文件由中国互联网协会提出并归口。

本文件起草单位：中国联通国际有限公司、南京南瑞信息通信科技有限公司、珠海市人民医院、南京航空航天大学、中国信息通信研究院、北京邮电大学、中科数测固源科技（安徽）有限公司、鸿彧智能科技（山东）有限公司、华兴中科标准技术（北京）有限公司。

本文件主要起草人：张航、刘书博、孙浩轩、朱世顺、魏兴慎、朱孟江、王涵、肖雨果、于向荣、张吉、陈文弢、静静、马若龙、邵彦华、刘欣然、董婧一、董坤、陈绍东、付帽林、张嘉欣、李华、任国静、丁月。

引　言

当前人工智能技术已进入规模化产业应用阶段，算法黑箱、决策逻辑不透明、可解释性不足等问题，已成为制约AI技术安全合规应用与保障用户合法权益的重要挑战。本文件旨在建立统一的AI算法可解释性与决策透明度技术框架与评估方法，适用于各类场景与部署环境。

AI 算法可解释性与决策透明度通用技术规范

1 范围

本文件规定了AI算法可解释性与决策透明度的技术要求，描述了评估方法。

本文件适用于各类AI系统的全生命周期过程，包括需求分析阶段的可解释性目标锚定、数据采集与预处理阶段的透明度基础构建、模型设计与训练阶段的可解释性嵌入、部署运行阶段的解释生成与透明度公开、监控迭代阶段的可解释性优化及透明度更新、评估审计阶段的合规性验证。

2 规范性引用文件

本文件没有规范性引用文件。

3 术语和定义

下列术语和定义适用于本文件。

3.1

可解释性 interpretability

AI系统能够以人类可理解的方式,清晰说明其决策过程逻辑、结果形成依据及不确定性边界的能力，包含局部解释（针对单个决策结果）和全局解释（针对整体模型行为）两类核心维度。

3.2

决策透明度 decision-making transparency

AI系统在决策全流程中，向相关方（含用户、监管机构、第三方审计机构）公开数据来源、算法逻辑、决策链路及结果影响等关键信息的程度与能力。

3.3

算法偏见 algorithmic bias

AI算法因训练数据分布不均衡、特征选择偏差或模型设计缺陷，导致对不同群体产生差异化决策结果的现象。

3.4

决策链路 decision chain

AI系统从数据输入、特征工程、模型推理到结果输出的完整决策执行路径，包含所有影响最终结果的关键节点与逻辑规则。

4 算法可解释性技术要求

4.1 基本要求

4.1.1 AI 系统应根据应用领域的风险等级提供相应级别的可解释性支持。高风险领域应提供全面可解释性，中风险领域应提供关键决策因素解释，低风险领域可提供基础可解释性。

4.1.2 算法可解释性应贯穿 AI 系统全生命周期，核心遵循“风险适配、精准匹配、全程可溯”原则，具体要求如下。

a) 全生命周期嵌入性：

1) 需求分析阶段应结合业务场景、风险等级、合规要求，明确可解释性的核心指标、目标受众、交付形式，形成《AI 系统可解释性需求规格说明书》；

2) 数据采集与预处理阶段应同步完成特征可解释性标注，明确每个输入特征的业务含义、取值范围、对决策结果的影响逻辑；

3) 模型设计与训练阶段应优先选择具备内在可解释性的模型架构，复杂深度学习模型应同步设计配套的事后解释方案，模型训练过程中应同步验证解释准确率、特征覆盖率等核心指标；

4) 部署运行阶段应将解释生成模块与模型推理模块同步部署，解释模块的运行不得影响主模型的推理性能与业务稳定性；
5) 监控迭代阶段应持续监控解释模块的运行状态、解释准确率波动、用户对解释结果的异议率，建立可解释性指标的常态化监控告警机制，当指标超出阈值时及时触发优化流程；
6) 评估审计阶段应将可解释性指标纳入 AI 系统常态化评估审计范围，留存完整的测试数据、评估报告、整改记录，支撑监管核查与第三方审计。

b) 结果准确性：
1) 解释结果应与模型真实决策逻辑一致；
2) 特征重要性量化误差应控制在±5%以内；
3) 风险领域应增加基于人类专家共识的校验机制，通过双盲专家评估进行验证。

c) 时效匹配性：自动驾驶等实时决策场景解释生成耗时不超过 1 s，批量信用评估等非实时场景不超过 5 min。

d) 无偏见性与地域适配：解释过程应规避放大模型固有偏见，对涉及不同用户群体的决策，应说明敏感属性的影响权重，解释内容应以受众可理解的形式呈现。

注：内在可解释性指模型自身架构具备天然可理解性，无需额外解释方法即可直接解读决策逻辑的属性，常见于线性回归、决策树等结构简单的模型。

4.1.3 技术方法选型应遵循性能适配、场景匹配、成本可控的原则，优先选择与模型架构、业务场景、风险等级高度适配的方法。针对多部署环境的 AI 模型，应选择支持多样化数据分布与用户群体特征的通用解释技术。

4.2 技术方法

算法可解释性技术方法分为全局解释与局部解释两类，应根据模型类型、风险等级及应用场景选择适配方案，具体参数见表1。

表 1 算法可解释性技术方法

方法类型	具体方法	核心参数要求	优势与局限
全局解释方法	内在可解释模型	模型参数可直接解读（如回归系数绝对值不低于0.1为关键特征）	优势：原生可解释，无额外计算成本； 局限：不适用于复杂深度学习模型
	SHAP全局摘要图	特征重要性排序一致性不低于90%，样本覆盖度不低于95%	优势：量化特征全局贡献，支持多模型适配； 局限：高维数据计算耗时较长，应优化算力
局部解释方法	部分依赖图（PDP）	特征边际效应曲线平滑度不低于0.8，异常点占比不超过3%	优势：直观展示单特征影响趋势； 局限：无法体现特征间交互作用
	累积局部效应图（ALE）	效应值计算误差不超过5%，局部区域样本量不低于50个	优势：消除数据分布干扰； 局限：可视化对高维数据适配性差
	LIME（局部可解释模型）	局部模型拟合度不低于0.85，解释文本简洁性评分不低于3.5分（5分制），多语种翻译准确率不低于98%	优势：模型无关，适配性强； 局限：解释结果受局部样本分布影响
	SHAP值局部解释	单样本特征贡献量化误差不超过4%，正负影响区分准确率不低于98%，决策ID与解释结果绑定率100%	优势：理论严谨，支持正负影响量化； 局限：需要专业工具支撑（如SHAP库）
	Grad-CAM（梯度加权类激活图）	关键区域定位准确率不低于90%，激活图信噪比不低于3:1，跨设备适配性不低于95%	优势：直观定位视觉关键区域； 局限：仅适配CNN类模型
	注意力机制解释	关键Token注意力权重不低于0.2，语义关联度不低于	优势：适配Transformer模型，贴合语义理解；

方法类型	具体方法	核心参数要求	优势与局限
		0.85，多语种语义一致性不低于96%	局限：存在“伪注意力”现象，注意力权重不等同于因果解释，局限性等级：中高

4.3 动态调整要求

算法可解释性应随系统迭代、场景变化更新动态优化，确保持续适配业务需求，具体调整规则如下。

e) 模型迭代调整：
   1) 模型版本迭代（如架构升级、训练数据更换）后，应验证原有解释方法的适配性；
   2) 若模型核心逻辑未变更，应测试解释准确率变化；
   3) 若核心逻辑变更，应重新选型解释方法并完成全流程验证。

f) 偏差问题调整：监测到模型存在显著偏见时，应启动解释优化，补充偏见影响量化解释模块，说明偏见来源及对决策结果的具体影响程度，并同步至透明度披露内容。

g) 合规要求变更调整：当业务覆盖区域的法律法规、监管要求发生变更，应进行可解释性方案的合规适配调整，补充对应区域的专项解释模块、多语种支持、合规性说明文档，确保符合当地监管要求。

h) 用户反馈调整：当单月内同一决策场景的用户解释异议率不低于 5%时，应启动解释方案优化，分析异议集中原因，优化解释呈现形式与内容深度。优化后应完成用户理解度验证。

i) 技术升级调整：当可解释性技术出现新型高效解释算法发布等重大突破，应在 3 个月内完成技术适配测试，结合业务规划逐步升级。

4.4 解释结果呈现要求

4.4.1 解释结果应根据目标受众类型适配呈现形式与内容深度，确保“按需易懂”，具体要求见表 2。

表 2 解释结果呈现要求

目标受众	呈现形式	内容深度要求	示例
普通用户	自然语言描述+极简可视化（如环形图）	规避技术术语，聚焦“决策结果+核心影响因素+通俗原因”，单条解释文本不超过200字	信贷审批拒绝：“您近6个月信用卡逾期2次（规则阈值不超过1次），是本次审批未通过的主要原因，月收入稳定性（当前评分B）也有一定影响”
技术开发/运维方	结构化报告+专业图表（如SHAP热力图、决策树可视化）	包含“模型版本+特征重要性排序（含权重值）+解释方法参数+误差范围”	医疗影像诊断AI解释报告：“模型V3.2，肺结节识别关键特征权重：结节边缘清晰度（0.38）＞大小（0.25）＞密度（0.19），解释方法LIME，误差±3.2%”
监管机构	可追溯文档+验证数据	覆盖“解释方法合规性说明+解释结果与模型逻辑一致性验证数据+历史解释记录索引”	自动驾驶决策解释：“采用Grad-CAM+决策路径图，近3个月解释结果与模型实际决策逻辑一致性不低于98.7%，历史解释记录可通过决策ID（如AD20240501-001）查询”

4.4.2 解释结果应满足隐私保护要求：

a) 不得泄露训练数据中的敏感信息，应对涉及个人信息的内容进行脱敏；

b) 技术方获取的解释数据不得用于模型训练以外的用途，应留存访问日志。

4.5 解释可验证性要求

4.5.1 技术验证机制

应提供解释过程的复现接口/工具，支持技术方通过输入相同测试样本，复现解释结果，每季度开展解释方法有效性验证，由内部技术团队或第三方机构出具《解释可验证性报告》，记录验证样本量、复现率、偏差原因。

4.5.2 用户可验证渠道

普通用户对解释结果存疑时，可申请解释复核，系统应提供补充解释材料，或人工复核反馈。

4.5.3 特定场景算法的可解释性补充要求

针对当前主流 AI 场景，应额外满足以下场景化解释要求。

a) 生成式 AI：
   1) 应解释生成内容与输入 Prompt 的关联逻辑；
   2) 若生成内容涉及事实性信息，应标注信息来源可信度及不确定性范围；
   3) 若生成内容涉及版权素材，应标注素材来源、授权范围及使用合规性说明。
b) 大语言模型（LLM）AI：
   1) 应优先采用思维可解释性技术，公开模型推理的分步逻辑与中间结论，解释内容应与模型实际推理步骤一致；
   2) 采用检索增强生成技术的大模型，应完整追溯生成内容的来源，标注引用片段对应的知识库条目 ID、原文位置及可信度评级；
   3) 应披露模型推理过程中的截断规则、采样策略及温度参数对生成结果的影响，明确模型输出的不确定性边界；
   4) 风险场景应用的大模型应同步部署事后可解释性验证模块，对关键决策的推理链路进行交叉核验。
c) 多模态融合决策 AI：
   1) 应分别解释各模态输入对最终决策的贡献权重，量化不同模态信息的影响程度；
   2) 应公开多模态特征融合的核心逻辑，说明跨模态信息对齐、冲突消解的规则与过程；
   3) 应标注决策结果对单一模态输入的敏感性阈值，明确关键模态缺失或异常时的决策降级机制；
   4) 风险场景的多模态决策应提供各模态独立的解释结果及融合后的综合解释报告。
d) 强化学习 AI：
   1) 应解释决策动作与环境状态的关联；
   2) 应记录关键决策的“探索—利用”权衡过程；
   3) 应披露决策过程中的安全约束规则。

5 决策透明度技术要求

5.1 数据透明度

5.1.1 数据透明度应覆盖数据全生命周期，重点适配数据流转场景，具体要求如下。

a) 公开数据质量指标：训练数据准确率不低于 98%应额外提供数据质量一致性报告。
b) 向监管机构、合作方提供核心特征分布报告，包括数值型特征的均值、中位数、分位数以及分类特征的类别占比。场景应按敏感属性或业务维度拆分分布数据。
c) 建立 AI 系统数据全生命周期台账，完整记录数据采集、存储、预处理、训练、验证、归档、销毁的全流程信息。台账内容应包含数据批次、来源、数量、处理规则、操作人员、操作时间、流转记录，台账数据不可篡改，留存期限不短于 AI 系统全生命周期+5 年。

5.1.2 数据分布与质量说明要求如下。

a) 数据脱敏应留存脱敏前后的映射关系备案（仅对监管机构开放）。敏感数据应分类说明脱敏方法：
   1) 个人信息类采用前 6 后 4 脱敏、模糊化处理；
   2) 商业秘密类采用聚合统计脱敏。
b) 应公开关键预处理步骤，包括缺失值处理、异常值剔除、特征编码。不同数据源的预处理差异应在透明度信息中说明。

5.1.3 数据预处理与脱敏说明要求如下。

a) 涉及数据流转时，应详细披露数据出境目的、流转路径、安全保障措施。风险场景应提供数据流动安全评估报告编号。

b） 公开训练/验证数据的核心来源，分类标注公开数据集、自有采集数据、第三方采购数据：
1） 公开数据集应标注名称及版本；
2） 自有采集数据应说明采集场景及合规依据；
3） 第三方数据应标注供应商资质及数据授权证明编号。

5.1.4 数据质量监控的透明度要求：应按月公开数据质量监控报告，披露训练/验证数据的准确率、缺失率、重复率、标签一致性等核心指标的波动情况。针对数据质量异常问题，应披露异常原因、整改措施及整改效果，风险场景应按周开展数据质量监控并向监管机构报送。

5.2 算法透明度

5.2.1 算法透明度应确保相关方清晰了解算法核心信息，兼顾技术细节与可读性。

5.2.2 算法基础信息公开：
a） 公开核心算法类型及模型架构；
b） 应补充关键参数范围；
c） 相关算法应标注关键环境适配参数。

5.2.3 算法性能与局限性的透明度要求：
a） 应公开 AI 系统的核心性能指标，包括准确率、召回率、精确率、F1 值、鲁棒性等，明确不同场景、不同环境下的性能波动范围；
b） 应如实披露算法的固有局限性，包括适用场景边界、环境约束条件、性能衰减场景、潜在决策风险。

5.2.4 算法训练过程的透明度要求：风险领域 AI 系统应公开模型训练的核心流程，包括训练环境、训练轮次、损失函数设计、优化器选型、正则化策略、验证集划分规则。应补充训练过程的关键节点日志。

5.3 决策过程透明度

5.3.1 决策过程透明度应全链路可视、可追溯、可干预。

5.3.2 流程可视化：
a） 以流程图与文字说明相结合的形式，公开从数据输入至结果输出的决策全流程；
b） 应标注关键节点的人工干预机制。

5.3.3 决策追溯：
a） 为每一项决策生成唯一 ID，支持查询：输入数据快照、调用的模型版本、特征重要性结果、决策时间；
b） 应建立决策链路的全流程审计机制，针对每一条决策链路，均可逆向追溯从数据输入到结果输出的每一个关键节点的计算逻辑、参数取值、规则应用情况。

5.3.4 人工干预全流程记录要求：所有人工干预 AI 决策的行为，均应生成完整的干预记录，包含干预人、干预时间、干预触发条件、干预前的 AI 决策结果、干预后的最终决策结果、干预原因、审批记录，干预记录与对应决策 ID 绑定，不可篡改，留存期限与决策追溯期限一致。

5.4 信息公开范围与权限控制

5.4.1 相关要求见表 3。

表 3 信息公开范围与权限控制

相关方	可获取的透明度信息	权限控制方式
普通用户	自身决策结果的解释、数据来源合法性说明、异议处理渠道	账号登录验证（如手机号验证码、人脸识别）
技术开发方	完整的算法参数、数据预处理细节、模型迭代记录	企业内部权限分级（如开发岗、运维岗）
监管机构	全量透明度信息（含用户数据分布、算法局限性报告）	专用监管接口+身份认证（如CA证书）
公众	算法类型、应用场景、整体决策偏差率（匿名化）	官网公示（无需登录）

5.4.2 权限分级管理细则：应建立最小权限原则的分级授权体系，针对不同岗位、不同主体设置精细化的信息访问权限，严禁超权限访问；所有访问透明度信息的行为均应生成不可篡改的访问日志，包含

访问主体、访问时间、访问内容、访问 IP、操作行为，日志留存期限不低于 7 年。

5.5 决策结果反馈与解释获取机制

主动反馈要求：

a) AI 系统做出决策后，应主动向用户推送决策结果及解释获取入口；
b) 风险领域决策应同步推送结果和基础解释，基础解释应包含核心影响因素、异议申请渠道。

5.6 透明度信息动态更新要求

明确场景限制、公开数据与环境限制。应声明 AI 系统的合规边界、数据层面、算法层面、决策规则层面、更新告知义务、更新记录留存、建立透明度信息的版本管理体系。

5.7 算法偏见的透明度披露要求

5.7.1 偏见评估与披露：

a) AI 系统应每季度开展算法偏见评估，计算不同群体的决策结果差异；
b) 偏见评估应覆盖单一敏感属性与多敏感属性交叉的场景,针对面向多样化用户群体的 AI 系统，应额外评估不同用户特征间的决策差异，确保决策公平性；
c) 若评估发现显著偏见（差异率不低于 15%），应在透明度信息中披露偏见表现、可能原因、改进计划。改进计划应明确时间节点。

5.7.2 偏见缓解的透明度：实施偏见缓解措施后，应公开缓解措施细节、缓解前后的偏见差异率变化、验证数据，接受监管机构监督。

6 评估方法

6.1 可解释性评估

6.1.1 评估维度与指标

具体要求见表4。

表 4 评估维度与指标

评估维度	核心指标	计算方式
准确性	解释准确率	对比解释结果与模型真实决策逻辑的一致性
可理解性	用户理解度评分	问卷调查
完整性	关键特征覆盖率	（被解释的关键特征数量/模型实际依赖的关键特征数量）×100%
时效性	解释生成耗时	多次测试取平均值

6.1.2 评估流程

评估流程应符合以下规定。

a) 评估准备：
   1) 确定评估对象、风险等级及目标受众；
   2) 收集资料：模型架构文档、训练数据样本、可解释性技术方法说明；
   3) 制定评估方案：明确指标权重、测试样本范围、测试工具与方法。风险场景的评估方案应邀请外部合规专家、技术专家进行外部评审，评审通过后方可启动正式评估。
b) 解释生成与测试：
   1) 选取测试样本（覆盖正常/异常场景，样本量不低于 1000），用目标可解释性方法生成解释结果；
   2) 借助自动化工具计算解释准确率、关键特征覆盖率、解释生成耗时；
   3) 采用双盲测试方式开展验证。
c) 用户调研：
   1) 针对目标受众设计问卷，回收有效问卷不低于 100 份；
   2) 计算用户理解度评分，取平均分。

d) 结果判定：
   1) 所有指标满足对应风险等级要求，判定为可解释性合格；
   2) 存在不满足项，应分析原因并制定改进方案。

6.2 透明度评估

6.2.1 评估维度与指标

透明度评估围绕信息公开质量、可获取性、合规性构建指标体系，重点核查数据披露、多区域权限控制等场景，评估周期与可解释性评估同步。

表 5 评估维度与指标

评估维度	核心指标	计算方式
信息公开质量	信息覆盖度	（已公开的透明度信息项数/本文件第5章节要求的信息项数）×100%
信息准确率	不低于98%	—
信息可获取性	获取时效达标率	（在规定时限内获取到目标信息的次数/总尝试次数）×100%
权限控制准确率	不低于100%	权限同步误差不超过24 h
合规与响应能力	异议处理达标率	（在规定时限内处理完成并反馈的异议数/总异议数）×100%
审计资料完整性	不低于100%	资料格式适配国际审计标准（如ISO/IEC 42005要求）

6.2.2 评估流程

评估流程应符合以下规定。

a) 评估启动阶段：
   1) 确定评估范围：涵盖数据、算法、决策过程全环节，明确评估重点；
   2) 组建评估小组：包含技术专家、合规专员、业务负责人，第三方审计应额外配备注册审计师；
   3) 制定评估方案：明确指标权重、测试工具。
b) 信息核查阶段：
   1) 公开信息核查：对照第 5 章要求，核查官网、公示平台等渠道的信息覆盖度与准确率，不同用户群体应能获得与其认知水平相适应的信息披露；
   2) 后台数据验证：调取系统后台数据，与公开信息比对，验证一致性；
   3) 权限与追溯测试：模拟不同相关方登录系统，测试信息获取流程及时效，验证权限控制准确性；通过决策 ID 测试追溯功能完整性；
   4) 专项核查：对已部署的 AI 系统，专项核查信息披露的合规性、准确性、权限控制一致性与数据流转披露完整性。
c) 响应能力验证阶段：
   1) 异议处理模拟：选取不同风险等级的异议案例，测试处理流程及时效，评估反馈内容的合理性；
   2) 审计支撑验证：向被评估方索要近 1 年审计资料，核查完整性与规范性，场景应验证区块链存证数据的不可篡改性。
d) 结果形成阶段：
   1) 指标计分：按评估方案权重计算各指标得分，综合得分达到规定评分标准判定为优秀、合格或不合格；
   2) 问题定位：对不合格指标，分析根源；
   3) 出具报告：明确评估结论、问题清单及整改建议，场景报告应经监管机构复核。

6.3 评估结果应用

6.3.1 评估结果应与 AI 系统准入、迭代审批挂钩：系统评估不合格不得上线，已上线系统应暂停使用

并限期整改；中风险系统评估不合格应暂停核心决策功能，限期整改并完成复核后方可恢复使用；低风险系统评估不合格应进行优化整改，整改期间应向用户公示评估不合格项及整改进度。

6.3.2 建立问题整改跟踪机制：针对评估发现的不合格项，应明确整改责任人、整改时限、整改措施，建立“问题发现—整改实施—复核验证—闭环归档”的全流程管理机制。整改完成后应提交佐证材料，由评估机构抽样复核；对于风险场景的重大不合格项，应实施挂牌督办，整改进度每周向监管机构报送。

6.3.3 评估结果公示与共享要求：AI 系统可解释性与透明度评估结果（脱敏后）应在公示平台披露，接受社会监督。鼓励行业内建立评估结果共享机制。

6.3.4 留存评估档案：所有评估资料应采用区块链存证、数字签名或 WORM 存储等可验证的防篡改技术。

6.3.5 持续改进要求：应基于历次评估结果，建立 AI 可解释性与透明度管理的持续改进机制，每年度复盘评估中发现的共性问题、高频风险点。应跟踪国内外可解释性技术发展、监管政策更新。

6.3.6 评估结果互认与适配要求：

a) 针对运营的 AI 系统，评估结果应适配监管规则，可根据要求出具专项评估报告；
b) 应积极对接 AI 治理标准体系与监管互认机制；
c) 针对差异化的监管要求，应基于评估结果建立适配方案。

6.3.7 人员能力与体系建设要求：

a) 应基于评估结果，针对 AI 研发、运营、合规、审计等相关岗位人员开展专项培训，培训内容包括可解释性技术要求、透明度管理规范、评估指标体系、合规要求等。培训考核结果与岗位准入挂钩；
b) 应根据评估中发现的体系短板，完善企业内部 AI 治理体系，将可解释性与透明度管理纳入企业 AI 伦理规范、数据安全管理体系、算法全生命周期管理制度，形成闭环管理；
c) 应建立专业的评估人才队伍，配备具备 AI 技术、合规管理、审计监督、业务等专业能力的专职人员，负责内部评估工作的落地实施，第三方评估机构应具备相应的专业资质与技术能力。

6.3.8 监督与问责机制：

a) 企业应建立内部监督机制，由合规部门、审计部门对可解释性与透明度评估工作的真实性、客观性、规范性开展常态化监督，对评估过程中的弄虚作假、瞒报漏报行为严肃追责；
b) 应明确各环节的主体责任，AI 研发团队对可解释性技术落地负责，运营团队对透明度信息公开与用户反馈处理负责，合规团队对合规适配与评估工作负责，评估结果与相关团队、人员的绩效考核直接挂钩；
c) 因可解释性与透明度不达标导致合规风险、用户权益受损、安全事件的，应严肃追究相关责任人的管理责任与直接责任；造成重大损失或恶劣影响的，应按规定向监管机构报告，并依法承担相应法律责任。

6.3.9 应急处置要求：

a) 评估过程中发现 AI 系统存在重大可解释性缺陷、透明度严重缺失，可能引发重大合规风险、安全风险的，应立即启动应急处置预案，暂停相关 AI 系统的决策功能，第一时间向监管机构报送相关情况；
b) 针对评估发现的系统性、行业性风险，应及时向行业协会、监管机构反馈，配合开展行业风险排查与治理工作，同步优化自身 AI 系统的可解释性与透明度方案；
c) 应建立评估结果与 AI 系统应急处置的联动机制，将评估中发现的高频风险点、缺陷纳入 AI 系统应急预案，明确应急触发条件、处置流程、责任主体，确保风险快速响应、闭环处置。

附 录 A
（资料性）
跨区域/多场景部署下的AI算法可解释性与透明度应用指南

A.1 概述

本文件正文部分规定了AI算法可解释性与决策透明度的通用技术要求和评估方法。附录A旨在为部署于跨区域、多场景环境下的AI系统提供应用参考。本附录不构成强制性要求，仅作为实施参考。

A.2 跨区域部署下的参考要点

A.2.1 数据来源与合规差异

AI系统在跨区域部署时，不同司法管辖区的数据法律法规存在差异（如中国《个人信息保护法》、欧盟《通用数据保护条例》（GDPR）等）。本文件要求的信息披露应兼顾区域合规性要求，具体适配可根据适用法规调整。建议实施方参考ISO/IEC 27701及ISO/IEC 38507的相关指引。

A.2.2 语言与文化适配

当AI系统服务于多语言用户群体时，本文件中的“可理解性”要求应进一步细化：解释内容的语言选择、表达方式、复杂度应适配目标受众的语言能力和认知水平。宜参照ISO 9241-110中“适合任务、适合用户”的交互设计原则。

A.2.3 多地监管报送

AI系统在跨区域部署时，可能需要向多个监管机构提交不同格式、不同详略程度的可解释性与透明度报告。实施方宜建立统一的信息披露框架，按监管要求分层输出。本文件“区分监管报送与公众公开”的分层原则可拓展至多监管机构场景。

A.3 多场景复用模型的应用参考

A.3.1 模型复用的一致性验证

同一AI模型部署于不同业务场景时，应验证解释方法在各场景中的有效性。推荐采用“基准场景+扩展校验”策略：选择一个代表性场景作为基准验证，再通过抽样测试确认迁移场景中的解释一致性。

A.3.2 风险映射调整

同一AI系统在不同场景下的风险等级可能不同。宜建立场景-风险映射表。

参 考 文 献

[1] ISO/IEC 22989 信息技术 人工智能 人工智能概念与术语（Information technology-Artificial intelligence-Artificial intelligence concepts and terminology）

[2] ISO/IEC TR 24028 信息技术 人工智能 人工智能可信度概述（Information technology-Artificial intelligence-Overview of trustworthiness in artificial intelligence）

[3] ISO/IEC 27701 安全技术 隐私信息管理体系 要求与指南（management-Requirements and guidelines）

[4] ISO/IEC 38507 信息技术 IT治理 组织使用人工智能的治理含义（Information technology-Governance of IT-Governance implications of the use of artificial intelligence by organizations）

[5] ISO/IEC 42001:2023 信息技术 人工智能 管理体系（Information-technology Artificial intelligence-Management system）