ICS 35.240.01
CCS L 77

T/ISC

团　　体　　标　　准

T/ISC XXXX—XXXX

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

（征求意见稿）

在提交反馈意见时，请将您知道的相关专利连同支持性文件一并附上。

XXXX－XX－XX 发布　　XXXX－XX－XX 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。

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、中科数测科技有限公司、华兴中科标准技术（北京）有限公司。

本文件主要起草人：刘书博、朱世顺、魏兴慎、张浩天、王涵、肖雨果、于向荣、张吉、陈文弢、静静、马若龙、邵彦华、刘欣然、董坤、董婧一、成瑾、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与跨境落地阶段，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、保障用户合法权益、防范跨境合规风险的核心瓶颈。为规范AI算法可解释性与决策透明度的技术实现、管理要求与评估方法，填补跨行业、跨场景、跨境运营场景下的通用技术规范空白，支撑各类主体落实AI安全治理主体责任，适配国内外AI监管合规要求，特制定本文件。

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和评估方法。

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。

本文件不替代高风险领域专项法规及国际区域性合规要求中关于可解释性与透明度的更严格规定；当行业专项法规或跨境合规要求有更具体要求时，优先遵循专项法规及区域合规要求，同时本文件提供技术适配支撑。

2 规范性引用文件

下列文件中的内容通过文中的规范性引用而构成本文件必不可少的条款。其中，注日期的引用文件，仅该日期对应的版本适用于本文件；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 信息安全技术 个人信息安全规范

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 explainability

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，如某用户信贷审批拒绝的具体原因）和全局解释（针对整体模型行为，如推荐系统整体特征权重分布）两类核心维度。

3.2

决策透明度 decision transparency

AI系统在决策全流程中，向跨境相关方（含用户、监管机构、跨国运营主体、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

风险等级 risk level

根据AI系统应用领域的潜在危害程度、影响范围（含跨境影响）及受众规模划分的层级，用于确定可解释性与透明度的适配要求，具体分为：

a) 高风险：应用于医疗诊断、金融信贷风控、L3 及以上自动驾驶、国际骨干网 AI 调度等领域，决策失误可能导致人身伤害、重大财产损失或跨境通信安全风险；

b) 中风险：应用于推荐系统、跨境云服务资源分配、企业 SaaS 服务 AI 决策等领域，决策失误可能导致用户体验下降或业务效率损失；

c) 低风险：应用于娱乐内容生成、普通用户行为分析等领域，决策失误对用户或业务影响较小。

3.4

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（如不同地域、年龄、职业的跨境用户）产生差异化决策结果的现象。

3.5

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（如医疗诊断、自动驾驶）应提供全面可解释性，中风险领域（如推荐系统、信用评估）应提供关键决策因素解释，低风险领域（如娱乐应用）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) 风险适配性：
   1) 高风险场景（如国际骨干网 AI 调度、医疗诊断、自动驾驶）应采用“全局+局部”双维度解释，解释结果需通过跨境第三方验证；
   2) 中风险场景（如跨境云资源分配、企业 SaaS 服务 AI 决策）可采用“局部优先+全局摘要”模式；
   3) 低风险场景（如娱乐内容生成）可简化为基础局部解释。
b) 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、跨境合规要求，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》，作为后续研发、测试、评估的核心依据；
   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，规避不可解释的黑盒特征工程，确保特征溯源可查；
   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，确保模型性能与可解释性同步达标；
   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，确保决策结果与解释内容同步生成、同步存储，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。
c) 结果准确性：
   1) 解释结果应与模型真实决策逻辑一致，高风险领域解释准确率不低于 95%，中风险不低于 90%，低风险不低于 85%；
   2) 特征重要性量化误差应控制在±5%以内（高风险场景±3%），跨境通信 AI 调度场景应额外满足“跨区域解释一致性不低于 98%”；
   3) 高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，即由不少于 3 名领域专家对解释结果的合理性进行评分，评分均值应不低于 4 分（5 分制）。
d) 时效匹配性：实时决策场景（如自动驾驶、国际通信质量实时优化）解释生成耗时不超过 1 s，非实时场景（如批量信用评估、跨境云资源月度调度）不超过 5 min，避免因解释过程影响业务流程效率。
e) 无偏见性与跨境适配：解释过程应规避放大模型固有偏见，对涉及不同地域、语种、合规体系的跨境用户决策，应额外说明地域等敏感属性的影响权重（如“地域因素对本次跨境云资源分配决策影响权重不超过 2%”），解释文本应支持多语种适配（至少含中、英双语）。

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”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，避免过度追求复杂技术导致的算力浪费与业务延迟；针对跨境多场景复

用的 AI 模型，应选择支持跨区域数据分布适配、多语种解释生成的通用解释技术。

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、风险等级及跨境应用场景选择适配方案，具体参数如下：

表 1 算法可解释性技术方法

方法类型	具体方法	适用场景	核心参数要求	优势与局限
全局解释方法	内在可解释模型	低/中风险、简单任务（如基础分类、线性预测）	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	中/高风险、复杂模型（如随机森林、Transformer，含跨境通信AI调度模型）	特征重要性排序一致性不低于90%，样本覆盖度不低于95%，跨境场景跨区域一致性不低于98%	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）	中风险、特征交互少的场景（如信用评分、跨境云资源基础分配）	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）	高风险、数据分布不均衡场景（如医疗影像分类、跨境用户行为分析）	效应值计算误差不超过5%，局部区域样本量不低于50个，跨境数据分布适配性不低于92%	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）	全风险等级、任意模型（如推荐系统、跨境客服对话AI）	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	高风险、复杂决策场景（如自动驾驶、国际骨干网AI调度）	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）	高风险、计算机视觉场景（如病灶检测、跨境安防监控）	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	中/高风险、自然语言处理场景（如智能合同审核、跨境多语种客服）	关键Token注意力权重不低于0.2，语义关联度不低于0.85，多语种语义一致性不低于96%	优势：适配Transformer模型，贴合语义理解；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。 **注：**该方法不应作为高风险场景的单一解释手段，需与其他可解释性方法结合使用

4.3 动态调整要求

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

a) 场景风险变更调整：
   1) 当 AI 系统从低风险场景迁移至中高风险场景（如从娱乐推荐转为跨境金融营销推荐），应在 15 个工作日内完成解释方案升级，新增全局解释维度及第三方验证环节；
   2) 从高风险转为中低风险时，可简化解释流程，但应留存简化依据备案；
   3) 当 AI 系统出现以下任一触发条件时，应在 7 个工作日内启动风险等级重新评估，服务用户规模较上线时增长超过 10 倍或单月活跃用户突破 1000 万，连续 30 天决策失误率超

过对应风险等级阈值的 2 倍，发生 1 起及以上因算法决策导致的重大安全事件或合规事件。业务覆盖范围新增高风险应用领域或跨境高风险监管区域。监管机构或第三方审计机构提出风险等级调整建议。评估完成后应在 3 个工作日内更新风险等级标识，并同步调整可解释性与透明度技术方案，评估报告及调整记录留存期限不低于 AI 系统全生命周期+5 年。

b) 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，应在 7 个工作日内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（波动不超过 5%为合格）；
   3) 若核心逻辑变更（如从 CNN 改为 ViT），应重新选型解释方法并完成全流程验证。
c) 偏差问题调整：监测到模型存在显著偏见（如不同地域用户决策差异率不低于 15%）时，应在 3 个工作日内启动解释优化，补充“偏见影响量化解释”模块，说明偏见来源（如训练数据地域分布不均衡）及对决策结果的具体影响程度，并同步至透明度披露内容。
d) 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（如区域 AI 监管法案正式实施、新增数据跨境合规要求），应在法规生效前 30 个工作日内完成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。
e) 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，应在 5 个工作日内启动解释方案优化，分析异议集中原因（如解释内容不易懂、解释逻辑与用户认知不符），优化解释呈现形式与内容深度，优化后需完成用户理解度验证，确保用户理解度评分提升至对应风险等级要求以上。
f) 技术升级调整：当可解释性技术出现重大突破（如新型高效解释算法发布），高风险场景应在 3 个月内完成技术适配测试，验证通过后落地应用；中低风险场景可结合业务规划逐步升级，升级周期不超过 6 个月，跨境场景应优先适配支持多语种、跨合规体系的解释技术。

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”，具体要求如下：

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（如用户身份证号、医疗隐私数据），应对涉及个人信息的内容进行脱敏（如“用户 A”替代真实姓名，“收入区间 5k-8k”替代具体收入）；
b) 技术方获取的解释数据（如特征权重）不得用于模型训练以外的用途，应留存访问日志（含访问人、时间、内容）。

4.5 解释可验证性要求

4.5.1 技术验证机制

4.5.1.1 中高风险领域 AI 系统，应提供解释过程的复现接口/工具（如开源解释库适配包、API 调用文档），支持技术方通过输入相同测试样本，复现解释结果（复现准确率：高风险不低于 99%，中风险不低于 95%）。

4.5.1.2 高风险领域（如医疗、自动驾驶）应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”，系统应在规定时限内（高风险场景不超过24 h，中低风险不超过72 h）提供补充解释材料（如更详细的特征影响说明），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（如文本生成、图像生成）：
   1) 应解释“生成内容与输入 Prompt 的关联逻辑”（如通过注意力权重可视化展示 Prompt 关键词对生成内容的影响，如“科幻风格”关键词影响生成图像的色彩占比 35%）；
   2) 若生成内容涉及事实性信息（如新闻摘要、知识问答），应标注“信息来源可信度”（如“基于公开文献[文献 ID]生成，可信度评级 A”）及“不确定性范围”（如“该历史事件描述的准确率约 92%，存在 1～2 处细节应进一步验证”）；
   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，如“生成图像包含来自[素材库名称]的授权素材，授权编号 XXX，商用授权范围覆盖全球”。
b) 大语言模型（LLM）AI（如基于 GPT、LLaMA 等架构的通用大模型、行业大模型）：
   1) 应优先采用思维链（Chain-of-Thought）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（RAG）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) 高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，避免单一解释方法的局限性。
c) 多模态融合决策 AI（如融合文本、图像、语音、传感器数据的综合决策系统）：
   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) 高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（如自动驾驶、机器人控制）：
   1) 应解释“决策动作与环境状态的关联”（如通过状态—动作价值图展示“前方出现行人”状态下，“减速”动作的价值评分（0.92）高于“避让”（0.78）的原因）；
   2) 应记录关键决策的“探索—利用”权衡过程（如“本次选择‘新路线规划’是基于探索策略，历史该场景探索成功率不低于 68%，后续将根据反馈优化”）；
   3) 应披露决策过程中的安全约束规则，如“本次减速决策严格遵循自动驾驶安全规范，最小安全车距阈值 2 m，制动加速度不超过 4 m/s^2”。

5 决策透明度技术要求

5.1 数据透明度

5.1.1 数据透明度是决策透明度的核心基础，应覆盖数据全生命周期，重点适配跨境数据流转场景，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%、缺失率不超过 5%、重复率不超过 2%，高风险场景应额外提供“跨区域数据质量一致性报告”，确保数据分布均衡性（单一地域样本占比不超过 40%）。
b) 向监管机构、跨境合作方提供核心特征分布报告：数值型特征的均值、中位数、分位数；分类特征的类别占比，跨境场景应按地域维度拆分分布数据（如“东南亚用户样本占比 35%，欧洲用户样本占比 28%”）。
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、跨境流转记录，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) 敏感数据应分类说明脱敏方法：
   1) 个人信息类（如身份证号）采用“前 6 后 4 脱敏”“模糊化处理”；
   2) 商业秘密类（如跨境企业客户带宽需求）采用“聚合统计脱敏”；
   3) 跨境通信数据应符合数据所在地法规的脱敏要求，留存脱敏前后的映射关系备案（仅对监管机构开放）。
b) 公开关键预处理步骤：缺失值处理（如均值填充、删除）、异常值剔除（如 3σ 原则）、特征编码（如 One-Hot、LabelEncoder），跨境场景应额外说明不同区域数据的预处理差异（如欧盟数据采用“差分隐私脱敏”，东南亚数据采用“掩码脱敏”）。

5.1.3 数据预处理与脱敏说明要求如下。

a) 涉及跨境数据流转时，应详细披露“数据出境目的、流转路径（如从中国香港至新加坡数据中心）、安全保障措施（如加密传输标准 AES-256）”，高风险场景应提供跨境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（如 ImageNet-1K 2012 版）；
   2) 自有采集数据应说明采集场景及合规依据（如符合 GDPR 第 6 条、《个人信息保护法》第 13 条）；
   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，高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 算法透明度应确保跨境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（如监督学习—分类、强化学习—路径规划）及模型架构（如 CNN：5 层卷积+2 层全连接；Transformer：12 层编码器+6 层解码器）；
b) 高风险领域应补充关键参数范围（如随机森林的树数量：100 棵～200 棵，学习率：0.01～0.1）；
c) 跨境场景应标注算法的跨区域适配参数（如“欧洲区域通信延迟权重系数 1.2，东南亚区域 0.9”）。

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，不得夸大算法能力，不得隐瞒已知的算法缺陷。

5.2.4 算法训练过程的透明度要求：中高风险领域 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，高风险场景应补充训练过程的关键节点日志，支持监管机构追溯模型训练全流程。

5.3 决策过程透明度

5.3.1 决策过程透明度应实现“全链路可视、可追溯、可干预”，适配跨境多主体协同需求。

5.3.2 流程可视化：

a) 以流程图+文字说明形式公开全流程：输入数据（如用户信贷申请信息）→特征提取→模型计算→阈值判断（如信用分不低于600通过）→输出结果（如贷款审批通过/拒绝）；

b) 高风险领域应标注关键节点的人工干预机制（如医疗诊断结果应经医生复核）。

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) 追溯期限应符合行业法规（如金融领域不低于5年，医疗领域不低于10年）。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，确保决策过程无断点、无黑盒，审计追溯响应时间高风险场景不超过2 s，中低风险场景不超过5 s。

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，跨境场景应符合当地法规要求。

5.5 决策结果反馈与解释获取机制

5.5.1 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”（如APP弹窗、短信链接），推送时效：实时决策场景不超过1 min，非实时场景不超过24 h；

b) 高风险领域决策（如医疗诊断、贷款审批拒绝）应同步推送“结果+基础解释”，基础解释应包含“核心影响因素+异议申请渠道”（如“您的CT诊断结果为‘轻度肺炎’，关键依据是‘右肺下叶炎症浸润影’，若有异议可通过【我的—异议申请】提交”）。

5.5.2 解释获取便捷性：

a) 普通用户获取详细解释的操作步骤不得超过3步（如“登录APP→进入‘我的决策’→点击‘查看解释’”）；

b) 应支持多形式获取（如文字版、语音版（适配视障用户）、PDF下载版（适配监管存档））。

5.5.3 为每一项决策生成唯一“跨境决策ID”（如INT-AI-20240601-0001），支持相关方通过ID查询：输入数据快照（脱敏后）、调用的模型版本、特征重要性结果、决策时间、参与节点、操作人员（若有人工干预）。

5.5.4 追溯期限应符合5.3.3 b）要求。追溯系统应支持多区域访问，响应时间不超过3 s。

5.5.5 用户知情权保障要求：针对涉及用户重大权益的AI决策，除推送决策结果与基础解释外，还应主动告知用户享有申请详细解释、异议复核、人工干预的权利，以及对应的申请渠道、处理流程、时限

要求，确保用户知情权与救济权得到充分保障。

5.5.6 异议处理与反馈要求如下。

a) 公开异议申请渠道：线上申诉入口（支持多语种）、跨境客服电话（24h×7）、区域专属申诉邮箱；明确处理流程：申请（提交 ID 及异议理由）→核查（技术团队 + 合规团队联合核查）→反馈（出具书面解释报告）；

b) 设定差异化处理时限：高风险场景不超过 1 个工作日，中风险不超过 3 个工作日，低风险不超过 7 个工作日；跨境异议应明确“区域责任划分”（如欧洲区域异议由法兰克福运营中心处理）；

c) 异议处理结果应同步更新至透明度信息：如“针对决策 ID INT-AI-20240601-0001 的异议已处理，调整依据为补充用户跨境信用记录，重新评估后决策结果变更为通过”；

d) 应针对不同区域用户的使用习惯，提供本地化的解释获取与异议反馈渠道，包括当地主流的即时通讯工具、线下服务网点、区域专属客服团队，确保跨境用户可便捷、无障碍地获取相关服务。

5.6 高风险领域透明度专项要求

5.6.1 高风险领域（医疗、自动驾驶、金融风控）AI 系统，除满足基础透明度要求外，还应符合以下专项要求。

a) 公开算法开发与迭代主体：如“由 XXX 公司 AI 实验室联合 XX 高校开发”，明确责任主体及技术支持渠道。

b) 应在官网、产品服务界面的显著位置，持续公示高风险 AI 系统的责任主体、法定代表人、技术负责人、合规负责人、投诉举报渠道，明确主体责任，确保监管机构与用户可快速对接责任方。

c) 算法迭代记录与追溯：

1) 每年度应由具备资质的第三方机构（如国家认可的 AI 检测实验室、行业协会指定机构）开展透明度审计，审计内容包括“数据来源合规性、算法逻辑可追溯性、决策过程规范性”，审计报告应向监管机构备案并向社会公开（脱敏后）；

2) 若系统发生重大迭代（如模型架构调整、核心数据更换），应在迭代后 3 个月内补充专项审计；

3) 建立“算法版本迭代台账”，公开各版本信息：迭代时间、核心改进点（如 V2.0 优化损失函数为交叉熵+L2 正则）、性能变化（如准确率从 85%提升至 92%）、解释方法同步更新情况（如模型更换后，解释方法从 LIME 转为 SHAP）；

4) 迭代记录应关联决策案例：如“V3.0 版本优化后，跨境通信中断修复决策准确率提升 7%，典型案例见[案例 ID：INT20240601]”，支持相关方追溯版本变更影响。

5.6.2 应急透明度要求：

a) 发生决策失误或安全事件（如自动驾驶碰撞、医疗诊断偏差）时，应在 24 h 内启动“透明度应急响应”，向监管机构提供“事件相关的决策日志（含输入数据、模型状态、决策链路）、解释结果追溯报告”，并在 72 h 内公开事件初步分析（含“透明度信息是否完整”的说明）；

b) 发生 AI 决策重大安全事件，造成人身伤害、重大财产损失、跨境安全风险的，应在事件处置完成后 15 个工作日内，向社会公开完整的事件调查报告，包含事件发生原因、决策链路追溯结果、算法存在的缺陷、已采取的整改措施、后续风险防范机制，接受社会监督。

5.7 透明度信息动态更新要求

5.7.1 明确场景限制：如“自动驾驶算法在暴雨天气下准确率下降不低于 30%”“跨境通信 AI 调度在海底光缆中断场景下应人工干预”。

5.7.2 公开数据与环境限制：如“仅适配 18～60 岁用户的行为数据”“在网络延迟不低于 100 ms 的区域决策效率下降 20%”。

5.7.3 跨境场景应额外声明“合规边界”：如 “在欧盟区域不支持基于宗教、种族的特征决策”，避免算法应用越界。

5.7.4 当 AI 系统发生以下变化时，应在变化生效后 72 h 内更新透明度信息：

a) 数据层面：训练/验证数据来源新增/更换、数据预处理规则调整（如缺失值处理方法从“均值填充”改为“中位数填充”）；
b) 算法层面：模型版本迭代（如从 V2.1 升级至 V3.0）、核心参数调整（如随机森林树数量从 100 棵改为 200 棵）；
c) 决策规则层面：决策阈值调整（如信贷评分通过阈值从 600 分改为 620 分）、人工干预流程优化。

5.7.5 更新告知义务：当透明度信息发生重大更新，可能影响用户权益、决策逻辑、合规性的，应在更新生效前 7 个工作日，通过官网公告、产品内推送、短信通知等方式，向相关用户、合作方、监管机构告知更新内容、更新原因、生效时间，确保相关方及时知悉变化情况。

5.7.6 更新记录留存：

a) 留存透明度信息的“历史版本”，记录“更新时间、更新内容、更新原因、更新责任人”，留存期限不得少于该 AI 系统的生命周期（若系统停用，应额外留存 5 年）；
b) 用户可查询历史透明度信息（如“2024 年 3 月更新前的算法参数”），查询入口应显著标注（如“透明度历史版本”栏目）；
c) 应建立透明度信息的版本管理体系，每个更新版本均应生成唯一的版本号，明确版本生效时间、废止时间，不同版本的透明度信息应相互隔离、完整留存，支持按版本号、时间维度快速查询追溯。

5.8 算法偏见的透明度披露要求

5.8.1 偏见评估与披露：

a) 中高风险 AI 系统应每季度开展“算法偏见评估”，评估维度包括“性别、年龄、地域、职业”等敏感属性，计算不同群体的决策结果差异（如“25～30 岁群体信贷通过率 78%，55～60 岁群体通过率 62%，差异率 16%”）；
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，如“地域+性别”“年龄+职业”等交叉维度，避免遗漏交叉性偏见问题；针对跨境场景，应额外评估不同国家/地区、不同语种、不同文化背景用户之间的决策差异，确保跨区域决策公平性；
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（如训练数据中某群体样本占比过低）、改进计划（如补充该群体样本再训练）”，改进计划应明确时间节点（如“3 个月内完成样本补充”）。

5.8.2 偏见缓解的透明度：

a) 实施偏见缓解措施（如重新加权训练样本、调整模型损失函数）后，应公开“缓解措施细节、缓解前后的偏见差异率变化（如从 18%降至 9%）、验证数据”，接受监管机构与公众监督；
b) 高风险领域应标注关键节点的人工干预机制：如“医疗诊断结果应经执业医师复核”“跨境通信重大调度决策（影响用户不低于 10 万）应经区域运营总监审批”，明确干预触发条件及时限；
c) 中高风险 AI 系统应每半年向社会公开一次算法偏见评估与缓解报告，披露全量敏感属性的决策差异率、偏见问题整改情况、缓解措施落地效果，高风险场景应每季度向监管机构报送专项偏见评估报告。

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	指标要求（按风险等级）	计算方式
准确性	解释准确率	高风险不低于95%，中风险不低于90%，	对比解释结果与模型真实决策逻辑的一致性（如

评估维度	核心指标	指标要求（按风险等级）	计算方式
		低风险不低于85%	SHAP值与模型权重的偏差率）
可理解性	用户理解度评分	高风险不低于4.5分，中风险不低于4 分，低风险不低于3.5分	问卷调查（样本量不低于100，5分制：1-完全不理解，5-完全理解）
完整性	关键特征覆盖率	高风险不低于98%，中风险不低于95%，低风险不低于90%	（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	实时场景不超过1 s，非实时场景不超过5 min	多次测试取平均值（测试样本量不低于500）

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（如某版本的医疗诊断 AI 模型）、风险等级及目标受众（如医生/患者）；
   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，高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) 计算“解释准确率”“关键特征覆盖率”“解释生成耗时”（借助自动化工具如 SHAP 库、LIME 库）；
   3) 采用双盲测试方式开展验证，避免评估人员主观因素影响评估结果的客观性与公正性。
c) 用户调研：
   1) 针对目标受众设计问卷（含解释结果的易懂程度、有用性评分），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→可解释性合格；
   2) 存在不满足项→分析原因（如方法适配性不足），制定改进方案（如更换解释方法），15 个工作日内重新评估。

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，重点核查跨境数据披露、多区域权限控制等场景，评估周期与可解释性评估同步，高风险场景可叠加专项审计。

表 5 评估维度与指标

评估维度	核心指标	指标要求（按风险等级）	计算方式	备注
信息公开质量	信息覆盖度	高风险不低于98%，中风险不低于95%，低风险不低于90%	（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%	覆盖跨境数据流转、区域合规依据等专项信息
信息准确率	所有风险等级不低于98%	（公开信息与系统实际情况一致的项数/抽查信息项数）×100%，抽查比例不低于20%	跨境数据来源、脱敏方法等信息准确率不低于99%	—
信息可获取性	获取时效达标率	普通用户不低于98%，监管机构不低	（在规定时限内获取到目标信息的次数/	跨境用户获取信息平均耗时不超过3 min

评估维度	核心指标	指标要求（按风险等级）	计算方式	备注
		于100%	总尝试次数）×100%	
权限控制准确率	所有风险等级不低于100%	（权限匹配符合5.4要求的访问记录数/总访问记录数）×100%	跨区域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	高风险不低于98%，中风险不低于95%，低风险不低于90%	（在规定时限内处理完成并反馈的异议数/总异议数）×100%	跨境异议处理应标注区域责任主体
审计资料完整性	高风险不低于100%，中风险不低于95%	（审计所需资料（如迭代记录、追溯日志）完整提供的项数/审计要求资料项数）×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（1-2 个工作日）：
   1) 确定评估范围：涵盖数据、算法、决策过程全环节，明确跨境评估重点（如欧盟 GDPR 合规披露）；
   2) 组建评估小组：包含技术专家、合规专员、跨境业务负责人，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（高风险场景“审计资料完整性”权重不低于 30%）、测试工具（如权限测试系统、追溯查询模拟器）。
b) 信息核查阶段（5-7 个工作日）：
   1) 公开信息核查：对照第 5 章节要求，核查官网、公示平台等渠道的信息覆盖度与准确率，跨境场景应核查多语种披露情况；
   2) 后台数据验证：调取系统后台数据（如数据预处理日志、算法迭代台账），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（普通用户、监管机构）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) 跨境专项核查：针对跨境运营的 AI 系统，专项核查不同区域的信息披露合规性、多语种披露准确性、跨区域权限控制一致性、数据跨境流转披露完整性。
c) 响应能力验证阶段（3-5 个工作日）：
   1) 异议处理模拟：随机生成 50 条不同风险等级的异议案例（含 20%跨境案例），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，高风险场景应验证区块链存证数据（如决策日志）的不可篡改性。
d) 结果形成阶段（2-3 个工作日）：
   1) 指标计分：按评估方案权重计算各指标得分，综合得分不低于 90 分为“优秀”，80～89 分为“合格”，低于 80 分为“不合格”；
   2) 问题定位：对不合格指标（如跨境异议处理达标率 85%），分析根源（如区域协同机制不足）；
   3) 出具报告：明确评估结论、问题清单及整改建议，高风险场景报告应经跨境监管机构复核。

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：高风险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在30个工作日内完成优化整改，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（如优化后的异议处理流程、补充的透明度信息截图），由评估机构抽样复核，确保所有问题全部整改到位；对于高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）应在企业官网、行业协会平台进行公示，公示期不少于15个工作日，接受社会监督；评估结果可作为企业AI治理能力成熟度评估、行业评优、跨境业务准入、政府采购招投标的重要参考依据；鼓励行业内建立评估结果共享机制，在合规前提下实现跨企业、跨区域的评估数据互通，避免重复评估。

6.3.4 留存评估档案：所有评估资料（含测试数据、评估方案、问卷样本、评估报告、整改记录、复核材料）应留存至少5年，高风险场景留存期限不少于10年，跨境场景应按业务覆盖区域法规要求延长留存期限（如欧盟不低于7年）；评估档案应采用不可篡改方式存储（如区块链存证、异地备份），支持监管机构、第三方审计机构随时调阅核查。

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，优化企业内部AI算法管理规范、技术标准、操作流程；应跟踪国内外可解释性技术发展、监管政策更新，将评估结果作为技术升级、制度优化的核心输入，不断提升AI系统可解释性与透明度管理水平。

6.3.6 分级分类管控要求如下。

a) 高风险AI系统：
   1) 评估结果应作为系统上线、版本迭代、跨境部署的一票否决项，评估优秀方可开展跨区域规模化应用；
   2) 应基于评估结果每季度开展一次专项优化，持续完善可解释性方案与透明度披露机制；
   3) 评估结果应同步报送至对应行业监管机构、跨境业务属地监管部门，作为合规备案的核心材料。
b) 中风险AI系统：
   1) 评估结果应纳入企业内部AI系统运营考核体系，评估合格方可开展业务放量推广；
   2) 应基于评估结果每半年开展一次全面优化，重点提升用户端解释易懂性、异议处理响应效率；
   3) 评估结果应在企业内部公示，接受内部审计与合规监督。
c) 低风险AI系统：
   1) 评估结果应纳入AI系统全生命周期管理台账，评估合格方可完成版本正式发布；
   2) 应基于评估结果每年开展一次优化迭代，保障基础可解释性与透明度要求落地；
   3) 评估结果应留存归档，接受监管机构抽查。

6.3.7 监管合规对接要求：

a) 应建立常态化监管报送机制，高风险场景每季度、中低风险场景每半年向监管机构报送可解释性与透明度评估报告，重大合规问题、重大安全事件相关评估结果应及时报送；
b) 应配合监管机构的现场核查、非现场监管工作，根据监管要求实时提供评估相关的全量资料，开放对应的监管查询接口，确保监管机构可实时核验AI系统可解释性与透明度落地情况；
c) 针对监管机构在核查、检查中提出的整改要求，应在规定时限内完成整改并提交评估复核报告，整改情况纳入下一次正式评估的核心考核维度。

6.3.8 跨境场景评估结果互认与适配要求：

a) 针对跨境运营的AI系统，评估方案应覆盖业务落地所有国家/地区的合规要求，评估结果应适配当地监管规则，可根据属地要求出具多语种专项评估报告；
b) 应积极对接国际AI治理标准体系与跨境监管互认机制，在符合我国法律法规的前提下，推动评估结果在跨境合作场景、跨国监管场景中的互认应用，减少跨境重复评估与合规成本；
c) 针对不同国家/地区的差异化监管要求，应基于评估结果建立区域化适配方案，明确不同区域的可解释性与透明度调整内容，确保AI系统在各属地均满足当地合规要求，相关适配情况纳入专项评估维度。

6.3.9 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、跨境合规要求等，每年培训时长不少于 16 学时，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、跨境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。

6.3.10 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。

6.3.11 应急处置联动要求：

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、高风险缺陷纳入 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。