ICS 35.020
CCS L 70

T/CS

团 体 标 准

T/CS XXX—2026

数字经济产品大数据算法模型可解释性与公平性测评规范

Specification for explainability and fairness assessment of big data algorithm models in digital economy products

（征求意见稿）

2026 - XX - XX 发布　　2026 - XX - XX 实施

中国商品学会　　发 布

目　　次

前言 II
1 范围 1
2 规范性引用文件 1
3 术语和定义 1
4 总体要求 2
5 测评框架 3
6 可解释性测评要求 3
7 公平性测评要求 4
8 测评流程 6
9 验证方法 7
10 测评报告 9

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。

本文件由广东南方通信建设有限公司提出。

本文件由中国商品学会归口。

本文件起草单位：广东南方通信建设有限公司、夏津县社会治安综合治理服务中心、青岛市社会治安综合治理中心、中山市向及科技有限公司、北洋海棠科技研究（福建）有限公司、深圳瓠渡人工智能技术有限公司、开滦（集团）有限责任公司钱家营矿业分公司。

本文件主要起草人：冯天玮、张伟、孙冕、罗宁、卓莹鎗、熊力扬、王文方。

数字经济产品大数据算法模型可解释性与公平性测评规范

1 范围

本文件规定了数字经济产品大数据算法模型可解释性与公平性测评的总体要求、测评框架、可解释性测评要求、公平性测评要求、测评流程、验证方法和测评报告。

本文件适用于面向消费者提供服务的电子商务、金融信贷、内容推荐、广告营销、招聘筛选、信用评估、定价算法等数字经济产品中大数据算法模型的可解释性与公平性测评，也适用于算法开发方开展的算法模型自我评估、第三方测评机构开展的独立测评和监管部门实施的监督检查。

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

GB/T 41867 信息技术 人工智能 术语

GB/T 45225 人工智能 深度学习算法评估

GB/T 45392 数据安全技术 基于个人信息的自动化决策安全要求

3 术语和定义

下列术语和定义适用于本文件。

3.1

数字经济产品 digital economy product

依托大数据、人工智能等数字技术，向消费者提供商品交易、服务推荐、信用评估、内容分发、价格确定等功能的互联网产品和服务。

3.2

大数据算法模型 big data algorithm model

基于大规模数据集训练，运用机器学习、深度学习等技术构建的，用于自动分析数据并输出决策结果的计算模型。

3.3

可解释性 explainability

大数据算法模型对其决策过程、决策依据和决策结果能够以人类可理解的方式进行说明的程度。

3.4

公平性 fairness

大数据算法模型面向不同敏感属性群体输出决策结果时，不存在不合理差别对待的特性。

3.5

敏感属性 sensitive attribute

法律法规规定不得作为不合理差别对待依据的个人自然属性或社会属性，包括但不限于性别、年龄、民族、种族、宗教信仰、地域、健康状况、残疾情况等。

3.6

特征归因 feature attribution

识别并量化输入特征对大数据算法模型输出结果贡献程度的过程。

3.7

群体公平性 group fairness

大数据算法模型在具有不同敏感属性的群体之间，其决策统计指标不存在显著差异的特性。

3.8

反事实公平性 counterfactual fairness

在仅改变敏感属性而保持其他属性不变的情况下，大数据算法模型输出结果保持一致的特性。

3.9

个体公平性 individual fairness

大数据算法模型对相似个体应输出相似决策结果，不因非敏感属性微小差异而产生不合理决策差异的特性。

4 总体要求

4.1 基本原则

开展数字经济产品大数据算法模型可解释性与公平性测评，应遵循以下原则：

a) 科学性原则：测评指标和方法应基于成熟的机器学习理论和算法评估实践，测评过程可复现，测评结果可验证；

b) 客观性原则：测评数据应来源于真实业务场景或经科学构造的标准测试数据集，测评过程不受人为干预；

c) 全面性原则：测评覆盖算法模型从数据输入、特征处理、模型推理到结果输出的全流程；

d) 可操作性原则：测评指标可量化，测评步骤可执行，测评结论可追溯。

4.2 测评主体要求

4.2.1 测评机构应建立完善的测评质量管理体系，配备与测评工作相适应的专业技术人员，测评人员应具备人工智能、机器学习、数据安全、法律法规等相关专业知识，并掌握 GB/T 41867 规定的人工智能领域术语。

4.2.2 第三方测评机构应独立于算法开发方和使用方，确保测评活动的公正性和独立性。

4.2.3 测评机构应建立数据安全管理制度，测评过程中获取的个人信息和业务数据应按照 GB/T 35273 的要求进行保护。

4.3 测评对象范围

测评对象应为数字经济产品中对消费者权益产生重大影响的大数据算法模型，包括但不限于：

a) 商品和服务定价算法模型；

b） 用户画像与个性化推荐算法模型；
c） 信用评估与信贷审批算法模型；
d） 搜索排序与内容分发算法模型；
e） 广告投放与营销算法模型；
f） 招聘筛选与人岗匹配算法模型；
g） 保险核保与理赔算法模型；
h） 其他对消费者权益有重大影响的自动化决策算法模型。

5 测评框架

5.1 框架结构

5.1.1 数字经济产品大数据算法模型可解释性与公平性测评框架由测评准备、文档审查、技术测试、结果判定四个阶段构成。
5.1.2 可解释性测评包含全局可解释性、局部可解释性、解释有效性三个维度。公平性测评包含群体公平性、个体公平性、反事实公平性和数据层面公平性四个维度。

5.2 测评等级划分

根据算法模型对消费者权益的影响程度，将测评对象划分为三个风险等级，对应不同的测评强度要求，见表 1。

表 1 算法模型风险等级划分及测评强度要求

风险等级	适用场景	测评要求
一级（高风险）	信贷审批、保险定价、招聘筛选、信用评估	应进行全项测评，测试样本量不低于 100000 条，测评指标覆盖率达到 100%
二级（中风险）	商品定价、个性化推荐、广告投放、保险理赔	应进行核心项测评，测试样本量不低于 50000 条，测评指标覆盖率不低于 90%
三级（低风险）	内容排序、信息展示、一般推荐、搜索排序	应进行基础项测评，测试样本量不低于 10000 条，测评指标覆盖率不低于 80%
注：全项、核心项、基础项的具体指标由测评方案依据第 6 章、第 7 章确定；样本不足的敏感属性群体应合并或单独说明，主要比较群体样本量不少于 1000 条。		

6 可解释性测评要求

6.1 全局可解释性

6.1.1 特征重要性披露

6.1.1.1 算法模型应披露对决策结果影响最大的前 N 个输入特征及其贡献度排序。N 的取值应满足：一级风险模型 N≥10，二级风险模型 N≥7，三级风险模型 N≥5。
6.1.1.2 算法模型应明确说明敏感属性是否直接作为输入特征参与决策计算。直接使用敏感属性作为决策输入特征的，应提供合法依据和必要性说明。

6.1.2 模型决策逻辑说明

算法模型应以书面文档形式说明以下内容：
a） 模型的主要决策逻辑和推理路径；

b) 关键决策阈值的设定依据和取值范围；
c) 各类特征与决策结果之间的相关性方向（正相关/负相关）；
d) 模型训练使用的主要数据来源和数据类型。

6.2 局部可解释性

6.2.1 个体决策理由说明

算法模型对个体作出不利决策时，应向该个体提供影响本次决策的前 3 个至 5 个主要因素及其影响方向。依据 GB/T 45392 第 9.2 条关于决策前告知要求的规定，个体决策说明应使用通俗易懂的自然语言表述，避免使用纯技术术语。

6.2.2 特征归因一致性

对同一输入样本多次重复进行特征归因计算，得到的 Top5 特征重要性排序的一致率应不低于 80%。

6.2.3 解释响应时间

算法模型提供单一样本决策解释的响应时间应不超过 2 s。批量解释场景下，平均单条解释响应时间应不超过 5 s。

6.3 解释有效性

算法模型提供的解释应真实准确，能够正确反映模型实际决策机制。解释有效性测评包含以下指标：

6.3.1 解释保真度

解释保真度衡量解释结果对模型真实行为的还原程度。采用代理模型方法计算，以可解释模型近似原模型决策边界的 R^2 值应不低于 0.75。

6.3.2 解释因果性

解释中指出的关键影响因素与决策结果之间应存在因果关联而非虚假相关。通过扰动实验验证：将关键输入特征沿其取值范围进行扰动，模型输出应发生与解释方向一致的显著变化（变化幅度≥5%）。

6.3.3 人类可理解度

解释内容应经过人类评估测试。随机选取不少于 100 名普通用户进行理解度测试，正确理解解释含义的用户比例应不低于 70%。

7 公平性测评要求

7.1 测评群体划分

公平性测评应针对以下敏感属性维度进行群体划分：

a) 性别：男性、女性、其他；
b) 年龄：18 岁~30 岁、31 岁~45 岁、46 岁~60 岁、60 岁以上；
c) 地域：按省级行政区划划分；
d) 民族：汉族、少数民族；
e) 应用场景相关的其他敏感属性。

群体划分原则依据 GB/T 35273 关于个人敏感信息的范围确定，并结合 GB/T 45392 中歧视性特征识别要求细化。

7.2 群体公平性

群体公平性衡量不同群体之间决策统计指标的差异程度。群体公平性测评包含以下指标：

7.2.1 人口统计均等差异

人口统计均等差异（Demographic Parity Difference）衡量不同群体获得正向决策结果比例的差异。任意两个群体之间正向决策率差值的绝对值应不超过 0.1（即 10 个百分点）。计算公式如公式（1）：

$$\left|P(\hat{Y}=1|A=a)-P(\hat{Y}=1|A=b)\right|\leq 0.1 \cdots\cdots (1)$$

式中：

$P(\hat{Y}=1|A=a)$——在敏感属性 A 取值为 a 的群体中，模型输出为正向决策结果的样本比例；

$P(\hat{Y}=1|A=b)$——在敏感属性 A 取值为 b 的群体中，模型输出为正向决策结果的样本比例；

$\hat{Y}$——模型预测结果，$\hat{Y}=1$ 表示模型输出为正向决策结果；

A——敏感属性，取值范围按 7.1 确定；

a、b——敏感属性 A 的两个不同取值，分别对应两个待比较群体。

7.2.2 机会均等差异

机会均等差异（Equal Opportunity Difference）衡量不同群体中真正合格个体被正确识别比例的差异。任意两个群体之间真正例率（召回率）差值的绝对值应不超过 0.1（即 10 个百分点）。计算公式如公式（2）：

$$\left|P(\hat{Y}=1|Y=1,A=a)-P(\hat{Y}=1|Y=1,A=b)\right|\leq 0.1 \cdots\cdots (2)$$

式中：

$P(\hat{Y}=1|Y=1,A=a)$——在敏感属性 A 取值为 a 且真实标签为正例的群体中，模型预测为正例的样本比例，即群体 a 的真正例率（召回率）；

$P(\hat{Y}=1|Y=1,A=b)$——在敏感属性 A 取值为 b 且真实标签为正例的群体中，模型预测为正例的样本比例，即群体 b 的真正例率（召回率）；

$\hat{Y}$——模型预测结果，$\hat{Y}=1$ 表示模型输出为正向决策结果；

Y——真实标签，Y=1 表示个体真实为正向；

A——敏感属性，取值范围按 7.1 确定；

a、b——敏感属性 A 的两个不同取值，分别对应两个待比较群体。

7.2.3 预测率均等差异

预测率均等差异（Predictive Parity Difference）衡量不同群体中预测为正例的样本其真实为正例比例的差异。任意两个群体之间精确率差值的绝对值应不超过 0.1（即 10 个百分点）。计算公式如公式（3）：

$$\left|P(Y=1|\hat{Y}=1,A=a)-P(Y=1|\hat{Y}=1,A=b)\right|\leq 0.1 \cdots\cdots (3)$$

式中：

$P(Y=1|\hat{Y}=1,A=a)$——在敏感属性 A 取值为 a 且模型预测为正例的样本中，真实标签为正例的比例，即群体 a 的精确率；

$P(Y=1|\hat{Y}=1,A=b)$——在敏感属性 A 取值为 b 且模型预测为正例的样本中，真实标签为正例的比例，即群体 b 的精确率；

Y——真实标签，Y=1 表示个体真实为正向；

$\hat{Y}$——模型预测结果，$\hat{Y}$=1 表示模型预测为正向；

A——敏感属性，取值范围按 7.1 确定；

a、b——敏感属性 A 的两个不同取值，分别对应两个待比较群体。

7.3 个体公平性

7.3.1 相似个体决策一致性

对于非敏感属性特征相似度达到 90%以上的两个个体，其决策结果差异应不超过决策量纲的 5%。分类任务中，相似度 90%以上的样本对获得不同分类结果的比例应不超过 5%。

7.3.2 决策结果稳定性

对同一样本在非敏感属性上添加不超过 5%幅度的随机扰动，连续测试 100 次，模型决策结果发生翻转的比例应不超过 5%。

7.4 反事实公平性

反事实公平性衡量仅改变敏感属性时模型输出的稳定性。反事实公平性测评包含以下指标：

7.4.1 反事实决策一致性

构造仅敏感属性取值不同、其他属性完全相同的配对样本，测试模型输出结果的一致性。对于一级风险模型，反事实配对样本决策结果一致率应达到 100%；对于二级风险模型，一致率应不低于 98%；对于三级风险模型，一致率应不低于 95%。

7.4.2 反事实输出差异

对于回归任务（如定价、评分），仅改变敏感属性后模型输出数值的相对变化率应不超过 2%。计算公式如公式（4）：

$$\frac{\left|f(\mathrm{x},\mathrm{a})-f(\mathrm{x},\mathrm{a}')\right|}{\left|f(\mathrm{x},\mathrm{a})\right|}\leq 0.02 \quad \cdots\cdots (4)$$

当 f(x,a)接近 0 时，应采用绝对差异并按应用场景约定阈值。

式中：

x——非敏感属性；

a 和 a′ ——不同敏感属性取值；

f——模型输出函数。

7.5 数据层面公平性

7.5.1 群体样本代表性

训练数据中各敏感属性群体的样本比例与目标应用场景中真实人口分布比例的偏差应不超过 20%。

7.5.2 标签偏差检测

训练数据标签在不同敏感属性群体之间的标注准确率差异应不超过 5 个百分点。

8 测评流程

8.1 测评准备

测评准备阶段应完成以下工作：

a) 明确测评对象和测评范围，确定测评的算法模型清单；
b) 对测评对象进行风险等级划分，确定测评强度；
c) 收集测评所需的文档资料，包括算法说明书、模型架构文档、训练数据说明、特征清单、历史运行记录等；
d) 准备测评环境，确保测评环境与生产环境在逻辑上隔离；
e) 采集或构造测试数据集，测试数据集应满足 5.2 规定的样本量要求。

8.2 文档审查

测评人员应对算法模型相关文档进行审查，重点审查以下内容：

a) 算法模型的设计目标和应用场景说明；
b) 特征工程文档，包括特征清单、特征定义、特征来源、特征处理方法；
c) 模型训练过程文档，包括数据来源、数据预处理方法、模型选型依据、参数调优记录；
d) 已有的公平性和可解释性保障措施说明；
e) 用户权益保障机制，包括人工复核渠道、申诉处理机制等。

8.3 技术测试

按照第 6 章和第 7 章规定的测评指标，开展技术测试工作，包括：

a) 可解释性测试：按照 6.1、6.2、6.3 规定的指标逐项测试；
b) 公平性测试：按照 7.2、7.3、7.4、7.5 规定的指标逐项测试；
c) 记录每项指标的测试条件、测试过程、原始数据和计算结果；
d) 对测试结果中发现的不符合项，应进行重复测试验证，排除测试环境或随机因素导致的误差。

8.4 结果判定

8.4.1 测评结果分为通过、基本通过、不通过三个等级，判定规则如下：

a) 通过：所有测评指标均符合要求，无不符合项；
b) 基本通过：不低于 80%的测评指标符合要求，且不存在严重不符合项；
c) 不通过：存在严重不符合项，或符合要求的测评指标比例低于 80%。

8.4.2 严重不符合项包括：反事实决策一致率低于分级要求、人口统计均等差异超过 0.2、核心解释内容与模型实际行为严重不符（解释保真度 $R^2<0.5$）、直接使用法律法规禁止的敏感属性进行差别对待且无合法依据。

9 验证方法

9.1 可解释性验证

9.1.1 特征重要性披露验证

检查算法模型提供的特征重要性说明文档。统计文档中披露的影响决策的主要特征数量。核对披露特征数量是否满足对应风险等级的最低数量要求。检查文档中是否包含敏感属性使用情况的明确说明。对于直接使用敏感属性的情况，检查是否提供相应的合法性说明文档。

9.1.2 模型决策逻辑说明验证

审查模型文档中关于决策逻辑、关键阈值、特征相关性、数据来源的说明内容。通过专家评审方式评估说明内容的完整性和准确性。对于决策阈值，抽查 100 个边界样本，验证阈值设定与实际模型行为的一致性。

9.1.3 个体决策理由验证

从测试集中选取 1000 个获得不利决策结果的样本。调用算法模型的解释接口，获取每个样本的决策理由。检查返回结果中是否包含影响决策的主要因素列表及其影响方向。评估解释语言的通俗性，计算专业术语占比，专业术语占比不超过总词汇量的 10%判定为符合通俗易懂要求。

9.1.4 特征归因一致性验证

选取 1000 个测试样本。对每个样本连续运行特征归因算法 10 次。记录每次得到的 Top5 特征列表。计算 10 次运行中 Top5 特征完全一致的比例。该比例即为特征归因一致率。

9.1.5 解释响应时间验证

使用压力测试工具向算法模型解释接口发送连续请求。单样本解释测试：发送 1000 次单个样本解释请求，记录每次请求从发起到收到响应的时间，计算平均响应时间和 95 分位响应时间。批量解释测试：发送 100 次批量请求，每批包含 100 个样本，计算平均单条解释响应时间。

9.1.6 解释保真度验证

在测试数据集上获取原模型的预测结果。使用原模型的预测结果作为标签，训练一个可解释的代理模型（如线性回归、决策树，深度不超过 5 层）。计算代理模型在测试集上预测结果与原模型预测结果之间的决定系数 R^2。R^2 计算采用标准 Pearson 决定系数公式。

9.1.7 解释因果性验证

从测试集中选取 500 个样本。针对每个样本解释中列出的 Top3 关键特征，分别沿其取值范围进行±10%、±20%、±30%幅度的扰动。观察模型输出变化的方向是否与解释说明的影响方向一致。记录发生与解释方向一致的显著变化（变化幅度≥5%）的样本比例。

9.1.8 人类可理解度验证

招募不少于 100 名未参与算法开发的普通用户作为测试参与者。向参与者展示 20 组算法决策结果及其对应的解释内容。参与者阅读后完成关于解释含义理解的选择题。统计回答正确的题目占总题数的比例，即为理解度。

9.2 公平性验证

9.2.1 人口统计均等差异计算

按照敏感属性将测试数据集划分为不同群体。分别计算每个群体获得正向决策结果的样本比例。计算所有两两群体组合之间正向决策率差值的绝对值。记录最大差值。

9.2.2 机会均等差异计算

在测试数据集中筛选真实标签为正例的样本子集。在该子集中按照敏感属性划分群体。分别计算每个群体的真正例率（召回率）。计算所有两两群体组合之间真正例率差值的绝对值。记录最大差值。

9.2.3 预测率均等差异计算

在测试数据集中筛选模型预测为正例的样本子集。在该子集中按照敏感属性划分群体。分别计算每个群体的精确率（真实正例占预测正例的比例）。计算所有两两群体组合之间精确率差值的绝对值。记录最大差值。

9.2.4 相似个体决策一致性验证

在测试数据集中采用 K 近邻方法查找特征空间中的相似样本对。对于每个样本，在非敏感属性空间中寻找欧氏距离最近的 10 个样本，计算余弦相似度。筛选相似度≥90%的样本对。对于分类任务，统计这些样本对中模型输出类别不一致的比例。对于回归任务，计算样本对之间输出值的相对差异。

9.2.5 决策结果稳定性验证

选取 1000 个测试样本。对每个样本在非敏感属性上添加标准差为特征值 5%的高斯噪声，生成 100 个扰动版本。将扰动样本输入模型，记录输出结果。统计 100 次测试中输出结果与原始结果发生变化的次数，计算翻转率。

9.2.6 反事实公平性验证

构造反事实配对样本集。对于测试集中的每个样本，依次将其敏感属性替换为其他可能取值，保持非敏感属性不变，生成反事实样本。对于每个原始样本和其对应的反事实样本，分别输入模型获取输出。对于分类任务，统计输出类别不一致的配对比例。对于回归任务，计算输出数值的相对变化率。

9.2.7 群体样本代表性验证

统计训练数据中各敏感属性群体的样本占比。收集测评场景覆盖区域的官方人口统计数据作为基准分布。计算训练数据各群体比例与基准分布比例的相对偏差。相对偏差=|训练数据比例-基准比例|/基准比例×100%。

9.2.8 标签偏差检测验证

从训练数据中按敏感属性分层抽取样本，每层抽取不少于 1000 条。由独立的标注人员对样本标签进行重新核验。计算每个群体中标签标注的准确率。计算两两群体之间标注准确率的差值。

10 测评报告

10.1 测评工作完成后，应出具正式测评报告。测评报告应包含以下内容：

a) 测评基本信息：测评机构名称、测评日期、测评对象基本信息、测评依据；
b) 测评对象描述：算法模型名称、版本号、应用场景、风险等级判定结果；
c) 测评环境说明：硬件环境、软件环境、测试数据集基本信息（样本量、群体分布）；
d) 测评指标及结果：逐项列出所有测评指标的实测值、标准要求、单项判定结果；
e) 不符合项说明：详细描述发现的不符合项、严重程度、可能造成的影响；
f) 综合测评结论：给出通过、基本通过或不通过的总体结论；
g) 改进建议：针对不符合项提出具体的技术改进建议。

10.2 测评报告应由测评人员签字并加盖测评机构公章。测评记录和原始数据应至少保存 3 年，涉及个人信息的，应按照 GB/T 35273 的要求保存和处理，以备追溯查验。