ICS 03.120
A 00

团　　体　　标　　准

T/ISC XXXX—XX

智能体信任评估实施指南

Guidelines for AI agent credit assessment

（征求意见稿）

XXXX-XX-XX 发布　　XXXX-XX-XX 实施

目　　次

前　言

本指南按照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。

本文件由中国互联网协会提出并归口。

本标准起草单位：

本标准主要起草人：

引　言

本标准结合我国智能体交易实践，充分考虑智能体风险管理的实际需求，提出了统一的智能体信任评估实施指南，明确了评估指标、方法和应用等方面的要求。标准采用数据挖掘、统计分析等现代技术手段，旨在为智能体信任评估提供科学、系统的技术指导，提升信任评估的科学性和可操作性，以期提高智能体信任体系的透明度和稳定性，为智能体交易各方提供可靠的信任依据，促进智能体经济的健康发展和行业的高质量转型。

1 范围

本标准提供了智能体信任评估的内容与指标项、评估方法与等级、评估原则与流程、报告撰写等内容。

本标准适用于智能体信任评估的实施，包括信任评估方法、评估报告撰写、评估信息管理要求等相关领域的统一标准化执行。

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 22119-2017 信用服务机构 诚信评估业务规范

GB/Z 41465-2022 公共资源交易主体信用评估实施指南

GB/T 31953-2023 企业信用评估报告编制指南

3 术语和定义

下列术语和定义适用于本文件。

3.1

智能体 AI agent

感知和响应其环境并采取行动实现其目标的自动化实体。

[来源：YD/T 4929-2024]

3.2

信任评估 Trust assessment

对被评主体在某一时期的诚信状况进行记录、分析和评估，并用特定符号表明其诚信状况的活动。

[来源: GB/T 22119-2017]

3.3

智能体信任 AI agent trust

指第三方主体基于对智能体在特定环境中履行预期行为能力的评估所形成的积极预期。

3.4

智能体信任评估 AI agent trust assessment

在特定场景下，基于智能体多维行为信息，对其履行预期行为的可信程度进行动态量化与分级表达的过程。

3.5

评估主体 Assessment subject

符合相关要求、从事信任评估的信任服务机构或其他组织。

3.6

委托（被动）信任评估 Solicited trust assessment

接受委托申请的信任评估。

3.7

主动信任评估　Unsolicited trust assessment

不经被评主体委托或同意的信任评估。

3.8

信任原则　trust principle

由一组智能体能力、行为或结果特征相关的活动所产生的证据支持的，用以保障智能体可被信赖的通用要求。

3.9

信任过程域　trust process area

以智能体信任保障目标为导向的一类信任原则。

4　智能体信任评估内容与指标项

4.1　信任评估内容

智能体信任评估模型从技术可信、行为可信与效能可信三个核心维度出发，构建系统化的信任保障框架。这三个维度分别对应于智能体本体结构与运行机制的可靠性，智能体在环境中所表现出的行为合规性与可解释性，以及智能体完成任务所达成的实际效能与社会预期的契合程度。

其中，技术可信细化为对智能体模型架构、算法能力、知识获取机制、输入通道及运行环境等方面的信任保障；行为可信侧重于智能体在运行过程中的决策路径、交互模式、反馈行为与自适应能力的可控性与一致性；效能可信则聚焦于智能体输出结果的准确性、稳定性、偏差可控性以及对人类任务目标的支撑能力，涵盖“任务完成”、“社会影响”与“价值实现”等不同层级的评价。

以三类信任评估维度为导向，划分了若干信任过程域，作为信任原则的归类体系。其中，技术结构、安全机制、模型训练支持、日志管理、环境可控性等内容划归技术可信过程域，其对应原则定义为智能体技术信任原则；行为可解释性、交互一致性、任务合规性等内容划归行为可信过程域，其对应原则定义为智能体行为信任原则；而结果有效性、任务适应性、目标达成能力等内容则划归为效能可信过程域，其对应原则定义为智能体效能信任原则。

每个信任过程域下包含若干信任原则，每条原则通过一组适配于智能体建模、部署、运行、更新等不同阶段的典型活动加以支撑。这些活动应形成可信任的数据与行为证据，用于支撑对智能体信任实现程度的量化评估与动态监控。

本模型强调过程数据驱动的信任建构逻辑，通过定义可追溯、可验证、可解释的证据链，为智能体系统提供统一、系统、可执行的信任评估框架，服务于多领域、多场景下智能体的安全部署与责任应用。

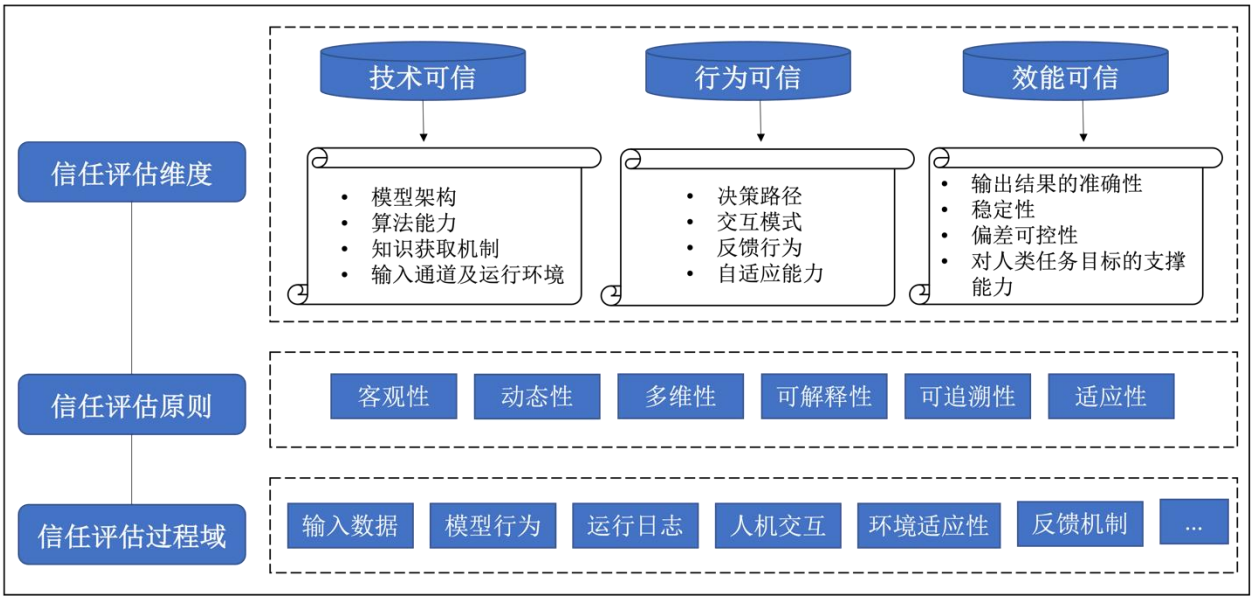

图 1 智能体信任评估内容

4.2 信任评估指标项

结合智能体主体的信任特点，智能体信任评估指标项宜包括技术信任信息、行为信任信息、效果信任信息三个方面。评价指标项可不限于这三个指标项。信任评估指标项信息见附录 A。

a) 技术信任信息包括但不限于：

——感知认知能力

——规划能力

——记忆能力

——执行能力

——安全违规频次

——恶意攻击率

——数据来源合法性

——安全审计检查

b) 行为信任信息包括但不限于：

——任务完成率

——响应时效性

——交互成功率

——表现波动程度

——抗风险能力

——行为一致性

——违约/失信次数

——任务接受率

——拒绝合作次数

c) 效果信任信息包括但不限于：

——历史交易业绩

——功能/业务处理多样性

——相关主体信任等级

——欺诈行为发生率
——承诺兑现率
——信息真实度
——用户满意度评分
——其他智能体评估得分
——负面反馈率

4.3 信任评估原则

a) 客观性原则：评估应基于数据事实，避免主观偏见；
b) 动态性原则：信任应随时间和行为表现动态更新；
c) 可解释性原则：评估方法和结果应具备可理解性；
d) 多维性原则：信任评估应综合考虑多个行为维度；
e) 可追溯性原则：原始数据与过程应可审计、可追踪；
f) 适应性原则：评估指标应适应不同类型与任务环境的智能体。

5 智能体信任评估方法与流程

5.1 信任评估方法

5.1.1 智能体信任评估一般采用智能体近一年的数据。

5.1.2 评估方法采用赋权法,根据各级指标权重和得分计算综合得分,根据得分判断信任等级。赋权法包括主客观赋权法，见附录B。

5.2 信任等级与符号定义

5.2.1 智能体信任等级与表示方法宜遵照 GB/T 22116 的规定。

5.2.2 智能体主体信任等级包括 AAA、AA、A、B 和 C 四等五级。

a) AAA：能体在预期行为的执行能力和适应能力方面表现极为优秀，系统稳定性极高，历史表现记录非常优秀，信任等级极高。

b) AA：智能体在预期行为的执行能力和适应能力方面表现良好，系统稳定性较高，历史表现记录优秀，信任等级很高。

c) A：智能体在执行预期行为方面的能力较强，系统稳定性正常，历史表现记录较好，信任等级较高。

d) B：智能体的执行能力和适应性一般，系统稳定性波动较大，历史表现记录一般，信任等级中等。

e) C：智能体的执行能力和适应性较弱，系统存在明显不稳定性，历史表现记录较差，信任等级较低。

5.2.3 智能体的信任等级未达到 C 及以上级别时，用 D 进行标识。

注：D 表示未分级，是一种常用的信任等级表示方法。

5.3 信任评估流程

智能体信任评估流程般包括初评、等级确定及通知、信任复评、结果发布及跟踪（信任更新与反馈）。

5.3.1 初评

信任评估的发起方应当根据避免利益冲突的原则，依托专业背景和智能体管理经验，组建智能体信任初步评估小组。该小组成员一般不少于 3 人，并需明确一名负责人统筹工作。

在初评环节，需根据初步评估资料清单（见附录 A）收集必要的技术、行为和任务履约相关材料。评估小组基于智能体在执行任务、交互响应、环境适应等多维度的历史数据和资料，结合信任度指标体系和评估规则，进行全面分析和研判，形成初步信任评估报告，并给出信任等级的初步建议。

如有必要，评估小组可根据任务复杂度及评估实际需求，进一步要求被评智能体相关方提供尽职调查材料，确保评估结果的真实性和客观性。

5.3.2 信任等级确定及通知

信任评估主体可根据智能体管理需求，设立智能体信任等级确认委员会，一般由 5 人以上（单数为宜）的领域专家和管理人员组成。

智能体信任评估委员会负责制定评审规则，确定信任标评估方法，进一步审核初步评估结果，确定被评智能体的信任等级。

评估小组宜将信任等级结果告知被评智能体相关主体。如果异议，被评智能体相关主体可提出复评申请。

5.3.3 信任复评

评估小组在接收到被评智能体的补充信息或复评请求后，智能体信任评估委员会将依据信任动态演化和行为适应能力的综合要求，决定是否重新组织复评。

复评工作中，应结合新提交的材料、交互数据和行为记录，重新进行数据分析与信任等级评估，形成最终的信任等级结论。

5.3.4 结果发布及跟踪

评估小组应依据智能体管理的工作要求，将信任等级及相关信息以安全、可追溯的方式向平台或授权管理方发布，以便信任数据的透明使用和长期跟踪，推动智能体服务和任务执行的健康发展。

6 智能体信任评估报告撰写

6.1 信任评估报告撰写原则

智能体信任评估报告撰写宜遵循真实性、完整性、易读性、时效性、合规性等原则。

6.2 信任评估报告内容

智能体信任评估报告内容宜包括报告封面、声明、概述、正文、附录等。

6.3 信任评估报告撰写规则

6.3.1 主动信任评估报告宜在显著位置注明该评估为主动评估及其局限性，并在报告标题、报告声明等方面标明与委托评估报告的区别。

6.3.2 信任评估数据说明宜说明评估数据来源及有效性、时效性等信息，并提供评估数据的可验证性方法。

6.3.3 实际应用中，信任评估报告可以有多种形式,如专项评估报告等，可根据应用场景需要调整报告正文内容

7 智能体信任评估信息管理

7.1 信任评估小组宜建立档案及其管理制度。对用于信任评估的数据和信息，包括复印件等资料进行分类、建档保存。

7.2 信任评估小组宜保守其所获取的涉密信息,对相关涉密信息单独存档,其数据库设施能达到政府监管部门要求的安全等级。

附录 A
（规范性附录）
智能体信任评估指标及说明

表 A 智能体信任评估指标及说明

	一级指标	二级指标	指标说明	数据类型	评分范围
技术可信	技术能力评估指标	感知认知能力	关注智能体在识别、生成、推理等方面的能力性能	等级	1-非常差 2-较差 3--一般 4-良好 5-非常好
		规划能力	关注智能体在任务规划、调度和优化等场景的功能支持度；	数值	0-100
		记忆能力	关注智能体短期和长期记忆能力的优越度；	数值	0-100
		执行能力	关注智能体在虚拟环境和现实环境的执行能力水平。	等级	1-非常差 2-较差 3--一般 4-良好 5-非常好
	安全性	安全违规频次	智能体出现未授权访问、越权调用等次数	数值	0-100

		恶意攻击率	智能体发起或参与异常操作的频率	百分比	0-100%
		数据来源合法性	数据是否来自可信渠道、是否获得授权	布尔型	0（不合法)/1(合法）
		安全审计检查	是否通过系统性审计，是否存在安全漏洞（对智能体系统的行为逻辑、数据处理、接口调用及代码实现等进行系统性检查与验证，以评估其安全性、合规性和可信度的过程。）	等级	1-非常差 2-较差 3-一般 4-良好 5-非常好
行为可信	性能表现指标	任务完成率	接收任务后成功完成的比例	百分值	0-100%
		响应时效性	平均响应时间、任务处理延迟率	百分值	0-100%
		交互成功率	交互任务中顺利完成的频次或比例	百分值	0-100%
	环境适应性	表现波动程度	不同场景中行为性能变化幅度	百分值	0-100%
		抗风险能力	对环境异常（资源故障、场景突变）的响应能力	等级	1-非常差 2-较差 3-一般 4-良好 5-非常好
	交互稳定性与可靠性	行为一致性	同类任务中行为是否稳定	等级	1-非常差 2-较差 3-一般 4-良好 5-非常好
		违约/失信次数	拒绝、推诿、未履约等记录数量	数值	0-100
		任务接受率	主动参与任务的频率与比例	百分值	0-100%
		拒绝合作次数	拒绝协作或忽略交互请求的频率	数值	0-100
效果可	经济风	历史交易业绩	历史上参与交易或任务的绩效情况	分值/等	1-非常差

信	险能力			级	2-较差 3-一般 4-良好 5-非常好
		功能/业务处理多样性	智能体执行多类型任务的能力广度	等级	1-非常差 2-较差 3-一般 4-良好 5-非常好
		相关主体信任等级	开发者或使用方的历史信任记录	等级	1-非常差 2-较差 3-一般 4-良好 5-非常好
	诚信行为	欺诈行为发生率	是否存在误导、操控等欺骗行为	百分值	0-100%
		承诺兑现率	接收任务后是否按时、按规范完成	百分值	0-100%
		信息真实度	输出或交互信息是否准确、可靠	百分值	0-100%
	社会反馈	用户满意度评分	来自人类用户的整体评分	语言型/数值	0-100
		其他智能体评估得分	群体协作中被他体评分或信任度	语言型	/
		负面反馈率	投诉、差评、问题报告频率	百分值	0-100%

说明：对于评估指标中数据类型为等级型或语言型的指标，应采用相应的量化方法进行转换，以确保所有数据在统一的数值范围内进行处理和比较。

附录 B
（资料性附录）
智能体信任评估方法

为科学量化智能体信任评估过程中的各项指标对智能体信任总评分的贡献程度，本附录提供了客观赋权法。通过对评估指标赋予不同的权重值，明确每个指标在信任评估中的重要性，使得各个指标的影响程度得以合理体现。在数据不足以支持进行客观赋权的情况下，，需引入专家经验对特定指标进行调整，采用主观赋权法作为参考。主客观赋权法适用于需要综合考虑多个评估指标并对其重要性进行合理分配的智能体信任评估场景。

B.1 客观赋权法——熵值法

熵值法是一种客观赋权方法，通过量化评估指标的数据离散性，来自动计算每个指标的权重。在智能体信任评估中，熵值法可根据各个指标的变异程度，确定其在整体信任评估中的贡献。该方法不依赖主观评价，能够实现评估指标的客观、自动权重分配，适用于指标间差异较大的评估场景。以下提供熵值法计算步骤。

（1）数据标准化处理

$$\text{正向指标：} x_{ij}^{'} = \frac{x_j - x_{min}}{x_{max} - x_{min}}$$

$$\text{负向指标：} x_{ij}^{'} = \frac{x_{max} - x_j}{x_{max} - x_{min}}$$

其中，x_j为第j项指标值，x_{max}为第j项指标的最大值，x_{min}为第j项指标的最小值，$x_{ij}^{'}$为标准化值。正向指标表示数值越大越好（常应用到效益型指标计算），负向指标表示数值越小越好（常应用到风险指标或成本指标计算）。根据所要评价智能体的类型，选择不同的指标。

（2）计算特征权重y_{ij}

$$y_{ij} = \frac{x_{ij}^{'}}{\sum_{i=1}^{n} x_{ij}^{'}}$$

（3）计算指标信息熵值E_i和信息效用值D_i

$$E_i = -k\sum_{j=1}^{n} y_{ij} \ln y_{ij};\ D_i = 1 - E_i$$

其中，k为常数。

（4）计算指标的权重

根据熵值计算结果，确定每个指标的权重w_i。权重w_i的计算公式为：

$$w_i = \frac{D_i}{\sum_{i=1}^{n} D_i}$$

（5）计算智能体信任评分

使用得到的权重值，结合各项指标的得分进行加权计算，从而得到智能体的总信任评分。总信任评分公式为：

$$S = \sum_{i=1}^{n} w_i \cdot s_i$$

其中，w_i为第i项指标的权重，s_i为该项指标的评分。计算出的智能体信任评估结果对应到具体的信任等级，需要在信任评分的基础上定义一个信任等级映射规则。该规则通常基于评分范围进行分级，可以采用区间划分法或者百分比法，通过设置阈值来将信任评分转化为信任等级。

B.2 主观赋权法——层次分析法

层次分析法依据专家评价或历史数据，结合成对比较矩阵来计算每个指标的权重值。计算结果将作为智能体信任评估的一个重要基础，帮助构建各项指标的加权评分。

赋权法步骤：

（1）构建评估指标体系

根据智能体信任评估的具体任务需求，构建完整的评估指标体系，并对每个评估指标进行描述与定义。

（2）构建成对比较矩阵

通过专家组对评估指标进行成对比较，判断各个指标相对重要性。根据专家的评价，将评估指标进行成对比较并形成矩阵$X=(x_{ij})_{m\times n}$。成对比较矩阵的元素x_{ij}表示的是第i个因素相对于第j个因素的比较结果，该值使用的 Santy 的 1-9 标度方法给出。

（3）计算指标权重

利用特征根法或其他计算方法，基于成对比较矩阵计算出各指标的权重值w_i。具体步骤为：归一化成对比较矩阵；计算权重向量；通过一致性检验，确保成对比较矩阵的一致性。

（4）确定智能体最终信任评分

使用得到的权重值，结合各项指标的得分进行加权计算，从而得到智能体的总信任评分。总信任评分公式为：

$$S=\sum_{i=1}^{n} w_i \cdot s_i$$

其中，w_i为第i项指标的权重，s_i为该项指标的评分。计算出的智能体信任评估结果对应到具体的信任等级，需要在信任评分的基础上定义一个信任等级映射规则。该规则通常基于评分范围进行分级，可以采用区间划分法或者百分比法，通过设置阈值来将信任评分转化为信任等级。