ICS 35.240
CCS L 70

中华人民共和国国家标准

GB/T 47507—2026

人工智能　可信赖　通则

Artificial intelligence—Trustworthiness—General rules

2026-04-30 发布　　2026-08-01 实施

国家市场监督管理总局
国家标准化管理委员会　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。

本文件由全国信息技术标准化技术委员会(SAC/TC 28)提出并归口。

本文件起草单位：中国电子技术标准化研究院、中国科学院软件研究所、中科南京软件技术研究院、中建科技集团有限公司、中电信数智科技有限公司、上海文鳐信息科技有限公司、上海商汤智能科技有限公司、浙江大华技术股份有限公司、杭州海康威视数字技术股份有限公司、北京瑞莱智慧科技有限公司、北京百度网讯科技有限公司、浪潮电子信息产业股份有限公司、北京航空航天大学、青岛港国际股份有限公司、北京浩瀚深度信息技术股份有限公司、中国民航信息网络股份有限公司、中国工商银行股份有限公司、电装智能科技(上海)有限公司、超越科技股份有限公司、浪潮软件科技有限公司、海信集团控股股份有限公司、西南科技大学、华北电力科学研究院有限责任公司、大连理工大学、马上消费金融股份有限公司、科大讯飞股份有限公司、中移九天人工智能科技(北京)有限公司、重庆国科础智信息技术有限公司、中科知道(北京)科技有限公司、西安交通大学、山东大学、浪潮通用软件有限公司、北京人形机器人创新中心有限公司、深圳地瓜机器人有限公司、京东方科技集团股份有限公司、中国移动通信集团有限公司、蚂蚁科技集团股份有限公司、上海市人工智能行业协会、天翼云科技有限公司。

本文件主要起草人：叶珩、范科峰、李斌斌、孟令中、高卉、董建、徐洋、曾涛、马骋昊、高峰、刘张宇、吴庚、邸贺亮、沈芷月、王珂琛、胡嵩智、杨光、刘祥龙、马艳军、王嘉凯、支阿龙、刘常昱、薛云志、赵剑、刘艾杉、徐小天、田丰、孔维生、邓志吉、金昕、林冠辰、仲凯韬、俞文心、夏知渊、秦日臻、庞韶敏、鄂磊、张磊、苏建光、任容玮、朱健、高雪松、童俊艳、谭拢、姜幸群、吕晓婷、隋伟、赵春昊、朱晓芳、陈乐然、陈利明、吴宇震、王维强、郝立强、王士宁、杨彤晖、王金超、樊则森、乔玉平、张志强、曹迎军、郭乙运、刘祥元、孟祥菊、任凯来、王帅、武斌、曹汐、吴畏甫、邓克俭、王宇晨、罗勇军、王俊杰、施展、苏瑾、王超、姜梦岑、温晓玲、艾笑天、刘永毅、孙香云、纪守领。

人工智能　可信赖　通则

1　范围

本文件确立了人工智能系统的可信赖核心要素，规定了相应的通用要求。

本文件适用于人工智能系统、含人工智能的计算机系统相关的设计、开发、应用、测试和监测。

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

可信赖　trustworthiness

＜人工智能＞满足利益相关方期望并可验证的能力。

注1：依赖于语境或行业，也依赖于具体的产品或服务、数据以及所用技术，应用不同的可信赖特征并对其进行验证，以确保利益相关方的期望能得到满足。

注2：可信赖的特征包括：可控性、安全性、鲁棒性、可靠性、准确性、韧性、可复现性、可泛化性、实时性、可备份性、数据合规性、隐私保护、保密性、透明性、可追溯性、可解释性、公平性、伦理符合、易用性、可问责性。

注3：可信赖作为一种属性用于描述服务、产品、技术、数据和信息，在所管辖语境中也用于组织。

[来源：GB/T 41867—2022，3.4.2]

3.2

人工智能系统　artificial intelligence system

针对人类定义的给定目标，产生诸如内容、预测、推荐或决策等输出的一类工程系统。

注1：该工程系统使用人工智能相关的多种技术和方法，开发表征数据、知识、过程等模型，用于执行任务。

注2：人工智能系统具有不同程度的自主性。

[来源：GB/T 41867—2022，3.1.8]

3.3

可控性　controllability

＜人工智能＞系统被人类或其他外部主体干预的性质。

[来源：GB/T 41867—2022，3.4.5]

3.4

安全性　safety

人工智能系统完成给定目标任务期间，能确保不会对用户、资源或环境造成无法接受风险的性质。

3.5

鲁棒性　robustness

人工智能系统在对抗条件下，能保持其性能水平的性质。

3.6

可靠性　reliability

人工智能系统在非对抗情况下，实施一致的期望行为并获得结果的性质。