认证技术规范

人工智能风险管理体系要求

CTS-BQC-AIRMS-2026

编制：王雷、刘凤栋、吕本斋、刘新

审批：杨小涛

北京质信认证有限公司

2026 年 5 月 5 日

认证技术规范-人工智能风险管理体系要求

本规范等同采用 ISO/IEC 42001:2023 的高阶架构（条款编号与标题一致），将 ISO/IEC 23894:2023 中关于 AI 风险管理的具体方法、风险源、影响评估等内容，以“要求”和“指南”的形式系统性地嵌入各相关条款。本规范可作为认证机构对组织的人工智能风险管理能力进行审核与认证的依据。

0. 引言

0.1 总则

人工智能（AI）系统的开发、提供和使用，为组织带来重大机遇，同时也引入区别于传统信息系统的特有风险，包括但不限于：算法偏见与不公平、缺乏透明度和可解释性、模型漂移与鲁棒性不足、数据投毒与对抗性攻击、自动化决策对个人和社会的深远影响等。

组织需要建立系统化的管理体系，以识别、评估、处置并持续监视这些 AI 特有风险，确保 AI 系统在全生命周期内得到负责任的管理。

0.2 与其他管理体系标准的关系

本规范采用 ISO 通用的**高阶架构（Annex SL）**（相同条款编号、标题、文本及通用术语），以便与 ISO/IEC 42001（人工智能管理体系）、ISO/IEC 27001（信息安全管理体系）、ISO 9001（质量管理体系）等协同实施。

0.3 本规范与 ISO/IEC 42001 及 ISO/IEC 23894 的关系

ISO/IEC 42001：规定了 AI 管理体系的“框架要求”，但未详细展开风险管理方法论。

ISO/IEC 23894：提供了 AI 风险管理的“技术指南”，但本身不可用于认证。

本标准：将 ISO/IEC 42001 的管理体系框架与 ISO/IEC 23894 的风险管理方法论整合为一个**可认证的体系要求标准**。

组织符合本规范，即证明其：

建立了符合国际共识的 AI 风险管理体系；

系统地识别并处置了 AI 系统特有的风险；

有能力对 AI 系统的开发、提供和使用履行责任与问责义务。

1. 范围

本规范规定了在组织环境内建立、实施、保持和持续改进**人工智能风险管理体系（AIRS， AI Risk Management System）** 的要求，并提供了实施指南。

本规范适用于开发、提供或使用 AI 系统或 AI 赋能产品的组织，无论其规模、类型与性质。

本规范旨在帮助组织：

系统化地识别、分析、评价和处置 AI 系统全生命周期内的特有风险；

将 AI 风险管理整合至组织的整体管理体系和业务流程；

满足适用法律要求、相关方义务及期望。

2. 规范性引用文件

下列文件的内容构成本文件的要求。凡注日期的引用文件，仅该版本适用；凡不注日期的引用文件，其最新版本（包括所有修改单）适用。

ISO/IEC 42001:2023，信息技术 — 人工智能 — 管理体系

ISO/IEC 23894:2023，信息技术 — 人工智能 — 风险管理指南

ISO/IEC 22989:2022，信息技术 — 人工智能 — 人工智能概念与术语

ISO 31000:2018，风险管理 — 指南

ISO Guide 73:2009，风险管理 — 术语

3. 术语和定义

ISO/IEC 22989:2022、ISO Guide 73:2009 界定的以及下列术语和定义适用于本文件。

3.1 人工智能风险

与 AI 系统的开发、提供或使用相关的不确定性对组织目标、个人或社会的影响。

注 1：AI 风险可能影响组织的战略、运营、财务、合规、声誉等目标。
注 2：AI 风险对个人和社会的影响可涉及公平性、隐私、安全、人权、环境等方面。

3.2 人工智能风险管理体系（AIRS）