ICS 35.240.01
CCS L 70

DB37

山　东　省　地　方　标　准

DB37/T 4845—2025

人工智能技术应用伦理风险的治理要求

Governance requirements of ethical risks in the application of artificial intelligence technology

2025－05－24 发布　　2025－06－24 实施

山东省市场监督管理局　　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。

本文件由山东省工业和信息化厅提出并组织实施。

本文件由山东省人工智能标准化技术委员会归口。

人工智能技术应用伦理风险的治理要求

1 范围

本文件规定了人工智能技术应用伦理风险的治理框架、治理原则、治理参与方和治理核心要求。

本文件适用于人工智能技术应用伦理风险的研究以及人工智能系统的规划、设计等。

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

人工智能 artificial intelligence; AI

〈学科〉人工智能系统相关机制和应用的研究和开发。

[来源：GB/T 41867—2022，3.1.2]

3.2

伦理 ethics

〈人工智能〉开展人工智能技术基础研究和应用实践时遵循的道德规范或者准则。

[来源：GB/T 41867—2022，3.4.8]

3.3

风险 risk

不确定性对目标的影响。

注1：影响是指偏离预期，偏离可以是正面的和/或负面的，可能带来机会和威胁。

注2：目标可有不同维度和类型，可应用在不同层级。

注3：通常风险可以用风险源、潜在事件及其后果和可能性来描述。

[来源：GB/T 24353—2022，3.1]

3.4

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、过程等的模型，用于执行任务。

注2：人工智能系统具备不同的自动化级别。

[来源：GB/T 41867—2022，3.1.8]

3.5

偏见 bias

〈人工智能〉对待特定对象、人员或群体时，相较于其他实体出现系统性差别的特性。

注1：对待指任何一种行动，包括感知、观察、表征、预测或决定。

注2：歧视是带有贬义色彩的偏见。

［来源：GB/T 41867—2022，3.4.10，有修改］

4 治理框架

人工智能技术应用伦理治理框架由内向外分别由治理原则、治理参与方、治理核心要求三方面组成，人工智能技术应用伦理治理框架见图1，具体如下。

a) 治理原则：提出了实施人工智能技术应用伦理治理的基本指导原则，为识别提出人工智能技术应用伦理治理的参与方和核心要求提供基础。

b) 治理参与方：识别了人工智能技术应用伦理治理过程中涉及的对象和主体，以及每类主体主要的治理目标，包括人工智能技术应用管理者、人工智能技术研发者、人工智能产品与服务供应商、人工智能产品与服务使用者等。

c) 治理核心要求：提出了人工智能技术应用伦理治理的核心要求，是伦理治理框架的关键组成部分。

图 1 人工智能技术应用伦理治理框架

5 治理原则

人工智能技术应用伦理治理原则包括以下内容。

a) 以人为本原则。人工智能技术应用应遵循人类共同价值观，尊重人权和人类根本利益诉求，遵守国家或地区伦理道德，避免对人类造成伤害和歧视。

b) 公平公正原则。人工智能技术应用应遵循公平公正原则，尤其在提供人工智能产品和服务时，应充分尊重和帮助弱势群体、特殊群体，避免产生不公平的结果或歧视，确保每个人都有平等的机会和待遇。

c) 隐私保护原则。人工智能技术应用应充分尊重个人信息知情、同意等权利，依照合法、正当、必要和诚信原则处理个人信息，保障个人隐私与数据安全，避免非法收集利用个人信息和侵害个人隐私权。

d) 可控可信原则。人工智能技术应用始终处于人类控制之下，人类拥有充分自主决策权，有权选择是否接受人工智能提供的服务，有权随时退出与人工智能的交互，有权随时中止人工智能系统的运行。

e) 责任担当原则。坚持人工智能技术应用的最终责任主体是人类，明确利益相关者的责任，在人工智能全生命周期各环节自省自律，遵循人工智能问责机制，不回避责任审查，不逃避应负责任。

f) 素养提升原则。坚持客观认识伦理问题，不低估不夸大伦理风险，主动开展或参与人工智能伦理问题讨论，深入推动人工智能伦理治理实践，提升应对能力。

6 治理参与方

6.1 管理者

人工智能技术应用管理者，包括从事人工智能技术应用相关的战略规划、政策法规和技术标准制定实施，资源配置以及监督审查等工作的人员与组织。

其治理目标，包括但不限于：

a) 遵守人工智能技术应用相关法律法规、政策和标准，积极参与并推动人工智能技术应用伦理治理，主动积极有序推动人工智能技术应用健康和可持续发展；

b) 充分尊重并保障人工智能技术应用相关主体的隐私、自由、尊严、等权利及其他合法权益；

c) 建立有效的风险预警机制，提升人工智能伦理风险管控和处置能力；

d) 充分重视人工智能各利益相关主体的权益与诉求，促进包容开放，推动各方共同参与人工智能伦理治理，实现多元主体共治，推动形成具有广泛共识的人工智能治理框架和标准规范。

6.2 研发者

人工智能技术研发者包括从事人工智能相关的科学研究、技术开发、产品研制等工作的人员与组织。其治理目标，包括但不限于：

a) 在人工智能研发相关活动中具有自律意识，主动将人工智能伦理道德融入技术研发各环节，不从事违背伦理道德的人工智能研发；

b) 在追求技术创新的同时，注重技术的透明性、可解释性、可理解性、可靠性、可控性，增强人工智能系统的韧性、自适应性和抗干扰能力，逐步实现可验证、可审核、可监督、可追溯、可预测、可信赖；

c) 在数据收集、存储、使用、加工、传输、提供、公开等环节，严格遵守数据相关法律、标准与规范，不侵犯个人隐私，提升数据的完整性、及时性、一致性、规范性和准确性等；

d) 避免偏见歧视，确保所研发的人工智能技术符合伦理道德要求，不侵犯个人隐私，不产生偏见和歧视，实现人工智能系统的普惠性、公平性和非歧视性。

6.3 供应商

人工智能产品与服务供应商包括从事人工智能产品与服务相关的生产、运营、销售等的工作人员与组织。

其治理目标，包括但不限于：

a) 提供符合伦理标准的人工智能产品与服务，强化质量监测和使用评估，不经营、销售或提供不符合质量标准的产品与服务；

b) 建立健全售后服务体系，对系统使用过程中出现的伦理问题进行及时处理；

c) 尊重市场规则，严格遵守市场准入、竞争、交易等活动的各种规章制度，积极维护市场秩序，尊重并维护其他主体的版权和知识产权；
d) 保障用户权益，履行告知义务，保障用户知情、同意等权利，保障用户数据安全。为用户选择使用或退出人工智能模式提供简便易懂的解决方案，不应设置障碍阻碍用户平等使用人工智能产品与服务。

6.4 使用者

人工智能产品与服务使用者包括从事人工智能产品与服务相关的采购、消费、操作等的个人与组织。其治理目标，包括但不限于：

a) 能积极学习人工智能相关知识，主动掌握人工智能产品与服务的使用各环节所需技能，确保人工智能产品与服务安全使用和高效利用；
b) 合理使用人工智能系统，避免滥用、误用与违规恶用；
c) 及时反馈人工智能系统使用过程中出现的伦理问题，协助人工智能系统供应商和人工智能技术研发者的技术和系统改进。

7 治理核心要求

7.1 以人为本

人工智能技术应用中以人为本的要求包括但不限于：

a) 人工智能技术的开发者或使用者应尊重人的根本利益诉求，以保障公共社会安全、尊重人的权益为前提，不准许误用、任何程度的滥用和恶用；
b) 人工智能技术的开发者或使用者应尊重社会共同价值观，促进人机和谐，坚持公共利益优先，人工智能的设计、开发和应用应为人带来福利，以实现人类社会利益最大化为目标。

7.2 隐私保护

人工智能技术应用中隐私保护的要求包括但不限于以下内容。

a) 保护数据隐私。人工智能技术的数据收集、存储、处理、使用应遵循合法、正当、必要的原则，明确告知用户收集、存储、处理、使用的目的、范围和使用方式，并经过用户的明确同意。应建立数据使用的审计机制，对数据的流向和使用情况进行监控和记录。
b) 数据安全保障。人工智能技术的开发者或使用者应采取合理的技术和管理措施，确保用户数据的安全存储和传输，防止数据被非法获取、篡改或滥用，确保数据的完整性和保密性。对已获得的个人隐私数据应进行脱敏处理。

7.3 知情同意

在人工智能技术的应用过程中，应充分尊重用户和相关利益方的知情权和选择权，保障用户对人 工智能系统相关功能与局限的知情、同意权利。包括但不限于以下内容。

a) 信息披露义务。人工智能技术的开发者或使用者应向用户充分披露技术的功能、应用范围、可能带来的风险以及数据收集和处理方式等信息，用户在使用前应被告知并理解这些信息。
b) 建立同意机制。提供人工智能产品或服务时，应为用户建立明确的同意机制以表达其意愿，且这种同意应是自愿的、明确的，并且用户有权随时撤回其同意。对于涉及无民事行为能力人等个人信息处理情形，应取得其监护人的同意。

7.4 安全可控

人工智能的设计者和开发者应采取必要的技术措施，确保人工智能系统的安全性和可靠性，包括 但不限于以下内容。

a) 算法安全。采取必要的技术措施，如加密、防火墙等，确保算法不被恶意攻击或篡改；人工智能的设计者和开发者开展人工智能算法安全测试，对人工智能算法的数据使用、算法运行开展日常监测工作，并定期评估算法的技术准确性和安全性；建立相应的算法终结机制，在算法决策遇到无法判断结果时立即终止；设立个人数据被遗忘机制和更改机制，划定算法自动关联的隐私边界。

b) 风险控制。人工智能的设计者和开发者进行算法安全评估、个人信息安全影响评估及风险评估，对算法、数据等内容从完成性、可靠性、稳定性等内容进行风险识别、风险分析，进行风险评估。

7.5 公平无歧视

人工智能技术的设计和应用应遵循公平性和无歧视原则，确保不同群体能够平等地受益于技术发展。具体要求包括但不限于以下内容。

a) 算法公平性，包括但不限于：
   1) 人工智能系统的算法不应存在任何形式的偏见和歧视，确保决策结果的公正性和客观性；
   2) 开发者应对数据集进行充分的预处理和平衡，以消除潜在的偏见和歧视因素；
   3) 数据标注阶段，应明确识别并标注出可能引入偏见的数据点，以便后续算法调整和优化，并对标注数据进行多样性和均衡性的检查，确保数据集中不同群体、背景、观点的代表性和平衡性；
   4) 在人工智能技术应用的整个生命周期内尽量减少和避免强化或固化带有歧视性或偏见的应用程序和结果，确保人工智能系统的公平；
   5) 对于带有歧视性和偏见的算法决定，应提供有效的补救办法。

b) 用户权益平等，包括但不限于人工智能技术不应对任何用户群体设置不公平的限制或条件，确保所有用户都能享有平等的权益和机会。

c) 语料选择与质量控制，包括但不限于所选语料应涵盖广泛的背景、文化和观点，以减少偏见和歧视。

7.6 透明可追溯

为了应对可能出现的风险和纠纷，人工智能技术应具备可解释性和可追溯性，包括但不限于以下内容。

a) 透明化决策。对于涉及用户利益的决策，人工智能系统应提供充分的解释和理由，确保用户能够理解并接受决策结果。同时，应明确告知用户语料的来源、使用目的和处理方式，用户应有权要求查阅决策过程的相关记录。

b) 可追溯。人工智能的设计者和开发者应确保人工智能关键决策的数据集、过程和结果的可追溯性，保证人工智能决策结果可被人类理解和追踪。

注：关键决策是指对研发结果可能产生重大影响的决策，如数据集的选择、算法的选取等。

7.7 责任明确

人工智能技术应用中责任明确的要求包括但不限于以下内容。

a) 明确责任主体，包括但不限于：
   1) 人工智能技术的管理者、研发者、供应商应明确各自在伦理风险治理中的责任和义务，确保责任的落实和追究；

2） 应将人工智能技术应用生命周期的任何阶段以及与之有关的补救措施的伦理和法律责任归属于自然人或现有法人实体，人工智能系统不应取代问责；

3） 应对提供的数据进行记录说明，留存数据使用日志，以便于数据溯源及主体责任的界定。

b） 建立监管机制。相关机构应加强对人工智能技术应用的监管力度，建立在人工智能设计或应用过程中存在恶意造成伦理风险行为的惩罚制度，制定并执行相关法规和标准，对违规行为进行处罚和纠正。

c） 审查机制。建立独立的审查机制，对人工智能系统的决策过程进行定期审查和评估，确保其符合伦理规范和社会期望。

7.8 技术中立性

人工智能技术应用理论风险治理中的技术中立性要求包括但不限于：

a） 人工智能技术的开发和应用不应偏离其原始目的，不被用于危害社会、侵犯人权等不正当用途；

b） 人工智能技术的提供者本身不应以非法的目的去生产该人工智能产品，且不能有意识地制造侵犯他人的人身权利或财产权利的产品，或是给人工智能输入类似算法导致侵权。

7.9 监测与改进

人工智能技术应用伦理风险的治理需要建立有效的监测与改进机制。包括但不限于以下内容。

a） 风险评估。应建立人工智能伦理风险评估机制，识别、分析、评价、处置人工智能伦理风险。

b） 反馈机制。建立用户反馈渠道，收集用户对人工智能技术应用的意见和建议，及时改进和优化系统。

c） 技术更新与迭代。应对人工智能系统进行更新和迭代， 以降低伦理风险并提高系统的性能和稳定性。

7.10 教育与培训

应加强人工智能技术的伦理教育和培训是降低伦理风险的重要途径。包括但不限于以下内容。

a） 开发者培训。对人工智能技术的开发者进行伦理教育和培训，增强其伦理意识和责任感，确保技术的合规应用。

b） 用户教育。通过宣传、讲座等形式，提高公众对人工智能技术的认识和理解，增强其风险意识和防范能力。

参 考 文 献

[1] GB/T 20988—2007 信息安全技术 信息系统灾难恢复规范

[2] GB/T 31722—2015 信息技术 安全技术 信息安全风险管理

[3] GB/T 38736—2020 人类生物样本保藏伦理要求

[4] GB/T 41867—2022 信息技术 人工智能 术语

[5] ISO/IEC 22989:2022 Information technology—Artificial intelligence—Artificial intelligence concepts and terminology

[6] ISO/IEC TR 24368:2022 Information technology—Artificial intelligence—Overview of ethical and societal concerns