ICS 35.240.01
L 70

DB37

山　　东　　省　　地　　方　　标　　准

DB 37/ XXXXX—XXXX

人工智能伦理风险的治理要求

Governance requirements of artificial intelligence ethical risks

（征求意见稿）

XXXX－XX－XX 发布　　XXXX－XX－XX 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。

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。

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。

引　言

以人工智能技术的研发和应用为代表的新一轮科技革命，在方便人们生活的同时也带来了复杂的伦理问题和治理挑战，人工智能伦理已经成为各国各地区人工智能政策的核心内容之一。2017年国务院印发《新一代人工智能发展规划》，提出制定促进人工智能发展的法律法规和伦理规范作为重要的保证措施。2020年国家标准化管理委员会、中央网信办、国家发展改革委、科技部及工业和信息化部联合印发《国家新一代人工智能标准体系建设指南》，将人工智能伦理列入国家人工智能标准体系，提出重点开展基于人工智能技术的医疗、应急等涉及伦理道德范畴的标准研制。2022年中共中央办公厅、国务院办公厅印发《关于加强科技伦理治理的意见》，首次对我国科技伦理治理工作作出系统部署，填补了我国科技伦理治理的制度空白。

为解决人工智能技术给个人或社会带来的伦理风险，制定本文件。本文件提出的人工智能伦理风险治理要求，可减少人工智能技术在歧视、隐私等方面带来的负面影响，有助于推动人工智能技术在各领域的广泛应用，也有利于人工智能安全、可靠、可控发展。

人工智能伦理风险的治理要求

1 范围

本文件针对人工智能伦理风险的不同来源，提出了对数据、研发者、生产商、用户的治理要求，以及对人工智能利益相关组织的基本治理要求。

本文件适用于研发和使用人工智能产品和服务的组织针对人工智能伦理风险的治理。

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（缩略语）
表现出与人类智能（如推理和学习）相关的各种功能的功能单元的能力。
[来源：GB/T 5271.28-2001，28.01.02]

3.2

伦理 ethics
处理人与社会、人与自然相互关系时应遵循的具体行为准则。
[来源：GB/T 38736-2020，3.11]

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，可以是积极的，也可以是消极的，或者两者皆有，它可以解决、创造或导致机会和威胁。
注2：目标可以是不同的方面和类别，可以应用于不同的层次。
注3：通常用风险来源、潜在事件、事件后果和事件发生可能性来表示风险。
[来源：ISO 31000:2018，3.1]

4 人工智能伦理风险类别

a）算法歧视

在人工智能算法设计过程中，由于开发人员对事物的认知存在某种偏见，或者对算法进行训练时，使用了带有偏见的数据集，使得算法产生带有歧视性的结果。

b）算法不可解释

影响人工智能算法决策的因素无法以人类可以理解的方式表达，导致人工智能在人类未知的情况下做出违背人类利益的决策。

c）算法决策困境

人工智能算法的决策结果存在两难的可能，导致算法决策困境。例如自动驾驶汽车在执行任务过程中面临多个选项，并且每一种选项都将对人类造成损害。

d）算法安全风险

人工智能算法存在泄露、被恶意修改等问题，从而产生违背人类道德的结果。

e）算法滥用

利用人工智能算法进行分析、决策、协调、组织等一系列活动中，其使用目的、使用方式、使用范围等出现偏差并引发不良影响的情况。

f）隐私泄露

在个人信息收集过程中，存在过度收集、信息外泄等问题，对人的正常生活和工作产生负面影响，造成财产损失等。

g）算法责任

人工智能技术或产品涉及多方主体，在出现人身伤害等违法或违反伦理道德的行为时，难以划分和认定各方主体的责任。

5 治理要求

5.1 概述

人工智能伦理风险产生的原因包括用于训练的数据存在缺陷、研发人员存在不合规行为、组织监督管理不到位及用户缺乏伦理风险意识等。基于人工智能伦理风险的不同来源，对数据、研发者、人工智能产品生产商及用户提出治理要求，并提出了对人工智能利益相关组织的基本治理要求。

5.2 对数据的治理要求

对数据的治理要求包括但不限于：

a） 应保证数据收集的充分与均衡，以保障数据的公平性；

b） 应对提供的数据进行记录说明，以便于数据溯源及主体责任的界定；

c） 应设置信息收集同意机制，对于涉及未成年人等个人信息处理情形，应取得其监护人的同意。

5.3 对研发者的治理要求

对研发者（包括研发人员及研发组织）的治理要求包括但不限于：

a） 应对关键决策进行记录说明并建立回溯机制（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）；

b） 应全面了解所用数据的分布特征，从数据中发现知识和规律；

c） 在技术条件允许的情况下，应开发建立本身具备可解释性的算法模型；

d） 应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；

e） 对已获得的个人隐私数据应进行脱敏处理；

f） 应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；

g） 应主动学习伦理安全相关知识，提升自身伦理风险意识；
h） 应开展人工智能算法安全测试，定期评估算法的技术准确性和安全性；
i） 应明确人工智能算法的目的意图及使用方式；
j） 应明确人工智能算法的应用领域，严格限定其适用边界；
k） 应监督和审查人工智能算法实现的过程和结果，并进行持续改进；
l） 应对人工智能算法的数据使用、算法运行开展日常监测工作；
m） 应建立参与式的算法决策框架，接受社会监督。

5.4 对人工智能产品生产商的治理要求

对人工智能产品生产商的治理要求包括但不限于：
a） 应有义务满足用户对算法结果解释的要求，针对不同背景知识的用户，应提供个性化定制和常识结合的解释；
b） 应对人工智能产品或服务的使用开展日监测工作。

5.5 对使用人工智能产品或服务的用户的治理要求

对使用人工智能产品或服务的用户的治理要求包括但不限于：
a） 不应对人工智能算法盲目依赖，坚持在算法应用中的主体性地位；
b） 应以良好的目的使用人工智能，禁止违规恶用；
c） 应主动反馈人工智能产品或服务的伦理风险相关信息；
d） 应主动学习伦理安全相关知识，提升自身伦理风险意识。

5.6 对人工智能利益相关组织的基本治理要求

对人工智能利益相关组织的基本治理要求包括但不限于：
a） 应组建人工智能伦理委员会或人工智能伦理治理团队，明确责任人、角色和职责；
b） 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；
c） 应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；
d） 应定期开展伦理安全相关教育培训，提升相关人员的伦理意识；
e） 应提升各利益相关者的参与程度，开展多元主体共治；
f） 各组织之间应开展落实算法责任的相互合作与监督。

参考文献

[1] GB/T 5271.28-2001 信息技术 词汇 第28部分:人工智能 基本概念与专家系统
[2] GB/T 38736-2020 人类生物样本保藏伦理要求
[3] ISO/IEC TR 24368 Information technology — Artificial intelligence — Overview of ethical and societal concerns
[4] ISO/IEC 22989:2022 Information technology — Artificial intelligence — Artificial intelligence concepts and terminology
[5] ISO 31000:2018 Risk management — Guidelines
[6] 人工智能伦理风险分析报告（2019）
[7] 网络安全标准实践指南 — 人工智能伦理安全风险防范指引
[8] 新一代人工智能伦理规范（2021）