人工智能伦理风险与治理研究

中国电子信息产业发展研究院
政策法规研究所（工业和信息化法律服务中心）
二〇二五年四月

1 问题聚焦

随着人工智能技术向经济社会各行业各领域加速渗透，其引发的伦理争议已成为全球关注焦点，美欧等主要经济体均根据自身情况，开展了一系列伦理治理实践。目前，我国人工智能伦理治理的基本理念、规则体系、治理模式和国际主张初步形成。在新形势新要求下，亟待优化完善具有中国特色的人工智能伦理风险治理路径。本研究重点回答：人工智能伦理风险是什么、治理现状如何、未来如何治理等问题。

拟回答问题

是什么?

- 人工智能伦理风险的内涵特征是什么?
- 人工智能带来了哪些伦理风险？这些风险如何进行类型划分?

治理现状如何?

- 国际组织、欧盟、美国等开展了哪些伦理治理实践，探索了哪些模式?
- 我国已探索开展了什么样的伦理治理实践?

未来如何治理?

- 面临哪些新形势新要求?
- 应如何完善现有治理路径?

治理对象分析

- 内涵特征
- 分类体系
- 生成机制
- 典型场景/案例

治理实践探索

国外

- 国际组织
- 欧盟
- 美国

我国

- 治理理念
- 治理规则
- 治理模式
- 国际主张

面临形势要求

- 中央要求亟待落实
- 实践问题亟待解决
- 企业诉求亟待回应

优化路径建议

- 持续完善法律法规体系
- 强化多元协同治理机制
- 深化国际治理合作
- 统筹伦理治理与产业发展

3 人工智能五大伦理风险

人工智能伦理风险是指：在人工智能技术的开发、部署和应用过程中，因技术特性、使用方式或管理缺陷等而引发的侵犯个人权益、冲击社会秩序、违背人类价值观的伦理风险，具体可从五个维度来分析。

	侵权性风险	歧视性风险	社会性风险	责任性风险	失控性风险
具体含义	• 对个人权益的直接侵犯	• 对特定群体产生不公正的价值判断或决策结果	• 误用、滥用技术或技术局限，而对公共社会秩序造成负面影响	• 造成负面影响后的责任界定难题	• 人工智能行为及影响超出人类预设、理解和可控范围
主要类型	• 侵犯个人信息权 • 侵害个人生命健康权 • 侵犯知识产权	• 对特定群体产生歧视性决策结果 • 生成内容出现歧视性表达	• 生成不良信息助长错误价值观 • 技术漏洞导致生成虚假、错误信息 • 人为滥用技术导致财产损失、精神损害	• “主体”身份认同危机 • 责任界定困难 • 利益与责任分配不公	• 人类主体能力退化和自主意识弱化 • 人类丧失对人工智能的控制能力
生成机制 **技术+规制**	• 海量数据需求与过度收集 • 算法“黑箱”与透明度缺失 • 技术漏洞与安全缺陷 • “合理适用”“知情同意”机制难执行 • 缺乏有效的版权审查和授权机制	• 训练数据包含隐性偏见内容 • 模型内部运作机制难以理解和解释 • 内容语料标准模糊 • 生成内容的审核、监测和控制机制不够完善	• 训练数据含有不良内容 • 模型存在技术漏洞被外部输入干扰信息 • 平台内容审核与处理机制不完善 • 缺乏对失业等社会性风险的有效应对方案	• 技术和模型的决策过程难以被人类理解和预测 • 人工智能应用过程中各方责任关系复杂 • 数据资本垄断化、算法权力黑箱化	• 人类产生“认知卸载”“算法迷信” • 自主学习机制等导致人类无法预测其行为 • 伦理审查形式化，如缺乏对“主体性”危机的关注

发布文件—提出治理原则建议

- **总体原则**

联合国《为人类治理人工智能》（2024.09）、联合国教科文组织《人工智能伦理问题建议书》（2021.11）、二十国集团《人工智能原则》（2019.06）等。

图：《人工智能伦理问题建议书》4项核心价值观

- **具体行业**

通信领域：国际电信联盟（ITU）《人工智能治理日——从原则到落实》（2024.07）专题报告。

金融领域：国际清算银行（BIS）《中央银行使用人工智能》（2024.01）《生成式人工智能与中央银行网络安全》（2024.05）。

卫生领域：世界卫生组织（WHO）《多模态大模型人工智能伦理和管理问题指导文件》（2024.01）。

搭建平台—推动全球伦理共识

- **人工智能系列峰会**

2023年英国峰会，28个国家和欧盟签署《布莱切利宣言》；2024年首尔峰会，10个国家和欧盟签署《首尔声明》；2025年巴黎行动峰会，61个国家签署《巴黎人工智能宣言》，促使越来越多的国家和地区就人工智能全球治理问题达成共识。

- **人工智能造福人类全球峰会**

2024年5月，国际电信联盟（ITU）与40个联合国伙伴机构合作达成了推动人工智能负责任应用、制定相关标准、讨论全球人工智能治理框架等共识。

- **全球工业和制造业人工智能联盟**

2023年7月，联合国工业发展组织（UNIDO）在第六届世界人工智能大会上宣布成立，聚焦于制定并推广工业和制造业领域的人工智能伦理准则，推动形成更广泛的国际共识。

图：全球工业和制造业人工智能联盟成立仪式

制定标准—规范技术发展

- 世界数字技术院（WDTA）发布《生成式人工智能应用安全测试标准》和《大语言模型安全测试方法》（2024.04），是国际组织首次就大模型安全领域发布国际标准。

《生成式人工智能应用安全测试标准》

定义人工智能应用程序架构各层的测试和验证范围

《大语言模型安全测试方法》

提出大语言模型的安全风险分类、攻击的分类分级方法以及测试方法

- 国际电信联盟（ITU）已发布或正在制定超过200项与人工智能相关的规则标准。其中，ITU在2024年世界电信标准化全会上形成的第COM4/AI号决议，是ITU首份人工智能新决议，致力于将人工智能应用于电信/ICT。

5 美欧人工智能伦理治理现状

美国

柔性和自愿性：侧重产业促进和经验探索，而非监管

协调性和咨询性：以软性指导为主，尚无监管性机构

分散性：联邦无统一立法，各州进行自发零散的治理探索

企业界影响较大：具有影响力的企业往往积极进行游说

对内鼓励技术创新：避免不必要地阻碍人工智能创新

对外强调超前发展：将人工智能视为大国竞争的重要资产

强有力的企业规模：技术领先的人工智能公司积极争取权益

部门探索和经验驱动：新业态监管倾向分散探索和事后规范

治理方式

治理机构

治理体制

影响力量

深层原因

内部目标

外部目标

行业力量格局

治理路径依赖

欧盟

刚性和严格性：治理文件规制明确，并设有严格的罚则

成熟性和系统性：整体层面和成员国层面均建立监管机构

统一综合：在整体层面统一立法以确保法律稳定性

政府主导：主要依赖行政手段对人工智能进行监管

对内安全伦理优先：高度重视伦理和安全问题

对外重视规则影响力：试图构建安全可靠的国际治理范本

相对滞后的技术水平：企业规模普遍较小，技术发展不突出

统一治理和立法先行：新业态监管倾向风险预防和统一立法