

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

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？



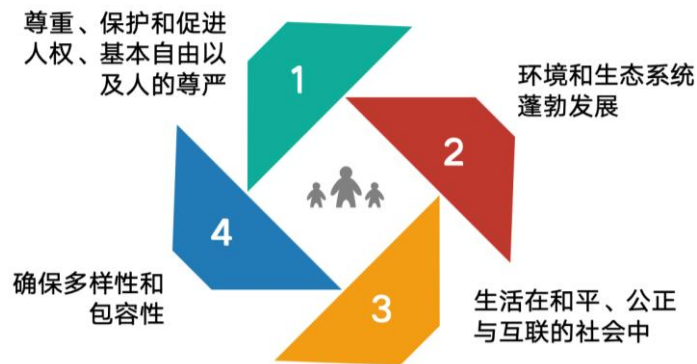
人工智能伦理风险是指：在人工智能技术的开发、部署和应用过程中，因**技术特性、使用方式或管理缺陷等**而引发的侵犯**个人权益**、冲击**社会秩序**、违背**人类价值观**的伦理风险，具体可从五个维度来分析。

	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。



美国

灵活性和自愿性：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

治理
理念

以人为本

“发展人工智能应坚持‘以人为本’理念，以增进人类共同福祉为目标，以保障社会安全、尊重人类权益为前提，确保人工智能始终朝着有利于人类文明进步的方向发展”
——习近平发布《全球人工智能治理倡议》

健康发展

“建立健全保障人工智能健康发展的法律法规、制度体系、伦理道德”
——习近平主持中共中央政治局第九次集体学习的讲话

开放协同

“鼓励通过国际合作和实际行动帮助各国特别是发展中国家加强人工智能能力建设，增强发展中国家在人工智能全球治理中的代表性和发言权”
——中国在联合国大会提出的加强人工智能能力建设国际合作决议

治理
规则

法律：顶层治理原则和框架

- 基础法律框架：《网络安全法》、《数据安全法》、《个人信息保护法》等法律分别从网络、数据、个人信息层面对人工智能进行伦理安全风险监管。
- 综合性立法提上日程：2025年3月，《全国人大常委会工作报告》提出围绕人工智能等新兴领域加强立法研究。2024年5月，《国务院年度立法工作计划》提到预备提请全国人大常委会审议人工智能法草案。

审议科学技术普及法修订草案。制定全民阅读促进条例、红色资源预备提请全国人大常委会审议人工智能法草案、商标法修订草案、制定历史街区与古老建筑保护条例、传统村落保护条例，预备修

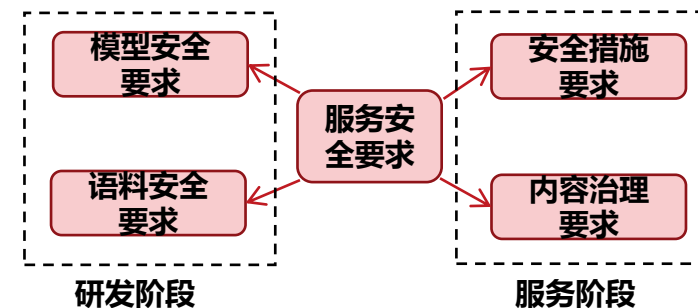
图：国务院2024年度立法工作计划（国务院官网）

部门规章：专门性治理要求

- 《生成式人工智能服务暂行办法》：世界首部应对生成式人工智能安全风险的专门立法。
- 《人工智能生成合成内容标识办法》：规范人工智能生成合成内容标识，重点明确“显式”和“隐式”标识要求，发挥监督溯源作用。
显式标识要求 + 隐式标识要求
- 《科技伦理审查办法（试行）》：规范科学研究、技术开发等科技活动的科技伦理审查工作，强化科技伦理风险防控，促进负责任创新。
- 《互联网信息服务算法推荐管理规定》：具有舆论属性或者社会动员能力的算法推荐服务提供者应当按照国家有关规定履行备案手续。

行业规范文件：重要合规参考

- 《生成式人工智能服务安全基本要求》：明确了生成式人工智能服务在语料安全、模型安全、安全措施等方面的基本要求。



图：《生成式人工智能服务安全基本要求》基本架构

- 《网络安全标准实践指南—人工智能伦理安全风险防范指引》：提出人工智能研究开发、部署应用等活动的伦理措施。
- 《新一代人工智能伦理规范》：提出管理、研发、供应等活动的18项具体伦理要求。



初步形成法律、部门规章、7 规范文件的阶梯式人工智能治理规则体系

治理
模式

01 政策引导+监管协调

制定和完善法律法规，明确人工智能伦理边界。及时发现并处理引发伦理风险的行为。

02 技术标准+行业规范

发挥专业性和影响力，制定人工智能技术标准和伦理规范，引导行业自律。

行业



企业



政府



用户



04 公众监督+共治参与

合理使用人工智能，并利用各种渠道对人工智能产品和服务进行监督。

03 落实要求+内部合规

建立健全企业合规制度，确保人工智能项目的各个环节符合法律法规规定的伦理要求。

国际
主张

输出中国治理方案

探索并推动中国治理方案走向世界。例如，发布《全球人工智能治理倡议》等文件，积极参与国际标准的制定。



搭建全球治理机制

积极搭建各国共建共治平台。例如，推动建设“一带一路”数字治理研究院，参与并推动以联合国为核心的多边合作机制。



参与多边治理活动

推动全球AI技术和治理合作。例如，在巴黎峰会主办专题边会，探讨如何促进各国共享人工智能发展成果。



中央要求亟待落实

- **健全完善人工智能监管机制。**党的二十届三中全会提出，“完善生成式人工智能发展和管理机制”“加强网络安全体系建设，建立人工智能安全监管制度”。
- **加强人工智能治理国际合作。**习近平总书记强调：“同各国加强交流和对话，共同促进全球人工智能健康有序安全发展”。



实践问题亟待解决

促进技术发展与保障伦理安全之间的关系日益复杂，平衡各维度矛盾成为**实践难题**：

- **技术普及 V.S. 社会公平。**如何更好处理技术快速发展中出现的资源分配不公、部分群体被边缘化等问题？
- **自主决策 V.S. 人类控制。**如何更好兼顾实现在提升人工智能技术先进性的同时，确保技术受人类控制？
- **商业利益 V.S. 公众利益。**如何更好地避免商业短期利益凌驾于人类利益之上？
- **全球治理 V.S. 区域规则。**如何妥善弥合全球人工智能伦理治理的鸿沟？



企业诉求亟待回应

通过调研华为、腾讯、360、京东、快手等国内十余家人工智能代表性企业，总结出企业对人工智能伦理治理优化的三方面**共性诉求**：

- **基于场景化开展分类监管。**应更有效区分不同的应用场景和风险等级。
- **提升规则、流程的明确性。**应更明确“舆论属性和社会动员能力”等概念界定，进一步明确大模型备案和算法备案之间的关系。
- **强化规则间的协同性。**应更好处理个人信息保护、算法公平、数据安全等相关治理工具和规则之间的关系。

通过阶段性推进、场景化适配和动态化调整的策略方法，持续优化我国人工智能伦理风险治理。

持续完善法律法规体系

- 完善和更新现有法律，针对新出现的问题制定专项法规，填补立法空白。
- 适时起草制定人工智能综合性法律，构建基础性、红线性制度规范以及科学性、系统性兼备的法律框架。



深化国际治理合作

- 加强国际合作和实践探索，积极为全球AI治理贡献“中国方案”。
- 灵活调整国际合作策略，减少单边主义对全球科技合作的干预，推动建立更加公正合理的全球科技治理新秩序。
- 引领全球人工智能发展方向，积极推动全球人工智能治理合作伙伴关系建设。

强化多元协同治理机制

- 政府加强执法监管，强化AI滥用风险监控。
- 行业组织加大自律力度，探索行业治理经验并推广最佳实践。
- 企业强化合规管理，将人工智能伦理要求嵌入企业管理和产品生命周期全流程。
- 提高公众对人工智能伦理问题的认知和关注，增强监督意识和能力。



统筹伦理治理与产业发展

- 建立技术与伦理协同创新体系。构建全链条伦理嵌入研发机制，将可信AI技术、伦理增强型芯片等产品纳入重点研发支持计划。
- 部署伦理治理类创新产品应用推广。推动企业部署治理合规自检系统，推广“治理即服务”。
- 探索新型激励机制。探索将企业伦理治理水平与融资、政府采购等挂钩的机制。

服务政府工作

支撑规则制定

法律法规层面：全国人大等人工智能立法支撑研究
规章制度层面：支撑工信部人工智能相关规章咨询意见支撑。如《标识》《伦理》等
行业标准层面：探索研究人工智能领域企业合规标准建设

支撑行业治理

工信部政法司：新业态监管、开源合规体系、企业合规标准等
工信部信管局：个人信息保护调研、加强人工智能监管课题研究等
工信部科技司：伦理相关工作支撑、国际治理模式比较联合研究等
国家发改委高技术司：大模型生态建设、人工智能相关工作方案等
国家数据局：人工智能与数据保护

提供市场咨询

助力企业合规

产品合规方案：为重点企业的人工智能产品提供合规方案
企业合规流程：为企业人工智能治理提供全流程合规方案
园区企业合规：为软件园区等提供园区企业合规路径咨询
其他专项合规咨询：针对特定问题提供合规咨询

服务地方政府

起草法规规章：为地方制定人工智能领域法规提供咨询服务
编制政策规划：为地方统筹人工智能产业发展与风险防范提供研究咨询
提供专项咨询：承接人工智能发展与安全相关专项课题委托

动态跟踪

事件解读

行业调研

专项研究

年度报告

底层研究能力

思想， 还是思想， 才使我们与众不同

研究， 还是研究， 才使我们见微知著

赛迪研究院政策法规研究所（工业和信息化法律服务中心）

通讯地址：北京市海淀区万寿路27号院1号楼25层

邮政编码：100036

联系人：彭健

联系电话：010-68209619 13810637353

传真：010-68209616

网址：<http://www.ccidthinktank.com>

电子邮件：pengjian@ccidthinktank.com



工信部机关刊《新型工业化》



联系方式:

(010)88558650 (010)88558656

- ◆ 2024年9月创刊，突出**理论性、权威性、指导性、实践性**。
- ◆ 宣传阐释习近平总书记关于新型工业化的重要论述的**重要阵地**，开展学理研究的**重要载体**，展示实践成果的**重要窗口**，跟踪交流新一轮科技革命和产业变革动态趋势的**重要平台**。
- ◆ 面向广大党员干部、理论工作者、企业家和高校、科研院所师生等读者群体，设置“重要论述”“特约文章”“探索实践”“专家论坛”等栏目。



期刊官微



扫码订刊