

人工智能风险治理报告

——构建面向产业的人工智能安全治理实践方案

(2024 年)

中国信息通信研究院人工智能研究所

2024年12月

版权声明

本报告版权属于中国信息通信研究院，并受法律保护。转载、摘编或利用其它方式使用本报告文字或者观点的，应注明“来源：中国信息通信研究院”。违反上述声明者，本院将追究其相关法律责任。

前 言

人工智能技术以其前所未有的发展态势引领新一轮科技革命和产业变革，深度驱动经济社会发展。技术跃迁释放价值红利，与此同时带来风险挑战。2023 年 10 月，习近平主席提出《全球人工智能治理倡议》，其中着重指出，做好风险防范，不断提升人工智能技术的安全性、可靠性、可控性、公平性。人工智能风险治理已经成为把握未来人工智能发展的核心要素，也是全人类需要共同应对的时代课题。

当前，全球各界积极探索人工智能风险治理之道。本报告广泛汲取国际行动、各经济体努力、产业实践等各方经验，勾勒新时代下人工智能风险认知的演进路径，提炼出人工智能风险认知正向全球性融合、前瞻性考量、系统性分析、交叉性联动不断深化。通过深入分析全球人工智能风险治理多元共治的实践经验，我们深刻认识到，将人工智能风险治理的方案落实到技术革新、产品开发和应用部署中去，是回应社会关切、解决突出矛盾、防范安全风险必然选择，也是关系到人工智能长远发展的重要议题。

本报告结合国际经验，立足我国产业实践，提出“系统治理-风险识别-风险评估-风险应对”的人工智能风险治理实践框架，实现穿透式风险管控与全链条流程管理的深度融合，为人工智能产业主体对于技术发展与安全保障提供解决方案。鉴于人工智能技术应用日新月异，本报告对人工智能风险治理的认识仍有未尽之处，恳请大家批评指正。

目 录

一、概述.....	1
（一）人工智能风险挑战引发全球关注.....	1
（二）人工智能风险治理模式亟需创新.....	3
（三）推动面向产业界的人工智能风险治理落地框架迫在眉睫.....	4
二、人工智能风险认知渐进深化.....	6
（一）人工智能内生安全与衍生安全复杂交织.....	6
（二）人工智能风险认知迈向多维度深度融合.....	9
三、人工智能风险治理全球实践.....	12
（一）国际层面：加速推进风险治理行动及合作.....	12
（二）主要经济体：迭代完善风险划分与管控举措.....	14
（三）产业组织：协同推进治理框架和安全工具.....	17
（四）企业主体：深化技管融合的风险应对方案.....	18
四、面向产业的人工智能风险治理实践方案.....	19
（一）打造全链条人工智能风险治理框架.....	19
（二）提升人工智能风险系统治理能力.....	21
（三）设立人工智能风险精准识别机制.....	23
（四）健全人工智能风险科学评估体系.....	25
（五）完善人工智能风险敏捷应对方案.....	27
五、展望建议.....	29
（一）促进全球合作，共筑人工智能治理新格局.....	29
（二）优化治理体系，夯实人工智能安全发展基础.....	30
（三）发挥桥梁作用，推进行业治理标准化与动态化.....	30
（四）强化内外协同，助力技术创新与责任共担.....	31

图 目 录

图 1 人工智能风险治理框架.....	5
图 2 人工智能内生安全和衍生安全	7
图 3 人工智能风险认知趋势	9
图 4 面向产业的人工智能风险治理实践方案	20

一、概述

（一）人工智能风险挑战引发全球关注

习近平总书记指出，人工智能是引领这一轮科技革命和产业变革的战略性技术，具有溢出带动性很强的“头雁”效应。当前，人工智能牵引工业、医疗、教育、交通、金融、文化等各领域提质增效，成为社会经济高质量发展的重要支撑和前进动能。然而，人工智能技术在重塑全球经济与社会格局的同时，也带来了多元风险挑战。根据美国斯坦福大学《2024 年人工智能指数报告》统计，2023 年全球共发生 123 起与人工智能滥用相关的重大事件，比 2013 年增长了近 20 倍，并呈现持续增长态势¹。与此同时，人工智能风险的波及范围和影响程度也显著加大。国际货币基金组织《生成式人工智能：人工智能与未来就业》报告指出，全球 40% 的工作将受到人工智能的影响²；奇安信《2024 人工智能安全报告》显示，2023 年基于人工智能的深度伪造欺诈暴增 3000%，钓鱼邮件增长 1000%³。

人工智能技术释放红利与应用风险显露并存，全球领域人工智能治理进程持续加速。国际组织和主要经济体在人工智能治理方面主动作为，联合国通过联合国大会将共识理念凝聚为决议文件，并通过组建人工智能高级别咨询机构以讨论形成全球合作的具体行动方案；经济合作与发展组织（OECD）研究、评查全球国家如何落实人工智能原则；金砖国家正推动成立人工智能研究小组以促进合作和负责任的

¹ <https://aiindex.stanford.edu/report/>

² <https://www.elibrary.imf.org/view/journals/006/2024/001/006.2024.issue-001-en.xml>

³ https://www.qianxin.com/threat/reportdetail?report_id=311