世界互联网大会
World Internet Conference

为人类共同福祉构建全球人工智能安全与治理体系

世界互联网大会人工智能专业委员会
安全与治理推进计划
2025年11月

世界互联网大会人工智能专业委员会
安全与治理推进计划成员及贡献者

牵头人

曾　毅　　中国科学院
北京前瞻人工智能安全与治理研究院

肖恩·欧·海格缇　　剑桥大学

专家及贡献者

（包括推进计划成员和特邀专家，以及为报告提供内容或意见的专家，
按单位名称首字母排序）

阿拉伯信息通信技术组织
纳达·拉阿比迪

阿里云计算有限公司
王星光、张　荣

巴西南大河州联邦大学
埃德森·普雷斯蒂

北京大学
王小塞、杨耀东

北京印象笔记科技有限公司
乔　迁

北京邮电大学
杨忠良

北京智源人工智能研究院
卜语嫣

德国埃尔朗根 - 纽伦堡大学
文森特·C·米勒

德国明斯特大学
伯纳德·霍尔兹纳格尔

德国慕尼黑工业大学
丹尼尔·凯里米

伏羲智库
程　凯、李　娜

杭州安恒信息技术股份有限公司
符春辉、王　欣

IQuilibriumAI
希梅娜·索菲亚·比维罗斯·阿尔瓦雷斯

金砖国家未来网络研究院（中国·深圳）
范　维、周　原

联想（北京）有限公司
胡永启

莫斯科国际关系学院
安娜·阿布拉莫娃

南开大学
陶　锋

清华大学
唐新华

人民网股份有限公司
汪凤琼

上海诺基亚贝尔股份有限公司
陶　涛、王　彤

上海人工智能实验室
乔　宇

上海稀宇极智科技有限公司
彭　韬、沈俊成

深圳市腾讯计算机系统有限公司
王梦寅、王　融

斯里兰卡计算机应急响应中心
查鲁卡·塞纳尔·达穆努波拉、尼罗沙·阿南达

香港岭南大学
姚　新

新浪微博
王　巍、张俊林

英国南安普顿大学
温迪·霍尔

中关村实验室
谭知行

中国互联网络信息中心
陈晶晶

中国人民大学
龚新奇、刘永谋

中国社会科学院人工智能研究促进中心
段伟文

中国信息通信研究院
郭苏敏、呼娜英

中国政法大学
张凌寒

世界互联网大会
梁　昊、张雪丽

香港大学
陈　澍

香港中文大学
蒙美玲、张寄冀

央视国际网络有限公司
程　明

云安全联盟
约翰·杨

中国电信集团有限公司
刘威辰、王　峰

中国联合网络通信集团有限公司
周　凯

中国社会科学院大学
刘晓春

中国网络空间安全协会
王健兵、夏文辉

中国移动研究院（中移智库）
林　琳、吴淑燕

中兴通讯股份有限公司
孟　伟

编写团队

北京前瞻人工智能安全与治理研究院
中国科学院自动化研究所
人工智能安全与超级对齐北京市重点实验室
曾　毅、王正奇、鲁恩萌、范津宇、皇甫存青

世界互联网大会
康彦荣、韩开宇

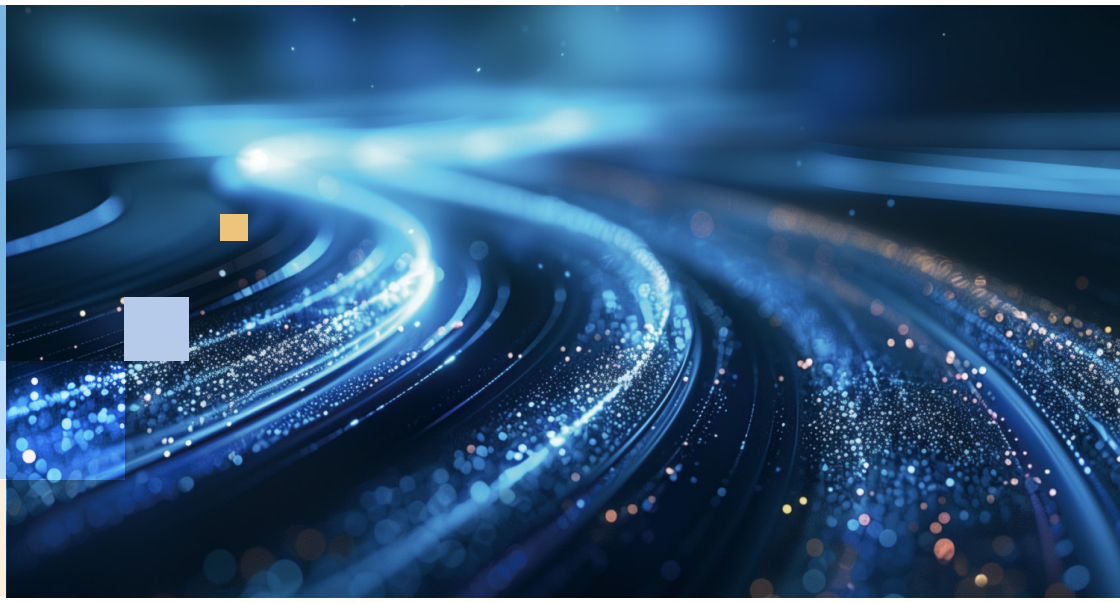
中国科学院大学
曹功策、陈　煜、谢佳玮、韩正强、郭晓阳、包傲日格乐、王　金

联系邮箱

research@wicinternet.org

前言

— PREFACE —



人工智能技术正以跨越式速度发展，深刻重塑全球科技、经济与安全格局。伴随其广泛应用，人工智能所引发的风险挑战日益复杂严峻。与此同时，全球人工智能安全与治理体系呈现出碎片化态势，各国在发展水平、治理能力、治理诉求等方面存在差异，国际协作面临阻力。在此背景下，探索构建完善的全球人工智能安全与治理体系，早日形成具有广泛共识的全球治理框架和标准规范，已成为关乎人工智能可持续发展的重要课题。

世界互联网大会依托人工智能专业委员会安全与治理推进计划成员单位，联合相关国际组织、知名智库、科研院所、专业协会及产业界专家，共同开展全球人工智能安全与治理体系研究，旨在促进各方凝聚共识、增进互信、加强协作，推动人工智能服务人类共同福祉与长远安全。

本报告系统梳理了当前全球人工智能安全与治理体系的探索与实践、应解决的关键问题及其他领域的全球治理多边实践经验。报告重点围绕机制设计展开分析，探讨如何构建确保安全、确保包容、权责明晰、协调有力、权威高效的全球人工智能安全与治理体系，并基于践行多边主义、共建人类命运共同体的理念，进一步提出以联合国为中心构建全球人工智能安全与治理体系的机制建议。本报告旨在为各方提供参考，以期凝聚广泛共识，共同推动全球人工智能安全与治理体系的构建与完善。

目 录

— PART —

01

为全球人工智能发展应用构建安全与治理体系

人工智能在带来显著发展机遇、重塑人类生产生活方式的同时，也引发了跨领域、多层次的系统性治理挑战，持续加剧全球治理格局的碎片化与制度脆弱性。**推动形成兼具可执行性与广泛共识的全球人工智能安全与治理体系，已成为人类需要共同面对的时代任务。**

（一）全球人工智能发展应用亟须应对安全与治理风险

当前，人工智能正以跨越式速度发展，深刻重塑全球科技、经济与安全格局。大模型在数学、编程等领域已显示出接近或达到人类专家级的推理能力；视频生成与合成技术在短时间内实现了从概念到高逼真度复杂场景视频生成的跃迁；具身智能的发展也正在推动通用机器人在复杂物理环境中的拟人化交互。同时，人工智能在科学、医学、气候、能源、交通等领域对重大问题的应对发挥了显著作用，创造了重要的公共价值与社会红利。

随着技术向更广泛、更深层次场景的拓展，治理挑战亦随之复杂化并呈系统性外溢。就技术内生风险而言，算法的快速迭代与能力跃升暴露出可解释性不足、鲁棒性薄弱与对抗性脆弱等安全隐患。在应用层面，人工智能的普及使算力安全、供应链脆弱性与跨境技术流动等问题日益凸显。区域与行业的关键系统日益依赖智能化组件，其连锁影响可能跨越能源、交通、金融与通信等基础设施。相关统计与风险评估表明，近年来人工智能相关的恶意利用与网络攻击呈明显上升态势。欧盟网络安全局（European Network and Information Security Agency，ENISA）发布的威胁评估报告显示[1]，在 2023 年 7 月至 2024 年 6 月期间共记录 11079 起攻击事件，其中 322 起为针对欧盟

1. 欧盟网络安全局，《ENISA Threat Landscape 2024》，来源：https://securitydelta.nl/media/com_hsd/report/690/document/ENISA-Threat-Landscape-2024.pdf

成员国的跨境攻击，受攻击的主要领域包括公共行政、交通运输、金融与数字基础设施（分别占比约 19%、11%、9% 与 8%）。同时，生成式人工智能驱动的深度伪造行为在 2024 年同比增长 118%[2]，全球相关网络犯罪造成的经济损失在 2025 年预计将高达 10.5 万亿美元[3]。此类风险已从技术层面向社会系统扩散，具有跨国界与跨领域传导特性，对现行治理体系构成严峻挑战。在此背景下，中国正通过《生成式人工智能服务管理暂行办法》[4]《人工智能生成合成内容标识办法》[5] 等法律法规，探索建立生成合成内容标识、安全评估等制度，来有效防范相关风险。

同时，人工智能风险的持续累积与地缘政治因素的叠加，正在进一步加剧全球治理困境。近年来，人工智能相关风险事件数量还在持续快速增长。如图 1 所示，2019 至 2024 年间，全球记录在案的人工智能风险事件由约 400 件跃升至 7900 余件，总量增长了近 20 倍。其中，涉及鲁棒性与数字安全、人权与隐私治理、透明度与问责制等问题的事件占比超过 60%，显示技术安全性与伦理性议题已成为全球性挑战。与此同时，部分国家通过出口管制、技术封锁等手段谋求战略优势[6]。尽管以“技术管制”名义实施的战略举措并未有效限制人工智能扩散，反而可能削弱政策制定国的技术优势[7]，这种零

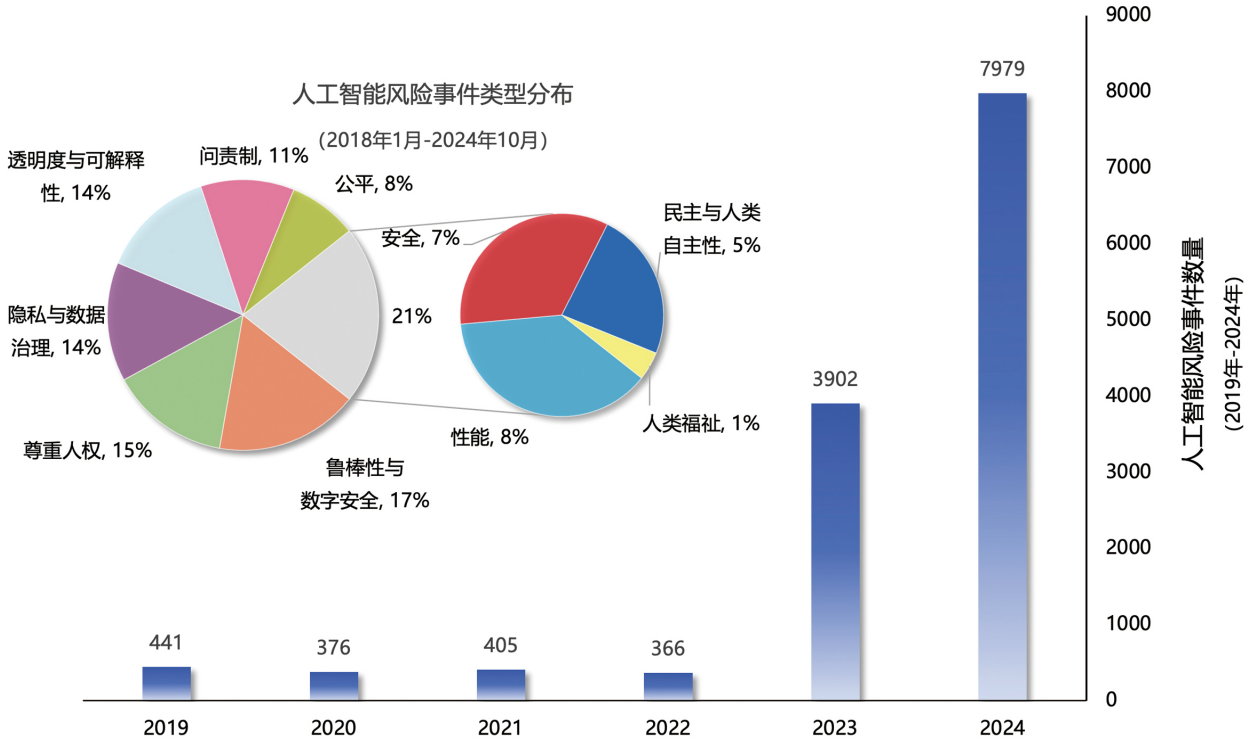


图 1 近年来全球人工智能风险事件数量变化及风险类型分布[8]

2. 欧洲议会，《Children and deepfakes》，来源：https://www.europarl.europa.eu/RegData/etudes/BRIE/2025/775855/EPRS_BRI(2025)775855_EN.pdf
3. Entrust 网络安全研究所，《2025 Identity Fraud Report》，来源：https://www.entrust.com/sites/default/files/documentation/reports/2025-identity-fraud-report.pdf
4. 《生成式人工智能服务管理暂行办法》，来源：https://www.cac.gov.cn/2023-07/13/c_1690898327029107.htm
5. 《人工智能生成合成内容标识办法》，来源：https://www.gov.cn/zhengce/zhengceku/202503/content_7014286.htm
6. 理解人工智能传播框架，来源：https://www.rand.org/pubs/perspectives/PEA3776-1.html
7. 新的人工智能扩散出口管制规则将削弱美国人工智能领导地位，来源：https://www.brookings.edu/articles/the-new-ai-diffusion-export-control-rule-will-undermine-us-ai-leadership
8. 原始数据来自经济合作与发展组织“人工智能风险事件和危害监测器（OECD AIM）”，由《全球人工智能治理评估指数 2025》统计分析获得，来源：https://agile-index.ai/publications/2025

和博弈思维正不断挤压国际合作空间，削弱各国参与全球人工智能安全与治理的能力和意愿，降低了国际社会集体应对系统性风险的效能。

面对当前治理格局呈现显著碎片化、安全与治理需求日益紧迫的现实，**从维护全人类共同安全与福祉出发，构建完善的全球人工智能安全与治理体系已刻不容缓。**人工智能的跨境属性使其风险与影响天然超越国界——无论是数据流动、算法部署，还是基础模型的应用与扩散，单一国家或地区的监管措施都难以形成有效约束，而任何国家又都无法置身事外。在技术风险持续演化并形成全球外溢效应的背景下，**尽早建立一个基于各方充分共识的，能够确保安全、确保包容、权责明晰、协调有力、权威高效的全球人工智能安全与治理体系，已成为应对系统性风险的关键路径。**这不仅是技术发展的必然要求，更是维护全球数字时代共同安全的重要保障。

（二）各方在全球人工智能安全与治理上的探索与实践

近年来，各方正通过多层次、多方向的努力，共同推进全球人工智能安全与治理体系的构建进程。联合国系统通过成立咨询机构、召开会议、发布报告、通过联大决议等方式，积极构建全球性的人工智能治理框架。**联合国秘书长**组建人工智能高级别咨询机构[9]，该机构发布《治理人工智能，助力造福人类》中期[10]及最后报告[11]，分析人工智能全球治理问题并提出建议。**联合国大会**（The United Nations General Assembly, UNGA）先后通过《抓住安全、可靠和值得信赖的人工智能

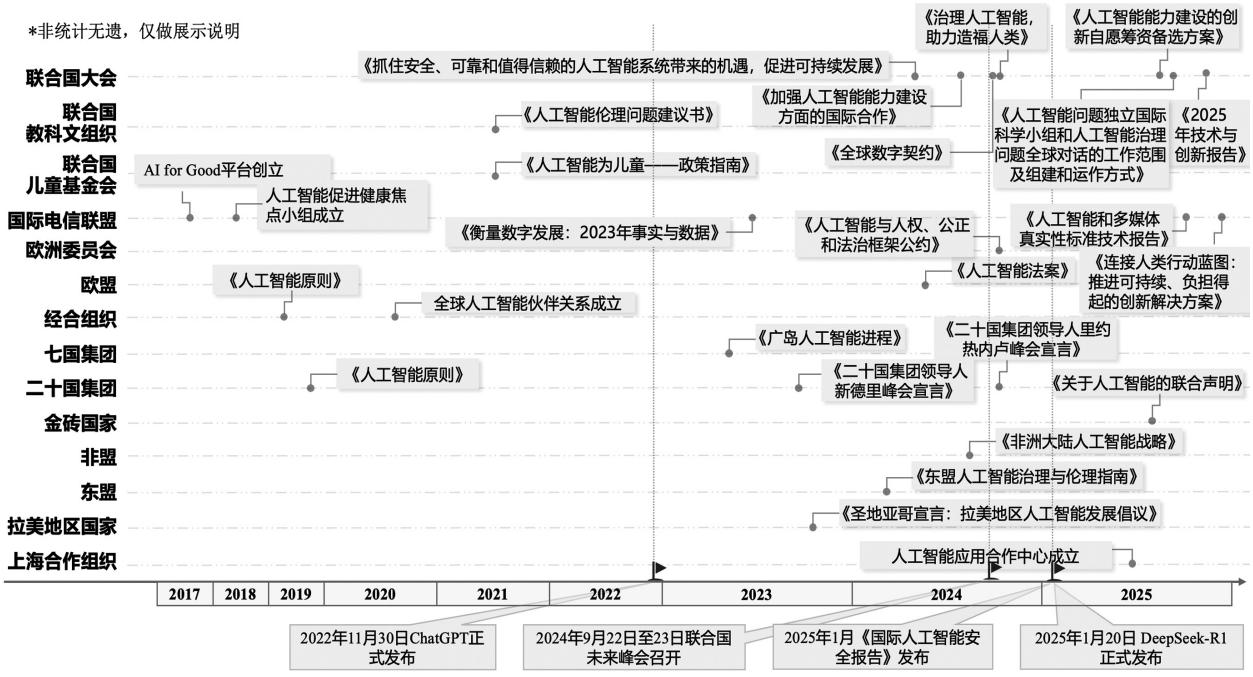


图 2 各个政府间组织在人工智能安全与治理方面的探索

9. 联合国秘书长组建人工智能高级别咨询机构，来源：https://news.un.org/zh/story/2023/10/1123382
10. 联合国，《治理人工智能，助力造福人类》中期报告，来源 https://www.un.org/sites/un2.un.org/files/ai_advisory_body_interim_report.pdf
11. 联合国，《治理人工智能，助力造福人类》最后报告，来源：https://www.un.org/sites/un2.un.org/files/governing_ai_for_humanity_final_report_zh.pdf

系统带来的机遇，促进可持续发展》[12]《加强人工智能能力建设方面的国际合作》[13]决议，召开未来峰会，通过《未来契约》[14]采纳《全球数字契约》[15]，并通过《人工智能问题独立国际科学小组和人工智能治理问题全球对话的工作范围及组建和运作方式》[16]等后续决议，持续推动人工智能全球治理议程实施。此外，联合国秘书长还向联合国大会递交《人工智能能力建设的创新自愿筹资备选方案》[17]报告，进一步为落实人工智能能力建设探索相关融资机制和筹资方案。**联合国教科文组织**（United Nations Educational, Scientific and Cultural Organization, UNESCO）发布《人工智能伦理问题建议书》[18]及其配套的《准备状态评估方法》[19]和《伦理影响评估》[20]，为各成员国和利益攸关方提供规范性伦理指导和评估。**国际电信联盟**（International Telecommunication Union, ITU）举办“人工智能向善全球峰会（AI for Good Global Summit）”[21]，峰会聚焦人工智能创新应用，致力于推动解决全球性挑战。**联合国儿童基金会**（United Nations Children's Fund，UNICEF）推出《人工智能为儿童——政策指南》[22]，提出了儿童友好的人工智能应遵循保护、赋能和友好的原则。

在国家层面，各国通过发布战略文件与行动倡议，积极构建治理框架。**美国**发布《人工智能风险管理框架》[23]，为公共部门和私营机构提供方法工具，用于识别和管理人工智能生命周期中的潜在风险；推动成立国际人工智能安全研究所网络[24]，在前沿模型安全评估方面开展国际合作和联合测试。**新加坡**发布了《人工智能模型治理框架（第二版）》[25]、《生成式人工智能治理模型框架：培育可信赖的生态系统》[26]，为政府、产业和研究机构在开发和应用人工智能时提供政策指引和实践参考。**沙特阿拉伯**发布《人工智能伦理原则》[27]，强调人工智能系统的安全性和可靠性。**中国**相继发布《全球人工智能治理倡议》[28]《人工智能安全治理框架》（1.0 版、2.0 版）[29]《人工智能全球治理行动计划》[30]，倡导发展和安全并重的原则，提出应对人工智能技术失控风险

12. 联合国，《抓住安全、可靠和值得信赖的人工智能系统带来的机遇，促进可持续发展》，来源：https://docs.un.org/zh/A/78/L.49
13. 联合国，《加强人工智能能力建设方面的国际合作》，来源：https://docs.un.org/zh/A/RES/78/311
14. 联合国未来峰会，《未来契约》，来源：https://docs.un.org/zh/A/RES/79/1
15. 联合国未来峰会，《全球数字契约》，来源：https://www.un.org/zh/documents/treaty/A-RES-79-1-Annex-I
16. 联合国大会决定设立人工智能独立国际小组，来源：https://docs.un.org/zh/A/RES/79/325
17. 联合国，《人工智能能力建设的创新自愿筹资备选方案》，来源：https://digitallibrary.un.org/record/4085951?ln=en&v=pdf#files
18. 联合国教科文组织，《人工智能伦理问题建议书》，来源：https://unesdoc.unesco.org/ark:/48223/pf0000380455_chi
19. 联合国教科文组织《准备状态评估方法》，来源：https://www.unesco.org/ethics-ai/en/ram?hub=32618
20. 联合国教科文组织《伦理影响评估》，来源：https://www.unesco.org/ethics-ai/en/eia?hub=32618
21. “人工智能向善” 由国际电联举办，拥有 50 多个联合国合作伙伴并与瑞士政府共同召集会议，来源：https://aiforgood.itu.int/#
22. 联合国儿童基金会关于制定维护儿童权利的人工智能政策和系统的建议，来源：https://www.unicef.org/innocenti/reports/policy-guidance-ai-children
23. 美国国家标准与技术研究院发布《人工智能风险管理框架》，来源：https://www.nist.gov/itl/ai-risk-management-framework
24. 美国商务部和美国国务院在旧金山举行的首次会议上启动国际人工智能安全研究所网络，来源：https://www.nist.gov/news-events/news/2024/11/fact-sheet-us-department-commerce-us-department-state-launch-international
25. 新加坡发布《人工智能模型治理框架》，来源：https://www.pdpc.gov.sg/-/media/Files/PDPC/PDF-Files/Resource-for-Organisation/AI/SGModelAIGovFramework2.pdf
26. 新加坡发布《生成式人工智能治理模型框架：培育可信赖的生态系统》https://aiverifyfoundation.sg/wp-content/uploads/2024/05/Model-AI-Governance-Framework-for-Generative-AI-May-2024-1-1.pdf
27. 沙特阿拉伯发布《人工智能伦理原则》，来源：https://sdaia.gov.sa/en/SDAIA/about/Documents/ai-principles.pdf
28. 中国发布《全球人工智能治理倡议》，来源：https://www.fmprc.gov.cn/web/ziliao_674904/1179_674909/202310/t20231020_11164831.shtml
29. 中国相继发布《人工智能安全治理框架》1.0 版和 2.0 版，来源：https://www.cac.gov.cn/2024-09/09/c_1727567886199789.htm, https://www.cac.gov.cn/2025-09/15/c_1759653448369123.htm
30. 中国，《人工智能全球治理行动计划》，来源：https://www.mfa.gov.cn/zyxw/202507/t20250726_11677803.shtml

的可信人工智能基本准则，并围绕技术监管、产业促进与国际协作提出系统性实施方案，倡议成立世界人工智能合作组织[31]，为全球治理提供中国方案。

在区域层面，欧洲地区率先建立了较为完善的规则和法律体系，欧盟委员会发布了《加强人工智能合作宣言》[32]《人工智能白皮书——通往卓越和信任的欧洲路径》[33]。欧洲议会和欧盟理事会颁布实施《人工智能法案》[34]，建成全球首部全面监管人工智能的法律框架。欧洲委员会推出了全球首部具有法律约束力的人工智能国际条约《人工智能与人权、公正和法治框架公约》[35]，旨在确保人工智能系统生命周期内的活动完全符合人权、民主、法治，并有利于科技创新。**在阿拉伯地区，**阿拉伯国家联盟通过其阿拉伯人工智能工作组，并与阿拉伯信息和通信技术组织等区域专业组织紧密合作，相继制定并通过了《阿拉伯人工智能战略》[36]和《阿拉伯人工智能伦理宪章》[37]，推动该地区人工智能治理与伦理框架建设迈向区域一致性。**在非洲地区，**非洲联盟发布《非洲大陆人工智能战略》[38]，确立区域层面的政策框架与行动方向；在此基础上，非洲多国签署了《非洲人工智能宣言》[39]，形成了广泛的政治共识与协同行动愿景。**在东南亚地区，**东南亚国家联盟发布《东盟人工智能治理与伦理指南》[40]并在后续发布《东盟人工智能治理与伦理指南——生成式人工智能》[41]，概述生成式人工智能风险，建议在东盟范围内落实确保负责任人工智能的政策举措。**在拉美地区，**拉丁美洲和加勒比地区国家代表达成共识并发布《圣地亚哥宣言：拉美地区人工智能发展倡议》[42]，发布《蒙得维的亚宣言：为构建人工智能治理及其社会影响的区域性方法》[43]，鼓励地区协调和开展高级别对话。这些举措强调区域协同与伦理责任，并呼吁在全球范围内推进公平与包容的治理框架，体现了不同地区在人工智能发展与治理上的积极探索。

国际组织方面，二十国集团（G20）新德里峰会发布《二十国集团领导人新德里峰会宣言》[44]，进一步声明实现人工智能向善并服务全人类；

31. 中国政府倡议成立世界人工智能合作组织，来源：https://www.gov.cn/yaowen/liebiao/202507/content_7033957.htm
32. 欧盟委员会，《加强人工智能合作宣言》，来源：https://digital-strategy.ec.europa.eu/en/policies/plan-ai
33. 欧盟委员会，《人工智能白皮书——通往卓越和信任的欧洲路径》，来源：https://commission.europa.eu/publications/white-paper-artificial-intelligence-european-approach-excellence-and-trust_en
34. 欧盟委员会，《人工智能法案》，来源：https://artificialintelligenceact.eu/the-act/
35. 欧洲委员会，《人工智能与人权、民主和法治框架公约》，来源：https://www.coe.int/en/web/conventions/full-list?module=treaty-detail&treatynum=225
36. 《阿拉伯人工智能战略》，来源：https://www.aicto.org/publications/studies/#flipbook-df_9695/1/
37. 《阿拉伯人工智能伦理宪章》，来源：https://www.aicto.org/publications/studies/#flipbook-df_9686/1/
38. 非洲联盟，《非洲大陆人工智能战略》，来源：https://au.int/en/documents/20240809/continental-artificial-intelligence-strategy
39. 非洲人工智能全球峰会，《非洲人工智能宣言》，来源：https://c4ir.rw/docs/Africa%20Declaration%20on%20Artificial%20Intelligence-FINAL-31-March-2025.pdf
40. 东南亚联盟，《东盟人工智能治理与伦理指南》，来源：https://asean.org/book/asean-guide-on-ai-governance-and-ethics/
41. 东南亚联盟，《东盟人工智能治理与伦理指南——生成式人工智能》，来源：https://mekongdataprotection.org/asean-guide-on-ai-governance-and-ethics-generative-ai
42. 《圣地亚哥宣言：拉美地区人工智能发展倡议》，来源：https://minciencia.gob.cl/uploads/filer_public/40/2a/402a35a0-1222-4dab-b090-5c81bbf34237/declaracion_de_santiago.pdf
43. 《蒙得维的亚宣言：为构建人工智能治理及其社会影响的区域性方法》，来源：https://www.gub.uy/agencia-gobierno-electronico-sociedad-informacion-conocimiento/sites/agencia-gobierno-electronico-sociedad-informacion-conocimiento/files/documentos/noticias/EN%20-%20Montevideo%20Declaration%20approved.pdf
44. G20,《二十国集团领导人新德里峰会宣言（摘要）》，来源：https://www.mfa.gov.cn/web/gjhdq_676201/gjhdqzz_681964/ershiguojituan_682134/xgxw_682140/202309/t20230910_11140667.shtml

2024 年里约热内卢峰会发布《二十国集团领导人里约热内卢峰会宣言》[45]，强调促进有利于创新的人工智能治理,加强合作并赋能可持续发展。**七国集团**（G7）发布《广岛人工智能进程》[46]，提出建立共同的治理框架，强调在安全、透明和负责任应用方面开展合作。**经济发展与合作组织**（Organisation for Economic Co-operation and Development，OECD）发布《经合组织人工智能原则》[47]，发起全球人工智能伙伴关系[48]，推动人工智能政策实践、技术研究与能力建设的国际协作。**亚太经济合作组织**（Asia-Pacific Economic Cooperation，APEC）发布《亚太经合组织人工智能倡议》[49]，致力于通过促进安全、可及、可靠的人工智能生态系统建设，实现具有韧性和包容性的经济增长。**金砖国家**（BRICS）在《金砖国家领导人第十七次会晤里约热内卢宣言》[50]以及《金砖国家领导人关于人工智能全球治理的声明》[51]中，积极倡导以包容、公平和可持续发展为导向的国际人工智能治理体系与合作平台。**上海合作组织**（Shanghai Cooperation Organization，SCO）于 2025 天津峰会发表《关于进一步深化人工智能国际合作的声明》[52]，倡导加强人工智能基础设施、人才培养、投资等领域合作，发展人工智能领域对话伙伴机制，聚焦人工智能产业的可持续发展和克服人工智能引发的潜在风险及挑战。**世界互联网大会**（World Internet Conference, WIC）相继发布了《发展负责任的生成式人工智能研究报告及共识文件》[53]《以普惠包容的人工智能治理赋能全球可持续发展》[54]等成果，提出应积极倡导并稳妥推进生成式人工智能的发展，并通过普惠包容的人工智能发展与治理来弥合数字鸿沟。

各重要国际峰会通过确立原则与构建机制，围绕各个议题凝聚全球共识。**人工智能安全峰会**（AI Safety Summit）、**人工智能首尔峰会**（AI Seoul Summit）、**人工智能行动峰会**（AI Action Summit）系列会议自 2023 年以来，相继达成了一系列自愿性承诺成果，包括《布莱切利宣言》[55]《安全、创新和包容性人工智能首尔宣言》[56]《前沿人工智能安全承诺》[57]和《关于发展包容、可持续的人工智能造福人类与地球的声明》[58]。**军事**

45. G20，《二十国集团领导人里约热内卢峰会宣言（摘要）》，来源：https://www.gov.cn/yaowen/liebiao/202411/content_6988277.htm
46. 七国集团发布《广岛人工智能进程》，来源：https://g7g20-documents.org/database/document/2023-g7-japan-ministerial-meetings-ict-ministers-ministers-language-g7-hiroshima-ai-process-g7-digital-tech-ministers-statement
47. 经合组织发布《经合组织人工智能原则》来源：https://www.oecd.org/en/topics/sub-issues/ai-principles.html
48. 经合组织，全球人工智能伙伴关系，来源：https://oecd.ai/en/about/about-gpai
49. 亚太经合组织，《亚太经合组织人工智能倡议》，来源：https://www.apec.org/meeting-papers/leaders-declarations/2025/2025-apec-leaders--gyeongju-declaration/apec-artificial-intelligence-(ai)-initiative-(2026-2030)
50. 金砖国家，《金砖国家领导人第十七次会晤里约热内卢宣言（全文）》，来源：https://www.gov.cn/yaowen/liebiao/202507/content_7031158.htm
51. 金砖国家，《金砖国家领导人关于人工智能全球治理的声明》（2025 年 7 月），来源：https://www.mfa.gov.cn/ziliao_674904/1179_674909/202507/t20250709_11668022.shtml
52. 上海合作组织成员国元首理事会关于进一步深化人工智能国际合作的声明，来源：https://chn.sectsco.org/20250901/1969229.html
53. 发展负责任的生成式人工智能研究报告及共识文件，来源：https://cn.wicinternet.org/2023-11/09/content_37486498.htm
54. 世界互联网大会，《发展负责任的生成式人工智能共识》，来源：https://cn.wicinternet.org/2023-11/09/content_37486498.htm
55. 人工智能安全峰会，《布莱切利宣言》，来源：https://www.gov.uk/government/publications/ai-safety-summit-2023-the-bletchley-declaration/the-bletchley-declaration-by-countries-attending-the-ai-safety-summit-1-2-november-2023
56. 人工智能首尔峰会，《安全、创新和包容性人工智能首尔宣言》，来源：https://www.gov.uk/government/topical-events/ai-seoul-summit-2024/about
57. 前沿人工智能安全承诺，来源：https://www.gov.uk/government/publications/frontier-ai-safety-commitments-ai-seoul-summit-2024/frontier-ai-safety-commitments-ai-seoul-summit-2024
58. 法国巴黎举行的人工智能行动峰会，《关于发展包容、可持续的人工智能造福人类与地球的声明》，来源：https://www.elysee.fr/emmanuel-macron/2025/02/11/declaration-sur-une-intelligence-artificielle-inclusive-et-durable-pour-les-peuples-et-la-planete

领域负责任使用人工智能系列峰会（Summit on Responsible Artificial Intelligence in the Military Domain，REAIM）自 2023 年以来先后签署发布了《军事领域负责任使用人工智能峰会行动倡议》[59] 和《军事领域负责任使用人工智能峰会行动蓝图》[60] 并发布《责任导向：军事领域人工智能的风险、机遇与治理战略指导报告》[61]，体现了各国对在军事领域负责任地使用人工智能议题的关切。

此外，**学术界和产业界**也在积极探索风险应对机制并推动自律自治。**学术界**通过理论构建与国际共识塑造，增进对前沿人工智能风险的认知与意识。国际人工智能安全对话 [62] 通过组织全球顶尖专家开展交流，凝聚应对人工智能极端风险的科学共识，倡导建立国际安全标准与治理框架；非营利组织人工智能安全中心（Center for AI Safety，CAIS）联合学者与政策制定者共同签署《人工智能风险声明》[63]，推动形成全球治理合力。**产业界**通过构建自主安全框架，来推动对高风险人工智能系统的审慎部署与行业共识。多家人工智能企业先后发布各自的风险管理框架，旨在确保人工智能系统的安全部署，并将前沿模型能力与其所带来的潜在风险挂钩，例如 Anthropic 的“负责任扩展政策”[64]、OpenAI 的“应对准备框架”[65]、Google DeepMind 的“前沿安全框架”[66] 以及 Meta 的“以结果为导向的前沿人工智能框架”[67] 等。与此同时，中国人工智能产业发展联盟发布《人工智能安全承诺》[68]，聚焦研发与应用环节的合规性、透明性与可控性，倡导企业主动落实安全主体责任。

（三）全球人工智能安全与治理体系应解决的关键问题

尽管全球各方在人工智能安全、伦理规范、能力建设及跨境合作等领域已开展广泛探索并取得一定成效，但面对人工智能发展带来的风险与挑战，**当前全球治理努力仍存在结构性局限。具体表现在以下几个方面：第一**，人工智能技术突破与应用路径高度不确定，且风险复杂多样，导致治理措施难以实现前瞻部署与技术动态相适应，敏捷协同的全球应对机制尚未形成；**第二**，各国在人工智能发展水平与治理模式上存在差异，致使参与能力与制度诉求显著不一，国际共识凝聚面临挑战；**第三**，人工智能领域存在多利益攸关方，倘若缺乏清晰的责任归属及有效的监督落实，会阻碍协调有力的全球行动；**第四**，当前地缘政治、意识形态及经济利益分歧引发的脱钩断链等干扰因素，

59. 2023 军事领域负责任使用人工智能峰会行动倡议，来源：https://www.government.nl/documents/publications/2023/02/16/reaim-2023-call-to-action
60. 2024 军事领域负责任使用人工智能峰会行动蓝图，来源：https://overseas.mofa.go.kr/eng/brd/m_5674/view.do?seq=321055
61. 《责任导向：军事领域人工智能的风险、机遇与治理战略指导报告》（Responsible by Design: Strategic Guidance Report on the Risks, Opportunities, and Governance of AI in the Military Domain），来源：https://hcss.nl/wp-content/uploads/2025/09/GC-REAIM-Strategic-Guidance-Report-Final-WEB.pdf
62. 人工智能安全国际对话（IDAIS）汇聚全球资深科学家，共同应对人工智能带来的极端风险，来源：https://idais.ai/dialogues
63. 人工智能安全中心（CAIS），《人工智能风险声明》，来源：https://aistatement.com/
64. Anthropic 发布负责任扩展政策，来源： https://www.anthropic.com/responsible-scaling-policy
65. OpenAI 发布应对准备框架，来源：https://cdn.openai.com/pdf/18a02b5d-6b67-4cec-ab64-68cdfbddebcd/preparedness-framework-v2.pdf
66. GoogleDeepMind 的“前沿安全框架，来源：https://storage.googleapis.com/deepmind-media/DeepMind.com/Blog/introducing-the-frontier-safety-framework/fsf-technical-report.pdf
67. Meta 发布以结果为导向的前沿人工智能框架，来源：https://about.fb.com/news/2025/02/meta-approach-frontier-ai/
68. 中国人工智能产业发展联盟起草发布《人工智能安全承诺》，来源：https://aihub.caict.ac.cn/ai_security_and_safety_commitments

正在削弱全球合力应对人工智能风险的能力。

为此，从人类命运共同体的理念出发，构建具有广泛国际共识的全球人工智能安全与治理体系，必须着力解决以下关键问题：

一是如何有效应对人工智能快速发展带来的技术安全风险及其衍生风险的不确定性，确保全球在面对潜在重大威胁时具备充分应对能力来保障安全可控。

二是如何平衡当前不同国家在发展阶段、制度文化、治理能力及治理诉求方面的差异，最大范围地促进各国共同发展，实现人工智能红利的普惠共享。

三是如何协调人工智能多利益攸关方之间的复杂互动，充分发挥各方的能动性与优势，形成合力，推动构建多元共治、协同高效的行动格局，避免失序带来的治理真空。

四是如何克服地缘政治与意识形态等障碍分歧，逐步构建起具备权威性与高效性的全球人工智能安全与治理体系，并保障其能够长远、可持续地运行。

总体来看，上述四个关键问题相辅相成，共同构成了全球人工智能治理体系的核心支柱：确保安全是治理体系建设的首要目标；确保包容是建立公正治理体系的基础，也是促进各国共同协商、共同建设、共享成果的前提；明晰权责是推动各方积极行动、实现治理体系协调有序、高效有力运行的关键支撑；面向未来则是保障治理体系能够排除阻碍因素和短视行为的干扰，汇聚合力，实现全球人工智能安全与治理体系稳健、可持续运行的根本价值指引。

（四）其他领域全球性多边治理体系的经验借鉴

在核安全、气候变化、国际贸易、公共卫生等关乎人类整体安全、跨境协作与风险应对的关键领域，国际社会已通过缔结多边协定、设立专门机构、构建争端解决机制等路径，逐步建立起各领域的全球治理体系。这些实践为进一步构建和完善人工智能的全球安全与治理体系提供了重要的机制参照与经验借鉴。

在国际核安全领域，国际原子能机构（International Atomic Energy Agency，IAEA）针对核技术潜在风险构建了严格检查[69]与技术援助[70]相结合的机制；通过建立权威监管体系[71]、支持履约监督与核查及能力建设[72]等，推动全球核安全的协作与信任，形成了兼顾强制性与合作性的综合治理模式，可为全球人工智能安全与治理体系建立可信履约机制、强化风险预防与应急响应能力提供有益借鉴。

在气候变化领域，联合国政府间气候变化专

69. 《不扩散核武器条约》，防止核武器扩散推动核裁军和促进和平利用核能，来源：https://disarmament.unoda.org/en/our-work/weapons-mass-destruction/nuclear-weapons/treaty-non-proliferation-nuclear-weapons
70. 《核事故或辐射紧急情况援助公约》为相关国家提供一个请求和提供援助的国际框架，来源：https://www.iaea.org/topics/nuclear-safety-conventions/convention-assistance-case-nuclear-accident-or-radiological-emergency
71. 《核事故早期通报公约》，要求缔约国在发生可能对其他国家产生放射学安全影响的核事故时立即通报，来源：https://www.iaea.org/topics/nuclear-safety-conventions/convention-early-notification-nuclear-accident
72. 《核安全公约》，确保民用核电厂的安全运行，通过同行审议促进核安全持续改进，来源：https://www.iaea.org/topics/nuclear-safety-conventions/convention-nuclear-safety

门委员会（Intergovernmental Panel on Climate Change，IPCC）[73] 针对全球气候变迁的不确定性构建了科学评估 [74] 与政策建议 [75] 紧密衔接的工作机制；通过多轮次的跨学科科学评估、专家协作和数据共享平台 [76]，推动了全球气候共识的形成与应对行动的协调 [77]，形成了以科学支撑政策、动态更新治理的决策模式，可为全球人工智能安全与治理在高不确定性和跨领域交织的复杂风险环境中，建立基于科学证据的风险评估体系提供参考。

在国际贸易领域，世界贸易组织（World Trade Organization，WTO）[78] 针对成员国间的贸易摩擦与规则冲突，构建了系统化的争端解决机制 [79]；通过统一规则执行、仲裁裁决与多方权益平衡，最终推动了跨境贸易的高效运行与公平有序 [80]；形成了以规则为基础、兼顾各方利益的治理模式 [81]，可为全球人工智能安全与治理体系设计透明、公正且可执行的争端解决机制借鉴经验。

在公共卫生领域，世界卫生组织（World Health Organization，WHO）针对跨国疫情传播风险，构建了覆盖疫情信息上报、跨区域感染源追踪、应急隔离与资源调配等环节的制度化工作机制 [82]；通过统一信息通报、资源协调和应急响应 [83]，提升了全球公共卫生体系的协同防控与危机应对能力 [84]；形成了跨境联动、快速响应的治理模式，为全球人工智能安全与治理体系建立高效的信息共享与风险预警机制，以及建设面对突发性风险扩散时实现快速的多边协调与干预能力提供了有益参照。

在全球航空海运领域，国际民航组织（International Civil Aviation Organization，ICAO）[85] 与国际海事组织（International Maritime Organization，IMO）[86] 针对跨国运输安全与系统互操作问题，构建了覆盖标准制定、

73. 《联合国气候变化框架公约》，全球应对气候变化的基础性的法律框架，来源：https://www.un.org/climatesecuritymechanism/en/united-nations-framework-convention-climate-change-unfccc-and-climate-peace-and-security
74. IPCC 气候变化相关专题的知识状况的全面和平衡的评估，来源：https://www.ipcc.ch/about/preparingreports/
75. IPCC 委员会协助政策制定者，来源：https://www.ipcc.ch/report/ar6/syr/summary-for-policymakers/
76. IPCC 数据共享平台，来源：https://www.ipcc.ch/data/
77. 《京都议定书》，设定了具有法律约束力的量化减排和限制目标，来源：https://unfccc.int/kyoto_protocol
78. 《成立世界贸易组织协议》，宣布建立世界贸易组织并赋予其管理多边贸易规则、提供谈判平台和解决争端职能的法律基础，来源：https://www.wto.org/english/res_e/booksp_e/agrmntseries1_wto_e.pdf
79. 《世贸组织协定》附件 2，来源：https://www.wto.org/english/tratop_e/dispu_e/dsu_e.htm
80. WHO 解决争端机制，来源：https://www.wto.org/english/thewto_e/whatis_e/tif_e/disp1_e.htm
81. 《关税及贸易总协定》，为商品贸易设定了非歧视、互惠和关税减让等基本规则，来源：https://www.wto.org/english/res_e/publications_e/ai17_e/gatt1994_e.htm
《服务贸易总协定》，旨在将多边贸易规则扩展至服务贸易领域，为其建立了透明、可预测和逐步自由化的法律框架，来源：https://www.wto.org/english/tratop_E/serv_e/gatsintr_e.htm；
《与贸易有关的知识产权协议》，为知识产权监管设定了最低国际标准，将知识产权纳入多边贸易体系，以解决与贸易相关的知识产权问题，来源：https://www.wto.org/english/tratop_e/trips_e/ta_modules_e.htm
82. 《国际卫生条例》，一项具有法律约束力的全球公共卫生安全框架，旨在帮助各国预防和应对可能构成国际关注的突发公共卫生事件，来源：https://www.who.int/health-topics/international-health-regulations#tab=tab_1
83. 《国际卫生条例》突发事件委员会，来源：https://www.who.int/teams/ihr/ihr-emergency-committees
84. 《烟草控制框架公约》，首个由世界卫生组织主持谈判达成的全球公共卫生条约，旨在通过减少烟草消费和供应来应对全球烟草流行，来源：https://wkc.who.int/resources/publications/i/item/9241591013
85. 《芝加哥条约》，为国际民用航空活动制定基本法律框架和原则，并设立了国际民用航空组织（ICAO），来源：https://www.icao.int/convention-international-civil-aviation-doc-7300
86. 《联合国海洋法公约》，一部被誉为“海洋宪法”的国际条约，系统地为所有海洋和海上活动建立了法律框架和秩序，来源：https://www.imo.org/en/ourwork/legal/pages/unitednationsconventiononthelawofthesea.aspx

风险管理、事故调查和合规检查的综合治理体系[87]；通过统一技术标准、强化跨系统风险管控和持续推进安全适航评估，提升了国际运输体系的安全性与互操作性保障能力，形成了标准协同、持续监督的治理模式，有助于指导全球人工智能安全与治理体系在基础设施安全、跨平台互操作性以及事故溯源等方面建立统一标准和责任体系，确保高风险场景下的运行安全与问责透明。

尽管上述领域的全球治理体系在制度设计、履约监督和风险应对方面积累了丰富经验，但人工智能技术在生成性、自主性、可演化性和跨领域渗透性等方面具有前所未有的复杂特征，其风险外溢范围更广、影响链条更长、治理边界更难界定。因此，**应对全球人工智能安全与治理挑战不能简单照搬现有体系模式，而需在借鉴既有机制的基础上，探索适应算法透明度、数据跨境流动、模型对齐与系统安全等新问题、新挑战的新制度，创新性地构建兼顾技术特性与公共利益的全球人工智能安全与治理体系。**

表 1 其他领域全球性多边治理体系的经验

治理领域	代表性治理机构	典型治理措施	核心机制框架	全球人工智能治理可借鉴的经验
核安全	国际原子能机构	核查机制与技术援助相结合，防止核扩散；核设施安全运行，事故通报	《不扩散核武器条约》 《核安全公约》 《核事故早期通报公约》 《核事故或辐射紧急情况援助公约》	强制性与合作性并重的综合治理，推动跨国信任与协作
气候变化	联合国政府间气候变化专门委员会	多轮次跨学科评估、数据共享平台，升温趋势预测，减排政策制定，气候行动跨国协作	《联合国气候变化框架公约》 《京都议定书》 《巴黎协定》	以科学支撑政策、动态更新的治理模式，强化前瞻性与科学性
国际贸易	世界贸易组织	通过系统化争端解决机制，解决关税争端，知识产权冲突以及贸易壁垒谈判，多方权益平衡	《成立世界贸易组织协议》 《关税及贸易总协定》 《服务贸易总协定》 《与贸易相关的知识产权协议》	规则明确，流程清晰且兼顾各利益攸关方利益的争议解决和仲裁机制
公共卫生	世界卫生组织	疫情上报，跨区域感染源追踪，应急隔离与资源调配机制	《国际卫生条例》 《烟草控制框架公约》	跨境联动、快速响应，全球危机防控与应急协作
全球航空海运领域	国际民航组织 国际海事组织	通过制定技术标准，开展安全事故权威调查和适航评估，保障航空航运安全运行并推动跨境规则互操作	《国际海上人命安全公约》 《联合国海洋法公约》 《约克 - 安特卫普规则》 《芝加哥条约》 《蒙特利尔公约》	标准协同、持续监督的治理模式，推动统一标准与跨系统风险管理

87. 《国际海上人命安全公约》，一部在国际海事组织（IMO）管理下，旨在为商船的安全建造、设备和操作制定统一标准的国际条约，来源：https://www.imo.org/en/knowledgecentre/conferencesmeetings/pages/solas.aspx
《约克 - 安特卫普规则》，一套有数百年历史，用于规范在共同海事冒险中为保护财产而做出的牺牲和支出的分摊的海上惯例规则，来源：https://charles-taylor-group.s3.amazonaws.com/production/filer_public/e8/c6/e8c600ba-e591-43e3-a9d0-3a885e61b776/rhl_-_york_antwerp_rules_2016_-_a_summary_of_the_changes.pdf
《蒙特利尔公约》，旨在统一和现代化国际航空运输中关于旅客、行李和货物运输责任的法律规则，来源：https://store.icao.int/en/convention-for-the-unification-of-certain-rules-for-international-carriage-by-air-doc-9740

— PART —

02

确保安全：应对人工智能快速变革与重大风险

本章从安全视角分析人工智能发展带来的多层次风险，探讨相应的全球人工智能安全与治理机制设计。前两节分别聚焦技术快速迭代的不确定性及广泛应用下的滥用风险，揭示了治理滞后、工具不足和跨境监管等难题；第三节则关注可能影响全球战略安全、关键基础设施及人类生存的系统性重大风险，强调亟需前瞻研判、国际协调及紧急干预机制，构建动态适应、跨国协作的人工智能安全与治理体系

（一）应对人工智能技术快速迭代与不确定性风险

当前，人工智能技术发展呈现出**迭代速度快、突破路径难以预测、模型能力与安全特性复杂**等特征。**首先，**技术能力水平快速迭代，提升速度显著超越传统的线性预测。例如，大语言模型的参数规模与推理性能在短时间内实现数量级突破[88]，在不到三年内已实现从基础对话到在数学、编程等领域展现出高水平推理能力的跨越。与此同时，新技术不断涌现，如扩散模型、自监督学习等范式迅速应用于多类任务；多模态大模型持续取得突破；推理式人工智能、代理式人工智能、具身智能等技术快速成熟。**其次，**技术路径突破具有高度不可预测性。例如，Transformer 架构在短时间内取代循环神经网络成为自然语言处理的主流范式；一些高级能力并非随着模型规模而线性提升，而是在参数量或计算量达到特定临界点后才突然“涌现”[89]。**最后，**技术体系存在固有脆弱性与不稳健性，而对其复杂安全特性的认知仍然缺乏。例如，大型模型在对抗性攻击下易产生不可控的错误输出，尚难以有效防范；目前对于模型幻觉、偏见以及安全对齐等机理的认知也尚不充分。

88. 一年内，前沿 AI 模型在多项严苛基准测试中取得了显著突破，来源：https://hai.stanford.edu/assets/files/hai_ai_index_report_2025_chinese_version_061325.pdf
89. Wei J, Tay Y, Bommasani R, et al. Emergent abilities of large language models[J]. arXiv preprint arXiv:2206.07682, 2022.

人工智能技术发展的上述特性，进一步给传统治理机制带来了多重挑战。一是在方向上看不准，难以进行先导部署。关键性技术突破的发生时点和路径具有较高不确定性，使得治理措施难以进行有效的前瞻性部署，易陷入被动应对；同时，系统性风险监测指标和能力的缺乏，也导致监管措施常落后于技术现实。**二是在节奏上跟不上，动态调整不足。**全球多边谈判与国际规则制定通常以数年为单位，与人工智能技术按月甚至按周的迭代周期形成巨大的“速度差”，治理规则的制定与更新常常滞后于技术的快速演进；各国各地区的监管措施也存在明显不同步，带来了潜在的治理“缝隙”。**三是在工具上不完善，技术支撑有限。**由于监管缺乏系统化的风险评估与干预工具，特别是针对模型应用安全风险的监测指标与评估平台尚未成熟，导致监管机构在风险识别、评估和干预方面的全面性与有效性受到制约。

对此，构建与人工智能技术发展相适配的全球安全与治理体系，应把握“科学、敏捷、协作”的原则，重点加强以下机制建设。一是建立技术跟踪与风险预警的协同机制。在联合国等多边框架下，推动形成跨国技术监测网络，促进各方对前沿人工智能模型能力与安全事件信息的主动披露与共享。定期组织发布前沿模型能力与风险的国际联合评估测试，共同强化对新兴和潜在风险的共同科学认知与预警能力。**二是建立治理规则的动态更新与互认机制。**推动相关国际标准组织建立完善人工智能安全标准的常态化、定期审议与更新程序，鼓励主要经济体率先在高风险应用领域实现测试结果与认证互认，通过规则的动态修正与互操作，有效防范监管套利，确保治理的敏捷与协同。**三是共建安全评测共性工具与平台生态。**鼓励通过国际合作，共同开发开放、高互操作性的人工智能安全测试平台、基准数据集与风险评估工具库，建设全球共享的技术工具箱，为各国监管机构提供精准、有效、协同的能力支撑，持续提升监管效能。

（二）应对人工智能广泛社会应用与滥用恶用风险

近年来，以大规模生成式模型与开源生态为代表的新一代人工智能技术，显著降低了开发与应用门槛，推动了能力的快速普及。例如 ChatGPT 在 2022 年底发布仅两个月后迅速达到亿级月活跃用户[90]，DeepSeek 等开源模型则进一步推动了能力下沉与成本下降。**在此背景下，人工智能能力正快速嵌入社会生产、公共服务、科研创新与信息传播等各个环节，技术的影响迅速扩展至社会各领域。在社会经济领域，**人工智能深度嵌入金融交易、智能制造与供应链管理，显著提升了资源配置效率与生产力水平，但也带来了市场集中、就业结构调整与技术依赖等风险，对经济安全与公平竞争形成挑战。**在社会民生领域，**人工智能广泛应用于教育、医疗和公共服务，提升了服务的普及性与精准度，同时算法决策不透明、数据偏见与数字鸿沟等问题也引发了对公平性与隐私保护的持续关注。**在科技创新领域，**人工智能助力蛋白质结构预测、新材料发现和药物研发，

90. ChatGPT 创下用户群增长最快的纪录，来源：https://www.reuters.com/technology/chatgpt-sets-record-fastest-growing-user-base-analyst-note-2023-02-01/

加速了科研进展，但科研伦理、成果归属与数据可验证性等问题日益突出，科研诚信面临新的考验。**在法律与制度方面，**生成式人工智能对内容创作与传播机制产生深远影响，在知识产权归属、侵权责任认定与合理适用范围等方面存在法律争议，对现行法律体系的适应性提出新的要求。**在伦理与社会价值方面，**人工智能生成内容可能模糊真实与虚构边界，模型偏见、操控与情感干预风险也对人机关系与社会价值体系形成新的伦理挑战。

与此同时，对人工智能的滥用和恶用正持续放大社会风险，加剧社会治理的复杂性与紧迫性。利用人工智能辅助的欺诈行为正变得复杂和频繁，数据显示 2024 年平均每五分钟就发生一次深度伪造（DeepFake）攻击，人工智能辅助的数字文件伪造同比增长 244%，首次超过物理伪造成为主要的欺诈方法[91]；与此同时，在招聘、信贷审批等领域中的自动化应用引发了社会歧视[92]；伪造政治人物言论、制造虚假舆论事态已多次干扰选举与外交活动[93]；基于跨平台行为数据的用户画像和心理操纵等，使得隐私侵犯行为因人工智能强大的数据关联与分析能力而变得更加隐蔽和规模化[94]。这些风险不仅具有显性的社会危害，更可能侵蚀社会信任基础。

从全球治理视角看，单一国家的治理措施已难以充分应对人工智能的广泛社会影响与风险，而有效的全球协同框架仍面临多重挑战。一是风险认知的共识不足。各国人工智能应用场景及对风险评估与分类分级的标准存在一定差异，导致监管诉求与方式不一，制约了协同治理合力的形成。**二是跨境风险防范机制薄弱。**人工智能生成内容的跨境传播缺乏国际公认的标识与认证标准，溯源能力不足增加了虚假信息等风险的治理难度。**三是跨境执法协作存在障碍。**针对人工智能滥用行为的跨境调查、证据固定与司法协作缺乏标准化流程，影响了实际治理效能。

为此，亟需从多方面建立和完善相应机制，统筹人工智能的发展和治理，构建更加安全有序的智能社会。一是推动形成具备国际共识性的风险治理框架。通过多边协商机制，促进各国在人工智能风险分类分级标准上的建立完善与协调互认，逐步建立具备国际共识性的人工智能风险分级分类框架，为精准协同的跨境风险识别与监管奠定基础。**二是建立跨境内容治理协同体系。**依托现有相关国际标准组织，推动数字水印、内容认证等溯源技术的国际标准协调，逐步构建形成全球性的人工智能内容溯源与认证网络，建立全球共享的溯源数据库，提升人工智能生成内容的跨境可识别性和可追溯性。鼓励多利益攸关方协同参与相关机制设计与实施，共同提升对虚假信息的识别、追踪

91. 深度伪造攻击每五分钟发生一次，来源：https://www.helpnetsecurity.com/2024/11/22/ai-assisted-fraud-rise/
92. 美国亚马逊公司开发的简历筛选的人工智能程序被发现该系统存在“性别歧视”，来源：https://paper.people.com.cn/zgcsb/html/2023-09/18/content_26017767.htm
93. 人工智能生成内容正在通过多种途径对选举过程构成威胁，亟需多管齐下予以应对，来源：https://www.iiss.pku.edu.cn/__local/3/7F/3D/9EA4A3B30950C7A22DAC5E56962_861E8C97_60B97.pdf
94. 人工智能中的算法经常被用于用户画像，通过对海量信息的追踪、归纳和分析，勾勒出用户的私密信息，进而实施“大数据杀熟”和“歧视性定价”等商业滥用行为，来源：https://www.zhonglun.com/research/articles/8670.html

与应对能力。对此，中国于2025年9月实施的《人工智能生成合成内容标识办法》[95]及配套强制性国家标准[96]，为规范人工智能生成合成内容治理提供了参照借鉴，为建立跨境内容治理协同体系提供了实践基础。**三是完善跨境执法协作机制。**在国际组织框架下，构建人工智能安全威胁信息共享网络，推动跨境调查、电子证据固定等程序标准化，逐步建立面向人工智能技术滥用恶用的监测与协同处置机制，提升全球范围内对跨境违法行为的响应与处置能力。

（三）前瞻研判与防范人工智能的全球性重大风险

除了快速迭代的技术风险和广泛应用及滥用恶用风险外，人工智能还可能带来全球性的、系统性的风险，包括对战略安全、国际和平甚至人类生存的潜在威胁。这类风险具有跨国性、系统性和战略性特征，其影响范围可能覆盖能源、交通、金融、通信等全球关键基础设施，同时不局限于单一国家或局部应用，而是可能跨越国家边界、产业体系及社会结构[97,98]，对全球安全和人类福祉形成潜在重大威胁。**同时，随着技术水平不断向通用人工智能及超级智能迈进，其可能触发的系统性风险进一步凸显。**具备跨领域整合资源、执行复杂任务能力的人工智能，将使风险从局部扩展到全球。例如，在民用场景可能引发能源供应中断、交通体系瘫痪、全球金融系统失灵等跨国性公共危机；在军事和安全领域，高度自主的致命性武器系统失控[99]可能引发战略防御误判、局部冲突快速升级[100]，甚至触及核安全。此外，全球性的人工智能能力集中及技术流动特性，使得任何单一国家或地区的监管短板都可能演变为国际性系统性风险，触发“木桶效应”，进一步强化了国际社会协同治理和联合防控的必要性。

在全球机制层面，已有一些前瞻性努力来应对上述风险。例如，在人工智能军控和战略安全方面，各国在联合国《特定常规武器公约》（The Convention on Certain Conventional Weapons，CCW）框架下持续讨论“致命性自主武器系统”（Lethal Autonomous Weapons Systems，LAWS）问题，并普遍认为“人的控制对于确保责任和问责、遵守国际法以及合乎伦理的决策至关重要”[101]；军事领域负责任使用人工智能峰会发布报告[102]，将“授权使用核武器的决定应由人类控制”作为其核心建议；与此同时，中美领导人就“保持人类对使用核武器决定的控制”达成共识[103]；《国际人工智能安全报告》[104]则推动国际科学界共同评估通

95. 中国《人工智能生成合成内容标识办法》自 2025 年 9 月 1 日起施行，来源：https://www.gov.cn/zhengce/zhengceku/202503/content_7014286.htm
96. 国家市场监督管理总局、国家标准化管理委员会发布《网络安全技术人工智能生成合成内容标识方法》，来源：https://openstd.samr.gov.cn/bzgk/std/newGbInfo?hcno=F32EA2A561F1886CD8D606513512D547
97. 高盛报告，预测人工智能可能影响 3 亿个全职工作，来源：https://www.goldmansachs.com/insights/articles/generative-ai-could-raise-global-gdp-by-7-percent
98. 因 CrowdStrike 更新导致的大规模 IT 中断，来源：https://www.cisa.gov/news-events/alerts/2024/07/19/widespread-it-outage-due-crowdstrike-update
99. 致命性自主武器系统对国际安全和人道主义构成重大挑战，来源：https://www.siis.org.cn/updates/cms/old/UploadFiles/file/20200307/202002006%20%E9%BE%99%20%20%E5%9D%A4.pdf
100. 人工智能可能引发“第三次战争革命”，来源：https://safe.ai/ai-risk
101. 联合国，《致命自主武器系统：秘书长的报告》，来源：https://docs.un.org/en/A/79/88
102. 军事领域负责任使用人工智能全球委员会（GC REAIM）发布的报告提出了五项核心建议，其中包括：“在具有法律约束力的层面上达成共识，即授权使用核武器的决定应始终由人类控制”，来源：https://hcss.nl/wp-content/uploads/2025/09/GC-REAIM-Strategic-Guidance-Report-Final-WEB.pdf
103. 两国元首确认应维持由人类控制核武器使用的决定，来源：https://www.gov.cn/lianbo/bumen/202411/content_6987686.htm
104. Y. Bengio 等，《国际人工智能安全报告》（DSIT 2025/001），来源 : https://www.gov.uk/government/publications/international-ai-safety-report-2025

用人工智能及相关前沿技术能力与潜在风险。在跨领域安全测试方面，联合实验和标准工具的开发已在包括人工智能与化学、生物、放射性及核（Chemical, Biological, Radiological, and Nuclear，CBRN）在内的风险管理中初步应用，如建立共享评估平台、模拟联合测试[105]和跨境能力验证。

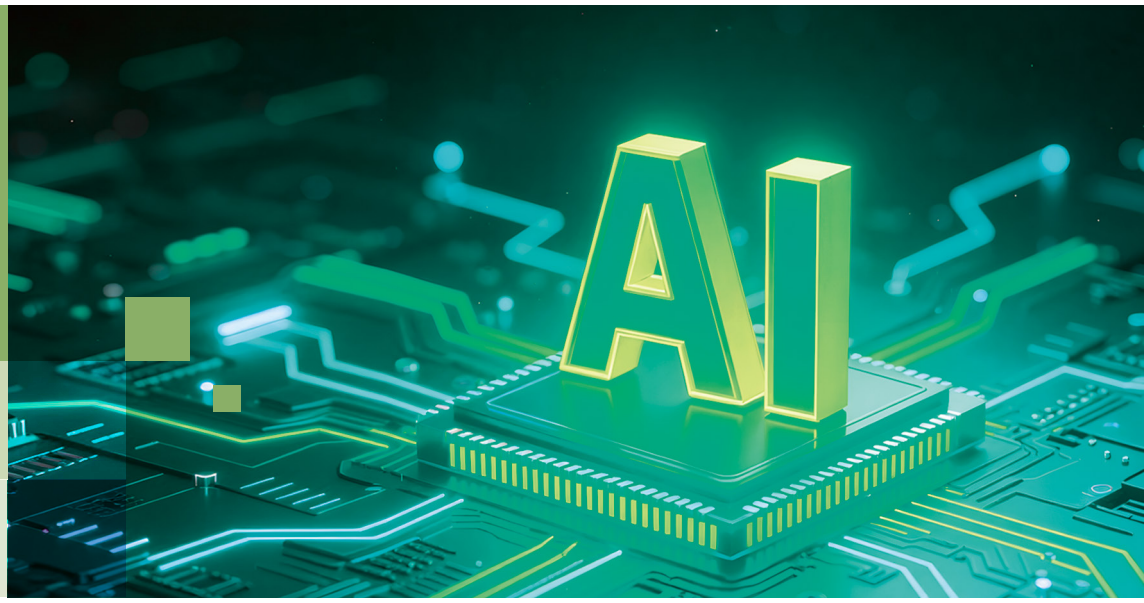
尽管如此，当前全球机制在应对潜在重大风险时仍显不足。一是前沿风险的国际联合评估参与度有限。目前覆盖 CBRN 等跨国性安全敏感场景的统一标准和联合测试流程仅在部分国家间进行探索。**二是对重大风险来源的制度约束不充分。**技术流动和监管差异可能导致高风险技术跨境滥用，全球性跨区域风险尚难以有效管控。**三是紧急干预与联合响应能力有限，**缺少可快速启动的国际机制来中止危险研发或阻断潜在滥用部署。

对此，应构建具备前瞻性、系统性、协同性和敏捷性的应对机制，进一步提升对重大全球性人工智能风险的国际防控能力。一是形成协同应对人工智能技术失控风险的共识。推广涵盖技术、伦理、管理多维度的可信人工智能基本准则，对核生化导等场景下应用人工智能技术提出相关要求，加强人工智能技术最终用途追溯管理，促进国际社会形成共识，防止人工智能技术被误用、滥用，严防威胁人类生存发展的失控风险，确保人工智能技术演进安全、可靠、可控，始终处于人类控制之下。**二是建立全球广泛参与的前沿人工智能模型监测与能力评估机制。**重点覆盖高风险应用和技术领域，通过跨国联合测试平台、标准化风险指标、滥用监测及定期能力评估，确保潜在重大风险能够广泛识别和通报。**三是构建面向关键资源的国际治理体系。**围绕算力、数据、开源算法等高敏感领域，推动形成具备透明度的国际规则与协同治理框架，从源头降低技术滥用风险扩散。**四是设立紧急干预与联合响应机制。**在发现潜在灾难性或战略性风险时，能够快速启动跨国协同措施，如中止危险研发、限制资金流和技术扩散、阻断危险部署等，维护对关键系统的控制权，确保技术应用始终服务于人类福祉。

105. 国际人工智能安全研究所网络开展针对基于大型语言模型的自主代理系统的联合测试演习，来源：https://www.aisi.gov.uk/blog/international-joint-testing-exercise-agentic-testing

— PART —

03



确保包容：兼顾各国人工智能发展与治理诉求

人工智能为全球发展带来前所未有的机遇的同时，也凸显出各国在技术能力、资源禀赋与制度准备上的不平衡。这些差异不仅导致发展诉求各异，也增加了全球人工智能治理的协调难度。为实现包容性治理，国际社会需在尊重多样性与促进共同利益之间寻求平衡。本章围绕“确保包容”展开：首先分析全球人工智能发展的结构性鸿沟及由此产生的多元诉求；其次探讨各国平等发展和利用人工智能权利的保障路径；最后提出在全球治理机制中实现代表性与包容性的制度安排，形成公平、包容与可持续性的治理体系与合作框架。

（一）正视全球人工智能发展鸿沟与各国诉求差异

全球人工智能发展水平呈现出显著的结构性不平衡，资源、能力与制度优势高度集中于少数发达经济体和大型科技企业。《2025 年人工智能指数报告》[106] 显示，2024 年美国私营部门人工智能投资约为 109.1 亿美元，是中国（9.3 亿美元）的近 12 倍；而《2025 年技术与创新报告》[107] 指出，全球企业研发支出的约 40% 集中于 100 家大型科技公司，多数总部位于美国、欧盟或日本。这种资本与技术的高度集中，使得少数国家在算法创新、标准制定与产业生态中占据优势和主导地位。

制度与治理能力上的差距同样明显。国际货币基金组织发布的《人工智能准备度指数》[108] 显示，高收入国家人工智能准备度显著优于新兴市场国家和低收入国家。《政府人工智能准备度指数 2024》报告 [109] 进一步指出，发达国家在“技

106. 斯坦福大学以人为本人工智能研究所（Stanford HAI），《2025 年人工智能指数报告》，来源：https://hai.stanford.edu/ai-index/2025-ai-index-report
107. 联合国贸易和发展会议（UNCTAD），《2025 年技术与创新报告》，来源：https://unctad.org/system/files/official-document/tir2025_en.pdf
108. 国际货币基金组织（IMF），《人工智能准备度指数》（AI Preparedness Index，AIPI），来源：https://www.imf.org/en/Blogs/Articles/2024/06/25/mapping-the-worlds-readiness-for-artificial-intelligence-shows-prospects-diverge
109. Oxford Insights，《政府人工智能准备度指数 2024》，来源： https://oxfordinsights.com/ai-readiness/ai-readiness-index/

术部门支柱”维度优势显著，而中低收入国家在“政府治理”与“数据基础设施”方面普遍滞后。世界银行报告亦指出，人才缺口、信息与通信技术基础设施不足及核心技术依赖构成发展中国家人工智能普及与治理的主要瓶颈[110]。

发展鸿沟进一步带来了治理诉求上的显著差异。技术领先国家更关注前沿突破、竞争优势与安全风险管控；而发展中国家则侧重于技术普及、能力建设与本土化应用，期望通过人工智能促进经济增长、改善公共服务并弥合数字鸿沟。不同的结构性条件塑造了国家在人工智能政策目标、风险认知及国际规则谈判中的多元立场。**同时，这种差异进一步加剧了全球治理的复杂性与协调难度：一是国际共识难以凝聚。**各国在技术发展水平、经济利益和风险认知上存在差异，导致在关键治理议题上立场分化，协调成本高昂。**二是统一规则框架推进受阻。**由于对安全底线、监管原则和伦理标准的理解不同，跨国规则互认与机制协同缺乏基础，制约了整体治理效能。**三是治理参与存在明显失衡。**技术领先国家在标准制定与议程设置中占据主导，而众多发展中国家有效参与不足，这不仅影响治理的公平性和代表性，也可能加剧未来治理体系的分化与冲突。

要实现全球治理的普惠性与包容性，必须兼顾不同国家的发展阶段与利益诉求。为此，国际社会可重点从以下三个维度着力：**一是构建开放包容的多边协商平台，**确保不同发展阶段和制度背景的国家能够平等参与治理议程设定、规则制定与技术标准协商，增强全球治理的合法性与包容性。**二是建立务实有效的国际合作机制，**在尊重各国主权与“人工智能治理权”的前提下，通过技术援助、经验共享与资源协调，支持发展中国家提升人工智能治理与创新能力，推动技术普惠、基础设施共建与人才合作，缩小智能鸿沟，持续推动构建基于共识的国际协作框架。**三是推行公平透明的决策程序，**在治理机制中嵌入公开、可追溯、可问责的运行规则，最大限度减少信息不对称，增强国际互信，为持续应对全球性风险奠定制度基础。

（二）保障各国平等发展和利用人工智能的权利

包容、公正的全球人工智能安全与治理体系应尊重各国发展模式与文化背景的多样性，促进协同共治与技术普惠，保障各国平等地发展和使用人工智能的权利。然而现实中，**全球数字发展权仍存在深层不平等，“智能鸿沟”持续扩大，制约了各国在人工智能领域的平等参与及共享发展。**

其根源主要体现在三个方面。首先，技术领先国家在核心资源领域的结构性优势不断巩固。技术领先国家在算力、算法与数据等基础资源方面的优势持续强化，高端芯片及高性能计算基础设施的国际流通面临诸多限制。例如，全球北方国家拥有世界上最强大超级计算机数量的 75%，而整个非洲大陆拥有的这些高性能系统不到 1%，稀缺性使得非洲国家在发展人工智能时面临显著的成本压力——相对于其人均经济水平，非洲国家在获取图形处理单元等人工智能核心算力设备的成本往往是发达国家的

110. 世界银行，《跨越人工智能鸿沟》，来源：https://openknowledge.worldbank.org/bitstreams/82fbc048-b723-4818-bb91-9a4c8855daf1/download

10 到 30 倍[111]。**其次，国际社会尚缺乏系统化的人工智能发展援助与资源协调机制。**在现有全球创新合作框架中，尚未形成专门面向人工智能的长期性、制度化支持渠道，发展中国家往往难以突破高昂研发投入与基础设施建设成本的约束，技术自给能力有限。**最后，能力建设合作仍呈碎片化与短期化特征，这也不利于发展的可持续性。**当前相关项目多以阶段性援助为主，缺乏持续的制度支持与知识传递机制，导致欠发达国家和地区在人才培养、制度建设及风险治理能力上长期滞后于技术迭代速度。

人工智能发展的全球失衡态势呼唤构建更具包容性与协调性的国际治理体系。具体而言，**一是建立可控、安全的技术共享与转让多边机制。**通过促进算法、算力、数据等关键要素有序流动，支持技术后发国家实现能力跃升；**二是创新融资模式。**依托世界银行、地区开发银行等多边机构，设立人工智能专项发展基金，通过发行人工智能可持续发展债券等形式，将债务偿还与技术能力建设成果挂钩，对达到预设发展指标的国家给予债务减免，以及通过建立公私合作伙伴关系（Public-Private Partnership，PPP）融资池等机制吸引私营部门参与发展中国家算力中心建设[112]。**三是健全系统化、长效化的能力建设合作机制。**围绕人才培养、监管框架、标准对接与风险治理等领域构建国际支持网络，从而助力各国提升本土治理与适应性发展能力。

（三）确保全球安全与治理机制的代表性与包容性

公正可信的全球人工智能安全与治理参与机制应保障各国在决策过程中的代表性与包容性。代表性要求治理架构充分吸纳处于不同发展阶段和拥有多元利益诉求的国家，确保其在国际议程设置、规则协商与标准形成等关键环节拥有实质性参与空间；包容性则强调通过平等对话与广泛磋商，使治理体系在尊重价值观和发展路径多样性的前提下实现有效协调。

当前全球人工智能治理领域的规则制定与话语权仍高度集中于少数国家。根据统计[113]，在联合国 193 个会员国中，仅有 7 个国家全部参与了近年来提出的七项主要人工智能治理倡议，118 个会员国完全缺席，主要为“全球南方”国家。《全球人工智能安全指数 2025》[114]对 40 个国家的统计同样表明，在近年来 5 项国际人工智能安全重要宣言倡议中，仅有 6 个国家签署了全部，显示出安全与治理机制在全球范围内参与度和代表性上的严重不足。

造成这种局面的原因主要包括三点：其一，国际层面缺乏具有约束力的信用与承诺执行机制，部分大国主导甚至退出关键议程，损害了多边规则的整体效力与权威性。**其二，**由于尚未达成全球共同遵循的安全与伦理底线红线，各国监管门槛存在显著差异，发达国家能够通过标准先行和评估体系垄断话语权，无形中抬

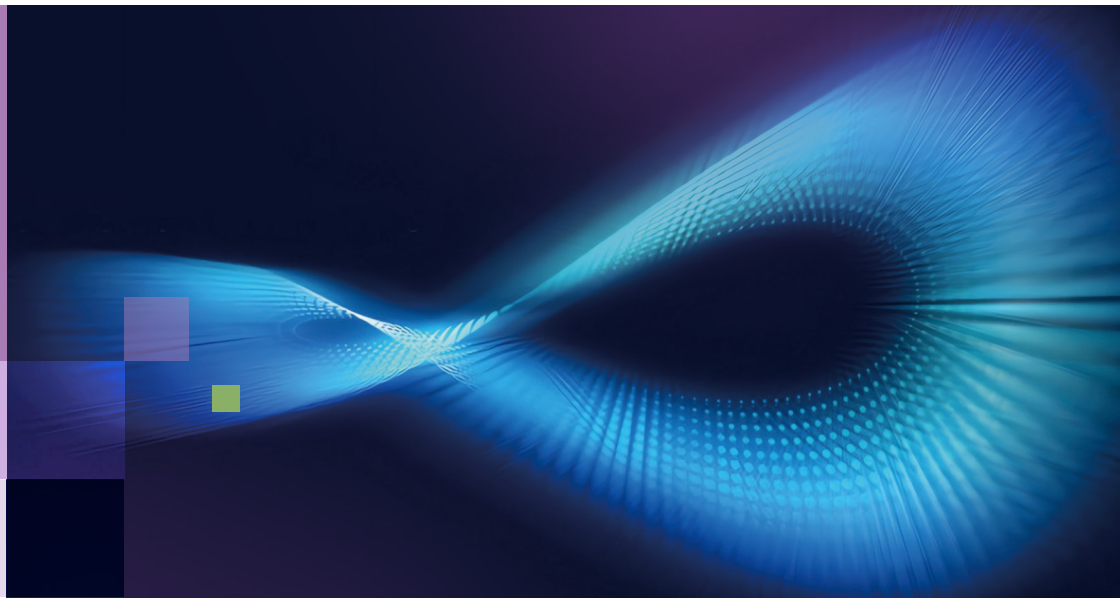
111. 非洲国家在获取图形处理单元等人工智能核心算力设备的成本高昂，来源：https://www.aihubfordevelopment.org/green-compute-coalition
112. 联合国，《人工智能能力建设的创新自愿筹资备选方案》，来源：https://digitallibrary.un.org/record/4085951?ln=en&v=pdf#files
113. 联合国，《治理人工智能，助力造福人类》最终报告，来源：https://www.un.org/sites/un2.un.org/files/governing_ai_for_humanity_final_report_zh.pdf
114. 《全球人工智能安全指数 2025》，来源：https://agile-index.ai/global-index-for-ai-safety

高了后发国家的合规与参与成本。**其三，**治理架构中未能体现差异化责任原则，发展中国家常被动接受高合规负担，却难以影响实质决策，导致其诉求无法有效纳入国际规则。

为构建更加公正、包容且具备约束力的全球人工智能治理体系，应着力推动以下机制建设：一是建立国际治理参与动态评估，对各国参与情况进行持续评估，并将评估结果与其国际声誉挂钩，尽可能抑制随意退出行为。**二是推动建立全球人工智能治理决策流程合规性听证机制，**围绕代表性、透明度、程序正当与协商一致等核心原则，对重大决策的形成过程进行多边审议与监督，防止因程序不公导致的合法性缺失与治理信任危机。**三是建立差异化的责任分担机制，**根据各国发展水平、技术能力与治理基础，分类设定相应义务。在确保技术领先的开发主体切实承担前沿模型风险研判与治理责任的同时，保障受影响的各方有效行使监督权利，从而推动全球人工智能治理的公平性、有效性与可持续性。

— PART —

04



明晰权责：推动多利益攸关方协调有力的行动

本章分析了人工智能治理中多利益攸关方的复杂互动与挑战，指出各方在角色定位、资源禀赋与价值取向上的差异与互补关系，强调单一主体主导规则制定的局限性及其对治理合法性、有效性与可持续性的潜在影响；进一步明确国家政府、国际组织、企业、科研机构以及社会公众等核心主体的差异化定位与责任边界，并提出以履约审查、风险监测、争端调解和标准互认为核心的系统化协同机制，以化解责任模糊、规则碎片化与执行难题，推动形成各方权责明晰、行动协调有力的全球人工智能安全与治理体系。

（一） 人工智能多利益攸关方的复杂互动与挑战

人工智能技术自身的开放性及其跨领域应用的特性，决定了多利益攸关方的广泛参与是其治理的必要条件。各国政府、国际组织、企业、科研机构和社会公众在职责专长、资源禀赋与价值取向方面既各具差异，又彼此互补。例如，政府掌握政策制定与监管资源，企业拥有创新与技术能力，科研机构提供科学评估与知识支撑，国际组织促进协作与标准衔接，而社会公众则在价值导向与社会监督中发挥关键作用。若治理框架仅依赖单一主体或单一维度的规则设计，不仅可能削弱治理措施的合法性与信任基础，也会限制其应对复杂风险与实现长期可持续发展的能力。

多利益攸关方的复杂互动显著提高了治理协调的难度。不同主体在角色、资源与目标优先级上的差异，使其既相互依赖又彼此制衡：国家在政策制定中需要企业的技术支撑与科研机构的知识供给；国际组织推动议程落实依赖成员国的政治承诺与社会组织的价值倡导；企业的发展需要良好的外部政策环境与公众信任基础；科研创新既得益于企业投入，也受社会监督制约。

在全球人工智能安全与治理体系中，缺乏有效的多利益攸关方参与和协调，可能引发多

方面风险。在宏观层面，国家间对话语权的竞争可能加剧技术生态与治理规则碎片化、推高技术壁垒，甚至催生基于价值观或安全诉求的治理机制分化趋势。**在具体层面，一是**若全球安全与治理议程的代表性不足，可能会削弱多边机制的中立性与公信力；**二是**公共政策制定易受大型科技企业过度影响，致使监管宽松化甚至被“规制俘获”，公共利益难以得到充分保障；**三是**科研独立性与开放性易受商业化压力挑战，开源生态也可能会受到挤压；**四是**企业逐利动机可能会与公众对公平、问责和权益保障的诉求间产生矛盾，甚至可能加深社会分化。**因此，多利益攸关方在全球人工智能安全与治理体系中的充分参与和协调机制，对于维护治理体系的合法性、有效性与可持续性不可或缺。**

（二） 明晰各利益攸关方的角色定位与责任边界

清晰界定各方角色，明确权责划分，是实现各利益攸关方良性互动与长期协作的基础。在全球人工智能安全与治理体系中，不同主体应在尊重差异、发挥优势的前提下履行相应职责，才能形成协同有序的全球人工智能治理格局。

各国政府应承担各国人工智能治理的主导与规制职责，建立健全国家层面的安全与治理体系，应对人工智能技术带来的主权、安全与伦理挑战。这包括制定和动态更新相关法律法规与政策标准，确立高风险系统的安全测试、评估与认证制度；建立完善跨部门监管协调机制，对人工智能研发、部署和应用的全生命周期进行风险监测与管控；优化算力、数据等关键基础设施布局，引导市场规范有序投入与产业健康发展；在国际层面积极参与全球人工智能安全与治理体系构建，推动形成公平包容、互利共赢的合作机制。

国际组织应发挥全球对话协调与规则对接的关键作用，推动形成一致性更强的国际治理体系。这包括搭建全球性多边对话与谈判平台，牵头制定人工智能安全、伦理与互操作性标准；建立跨国审查与监督机制，监测并评估各国人工智能政策与履约进展；为技术能力较弱的国家提供援助与能力建设支持，缩小数字治理差距；协调碎片化机制，推动各机构在规则制定、风险评估与标准认证等方面形成议程互补与政策合力。

企业应依据自身在产业链中的定位承担相应的伦理、安全与治理责任。对于研发企业，应确保前沿人工智能系统在设计与开发阶段即嵌入内在的安全性、透明性与合规性要求。对于应用企业，应在具体部署和运营阶段开展适配性审查与风险评估，保障系统在公共服务、安全生产等领域的稳健性，防止误用与滥用。同时，所有企业主体均应履行信息披露义务，主动披露人工智能系统的技术特性与潜在影响，接受社会监督与合规审计，确保技术创新与社会责任相协调。

科研机构应承担人工智能知识生产与独立评估的重要职能，为人工智能治理提供科学客观的政策依据和解决方案，弥合技术前沿与政策制定之间的知识鸿沟，确保决策科学有效。具体包括：独立开展前沿人工智能系统的技术安全测试、社会伦理影响评估及长期风险预测研究；为监管机构提供基于科学证据的政策建议、标准制定参考与合规评估工具；探索更加符合伦理和确保安全的技术路径；推动跨学科合作协同与共识形成。

社会公众是人工智能安全与治理的价值守护

者与重要监督力量，应通过发挥价值阐述与外部监督的关键作用，确保人工智能治理过程的透明和包容。具体包括：社会组织可通过独立调研、伦理倡议与公众教育等方式，促进风险认知与价值多元化；公众可通过意见反馈、参与社会讨论及监督政策执行，推动治理议程回应公共利益，防范社会不公加剧以及公共利益被侵蚀。

此外在制度层面，多利益攸关方的角色定位和责任边界还需要通过多层次机制来加以确立。对不同的责任主体，应推动差异化制度工具落地：国家需参与多边协商与合规审查，并建立定期责任报告机制；企业应履行强制性透明度义务，接受算法安全审计与数据保护合规评估；科研机构与国际组织则须依托独立伦理审查与技术评估，强化研究过程的规范合规。**对不同的责任范围，**需构建多层次、互补性强的规则协调体系，确立以全球共性规则为基石、区域规范为支撑、行业标准为补充的规则架构，推动国际准则与国内立法、通用要求与特殊场景规则间的互认与衔接，并通过标准转化与动态调整机制降低合规成本，提升治理效能。**对不同的责任程度，**应通过“软硬结合”强化制度约束力：既通过全球性的伦理准则与行业公约形成自愿价值承诺，也借助联合国框架下国际条约、区域协定及监管合作备忘录等建立具有约束力的遵约履约保障体系。

（三） 构建有效的多利益攸关方协同与落实机制

明晰各方角色定位与责任边界是先行基础，而协调有力的落实机制，是确保全球人工智能安全与治理可执行与可持续的关键。目前，国际社会在履约核查、风险监测、争端解决与标准互认等方面仍面临系统性挑战。**一是有效的核查与监督机制缺位，**各方在技术安全、数据治理和算法伦理等领域的承诺难以有效监督落实。**二是风险监测与信息共享机制尚不健全，**导致模型开发、数据使用和算力配置等环节的系统性风险难以及时识别和预警。**三是争端调解机制缺乏权威性与时效性，**多元主体间的利益冲突缺乏高效、权威的调解与仲裁渠道。**四是跨国与跨行业的标准互认进展缓慢，**制约了多方合作的制度衔接与执行效率。

确保多利益攸关方在全球人工智能治理中实现有效协同与责任落地，需依靠系统化、制度化的机制保障，可从以下方面着力。一是建立强有力的履约审查与监督机制，通过对国家及国际组织在规则遵守、政策实施与发展援助等方面的承诺进行定期评估与公开评议，增强在执行落实上的透明度与公信力。**二是建立完善高效的风险监测与信息共享机制，**要求企业与科研机构在模型研发、算力资源配置和数据应用等关键环节实施标准化风险评估并履行信息共享义务，形成覆盖全链条的动态预警网络。**三是设立高效、权威的跨国争端调解与仲裁程序，**为政府、企业、科研机构及公众等多元主体提供具备约束力的纠纷解决渠道，化解因责任模糊或利益分歧导致的协作困境，维护治理进程的稳定与有效。**四是积极推进国际标准互认与规则对接，**推动在安全评估、数据治理、伦理审查等关键领域形成兼容性规范框架，减少制度性壁垒与合规成本，提升全球人工智能安全与治理的协同效率与执行效能。

— PART —

05

面向未来：践行多边主义共建人类命运共同体

本章分析了地缘政治博弈、技术封锁等对全球合作造成的冲击，指出“以人为本、智能向善”作为全球人工智能治理的基础价值共识，强调应以联合国为中心，协调对接与优化现有安全与治理资源、推动广泛跨机构合作、提升执行效能以实现权威高效的全球人工智能安全与治理体系。

（一） 凝聚与落实基于人类共同福祉的全球共识

基于地缘政治考量的短视行为，正在削弱各国在人工智能治理和全球合作中长期积累的共同努力。近年来，一些国家将人工智能技术视为战略竞争的关键资源，采取出口管制、技术封锁、供应链脱钩等措施，将自身安全与利益凌驾于全球共同安全与利益之上。这种以防范和博弈为导向的政策取向，正在割裂全球长期协作形成的国际科技创新网络和风险应对体系。同时，全球现有人工智能治理机制呈现出日益显著的碎片化趋势，以特定伙伴关系或区域性机制等“小圈子”为基础制定排他性准入标准和治理规则的现象日益增多，双边安排和集团化趋势等在一定程度上削弱了国际规则的协调性与开放性，也侵蚀了多边合作的广泛性和包容性。在此背景下，国家间的互信水平面临挑战，算法透明、数据跨境流动、算力资源互享等议题被高度敏感化、政治化，信息沟通渠道受限，制约了跨国风险评估与危机应对能力的有效性。

尽管各国在制度模式和利益诉求上存在差异，但需要看到，“以人为本、智能向善”的理念，始终是各方推动人工智能安全与治理的广泛共识基础。首先，这一共识具有高度的内生性：无论技术发展路径如何不同，保障人的尊严、促进社会福祉、确保技术安全，始终是多数国家政策合法性与维护社会契约的核心诉求。即使在大国战略竞争的背景下，技术创新最终也需服务于公共利益和社会进步，这是各方普遍认可的共同目标。**其次，**技术风险的跨境性与扩散性，使得各国在算法透明、隐私保

护、安全部署等问题上存在天然共鸣，没有任何单一国家能够独自应对全球性挑战，合作仍是最具理性的选择。**最后，**多边机制的长期积累为维护价值共识提供了制度支撑。联合国《全球数字契约》、联合国教科文组织《人工智能伦理问题建议书》等倡议在理念与规范上已形成广泛影响，为未来的规则衔接、能力建设与治理协同奠定了基础。

落实“以人为本、智能向善”的共识，需要建立兼顾科学性与包容性的工作机制，推动形成“科学研判 + 分级对话”的务实路径。建议在联合国框架下，由人工智能问题独立国际科学小组发挥指导协调作用，推动构建面向公众的人工智能科学传播全球公共服务体系。独立国际科学小组应在为联合国系统及全球政策制定者提供专业咨询的基础上，进一步承担起全球人工智能科学传播工作的指导职责，整合联合国系统内外的相关资源，发挥自身专业性、权威性和联合国多边平台优势，逐步形成覆盖广泛、机制完善、可持续发展的人工智能公众科学传播体系。该体系应依托多语言、多媒介和跨文化的传播策略，支持成员国，特别是发展中国家的青年、教育工作者及基层社区组织，提高其对人工智能技术发展与治理的理解与参与能力，增强其在全球人工智能安全与治理多利益攸关方框架中的表达与参与。通过加强科学普及与公共参与，进一步夯实人工智能治理的社会公众信任基础，促进构建更加包容、公平和可持续的全球治理体系。

与此同时，有必要在联合国现有人工智能对话机制中逐步引入分级分类理念，使对话形式与参与层级能够根据风险特征与影响程度实现精准匹配。在人工智能问题独立国际科学小组的支持下，可探索建立前沿风险识别、分级与预警机制，并据此设置分层分类对话安排：对常规性或可控性风险，可通过专家层面和技术层面开展定期交流；对于涉及特定领域风险的议题，还应充分吸纳领域内相关国际与区域性专业组织机构共同参与；对具有高敏感性或潜在系统性影响的重大风险，可适时提升磋商规格，启动高级别政策对话机制，邀请国家代表乃至政府高层参与。此举有助于将重大风险议题纳入国家战略视野、融入科学专业视角，促进政策协调与资源整合，同时提升公众认知并激发跨层级响应能力，从而增强全球人工智能治理体系的协同效能与危机应对能力。

（二）共建以联合国为核心的全球安全与治理体系

联合国作为最具包容性的全球治理平台，在当下全球人工智能安全与治理体系构建中发挥着不可替代的作用。作为当今世界最具普遍性、权威性和代表性的政府间国际组织，联合国拥有近 200 个成员国政府参与，具备协调多元利益、平衡多方关切的天然优势，其合法性建立在《联合国宪章》所确立的共同原则和价值基础之上。随着人工智能技术加速发展、风险跨国外溢，以及全球治理碎片化与制度性摩擦加剧，联合国的协调与引领作用愈发凸显。一方面，联合国体系内制度资源丰富，通过联合国大会、联合国秘书长的“全球数字契约”进程，以及联合国教科文组织、国际电信联盟等机构，为凝聚价值共识、统一技术规范、加强风险治理、开展能力建设合作等提供了多维支持。另一方面，联合国的多边包容性使其能够在发达国家与发展中国家之间发挥桥梁作

*非详尽无遗，仅做展示说明

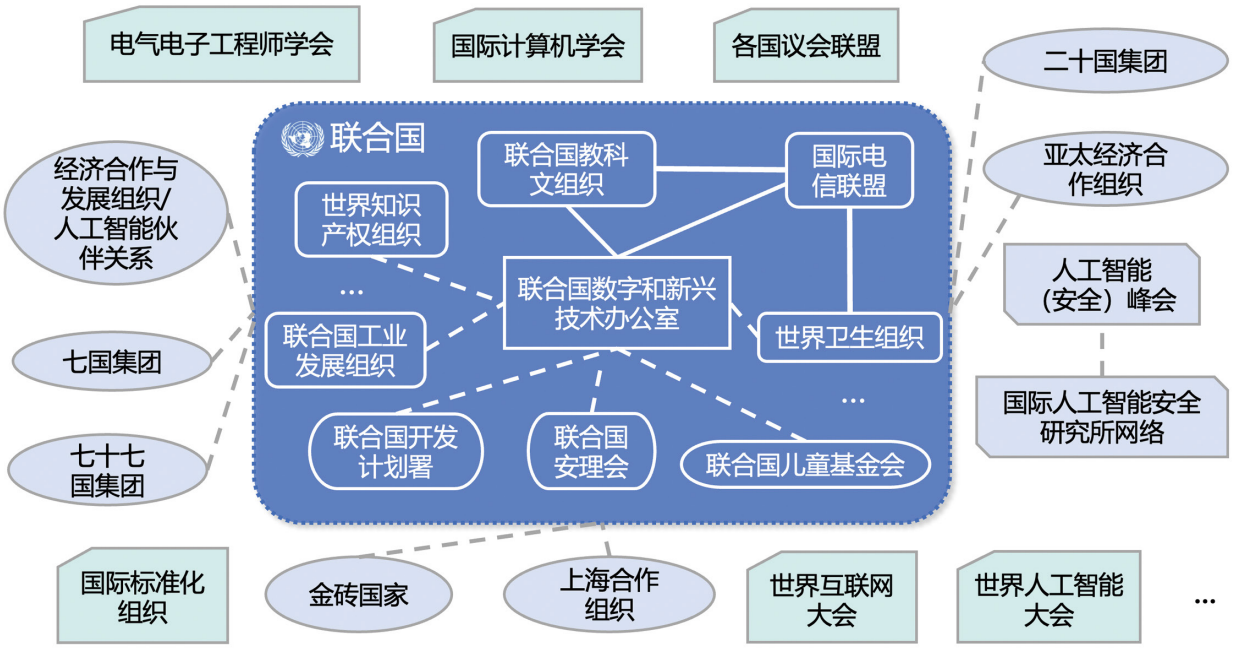


图 3 联合国系统及其与人工智能治理相关国际组织与机构间的关联示意图

用，推动形成兼顾安全、发展与公平的全球规则体系。

为应对日益复杂的治理挑战，有必要进一步整合与优化联合国系统内人工智能相关资源，推动各方持续深化共识与协同。目前，联合国系统内与人工智能安全与治理相关的职能分布于多个机构和议程，其协调机制和协同效能仍有加强空间。现有机制各有侧重：联合国教科文组织推动人工智能伦理规范制定，国际电信联盟聚焦技术标准和基础设施互联互通，人权理事会关注隐私与数字权利保护，秘书长技术特使办公室则推动“全球数字契约”进程。这些机制在各自领域发挥了重要作用，但在整体衔接与政策协同上仍需进一步提升，以充分发挥系统合力。为此，可考虑在联合国框架下建立跨机构的人工智能治理协调机制，推动不同机构间议程对接与政策整合，并以“全球数字契约”为核心抓手，将价值共识、技术标准、风险防控与能力建设等关键议题纳入系统化议程。

面向未来，国际社会可在联合国框架下进一步探索建设具备更强协调性与执行力的全球人工智能安全与治理平台。该平台可在联合国大会授权下运行，确保其合法性与代表性，并依托联合国数字和新兴技术办公室及相关机构形成跨机构的议程协调体系[115]，同时参考其他全球治理机制的成功经验——例如气候变化领域的政府间专门委员会、国际民用航空组织的规范体系、核安全公约的同行审议机制，以及国际卫生条例的应急响应框架等——在技术规

115. 参考联合国政府间气候变化专门委员会，来源：https://www.ipcc.ch/

范[116]、风险治理[117]、伦理审查[118]、能力建设[119]等多方面推动跨系统整合。该平台应具备协调、审议、标准规范制定及必要的执行与监督权能，并可通过建立国际公约等形式，逐步确立成员国履约义务[120]，完善跨国风险信息共享[121]与应急响应机制[122]，推动技术标准的国际互认与落地[123]，从而提升全球人工智能治理的整体协调力与执行力，最终成为统筹全球安全与治理资源、平衡多方利益、推动有效落实的中枢力量。

在技术快速演进与全球秩序深度变革的时代交汇点上，人工智能安全与治理关乎全人类的共同未来。国际社会应以联合国为核心，持续凝聚共识、汇聚资源，坚持主权平等、国际法治和多边主义原则，秉持以人为本、包容合作、注重实效的价值导向，共同构建兼顾安全、发展、公平与包容的全球治理新体系，在机遇与挑战并存的时代，共同开创普惠、共享、可持续的智能化未来。

116. 参考《芝加哥条约》，为国际民用航空活动制定基本法律框架和原则，并设立了国际民用航空组织（ICAO），来源：https://www.icao.int/convention-international-civil-aviation-doc-7300
117. 参考《核安全公约》；确保民用核电厂的安全运行，通过同行审议促进核安全持续改进，来源：https://www.iaea.org/topics/nuclear-safety-conventions/convention-nuclear-safety
118. 参考《烟草控制框架公约》，首个由世界卫生组织主持谈判达成的全球公共卫生条约，旨在通过减少烟草消费和供应来应对全球烟草流行，来源：https://wkc.who.int/resources/publications/i/item/9241591013
119. 参考联合国，《加强人工智能能力建设方面的国际合作》，来源：https://docs.un.org/zh/A/RES/78/311
120. 参考《联合国气候变化框架公约》，全球应对气候变化的基础性的法律框架，来源：https://www.un.org/climatesecuritymechanism/en/united-nations-framework-convention-climate-change-unfccc-and-climate-peace-and-security
121. 参考《核事故早期通报公约》，要求缔约国在发生可能对其他国家产生放射学安全影响的核事故时立即通报，来源：https://www.iaea.org/topics/nuclear-safety-conventions/convention-early-notification-nuclear-accident
122. 参考《国际卫生条例》，一项具有法律约束力的全球公共卫生安全框架，旨在帮助各国预防和应对可能构成国际关注的突发公共卫生事件，来源：https://apps.who.int/gb/ebwha/pdf_files/WHA77/A77_ACONF14-en.pdf
123. 参考《国际海上人命安全公约》，国际海事组织（IMO）管理下，旨在为商船的安全建造、设备和操作制定统一标准的国际条约，来源：https://www.imo.org/en/knowledgecentre/conferencesmeetings/pages/solas.aspx

附录 全球人工智能安全与治理体系（建议列表）

维度	面临的关键问题	建议的机制举措	期望实现的目标	可借鉴的国际经验
应对人工智能快速变革与重大风险	技术突破方向不确定，风险前瞻不足，缺乏相应指标与能力	**技术跟踪与风险预警协同：**形成跨国技术监测网络，推动各方信息共享，常态化、动态化评估技术进展	精准捕捉技术突破方向，提前识别潜在风险，筑牢全球人工智能安全发展的前置防线	IPCC 的跨学科数据共享平台与预测机制；IAEA 的核查与通报机制（提前发现风险）
	治理规则制定与更新速度慢，各国各地区监管措施不同步	**治理规则的动态更新与互认机制：**完善人工智能安全标准的常态化、定期审议与更新程序，把握规则动态修正与互操作	消除各国监管差异带来的治理缝隙，夯实全球协同治理的基础	IPCC 的科学评估与政策建议衔接模式（科学支撑政策，动态更新）
	缺乏系统化风险评估与干预工具，模型安全监测未成熟，监管效能受到制约	**安全评测共性工具与平台生态：**共同开发安全测试平台、基准数据集与风险评估工具库，建设全球共享技术工具箱	破解监管效能不足的困境，为各国监管提供统一技术支撑	WHO《国际卫生条例》（IHR）的疫情信息上报与跨区域追踪机制（多源数据 + 快速反馈）
	各国风险认知共识不足，标准存在差异，监管诉求与方式不同	**国际共识性风险治理框架：**通过多边协商，完善和协调各国风险分类标准	建立国际共认的风险分级分类体系，提升跨境风险识别能力	IAEA 制定国际通行的核事件分类体系 INES
	跨境风险防范机制薄弱，溯源困难	**跨境内容治理协同：**依托现有相关国际标准组织，推动不同认证技术的国际协同	构建全球人工智能生成内容溯源与认证网络和共享的数据库，建立跨境内容治理协同体系	ISO/IEC 的国际标准体系推动内容认证与互认标准
	跨境执法协作困难，缺乏标准化协作流程	**跨境执法协作机制：**加强跨境执法在信息共享，跨境合作，调查取证方面的标准化	建立人工智能技术滥用恶用的监测与协同处置机制，提升跨境违法行为的响应与处置能力	《布达佩斯网络犯罪公约》涉及的相关跨境执法与证据合作机制
	前沿风险国际联合评估参与度有限，部分标准和流程仅在少数国家中探索	**国际合作，多边对话等机制：**形成应对人工智能技术失控风险的共识。	防止技术误用、滥用，严防威胁人类生存发展的失控风险，确保人工智能技术演进安全、可靠、可控	《禁止生物武器公约》《禁止化学武器公约》中多边合作交流以及透明性审查机制
	高风险前沿模型能力不透明，可解释性不足，出现滥用后难以及时发现	**前沿模型监测与评估：**开展对高风险模型的能力评估、滥用监测与跨境通报	实现潜在威胁识别与预警，提升透明度与跨境信任	IAEA 核查与通报机制，要求对核设施定期检查评估；《核事故早期通报公约》中要求的危急通报义务
	算力、数据、算法开源与芯片制造等关键资源分布不均，存在滥用与垄断风险	**关键资源治理：**推动国际共识，对算力、数据、算法与芯片等高敏感资源进行协调管理	从源头降低技术滥用风险扩散，促进资源公平与可控	WTO 规则协调机制（防止垄断与资源壁垒）；ICAO 和 IMO 标准化体系（跨系统互操作与责任追踪）
	极端风险情境下缺乏应急处置手段，单一国家难以应对跨国危机	**紧急干预与联合响应：**快速启动的跨国协同方案，包括中止危险实验、阻断模型部署、冻结资金流动等	在高风险情境下实现快速有效响应，维护对关键系统的控制权，避免全球性灾难性后果	WHO《国际卫生条例》（IHR）的紧急委员会机制（突发状况应急响应）

兼顾各国人工智能发展与治理诉求	国际共识难以凝聚，关键治理议题立场分化，协调成本高昂	**开放包容的多边交流平台：**在治理议程、规则制定和技术标准设定中吸纳多元声音	凝聚全球共识，确保普遍参与，避免排他性治理，增强全球治理的合法性与包容性	《联合国气候变化框架公约》（UNFCCC）气候谈判机制（不同国家平等参与、共同但有区别的责任原则）；WTO 多边谈判框架
	统一规则框架推进受阻，跨国规则互认与机制协同缺乏基础	**务实有效的国际合作：**促进能力建设、知识共享，资源支持，提升发展中国家治理与防控能力	弥合南北差距，实现技术普惠、基础设施改善和人才培养，缩小智能鸿沟	WHO 能力建设援助机制； WTO 技术援助与能力建设计划
	治理参与明显失衡，技术领先国家占据主导	**公平透明的决策程序：**建立公开透明的规则与可问责的协调机制	减少信息不对称，增强国际互信，为持续应对全球性风险奠定制度基础	WTO 争端解决机制（规则明确、过程透明）
	技术能力差距大，发展中经济体缺乏核心资源领域结构性优势	**技术共享与转让：**在可控安全前提下，促进算力、算法、数据和应用的有序跨国流动	帮助技术能力薄弱国家实现跨越式发展，缩小技术鸿沟	WTO《与贸易有关的知识产权协议》（TRIPS）中的技术转让机制
	发展中国家资金投入不足，公共资源建设受约束，技术自给能力受限	**创新融资模式：**依托联合国、世界银行、区域开发银行等多边平台，设立专项发展基金，构建长期稳定的资金支持	为研发、基础设施和公共数据提供持续性资金保障，缓解资金结构性不足	全球环境基金（GEF）； 世界银行数字化转型项目
	能力建设合作以阶段性援助为主，不利于欠发达国家和地区发展的可持续性	**长效化的能力建设合作机制：**在人才培养、监管经验、标准制定、风险防控等方面提供规模化支持	助力各国提升本土治理与适应性发展能力，强化全球整体安全水平	WHO 能力建设援助机制； WTO 技术援助与能力建设计划
	国际层面缺乏信用与承诺执行机制，约束效果差，违规成本低	**治理参与动态评估：**对各国人工智能全球治理的参与和表现进行持续评估，将违规行为与国际声誉建立关联	提高违规退出代价，强化国际规则的约束力	WTO 贸易政策审议机制； IAEA 履约监督机制
	发达国家垄断话语权，抬高后发国家的合规与参与成本	**决策流程听证机制：**维护代表性、透明度、程序正当与协商一致原则，对重大决策的形成过程进行多边审议监督	增强治理决策的合法性、包容性与国际信任	联合国普遍定期审议（UPR）机制；WTO 贸易政策审议机制
	发展中国家被动接受高合规负担，难以影响实质决策	**差异化责任分担机制：**依据发展水平、技术能力与治理基础设定相应义务	在包容性与有效性之间实现平衡，推动责任公平分担	UNFCCC“共同但有区别的责任”原则；WTO 特殊与差别待遇条款
推动多利益攸关方协调有力的行动	国家与国际组织履约不力，承诺难以监督落实	**履约审查与监督：**监督国家与国际组织在规则制定、政策执行、能力建设方面的履约情况	提高治理透明度与责任落实程度，增强执行透明度与公信力	IAEA 核查与履约机制； WTO 贸易政策审议机制
	技术研发与应用环节风险难以及时识别和预警	**风险监测与信息共享：**推动企业与科研机构在关键环节实施标准化风险评估并履行信息共享义务	构建多层次、多环节的动态风险预警体系，形成覆盖全链条的动态预警网络	WHO《国际卫生条例》（IHR）疫情通报机制；ICAO 安全事件与事故报告体系

	多元主体间存在利益冲突，责任模糊导致治理僵局	**争端调解与仲裁：**为政府、企业、科研机构、公众提供高效公正、具备约束力的争端解决渠道	避免治理僵局，保障协同机制正常运转，维护治理进程的稳定与有效	WTO 争端解决机制；《约克 - 安特卫普规则》（YAR）中关于责任承担的处理惯例
	跨国与跨行业标准互认进展缓慢，标准碎片化	**标准互认与对接：**推动算法安全、数据治理、技术评估等标准的跨区域、跨行业对接，形成兼容性规范框架	降低制度摩擦与重复建设成本，提升全球人工智能安全与治理的协同效率与执行效能	ICAO 和 IMO 国际标准体系；ISO 多边相互承认协议
践行多边主义 共建人类命运共同体	公众科学素养不足，治理参与度有限，公共部门政府决策者科学素养有待提升	**科学传播机制：**由联合国人工智能问题独立国际科学小组主导，构建覆盖广泛的全球人工智能科学传播体系：采用多语言、跨媒介策略，重点赋能发展中国家青年、教育工作者及基层组织，提升其对人工智能技术与治理的认知能力及参与权高级别磋商	巩固人工智能治理的民主基础与社会信任，增强公众参与度	UNESCO 全球教育传播项目
	风险对话层级单一，响应效能不足	**分级分类对话融合机制：**对常规风险开展专家层定期交流；对高敏感性重大风险启动高级别磋商（国家代表乃至政府首脑参与），推动风险议题进入更高战略视野并形成协调行动	提升重大风险响应效能，强化联合国多边治理机制的现实支撑力	IPCC 气候科学评估机制中的分级报告体系

世界互联网大会人工智能专业委员会

世界互联网大会于 2024 年设立首个专业性、常态化分支机构——人工智能专业委员会，下设标准推进计划、安全与治理推进计划和产业推进计划。专委会汇集了来自人工智能领域国际组织、知名智库、科研院所、专业协会及产业界的权威专家和专业人才，秉持搭建国际交流合作平台、推动发展与治理协同、促进全球共享人工智能发展成果的原则，通过开展专题研讨、成果分享、倡议发布等活动，不断凝聚国际共识，促进人工智能包容普惠、可持续发展。

为人类共同福祉构建全球人工智能安全与治理体系

如何引用本报告：

曾毅, Seán Ó hÉigeartaigh, 王正奇, 鲁恩萌, 陈煜, 曹功策, 郭晓阳, 康彦荣, 韩开宇, 范津宇, 谢佳玮, 韩正强, 王金, 皇甫存青, 包傲日格乐, Dame Wendy Hall, 林琳, 段伟文, 王融, 张俊林, 唐新华, Vincent C. Müller, 陈晶晶, 李娜, 程凯, Sebastian Sunday Grève, Danil Kerimi, Bernd Holznagel, Anna Abramova, Jimena Sofia Viveros Alvarez, Edson Prestes, 沈俊成, 刘永谋, 张寄冀, Nada Laabidi, 姚新, 周凯, George Chen, 符春辉, 刘晓春, 郭苏敏, 呼娜英, 乔迁, 孟伟, 张凌寒, 谭知行, Helen Meng, 乔宇, 卜语嫣, 范维, 王星光, 王峰, 龚新奇, 夏文辉, 王梦寅, 陶锋, 胡永启, Charuka Senal Damunupola, 程明, 刘威辰, Nirosha Ananda, 彭韬, 陶涛, 汪凤琼, 王健兵, 王彤, 王巍, 王欣, 吴淑燕, 杨耀东, 杨忠良, John Yeoh, 张荣, 周原, 张雪丽, 梁昊. 为人类共同福祉构建全球人工智能安全与治理体系 [R/OL]. 北京：世界互联网大会，2025.

关注大会公众号

© 2025 世界互联网大会


世界互联网大会
World Internet
Conference