

全球人工智能治理研究报告

上海社会科学院

武汉大学

同济大学

中国现代国际关系研究院

中国信息通信研究院

中国社会科学院世界经济与政治研究所

北京邮电大学

北京航空航天大学

北京理工大学

中国人民公安大学

中国政法大学

复旦大学

南京邮电大学

浙江师范大学

伏羲智库

联合发起

前言

人工智能技术是新一轮科技革命和产业变革最具代表性的颠覆性技术，给人类社会带来巨大发展机遇的同时，也带来了诸多全球性挑战。人工智能技术的普及应用已惠及医疗、教育、交通、农业、工业、金融、文化、生态等诸多领域的发展；与此同时，人工智能的误用、滥用和恶用已危及个体、群体、社会、生态以及价值规范等不同层面的安全。2022年生成式人工智能技术取得突破性进展以来，新一轮全球人工智能治理浪潮席卷世界。

在此背景下，上海社会科学院、武汉大学、同济大学、中国现代国际关系研究院、中国信息通信研究院、中国社会科学院世界经济与政治研究所、北京邮电大学、北京航空航天大学、北京理工大学、中国人民公安大学、中国政法大学等中国智库和高校的专家围绕全球人工智能治理的基本现状、主要议题和体系建设等问题展开了研究。本报告中的“全球人工智能治理”，是指国际组织、各国政府、科技公司、非政府组织等行为体为实现人工智能在全球的安全发展与和平利用而共同制定共同实施一系列原则、规范、标准、政策、法律和制度的协作过程。需要指出的是，关于人工智能技术在军事领域的应用问题有独立的国际规制进程，本研究

报告不涉及为军事目的的开发或使用人工智能的全球治理进程。

目 录

前 言	I
一、全球人工智能治理现状	1
(一) 多边平台的进展	1
(二) 多方机制的成果	2
(三) 相关国家/地区的做法	3
(四) 科技企业的实践	5
二、全球人工智能治理的十大重要议题	6
(一) 国家主权原则和人工智能发展	6
(二) 社会变革和可持续发展	6
(三) 技术创新和产业发展	7
(四) 人机情感和生命伦理	8
(五) 内容安全风险	8
(六) 模型算法安全风险	8

(七) 数据安全和隐私保护	9
(八) 产品责任和风险	9
(九) 知识产权保护	10
(十) 智能鸿沟和国际协作	10
三、构建完善全球人工智能治理体系	11
(一) 目标宗旨	11
(二) 原则共识	11
1. 尊重国家主权	11
2. 统筹发展和安全	11
3. 坚持平等互利普惠	12
(三) 行动路径	12
1. 坚持以人为本	12
2. 坚守智能向善	12
3. 赋能千行百业	12

4.防范应对安全风险	13
5.加强国际合作能力建设	13
6.完善全球治理机制	14
附录：近年来全球人工智能治理重要文件概览	16

一、全球人工智能治理现状

1956年，达特茅斯会议正式确定了“人工智能”术语，开辟了人工智能科学的独立研究。此后，人工智能发展经历了数次起伏，并于上世纪九十年代取得了一定突破。1997年深蓝超级计算机战胜国际象棋世界冠军，它不仅引起了公众对于技术未来的思考，也促进了国际社会关于人工智能伦理的早期讨论。

随着人工智能技术向机器学习、强化学习、深度学习等方向演进，2016年AlphaGo战胜围棋世界冠军成为标志性事件，全球人工智能治理迈入制定原则方针的探索阶段。如2016年，电气电子工程师学会提出《人工智能的道德准则设计》，为自主和智能系统的设计和开发提供指引；2017年，全球行业领袖发起《阿西洛马人工智能原则》，提出技术发展“有益于人类”的守则等。

2022年以来，ChatGPT问世并风靡全球，加速推动人工智能从“小模型+判别式”向“大模型+生成式”转变，加强全球人工智能治理的呼声和行动也全面兴起。

（一）多边平台的进展

多边治理进程是全球人工智能治理的主要渠道，是推动国际社会在人工智能相关领域达成普遍共识的关键。

联合国在全球人工智能治理上发挥了重要引领作用。2023年，联合国秘书长安东尼·古特雷斯宣布成立人工智

能高级别咨询机构，这一机构在 2024 年发布了《为人类治理人工智能》最终报告，提出了加强全球合作的行动方案。2024 年，联合国大会通过了《加强人工智能能力建设国际合作》《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》两项决议，有力促进了人工智能国际合作；联合国未来峰会通过《全球数字契约》，提出“加强人工智能国际治理，造福人类”的目标，授权联合国设立人工智能国际科学小组和人工智能治理全球对话等机制。

此外，联合国框架下的多个机构也取得了进展。教科文组织提出的伦理框架已得到了 50 多个国家积极参与；国际电信联盟连续七年举办人工智能向善全球峰会，已成为全球人工智能交流对话的重要平台。

政府间组织也将全球人工智能治理纳入讨论议程，并取得了相应成果。二十国集团在新德里峰会重申了“以人为本，实现人工智能向善并服务全人类”的观点；金砖国家在 2022 年同意成立人工智能研究组，意在加强人工智能技术研发、标准制定、产业应用等领域务实合作；经济合作与发展组织等也更新了相关建议书。

（二）多方机制的成果

多方机制在技术标准、伦理规范等领域积极推动发展负责任的人工智能开发与应用。

电气电子工程师学会、国际标准化组织和国际电工委员

会等技术社群发布了人工智能前沿技术标准、人工智能设计的伦理准则、人工智能系统管理指南等。2023年，电气电子工程师学会推出了免费获取人工智能伦理和治理标准的计划；国际标准化组织发布的《信息技术 人工智能 管理体系》（ISO/IEC 42001）则为各类组织管理人工智能提供了框架。

此外，一些国际组织和会议论坛也在该领域加快行动。2023年，世界经济论坛成立了人工智能治理联盟，提出负责任地开发、开放创新和国际合作等建议；世界互联网大会人工智能工作组发布了《发展负责任的生成式人工智能研究报告及共识文件》，致力于推动发展负责任的人工智能。2024年，全球移动通信系统协会推出《负责任的人工智能成熟度路线图》，可帮助通信企业评估其负责任的人工智能成熟度水平。

（三）相关国家/地区的做法

相关国家和地区基于不同的政治文化、产业发展和目标诉求等因素提出了各具特点的治理方案。

中国遵循习近平主席提出的“坚持以人为本、智能向善”的理念和宗旨，统筹人工智能发展和安全，取得了积极进展。2023年，中国提出《全球人工智能治理倡议》，围绕人工智能发展、安全、治理三方面系统阐述了人工智能治理中国方案；国家互联网信息办公室等七部门联合公布《生成式人

工智能服务管理暂行办法》，此部立法系全球首部生成式人工智能专门立法。2024年，世界人工智能大会发布《人工智能全球治理上海宣言》，呼吁全球共同推动人工智能健康可持续发展；中国在第19届联合国大会期间发布《人工智能能力建设普惠计划》，倡议成立“人工智能能力建设国际合作之友小组”，提出“五大愿景”“十项行动”；中国全国网络安全标准化技术委员会发布《人工智能安全治理框架》1.0版。

美国强调创新发展，发布《关于安全、可靠和值得信赖地开发和人工智能的行政命令》等多部政策文件，加大对人工智能创新生态的资金、人才等支持，并在人工智能风险管理上引导建立最佳实践和标准，同时对先进芯片等实施出口管制，以维持美国在人工智能领域中的领先地位。

欧盟注重人工智能监管立法，相继制定出台人工智能相关战略、宣言、计划、指南等文件，特别是在2024年率先推出了《人工智能法》，积极引领人工智能全球治理。

英国将人工智能视为提升其全球影响力的机遇，发布《国家人工智能战略》文件，采取“支持创新”的监管方式，2023年举办首届全球人工智能安全峰会并推动签署首个人工智能安全国际性声明——《布莱切利宣言》。

发展中国家也在积极行动。肯尼亚将人工智能和区块链视为“关键的经济和商业支持技术”，2024年推出了《信

息技术-人工智能-人工智能应用实践守则》；埃及 2019 年组建国家人工智能委员会，相继发布《国家人工智能战略》等文件；哈萨克斯坦 2024 年发布《2024-2029 年人工智能发展构想（草案）》，加快构建本国人工智能生态系统。

（四）科技企业的实践

大型科技企业在技术、产业和市场领域具有优势，也是参与制定和推进落实全球人工智能规则的主要载体。

在组织建设层面，科技企业不仅设立了内部伦理委员会，还推出了企业人工智能原则，涵盖对社会有益、安全、保护隐私、公平、透明、可解释、可控、负责任等内涵。

在产业实践层面，科技企业在提升人工智能系统的安全性和稳健性、保护数据隐私、提升透明度、提高可解释性、保障公平包容、确保价值对齐、开展内容安全审核等方面提供了解决方案。

在国际合作层面，20 家科技企业在慕尼黑安全会议上宣布，将联合打击深度伪造信息，努力检测和抵制欺骗性的人工智能生成内容；16 家公司签署《前沿人工智能安全承诺》，承诺实施内外部测试、信息共享、网络安全投资、第三方漏洞报告机制等措施。

总体而言，全球人工智能治理在原则议题上达成了一定的共识。但是，各类机制仍然高度碎片化，在理念细化和方案落实方面存在较多分歧。发达国家在人工智能领域拥有较

多跨国企业，把人工智能议题与自由、民主、人权等议题挂钩，其主导的人工智能治理平台起步较早、影响力较大。相较而言，大多数发展中国家则缺少领先的人工智能科技企业，更多关注能力建设、智能鸿沟等发展议题，在全球治理进程中声势较弱、代表性和话语权不足。尤其是一些国家以意识形态划线，构建排他性集团或制造发展壁垒，阻碍了全球人工智能治理合作。

二、全球人工智能治理的十大重要议题

随着算法模型、算力和数据领域不断取得突破，人工智能技术战略性地位进一步凸显，正在对经济发展、社会进步、地缘政治等方面产生重大而深远的影响。目前，全球人工智能治理已形成十大重要议题。

（一）国家主权原则和人工智能发展

人工智能技术的发展与国家主权问题密切相关。一方面，人工智能技术的战略性发展可用于维护国家主权、安全、发展利益。另一方面，国家主权也面临着来自技术自主性、算法模型、数据、内容、算力、供应链等安全风险的挑战。为此，国际社会已围绕人工智能领域国家主权原则适用性问题、人工智能时代国家主权的内涵和外延变化以及如何维护国家主权、安全和发展利益等展开相关探讨。

（二）社会变革和可持续发展

人工智能技术在赋能社会经济发展的同时，也对原有社

会秩序带来了冲击。发展性议题包括如何利用人工智能技术帮助传统产业转型升级，提高生产效率，赋能智慧医疗，智慧教育和智慧城市建设，创造新的就业机会，助力推动经济社会发展绿色化、低碳化等；约束性议题包括人工智能训练和运行需要大量的计算资源和能源对环境造成的负担问题，人工智能在化学、生物、核能等领域给人类社会带来的安全风险问题，以及人工智能技术发展导致的结构性失业问题等。

（三）技术创新和产业发展

加快发展人工智能是事关能否抓住新一轮科技革命和产业变革机遇的战略问题。相关讨论包括：人工智能的技术研发，即如何发展人工智能前沿基础理论、关键共性技术、提升智能化水平；人工智能的算力问题，即如何加快部署算力基础设施，优化布局算力资源，高效分配和合理使用算力促进技术和产业发展；人工智能的创新应用，即如何将创新成果转化为各领域新的生产力；人工智能的产业政策，即如何利用体制、机制、资金、人才等各类资源来发展人工智能；人工智能的标准制定，即如何形成开发、部署和使用人工智能的行业标准；人工智能的产业链和供应链安全，即如何避免人工智能发展所需的先进芯片、关键软件等成为少数国家开展地缘政治博弈的工具，确保人工智能产业链供应链在最大限度嵌入全球分工体系的同时，增强抗风险的安全能力和

韧性。

（四）人机情感和生命伦理

人工智能技术可能引发根本性的伦理问题，危及人类自身生存和发展。全球各界主要关注以下议题：人机交互中的情感依附探讨，即如何应对在人机交互中出现的情感错觉，避免人际关系疏离与系统沉迷；人工智能决策的伦理抉择与道德判断问题，即如何制定人工智能伦理规则，避免人工智能失控或对人类生命、权利和利益造成损害；人工智能带来的结构性影响，即可能颠覆就业观、生育观、教育观等问题。

（五）内容安全风险

滥用人工智能技术生成错误、虚假信息等加剧了内容安全风险。相关治理议题包括：大模型基于训练数据进行模仿而非理解的特性，可能生成错误的、不准确的、不真实的“幻觉”内容；恶意使用者利用深度伪造等技术可合成逼真的图像或音视频，操纵舆论、编造与传播虚假信息等问题；智能算法推荐“投其所好”的特点造成或加剧“信息茧房”，致使个体认知窄化、加深刻板印象等。

（六）模型算法安全风险

模型算法是人工智能系统的核心组件，直接影响系统运行方式。主要涉及以下内容：模型算法的可解释性，即要求输出结果可预测、可归因、可修正和可追责；模型算法的可靠性，即要求避免输出偏见或歧视性内容以及“智能幻觉”；

模型算法的鲁棒性，即要求避免因复杂多边运行环境、恶意干扰或诱导影响等因素带来的性能下降、决策错误等问题；模型算法的公正性，即算法在决策过程中不偏袒某些群体或个人，避免产生歧视和不公平现象；模型算法的抗攻击性，即避免参数、结构、功能等核心信息被窃取、篡改、嵌入后门、攻击的风险等问题。

（七）数据安全和隐私保护

数据安全对于确保人工智能发展过程中的安全性、可靠性、可控性和公平性意义重大。全球关注的数据安全议题主要包括：在数据的收集和使用中，如何避免未经同意收集、不当使用个人数据和信息等问题；在数据训练中，如何确保数据质量，避免使用含有虚假、偏见、侵犯知识产权等信息，防范数据被篡改或污染以及数据标注不规范；在数据处理中，如何避免数据泄露，防止非授权访问、恶意攻击、诱导交互等。

（八）产品责任和风险

人工智能产品（如具身智能）涉及多种复杂技术和多元法律主体，出现问题后的责任归属难以明确，已成为治理面临的一大挑战。通常涉及以下内容：人工智能产品的设计者或运营者责任，即因模型或算法中的错误、疏忽或故意不当行为而应承担的责任；人工智能产品生产者责任，即因产品存在缺陷或者未能达到预期安全标准而应承担的法律责任；

人工智能产品使用者责任，即因不当操作需承担的法律责
任。

（九）知识产权保护

人工智能技术及其相关产品应用给知识产权保护带来
新挑战，需要政策和法律制定者不断做出调整。专利权方面，
人工智能算法、模型以及系统某些组件，可能成为专利权的
保护对象，但纯算法或数学方法在某些司法管辖区不能申请
专利。著作权方面，各国对人工智能生成的内容是否可获得
著作权保护尚无定论，对于权利归属于人工智能产品的开发
者、使用者或者其他相关方也存在争议。侵权行为方面，人
工智能系统使用受版权保护的数据训练具有侵权风险；人工
智能生成的品牌名称或标识等也可能涉及商标权保护问题。

（十）智能鸿沟和国际协作

国家间的智能鸿沟问题是人工智能技术发展在全球数
字鸿沟领域的新表现，迫切需要加强国际协作予以解决。智
能鸿沟问题已在人工智能的政策、技术、产业、应用、治理
等领域呈现全面加剧趋势。各国虽然都重视人工智能的国际
协作，包括国际标准制定、提高政策的互操作性、国际研发
合作等。但是，个别国家在人工智能领域大搞技术封锁，通
过阻断别国的人工智能技术发展来寻求在该领域的所谓领
导地位，该势头如果得不到制约，将严重干扰全球人工智能
的发展和治理进程。

三、构建完善全球人工智能治理体系

为更好地把握人工智能发展带来的机遇,应对人工智能发展带来的全球性风险挑战,提升全球人工智能治理的成效,应进一步推动完善全球人工智能治理体系建设,主要包括明确目标宗旨、凝聚原则共识、拓展深化行动路径等。

(一) 目标宗旨

加强信息交流和技术合作,共同做好风险防范,形成具有广泛共识的人工智能治理框架和标准规范,构建开放、公正、有效的治理机制,不断提升人工智能技术的安全性、可靠性、可控性、公平性,促进人工智能技术造福人类,推动构建人类命运共同体。

(二) 原则共识

1. 尊重国家主权

各国有权根据国情自主选择技术发展模式和治理方案。面向他国提供人工智能产品和服务时,应尊重他国主权、遵守他国法律,反对利用人工智能技术和应用干涉他国内政。

2. 统筹发展和安全

全球人工智能治理应坚持发展与安全并重原则。一方面,应尊重技术发展规律,鼓励和推动人工智能技术创新,释放其在各领域的应用潜力;另一方面,应秉持基于风险的治理理念,将安全意识和监管措施贯穿人工智能发展研究、设计、开发、部署、使用等生命周期的各个阶段。

3.坚持平等互利普惠

发展人工智能应坚持平等、互利、普惠原则。一是促进各国在开发和利用人工智能技术方面权利平等、机会平等、规则平等。二是促进人工智能技术和知识共享，减少技术壁垒，缩小智能鸿沟。三是促进全球人工智能市场的开放性和竞争性，防止垄断行为，并避免技术问题政治化，使全球各国多边各方都能共享人工智能带来的福祉。

（三）行动路径

1.坚持以人为本

全球人工智能治理的价值取向应当始终坚持“以人为本”理念，以尊重人类权益为前提，以促进人类可持续发展为动力，以提高人类共同福祉为目标。积极支持以人工智能助力可持续发展，应对气候变化、生物多样性保护等全球性挑战。

2.坚守智能向善

遵守适用的国际法，符合和平、发展、公平、正义、民主、自由的全人类共同价值，共同防范和打击恐怖主义、极端势力和跨国有组织犯罪集团对人工智能技术的恶用滥用。

3.赋能千行百业

开展人工智能模型研发合作，促进人工智能关键技术和系统平台优化升级。完善全球可互操作的人工智能和数字基础设施布局，为各国特别是发展中国家开展技术应用和场景

创新提供算力和算法资源。加快人工智能创新链、产业链深度融合，形成丰富多样、健康向善人工智能发展生态。推进人工智能全方位全链条多场景赋能实体经济，支持以人工智能助力可持续发展，应对气候变化、保护生物多样性等。

4.防范应对安全风险

坚持与时俱进、敏捷应对策略，密切关注人工智能安全风险变化，快速动态精准调整治理措施。加强人工智能数据安全，合作推动数据依法有序自由跨境流动，探索构建数据共享的全球性机制平台。推动人工智能数据语料库平等多样，消除种族主义、歧视和其他形式的算法偏见，保护文明多样性。支持在联合国框架下，建立兼顾发展中国家利益的、全球可互操作的人工智能安全风险评估框架、标准和治理体系。共同研判人工智能研发与应用风险，完善应对人工智能安全风险的技术和政策。严厉打击滥用人工智能的犯罪行为。

5.加强国际合作能力建设

一方面，全球应发挥联合国机制性协调的关键作用，开展人工智能领域南北合作、南南合作和三方合作，共同落实联合国未来峰会成果；同时，提升发展中国家或技术落后国家的代表性、参与度和话语权，反对构建排他性组织、恶意阻挠他国技术发展的行为；并在普遍参与的基础上，就人工智能全球治理规则达成国际协议。另一方面，国际社会应强

化能力建设，弥合智能鸿沟，积极协助全球南方发展人工智能技术和服 务；提升公众人工智能素养，保障妇女和儿童数字和智能权益，共享人工智能知识成果和经验。

6.完善全球治理机制

联合国作为综合类国际组织，是全球人工智能治理的主渠道，应积极探索在联合国框架下成立人工智能国际治理机构，协调涉及人工智能发展、安全与治理的重大问题。区域性多边组织和国际标准化组织、世界经济论坛、世界互联网大会等各类专门性组织应发挥重要作用。其他人工智能的倡议机制、安全研究网络、企业联盟等组织形式是重要参与方。政治组织、行业组织和技术组织之间同样需要加强互动。主要大国应增强政治互信和合作意愿，建立应急联络与协作机制，合理管控人工智能领域的竞争。此外，在全球人工智能治理体系内，应逐步推动信息共享、知识分享、风险共担、利益共享机制的形成和健全。

人工智能技术仍在快速发展过程中，全球人工智能治理亦方兴未艾，攸关全人类命运，是世界各国面临的共同课题。推进全球人工智能治理是一个长期的过程。为协调全球人工智能发展、安全与治理重大问题，国际社会应积极推动建立人工智能治理机制，支持联合国在治理进程中发挥主渠道作用，加强南北合作和南南合作，提升发展中国家的代表性和

发言权,尽快形成具有广泛共识的人工智能治理框架和标准规范。各国政府应积极参与人工智能全球治理,遵守相关的国际法和国际规范,加强国际交流合作。国际组织、企业、研究机构、社会组织和个人等多元主体应积极发挥与自身角色相匹配的作用,参与人工智能治理体系的构建和实施。

展望未来,面对不同国家和地区间仍存在的数字鸿沟和智能鸿沟,各国政府、科技界、产业界等利益相关方要携手合作,充分发挥人工智能的潜力,共同推动人工智能健康发展,共同维护人工智能安全,赋能人类共同的未来,推动构建人类命运共同体。

附录：近年来全球人工智能治理重要文件概览

主要行为体	时间	相关文件
联合国大会	2024 年	《加强人工智能能力建设国际合作》《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》
联合国人工智能高级别咨询机构	2023 年	《为人类治理人工智能》临时报告
	2024 年	《为人类治理人工智能》最终报告
联合国教科文组织	2021 年	《人工智能伦理问题建议书》
	2023 年	《生成式人工智能在教育和研究中的应用指南》
世界卫生组织	2021 年	《世界卫生组织卫生健康领域人工智能伦理与治理指南》
二十国集团	2019 年	《二十国集团人工智能原则》
经济合作发展组织	2019 年 (2024 年更新)	《关于人工智能的建议书》
七国集团	2023 年	《广岛进程先进人工智能系统开发组织国际指导原则》《广岛进程先进人工智能系统开发组织国际行为准则》
欧洲委员会	2024 年	《人工智能与人权、民主和法治框架公约》
全球人工智能安全峰会	2023 年	《布莱切利宣言》
	2024 年	《关于安全、创新和包容性人工智能的首尔宣言》《首尔人工智能安全科学国际合作意向声明》
世界经济论坛	2023 年	《关于负责任的生成式人工智能的建议》
世界互联网大会	2023 年	《发展负责任的生成式人工智能研究报告及共识文件》

主要行为体	时间	相关文件
世界人工智能大会	2024 年	《人工智能全球治理上海宣言》
北京 AI 安全国际对话	2024 年	《北京 AI 安全国际共识》
全球移动通信系统协会	2024 年	《负责任的人工智能成熟度路线图》
国际标准化组织	2023 年	《信息技术 人工智能 管理体系》 (ISO/IEC42001)
中国	2023 年	《全球人工智能治理倡议》《生成式人工智能服务管理暂行办法》
	2024 年	《人工智能安全治理框架》1.0 版
美国	2023 年	《人工智能风险管理框架》《关于安全、可靠和值得信赖地开发和人工智能的行政命令》
欧盟	2024 年	《人工智能法》
英国	2023 年	《促进创新的人工智能监管方法》
日本	2024 年	《人工智能运营商指南（草案）》
新加坡	2024 年	《生成式人工智能治理模型框架》
东盟	2024 年	《东盟人工智能治理与伦理指南》