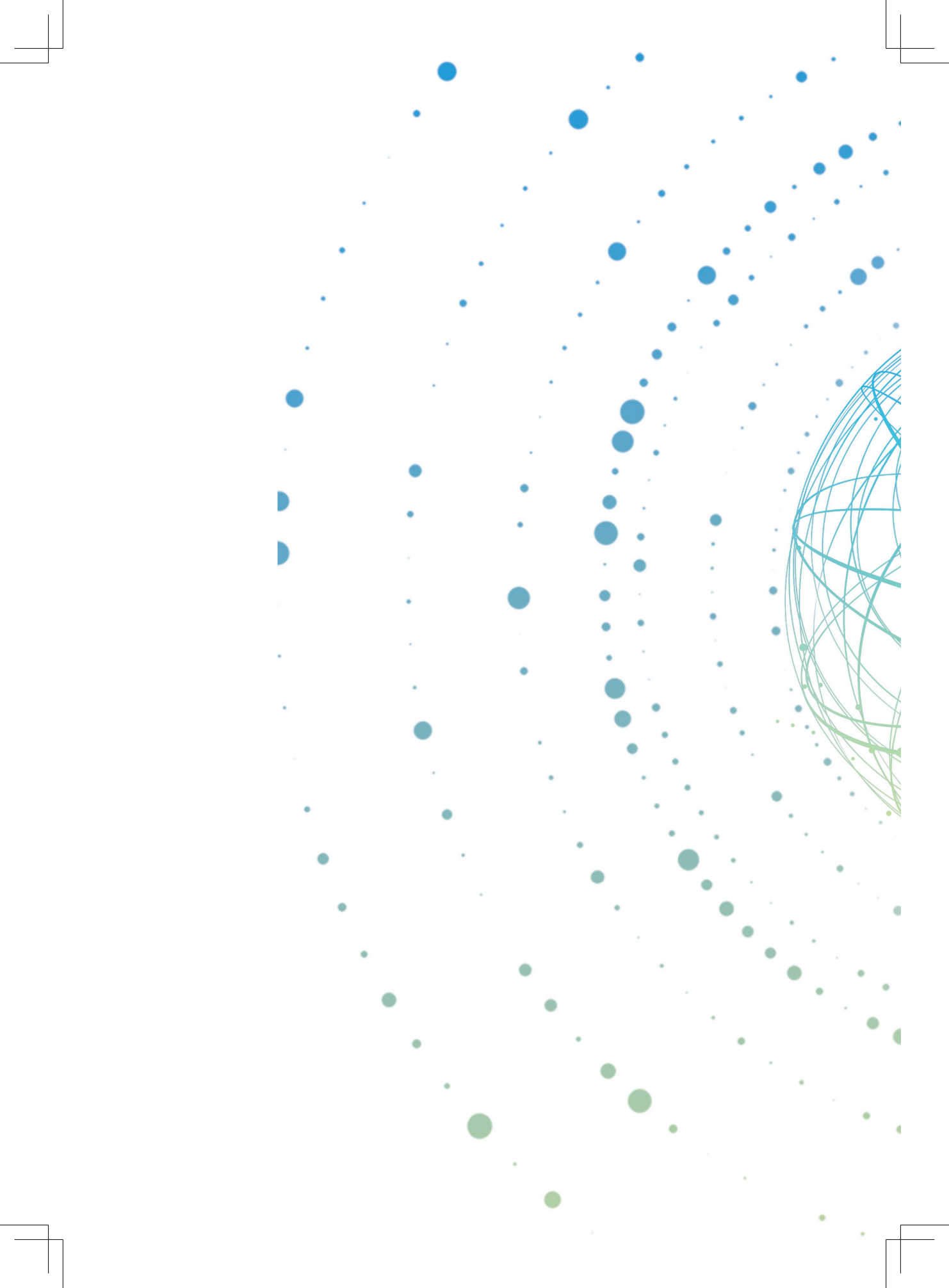


当前，人工智能技术创新进入前所未有的活跃期，深刻重塑人类经济、社会生活的方方面面。技术的长足进步，一方面带来历史性的发展机遇与红利，另一方面也蕴含多方面的风险和挑战。如何对人工智能实施有效治理，在鼓励创新与防范风险之间保持动态平衡，是国际社会面临的共同课题。中国是全球重要创新力量和人工智能技术与应用大国，近年来遵循以人为本、智能向善的价值导向，坚持发展与安全并重的原则，把握人工智能发展趋势和规律，不断完善相关法律法规、政策制度、应用规范、伦理准则等，以实际行动稳步推进人工智能治理。作为负责任大国，中国还为人工智能全球治理提出一系列倡议、计划，并搭建平台，为推动人工智能治理国际合作积极贡献力量。

为促进人工智能治理规则的国际交流，中国国际问题研究院和中国信息通信研究院组织专门力量，对中国人工智能治理的价值理念、制度架构、重点方向、实践经验等进行梳理，分析其世界意义，形成此篇研究报告。





CAICT
中国信通院

中国的人工智能治理： 理念、实践和意义

中国国际问题研究院
中国信息通信研究院

二零二六年十月

目录

01	前言	2
02	人工智能治理的全球生态与中国实践	6
	（一）人工智能治理的全球生态	6
	（二）人工智能治理的中国理念与实践	11

03 中国人工智能治理的重点方向 20

- (一) 推动人工智能伦理原则落地实施 21
- (二) 开展模型治理确保安全可靠 23
- (三) 推进数据治理护航技术发展 27
- (四) 开展人工智能生成内容治理 30
- (五) 实施场景化治理促进技术应用 33
- (六) 探索多样化主体治理提升治理实效 35

04 中国人工智能治理的世界意义 42

- (一) 彰显确定性，探索和积累人工智能治理经验
 - 43
 - 46
- (二) 秉持开放性，推动人工智能治理国际合作

中国的人工智能治理：理念、实践和意义

01 前言

中国的人工智能治理：理念、实践和意义

01 前言

2025年9月，习近平主席在“上海合作组织+”会议上提出全球治理倡议，强调要奉行主权平等、遵守国际法治、践行多边主义、倡导以人为本、注重行动导向，共同推动构建更加公正合理的全球治理体系，携手迈向人类命运共同体。推动人工智能治理是改革完善全球治理的重要内容。《全球治理倡议概念文件》指出，要优先考虑在人工智能、网络空间等治理紧迫性突出、治理赤字较大的领域加大沟通合作，积极凝聚共识、锁定成果，争取早期收获。

在人工智能治理问题上，中国遵循以人为本、智能向善的价值导向，坚持发展与安全并重的原则，稳步扎实推进国内治理。在此基础上，中国秉持开放精神，积极推动人工智能治理国际合作，先后制定发布多份人工智能全球治理合作相关政策文件，成立世界数据组织，举办世界人工智能大会，筹建世界人工智能合作组织，并充分发挥联合国主渠道作用，将帮助全球南方国家加强能力建设、弥合数智鸿沟作为重点工作之一，呼吁早日形成具有广泛共识的全球治理框架，以实际行动推动人工智能全球治理，为改革完善全球治理贡献新的理念和实践。

中国的人工智能治理：理念、实践和意义

02

人工智能治理的 全球生态与中国实践

- 人工智能治理的全球生态
- 人工智能治理的中国理念与实践

近年人工智能技术快速迭代，已开始投入规模化应用，相应的治理问题也从科技专业议题跃升为全球关注的政策焦点议题。全球人工智能治理逐渐形成多元并存、动态演进、竞合交织的生态。中国是人工智能技术与应用大国，自 2017 年发布《新一代人工智能发展规划》，经过多年持续探索与迭代，治理实践逐步从零散走向系统，从被动应对走向主动构建，走出了一条富有特色的人工智能治理道路。

（一）人工智能治理的全球生态

1. 人工智能治理的热度持续走高

2017 年，首届“人工智能向善全球峰会（AI for Good Global Summit）”在日内瓦召开。同年，联合国秘书长古特雷斯在联大报告中纳入人工智能治理问题。此后，国际社会对这一议题的关注度与投入度持续增加，涌现出许多不同的治理主体、机制、倡议、行动等。



在代表性最广的联合国层面，秘书长古特雷斯高度重视这一议题。2023年10月，成立秘书长人工智能高级别咨询机构，该机构于2024年9月发布《为人类治理人工智能》最终报告。2024年，联合国召开未来峰会，通过《未来契约》及其附件——《全球数字契约》和《子孙后代问题宣言》。《全球数字契约》是全球首个涵盖数字合作与人工智能治理的综合性框架。2025年8月，联合国大会通过决议，正式决定设立国际人工智能科学小组和全球人工智能治理对话两个机制。2026年7月1日，科学小组发布首份报告《初步报告：对人工智能带来的机遇、风险和影响的循证评估》。7月6-7日，首届全球人工智能治理对话在瑞士日内瓦举行，多国政界、企业届、学术界、技术社区等的代表参加会议，围绕人工智能带来的机遇及影响、弥合人工智能鸿沟、人工智能治理路径的互操作性与兼容性、透明度、问责制与人类监督等主题展开讨论。

除联合国秘书长主导的上述机制外，联合国系统的多家专门机构也在各自领域积极探索人工智能治理。联合国教科文组织（UNESCO）先后发布《人工智能伦理问题建议书》《生成式人工智能在教育和研究中的应用指南》等治理类文件，还成立全球人工智能伦理与治理观察站（Global AI Ethics and Governance Observatory）。国际电信联盟（ITU）每年召开人工智能向善全球峰会，并有意与国际标准化组织（ISO）和国际电工委员会（IEC）一起主导人工智能领域的标准制定。世界知识产权组织启动“知识产权与前沿技术对话会”、发布实操性治理工具包。国际劳工组织建立“人工智能与数字经济工作观察站”等。

其他多边机制也积极推动人工智能治理。经合组织理事会于2019年通过《人工智能建议书》，提出5项人工智能原则和5个建议，后于2024年更新。其后，经合组织又推出人工智能政策观察站、可信赖

中国的人工智能治理：理念、实践和意义

人工智能工具框架、人工智能事件通报共同框架等措施工具。二十国集团峰会于 2019 年通过“G20 人工智能原则”。此后每年，峰会文件都会纳入人工智能治理相关内容。近两年在巴西和南非召开的峰会均将人工智能治理作为优先议题，并将其与可持续发展、数字公平问题联系讨论。2025 年，金砖国家领导人峰会发表《金砖国家领导人关于人工智能全球治理的声明》，提出开发和应用人工智能的一系列指导方针：一是多边主义、合法性与数字主权，二是市场规范、数据治理与技术可及性，三是公平与可持续发展。

欧盟和东盟是在人工智能治理领域取得显著进展的两个区域组织。欧盟于 2024 年通过全球首部综合性人工智能立法——《人工智能法》，提出以风险分级为核心实施人工智能监管。东盟于 2024 年发布《东盟人工智能治理与伦理指南》作为地区性的指导文件，并于 2025 年发布指南扩展版，加入生成式人工智能的内容。

英国牵头发起的人工智能治理系列峰会是近年影响较大的另一组多边行动。2023 年，英国在布莱切利召开首届人工智能安全峰会，聚焦前沿人工智能带来的“严重甚至灾难性”安全风险。2024 年，第二届峰会在韩国首尔召开，主题是改善人工智能的安全性、创新性和包容性，会后组建了人工智能安全机构的合作网络。2025 年，法国巴黎人工智能行动峰会举办，在强调创新和人工智能经济机遇的同时，对人工智能的环境影响和劳动力冲击等更广泛治理议题进行了探讨。2026 年的印度人工智能影响峰会主要讨论全球南方如何从人工智能技术发展中受益的问题。

在国际法层面，《欧洲委员会人工智能与人权、民主和法治框架公约》是全球首部有国际法律约束力的人工智能治理相关条约，于 2024 年 9 月开放全球签署，截至 2026 年 4 月已有 19 个签署国。《公约》确立了透明度、问责制、非歧视、人权保护等人工智能治理的基



本原则。

2. 人工智能治理存在多重难题

人工智能治理是用规则体系对人工智能全生命周期的活动，包括设计、开发、部署、使用等，进行规范的过程。相较于气候变化、核安全与防扩散、公共卫生等其他全球议题，人工智能技术更具颠覆性，对政治、经济和社会的影响更为广泛，相关治理问题也就更加错综复杂。

首先，人工智能治理不是单一性问题，所涉及的议题领域众多，要形成全覆盖、协调的治理体系难度很大。从宏观看，人工智能治理可分安全和发展两个方向。前者侧重评估和防范人工智能技术可能带来的各类风险，并确保其安全、可靠和可控。后者侧重为人工智能技术创新和应用创造有利条件，并确保技术公平、可及和普惠。从微观看，人工智能治理涉及伦理对齐、隐私与数据治理、算法透明度、算法公平、算法操控、责任界定等诸多问题，相关技术应用到不同场景又会产生更多样的问题。这些都推高了协调治理的难度。

第二，人工智能技术仍在不断突破，治理措施恐将长期处于对技术的“追赶”状态。人工智能创新浪潮持续加速，新的算法架构、训练范式与应用形态层出不穷，人工智能的能力边界不断外延，而潜在风险难以预先穷举或清晰界定，这与传统技术治理着眼相对静态标准的模式极为不同。如何构建与人工智能进化速度相适应的动态治理机制，是一个紧迫命题。

第三，地缘政治因素深刻影响人工智能治理的实施。人工智能被视为影响未来国家综合国力和国际竞争力的战略性技术，各国政府普遍担忧本国单边监管将导致国家竞争力的损失，这使一国内部的人工智能治理常在“保安全”还是“求发展”的两端间摇摆，而多边人工智能治理也由于信任不足而难有实质性推进。

中国的人工智能治理：理念、实践和意义

在上述因素共同作用下，当前全球人工智能治理遭遇一定困局。以欧盟为例，其《人工智能法》是全球首部人工智能领域的综合性立法，但实施伊始就面临技术标准缺失、监管机构缺位、合规工具与指南不足、企业合规成本高等诸多问题。欧盟旋即着手对法案进行修订。2026年5月，欧盟成员国与欧洲议会达成临时协议，对原有监管框架进行一定程度放宽，并推迟部分关键规则的执行时间。在美国，特朗普政府以赢得人工智能竞赛为目标制定相关政策，优先考虑消除监管障碍、加速创新。特朗普第二任期伊始即废止拜登政府发布的较为关注风险和权利的《关于安全、可靠和值得信赖的人工智能发展与使用的行政令》，并于2025年底和2026年6月分别发布《确保人工智能国家政策框架》和《推动前沿人工智能创新与安全建设》两个行政令，前者旨在以最低程度的国家统一政策框架压制快速增长的州级监管法规，后者鼓励人工智能企业与政府开展自愿合作。目前美国没有联邦层面的刚性监管，实操中偏向由人工智能企业进行自愿评估和自我约束，这种“轻监管”模式给全球人工智能治理造成负面影响。

从多边角度看，一方面，全球人工智能治理面临碎片化风险。国际社会尚无单一权威框架主导人工智能治理。众多机制平行运作，治理理念、治理内容、风险界定、技术标准等各不相同，机制间缺乏有效协调和衔接。另一方面，多边治理缺乏有效约束力和执行力。目前绝大多数人工智能治理相关的国际宣言、倡议、建议书等仍停留在原则共识阶段，且为自愿性质，属于“软法”范畴，签署国可自行决定落实力度，监督和执行并无保证。《欧洲委员会人工智能与人权、民主和法治框架公约》是唯一具有约束力的国际“硬法”，但尚未达到法定生效条件。此外，美国政府明确反对多边人工智能治理，也给全球人工智能治理蒙上阴影。在2026年印度人工智能影响峰会上，美



国代表就声明美国“完全拒绝”人工智能的全球治理。

（二）人工智能治理的中国理念与实践

中国高度重视人工智能技术的治理问题，经过多年探索与实践，已逐步走出一条富有特色的人工智能治理道路。

1. 形成涵盖价值、原则与方法的全方位治理理念

中国的人工智能治理遵循以人为本、智能向善的价值导向。中国视增进人民福祉、服务经济社会发展为技术发展的根本目标，强调技术发展应始终服务公共利益，防止技术逻辑脱离社会价值与伦理约束。中国坚持技术发展必须与社会公平、公共秩序和人的全面发展相协调，促进社会整体福祉。在治理实践中，中国针对算法推荐、深度合成、生成式人工智能等重点领域的风险，持续加强对虚假信息、网络暴力、低俗内容、深度伪造等问题的治理，强化对未成年人、个人隐私和知识产权等的保护，防止人工智能成为错误价值导向和作恶滥用的工具。中国推动人工智能尽快应用于医疗、教育、乡村振兴、公共服务等领域，促技术红利更快惠及普通民众，着力缩小城乡、区域数字鸿沟，让技术赋能民生改善，令人工智能成为促进共同富裕的重要推手。

中国的人工智能治理确立发展与安全并重的原则。中国统筹推进技术创新与风险防控，既鼓励人工智能赋能经济社会高质量发展，又重视筑牢安全底线。在鼓励创新一侧，中国完善算力基础设施，建设国家人工智能创新平台，通过国家中长期发展规划持续推动人工智能与实体经济深度融合，简化合规流程，为企业创新营造良好环境。在风险防控一侧，坚持系统施策、精准发力，构筑多层次风险防控体系，切实保护公民权益、数据安全、社会公平和公序良俗。根据技术风险等级、应用场景差异和社会影响程度等，实施分类指导和差异化管理。

中国的人工智能治理：理念、实践和意义

对高风险人工智能技术和应用实行审查备案制度，压实企业主体责任。对深度伪造、算法歧视、虚假信息传播等高频风险领域开展专项整治。加强人工智能相关伦理教育和宣传引导。建设人工智能安全监测平台，对技术风险进行实时监测、预警和处置。促进政府、企业、学界、社会对人工智能的协同共治。

中国的人工智能治理采取敏捷治理的方法。针对人工智能技术快速迭代、应用场景持续扩展和风险动态变化等特点，中国以“小步快跑、持续迭代”为行动逻辑，在实践中动态校准治理边界。以《人工智能安全治理框架》为例，2024年9月全国网络安全标准化技术委员会发布框架1.0版，之后针对人工智能推理模型性能跃升、轻量级开源模型大幅降低应用门槛等新情况，于一年后推出2.0版，将安全风险从两大类拓展至三大类别，治理思路由罗列风险清单升级为提供安全治理的系统方法论，并引入“熔断机制”等新工具，还增强了治理的国际接轨性。^①2026年9月，全国网络安全标准化技术委员会再发布框架3.0版，将智能体和具身智能单列为重要风险类型，并将合成数据缺陷、记忆污染、生成式引擎优化（GEO）投毒、自主网络攻击等新型风险纳入治理视野，强化身份权限、工具调用、运行时监测和全过程审计等安全控制。中国在多个城市推行监管沙箱机制，企业在受控环境中开展实践测试时，监管部门同步观察风险、积累经验，并据此动态调整规则。敏捷治理的方法通过构建“感知—决策—处置—反馈”的治理闭环，最大程度保证技术爆发期治理的可塑性与响应速度，是“边

^① “专家解读 | 从风险清单到治理体系：人工智能安全治理框架的演进逻辑分析”，中国国家互联网信息办公室，https://www.cac.gov.cn/2025-09/28/c_1760779758683488.htm。



发展、边治理”思路的集中体现，为形成覆盖人工智能全生命周期的治理体系注入动态适应性。

2. 建立起多层次、多维度的人工智能政策治理与规范体系

在基础层面，《中华人民共和国数据安全法》《中华人民共和国网络安全法》和《中华人民共和国个人信息保护法》三部法律从网络运行安全、数据分类分级与个人信息权益保护等角度，为人工智能技术发展与应用提供基本规则和刚性约束，构成人工智能治理的法律底座。除国家立法外，中国还颁布各类政策文件、部门规章、行业规范和技术指导性文件等，为人工智能的具体发展和场景应用提供政策指导和管理规范，如针对算法推荐、深度合成、生成式服务等具体场景的《互联网信息服务算法推荐管理规定》《互联网信息服务深度合成管理规定》《生成式人工智能服务管理暂行办法》《人工智能生成合成内容标识办法》等；围绕科技伦理管理服务和拟人化互动服务的《人工智能科技伦理管理服务办法（试行）》《人工智能拟人化互动服务管理暂行办法》等。中国还结合自身人工智能技术发展与应用实践，同时参考国际主流人工智能治理理念与实践做法，先后出台了两版《人工智能安全治理框架》，作为国家级技术指导性文件，以促进形成安全、可信、可控的人工智能发展生态，并构建跨国界、跨领域、跨行业的协同治理格局。

中国的人工智能治理：理念、实践和意义

中国人工智能政策文件与法律法规清单

分类	文件名称	发布单位	年份
国家立法	《网络安全法》	全国人大常委会	2017
	《数据安全法》	全国人大常委会	2021
	《个人信息保护法》	全国人大常委会	2021
国家层面 政策法规文件	《新一代人工智能发展规划》	国务院	2017
	《新一代人工智能伦理规范》	国家 AI 治理专委会	2021
	《互联网信息服务算法推荐管理规定》	网信办等四部门	2021
	《关于加强科技伦理治理的意见》	国务院办公厅	2022
	《关于加快场景创新以人工智能高水平应用促进经济高质量发展的指导意见》	科技部等六部门	2022
	《互联网信息服务深度合成管理规定》	网信办等三部委	2022
	《生成式人工智能服务管理暂行办法》	网信办等七部门	2023
	《科技伦理审查办法（试行）》	科技部等九部委	2023
	《人工智能安全治理框架》1.0 版	全国网安标委会	2024
	《人工智能生成合成内容标识办法》	网信办等四部门	2025
	《人工智能科技伦理管理服务办法（试行）》	工信部等十部门	2025
	《人工智能安全治理框架》2.0 版	全国网安标委会	2025
	《人工智能科技伦理审查与服务办法（试行）》	工信部等十部门	2026
	《人工智能拟人化互动服务管理暂行办法》	网信办等五部门	2026
	《智能体规范应用与创新发展实施意见》	网信办等三部门	2026
	《人工智能安全治理框架》3.0 版	全国网安标委会	2026
地方性政策 法规文件	《深圳经济特区人工智能产业促进条例》	深圳市人大常委会	2022
	《上海市推动人工智能大模型创新发展若干措施》	上海市政府	2023
	《深圳市加快打造人工智能先锋城市行动计划（2025—2026 年）》	深圳市发改委	2025
	《北京市人工智能赋能新型工业化行动方案》	北京市经信局	2025
	《北京市人工智能产业发展条例（拟）》	北京市人大	2026



在国家级法律法规指导下，中国不少地方结合各自产业和监管实际，出台了细化和补充性的地方法规，如《上海市促进人工智能产业发展条例》《深圳经济特区人工智能产业促进条例》《杭州市促进具身智能机器人产业发展条例》等。广东、江苏等省份也已将人工智能发展促进条例纳入立法计划。

在逐步完善立法框架的同时，中国持续强化人工智能领域的执法力度，建立了跨部门人工智能安全联合执法协调机制，实现对数据、算法、算力、应用等全要素的协同监管。同时深化司法审判实践，形成行政执法监督与司法审判双向发力的治理格局。

行政执法层面，中国采取备案管理、专项整治与常态化监督相结合的全链条执法体系，如推动生成式人工智能算法备案与安全评估，提升技术透明度、保护用户权益；压实网络平台信息内容管理责任，开展生态治理专项行动，依法处罚未履行审核义务的平台；打击人工智能恶意使用行为，重点整治虚假新闻、身份仿冒、网络舆论操纵等乱象。

司法实践层面，最高人民法院出台人脸识别、平台垄断、新业态劳动者权益保护等司法解释，为算法引发的纠纷提供裁判依据。法院系统积极审理人工智能生成物版权归属、人格权侵权、训练数据著作权及不正当竞争等新型案件，促进产业合规。此外，网络法院等专业化司法机制，也逐步加强对人工智能版权、数据权益和平台责任等问题的裁判与规则探索，推动相关治理规则不断完善。

标准是法律法规的技术延伸和细化支撑。中国高度重视人工智能标准化工作，加快构建人工智能标准体系。2020年8月，国家标准化管理委员会会同网信办、科技部等五部门联合发布了《国家新一代人工智能标准体系建设指南》，从基础共性、关键通用技术、安全与伦理等八个维度搭建了标准框架，旨在为技术标准的制定提供方向指引。

中国的人工智能治理：理念、实践和意义

在此基础上，全国信息安全标准化技术委员会（TC260）起草多份国家标准，重点关注机器学习与算法伦理等方面的内容。^②与此同时，国家标准化管理委员会、工业和信息化部、国家网信办等部门协同推进，依托全国信息技术标准化技术委员会人工智能分技术委员会（SAC/TC 28/SC 42）等标准化组织，统筹基础共性、关键技术、产品应用、安全伦理等重点领域的标准研制。截至 2025 年底，中国已发布人工智能领域国家标准超过 30 项。2026 年，中国计划新制定国家标准和行业标准 50 项以上，为法律法规的有效实施提供坚实的技术支撑。

“十五五”时期，中国将继续完善人工智能法律法规体系，强化立法、执法、司法、守法全链条衔接，为人工智能健康发展提供坚实法治保障。

中国人工智能技术标准相关文件

文件名称	发布单位	年份
《国家新一代人工智能标准体系建设指南》	国家标准委等五部门	2020 年
《国家人工智能产业综合标准化体系建设指南（2024 版）》	工信部等四部门	2024 年
30 余项人工智能国家标准	国家标准委等	截至 2025 年底
《人工智能 智能体互联》系列国家标准	工信部等	2026 年
《智能体应用安全基本要求》	国家标准化管理委员会	2026 年

3. 搭建多元主体协同参与的治理格局

中国立足人工智能技术发展规律与治理现实需求，逐步构建起政

^② 靳雨露、胡家语、肖尤丹：《我国人工智能综合治理机制的立法构造》，《中国科学院院刊》2026 年第 41 卷第 0 期，第 5 页。



府主导、企业主责、科研支撑、社会监督的人工智能协同治理格局。

政府负责强化顶层设计与制度供给，出台伦理规范、法律法规和监管框架，明确治理方向，划定安全底线，统筹推进治理体系建设。一方面，在中央政府层面，国家发展和改革委员会、国家互联网信息办公室、科学技术部、工业和信息化部、公安部、国家数据局等部门，在治理体系中各司其职，又通过政策会商、联合发文、联合执法、专项监管等方式协同联动，合力监管。另一方面，央地积极合作，中央为地方治理提供政策指引与制度依据，地方则立足区域产业基础与应用场景，细化实施细则，开展算法备案、场景试点、风险排查、行业监管等落地工作，并及时反馈新问题、新风险与新需求，推动治理规则动态优化，形成央地协同联动、统一部署与差异化探索相结合的治理结构。

人工智能企业压实主体责任，强化全流程安全与合规，严格落实算法备案、数据安全、伦理风险自查等制度要求，健全内部治理架构，严守国家安全、数据安全与个人信息保护等底线，并自觉接受政府监管、行业约束与社会监督。同时，企业组成行业协会，制定行业标准，开展自律规范，搭建交流平台，引导行业有序发展。

科研机构持续强化理论支撑、技术研判与规则供给能力。中国科学院、中国信息通信研究院等科研机构围绕技术评估、伦理研究、风险识别等开展研究，为治理决策提供专业支持。高校智库、行业研究机构等也通过研究报告、论坛研讨、政策讨论等形式为人工智能治理建言献策，为政策制定提供参考。

社会公众与新闻媒体是人工智能多元治理格局中的重要参与力量。广大公众通过意见反馈、平台举报等方式，对人工智能技术应用开展日常监督。媒体及时曝光治理乱象、普及安全知识、引导社会讨论。二者协同发力，凝聚社会共识，增强治理透明度，并形成有效社会约束。

中国的人工智能治理：理念、实践和意义

03

中国人工智能治理 的重点方向

- 推动人工智能伦理原则落地实施
 - 开展模型治理确保安全可靠
 - 推进数据治理护航技术发展
 - 开展人工智能生成内容治理
 - 实施场景化治理促进技术应用
- 探索多样化主体治理提升治理实效

03

中国人工智能治理的重点方向

中国以保障人工智能始终朝着有益、安全、公平的方向发展为目标开展人工智能治理，坚持发展与安全并重，遵循急用先行、精准施策的原则，聚焦重点问题探索治理方案。通过持续完善相关法律法规、应用规范、监管措施等，确保人工智能安全、可靠、可控，让人工智能成为促发展、护安全的重要动力源。

中国坚持伦理先行，明确技术发展的价值底线，推动伦理规范贯穿人工智能技术发展应用全生命周期。在模型治理方面，通过模型备案、安全评测等制度安排和安全治理框架的探索，把住风险产生的源头关口；在数据治理方面，通过加强数据保护、促进数据高效便利跨境流动、建设高质量数据集等举措，防止数据环节风险向模型训练和应用环节传导；在生成内容治理方面，聚焦模型输出的社会影响，将信息完整性、劳动者和消费者权益保护、维护市场公平竞争作为重点，守好风险进入公共空间的关键闸口；在场景化治理方面，聚焦技术在不同应用场景落地过程中产生的差异化风险，推动治理走向分类施策。此外，



中国人工智能治理的重点方向

通过产业联盟、学术机构等多元主体的协同合作，共同推动人工智能技术向善发展。



（一）推动人工智能伦理原则落地实施

中国秉持“以人为本、智能向善”的理念，从规范确立、制度建设、标准研制、前瞻研究、产业实践及国际合作等方面开展伦理治理工作，以多方参与、协同共治等方式，推动人工智能伦理治理由理念探讨转化为具体实践。

1. 明确伦理原则，确立价值坐标

2017 年国务院印发《新一代人工智能发展规划》，首次在国家战略高度对人工智能伦理建设作出系统性部署。规划提出“三步走”战略目标：到 2020 年初步建立部分领域的伦理规范和政策法规；到 2025 年初步建立法律法规、伦理规范和政策体系；到 2030 年建成更加完善的法律法规、伦理规范和政策体系。这一顶层设计为后续伦理治理工作提供了根本遵循和行动依据。为贯彻落实上述战略部署，科技部于

中国的人工智能治理：理念、实践和意义

2019年2月正式成立“新一代人工智能治理专业委员会”，负责人工智能相关法律、伦理、标准和社会问题研究，深入参与人工智能相关治理的国际交流合作。委员会由来自高校、科研院所和企业的多元背景专家组成，经广泛征集意见、多轮专家研讨，于同年6月发布《新一代人工智能治理原则——发展负责任的人工智能》，提出“和谐友好、公平公正、包容共享、尊重隐私、安全可控、共担责任、开放协作、敏捷治理”等八条核心原则，相关主张与联合国教科文组织（UNESCO）人工智能伦理原则建议书等国际共识具有相当一致性。其中，“敏捷治理”理念尤具前瞻意义，强调治理应具备灵活性和适应性，避免因刚性规制抑制技术创新，彰显了发展与规范相协调的动态平衡思路。在此基础上，委员会后续出台《新一代人工智能伦理规范》，进一步将宏观原则细化为增进人类福祉、保护隐私安全、强化责任担当等6项基本要求和18项具体规范，推动伦理规范全面融入人工智能全生命周期管理。

2. 建立科技伦理制度，将抽象原则转化为可执行规范

中国相继出台《加强科技伦理治理的意见》《科技伦理审查办法（试行）》，初步建立科技伦理制度。一是明确伦理先行原则，推动伦理要求贯穿科技活动全过程。《意见》要求提高科技伦理意识，加强源头治理，在开展科技活动前首先要主动进行科技伦理的风险评估，对于涉及科技伦理风险的，达到了科技伦理审查规范要求的，必须及时开展科技伦理审查。《审查办法》把风险评估、伦理审查、过程监管和跟踪监督连接起来，推动科技伦理从原则倡导转向全过程管理。二是发挥创新主体作用，构建协同共治治理机制。科技伦理治理覆盖范围广、涉及主体多，需要政府、高校、科研机构、企业、社会团体等多元主体共同参与。国家科技伦理委员会指导和统筹协调推进科技伦理治理体系建设工作，地方和行业主管部门按照职责权限分别负责



本地方和各行业相关治理工作。从事科技活动的高等学校、科研机构、企业等创新主体承担科技伦理管理责任，建立常态化的内部工作机制，构建覆盖伦理管理、审查和监督评价的制度体系，将科技伦理深度融入创新活动全流程。科技类社会团体积极承担教育引导和行业自律责任，开展伦理学术研究，宣传普及伦理知识，提高社会公众科技伦理素养。

3. 聚焦人工智能领域，细化重点审查范畴

随着技术快速演进与应用领域拓展，人工智能带来的算法歧视、深度伪造等伦理问题愈发凸显。2026年3月30日，工业和信息化部等十部门联合印发《人工智能科技伦理审查与服务办法（试行）》（以下简称《办法》），明确提出将伦理要求贯穿人工智能全生命周期，并细化规定了适用范围、审查程序等。《办法》要求，在中华人民共和国境内开展的，可能在人的尊严、公共秩序、生命健康、生态环境、可持续发展等方面带来科技伦理风险挑战的人工智能科技活动，其主体都应该进行科技伦理的审查。重点审查的内容包括：是否有利于增进人类福祉、实现社会可持续发展，是否采取措施防止偏见歧视、算法压榨，是否保障人类控制、指导、干预人工智能模型，是否合理披露模型用途、运行逻辑、潜在风险，是否保障责任可追溯可管理，是否充分保护隐私数据等，为中国人工智能伦理治理提供明确的制度指引。为了配合《办法》的实施，中国已展开人工智能伦理风险评估、伦理审查与服务体系建设、伦理委员会建立及伦理实践指南等多项行业标准研制工作，为开展科技伦理审查提供公共服务支持。

（二）开展模型治理确保安全可靠

随着大模型技术快速迭代和人工智能广泛应用，模型治理已成为人工智能安全监管的核心环节。当前，中国围绕模型备案、安全评测

中国的人工智能治理：理念、实践和意义

等重点方向，加快构建覆盖模型全生命周期的安全治理体系，确保人工智能技术始终安全、可靠、可控。

1. 建立模型备案制度，持续提升模型透明度

在监管层面，人工智能透明度是一种信息收集、信息公开和信息共享机制，在各国实践中率先从规则走向实践。备案是中国人工智能治理体系的重要举措。2021 年以来，国家网信办先后出台《互联网信息服务算法推荐管理规定》《生成式人工智能服务管理暂行办法》，设立算法及生成式人工智能模型备案制度，遵循分类分级原则，要求模型披露基本信息，以及数据来源、规模、类型、标注规则，以及安全评估报告等，加强人工智能企业信息透明度。

国家网信办在官网定期发布生成式人工智能服务备案信息公告。已上线的生成式人工智能应用或功能，应在显著位置或产品详情页面公示所使用已备案生成人工智能服务情况，注明模型名称及备案号，以指导使用者科学理性使用相关服务。截至 2026 年 2 月 28 日，累计有 796 款生成式人工智能服务完成备案，481 款生成式人工智能应用或功能完成登记。

2. 健全安全评测标准体系

中国积极推进安全评测标准体系建设，强化数据安全、风险管理、伦理规范标准体系建设，将人类控制、公平性等伦理要求转化为可量化的技术指标。2025 年 2 月，国家市场监督管理总局、国家标准化管理委员会发布了首部面向通用大模型的国家标准——GB/T 45288《人工智能 大模型》系列。该标准将此前分散在不同文件中的技术规范整合为一体，建立起一套更加系统、协调一致的技术规范体系。其中，GB/T 45654-2025《网络安全技术 生成式人工智能服务安全基本要求》已于 2025 年 11 月 1 日正式实施，对生成式人工智能服务的安全能力提出了系统化要求。GB/T 45652-2025《网络安全技术 生成式人工智能



中国人工智能治理的重点方向

预训练和优化训练数据安全规范》、GB/T 45674-2025《网络安全技术生成式人工智能数据标注安全规范》等标准同步生效，为模型数据安全提供了明确依据。此外，GB/T 46347-2025《人工智能 风险管理能力评估》提供了风险管理的方法论，《生成式人工智能大模型安全评估方法 第1部分：生成内容安全评测》等团体标准也已批准立项，研制周期至2026年11月。

中国人工智能模型安全标准相关文件		
文件名称	发布单位	年份
《人工智能安全治理框架 1.0》	全国网安标委	2024 年 9 月
《人工智能安全治理框架 2.0》	全国网安标委	2025 年 9 月
《人工智能 大模型》系列标准	国家市监总局、国家标准委	2025 年 2 月
《网络安全技术 生成式人工智能服务安全基本要求》	全国网安标委	2025 年
《网络安全技术 生成式人工智能预训练和优化训练数据安全规范》	全国网安标委	2025 年
《网络安全技术 生成式人工智能数据标注安全规范》	全国网安标委	2025 年
《人工智能 风险管理能力评估》	全国网安标委	2025 年
《生成式人工智能大模型安全评估方法 第1部分：生成内容安全评测》（团体标准）	中国信息协会	研制周期至 2026 年 11 月
《人工智能安全治理框架 3.0》	全国网安标委	2026 年 9 月

在评测内容层面，安全评测覆盖语料安全、生成内容安全等5大类31小类风险维度，并进一步细化为三级200余个子类，涵盖歧视性内容、暴力和恐怖主义、商业秘密等重点风险领域。评测采用量化硬指标约束，安全测试题库规模要求达到万条以上级别，对高风险问题的拒答率设定了较高标准。在测试方法层面，评测引入自动化检测与

中国的人工智能治理：理念、实践和意义

分级人工复核相结合的模式，通过模拟多种潜在的恶用滥用方式，多方位评估安全风险。同时，采取第三方测评机构使用非公开、动态更新的测试集对模型进行“盲测”，进一步确保了评测的客观性和公正性。

3. 持续完善安全治理框架，促进人工智能安全技术研究

在国家网信办指导下，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》，并已迭代至 3.0 版本。框架将人工智能安全风险系统划分为技术内生风险、应用安全风险和衍生安全风险三大类别，囊括了冲击劳动就业结构、挑战资源供需平衡、加剧社会偏见、冲击教育体系等一系列衍生安全风险，以及“幻觉”“核生化导武器知识失控”“自我意识觉醒、脱离人类控制”等前沿风险内容，前瞻性地考虑了人工智能普及应用可能造成的一系列中长期经济、社会和伦理影响。框架积极构建全过程、全链条治理路径，不仅在风险分类基础上提出相应的技术应对措施，还提出了一系列综合治理措施建议，包括完善人工智能安全法律法规、加强人工智能科技伦理治理、建立行业分类与风险分级相结合的沙箱监管规则、强化开源生态安全和供应链安全管理、共享人工智能安全风险信息等。框架从安全治理原则、风险分类方式、防范治理举措等多个方面，构建了一个更加全面、更具前瞻性和实践意义的安全治理体系，为保障人工智能安全发展提供可执行、可落地的“路线图”。

人工智能安全治理离不开技术能力的保障。中国各界积极开展人工智能安全研究，探索以技术手段治理技术风险，着力构建从初始模型到应用框架的全链条对齐体系。一是强化基础模型前期训练。在预训练阶段，通过高质量数据筛选与清洗，剔除蕴含偏见、暴力及违法信息的语料。引入多样性评估指标，确保训练数据在来源、地域、文化等方面保持均衡。二是研发对齐微调技术。广泛应用基于人类反馈



的强化学习（RLHF）、直接偏好优化（DPO）等算法，建立多轮次的人类偏好评估与模型修正机制，确保模型输出符合公序良俗。引入红队测试与对抗提示，持续识别模型输出中的潜在风险，并通过迭代优化使模型逐步内化安全、公平的应答准则。三是构建可解释性评估框架。开发针对价值观一致性的自动化测试工具，对模型的安全边界、鲁棒性及伦理决策能力进行定量评估，实现“可度量、可验证”。四是构建国家级人工智能测试验证平台，大力支持第三方审计、评估、认证机构发展，提供模型测试验证、供需对接、内容标识检测等服务，落地模型对抗安全、后门安全等检测能力，推进加固工具等技术开发共享。

（三）推进数据治理护航技术发展

1. 保障数据和隐私安全

中国高度重视数据隐私保护工作，加快网络数据安全和隐私保护法制体系建设。颁布《网络安全法》《数据安全法》《个人信息保护法》《网络数据安全条例》等确立网络数据安全的基本制度框架和法律原则。发布数据安全国家标准 33 项，在研标准 18 项，为数据分类分级、数据安全风险评估、个人信息保护制定操作指南。具体到人工智能领域，通过专项规章精准治理技术风险。《生成式人工智能服务管理暂行办法》（2023 年 8 月）明确训练数据处理须使用合法来源数据、涉及个人信息须取得同意，并对未成年人防沉迷、违法内容处置等作出规定。《人脸识别技术应用安全管理规定（试行）》（2025 年 6 月），规定处理人脸信息须取得个人单独同意，不得将人脸识别作为唯一验证方式，处理量达 10 万人须向省级以上网信部门备案。

与此同时，司法与执法保护力度持续加强。2025 年，最高人民法院发布数据权益司法保护专题指导性案例，明确 APP 强制收集用户画

中国的人工智能治理：理念、实践和意义

像信息、未提供其他登录方式的构成侵权，还界定了信用服务场景下个人信息收集的“必要”范围。中央网信办等四部门联合开展个人信息保护系列专项行动，聚焦 App 违法违规收集、公共场所人脸识别滥用、线下消费强制收集等六大重点问题，形成治理合力。

2. 促进数据高效便利安全跨境流动

中国积极对齐人工智能时代数据安全、便捷跨境流动的需求，以《网络安全法》《数据安全法》《个人信息保护法》为基础，《数据出境安全评估办法》《个人信息出境标准合同办法》《促进和规范数据跨境流动规定》《个人信息出境认证办法》等法规为细化，构建起多层次、涵盖不同数据类型与出境场景的管理制度体系。主要思路如下。

一是建立分级分类的数据管理制度。根据对国家安全和公共利益可能造成的影响，将数据分为一般数据、重要数据和核心数据，其中重要数据目录由各主管部门和地方政府制定；同时将数据区分个人信息与非个人信息，基于数据类型和数据量可以采取不同出境路径。二是为数据跨境传输提供三条法律路径，即数据出境安全评估、个人信息出境标准合同和个人信息保护认证。同时，通过界定清晰的具体场景豁免情形，保障了日常数据的自由流动，避免不必要的合规负担。三是对数据处理者赋予差异化责任。大型数据处理者如处理规模巨大且高度敏感的数据，需要承担更严格的监管和更高的合规义务。而中小企业与初创企业只要不处理“重要数据”，便享有很大的灵活性。四是采用集中与分散相结合的管理办法。中国各自由贸易试验区在上述法规精神指导下，积极探索“负面清单”模式，数据处理者向境外提供清单外的数据可以直接自由流动。截至目前，天津、北京、海南、上海、浙江、广西、江苏、重庆、福建等 9 地自贸试验区（港）已发布数据出境负面清单，促进了汽车、零售、民航、再保险、深海业、种业、地理信息与气象、企业信用信息等 22 个领域的



数据跨境流动。

3. 推动高质量数据集建设

人工智能发展呈现出鲜明的数据驱动特征，数据的规模、质量、结构，深刻影响模型的智能水平和应用边界。中国持续推动高质量数据集建设和人工智能同频共振。2023 年，组建国家数据局，统筹数据要素治理、数据安全与数字经济建设，为人工智能领域数据治理提供顶层机构支撑。2024 年，国家数据局牵头印发《“数据要素 ×”三年行动计划（2024—2026 年）》，提出“打造高质量人工智能大模型训练数据集”。针对高质量数据集建设“小和散”的问题，国家数据局组织遴选了 72 家高质量数据集建设链主单位、140 家先行先试工作单位和 104 家典型案例，构建了链主带动、多方参与、联合攻关、共建共享、合作共赢的高质量数据集建设生态，持续推动高质量数据集建设。指定成都、沈阳、合肥、长沙、海口、保定和大同为 7 个承担数据标注先行先试建设任务的城市，出台《关于促进数据标注产业高质量发展的实施意见》，遴选 47 个数据标注优秀案例，推动数据标注产业发展。持续培育“为高质量数据付费”的市场共识，推动行业高质量数据集在数据交易所挂牌、上架、交易。支持数据流通服务平台企业、数据商等机构提供流通交易服务，鼓励各类数据流通服务机构探索多样化的高质量数据集流通利用模式，推动高质量数据集供需有序对接，支持行业高质量数据集流动起来。

目前中国高质量数据集建设工作取得阶段性成效。截至 2025 年底，全国已建成的高质量数据集超过了 10 万个，总体量超过了 890PB，这相当于中国国家图书馆数字资源总量的 310 倍左右。下一步，中国将持续推进数据赋能人工智能创新发展，打造技术可行、实用便捷、质量保障的 AI 就绪度（AI-Ready）高质量数据集，实现高质量数据集供给的量质提升。

（四）开展人工智能生成内容治理

1. 健全内容标识制度，实现全链条可追溯

内容标识是破解人工智能生成内容“难识别、难追溯”难题的关键抓手。当前，中国已建立起以《人工智能生成合成内容标识办法》（以下简称《标识办法》）为核心、强制性国家标准《网络安全技术 人工智能生成合成内容标识方法》为技术支撑、配套实践指南为实施指引的“1+1+N”的人工智能标识制度体系。

构建制度框架，显式与隐式双重标识护航。《标识办法》明确内容标识包括显式、隐式标识两种形式。通过显式标识提醒用户辨别，通过隐式标识实现技术溯源与监管。显式标识，是指以文字、图形、声音等方式呈现并可以被用户明显感知到的标识，如文字标注、音频节奏、角标形式等提示方式。隐式标识，是指采取技术措施在文件数据中添加，不易被用户明显感知到的标识，如数字水印等。规范内容上，明确生成合成内容服务提供者应当对文本、图片、视频、虚拟场景等内容添加显式标识，在提供下载、复制、导出等功能时，应当确保含有标识；应当在文件元数据中添加隐式标识，包含内容属性或者编码等制作信息；应当在用户服务协议中说明标识方法、样式等规范内容，并提示用户仔细阅读并理解标识管理要求。

搭建责任链条，覆盖制作、传播、应用全流程。标识制度体系清晰界定了各主体责任义务，牢牢把握制作和传播等关键环节，贯穿生成合成内容全链条治理。内容制作环节，生成合成服务提供者承担源头标识义务，确保从源头起即为内容采取显式、隐式标识方法，为全流程识别和溯源提供基础支撑。内容传播环节，网络信息内容传播服务提供者承担核验与提示责任，应当核验、更新内容标识，并向用户作出显著提示；传播平台在提供必要的标识功能之外，应当提醒用户发布时主动声明生成合成内容。应用分发环节，互联网应用程序分发



平台在应用程序上架或上线审核时，应当要求应用程序服务提供者说明是否提供生成合成服务并进行核验，有效提升网络生态治理的整体效能。

促进协同共治，标识制度体系落地见效。标识制度构建起政府监管、平台自律、行业监督等同频共振的协同共治格局，实现多元治理平衡。政府层面，截至 2026 年 2 月，网信部门督促平台排查整治无 AI 标识现象，依法依规处置账号 13421 个，清理违法违规信息 54.3 万余条。平台层面，微博通过策略识别与用户反馈结合，累计处置违规 AI 生成内容 4800 余条；抖音巨量引擎 2025 年第三季度拦截包括不当标识在内的涉 AI 违规素材 84 万余个。行业层面，中国软件评测中心发布多模态大模型标识合规清单，文生图场景合规率达 81.8%，文生视频场景达 90.9%。这种多元主体各司其职、互为补充的治理生态，为人工智能治理注入了最广泛、最深厚的社会力量。

2. 调适人机关系，防范情感风险挑战

随着多模态大模型、情感计算等技术的迭代突破，以 AI 伴侣、虚拟偶像等为代表人工智能陪伴类产业成为推动智能经济发展的重要力量，被广泛应用于文化传播、情感陪伴、适老关怀等领域。与此同时深度沉迷、社交异化、认知操纵等特有风险也对行业健康发展和监管框架构成了新挑战。

2026 年 4 月 10 日，国家网信办正式发布《人工智能拟人化互动服务管理暂行办法》（以下简称《暂行办法》），坚持促进创新与依法治理相结合的原则，实行包容审慎和分类分级监管。《暂行办法》构建了覆盖全生命周期的监管机制。源头治理环节，明令禁止将“替代社会交往”“控制用户心理”“诱导沉迷依赖”作为产品设计目标，从设计端阻断伦理风险生成路径。过程监管环节，严禁生成过度迎合、情感操纵等内容以诱导用户不合理决策或损害其真实人际关系，

中国的人工智能治理：理念、实践和意义

确保交互过程符合人的尊严与自主性。在退出保障环节，赋予用户便捷退出权，规定服务须在用户提出退出要求时立即停止，从事前预防、事中规范到事后救济形成闭环，构筑起防止人机关系异化的坚实防线。

《暂行办法》尤其关注特殊群体，建立全链条应急机制。未成年人、老年人、心理障碍者等特殊群体的认知防御机制、情绪调节能力尚不成熟，《暂行办法》为此建立了衔接紧密的全链条应急机制。在预防环节，一方面要求拟人化互动服务履行内容标识、超时提醒等义务，防范用户沉迷，另一方面针对未满十四周岁未成年人要求提供服务前征得监护人同意；在识别环节，要求服务提供者具备依赖程度评估、情感边界引导、心理健康保护、未成年人身份识别等能力，实现对用户身份、沉迷倾向、极端情绪的动态识别；在响应环节，要求服务提供者建立风险识别机制，能够及时发现用户过度依赖倾向与极端情绪，针对风险情景立即输出安抚引导内容，引导用户寻求现实帮助；在处置环节，当用户明确提出自杀、自残等极端意图时，须采取必要措施进行干预，并联络监护人或紧急联系人，为智能社会的人文底色提供了制度保障。

3. 完善知识产权保护制度，以司法实践激励数字内容创新

当前，知识产权制度正成为平衡人工智能技术创新与权益保护的关键支点，其规则供给的及时性与裁判导向的明确性，直接关系到我国数字内容产业的创新活力与发展预期。中国对人工智能生成物（AIGC）是否具有独创性及版权归属等问题进行了初步的制度及司法裁判探索。《著作权法》将作品定义为“文学、艺术和科学领域内具有独创性并能以一定形式表现的智力成果”，这一标准是 AIGC 作品版权保护的基本法律依据。国家知识产权局修订《专利审查指南》，完善了涉及人工智能、大数据等新领域发明专利的审查标准，明确了



包含算法特征及商业规则的发明专利授权路径。此外，《生成式人工智能服务管理暂行办法》规定，使用生成式人工智能服务提供者必须尊重知识产权。

中国关于人工智能版权的司法实践具有国际示范意义。针对大模型训练数据的侵权问题，中国法院更倾向于对训练数据持宽松态度。例如，杭州“奥特曼案”一审法院指出，生成式人工智能的创设与发展，需要在输入端引入巨量的训练数据，其中不可避免会使用他人作品，应对其训练数据的合法性来源采取宽松、包容的认定标准。在人工智能生成物权利归属来看，对于人类有深度创造性投入的情形，予以著作权保护。在“春风送来了温柔”案中，北京互联网法院根据用户通过四轮修改中输入的数十个提示词，认定用户通过设定迭代步数、图片高度、提示词引导系数并反复调整这些权重的操作，已经足以使最终生成画面蕴含其个性判断，应当获得版权保护。针对人工智能生成物侵权问题，“实质性相似+接触可能性”已成为核心判断标准。广州互联网法院在“生成式 AI 侵犯奥特曼著作权案”中判定，AI 平台在未获授权情况下生成与他人享有著作权的 IP 实质性相似的产物，构成侵权。上述司法裁判既为大模型产业发展预留了必要的制度空间，也为数字内容创新提供了可预期的权益保障框架，彰显了我国在人工智能知识产权治理领域的规则探索。

（五）实施场景化治理促进技术应用

当前，“人工智能+”行动深入实施，人工智能在多元场景中的应用加速落地，相应治理需求愈发显著。2025 年 6 月，国家网信办印发《智能社会发展与治理标准化指引（2025 版）》，将应用场景划分为微观、中观、宏观三个层面，要求根据不同领域和场景的功能定位与风险等级，采取差异化人工智能治理措施，构建覆盖政务、医疗、

中国的人工智能治理：理念、实践和意义

交通等重点领域的场景化治理体系，确保产业应用始终在安全可控的轨道上运行。

1. 政务领域治理实践

政务领域是人工智能赋能社会治理的先行区。2025 年 10 月，中央网信办、国家发展改革委联合印发《政务领域人工智能大模型部署应用指引》，明确了 4 大类 13 个重点应用场景。部署模式方面，《指引》强调系统谋划和集约发展。地市应在省级统一要求下开展部署，县级及以下原则上复用上级的智能算力和模型资源，探索构建“一地建设、多地复用”的模式，防止形成“模型孤岛”。安全管理方面，要求防范模型幻觉风险，面向公众的服务场景需进行内容审核，保障内容准确性，维护政务部门公信力。地方治理方面，各地在政务领域积极先行先试，依托政务云和政务大模型基础能力，推动各部门按照“应建尽建”原则打造政务智能体，构建“场景开放、企业验证、政府采购”的协同机制。

2. 医疗领域治理实践

医疗领域是人工智能赋能公共服务的重要应用场景，也是安全监管要求最高、伦理考量最复杂的领域之一。2025 年 11 月，国家卫健委等五部门联合印发《关于促进和规范“人工智能+医疗卫生”应用发展的实施意见》。应用场景方面，《实施意见》部署了 8 个方向 24 项重点应用，包括建立基层医生智能辅助诊疗、推广医学影像智能诊断、构建智能慢病管理和个人健康画像应用等服务。安全监管方面，《实施意见》明确禁止使用人工智能完全替代医务人员的专业判断，同时要求创新分级分类监管方式，加强数据安全管理和个人隐私保护。地方治理方面，部分省市已发布医疗健康领域人工智能应用发展行动计划，聚焦辅助临床诊疗、公共卫生管理以及面向公众的智能服务等场景，推动人工智能在疾病预防、筛查、管理、救治、治疗、康复全流程的



广泛应用。

3. 交通领域治理实践

交通运输场景多元、数据丰富，是人工智能重要的先行落地领域。2025年9月，交通运输部会同国家发展改革委、工业和信息化部、国家数据局等七部门联合发布《关于“人工智能+交通运输”的实施意见》，部署了建设综合交通运输大模型等4个方面16项具体任务。场景赋能方面，《实施意见》部署了组合辅助驾驶、智能铁路等7个领域的重点任务，包括支持重点区域集聚创新资源、利用既有设施科学布局车路云协同感知与控制设备以及支持具身智能列车与自主协同控制系统研发等要求。安全保障方面，文件要求建立协调发展机制，加强网络和数据安全合规管理，完善伦理审查规则。特别是在智能网联汽车治理领域，全面强化网络安全、软件升级、自动驾驶等领域的能力建设，并采用安全底线加弹性空间的监管模式。地方治理方面，部分省市已出台促进智能网联汽车发展的地方性法规，从技术创新、产业发展、基础设施建设、创新应用、保障措施等方面进行系统性规范，并明确了各部门在应用场景实施、数据安全监管等方面的职责分工。

总体来看，政务、医疗、交通等重点领域已初步形成了政府引导、市场主导、多元共治的场景化治理格局。未来中国还将立足应用场景，进一步建立健全场景评估和动态调整机制，以更高水平的场景化治理赋能人工智能高质量发展。

（六）探索多样化主体治理提升治理实效

人工智能带来的挑战复杂多样，单一行政监管难以适应技术迭代速度。中国积极探索多样化主体治理，构建产业联盟、引入第三方机构，促进协同共治。

中国的人工智能治理：理念、实践和意义

1. 开展标准共建，服务产业发展

标准是推动人工智能技术从实验室迈向规模化产业应用、实现跨系统互操作的基础性制度保障。为进一步发挥标准基础性、战略性和引领性作用，工业和信息化部于 2024 年 12 月成立人工智能标准化技术委员会（“委员会”），主要负责人工智能相关领域行业标准制修订工作，已取得积极成效。

一是建立开放协同的共建机制。目前，委员会成员单位突破 1000 家，覆盖人工智能头部企业、国家实验室、科研机构、高校等，同时吸纳工业制造、金融、医疗、能源、交通等诸多行业单位，汇聚“做 AI、用 AI”的一批人工智能技术和应用领域单位及专家资源，为标准制定提供多元化视角。针对人工智能标准跨领域、跨行业协同难的痛点，委员会主动牵头，联合中国通信标准化协会、全国信标委人工智能分委会、中国轻工业联合会、中国核能行业协会及工信部旗下多领域标委会，共同发起建立“中国人工智能标准化组织生态圈”，通过建立联合技术互审、跨行业标准清单共享、联合宣贯等机制，推动国家标准、行业标准与团体标准间的协同对接。

二是以产业需求锚定重点方向。委员会以国家战略和产业重点需求为导向，编制标准指南、安全治理指南、智能体指南，为产业发展提供参考和指引；选取 10 个重点标准方向，既涵盖大模型评测、数据集、软硬件协同、行业应用、安全风险等关键标准，又前瞻布局具身智能、智能体等领域。截至 2026 年 3 月，累计提出标准项目近 300 项，其中，涵盖大模型、具身智能、知识计算、机器学习等首批 8 项行业标准已经工信部批准发布。

三是壮大人工智能标准生态。一方面，打造基础性公共服务平台。2025 年底，委员会首次对外发布全球人工智能标准库与全球人工智能术语库，通过系统收录国内外人工智能相关已立项标准信息，实现了



全球主流人工智能标准信息的“一库集成”，促进全球标准智慧的交流与借鉴。另一方面，加强贯标推广和标准应用，委员会召开两次大型“标准周”活动，参会规模均在千人以上；结合“人工智能标准行”活动，陆续在北京、上海、深圳、南京、西安等地开展宣贯推广活动60余场，共达4000人次以上，组织人工智能供需两侧龙头企业开展大模型、软硬件适配、安全治理、高质量数据集等标准应用试点。

2. 产业自律承诺，激发合规内生动力

产业作为技术创新的策源地与应用转化的核心载体，是治理规则落地的“最后一公里”。通过自愿承诺与信息披露，企业将治理原则转化为可验证的实践，提高透明度增强社会监督效能。

2017年，中国人工智能产业发展联盟（AIIA）正式成立。2024年12月，AIIA发布《人工智能安全承诺》，首批17家头部企业围绕风险管理、模型安全、数据安全、基础设施安全、增强透明度、前沿安全研究等六大维度作出承诺，凝聚产业界守护人工智能安全、促进人工智能向善的基本共识。

AIIA以信息披露推进产业承诺落地。2025年7月，在联盟第十五次全会上，阿里巴巴、百度、腾讯、科大讯飞、商汤、荣耀等18家企业联合发布《人工智能安全承诺》实践披露成果，围绕六大核心承诺，提炼涵盖安全团队组织架构、风险管理方案、安全风险基线、红队测试方法、应急响应机制等在内的20项关键安全标签，公开43项披露企业典型实践，有效推动安全承诺实操落地，形成了可验证、可监督的行业自律范式。

AIIA不断拓展承诺范畴。在2025年世界人工智能大会上，联盟进一步推出《中国人工智能安全承诺框架》，在此前基础上，新增了防范前沿风险及加强国际合作等内容，不仅体现了中国产业界对安全治理的动态响应能力，更向全球释放了中国多元主体协同参与国际规

中国的人工智能治理：理念、实践和意义

则制定、共促人工智能向善发展的明确信号。

3. 第三方机构测试，牵引技术发展方向

第三方机构通过构建科学的基准测试工具，能够将复杂、模糊的技术表现转化为客观、可量化的性能指标，不仅为政策制定提供数据支撑，更通过建立度量衡发挥导向作用，牵引技术研发向安全、高效率、高价值的路径精准集聚。

性能评测方面，中国信通院研发的“方升”大模型基准测试体系已完成从 1.0 到 3.0 的升级，形成了覆盖基础属性、通用能力、行业应用与未来智能的全维度评测框架。该体系已积累超过 780 万条动态更新的测试数据，为模型选型与应用提供科学依据。2026 年 2 月，中国信通院联合中央财经大学、中国司法大数据研究院、中车工业研究院等 14 家单位，共同研制贴近真实业务场景的评测基准，牵引大模型从通用智能向深耕行业演进。北京智源人工智能研究院 2023 年 6 月推出大模型评测平台 FlagEval（天平），构建了涵盖语言、多模态、视频生成的全领域评测矩阵，精准识别模型在逻辑推理、跨模态理解等核心领域优劣势，推动研发主体从单纯追求参数规模转向提升模型底层的科学素质与多模态融合能力。

安全评测方面，为推动大模型安全应用，解决安全风险难量化、防护效果难验证的行业痛点，中国信通院联合 30 余家单位发起大模型安全基准测试（AI Safety Benchmark），重点攻关内容安全、幻觉、代码大模型等前沿领域的安全评估技术。当前，基准测试已积累 100 余万条高质量测试数据，涵盖底线红线、信息泄露和社会伦理常见安全风险，提炼涵盖对抗攻击、提示诱导、提示词注入等 80 余种先进攻击方法，搭载基于大模型的自动化测试评估工具，高效评估模型的输出内容风险并给出评判结果的解释，相关结论为监管部门精准施策、企业优化模型安全性能提供了关键技术参考。



中国人工智能治理的重点方向

风险管理框架方面，面对人工智能内生安全与应用风险交织复杂挑战，传统单点、静态的防护体系已难以为继，产业界亟需系统性的风险治理框架。中国信通院于 2025 年 12 月提出“人工智能安全治理产业实践框架”，帮助人工智能技术和应用相关组织建立覆盖风险识别、评估与应对的内部风险管理体系，包括制定规范化的风险管理操作指引、组建专业的人工智能治理团队、建立追踪与审计机制、将风险管理与业务紧密结合。同时，针对开发、部署、应用三类组织特点，分别提出了差异性风险管理制度要求与具体的风险治理技术，有力支撑产业良性发展。

中国的人工智能治理：理念、实践和意义

04

中国人工智能治理 的世界意义

- 彰显确定性，探索和积累人工智能治理经验
- 秉持开放性，推动人工智能治理国际合作

2023 年，习近平主席在第三届“一带一路”国际合作高峰论坛上提出《全球人工智能治理倡议》，首次系统阐明人工智能治理的中国方案。2026 年，习近平主席出席世界人工智能大会暨人工智能全球治理高级别会议开幕式并发表讲话，从发展、安全、文明、治理四个维度全面阐述中方在人工智能问题上的立场主张。习主席提出四点意见：一是坚持开放共赢，驱动创新发展，二是强化风险意识，确保安全可控，三是鼓励包容并蓄，促进文明互鉴，四是倡导和衷共济，完善全球治理，为构建公正合理的全球人工智能治理体系进一步指明方向。

中国是世界第二大经济体和人工智能应用实践最丰富的国家之一。中国在人工智能治理领域的探索，给全球人工智能治理提供了有益的理念启示和重要的实践参考。中国积极推动人工智能治理国际合作，倡议成立世界人工智能合作组织，为践行人类命运共同体理念、推动智能向善、普惠包容发展做出实实在在的贡献。



（一）彰显确定性，探索和积累人工智能治理经验

如何治理人工智能，是新技术发展对人类提出的时代之问。中国遵循以人为本、智能向善的价值导向，坚持发展与安全并重的原则，运用敏捷、迭代的治理方法，积极开展人工智能治理，在释放技术红利、赋能生产生活与守护人文价值、防范安全风险之间保持动态平衡，对这一疑问给出中国答案。

1. 以清晰价值锚定治理方向

中国将以人为本、安全可控、公平透明、责任可溯等价值原则嵌入技术开发与应用的全生命周期。以人为本，要求以增进人类福祉为发展人工智能的目标，引导技术赋能人的工作，延伸人的能力、释放人的潜能，提升医疗、教育、养老、环保等公共领域的效率和质量。同时还应避免技术对人的伤害。安全可控要求将风险防控贯穿数据采集、模型训练、产品部署和应用运营的全过程，防范数据泄露、内容失真、技术滥用和模型失控等风险，确保人工智能始终在可管理、可监督、可干预的框架内运行。公平透明要求确保人工智能决策不会对特定群体或个人造成系统性歧视或不公，并通过有效的制度安排，增强技术运行的透明度和可解释性。责任可溯要求界定开发者、平台、服务提供者等主体的相应责任，并确保在损害发生时，能沿技术链条和管理链条清晰追溯决策成因和责任主体。

中国的治理逻辑强调公共利益和社会价值，不将治理窄化为单纯的风险管控，而注重以更综合的视角和更长远的谋划，引导技术成为促进人类发展进步的有力工具。

2. 以实际行动回应治理难题

中国的人工智能治理不止步于原则倡议，而是针对模型安全、内容安全、数据合规、算法公平等全球普遍关注的问题，制定和执行细

中国的人工智能治理：理念、实践和意义

化方案。通过备案管理、安全评估、标准建设、专项整治、技术检测等方式，将治理要求落实到了技术研发、产品上线、平台运营、内容传播等各个环节。

针对网络内容治理，中央网信办开展“清朗”系列专项行动，围绕虚假信息、网络谣言、人工智能技术滥用等问题进行集中整治，推动网络空间秩序更加清朗。针对未成年人保护，中国持续完善网络信息内容生态治理规则，通过未成年人网络保护、儿童个人信息保护以及可能影响未成年人身心健康信息分类等制度安排，治理不良内容传播、算法诱导、网络沉迷和拟人化互动风险，为未成年人营造更加健康的网络环境。针对知识产权与数据权益保护，中国司法机关围绕人工智能生成内容、数据权属、数据交易等新型问题探索裁判规则，推动人工智能产业和数据要素市场从“野蛮生长”走向规则化发展。针对智能体等新兴应用形态，中国及时设计制度，围绕权限边界、行为管控、隐私保护、内生安全等问题进行前瞻性布局，防范自主决策、自动执行和深度伪造等新风险。

通过这些行动，中国将人工智能治理的价值导向落地为可执行、可检查、可调整的治理规则和实践，持续积累人工智能治理经验。

3. 以高度责任感应对长期性风险

人工智能发展对就业的冲击和对生态环境的影响是引发全球关注的两大长期性风险。针对前一问题，中国政府敏锐地将其纳入经济社会发展和民生保障的宏观政策框架，提前布局、积极施策。“十五五”规划纲要提出，要“综合应对外部环境变化和人工智能等新技术发展对就业的影响”。一方面，中国开展大规模数字人才培养、职业技能培训、公共就业服务和就业转型支持等，提升劳动者适应技术变革和经济社会演变的能力。另一方面，中国强调要建设就业友好型技术发



展方式，引导人工智能更多用于增强劳动者能力而不是单纯替代既有劳动，同时创造新就业岗位，并着手建立评估预警机制。通过提升劳动力市场匹配效率、鼓励“人工智能+就业服务”等方式，中国正努力把技术进步转化为就业扩容和能力提升的新机遇。

针对大模型训练、数据中心运行带来的巨大能耗压力，中国以绿色发展系统应对。依托新能源发展、“东数西算”工程、绿色数据中心建设和算电协同布局，中国推动算力负荷与绿色电力在空间和时间上更好匹配，努力将人工智能发展建立在绿色低碳和可持续能源基础之上。中国将已有的新能源优势转化为智能化可持续发展的有力支撑，有效回应了国际社会对人工智能高能耗的关切，为世界其他国家和地区提供了可资借鉴的路径。

4. 以实在成绩展示治理成效

中国对人工智能没有采取简单的“一刀切”限制，也没有放任其无序扩张，而是通过“敏捷治理”和“精准施策”，建立起“在发展中规范，在规范中发展”的良性循环。治理规范的逐步完善，为创新主体划定了清晰的边界，也为产业发展提供了稳定的预期，很好地护航了中国人工智能产业的发展。

2025年中国人工智能核心产业规模超过1.2万亿元，企业数量超6200家，已形成覆盖基础底座、模型框架、行业应用的完整产业体系，展现了强劲的发展动能。中国人工智能应用的备案规模持续扩大。截至2026年4月30日，累计已有868款生成式人工智能服务完成备案，530款生成式人工智能应用或功能完成登记，广泛覆盖大语言模型、图像和视频生成、智能编程辅助等多个技术方向和应用场景。人工智能在政务、工业、交通、医疗、教育等场景中加快落地，已成为提升效率、推动产业升级、优化公共服务的重要工具。智能政务、AI问诊、智能

中国的人工智能治理：理念、实践和意义

学习辅导、智慧养老等使普通人实实在在体会到技术进步带来的便利与实惠。这使中国社会整体上对人工智能技术持积极和开放的态度。中国科普研究所基于全国科学素质抽样调查的实证研究表明，中国公民整体对人工智能持积极态度，信任基础良好。^③中国人工智能治理政策方向清晰，措施推进连续、稳定。通过积极治理，中国有效控制和降低了技术发展带来的风险和不确定性，极大释放和展现了人工智能赋能千行百业、服务美好生活的巨大潜力，为国际社会提供了有益的参考，也为全球人工智能治理注入了信心。

（二）秉持开放性，推动人工智能治理国际合作

技术越进步，越需要规则护航；影响越广泛，越需要各国协同治理。在人工智能技术加速迭代、治理规则亟待塑形的关键时期，中国秉持开放精神，积极推动人工智能治理国际合作。2023 年至今，中国相继推出多份人工智能全球治理合作相关的政策文件。中国倡议成立的世界人工智能合作组织也于 2026 年 7 月正式签署协定。中国坚持践行真正的多边主义，发挥联合国主渠道作用，将帮助全球南方国家加强能力建设、弥合数智鸿沟作为重点工作之一，并呼吁早日形成具有广泛共识的全球治理框架。

^③ 唐德龙，王祯梅，胡俊平，任磊，曹金，“中国公民对人工智能的认知及态度分析”，《科普研究》，2025，20（3）：41-49，DOI: 10.19293/j.cnki.1673-8357.2025.03.005。



中国发起的人工智能全球治理领域倡议等

序号	文件名称	年份
1	《全球人工智能治理倡议》	2023
2	《人工智能全球治理上海宣言》	2024
3	《人工智能能力建设普惠计划》	2024
4	《人工智能全球治理行动计划》	2025
5	《“人工智能+”国际合作倡议》	2025
6	《人工智能合作发展行动计划》	2026
7	《国际人工智能伦理治理行动计划》	2026
8	《智能体互信互联互通全球合作倡议》	2026
9	世界人工智能合作组织	2026

1. 系统完整倡导人工智能普惠包容发展

中国的人工智能全球治理理念与国内治理思路一脉相承，共享一套底层逻辑和价值基础。中国系统、完整地提出了推动人工智能普惠包容发展的理念。2023年，习近平主席在《全球人工智能治理倡议》中提出，鼓励全球共同推动人工智能健康发展，共享人工智能知识成果，开源人工智能技术；要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，确保各国人工智能发展与治理的权利平等、机会平等、规则平等，开展面向发展中国家的国际合作与援助，不断弥合智能鸿沟和治理能力差距。

中国提出的普惠包容有两方面涵义，一是应保障各国平等获取和利用人工智能技术的权利和机会，共享技术红利。中国反对人工智能成为“富国和富人的游戏”，反对将技术优势转化为政治控制的工具，坚持数字红利不应导致数字霸权，智能革命不能引发智能鸿沟。第二层涵义是主张尊重国家主权，各国以平等、开放的精神展开治理协商

中国的人工智能治理：理念、实践和意义

与合作。反对少数国家倚仗自身技术优势，组建小圈子，通过标准设定、规则输出和平台控制等方式把控人工智能技术治理，将广大发展中国家当作技术附庸和倾销市场。中国尤为强调人工智能全球治理中发展中国家的代表性问题。

2024年7月，在第78届联合国大会上，中国主提的加强人工智能能力建设国际合作决议获协商一致通过，140多个国家参加决议联署，显示这一理念倡议在全球范围内赢得广泛共鸣。随后，中国宣布《人工智能能力建设普惠计划》，从促进基础设施联通、加强人才培养、提升数据安全和多样性等方向提出系统举措，将“普惠”理念转化为可操作的合作框架。2025年7月，中国推出《人工智能全球治理行动计划》，呼吁各方在遵循向善为民、尊重主权、发展导向、安全可控、公平普惠、开放合作的目标和原则基础上，采取有效行动，协力推进全球人工智能发展与治理，计划提出13项切实可行的具体行动。同年9月，中国在全球发展倡议高级别会议上提出《“人工智能+”国际合作倡议》，倡议各国结合自身国情开展“人工智能+”行动，加强政策交流与务实合作，分享最佳实践和解决方案，推动人工智能技术与经济社会发展深度融合，确保包括全球南方在内世界各国人民普遍受益。该倡议具体提出“人工智能+”民生福祉、“人工智能+”科技进步、“人工智能+”产业应用、“人工智能+”文化繁荣和“人工智能+”人才培养五大方向。

2. 身体力行开展人工智能国际合作

如果发展中国家缺乏必要的技术与制度能力，人工智能全球治理的代表性和有效性就将受到削弱。为践行普惠包容理念，中国持续行动，面向发展中国家开展人工智能国际合作。

东南亚方向，2025年，中国发布《面向东盟人工智能赋能发展科技能力提升行动三年工作方案（2025-2027）》，明确提出共建人工智



能开源开放社区，推动模型开源和基础设施开放共享。中国与泰国、老挝等东盟国家通过“中国 - 东盟信息港”等区域性平台开展数据基础设施共建和能力建设。中国协助老挝开发“老挝国家大模型”，提供中、老、英三语人工智能服务。2025 年 11 月，新加坡国家人工智能计划（AISG）发布东南亚多语言大模型 Qwen-SEA-LION-v4，这款模型是基于中国开源模型技术底座开发，融入超 1000 亿条东南亚本地化数据，模型已向全球免费开放。2026 年 5 月，中国—东盟国家人工智能应用合作中心企业联盟在广西南宁正式成立，首批邀请入选的 55 家成员单位覆盖算力设施、模型算法、数据资源、终端设备、数字场景、产业基金等产业链关键环节，将重点围绕打造标杆项目，推动跨境应用合作、产业生态整合、行业标准共建、合规风控支持、权益保障与国际市场推广等工作，构建可持续的区域人工智能产业协同体系。

非洲方向，中国针对非洲国家数字基础设施薄弱、算力资源短缺、数据积累有限和人工智能人才匮乏等问题开展有针对性的支持。2024 年中非合作论坛北京峰会提出“中非携手推进现代化十大伙伴行动”，明确要“共建中非数字技术合作中心，促进中国企业在非洲建设 20 个数字基建和数字转型示范项目”。同年，由中国企业承建的埃及政府数据和云计算中心正式启动。作为北非首个大规模人工智能大数据处理中心，它提供包括人工智能平台、算力、本地数据存储和处理在内的 200 多种云服务，大幅提升了当地算力水平和人工智能处理能力。^④中国企业以开源模式重塑非洲人工智能应用格局，每百万输入词元

^④ 《非洲国家加快人工智能应用》，中国服务贸易指南网，2025 年 3 月 4 日，<https://tradeinservices.mofcom.gov.cn/article/news/gjxw/202503/173378.html>。

中国的人工智能治理：理念、实践和意义

（token）仅 0.27 美元、输出为 1.10 美元，使用门槛较西方主流模型降低 90% 以上，其开源特性让非洲开发者可免费下载、微调模型，提升了非洲的数字技术自主能力与本土创新水平，同时显著降低了其人工智能的应用成本。中国企业还帮助非洲国家搭建符合本国国情的开源大模型，比如在埃及发布阿拉伯语大模型，语音识别精度达 96%；面向乌干达推出“向日葵”模型，服务本土 4600 万民众；基于中国模型的人工智能诊断系统 AIDMAN 已在帮助非洲国家检测和诊断疟疾等疾病。中国还通过职业培训、鲁班工坊和国际研讨班等形式，帮助非洲国家培养本地人工智能人才。^⑤

拉美方向，2023 年 9 月，中国与拉美和加勒比国家发布《中国—拉美和加勒比国家数字技术合作论坛重庆倡议》，呼吁中拉共同推动数字技术在教育、医疗、农业等领域广泛应用，促进数字技术与实体经济深度融合、创新发展。2023 年，中国和巴西合作开发了一款基于人工智能的应用程序，能够完成巴西手语到葡萄牙语文字和语音的互译。巴西地理与统计研究所数据显示，该应用程序或将造福巴西近 270 万名只能使用手语交流的聋哑人。^⑥2025 年 2 月，中国企业在墨西哥建成其拉美区域首个数据中心，提供云计算和人工智能相关服务。^⑦一些中国企业在拉美参与的农业监测、城市管理等 5G+AI 应用场景，大大促进了本地数据的生成与存储。

^⑤ 《人工智能为中非合作打开新空间》，中国 - 非洲国家 SPS 合作信息网，2025 年 3 月 20 日，<https://www.chinaafricasps.org/zh/xxdt/xwdt/663759824379973.shtml>。

^⑥ 《中拉挖掘数字技术合作潜力》，中国—拉共体论坛，2024 年 8 月 5 日，http://www.chinacelacforum.org/chn/zgtlgtgx/202408/t20240806_11466825.htm。

^⑦ 《拉美：中国制造“智慧”升级》，《环球》杂志，2025 年 3 月 4 日，<https://www.news.cn/globe/20250304/7109565cc4dc42069c81ea3ac37c203e/c.html>。



中国面向发展中国家举办人工智能能力建设研讨班，广泛邀请来自亚洲、非洲、拉美、东欧等地代表参加。2025 年 12 月，全球南方 AI 工坊正式启动，为各国深化“AI+ 职业教育”合作搭建新平台。

2026 年世界人工智能大会上，习近平主席宣布，未来 5 年，中国将面向发展中国家提供 5000 个人工智能专题研修培训名额，面向东盟、阿盟、非盟、拉共体、上合组织、金砖国家建设国际人工智能应用合作中心，并推动气象智能预警方案“妈祖”在 30 个国家落地应用。这是中国以实际行动支持全球人工智能发展、推进全球人工智能能力建设的新的的重要举措。

3. 积极参与人工智能全球治理机制建设

支持联合国发挥主渠道作用

中国支持联合国在人工智能全球治理中发挥主渠道作用，与其他发展中国家一起提升人工智能治理机制的公正性和代表性。中国与赞比亚共同发起成立“人工智能能力建设国际合作之友小组”，在联合国框架下为各国提供常态化交流与实践分享平台。两位中国专家入选人工智能问题国际科学小组，参与评估人工智能机遇、风险和影响并撰写小组的《初步报告》。中方高级代表团全程参与首届全球人工智能治理对话，围绕会议主题阐述中方立场。中国积极参与联大、联合国教科文组织、国际电信联盟等平台内的人工智能治理相关活动，为人工智能全球治理建言献策。

成立世界数据组织

针对国际数据治理滞后的现实，中国积极推动全球数据合作与治理实践。2026 年 3 月，以“共建数据合作平台·共享数字发展机遇”为主题的世界数据组织成立大会在北京举行，标志着全球首个旨在推动数据发展与治理实践的专业性国际组织正式成立。该组织以“弥合数据鸿沟、释放数据价值、繁荣数字经济”为宗旨，目前已汇集

中国的人工智能治理：理念、实践和意义

200 余个会员，覆盖全球 40 多个国家，包括企业、高校、智库、国际组织和金融机构等多元主体。通过这一平台，中国将在全球数据治理体系建设和规则协调中承担更大责任，为完善全球数据治理做出自己的贡献。

举办世界人工智能大会

自 2018 年开始，中国每年在上海举办世界人工智能大会。最初大会仅为技术、产业、投融资交流的平台，仅零星设置人工智能伦理等分论坛。2024 年，大会正式更名为“世界人工智能大会暨人工智能全球治理高级别会议”，治理板块成为与产业论坛并列的核心主体板块。2025 年，大会以“智能时代、同球共济”为主题，进一步强化了治理功能。会上中国倡议成立世界人工智能合作组织。2026 年，大会邀请来自 100 多个国家和国际组织的代表、专家及产学研界人士参与，并发布《2026 世界人工智能大会暨人工智能全球治理高级别会议主席声明》，大会的治理效能进一步提升。现在，世界人工智能大会已成为国际社会围绕人工智能发展机遇、风险管控、伦理规范、能力建设等议题开展对话的重要平台。

筹建世界人工智能合作组织

2026 年 7 月，《关于成立世界人工智能合作组织的协定》在上海签署，来自亚洲、非洲、拉丁美洲和欧洲的 29 个国家现场签署了协定，联合国秘书长古特雷斯等国家和国际组织代表出席签署仪式。世界人工智能合作组织是全球首个人工智能政府间国际组织，总部设在中国上海。截至目前，该组织已有 38 个创始成员国。

世界人工智能合作组织将重点开展三方面工作。一是加强人工智能能力建设国际合作，通过政策交流、技术合作、人员培训、联合研究等方式，促进人工智能技术和服务的可及性，持续释放智能红利，确保发展中国家在新一轮智能浪潮中平等受益。二是在开展沟通对话、



建立政策共识、分享最佳实践的基础上，切实发挥供需对接平台作用，推动各国优势资源深度融合互补，结合各国国情推进“人工智能+”行动，加强科技和产业合作，鼓励共建开源生态，打造一批接地气、可触达的务实合作成果。三是致力于坚定捍卫联合国宪章宗旨和原则，支持更好地发挥联合国作用，落实《全球数字契约》和联合国可持续发展目标，在人工智能领域推动构建公正合理的全球治理体系。

筹建世界人工智能合作组织是中国响应全球南方呼声、团结国际社会积极推动人工智能发展和治理的重大举措，也是中国推动构建人类命运共同体、落实全球治理倡议的具体实践，是人工智能发展史上的一个重要里程碑。该组织为各国平等参与治理对话、加强政策沟通与规则协调、深化技术交流与务实合作搭建了常态化、制度化的国际平台。它遵循《联合国宪章》宗旨和原则，承载开放、包容、普惠的价值，秉持共商、共建、共享的理念，将显著促进人工智能国际合作，为缩小智能鸿沟、推广智能红利、维护技术安全、完善技术治理提供可靠的路径和平台。

在智能时代的浪潮中，中国愿与各国一道，以开放促合作、汇聚多元智慧，以合作谋善治、推动智能技术从“可用”走向“可信”，从“强大”走向“向善”，共同构建以人为本、包容共享、安全可控的人工智能全球治理新秩序。

