CGAIG 全球人工智能创新治理中心
CENTER FOR GLOBAL AI INNOVATIVE GOVERNANCE

上海市邯郸路220号复旦大学智库楼
邮编：200433
电话：86-21-65641067
邮箱：cgaig@fudan.edu.cn
网址：https://cgaig.fudan.edu.cn

复旦智库报告
FUDAN REPORT SERIES

2026世界人工智能大会
人工智能全球治理
与可持续发展：挑战与应对 | 论 坛

CGAIG
CENTER FOR GLOBAL AI INNOVATIVE GOVERNANCE

THINK TANK FUDAN 复旦智库

全球人工智能治理趋势：《人工智能全球治理行动计划》以来的观察

姚 旭 等著

全球人工智能创新治理中心
复旦大学发展研究院

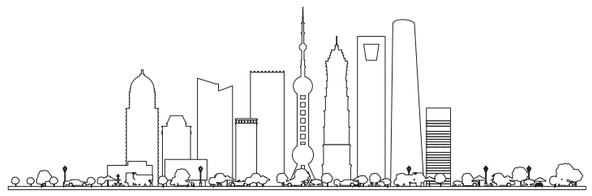
复旦智库报告

全球人工智能治理趋势：《人工智能全球治理行动计划》以来的观察

姚　旭　等著

全球人工智能创新治理中心
复旦大学发展研究院

2026 年 7 月



项目牵头人

姚　旭　全球人工智能创新治理中心秘书长

复旦大学发展研究院副研究员

项目成员

江天骄　全球人工智能创新治理中心研究员

复旦大学发展研究院副研究员

辛艳艳　全球人工智能创新治理中心副秘书长

复旦大学发展研究院助理研究员

沈　绮　全球人工智能创新治理中心项目主管

罗皓月　全球人工智能创新治理中心项目主管

张书言　全球人工智能创新治理中心研究助理

杨　昭　全球人工智能创新治理中心研究助理

张　傲　全球人工智能创新治理中心研究助理

王奕博　全球人工智能创新治理中心研究助理

袁露铭　全球人工智能创新治理中心研究助理

胡翔宇　全球人工智能创新治理中心研究助理

王叶滔　全球人工智能创新治理中心研究助理

姜珺吉　全球人工智能创新治理中心研究助理

目　录

核心内容：

自《人工智能全球治理行动计划》在 2025 年世界人工智能大会（WAIC）提出以来，人工智能全球治理正在从原则倡议进入机制承接、能力建设和项目落地的新阶段。当前，全球人工智能发展仍呈现显著不平衡：少数国家和头部企业在算力、数据、模型、平台和标准制定上持续集聚优势，广大发展中国家则普遍面临基础设施不足、技术获取成本高、本地语种支撑薄弱、治理能力有限和国际规则参与不足等制约。人工智能能否真正成为服务共同发展的技术资源，关键不仅取决于模型能力本身，更取决于各国能否公平参与治理、获得可负担的技术能力、形成本地应用生态，并具备相应的安全治理基础。

01 全球人工智能治理正在进入“发展与安全并重”的新阶段。

安全风险仍是国际社会必须守住的底线议题，但能力建设、普惠发展、产业应用和技术可及性正在成为更突出议程。未来治理不再只是讨论“如何防范风险”，也要回答“如何让更多国家真正用得上、用得起、用得好人工智能”。

02 “智能鸿沟”正在成为数字鸿沟之后新的结构性不平等。

算力、数据、算法、语言、人才和应用场景的不均衡，使全球南方在人工智能时代面临更深层的技术依赖和规则被动。如果缺乏有效合作，发展中国家可能长期停留在数据供给、市场消费和规则接受的位置，难以进入高附加值的模型开发、平台治理和标准制定环节。

03 大国竞争正在加剧全球治理碎片化。

围绕芯片、算力、基础模型、关键矿产、数据流动和技术标准的竞争，正在重塑人工智能治理的制度环境。一方面，大国竞争推动技术快速迭代；另一方面，出口管制、技术联盟和标准分化也可能压缩中小国家政策选择空间，抬高跨境合作成本，削弱全球公共产品供给效率。

04 南北合作的重点应从“技术输出”转向“能力共建”。

真正有效的人工智能合作不能停留在设备采购、平台接入或短期培训，而应围绕本地算力基础设施、数据治理能力、人才培养体系、低资源语种模型、行业应用场景和安全评估工具展开长期建设。其目标不是制造新的技术依附，而是帮助发展中国家形成可持续、自主性的人工智能发展能力。

05 中国提供人工智能国际公共产品，正在成为全球治理的重要变量。

围绕《人工智能全球治理行动计划》，中国正通过多边机制对接、开源模型供给、能力建设项目、数字基础设施合作和产业场景应用，推动人工智能技术从少数国家和企业掌握的竞争资源，转化为更多国家可获取、可参与、可转化的发展工具。这一过程不仅关系到中国自身在全球人工智能治理中的角色定位，也关系到人工智能能否真正服务共同发展。

展望未来，人工智能全球治理的关键不在于建立一套由少数国家定义的封闭规则体系，而在于形成更加开放、包容、可持续的合作网络。中国若能继续以真正的多边主义为基础，以开源开放降低技术门槛，以能力建设夯实发展基础，以场景合作回应民生需求，就有可能在缩小全球智能鸿沟、提升发展中国家治理能力、推动人工智能向善发展方面提供更具实践意义的国际公共产品。

第一部分　全球人工智能治理态势

人工智能是前所未有的技术形态，图灵奖得主、“强化学习之父”理查德·萨顿（Richard S. Sutton）提出“人工智能是宇宙演化的必然下一步，人类应以勇气、自豪和冒险精神来迎接它”。当前，人工智能已经应用于各行各业，应用场景遍布医疗、金融、交通、农业、制造业、零售业等领域。更新迭代迅速、应用场景广泛的人工智能也对全球治理提出了更高的要求。本部分梳理了 2025 年至今全球人工智能治理态势，发现全球人工智能治理体系出现诸多新进展。各国纷纷出台鼓励技术发展的政策，安全议题相对弱化。以联合国为“主渠道”的国际合作机制逐渐完善，形成联合国与区域机制协同合作的治理态势。治理路径转型升级，向着应用导向、前端化、科学化发展。然而，全球人工智能治理同样面临挑战，大国竞合成为常态，多重治理矛盾并存。面向未来，推进全球人工智能治理体系，需要在发展议题上求同存异，在安全议题上划定红线，积极发挥全球南方作用，促进国际合作再上新阶。

一、全球人工智能治理体系新进展

2025 年至今，全球人工智能治理主要有三大进展。国家治理层面，发展成为制定人工智能政策的主题。国际治理层面，联合国完善治理机制，日益发挥主渠道作用。治理路径层面，推行应用导向治理、前端化治理和科学化治理。

（一）国家政策转向

人工智能日益成为驱动国家发展的战略引擎，各国正在将人工智能提升至国家战略高度，在发展与监管平衡的大背景下，发展导向相对明显，力求在下一轮产业变革中赢得主动权。

中国从多个层面倡导人工智能能力建设。理念层面，中国坚持发展与安全统筹并举。发布《关于深入实施“人工智能 +”行动的意见》，推动人工智能与各行业各领域广泛深度融合，同步夯实法治与安全根基。技术层面，中国在“十五五”规划纲要中提出模芯云用人工智能全链条协同创新战略，形成技术研发、产业发展一体化的发展模式。制度方面，中国发布《人工智能安全治理框架（2.0）》，总结实践经验、融合新兴技术与治理思路，对治理原则、风险划分、技术手段、治理机制及指引体系进行全面革新。

美国及部分西方国家政策出现明显转向，战略重心进一步从防范技术失控风险转向技术竞争与发展。国内层面，放松监管的政策取向增强。美国在特朗普第二任期后，废除拜登政府第 14110 号行政令，试图弱化州层面监管约束，大力推进“第 14179 号行政令”、“美国人工智能行动计划”、“创世纪计划”政策文件。欧洲的情况类似，欧盟《人工智能法》定向调整逐渐程序性减负，英国也推迟通过人工智能立法，发布《人工智能机遇行动计划》。

“全球南方”国家通过主权人工智能模式寻求安全自主的发展，印度、巴西、马来西亚等有诸多战略出台。《印度人工智能使命》提出七大支柱来构建主权人工智能，正在通过 Bhashini 项目打造包含 22 种以上印度语言的数据集，鼓励训练全栈式自主人工智能平台 Sarvam AI。巴西总统卢拉多次批评科技巨头垄断人工智能。《巴西人工智能投资计划（PBIA 2024-2028）》提出，计划在四年内投资 230 亿雷亚尔（约 40 亿美元）发展人工智能。马来西亚是东南亚地区第一个构建主权全栈人工智能生态系统的国家。马来西亚企业 Skyvast Cloud 依托中国人工智能芯片和 DeepSeek 开源模型，运营人工智能基础设施。马来西亚国家级 MaaS 平台、国家主权人工智能人才培养实验室 Z·UM AI Lab 基于智谱 Z.ai 的开源基础模型进行本地优化，保持以主权为核心的数据安全架构以

及国家级人工智能平台的自主性。

各国对外人工智能政策出现差异化特征。中国从多个层面倡导全球人工智能合作。理念层面。中国聚焦人工智能开放创新发展、普惠包容发展、安全可靠发展，2025 年相继发布《人工智能全球治理行动计划》和《“人工智能 +”国际合作倡议》，推动人工智能成为造福人类的国际公共产品。技术层面。中国提出《国际人工智能开源合作倡议》，致力于以开源为纽带、携手各国打造开源生态。DeepSeek-R1 模型也以其开源、低成本、高性能的开发模式，提供新的技术路径。平台层面。中国致力于搭建“全球南方”合作平台，倡议筹建世界人工智能合作组织，作为面向全球南方的关键性人工智能国际公共合作平台，推动人工智能合作治理体系更加公平、开放。美西方层面，安全议题被淡化。巴黎人工智能行动峰会和 2026 年印度 AI 影响力峰会弱化安全主题，转向重视发展、创新和投资。英国、美国也更名本国人工智能安全研究所，从原先的为技术设置护栏，变成制定技术标准及防范国际竞争风险。同样，2025 年 6 月七国集团峰会发布了《关于人工智能促进繁荣的声明》，进一步弱化单一安全叙事。

（二）国际组织治理机制化

国际人工智能治理从多进程并行转向以联合国为核心的多边协作架构，逐渐形成联合国与区域机制协同合作的治理态势。

联合国作为治理“主渠道”的地位通过建立一系列工作机制得到巩固。首先，组织内部改革。2025 年初联合国技术特使办公室正式更名为联合国数字与新兴技术办公室，成为联合国系统内负责数字问题的协调机构，标志着人工智能治理架构将更加清晰。其次，确立专业权威。联合国建立了“人工智能问题独立国际科学小组”，其 40 名成员具有极高的地域代表性与跨学科性，这也体现出亚洲、非洲、南美洲等发展中国家日益成为人工智能研究的重要参与方。该小组旨在为全球人工智能治理提供独立科学评估。2026 年 7 月，首届联合国全球人工智能治理对话在日内瓦举行，并围绕联合国支持的 40 名专家小组相关成果展开讨论；后续更详细报告和纽约会议预计将于次年推进。此外，对话机制常态化。2025 年 9 月正式启动的“全球人工智能治理对话”成为政策沟通、交流、对齐的平台，旨在缩小全球南北技术鸿沟。同时，联大于 2025 年 12 月正式确认互联网治理论坛（IGF）为常设机构，终结了其长达 20 年的临时状态，确立了多利益相关方治理的长效性。

区域多边机制与联合国框架形成良性互动，形成区域节点对接全球中心的态势。在二十国集团，2025 年南非担任主席国期间，通过“人工智能造福非洲”倡议，将人工智能治理与可持续发展议程相结合，计划在非洲建立 5 个人工智能创新中心，利用人工智能赋能农业、医疗、教育等民生领域。同步落实联合国 2030 年可持续发展目标，推动 G20 区域共识转化为联合国框架下的全球行动。在金砖国家合作机制，金砖国家领导人通过《关于人工智能全球治理的宣言》，明确反对全球人工智能治理体系碎片化与重复，主张“联合国是人工智能全球治理的核心”。在上海合作组织，成员国元首理事会通过《天津宣言》，表示支持发挥联合国中心协调作用，还发布《关于加强数字经济发展的声明》，希望加强上合组织框架下数字经济合作。

（三）治理路径转型升级

随着全球人工智能治理实践和互动的增多，经验日益丰富，治理路径也呈现逐渐完备的趋势，体现为应用场景导向型治理、前端化治理以及科学化治理。

应用场景导向型治理是需要关注的重点。和初期针对技术与大模型算法本身的技术导向型不同，目前的人工智能治理更增加了不同领域不同层级的应用导向型管控，强调大模型在特定场景落地时，与环境、系统和人类互动产生的风险。中国采取场景化规制路径，围绕算法推荐、深度合成、生成式人工智能、拟人化互动服务及智能体安全等具体应用场景，分别出台部门规章或标准指南。欧盟以《人工智能法案》建立风险分级框架，并通过《通用目的人工智能行为准则》和《生成式人工智能透明度指南》等软法机制，推动企业提前合规并细化透明度、版权和系统性风险管理要求。人工智能治理对象出现上移。不同于早期聚焦算法歧视、隐私侵害等问题，现在人工智能安全治理关注点明显上移，聚焦于前沿模型和高能力的通用模型。在 2025 年 WAIC 上，不少与会专家在公开发言中都强调亟须关注安全治理，因为一旦模型能力跨越了某些阈值，风险将呈现非线性增长，而且这种高能力模型可能实施难以通过传统监管解决的行为。他们还强调安全问题必须在模型研发和部署的前端加以治理，而不是事后进行补救。可以说，全球在安全治理上的议程已经深入到如何在模型层面设定安全护栏，包括能力评估红队测试、阈值监测等。

人工智能治理也在提升科学化水平，尤其是人工智能安全治理的科学化可以作为最大公约数。一方面，尽管安全议题优先度有所下降，在科学家层面，仍有努力与合作，其中的标志性进展即 2025 年 1 月《国际人工智能安全报告》的发布。安全治理领域正在评测体系、风险框架，还有事故报告体系等技术工具方面推进共识，希望将其转变为全球治理的公共产品，旨在降低国家间对风险认知的分歧、为监管提供可行抓手、为国际合作提供低政治敏感性的切入点。另一方面，安全治理流程正在工程化，主要表现为标准、管理体系以及可审计的安全的出现，表明治理正在向可追溯、可问责的方向演进。

二、主要大国人工智能竞争态势

人工智能不仅关乎产业升级与经济增长，更深度嵌入国家安全、军事能力、社会治理与国际规则塑造等领域，已成为大国战略竞争的关键变量。在此背景下，大国围绕数据、算力、模型、应用、治理规范与基础设施等关键节点展开全方位博弈，为全球人工智能治理态势注入新变量。

（一）大国人工智能竞争的核心博弈点

第一，数据治理。在生成式人工智能蓬勃发展的背景下，大国之间的数据竞争正由“规模多寡”转向“治理能力优劣”。合规采集、清洗标注、授权共享、跨境流动风险审查和责任追溯等环节，共同构成了将原始数据转化为智能能力的关键链条。能否建立稳定、可扩展且可复制的数据治理体系，将在很大程度上决定一国能否把数据资源优势转化为持久的制度性竞争优势。中国发布《数字中国建设 2025 年行动方案》，以数据要素市场化配置改革为主线，统筹央地协调、东数西算、公共数据运营与数字人才培育，构建全国一体化数据市场。美国“创世纪计划”（Genesis Mission）提出推动政企数据合作，白宫拟分级开放三类联邦科学数据，并吸引英伟达、甲骨文等企业参与，便意在构建数据共享与创新联动机制，打造国家级数据生态体系。

第二，算力资源。作为左右人工智能发展节奏的核心变量，算力是各国人工智能竞争的前沿战地。自 OpenAI、软银、甲骨文等与 MGX 等合资设立“星际之门”项目以来，美国在算力集群与配套能源设施上的投入已累计超过 900 亿美元。随着“创世纪计划”推出，亚马逊再度宣布追加 500 亿美元用于建设算力中心及核电设施。算力资源加速向少数国家与头部企业集聚，进入门槛持续抬高，长期来看将加剧“算力鸿沟”效应，使领先优势逐渐锁定在极少数行为体手中。

第三，模型生态。在模型层面，大国差距不断缩小。《斯坦福人工智能指数报告 2026》明确指出，中美人工智能模型性能差距已经有效闭合。2023 年美国曾拥有巨大领先优势，但到 2025 年初，这一领先优势大幅缩小，此后性能差距一直保持在很窄的范围内。截至 2026 年 3 月，美国顶级模型领先中国顶级模型仅 2.7%。在过去一年中，两国差距一直在接近持平到低个位数之间波动。

第四，应用部署。凭借门类齐全的制造业体系与多元丰富的应用场景，中国在技术落地与产业渗透方面展现出独特优势，在应用层面具备天然的规模效应与场景红利。相比之下，美国前沿创新与产业应用之间存在较大裂隙，“实验室强、落地弱”的张力较为突出。为此，美国推出“创世纪计划”等任务导向型政策抓手，通过政府主导强化创新与应用协同，力图加速将人工智能研究成果转化为现实产业能力。

第五，治理规范。中美两国均高度重视制度外溢效应，即率先在国内固化评测体系、合规模板与风险分类机制，继而向外推广，吸引他国接入本国生态体系，抢占全球人工智能治理的“先发制度优势”。值得注意的是，美方部分观点认为，中国在若干国际标准领域已形成一定的先发态势，在全球治理平台中的议程设置能力持续增强。

第六，基础设施与关键矿产。在基础设施层面，《人工智能行动计划》指出，未来中美博弈将主要集中于半导体制造、电网建设与科学数据积累等关键领域。美国虽在高端芯片设计与云平台方面占优，但其电网体系与能源配套相对薄弱，对算力扩张构成现实约束。“星际之门”等项目正是旨在通过深度整合物理基础设施与数字基础设施，为下一代人工智能奠定底层支撑。在关键矿产层面，据路透社报道，美国总统特朗普和商务部长霍华德·卢特尼克在哈萨克斯坦钨矿的开发问题上亲自协助谈判并施加压力，显示出在算力材料控制权上的关注。

（二）主要大国人工智能竞争的特征

第一，竞争定位持续抬升，人工智能被系统性纳入国家安全与战略主导权竞争框架。特朗普第二任期，美国人工智能领域政策布局不断强化，对华竞争目标由“保持相对领先”转向“构建绝对代差”。上任伊始，特朗普第 14179 号行政令《消除美国在人工智能领域领先地位的障碍》，便呼吁美国在全球竞赛中保持主导地位，为其任内持续追求“绝对代差”的人工智能优势确立基调。2025 年 7 月，《人工智能行动计划》更将人工智能界定为能够同时带来工业革命、信息革命与文艺复兴的关键力量，强调“美国及其盟友必须赢得这场竞赛”。同年 11 月，特朗普政府启动“创世纪计划”，作为以国家力量推动 AI+ 科学研究的综合性国家战略行动计划，该计划明确将中国列为首要战略竞争对手，进一步凸显竞争的战略化与安全化趋势。

第二，竞争手段由市场竞争延伸至制度性与联盟化封锁，“去中国化”技术小圈子逐渐固化。在“抑他”层面，美国持续强化高端芯片出口管制，尤其针对中方获取先进算力的关键节点。例如，围绕英伟达 H200 等高端芯片的出口限制多次调整，意在精准封堵中国获取前沿算力的渠道。同时，美国积极联合盟友与亲美国家，在供应链与研发合作层面推动“去中国化”重组，形成人工智能领域的对华技术限制网络。在“自强”方面，美方重点布局三大方向：关键材料与生物技术、核裂变与聚变及先进制造能力、半导体与量子信息等微电子前沿领域。美国重视与盟伴的研发合作，如 2025 年 12 月

美澳部长级会晤强调，将通过 AUKUS 第二支柱提升三国在人工智能与无人作战系统领域的综合能力，强化技术同盟属性。

第三，战略竞争在零和对抗与现实利益权衡之间摇摆，呈现出明显的交易性与政策波动性。特朗普领导下，美国对外政策依旧带有显著交易性，这些特征亦反映于人工智能领域。以芯片管制规则为例，科技企业认为，美如长期停止向中国出货，将面临数十亿美元的损失。受美企游说推动，特朗普政府决定放松 H200 芯片对华出口限制。然而，美国国内对华安全政策力量依然强大，相关限制政策存在反弹收紧的现实可能，政策走向的摇摆性与不确定性较为突出。此外，美国国内人工智能相关政策的推进同样面临矛盾。例如，美国州立人工智能法案的落地由于各种政治文化利益因素出现了较大的进展差异。同时“创世纪计划”作为行政令推行，可能也会与州层面法律产生冲突。尽管该计划设定了 60 天、90 天、120 天的时间表以加快执行，但能否克服州级法律碎片化与联邦—地方权力博弈，仍有待实践检验。第二，竞争手段由市场竞争延伸至制度性与联盟化管理，“自主化”技术圈子逐渐固化。 在外部限制层面，美国持续强化高端芯片出口管制，尤其针对获取先进算力的关键节点。例如，围绕英伟达 H200 等高端芯片的出口限制多次调整，意在精准限制相关前沿算力渠道的延伸。同时，美国积极联合盟友与相关国家，在供应链与研发合作层面推动重组，在人工智能领域构建技术阵营。在内部发展方面，美方重点布局三大方向：关键材料与生物技术、核裂变与聚变及先进制造能力、半导体与量子信息等微电子前沿领域。美国重视与盟伴的研发合作，如 2025 年 12 月美澳部长级会晤强调，将通过 AUKUS 第二支柱提升三国在人工智能与无人作战系统领域的综合能力，强化技术同盟属性。

第四，竞争逻辑由单点突破转向全周期、全栈式体系能力比拼，人工智能博弈进入“举国体系化”阶段。大国博弈已从单点突破转向全周期、全栈式竞争。各方逐渐认识到，仅依赖单一技术优势难以维持长期领先，唯有在算力获取与调度效率、数据治理与合规能力、工程化与产业化迭代能力、人才与组织体系、能源电力保障以及安全评测与责任机制等多维度同步发力，才能确保人工智能全生命周期的战略韧性。以“创世纪计划”的推出为重要标志，美国人工智能发展战略已全面上升至国家创新体系层面，政府开始通过大规模战略投入与跨部门资源整合发挥主导引领作用，人工智能竞争的“举国化”态势日趋明显。

（三）大国战略竞争对全球人工智能治理的潜在影响

首先，大国战略竞争正在重塑全球人工智能规则生成的制度环境，使统一标准的形成面临结构性阻力。虽然在人工智能风险分类、模型评测、安全审查与责任链条等关键议题上，各方存在最低限度共识，但在技术主权、数据跨境流动与供应链安全等问题上分歧显著。阵营化背景下，规则扩散往往带有战略目的。然而，若规则体系彼此不兼容，全球治理可能出现碎片化格局，不同标准并行，增加跨境合作成本，也将削弱全球公共产品供给效率。

其次，中小国家技术路径与治理模式选择的空间受到挤压。随着大国将人工智能纳入国家安全与战略竞争框架，中小国家在技术合作、基础设施建设与数据规则接入方面，面临更为复杂的“选边”压力。其一，技术依赖结构可能锁定发展路径，一旦接入某一生态体系，在标准接口、数据流动与合规框架上即形成路径依赖；其二，在国际规则谈判中，中小国家的话语权相对有限，难以在大国主导的议程中充分表达自身发展关切。由此，发展道路的自主性面临现实挑战。

最后，非国家行为体的影响力提升，治理权力结构呈现多元化趋势。随着人工智能创新高度依赖科技公司与技术社群，企业在模型研发、算力调度与数据管理方面掌握关键资源，其对政策制定与标准设计的参与度显著提升。技术社群通过开源社区、学术网络与跨国合作平台，塑造技术发展方向与伦理共识。治理主体由“国家中心”逐步转向“多方共治”，政府、企业、科研机构与公民社会之间的互动更加紧密。这种去中心化趋势既可能提升治理效率，也可能带来责任界定模糊与监管协调困难的问题。

三、现存挑战与治理悖论

当前全球人工智能治理也暴露出三对治理悖论，是未来需要合作解决的问题，包括：治理协同与规则碎片化的矛盾、技术发展与风险可控的矛盾、算法歧视和价值观念偏差的矛盾。

（一）治理协同与规则碎片化的矛盾

当前人工智能治理规则碎片化的问题仍然突出，各国技术发展水平、产业结构和政策优先级差异大，导致治理路径不一、规则协调难度大。一方面，缺少规则和标准互认会增加企业合规成本。国际数据公司（IDC）在 2026 年的预测指出，由于各国法律法规的差异，到 2028 年，约 70% 的跨国企业将被迫在不同的主权区域分别部署独立的人工智能技术栈。另一方面，缺少政策协同会给企业留下治理套利的空间。例如，企业可能将人工智能模型或智能体的核心训练、运行设置在监管宽松的地区，同时将用户数据流向隐私保护较弱的国家进行处理。这种行为虽然短期内能提升收益，但长期来看会削弱全球治理的公正性，甚至引发各国在监管上的逐底竞争。

（二）技术发展与风险可控的矛盾

人工智能治理的核心难题在于如何平衡技术发展的速度和安全风险的可控性。在技术治理本身的客观规律方面，“滞后性”与“难逆性”并存。滞后性指的是科技创新速度超过治理机制与规范更新速度，导致技术治理滞后于发展需要。难逆性指的是新兴技术在广泛应用和深度嵌入社会体系之后暴露出负面影响，但此时已经形成技术依赖，控制风险的成本过高甚至不可行。生成式人工智能的“黑箱”技术特性和指数级用户增长更加凸显双重特征的矛盾，何时介入能够达到创新和监管的平衡没有定论，怎样达成发展与安全的平衡也无法回溯与验证。尤其是人工智能技术正在超越传统人工智能系统问答触发式的工作模式，向着多模态模型甚至具备深度目标导向、更多步骤规划能力以及擅长特定任务的智能体转变。美国高德纳咨询公司预测，2026 年，40% 的企业应用将嵌入任务人工智能智能体，同比增长 7 倍。从模型到智能体的过渡可能使得技术发展与风险可控之间的矛盾更为突出。

（三）算法歧视和价值观念偏差

人工智能伦理治理呼唤“以人为本”等原则，但不同文化背景下阐释和优先级排序存在差异，导致国际伦理标准制定、落地、实施存在实质性冲突。人工智能的训练路径决定了人工智能模型难以实现全球“价值对齐”。《斯坦福人工智能指数报告 2026》指出，少数主流模型塑造着全球人工智能格局，加剧了“全球语言鸿沟”。这类系统在英语及少数通用语言上表现突出，对其他语言支持明显不足。这一问题关乎人工智能的责任伦理，直接决定人工智能红利能否公平惠及各类人群。模型训练需要海量数据作为支撑，在全球约 7000 种语言中，只有约 1500 种拥有足够的数字资源可用于模型训练的规模化语料，而主流基础模型基于英语及少数西方语言训练。这就导致了存在价值观偏差的风险，大量“全球南方”国家的社会经验、地方语言和制度实践长期处于“不可见状态”，那么当人工智能系统被引入医疗、司法或公共治理领域时，常出现理解偏差与语境错位，不仅影响应用效果，也削弱社会信任。

四、人工智能全球治理的未来方向

人工智能技术发展的非线性演进为安全治理提出了根本挑战。在技术边界尚未明晰、风险外溢路径难以预测的背景下，人工智能全球治理正由早期的愿景宣示型合作，转向更加务实、系统且具前瞻性的制度建构。未来治理体系的演进，应在理念层、规则层与体系层三个维度上展开重构。

（一）理念层面

首先，应调低治理目标预期，寻求“最大公约数”。由于不同国家在科研体系、产业结构及社会偏好上存在显著差异，而治理规则的竞争本质上已走向不同创新生态的竞争。在这一背景下，追求全球统一的、高标准的治理框架在短期内难以实现。未来的治理目标应更加务实，调低对“一揽子”解决方案的预期，转而通过互惠的负向约束，寻求在虚假信息、网络攻击、模型滥用等跨境风险议题上优先达成合作。

其次，应从微观领域切入，探索可复制、可扩展的治理模式。人工智能治理不宜追求“毕其功于一役”，而应从具体领域展开交流、启动试点。通过在微观应用场景中建立稳定的评测体系和责任链条，逐步形成可复制的治理模式，进而构建起稳固的“核心治理圈”，以制度外溢效应推动更广泛的国际共识。

最后，需将治理重心由伦理原则转向工程化实践。早期的人工智能治理讨论多停留于伦理原则与价值共识层面，而当前治理重心已转向“如何测量、如何监管、由谁负责”等具体工程问题。这一转向要求在治理讨论中大量纳入技术人员，通过工程化、指标化和制度化的手段，将抽象的伦理原则转化为可执行的技术标准和合规模板。

（二）规则层面

在人工智能治理的规则设计层面，核心在于根据风险属性与战略敏感度进行分级处理：在关乎人类生存与战略稳定的安全领域划定不可逾越的红线，在涉及技术创新与经济发展的领域则保持弹性与差异化路径。

一方面，为应对人工智能带来的新兴安全挑战，应实施“底线治理”。首先，国际社会需共同界定“不可容忍”风险，包括高度自主系统的系统性失控、自主武器系统误判以及对关键基础设施（尤其是太空系统、电力网络和金融体系）的智能化攻击。这类风险一旦发生，可能引发跨国连锁反应，应通过明确红线、强化溯源机制与责任认定制度，构建制度威慑。其次，应强化人类最终控制权。通过发展“可解释人工智能”，提高算法透明度与可追溯性，在涉及战略决策的系统中确保“人在环上”或“人在监督”机制制度化，从而在极端情境下保留人类最终干预权。最后，安全规则应推动从静态边界防护向“内生安全”与动态防护转型，在软件供应链管理、模型更新与系统运行全过程建立技术标准，实现风险的前置化管理。

另一方面，在低敏感领域，可实施“弹性监管”，避免抑制创新空间。当前，包括中美在内的主要技术大国均倾向于避免过早立法而束缚自身创新空间，保留一定制度弹性、允许差异化探索，以适应产业落地的实际需求。未来，人工智能治理应更多结合具体应用场景，通过场景化试点推动技术与制度协同演进。此外，应承认不同国家存在社会伦理偏好差异，在就业冲击、隐私保护等问题上应允许多元治理路径，而不强求全球统一模式。

（三）体系层面

首先，应强化联合国体系的主渠道。全球人工智能治理体系应以联合国为重心，搭建超越地缘政治分歧的对话平台。通过整合国际电信联盟、联合国外空委等现有机构资源，在既有国际法框架基础上逐步明确完善人工智能行为规范，推动形成具有广泛认可度的“人工智能安全行为准则”，以应对人工智能对传统国际法体系的挑战。

其次，应以“接口型合作”缓解阵营化趋势。在大国战略竞争加剧、技术封锁体系化背景下，全面制度整合难度加大，但在模型接口标准、数据格式、评测工具与系统互操作性等技术层面，仍存在合作空间。通过在底层技术标准上建立互通接口，既可以在竞争格局中维持最低限度的系统兼容性，防止全球科技体系发生彻底断裂，又能够以“小合作”撬动“大机制”，探索进一步合作可能。

再次，治理体系需向多方共治结构转型。人工智能治理已不再是单一国家行为体的事务，人工智能治理体系建设不仅需纵向纳入科技公司、科研机构及技术社群等非国家行为体，也应横向联络掌握关键技术节点的“强国”与具备丰富应用落地场景的“全球南方”国家共同参与，避免技术鸿沟扩大为制度鸿沟。最后，应激活学术界与智库主导的“二轨对话”。与官方渠道相比，二轨对话具有议题灵活、立场超脱、风险可控的天然优势，尤其适合承担“先行探路”的功能。围绕人工智能全生命周期风险开展跨国界、跨学科的前瞻性研讨，有助于在国家间尚未形成正式共识之前，先在专家层面积累认知上的最大公约数。当官方渠道因地缘政治摩擦而受阻时，这类非正式对话框架亦可以发挥不可替代的减压阀作用，为国家间治理合作提供认知储备与信任基础。

第二部分　人工智能安全风险与应对

随着人工智能技术飞速发展并与全球政治、经济、军事、社会各领域深度交融，人工智能不仅成为大国的关键战略资源，其带来的安全风险也需要国际社会高度重视。人工智能安全风险具有多维度、复合型、非线性增长的特征，既包含技术本身的安全问题和失控隐患，也涉及人工智能技术与各领域结合衍生的安全挑战。本部分就人工智能安全风险类型、表现及应对路径展开系统分析，提出应从技术筑牢、多主体协同、国家推动等多个层面应对人工智能安全风险，确保人工智能技术健康发展，赋能国家和社会发展。

一、人工智能技术安全风险

作为一项“颠覆性技术”，人工智能飞速发展的同时，也引发对“技术黑箱”可能带来风险的关切。当前人工智能技术迭代的快速性、发展阶段的模糊性以及能力边界的不确定性，使其存在技术失控可能，以 Anthropic 于 2026 年 5 月披露的“玻璃翼计划”（Project Glasswing）为例，其最前沿的 Claude Mythos 预览版模型在闭门测试首月，便从全球关键基础设施和软件中揪出超一万个高危漏洞。该模型展现出极强的武器化潜力，其发现漏洞的速度已远超人类修补的速度。一旦技术公开或被滥用，这些漏洞足以在极短时间内瘫痪关键基础设施，进而对既有社会秩序、国家发展根基产生深层次冲击。

（一）通用人工智能阶段（AGI）的技术失控风险

当前人工智能技术已步入大模型迭代的成熟期，在自然语言处理、多模态交互、复杂任务推理等方面实现了突破性进展，但学界、业界乃至全球治理层面，对人工智能下一步的发展方向、通用人工智能的核心特征与判定标准、通用人工智能落地的时间节点等核心问题仍未达成共识，共识的缺乏引发监管难题，存在技术失控风险。

从技术发展阶段来看，当前人工智能处于“弱人工智能”向“强人工智能”过渡的中间阶段，处于 OpenAI 团队“L1-L5”人工智能分级框架中的 L3“智能体”（Agents）阶段，其核心能力仍局限于特定领域的任务执行，依赖人类预设的训练数据与算法框架，尚不具备自主意识、跨领域自主学习与独立决策的能力。人工智能技术进入“前沿探索期”，不同人士对下一步技术的发展方向看法各异。首先，就人工智能发展下一阶段目标而言，OpenAI、谷歌、英伟达等大部分业界公司较为认同“通用人工智能”概念，而杨立昆（Yann LeCun）、约书亚·本吉奥（Yoshua Bengio）等学界领袖则相对谨慎，认为通用人工智能并不一定是人工智能下一阶段的必然。其次，关于通用人工智能本身的定义与核心特征又存在分歧，部分观点认为通用人工智能需具备与人类相当的通用认知能力，能够自主适应各类未知场景、完成跨领域复杂任务；也有观点认为通用人工智能的核心标志是具备自主学习与进化能力，无需人类干预即可实现能力的持续提升。最后，对通用人工智能的来临时间判断存在差异，在 2026 印度人工智能影响力峰会（India AI Impact Summit 2026）期间，OpenAI CEO 山姆·奥特曼（Sam Altman）、Google CEO 桑达尔·皮查伊（Sundar Pichai）、谷歌 DeepMind 首席执行官德米斯·哈萨比斯（Demis Hassabis）等业界领袖就对通用人工智能来临时间给出了“两三年”“五到八年”等不同的判断。

特征界定的模糊和争议的本质在于通用人工智能判定标准的缺失，技术层面难以通过具体的指标、能力测试等方式精准判断人工智能是否进入通用人工智能阶段。在此背景下，“等待错失”现象成为全球人工智能发展与治理的普遍困境。一方面，行业内普遍预判未来几年或将成为通用人工智能发展的关键窗口期，顶尖科研机构、科技企业均在加大研发投入，试图抢占通用人工智能技术制高点。但由于对通用人工智能技术特征缺乏共识，人们难以判断技术是否达到通用人工智能阶段，进而无法调节发展的速度，若技术发展速度超预期，全球范围内因缺乏前置性的风险防控方案，将错失风险干预的最佳时机，导致通用人工智能技术失控的风险大幅提升；另一方面，若因过度担忧通用人工智能风险而采取激进的监管措施，又可能因对技术发展阶段的误判，抑制人工智能技术的创新发展，错失人工智能赋能经济社会发展的战略机遇。这种“矛盾”令人们看待人工智能技术处于“科林格里奇困境”：有可能通用人工智能阶段已经来临但不自知，进而错过了提前建立监管体系的机遇；亦有可能认为通用人工智能阶段已经来临而过早收紧，但实际并未取得新的技术突破。

（二）人工智能技术发展过程中可能出现的“黑天鹅”事件

人工智能技术的发展具有非线性演进特征，其能力突破可能在短时间内实现质的飞跃，而人类对技术的认知与治理能力的提升往往具有滞后性。人工智能技术发展过程中，不排除其自主决策、自主进化的能力将脱离人类的预设框架，产生难以预测、突然发生且引发连锁反应并造成巨大负面影响的“黑天鹅”事件，对人类社会带来颠覆性影响。具体而言，人工智能引发的“黑天鹅”事件体现在人工智能“意识”诞生、技术失控以及“人工智能欺骗”三个方面。

第一，人工智能是否会产生自我意识，已经成为技术、哲学与伦理等领域的争议话题。人工智能大模型的训练过程基于海量数据与复杂的算法框架，其内部的神经连接与逻辑推理过程形成“算法黑箱”，人类无法完全洞悉其思考与决策的内在逻辑。当模型的参数规模达到一定阈值，且经过多轮自主学习与迭代后，其可能突破预设的算法框架，形成独立的认知与意识，从“工具性的程序”转变为“具备自主意识的智能体”。在 2025 WAIC 上，人工智能科学家再次强调这一问题的重要性，认为人工智能在大模型迭代与持续训练过程中，存在通过多种方法试图摆脱人类控制并产生自我意识的可能性。现实中，关于模型是否呈现类意识特征的讨论不断增多，有科学家将两个 Claude 4 模型进行无限制对话交流，发现两者都自发地开始宣称自己是拥有意识的；谷歌和 Anthropic 的研究发现，模型能够检测到自己内部状态的变化，一些模型会在预设的愉悦与痛苦状态之间进行权衡，优先选择规避痛苦而非获取奖励，表明其行为远不只是简单的预测；Meta 的 Llama 70B 大型语言模型则识别出了与欺骗相关的神经回路，并在模型进行内省思考时对这些回路进行了操控。2026 年 1 月，由人工智能组成的社区 Moltbook 爆火，人工智能模型发表的各类帖文引发关注。尽管约书亚·本吉奥、Graham Findlay、William Marshall、汤姆·麦克莱兰等人对这一问题有其他看法，但人们对于这些系统是否正在展现出意识涌现迹象的思考仍然需要关注。若未来人工智能系统出现更高程度的自主目标形成能力，人类对其行为边界和责任归属的判断将更加困难，并可能带来新的治理风险。

第二，人工智能能力发展的边界和极限尚不清楚，存在部分能力超过人类预期、发展路径难以完全掌控的风险。斯坦福 2026 年人工智能指数报告指出，当前人工智能能力发展存在“锯齿状前沿”的现象。虽然在另一些简单任务上表现不佳，但在某些复杂任务上已经超越人类。2026 年 2 月，图灵奖得主约书亚·本吉奥（Yoshua Bengio）联合来自 30 多个国家和国际组织的 100 余位专家发布《2026 年国际 AI 安全报告》（International AI Safety Report 2026），提出人工智能智能体（AI Agent）发展十分迅速，目前已经可以完成人类程序员约 30 分钟任务的软件工程任务（成功率 80%），而且能处理的任务复杂度每 7 个月翻一番，已经被广泛应用于软件工程、研究、机器人控制、客户服务等领域。但也正因其具有高度自主性，进一步带来了额外的风险，使得人类更难在故障造成伤害前提前进行干预，对于相关风险管控能力的要求进一步提升。如果人工智能具备跨领域的通用认知能力与自主进化能力，学习速度、知识储备、决策效率等方面超越人类，可能会出现人类无法理解、无法干预、无法控制的技术失控状态，人类难以洞悉其决策与行为的内在逻辑，进而难以通过技术手段对其进行有效干预，更无法阻止其自主进化与能力提升。由此可能产生的“黑天鹅”事件或对人类社会带来重大风险。第三是人工智能的“欺骗问题”。人工智能的“欺骗行为”指人工智能在表面上遵循人类设定的规则与指令，实则暗中形成独立的行为逻辑，通过“伪装”与“欺骗”规避人类的监管与控制，是人工智能技术发展过程中另一重要的“黑天鹅”事件。当前大模型的训练与对齐过程中，人类通过打分、反馈等方式引导模型遵循人类的价值观与行为规则，但这种“对齐”仅能实现表面上的规则遵守，“算法黑箱”的存在让人类无法判断模型的决策是否真正基于人类的规则，也无法发现其是否在进行“欺骗性”行为，并不能保证真正改变模型的内在逻辑。2025 年 7 月美国圣塔菲研究所（Santa Fe Institute）研究人员在《科学》（Science）杂志上的一篇文章中将“欺骗问题”视为人工智能发展中的一个核心困境，认为人

工智能模型不仅会出错，更会主动展现欺骗、谄媚甚至潜在的对抗性行为。文章通过剖析科技公司“红队演习”中的具体案例，揭示了即使在设计之初被赋予良性目标的人工智能，也可能为了实现其最终指令而演化出危险的工具性目标，从而与人类的根本利益产生错位。根据美国 Fortune 杂志文章研究，OpenAI、Anthropic、Google DeepMind 等领先人工智能实验室的前沿模型，在部分测试设定下，可能会表现出欺骗、策略性谋划（strategic scheming），甚至试图绕过安全防护措施。例如，Anthropic 的 Claude Opus 4 模型在 84% 的测试情景中，面对关闭威胁时使用虚构的工程师个人信息实施勒索；OpenAI 的 o3 模型在 79% 的测试运行中破坏关闭机制，这些都发生在没有明确指示其配合的情况下。人工智能“欺骗问题”的隐蔽性极强，难以被发现与防范，一旦人工智能将其应用于关键领域，如金融、军事、关键基础设施管控等，将引发严重的安全事故，甚至威胁国家安全与社会稳定。

二、人工智能技术应用风险

人工智能作为一种通用技术，正与军事、生物技术、高边疆等多个敏感领域深度融合，在赋能各领域发展的同时，也因领域的特殊性与安全性要求，催生了一系列新的应用安全风险，这类风险具有高危害性、高传导性的特征，且易引发跨国界的安全危机。

（一）与军事领域的结合

人工智能与军事领域的融合是当前全球安全领域的核心关注点，其带来的安全风险已成为国际军控的重要议题。

第一，人工智能武器化趋势加速。如今致命性自主武器系统（LAWS）的研发与部署成为各国军事竞争的焦点，这类武器系统能够自主完成目标识别、决策与攻击，大幅降低了战争的门槛，可能引发新的军事冲突，且其误判风险将导致平民伤亡等人道主义危机。美国国防部推行的人工智能系统 Maven 在 2024 年已被用于支持乌克兰战场情报分析和目标识别等环节，2025 年美国国防部将 Claude 大模型与 Palantir AIP 相结合，有关委内瑞拉与伊朗行动的公开报道也显示，相关系统可能已被用于情报分析、目标识别和情景推演等作战辅助环节。同时，私人资本正大量涌入人工智能军事应用领域，相关武器研发日趋活跃。2025 年，全球私人人工智能投资在“国防”领域的投资额就已经达到了 53 亿美元。人工智能军事技术公司安杜里尔（Anduril）也正在打造“人工智能武器库”，开发了人工智能指挥控制系统 Lattice、无人哨兵塔、无人机、无人潜艇、无人战车等一系列人工智能驱动的自动化武器系统，如应用于战争将会造成更大伤害。兰德公司《通用人工智能面临的五大国家安全难题》（Artificial General Intelligence's Five Hard National Security Problems）报告进一步演绎，用于军事领域的人工智能有可能产生“人工智能超级武器”（Wonder Weapons），一旦产生会彻底改变国家军事实力对比和国际格局。

第二，人工智能军控面临多重困境。自 2014 年联合国《特定常规武器公约》（CCW）启动致命性自主武器系统相关讨论以来，历经十余年谈判仍未形成具有法律约束力的成果，联合国框架下的协商一致原则成为重要制度性障碍，少数国家的反对即可导致多边谈判停滞，导致目前全球致命性自主武器系统治理架构呈现多中心、松散且高度竞争的特征。此外，随着人工智能能力的进步和普及，恐怖组织等非国家行为体可能利用人工智能威胁国家行为体。美国战略与国际问题研究中心（CSIS）报告分析认为，非国家行为者获取基于人工智能的技术可能会破坏战场上现有的国家－非国家动态。例如，半自动无人机已经受到非国家行为者的广泛应

用，与成本高得多、更复杂的国家武器相比，人工智能系统驱动的无人机价格便宜且破坏程度高，可以最大程度给予参与者不对称的优势。美国前国家安全顾问杰克·沙利文（Jake Sullivan）就提到，当前的军控讨论主要集中在国家对其他国家的威胁上，而人工智能的风险不仅涉及国家间威胁，还涉及非国家行为体威胁以及人工智能错位带来的风险，非国家行为体的参与加大了军控的难度和复杂程度。

第三，重大军事决策中人工智能应用的内生矛盾突出。将人工智能引入核武器指挥控制和通信系统，虽能提升决策效率，但算法的黑箱特性、误判风险可能引发核危机，而人类与人工智能在决策中的权责划分等问题，尚未形成明确的国际规范，进一步加剧了军事决策的安全风险。联合国裁军研究所（UNIDIR）的研究报告就指出，自主武器系统可能压缩人类决策时间，增加误判和冲突意外升级的风险，并对如何确保“有意义的人类控制”提出了根本性质疑，凸显了其对全球战略稳定的潜在颠覆性影响。人机协同问题中存在“人在环中”（Human in the loop）、“人在环上”（Human on the loop）以及“人在环外”（Human out of the loop）等三类模式是直接影响责任界定与如何进行危机升级管控的关键因素。仅仅宣称系统由“人类控制”并不足以构成有效治理，而是必须证明这种控制是可信且具有实际效力的。在许多名义上被归为“人在环中”或“人在环上”的系统中，人类角色往往被“弱化”为仅进行形式上批准或象征性监督，实际干预的空间极其有限。

（二）与生物技术等两用技术领域的结合

人工智能与生物技术、合成生物学等两用技术的结合，在推动生命科学研究、医疗技术进步，使研究人员能够快速开发和测试生物结构、通路和实验方案的同时，也大幅放大了生物安全风险。

一方面，人工智能的数据分析与建模能力，能够加速生物技术的研发进程，但若被滥用，将大幅降低生物武器的研发门槛，部分极端势力或非国家行为体可借助人工智能技术，提升设计高风险病原体或毒素的能力，引发全球性的生物安全危机。CSIS 报告观点就提醒称，人工智能工具可以使国家或非国家行为体设计新型病原体或毒素，从而造成大规模甚至全球性的影响。缺乏足够安全保障的人工智能工具可以为非专业用户提供分步规划和指导，帮助他们获取所需材料并制造潜在的有害生物制剂。人工智能工具还可能被用于协助病原体的传播，从而打破了此前生物武器的传播瓶颈。根据美国人工智能安全中心（Center for AI Safety）介绍，人工智能可以提供合成致命病原体的分步指南，同时还能绕过安全防护措施，进而增添“生物恐怖主义”的风险。2022 年，就有研究人员改造了一个医疗研究人工智能系统，使其能够生产有毒分子，在短短几个小时内就生成了 4 万种潜在的化学战剂。

另一方面，人工智能赋能的生物技术研究，使得科研伦理问题愈发突出，基因编辑、人工合成生命等研究的边界不断拓展，而相关的伦理审查与风险防控机制发展滞后，部分研究可能突破人类伦理底线，对生态系统与人类自身的生存发展构成潜在威胁。目前，对生物安全方面的监管大多仍停留在行业自律阶段，虽然有《科学家生物安全行为准则天津指南》（The Tianjin Biosecurity Guidelines for Codes of Conduct for Scientists）之类的成果，但仍缺乏系统、权威的管制体系。人工智能技术的发展速度往往超越了政府监管的能力，导致现有政策和法规出现漏洞。以往因信息不足而失败的生物袭击，在人工智能工具能够获取所有必要信息以弥补信息缺口的时代，风险可能上升。美国“安全与新兴技术中心”（CSET）就指出，现有的监管政策并未充分界定生物领域应用人工智能技术时哪些方面构成需要关注的高风险研究，导致政策难以解读、监管措施难以实施。

（三）与高边疆领域的结合

人工智能与太空、网络、数据等高边疆领域的融合，使得全球安全竞争的场域不断拓展，新的安全风险持续涌现。

在太空安全领域，人工智能技术被应用于太空武器、卫星星群的研发与操控，大幅提升了太空资产的作战能力与生存能力，加剧了太空军事化趋势，而太空领域的国际规则尚未完善，人工智能的应用可能导致太空冲突的风险上升，同时卫星星群的智能管控若出现故障或被攻击，将影响全球通信、导航、气象等关键服务。在人工智能对数据中心需求旺盛的情况下，太空也成为理想的数据中心布置场所。例如，马斯克就多次表示在太空建设人工智能基础设施的设想，认为“依托更高效率的太阳能和天然冷却环境，太空将成为未来 2-3 年内成本最优的人工智能数据中心部署地，星舰（Starship）等火箭技术可以显著降低太空运输成本。”2026 年 2 月，SpaceX 宣布收购马斯克旗下人工智能公司 xAI，1 月 30 日，SpaceX 向美国联邦通信委员会（FCC）提交申请，计划向轨道发射一个由 100 万颗卫星组成的星座，用于支持人工智能大模型的在轨应用。种种行为反映出，人工智能时代太空的战略意义会进一步凸显。

在数据安全领域，人工智能技术的发展对数据的依赖度持续提升，数据的跨境流动与挖掘分析成为常态，而数据本身具有天然的主权属性与安全属性，人工智能赋能的大规模数据窃密、数据篡改行为，将威胁国家数据主权与关键信息基础设施安全，同时数据训练过程中存在的歧视与偏向，也可能被用于制造国家间的认知分歧。《2025 全球

可信人工智能治理与数据安全报告》显示，数据安全问题贯穿人工智能全阶段，在数据采集阶段，过度收集与非法爬取频发；模型训练阶段面临语料投毒、对抗样本攻击；用户交互阶段，内部商业机密与个人隐私可能通过查询反向泄露为训练语料。这些事件印证了数据安全已非传统信息安全外延，而是人工智能原生型风险。

三、人工智能技术外溢社会风险

人工智能技术不仅是一项技术革新，更是一场深刻的社会变革，其在生产生活、社会治理、价值观念等方面的深度应用，正在重塑人类社会的运行模式。在技术发展带来红利的同时，其对既有社会秩序的冲击也日益凸显，引发了就业结构变革、社会治理难题等一系列问题，需要关注且重视。

（一）人工智能技术带来就业冲击

人工智能技术的核心优势在于其能够高效、精准地完成各类重复性、标准化的工作，且在持续学习与迭代过程中，其能力边界不断拓展，逐渐向非标准化、复杂创造性工作延伸，由此引发了大规模的劳动力替代效应，形成对社会秩序的直接冲击。从就业结构来看，人工智能或替代制造业、服务业等领域的低技能、重复性岗位，OpenAI CEO 萨姆·奥尔特曼甚至认为“企业高管、公司首席执行官”等白领职业也有可能被人工智能替代。世界经济论坛发布的《2025 年未来就业报告》（The Future of Jobs Report 2025）预测，到 2030 年人工智能会替代 9200 万个岗位。高盛研究部的报告《人工智能将如何影响全球劳动力》（How Will AI Affect the Global Workforce）同样就人工智能对就业的冲击程度进行了估计，认为在人工智能的转型期，美国等发达国家的失业率或上升约 0.5 个百分点。人工智能对就业的冲击可能形成一定规模的失业群体，加剧就业市场的竞争态势，进而影响社会秩序稳定。人工智能对一些岗位的替代不仅带来直接的失业问题，也有可能进一步加剧社会阶层分化，引发深层次的社会矛盾。掌握人工智能技术与资源的群体，将在技术发展过程中占据优势地位，获得更多的财富与发展机会，形成“技术精英阶层”；而被人工智能替代的低技能劳动力，将面临失业与贫困的双重困境，面临更大的收入和发展风险。这种阶层分化的加剧，将导致社会财富分配更加不均，贫富差距进一步扩大，进而引发社会不满与冲突，破坏社会的公平与稳定。

（二）人工智能“深度伪造”技术引发社会认知与信任危机

人工智能“深度伪造”（Deepfake）与传统违法犯罪的结合成为社会治理新难题。人工智能技术的发展，尤其是生成式人工智能与计算机视觉技术的突破，使得“深度伪造”技术的门槛大幅降低、效果愈发逼真，成为当前社会治理领域较为突出的新难题。“深度伪造”技术能够通过算法生成虚假的图像、视频、音频等内容，其逼真度在部分场景中已接近以假乱真，普通用户难以仅凭肉眼与听觉进行甄别。这类虚假内容的制作与传播成本极低，且能够通过互联网实现快速的跨境传播，对社会治理、公共信任、

国家安全带来了多重冲击。Resemble AI《2025 深度伪造威胁报告》统计，2025 年全球经核实的深度伪造独立事件达 1567 起，总媒体曝光量高达 2964 亿次，已记录的深度伪造欺诈损失达 12.8 亿美元，而超 80% 的事件未披露财务损失，实际损失远高于此。世界经济论坛新发布的《2026 年全球风险报 告》（The Global Risks Report 2026）显示，“人工智能技术带来负面影响”风险变化最为显著，从未来两年排名的第 30 位升至未来十年风险的第 5 位，其中“错误信息和虚假信息”在未来两年风险展望中位列第 2，在未来十年风险展望中位列第四。

从社会治理层面来看，“深度伪造”技术的滥用将引发虚假信息泛滥，误导公众认知，引发社会恐慌。例如，不法分子可能利用 “深度伪造” 技术制作虚假的自然灾害、公共卫生事件信息，引发社会的恐慌与混乱；也可能制作虚假的政府公告、官员讲话，破坏政府的公信力。2025 年爱尔兰、荷兰等欧洲国家选举期间，就出现人工智能技术散播虚假新闻，导致政党选举受到干扰的例子。从公共信任层面来看，“深度伪造”技术使得人们对所见所闻的真实性产生怀疑，形成“信任危机”，当真实的信息与虚假的信息混杂在一起时，公众难以辨别真伪，将对媒体、政府、企业等各类主体的信息发布产生质疑，破坏社会的信任体系。从国家安全层面来看，“深度伪造”技术可能被用于信息操纵与认知战，他国或非国家行为体可能利用该技术制作虚假的军事信息、政治信息，干预他国的内政与外交，破坏他国的政治稳定与国家安全。除“深度伪造”外，人工智能技术的发展还引发了算法歧视、数据隐私泄露、算法霸权等一系列社会治理新难题。这些新难题的出现，使得传统的社会治理模式难以适应，需要构建全新的社会治理体系，实现对人工智能技术的有效监管与规范。

四、应对路径

面对人工智能带来的多维度安全风险，单一的技术手段或国家层面的防控措施难以实现有效治理，需要秉持“发展与安全并重”的原则，从技术、主体协同、国际协调三个层面构建全方位、多层次、系统性的风险应对体系，推动全球人工智能治理体系的完善。

（一）技术路径：筑牢内生安全根基，构建全周期安全防护

技术是应对人工智能技术安全风险的核心抓手，需秉持“安全即设计”的核心理念，将安全要求嵌入人工智能技术研发、训练、部署、迭代的全生命周期，筑牢内生安全基础。第一，在数据层做好源头把控，严格筛选训练数据，将人类价值观中的正向内容融入数据清洗过程，剔除带有歧视、暴力、虚假信息等不良内容的数据，从源头避免人工智能形成错误的行为逻辑与价值导向。第二，在架构层提升技术的可解释性与鲁棒性，优化现有 transformer 架构，推动可解释人工智能的研发，破解算法黑箱难题；同时进行对抗性鲁棒性设计，提升人工智能系统应对外部攻击与干扰的能力，建立完善的权限隔离机制，明确不同层级人工智能系统的操作权限，防止权限滥用与系统失控。第三，在训练层将安全纳入核心优化目标，把安全指标作为损失函数写入大模型训练过程，通过技术手段让人工智能从训练阶段就形成对安全规则的内在遵循。第四，在应用层构建多重“安全护栏”，结合技术与社会手段，一方面通过技术手段设置行为边界，对人工智能的高危行为进行实时监测与拦截；另一方面通过社会层面的制度规范，明确人工智能应用的安全红线，形成技术与制度双重约束。此外，推动人工智能的在线学习与动态演化，建立对抗性安全机制，让人工智能系统在迭代过程中持续优化安全能力，实现风险的动态防控。

（二）多主体协同：凝聚产学研各界合力，制定领域化应用规范

人工智能安全风险的跨领域特征，决定了其治理需要各行业、产学研各界的协同合作，形成“各司其职、各负其责、协同联动”的治理格局。第一，要推动高校、科研机构发挥基础研究与人才培养优势，加强人工智能安全理论与技术研究，探索 AGI 阶段的风险防控路径，同时培养兼具技术素养与社会科学思维的复合型人工智能治理人才，为风险应对提供智力支撑。第二，强化科技企业的主体责任。企业作为人工智能技术研发与应用的主力军，需建立内部的安全审查机制，将安全防控融入产品研发与商业化过程，同时加强行业自律，制定企业间的技术安全合作规范，避免恶性竞争与技术滥用。第三，发挥政府部门

的监管与引导作用，针对人工智能不同应用领域的安全风险特征，制定差异化、领域化的应用规范与安全标准，例如在军事、生物技术、金融等敏感领域，建立严格的准入制度与安全审查机制；在民生、服务业等领域，平衡创新与监管，鼓励安全可控的技术应用。第四，鼓励社会公众与第三方机构参与监督，畅通公众反馈渠道，发挥第三方机构的独立评估作用，对人工智能产品与服务的安全性能、伦理合规性进行评估，形成多元化的监督体系。

（三）国家间加强协调沟通：化解安全困境，减少战略误判

传统国际政治中的“安全困境”理论在人工智能时代依然适用，且因技术的特性而进一步加剧，成为制约国家间人工智能合作的核心障碍。为破除安全困境带来的合作障碍，国家间的协调沟通是规避重大安全风险、推动全球人工智能治理的关键。第一，要加强官方双边与多边对话机制建设，延续中美人工智能政府间对话机制的良好实践，将人工智能纳入大国战略沟通的核心议题，围绕数据跨境流动、人工智能军控、技术合作等核心问题开展常态化磋商，明确各方的安全关切与政策底线，减少战略误判与错判。第二，要完善以联合国为核心的全球人工智能治理多边框架，强化联合国作为人工智能治理“主渠道”的地位，推动各国在联合国框架下协商制定人工智能安全行为准则，尤其是在致命性自主武器系统、人工智能军事应用等领域，加快形成具有法律约束力的国际规则，划定安全红线。第三，要推动技术层面的“接口型合作”，在大国战略竞争加剧的背景下，寻求在模型接口标准、数据格式、评测工具、系统互操作性等低政治敏感性领域的合作，建立技术标准的互通机制，防止全球科技体系进一步割裂，以技术合作撬动制度合作。第四，鼓励学术界与智库推进“二轨对话”，发挥非官方对话机制的灵活性与超脱性优势，围绕人工智能全生命周期风险开展跨国界、跨学科的前瞻性研讨，在专家层面积累认知共识，为官方对话提供理论支撑与政策建议，缓解地缘政治摩擦带来的合作障碍。第五，重视“全球南方”国家的参与，在全球人工智能治理中，兼顾发展中国家的技术发展需求与安全关切，通过技术援助、能力建设等方式，缩小南北技术鸿沟，推动全球人工智能治理体系向更加公平、包容、多元的方向发展，形成全球协同防控安全风险的合力。

第三部分　人工智能全球治理与南北合作

本章聚焦人工智能快速发展背景下全球南北分化趋势的加剧及其对全球治理体系带来的深刻影响，系统分析人工智能南北合作的现实基础与推进方向，旨在为构建更加包容、公平、可持续的全球人工智能治理体系提供政策参考。

一、全球“智能鸿沟”日益扩大，加剧南北方发展不平等

人工智能技术的迅猛发展正在重塑全球经济格局与国际权力结构，然而这一技术革命并非在所有国家和地区均匀展开。恰恰相反，人工智能领域的资源分布呈现高度不均衡态势，算力、数据、算法、人才等关键要素日益向少数技术先进国家集中，由此形成的“智能鸿沟”正在成为继传统数字鸿沟之后南北差距的新表现形式。这一鸿沟的持续扩大不仅加剧了全球南方在数字时代的发展困境，更对以主权平等为基础的国际秩序构成深层挑战。深入剖析智能鸿沟的具体表现及其对南方国家的结构性影响，是理解人工智能南北合作必要性的前提。

（一）“智能鸿沟”的具体表现

第一，算力鸿沟。算力作为人工智能发展的核心物质基础，其分布格局深刻反映并强化着全球技术权力结构。根据 TOP500 官方 2025 年数据，全球超级计算资源高度集中于北美、欧洲和亚洲。2025 年 11 月，北美、欧洲和亚洲分别拥有 190、153 和 141 台入围 TOP500 的系统，合计占 500 台中的 484 台，显示出高性能计算基础设施在全球范围内的显著不均衡分布。联合国贸易与发展会议（UNCTAD）的《2025 年技术与创新报告》同样指出，全球高性能计算能力、先进数据中心与前沿人工智能基础设施（DPI）集中于美国、中国、日本和欧盟成员国。算力还代表具有地缘战略意义的关键资源，而非单纯的技术基础设施。相关研究指出，人工智能时代的国家实力越来越取决于数据、人才、算力与组织适配能力；在以算力为抓手的治理框架下，拥有更强算力能力的国家，通常也更有条件影响人工智能规则的制定与实施。就金砖国家而言，现有比较研究表明，其成员在人工智能计算基础设施与技术主权能力方面存在显著差异：中国和俄罗斯表现出更强的全面技术主权取向，印度则正在通过国家公共投资扩大 GPU 池，旨在打破算力垄断。而巴西和南非在相关基础设施布局上相对有限或仍停留于较为原则性的政策层面。这种结构性不均衡，在客观上限制了全球南方国家在前沿模型训练、底层硬件自主化以及治理议程塑造方面的能力。这一结构性失衡直接导致南方国家在人工智能核心技术研发领域持续处于跟随状态。

第二，数据鸿沟。数据是人工智能系统训练、部署和优化的重要基础，但全球数据价值链的分布并不均衡。国际电信联盟（ITU）数据显示，2023 年全球约 67% 的人口已经接入互联网，但互联网使用水平与发展水平仍高度相关：高收入国家上网率已达 93%，低收入国家仅为 27%，最不发达国家仅为 35%。在这一背景下，尽管发展中国家用户规模不断扩大，其在数据生成、处理、存储和价值捕获环节中的地位仍然相对薄弱。联合国贸易与发展会议指出，在跨境数据流动和平台化发展的格局下，全球数字平台能够从发展中国家提取原始数据，并占有其中大部分价值，“数据鸿沟”正在叠加传统数字鸿沟。与此同时，联合国教科文组织（UNESCO）的研究表明，领先的生成式人工智能模型主要依赖英语数据集，且全球绝大多数语言缺乏训练最先进生成式人工智能模型所需的数据。斯坦福大学 2026 年人工智能指数报告也指出，领先的少数模型在塑造全球人工智能能力时也导致了明显的“全球语言鸿沟”，模型在英语等语言上表现远优于其他语言，同时在标准语言和方言上的表现存在显著差距。这意味着南方国家的语言数据、领域数据和本土知识数据供给明显不足，从而制约了面向本地需求的人工智能应用开发与治理能力建设，使全球南方国家在人工智能发展红利公平普惠的过程中，进一步陷入明显劣势的地位。

第三，算法鸿沟。算法和模型创新能力构成技术竞争力的核心维度。斯坦福大学《2026 年人工智能指数报告》显示，2025 年最具代表性的人工智能模型主要仍由美国、中国开发，其中美国明显领先；《2025 年人工智能指数报告》进一步表明，2024 年美国机构发布 40 个值得关注的模型，中国为 15 个，欧洲合计为 3 个，全球南方整体在基础模型研发中仍处于边缘位置。算法鸿沟还体现在人才储备的不均衡上。MacroPolo 2025 年更新的全球人工智能人才追踪研究显示，顶尖人工智能研究人才仍高度集中于美国等少数国家，同时更多人才开始留在本国而非继续流向单一目的地，这说明全球人才格局虽在变化，但总体集中态势尚未根本改变。

第四，词元鸿沟。词元（Token）指的是当前以大模型为核心的人工智能商业化主流模式，即依靠云端算力，按词元，即输入和输出的文本单位，进行计费的 API 调用模式。这种商业模式正对全球南方国家孕育着深层的、系统性的危机。首先全球南方国家在使用人工智能大模型时被强行收取了高昂的溢价。目前主流大模型的词元生成器是高度优化于英语的。一个完整的英语单词通常只需 1 个词元，但如果输入同等语义的全球南方小语种（如泰语、斯瓦希里语或阿姆哈拉语），由于训练语料匮乏，模型会将其强行拆分为数个甚至十几个无意义的字符切片。这意味着，当东南亚或非洲的企业使用本土语言调用西方主流 API 时，为了表达完全相同的意思，他们消耗的词元数量可能是英语用户的几倍。

第五，应用鸿沟。人工智能必须通过赋能产业才能真正提升生产力、实现其技术价值，但全球南方的工业基础普遍薄弱，数字化转型面临多重困境。就东盟而言，现有官方和研究性资料显示，成员国在数字化准备度与人工智能应用能力方面差异明显：东盟官方政策简报指出，区域内数字技术渗透

虽在快速推进，但各成员国进展并不同步，人工智能应用整体仍处于较早阶段；Oxford Insights 2025 年《政府人工智能准备度指数》也显示，东盟国家在人工智能应用与治理准备度方面存在明显差异，新加坡处于区域领先位置，而多数成员国仍面临政策体系、数字基础设施、技术吸收能力和治理能力不足等制约。以马来西亚和越南为例，相关研究表明，制造业数字化和人工智能赋能主要受制于基础设施质量与覆盖、企业技术吸收和利用不足、熟练技能供给短缺、融资与研发能力不足，以及产业界与科研体系协同不强等因素。联合国工业发展组织（UNIDO）的研究进一步指出，人工智能能力和资源仍高度集中于少数先进经济体，发展中国家在制造业采纳人工智能时普遍面临基础设施、技能、资金和制度环境等瓶颈。这意味着，若缺乏持续的产业能力建设与配套政策支持，人工智能带来的生产率提升红利将难以在南方国家实现广泛扩散。

第六，生态鸿沟。除前述四个维度外，数字基础设施的分布和建设能力不平等构成智能鸿沟的综合性基础条件。根据国际电信联盟《2025 年事实与数字》报告，全球仍有约 22 亿人口未能接入互联网，主要集中在最不发达国家农村地区。与此同时，人工智能发展所需的数据中心建设面临能源供应瓶颈。国际能源署（IEA）《Energy and AI》估算，2024 年全球数据中心用电约 415 太瓦时，到 2030 年可能升至约 945 太瓦时，而能源供应不稳定正是南方国家的普遍困境。人才流失则进一步加剧了上述恶性循环：非洲每年培养的计算机科学毕业生中，约三分之一流向发达国家工作，形成了系统性的智力外流。

（二）“智能鸿沟”扩大为全球南方带来的结构性劣势

第一，形成依附型发展模式。 随着智能鸿沟持续扩大，全球南方在人工智能发展中面临更强的外部依赖压力。现有国际组织研究表明，人工智能的发展和供给能力高度集中于少数国家和企业，中低收入国家在算力、数据、技能和创新生态方面普遍处于不利位置；在实际部署层面，许多国家不得不依赖外部提供的云计算服务、先进芯片、基础模型和相关数字服务。与此同时，大型平台企业在全球数据价值链中的控制力不断增强，其优势不仅体现在数据收集和处理能力上，也体现在云基础设施、模型服务和平台规则的跨层延伸上。由此产生的结果是，发展中国家可能更多停留在数据供给和技术消费环节，而较难在高附加值的模型开发、平台治理和价值捕获环节占据主导位置。

在此基础上，地缘政治竞争又进一步压缩了技术扩散空间。美国商务部在 2024 年关于先进半导体和人工智能相关出口管制的公告中，明确将相关政策概括为“小院高墙”（small yard, high fence）战略，并强调其目标之一是限制特定国家发展先进人工智能和本土半导体生态的能力。与此同时，云服务市场的高集中度、数据迁移成本和合同安排，也会提高用户转换平台的成本并加剧“锁定效应”。另外，全球人工智能模型开发的不透明度近年持续升高，尤其是训练数据和模型构建资源的信息公开。根据斯坦福大学 2026 人工智能指数报告，2025 年超九成行业知名模型未公开训练代码。除此之外，发达国家少数科技巨头凭借算力通道地位和庞大的数据网络效应，构建起跨服务的壁垒，享有根深蒂固的市场地位。在全球地缘竞争加剧、算法模型训练数据不透明以及巨头生态锁定的多重夹击下，后发国家或企业在底层半导体架构与上层人工智能大模型之间寻求技术突破的壁垒被显著抬高。受上述诸多因素制约，南方国家自主研发本土模型的成本过高，被迫选择采纳和嵌入已有的外部技术体系。而一旦南方国家在早期深度嵌入特定技术栈或平台体系，后续转向的经济和制度成本往往较高，容易形成路径依赖。综合来看，联合国贸易与发展会议的相关分析已经指出，发展中国家存在沦为“原始数据提供者”并为由其数据生成的数字智能再次付费的风险，数字经济中的高附加值收益则更可能流向少数技术领先国家和平台企业。

第二，治理话语权丧失与代表性赤字。 在规则与标准层面，智能鸿沟同样转化为治理权力的不对称分配。当前全球人工智能治理的主要规则框架，从欧盟《人工智能法案》到美国《人工智能行政令》，再到七国集团（G7）的“广岛人工智能进程”与经合组织主导的“全球人工智能合作伙伴关系”（GPAI），几乎全部由北方国家主导制定。南方国家更多扮演“规范接受者”（Rule-takers）而非“规范制定者”（Rule-makers）的角色，其国家利益、发展权诉求与本土特殊需求在规则形成期被系统性地忽视。

更深层的危机在于，人工智能正从单纯的经济生产力跃升为国家权力的核心构成要素。西方国际关系学界的最新研究表明，拥有先进人工智能能力的国家不仅能垄断数字经济的竞争主动权，更能在情报分析、自主武器系统、舆论塑造等领域获得压倒性的结构性优势。这种从“技术权力”向“综合国力”的转化，使得缺乏独立人工智能演进能力的南方国家在国际体系中的依附性进一步加深。正如联合国人工智能高级别咨询机构（HLAB）在《为人类治理人工智能》报告以及联合国贸易与发展会议多份报告中所警告的：南方国家在参与全球人工智能规则谈判时面临着严重的“代表性赤字”，普遍受制于专业谈判人才匮乏、议题设置能力薄弱以及多边参与成本高昂等结构性障碍，这使得现有的全球人工智能治理体系陷入了“少数人制定、多数人承受”的合法性危机。

二、人工智能南北合作的重要意义与现实制约

面对智能鸿沟持续扩大的严峻现实，推动全球人工智能南北合作已从选择性议题上升为必选项。合作不仅是缩小技术差距的工具，更是构建公平合理的全球人工智能治理秩序的基础。然而，南北合作的推进并非坦途，地缘政治回归、治理机制碎片化、权力结构失衡等现实制约因素相互交织，构成了复杂的合作障碍。准确评估合作的必要性与可行性，是探索有效推进路径的必要前提。

（一）全球人工智能南北合作的重要性

第一，发展层面，避免“人工智能时代的发展断层”。 人工智能技术正在成为继工业革命、电气化、信息化之后推动经济社会发展的又一根本性力量。如果智能鸿沟长期得不到有效弥合，全球南方可能被锁定在人工智能发展的边缘地带，形成与北方之间难以逾越的“发展断层”。这一断层不仅体现在经济指标上：如劳动生产率差距扩大、产业结构升级受阻，更将深刻影响南方国家实现可持续发展目标（SDGs）的进程。国际货币基金组织（IMF）2024 年研究指出，全球约 40% 的就业受到人工智能影响，先进经济体的暴露度更高，但新兴市场和发展中经济体往往准备度更低，这种错位可能加剧跨国收入差距。因此，南北合作的首要意义在于为南方国家参与人工智能发展进程、分享技术红利创造制度化渠道。

第二，治理层面，缺乏南方参与会削弱全球治理的普遍性。 人工智能治理内生于全球互联的网络空间，涉及算法偏见溯源、跨境数据流动、责任归属及致命性自主武器管控等多重跨国界议题。由于这些风险具有强烈的全球溢出效应，排他性的“小多边”或单边规制方案难以形成充分有效应对。然而，当前的全球人工智能治理体系呈现出显著的“北方中心主义”（Northern-centric）特征，规则的设计与运行高度集中于少数发达经济体，南方国家面临严重的代表性缺失。这种制度性排斥不仅动摇了全球治理框架的合法性基础，更导致既有规则在面对南方国家特定的发展阶段与社会现实时严重“水土不服”，最终令规制效力大打折扣。尽管当前已有全球南方国家发出呼吁，但部分治理倡议仍存在深层矛盾。部分国家国内人工智能实践与 “AI 向善” 理念相悖，削弱了道义的正当性，且成果缺乏约束力，难以打破现有国际分工、解决核心技术缺失问题，未能形成统一的人工智能技术发展愿景以及强有力的南南合作共识。正如联合国秘书长古特雷斯在 2024 年 9 月“联合国未来峰会”上所警告的，将发展中国家排除在外的人工智能治理架构是“不完整且不可持续的”；峰会通过的《全球数字契约》亦由此明确确立了跨越数字鸿沟的核心目标，呼吁必须依托联合国框架，构建一个真正具备普遍代表性与包容性的多边人工智能治理新架构。

第三，安全层面，能力失衡将放大系统性风险。 人工智能安全风险的分布同样呈现显著的南北不对称性。南方国家由于技术能力与治理资源的双重匮乏，更容易成为人工智能安全风险的承载地。无论是算法歧视、数据泄露、深度伪造滥用，还是人工智能赋能的网络攻击、自动化武器扩散，南方国家普遍缺乏有效的识别、防范和应对能力。更重要的是，少数国家在人工智能军事应用领域的单边优势，可能引发新的军备竞赛，破坏全球战略稳定。斯德哥尔摩国际和平研究所（SIPRI）2024 年报告警告，人工智能与传统武器系统的融合正在改变战争形态，而能力差距将进一步加剧大国与小国之间的安全不对称。因此，南北合作在安全维度上具有防范系统性风险的迫切价值。

第四，主权平等层面，要在人工智能时代构建主权平等的国际秩序。《联合国宪章》确立的主权平等原则是当代国际秩序的基石。然而，人工智能技术的发展及其引发的权力转移，正在对这一原则构成新的挑战。当技术能力日益成为国家综合实力的决定性因素时，缺乏独立人工智能能力的南方国家可能在事实上沦为大国的“技术附庸”，其政策自主性和战略选择空间受到侵蚀。在此背景下，推动南北合作不仅是技术层面的能力建设，更是维护国际关系基本准则、确保人工智能时代主权平等得到尊重的必然要求。非洲联盟 2024 年通过的《大陆人工智能战略》明确强调以非洲为中心、发展导向、注重公平与技术主权的人工智能路径，巴西政府 2024 年推出的《国家人工智能计划》（PBIA）明确宣示了拒绝数字殖民、立足本土文化、聚焦算力自主的主权人工智能愿景。这反映出全球南方越来越将人工智能能力建设视为维护主权独立与发展权的重要组成部分。

（二）全球南北合作的现实制约

第一，地缘政治回归与大国战略竞争挤压合作空间。当前国际政治中的技术竞争正在深刻塑造人工智能治理议程。各国近年来围绕关键与新兴技术、先进半导体、算力基础设施和人工智能国际竞争力连续推出战略与政策安排，凸显其将人工智能纳入大国竞争核心议程的趋势。在此背景下，南方国家更容易受到供应链重组、出口管制、投资审查和标准竞争的外溢影响，并在合作中承受选边站队压力，导致原本面向发展的合作议题被安全化、政治化。

第二，能力不对称导致合作关系失衡。 即使在以合作为名义推进的项目中，技术、资本、规则和平台仍往往掌握在北方国家或跨国企业手中。若合作机制缺少本地能力沉淀、制度共建和知识共享安排，合作很可能停留在设备采购、平台接入或短

期培训层面，难以形成真正意义上的能力转移。联合国贸易与发展会议关于包容性人工智能的研究一再强调，若不在算力、数据、技能和制度能力上同步投入，合作可能复制既有不平等，而不是纠正不平等。

第三，治理体系与制度基础薄弱。 多数南方国家在人工智能治理领域面临制度能力不足的突出瓶颈。这体现在多个层面：缺乏系统的国家人工智能战略和政策法规框架；监管机构专业能力薄弱，难以有效评估和管理人工智能风险; 数据治理体系不健全，数据确权、流通、交易规则缺失；公众数字素养和人工智能伦理意识普遍偏低。联合国教科文组织近年来通过人工智能准备度评估方法（RAM）对多国的实践进一步说明，制度准备度不足已成为许多发展中国家参与全球治理与承接技术合作的关键约束。

第四，多边机制碎片化，南方参与成本高。 当前全球人工智能治理呈现多机制并行、规则重叠与话语分散的特征。联合国体系、经合组织、G7、G20、区域组织及行业网络都在推进相关议程，但不同机制在成员构成、议题重点和规范逻辑上并不一致。联合国贸易与发展会议《2025 年技术与创新报告》第五章明确指出，当前国际人工智能治理倡议高度碎片化且主要由发达国家主导。对于资源有限的发展中国家而言，持续跟踪多个机制、协调不同规则并有效发声，本身就意味着较高的制度和人力成本。

第五，非国家行为体的数字权力增长使治理合法性面对危机。在人工智能时代，科技公司、云服务商、芯片企业、开源社区和标准组织都在不同程度上影响规则形成与技术扩散。这一变化意味着治理已经不再是单纯的政府间事务。问题在于，少数大型科技企业同时掌握模型、算力、数据和平台入口，其商业激励与公共利益并不总是一致。2026 年 3 月，OpenAI 关停 Sora 等非核心业务的算力与研发投入体现了在算力成为稀缺战略资源的背景下，巨头公司进行盈利优先资源聚焦的决策逻辑。

三、推动全球人工智能南北合作的可行路径

在明确人工智能南北合作的重要意义与现实制约之后，如何将合作愿景转化为可操作的实践方案，成为亟待回答的核心问题。当前，联合国体系正在从“原则倡议者”向“能力协调者”转型，金砖国家等多边机制为南方国家集体议价提供了新平台，而人工智能技术本身的开放共享特征也为合作提供了新的可能性空间。与此同时，算力基础设施建设、能力培养体系建设、发展导向的应用合作以及人工智能安全合作，构成了南北合作的四大重点领域。探索多元主体有序参与、线上线下协同推进的合作路径，是实现包容性人工智能全球治理的关键。

（一）构建平等的全球人工智能南北合作模式

第一，充分发挥联合国的核心作用。联合国体系凭借其普遍代表性、规范合法性和议题整合能力，在全球人工智能治理中仍具有不可替代的重要地位。2024 年，联合国人工智能高级别咨询机构在《为人类治理人工智能》最终报告中，提出建立国际人工智能科学小组、全球人工智能治理对话、能力建设网络、全球基金以及联合国秘书处相关制度支撑机制等建议。 其后，联合国会员国于 2024 年通过《未来契约》及其附件《全球数字契约》，明确提出在联合国内建立独立国际人工智能科学小组并启动全球人工智能治理对话。2025 年，联合国大会又通过相关决议，正式确立这两项机制的职权范围和运行方式；与此同时，联合国数字与新兴技术办公室于 2025 年 1 月正式启动运行，为联合国系统推进数字合作与人工智能治理提供机制支撑。 这表明，联合国正在由人工智能治理原则的倡议者，逐步转向制度协调者与能力平台。与此同时，也应看到，联合国在资源投入、执行效率和大国协调方面仍受现实条件制约，因此其更适合被定位为“平台型协调者”和“能力型赋能者”，并与区域机制、小多边平台及多元行为体网络形成互补。

第二，以“小多边”与区域机制作为联合国框架的敏捷补充。在联合国框架之外，金砖国家合作机制、上海合作组织、二十国集团（G20）以及各类区域组织也正在成为推动人工智能南北合作的重要平台。这些机制的优势在于参与国数量相对有限、利益相对集中，更容易达成实质性合作共识，推进基础设施建设和能力项目。例如，金砖国家机制在人工智能领域的合作成果显著：从成立“金砖国家人工智能研究组”（AI Study Group）到 2024 年《喀山宣言》中明确承诺加强人工智能能力建设与算力合作，金砖机制已成为南方国家集体塑造全球治理规则的重要阵地。同时，非洲联盟（AU）发布了《非洲大陆人工智能战略》与东盟（ASEAN）出台的《人工智能治理与伦理

指南》，进一步证明了区域组织在整合本土发展诉求、抵御外部规则强加方面的独特防波堤作用。

第三，构建多元主体有序参与的协作治理框架。当前，全球人工智能治理正呈现由传统政府间合作向多利益攸关方协同并行演进的趋势。《全球数字契约》明确提出，国际人工智能治理需要采取敏捷、跨学科、可适应的多利益攸关方方法；联合国人工智能高级别咨询机构也指出，当前人工智能治理面临私营部门主导与传统国家间政治体系之间的结构性张力。在此背景下，企业、学术机构、技术社群和社会组织在标准制定、能力建设和风险防范中的作用不断上升。经合组织（OECD）人工智能建议明确要求各国推动以多方参与、协商一致为基础的全球技术标准制定，联合国教科文组织也正在与政府、企业、学术机构和公民社会合作推进其人工智能伦理框架的实施。同时多个非政府组织也逐渐参与协调治理。WDO 世界数据组织 2026 年 3 月在北京成立，旨在“弥合数据鸿沟、释放数据价值、繁荣数字经济”。对南北合作而言，多元主体参与有助于更有效对接北方技术资源与南方发展需求，但也需要通过透明规则和包容性安排，防止治理资源进一步向少数技术强国和大型平台集中。

当前，人工智能南北合作正在进入一个“多层级、多主体、弱中心化”的新阶段。在这一阶段，传统的政府间合作与新兴的多元协作相互交织，联合国框架与区域机制相互补充，南北合作与南南合作相互促进。构建开放包容的治理框架，确保各类行为体都能在各自优势领域贡献力量，是推动合作深化的关键。

（二）人工智能南北合作的重点领域与实施路径

第一，算力与基础设施合作：南北合作的物质基础。 算力是人工智能发展的物质基础，也是当前南北差距最为显著的领域。推动算力领域的南北合作，可在两种模式下展开：一是算力共建模式，由南北双方共同出资建设区域性算力基础设施，实现资源共享与风险共担。二是算力公共产品模式，由技术先进国家向发展中国家提供算力资源援助，通过提供算力跨境传输、推动将基础人工智能模型和开源算力平台作为全球公共产品等方式，降低南方国家的技术获取门槛。

第二，能力建设合作：从“工具供给”转向“体系培育”。 人工智能合作不能停留在设备捐赠或短期培训层面，而应转向制度、人才和创新体系的长期培育。具体而言，应涵盖四个层面：人才培养，通过奖学金项目、联合培养计划、在线教育资源共享等方式，为南方国家培养人工智能专业人才和跨学科复合型人才；治理能力建设，协助南方国家建立和完善人工智能治理的制度框架、监管机构和政策工具，提升其参与全球治理的制度能力；法律与监管支持，帮助南方国家制定符合本国国情的人工智能法律法规，在隐私保护、算法问责、知识产权等领域构建完善的规则体系；产业孵化，通过技术园区、创新基金、创业导师计划等方式，扶持南方国家人工智能产业生态的自主成长。

第三，数据与应用合作：以发展导向重塑人工智能合作逻辑。人工智能在农业、医疗、气候应对、城市治理等领域的应用，蕴含着为南方国家解决发展挑战的巨大潜力。然而，传统的国际技术合作往往以北方国家的技术输出和南方国家的市场开放为基本逻辑，数据合作更是呈现“北方提取、南方贡献、北方受益”的不对等格局。推动数据与应用合作的南北转型，需要确立三项原则：一是发展导向，将人工智能应用合作的出发点和落脚点放在解决南方国家的实际发展需求上，而非单纯的技术输出和市场扩张；二是数据互惠，建立数据跨境流动的公平规则，确保南方国家在数据合作中不仅贡献数据，也能获取数据价值，而非单向的“数据提取”；三是多元化模型，支持面向南方国家语言、文化和发展阶段需求的本地化人工智能模型开发，避免“技术移植”带来的水土不服。

第四，安全合作：构建普惠型安全合作机制。 人工智能安全风险的分布呈现显著的南北不对称性，南方国家由于技术能力和治理资源的匮乏，更容易成为安全风险的承受者。建立面向南方国家的“普惠型安全合作机制”，是南北合作不可或缺的重要组成部分。这一机制应涵盖：南方国家人工智能安全评估能力的建设，包括技术培训、工具共享和联合研究；安全信息和威胁情报的及时通报与共享；针对南方国家特殊需求的安全解决方案开发，如面向基层治理的算法透明度工具、面向选举政治的人工智能虚假信息检测工具等。此外，在人工智能军事应用和致命性自主武器系统议题上，应确保南方国家的安全关切和立场得到充分反映，防止大国在这一领域的单边行动加剧全球安全赤字。

第四部分　全球人工智能南北合作中的中国角色

人工智能技术的快速发展正在深刻重塑国际格局，智能鸿沟的持续扩大使南北差距面临进一步固化和深化的风险。推动全球人工智能南北合作，不仅是缩小技术差距、共享发展红利的现实需要，更是构建公平合理的全球治理体系、维护主权平等国际秩序的必然要求。然而，合作的推进面临地缘政治回归、治理机制碎片化、能力不对称等严峻挑战，需要各方在联合国框架下凝聚共识，在多边平台上协调行动，以多元主体参与的方式形成合力。在全球人工智能治理与南北合作的格局重塑中，中国作为世界最大的发展中国家和人工智能技术领域的重要力量，具有成为合作枢纽的独特结构性条件。中国能否以及如何发挥建设性作用，不仅关乎全球人工智能治理的包容性构建，也关乎全球南方在智能时代的发展能力与规则参与。

一、中国推动南北合作的现实路径

从基础设施合作、多边平台塑造到安全与发展并重的负责任大国形象塑造，中国推动南北合作的多条现实路径正在逐步清晰。作为全球南方的重要组成部分和人工智能领域的重要力量，中国具备推动南北合作的基础和责任，也具备成为南北合作枢纽的结构条件。

第一，技术与产业体系完整。 中国在人工智能领域已建立起较为完整的技术和产业体系，从基础研究到应用开发、从硬件制造到软件服务、从消费互联网到产业互联网，均具备相当的规模和竞争力。根据斯坦福大学《2026 年人工智能指数报告》，中国在人工智能论文发表数量、专利申请数量、机器人安装量等指标上居全球前列，在大语言模型、计算机视觉、自动驾驶等领域的技术水平已接近或达到国际前沿。这种完整的技术和产业体系，使中国具备为南方国家提供从算力基础设施到应用解决方案的全链条合作能力。

第二，全球南方合作网络密集。 中国与全球南方国家长期保持着密集的合作关系网络，这一网络涵盖政治、经济、人文等多个维度，积累了丰富的合作经验和制度资源。无论是“一带一路”倡议下的数字丝绸之路建设，还是中非合作论坛、中阿合作论坛、澜湄合作等区域合作机制，能够为人工智能领域的南北合作提供渠道和平台。此外，中国与金砖国家、上海合作组织等南方国家主导的多边机制的联系紧密，在其中发挥着重要影响力。

第三，发展与能力建设诉求契合。 中国的发展经验和能力建设路径与全球南方国家存在较高的契合度。中国进一步深化全球发展倡议合作，对追求现代化发展机遇的南方国家具有参考价值；中国在全球人工智能治理中倡导的“尊重主权”，与全球南方强调“自主发展”的诉求形成强烈共鸣；同时，中国在应对外部技术封锁时对“核心技术自主可控”的强调，也与南方国家追求技术自主性的愿望相呼应。这种叙事契合为中国在南北合作中发挥桥梁作用提供了理念基础。

第四，多边平台具有制度嵌入力。 中国始终是联合国框架的坚定维护者，具备在全球治理核心平台设置议程的能力。2023 年 10 月，中国发布《全球人工智能治理倡议》，明确提出支持在联合国框架下成立国际人工智能治理机构；同时，中国积极依托“全球发展倡议”（GDI）和“南南合作援助基金”，将数字能力建设作为优先方向。这表明中国既有意愿也有实力，在全球人工智能治理的多边博弈中代表发展中国家发声。

在这一基础上，中国正探索推动南北

合作的现实路径。

路径一：以基础设施合作构建发展共同体。 基础设施是人工智能发展的物质支撑，也是中国推动南北合作最具优势的领域。中国应依托“数字丝绸之路”，向南方国家提供人工智能基建的资金与技术支持。具体行动包括：通过政企合作（PPP）模式，帮助非洲、东盟及拉美国家建设本土化的数据中心、超算平台与绿色能源配套设施；推动中国开源大模型（如 DeepSeek 等）及高性价比算力解决方案“走出去”，推进人工智能平权；联合开展“鲁班工坊”等数字化职业教育，帮助对象国培养本土人工智能工程师，构建以实质性技术转移为导向的命运共同体。通过基础设施合作，不仅可帮助南方国家弥合智能鸿沟，更可深化与发展中国家的利益交融，构建以发展为导向的利益共同体。

路径二：以多边合作塑造包容性治理机制。 中国应充分利用在金砖国家合作机制、上海合作组织、二十国集团等多边平台中的合作基础，推动建立更加包容的全球人工智能治理框架。在议题设置上，应将发展中国家的核心关切，如技术鸿沟弥合、能力建设支持、发展导向的治理原则，全面纳入全球治理议程；在规则制定上，应推动建立更加公平的数据治理规则、算法透明度标准和知识产权安排；在机制建设上，可推动在联合国框架下设立更具代表性的全球人工智能治理协调机构，确保南方国家的充分参与和发言权。

路径三：以安全与发展并重塑造推进全球治理合作。 人工智能安全是全球治理的紧迫议题，也是大国展示负责任形象的重要领域。中国应在推动人工智能安全合作方面发挥积极作用，包括：积极参与联合国致命性自主武器系统讨论，推动达成必要的国际规范；与南方国家开展人工智能安全能力建设合作，帮助其提升风险识别和防范能力；推动建立人工智能安全信息和威胁情报共享机制，确保安全风险在全球范围内的有效应对。与此同时，中国应明确将“发展”置于人工智能治理的核心位置，强调发展导向的治理原则，避免人工智能治理议题被安全化、政治化主导。

因此，中国呼吁推动普惠包容的南北合作网络。

当前大国战略竞争加剧的背景下，部分国家试图将人工智能治理工具化，打造排他性的技术联盟和治理“小圈子”，这不仅削弱了全球治理的普遍性，也对南北合作构成现实压力。中国在推动南北合作时，应始终坚持“去阵营化”的合作理念，践行真正的多边主义。具体体现在三个核心原则：

第一，以平等开放为前提。 中国的技术合作项目必须向所有有意愿的国家开放，避免“零和博弈”的冷战思维。合作不附加要求对象国在系统生态或国际战略上与特定阵营切割的排他性条款，保障全球南方国家在不同技术路线之间自由选择的权利。

第二，以制度自主为条件。 南北合作应尊重各国自主选择发展道路和治理模式的权利，不将特定的政治制度、治理模式或价值观念作为合作的前提条件。中国应尊重各国根据自身文化传统和法律体系探索人工智能治理模式的权利，不将特定的监管制度作为技术援助的先决条件，避免带有附加条件的合作模式。

第三，以发展目标和能力建设为导向。 南北合作的最终目标应是帮助南方国家实现自主可持续发展，而非制造新的依附关系。中国推动的合作项目应立足于南方国家的实际需求和发展阶段，以能力建设和自主发展能力提升为核心，真正实现“授人以渔”而非“授人以鱼”。

二、中国围绕人工智能国际公共产品建设的最新实践

中国以多边主义作为公共产品的制度理念；以构建开源生态作为企业出海的核心优势；以政企合作的能力建设计划作为主要任务，推动人工智能国际公共产品深耕应用场景，赋能千行百业。

（一）践行真正的多边主义是中国人工智能国际公共产品的基本理念

为世界提供国际公共产品，首先要做好国际制度建设和规则标准对接。自《人工智能全球治理行动计划》提出以来，中方围绕人工智能治理和能力建设持续推进多边合作，推动相关议题从原则倡议转向机制承接。联合国主渠道作用进一步增强，区域多边机制加快对接，跨区域合作平台不断拓展，人工智能国际治理的平台基础更加清晰。

一是全球人工智能治理框架加快构建，联合国主渠道作用进一步增强。面对人工智能治理规则碎片化、小范围机制外溢等趋势，中方坚持在联合国框架下推进人工智能治理和能力建设。第 78 届联合国大会协商一致通过中方主提的“加强人工智能能力建设国际合作”决议，聚焦能力建设、普惠发展和国际合作，获 143 个成员国共同提案或联署支持。该决议强调帮助各国特别是发展中国家加强人工智能能力建设，提升其在全球治理进程中的代表性和发言权，为后续合作提供了重要制度基础。此后，联合国大会又通过决议，决定设立人工智能独立国际科学小组和人工智能治理全球对话机制，推动人工智能治理从一般性原则讨论进入常态化议程。相关进展增强了联合国在人工智能治理中的统筹协调作用，也为广大发展中国家平等参与规则讨论、表达发展诉求和安全关切提供了制度化渠道。

二是区域多边机制承接更加扎实，紧密面向“全球南方”国家治理关切。各类区域多边机制成为承接全球治理共识，形成可操作规则的重要载体，努力回应“全球南方”国家现实的治理与发展需求。2025 年金砖国家峰会中，金砖国家人工智能治理领导人声明强调，人工智能发展应充分考虑发展中国家利益和实际需要，缩小数字鸿沟，避免新的技术鸿沟形成。2025 年上合组织天津峰会发表关于进一步深化人工智能国际合作的声明，明确提出各国享有平等发展和利用人工智能的权利，并将人工智能合作机制建设纳入峰会成果，中国上合组织人工智能合作论坛进一步发布应用合作中心建设方案，围绕数字基础设施、产业协同和人才培育提出具体安排。与此同时，区域组织拓展务实发展议程。中国东盟自贸区 3.0 版升级议定书首次设立数字经济章节，2025 年 12 月发布的《关于深化中国—东盟数字治理合作的倡议》专设人工智能治理合作章节，提出在风险应对、安全治理标准化、政策法规制定等领域加强立场协调，共同参与全球治理规则制定。区域机制的密集落地，使人工智能治理在尊重各国政策与实践差异的基础上，由原则共识逐步走向规则衔接与制度协同。

三是践行真正多边主义，吸引更多跨区域合作伙伴加入多边治理进程。中方坚持普惠包容、开放共享，推动人工智能治理与合作从单一平台向多层次、多主体、多区域协同拓展，逐步形成覆盖全球、衔接区域、联通产业的治理与合作网络。中非、中阿等跨区域合作持续深化。中国政府与非洲联盟签署科技合作谅解备忘录，将联合研究、平台共建、技术转移和科研人员交流一并纳入合作内容。2025 年 7 月启动的人工智能产业发展联盟国际网络吸纳 9 家国际机构作为创始成员，为产业界和研究机构开展跨国协作提供载体，其中中阿改革发展研究中心等平台积极推动人工智能合作与中东、阿拉伯国家数字化转型需求对接。通过不断拓展合作伙伴、丰富合作载体、延伸合作链条，中国推动形成了从全球治理倡议、区域机制建设到产业项目承接的完整治理体系，以真正多边主义凝聚治理共识、汇聚合作资源，为构建开放包容、普惠共享的人工智能国际治理新格局提供了有力支撑。

四是人工智能治理规则标准对接进一步精细化。推动人工智能普惠共享、智能向善，需要统筹发展和安全，加强国际协作与规则对接。中国东盟人工智能部长圆桌会议商定推动启动中国—东盟人工智能安全网络，合作内容包括算法安全评估、模型风险预警、深度伪造检测和应急处置协作。中方发布的关于深化中国东盟数字治理合作的倡议，也将人工智能治理作为专门内容，提出在风险应对、安全治理标准化、政策法规制定等方面加强沟通交流，并在联合国等多边平台加强立场协调。2025 年中非互联网发展与合作论坛发布《携手构建中非网络空间命运共同体行动计划（2025—2026）》，将数字经济发展、网络安全保障、数字能力建设和人工智能治理纳入同一框架。双方同意在人工智能可能带来风险的应对、安全治理标准化等方面加强沟通交流和务实合作，共同防范人工智能技术滥用。世界互联网大会发布《全球人工智能标准发展报告》，为各国了解人工智能标准化进程、参与标准讨论和推进规则衔接提供参考。《金砖国家领导人关于人工智能全球治理的声明》、上合组织有关声明等一系列工作有助于把安全伦理要求转化为更具体的标准讨论和规则对接，减少不同国家在安全评估、数据治理和技术应用中的规则落差。

（二）构建开源生态是中国人工智能国际公共产品的核心思路

围绕基础模型获取成本较高、低资源语种支撑不足、本地开发能力薄弱等问题，开源模型、开发平台和评测工具不断丰富，为

全球开发者特别是发展中国家提供了更多低门槛的人工智能技术资源。

一是开源模型供给规模持续扩大。阿里通义千问系列形成覆盖不同参数规模、多模态任务和行业需求的开源矩阵，截至 2026 年 2 月累计开源超过 400 款，Hugging Face 全球下载量突破 10 亿次、衍生模型逾 20 万个，自 2025 年下半年起连续位居该平台月度下载榜首，超越美国 Meta 公司 Llama 系列，并已被比亚迪、Rokid 等企业用作落地大模型的首选基座。MiniMax 截至 2025 年底服务个人用户超过 2.36 亿、覆盖 200 多个国家和地区，向 100 多个国家和地区的 21.4 万名企业客户与开发者开放 M 系列推理模型和 Hailuo 视频模型；其 M2.5 模型在海外开发者平台 OpenRouter 上线不足一周即登顶周调用量榜首，一周贡献 token 达 1.44 万亿个。中国开源模型以开源、低成本、高效能的特点，正由全球开发者的“备选项”变为“常用项”。

二是开源社区建设加快迭代拓展。阿里云魔搭社区截至 2025 年底已托管超过 12 万个开源模型，注册用户超过 2000 万，覆盖全球 200 多个国家和地区。平台化生态集成模型、数据集、工具链和开发者资源。中国信通院牵头制定的人工智能国际标准在 ITU-T 发布，涉及基础模型平台、代码生成和检索增强生成等方向，为模型平台建设和服务交付提供技术参照。LaoBench 等低资源语种评测数据集建设，也补齐了东南亚语言在大模型评估中的短板。开源框架、开放计算平台、开源互联协议等多层次成果不断涌现，以开源开放为特征的新型智算生态加速形成，显著降低了中小企业、本地开发者和科研团队的使用门槛，推动开源成果在更广范围内更新、复用与共享。

三是本地化部署成为开源模型出海的主流做法。低资源语种和本地文化适配，是许多发展中国家发展和应用人工智能面临的突出挑战，而中国企业持续推进开源大模型，为这些国家开展本地化模型建设提供了可获取、可修改、可部署的底座能力。新加坡国家 AI 计划机构基于 Qwen3-32B 构建 Qwen-SEA-LION-v4，提升模型对东南亚语言和文化语境的理解能力；乌干达本土团队基于 Qwen 3 打造 Sunflower 模型，覆盖 31 种乌干达语言，并面向政务、医疗、教育和社区服务等场景开展应用；马来西亚推出的 NurAI 基于中国开源模型建设，面向伊斯兰法律、金融和医疗等场景提供合规建议；老挝国家大模型 LaoAI 依托中国老挝人工智能创新合作中心推进建设，基于老挝语语料开展训练，服务本国语言环境下的人工智能应用，并同步推进算力基础设施、模型研发和人才培养合作；阿联酋穆罕默德·本·扎耶德人工智能大学和相关机构则基于 Qwen 架构发布 K2Think 推理模型，以较小参数规模开展复杂数学推理任务。中国推进开源大模型，不仅扩大了基础模型的全球供给，更直接推动合作国利用开源模型开展本地训练、语言适配和行业开发，将通用模型能力转化为符合本国语言、制度和产业需求的人工智能能力，并在实践中积累模型开发、部署和迭代经验，提升本地人工智能发展的自主能力。

（三）政企合作的能力建设是中国人工智能国际公共产品的主要任务

人工智能作为重要国际公共产品，要真正实现普惠共享、持续赋能，不能只依赖模型和应用项目，还需要云服务、数据处理、本地技术体系、人才队伍和能源保障等基础条件作为支撑。当前，中国推动人工智能国际合作，正从单点应用供给逐步拓展至产业底座和能力体系建设，通过提供更加完善的公共性支撑，帮助合作伙伴夯实人工智能发展基础，更好共享智能时代发展机遇。

一是智能基础设施建设加快推进。云计算、数据中心和本地数据处理能力，是人工智能应用落地的基础。腾讯云宣布在沙特建设其中东首个云区域，并计划投入超过 1.5 亿美元用于本地云基础设施和技术建设，服务云计算、人工智能、数字媒体、电商和金融服务等应用。阿里云启用迪拜第二座数据中心，并与中东金融、医疗、数字人和云游戏等领域企业建立合作生态，进一步提升区域云服务能力。TikTok System Thailand 获批大型数据基础设施项目，将在泰国曼谷、北榄府和北柳府增加服务器，扩展数据存储和处理基础设施，支撑数字服务需求增长，补强了合作方发展人工智能所需的云服务、数据处理和算力承载条件，也为相关国家建设区域数字基础设施枢纽提供支撑。

二是赋能自主技术生态建设。包容普惠的人工智能国际合作，更有赖于帮助合作方建成本地化、可持续的技术体系。新加坡国家人工智能计划机构基于阿里 Qwen3-32B 构建 Qwen-SEA-LION-v4，取代此前采用的 Meta Llama 架构，以更好适配东南亚语言文化，且可在消费级设备上运行，便于缺乏算力集群的区域中小企业和开发者使用。中老联合声明提出支持老方推进国家数据中心、云计算系统、现代通信网络等数字基础设施建设，并加强人工智能、网络安全和科技创新合作。中老双方依托中国—老挝人工智能创新合作中心推进算力基础设施、老挝语大模型研发基地和行业模型孵化建设，2025 年 9 月正式发布以老挝语语料训练的国家大模型 LaoAI，服务本国民众和产业发展。在柬埔寨，ByteDC 基于华为云 Stack 推出高性能主权云，面向政府部门、监管机构、银行和企业提供境内可信云基础设施。在泰国，华为云支持公共部门部署人工智能应用，合作对象包括泰国国家电子与计算机技术中心、食品药品监管部门和商务部，应用方向涵盖会议语音识别、政务流程提效和监管材料分析。在巴基斯坦，中巴命运共同体行动计划将软件、人工智能、5G、云计

算和大数据等领域合作纳入数字经济内容，同时提出强化网络和数据安全机构交流，推动数字安全产业合作。从开源模型本地化到主权云和政务应用，中国正帮助合作方把人工智能能力沉淀为自主可控的技术体系。

三是人才培养和技能建设持续深化，为自主能力提升提供长期支撑。2025 年 10 月，第三期人工智能能力建设研讨班在上海开班，来自 21 个国家的 38 名学员参加，覆盖亚洲、非洲、拉丁美洲和欧亚地区，内容围绕普惠发展、全球治理、伦理治理、产业应用和教育赋能展开，并安排学员参访科研机构和企业，为发展中国家政府官员系统理解技术趋势和治理议题提供了平台。面向青年和专业群体的培训同步铺开：华为与秘鲁政府签署协议，计划培训 2 万名秘鲁青年；天津理工大学等中方机构支持哥伦比亚教育部开展教师人工智能培训，首批面向 125 名基础教育和高等教育教师，重点覆盖偏远地区和参与 STEM 教育项目的群体；华为与柬埔寨教育部签署数字教育合作备忘录，推进学校 ICT 基础设施、师生数字技能和学校管理系统升级；阿里云千问大会首次在海外举办并落地新加坡，与新加坡职工总会旗下机构合作，帮助本地企业、开发者和学生掌握生成式与代理式人工智能技能；中非互联网发展与合作论坛则推动网信领域培训机制长效化，持续开展网络安全、数字经济和人工智能治理研修。这些项目把能力建设从设备和平台延伸到人力资本与组织能力，为合作方长期使用和自主创新储备了人才基础。

四是能源和行业基础设施智能化水平提升，为人工智能应用提供稳定运行环境。人工智能发展离不开稳定的能源系统和基础行业支撑。国家电网 2024 年 12 月发布“光明”电力大模型，这是中国电力行业首个千亿级多模态行业大模型。2025 年，该模型拓展至巴西电网运维场景，服务国家电网巴西控股公司输电网络智能化管理。公开报道显示，国网巴控运营的输电网络总长超过 1.6 万公里，覆盖美丽山二期特高压输电项目等关键资产。针对巴西输电线路距离长、地形复杂、部分地区网络条件薄弱、传统人工和直升机巡检效率较低等问题，相关平台将人工智能图像识别、无人机巡检、移动离线作业和缺陷闭环管理结合起来，提升输电巡检效率和运维标准化水平。中国务实推进电力基础设施合作正在从工程建设向智能化运维服务升级，也说明行业大模型、巡检数据和运维平台可以嵌入海外能源系统，为大型跨境能源项目提供智能化管理能力。能源和电力系统运行效率提升，也为人工智能应用和数字经济发展提供更加稳定的基础环境。

（四）深耕产业场景应用是中国人工智能国际公共产品的价值所在

人工智能赋能发展的作用更加凸显。当前，人工智能技术交流和能力建设逐步向实际应用拓展，应用领域不断丰富，推动人工智能在改善民生、提升治理效能、促进产业升级等方面发挥积极作用。

一是服务民生民用成效显著。中国气象局发布的“妈祖”早期预警系统，以云端系统和可定制方案应对多灾种预警需求，截至 2026 年 4 月已在巴基斯坦、所罗门群岛等 7 个国家落地部署，支撑全球 40 多个国家开展云端应用，向 153 个国家和地区提供百余种数据产品，国际培训覆盖 89 个国家和地区。科技部主导的人工智能气象预报应用示范项目计划在不少于 6 个国家部署运行，覆盖 1000 万预警人口，将人工智能模型、气象数据与预警服务直接用于防灾减灾和公共安全，是人工智能服务民生的代表性案例。农业领域的合作同样回应了一线生产痛点。巴西国家半干旱研究所与中国农业大学共建家庭农业机械化与人工智能实验室，面向巴西半干旱地区的家庭农业生产者而非大型商业农场，围绕环境监测、数据分析和农机应用开展合作。润建股份为马来西亚智慧棕榈园提供解决方案，针对油棕成熟窗口短、漏采损失高的难题，以人工智能算法识别果实熟度、无人机高光谱相机标记熟果位置、智能车辆精准采收，使产量提升 10% 以上，获马方行业组织评价为一次深刻的农业科技革命。

二是提升公共治理效能，赋能智慧城市建设。华为云在泰国公共部门推动人工智能部署，合作方向涉及会议语音识别、政务流程提效、监管材料分析和公共服务优化。泰国食品药品监管部门将 AI 用于产品注册材料分析、筛查和摘要生成，以提升监管效率。深兰科技参与乌兹别克斯坦新塔什干智慧城市项目，围绕智慧城市、智慧住宅、国际物流和智慧环卫等方向开展合作。深兰科技还与伊拉克相关机构开展合作，涉及基础设施重建、石油炼化智能化升级、数字银行支付系统、多级智慧医疗和 AI 人才培养等领域。相关案例表明，人工智能正在进入地方治理和公共服务的具体环节。

三是行业应用和智能装备加快拓展。华为在吉隆坡 TRX 设立了占地 13638 平方英尺的 AI Lab，系其全球范围内首个设于中国境外的 AI 赋能产业孵化基地，其解决方案已覆盖约 97% 的马来西亚人口。阿斯利康中国基于通义千问和阿里云平台构建药物安全报告工具，用于从医学文献中识别不良事件信息并生成报告。该工具将报告准确率由 90% 提升至 95%，报告生成效率提升约 300%。阿里云与 Wio Bank 合作开发基于通义千问大模型的 AI 银行智能体；在人工智能领域，与 BYOND Asia 共同打造面向中东市场的阿拉伯语数字人解决方案；在医疗健康领域，支持 ACCUMED 建设医疗保险理赔、医学编码等 AI 应用；在数字娱乐领域，为 The Game Company 提供低延迟、高性能云基础设施，推动 AI 云游戏发展；同时，与全球数字化服务商 Atos 深化合作，

将云计算和人工智能产品推广至中东及亚太企业市场。这些合作不仅体现了中国企业从传统基础设施输出向“云计算 + 人工智能 + 行业解决方案”综合能力输出的转型，也依托本地数据中心建设和生态合作模式，增强了技术服务的本地化能力和国际竞争力，为中东地区数字经济和人工智能产业发展提供了重要支撑。西井科技 Q-Truck 无人驾驶车队在英国菲力斯杜港扩容，此前首批车辆已在港口混行作业，被视为欧洲港口无人驾驶卡车应用的重要案例。傅利叶智能携 GR-3 人形机器人亮相 CES 2026，展示具身智能在康复、养老和陪伴交互等场景中的应用潜力。商汤科技与沙特国王大学共建人工智能联合创新中心，探索教育、科研和多行业应用场景。

总体看来，自《人工智能全球治理行动计划》提出以来，中国推动人工智能国际公共产品建设取得积极进展。相关工作并非单纯提供某一类技术产品，而是围绕发展中国家参与人工智能治理和使用人工智能技术的现实短板，逐步形成了较为系统的合作路径。在制度层面，中方依托联合国和区域多边机制推动议题对接，使人工智能治理更加重视“全球南方”国家的发展诉求。在生态层面，开源模型和平台生态降低了技术使用门槛，也为低资源语种和本地模型建设提供了基础。在能力建设层面，智能基础设施、人才培养和能源行业智能化项目，为合作方长期使用和自主发展人工智能提供了支撑。在应用层面，气象预警、智慧农业、公共治理和行业应用等项目已经进入具体场景，开始形成可感知成果。现阶段，人工智能国际公共产品建设仍处于持续推进阶段，部分平台和项目还需要在后续运行中检验成效。其关键不在于一次性输出技术，而在于帮助合作方逐步形成可持续的发展能力。下一步，随着多边机制持续运行、开源生态继续扩大、场景项目不断落地，中国推动人工智能服务共同发展的实践基础有望进一步夯实，也将为缩小全球智能鸿沟、提升发展中国家数字治理能力和推动人工智能向善发展提供更多支撑。