人工智能国际治理：“安全偏好”及其现实影响

李　艳

内容提要：当前，新一轮人工智能发展浪潮正席卷全球，虽然从技术成熟度、应用落地规模化等方面看，它仍处发展初期，但作为驱动第四次科技革命的核心技术与应用，其潜在的变革性力量已受到国际社会各方高度关注，尤其是安全关切格外突出、安全治理呼声日益高涨。由此，人工智能国际治理进程整体呈现显著的“安全偏好”，其背后的底层逻辑涉及发展历程、技术特质与地缘政治等诸方面。这种“安全偏好”在很大程度上决定了人工智能国际治理的认知理念与实践重心，并会给人工智能发展前景乃至国际力量格局带来深远影响。

关键词：人工智能 | 国际治理 | 安全偏好

作者介绍：李艳，中国现代国际关系研究院科技与网络安全研究所所长、研究员，主要研究方向为科技与网络安全战略政策及国际治理。

新一轮人工智能技术与应用方兴未艾，但安全关切不断攀升，国际社会各方未雨绸缪，积极投身应对潜在的风险隐患。这种关切投射到人工智能国际治理进程中，表现为明显的“安全偏好”，背后深层次根源在于人工智能技术与发展历程的特殊性及地缘政治竞争态势的影响。多层因素塑造下的“安全偏好”将对未来人工智能国际治理走向产生深远影响，同时也为更好参与相关进程带来启示。

一、人工智能国际治理的“安全偏好”

随着新一轮人工智能发展浪潮汹涌推进，人工智能国际治理亦成为一项新兴全球性治理议题。从促进发展维度看，“由于人工智能行业是全球性的，其网络和计算资源遍布许多国家，国际合作至关重要”。[①] 从确保安全角度看，“人工智能带来的风险可能造成各种社会、经济和道德风险，忽视这些风险将对全球产生重大影响并阻碍技术进步”。[②] 为此，“必须制定框架指导人工智能发展，以平等造福全人类，国际治理在确保这一点上发挥着至关重要的作用”。[③] 此轮人工智能技术与应用发展浪潮，无论在技术成熟度还是应用落地规模化等方面都如日方升，相应的国际治理亦刚刚起步。与历史上相应的国际治理经验与一般规律不同，相较于技术应用发展初期惯常的“发展偏好”，当前人工智能国际治理进程呈现鲜明的“安全偏好”。

国际治理“偏好”是指治理理念与实践推进的阶段性重心向某方面倾斜。全球性议题的国际治理核心目标均为实现发展与安全的平衡，但绝对的平衡只是一种理想状态。所谓的“平衡点”在实践中始终处于动态变化，必然会出现相对而言的实际“偏好”，即阶段性治理重心或更趋向发展、或更突出安全。值得注意的是，理解“国际治理偏好”有两个要点：其一，基于不同历史阶段，如技术与应用发展初期与成熟期的偏好必然不同；其二，基于相对视角，毕竟发展与“安全偏好”从来都是相对的，不是非此即彼。

新兴技术应用的国际治理具有一定规律性：初期多呈现“发展偏好”，即理念认知、机制建设与政策措施更多强调促进发展；随着技术成熟，尤其是社会应用的推广与普及，逐渐显露出更多“安全偏好”。这符合事物发展的客观

① Robert Trager et al., “International Governance of Civilian AI: A Jurisdictional Certification Approach,” August 31, 2023, https://arxiv.org/pdf/2308.15514.

② See Esmat Zaidan and Imad Antonie Lbrahim, “AI Governance in a Complex and Rapidly Changing Regularory Landscape: A Global Perspective,” *Humanities and Social Sciences Communications*, 2024, pp.1-18.

③ See David Leslie et al., “‘Frontier AI’ , Power, and the Public Interest: Who Benefits, Who Decides?” *Harvard Data Science Review*, September 9, 2024, pp.1-20.

规律：一方面，在技术与应用初期，为更快促进发展，安全顾虑往往让位于发展优先，各方也愿意承担更多安全风险；另一方面，很多具体的安全问题一般出现在技术与应用落地后，并且只有当这些问题的社会影响与后果足够引起关注时才会被纳入治理议程或成为治理重心。围绕互联网技术与应用展开的网络空间国际治理就是其中典型案例。20 世纪 90 年代互联网发展初期，国际社会各方聚焦尽快推进互联网技术与应用的全球普及，以期共享互联网红利、推进人类社会向信息社会迈进。在此阶段，对网络安全的认知与实践仅是确保技术架构本身的安全与稳定，核心目标是确保发展顺利推进。随着互联网的社会应用逐步深化，其带来的全球性安全问题开始显现并逐渐得到重视，标志性事件就是 2003 年联合国框架下的信息社会世界峰会（WSIS）进程开启。它推动网络安全治理从“以技术为中心”向应对更多社会问题的“综合治理”转型。2013 年“斯诺登事件”后，网络空间治理中的国家安全关切上升，地缘政治对网络安全的影响凸显，这不仅掀起各国制定与更新国家网络安全战略的热潮，也催生了更多安全议题的国际治理进程。

不同于上述一般性规律，当前尚处初期的人工智能国际治理进程表现出鲜明的“安全偏好”。一是学术界研究视角与内容呈现全面的安全观。他们既从技术应用方面入手，从不同影响领域维度探讨人工智能带来的各种安全风险与威胁，如政治、经济、社会以及意识形态安全等；也从国际政治视角切入，将人工智能视为未来国家实力乃至国际力量格局的重要因素或变量之一，并上升到大国竞争与博弈背景下更加高维的国家安全。2024 年 4 月斯坦福大学人工智能研究所（Stanford HAI）发布《2024 年人工智能指数报告》（Artificial Intelligence Index Report 2024），其中对 2019 年—2023 年人工智能领域学术会议安全类论文提交数量进行统计，结果显示，2023 年提交数量较 2019 年总体增长 70.4%。[①] 这说明，随着此轮人工智能技术突破与应用落地，学术界对人工智能安全问题的关注显著攀升，其潜在安全风险渐成学术研究的热点和

① *Artificial Intelligence Index Report 2024*, Stanford Institute for Human-Centered Artificial Intelligence, 2024, p.187.

技术应用中的重要挑战。

二是政策界重在提出安全治理原则与框架设计。国际组织及相关国家在推进人工智能治理的各项国际议程与政策文件中，除使用“安全”字样外，还包括诸如“可靠”“可信”“向善”“负责任”等高频词，充分表明各方对人工智能技术与应用发展应有目标的期待或规划，即都基于不同层次的安全考虑。从联合国层面，联合国秘书长古特雷斯 2023 年宣布成立人工智能高级别咨询机构（High Level Advisory Body on AI），[①] 并在 2024 年发布《为人类治理人工智能》（Governing AI for Humanity）最终报告，提出加强全球合作的行动方案。同年，联合国大会通过了“加强人工智能能力建设国际合作”（Resolution on Enhancing International Cooperation for AI Capacity Building）和“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”（Resolution on Seizing the Opportunities of Safe, Secure and Trustworthy Artificial Intelligence Systems for Sustainable Development）两项决议。一些重要的国际组织与机制亦将人工智能视为重要议题，例如，2023 年 6 月，世界经济论坛成立人工智能治理联盟（AI Governance Alliance），提出负责任地开发、开放创新和国际合作等建议；9 月，二十国集团在新德里峰会上重申“以人为本，实现人工智能向善并服务全人类”[②] 的观点，提出相关治理框架并呼吁全球监督。此外，国际社会各方还积极搭建新的人工智能治理平台与机制。例如 2023 年 11 月，人工智能安全峰会进程开启，首次会议在英国布莱切利举办，会上发表了《布莱切利宣言》（Bletchley Declaration），强调科技发展应促进人类共同福祉和环境可持续性的重要性，并呼吁在全球合作、确保技术积极影响最大化的同时，“减少潜在风险和负面

① 《秘书长组建高级别咨询机构，全球39名专家共商人工智能治理》，联合国新闻网，2023年10月26日，https://news.un.org/zh/story/2023/10/1123382。

② 《二十国集团领导人新德里峰会宣言（摘要）》，https://www.gov.cn/yaowen/liebiao/202309/content_6903173.htm。

影响”。[①] 2023 年 10 月 18 日，中国发布《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的中国方案。

三是产业界积极探索保障人工智能安全的最佳实践。在推进应用落地过程中，业界在人工智能产品与服务的设计与推出上更强调安全性。例如，全球移动通信系统协会（Global System for Mobile Communications Association，GSMA）2023 年 9 月推出《负责任的人工智能成熟度路线图》（Responsible AI Maturity Roadmap），为电信运营商提供适配工具和指导，助其评估负责任的人工智能成熟度水平。 2024 年 2 月，在德国慕尼黑安全峰会上，包括亚马逊、谷歌、IBM、Meta、Microsoft、OpenAI、TikTok 和 X 等在内的 20 家科技公司共同签署《打击在 2024 年选举中欺骗性使用人工智能的技术协议》（The Tech Accord to Combat Deceptive Use of AI in 2024 Elections），抵制欺骗性人工智能生成内容、减少其带来的风险，并同意在各自平台或产品中提出解决方案。该协议还承诺，将与全球组织和学术界合作，努力让公众和媒体意识到人工智能生成欺骗内容的危险性。[②] 2024 年 5 月，人工智能首尔峰会起草了《前沿人工智能安全承诺》（Frontier AI Safety Commitments），[③] 敦促签署者在开发和部署前沿人工智能模型及系统时负责任地管理风险，来自中国、美国、欧洲、中东和亚洲其他国家的 16 家人工智能公司或组织签署承诺。签署者自愿承诺实施与前沿人工智能安全相关的一系列最佳实践，包括内外部红队测试、信息共享、网络安全投资、第三方漏洞报告机制以及公开透明地报告自身安全框架和风险管理方法等。

① “AI Safety Summit 2023: The Bletchley Declaration,” November 2023, https://www.gov.uk/government/publications/ai-safety-summit-2023-the-bletchley-declaration.

② 《亚马逊谷歌IBM微软等巨头刚刚承诺一同对抗人工智能选举干扰》，2024年2月19日，https://36kr.com/p/2654057871035654。

③ 《中美欧等16家公司共同签署〈前沿人工智能安全承诺〉》，国际科研合作信息网，2024年5月29日，http://www.ircip.cn/bbx/993897-1892867.html?id=26645&newsid=5496323。

二、“安全偏好”形成的底层逻辑

此轮人工智能国际治理之所以在初期就呈现出明显“安全偏好”，背后原因多元复杂，主要涉及发展、技术与政治三大底层逻辑。发展逻辑上，此轮发展浪潮打通了人工智能技术走向广泛应用的重要环节，技术突破与应用落地同时铺开，客观上加速安全议程的前置。技术逻辑上，人工智能技术作为一项前所未有的特殊技术形态，高度的不确定性甚至不可控性带来诸多新型安全挑战，这些挑战也成为国际治理的重心。从政治逻辑看，地缘政治竞争态势加剧背景下，政治因素高度内嵌于人工智能技术与应用发展全过程，安全成为博弈的工具和打击对手的由头，不断影响并塑造着国际政策环境。

（一）发展逻辑：应用环节打通加速安全议程前置

历史上的新兴技术与应用多呈“线性”发展，即：技术与应用持续推进，安全议题相对置后；随着技术应用逐渐落地并渐进式呈现，安全问题才陆续进入国际治理议程。人工智能发展迥异于以往，呈现出“非线性”发展脉络。在不同历史时期，人工智能虽都曾有阶段性突破或进展，但这项高度融合性技术需要跨学科及众多关联技术支撑，如算力、算法以及神经网络等脑科学研究等，因此往往受制于技术条件而多轮整体停留在实验室或有限的特定领域，更没有打通从技术突破到社会广泛应用的路径，也未能对社会形成广泛、变革性影响。

人工智能的源起与发端可追溯到 20 世纪中叶：沃伦·麦卡洛克（Warren McCulloch）和沃尔特·皮茨（Walter Pitts）1943 年设计出第一个人工神经元模型；艾伦·图灵（Alan Turing）1950 年提出测试机器是否具备人类智能的图灵测试方法；1956 年达特茅斯会议首次提出“人工智能”一词。之后，人工智能经历了多轮发展小高潮，如 1960 年—1970 年出现了早期聊天机器人与专家系统的雏形。随着计算机技术的出现与飞速发展，人工智能技术进一步推进，标志性事件是 1997 年 IBM 的“深蓝”（DeepBlue）战胜了国际象棋世界冠军加里·卡斯帕罗夫（Garry Kasparov），展现了人工智能潜在的、强大的

思维分析能力。此后 2004 年，科学界开始流传有关生成式对抗网络（GAN）的想法；2006 年，杰弗里·辛顿（Geoffrey Hinton）将深度学习推向风口浪尖，人工智能技术创新开始加速。2011 年，IBM 公司研发的电脑“沃森”（Watson）战胜美国电视智力节目《危险边缘》的两位“常胜将军”，显示人工智能在理解和处理自然语言方面取得巨大进步。2014 年，伊恩·古德费洛（Ian Goodfellow）及其团队正式提出生成式对抗网络概念，创造了一种革命性工具，大幅促进了人工智能领域的创造力和创新。

近年来，在大数据、算法与算力等技术与应用加持下，人工智能技术出现质的飞跃，诞生了联通社会应用的技术路线。2022 年以 ChatGPT 为代表的通用式大模型横空出世，直接推动人机交互进入新时代，打通了技术创新到社会应用的“最后一公里”。短时间内，OpenAI 的 ChatGPT-4、谷歌的 Bard、微软的 Bing AI 以及中国深度求索公司发布的 DeepSeek-R1 等大模型如雨后春笋般纷纷露面，各种基于通用模型的应用场景也全面涌现，人工智能潜在的颠覆性、变革性力量开始显现。例如，AlphaFold 3 对生物学的颠覆，它以前所未有的“原子精度”一夜间预测出所有生物分子的结构和相互作用。[①] 再如，材料搜索图形网络（GNoME）对材料学的颠覆，它成功预测了 220 万种晶体结构，[②] 其中 38 万种特性最稳定的晶体结构有潜力成为未来变革性技术的材料，为超导体、电动汽车电池研发及超算供电等领域提供了动力。目前，科学家们已开始在 GNoME 辅助下进一步合成新材料。

此轮人工智能浪潮最大的特点是“技术突破—社会应用—重大影响”的闭环已然形成。事实证明，随着新一轮人工智能技术迅捷落地、社会应用场景全面铺开，其所带来的安全风险与问题亦全面呈现，涉及个体隐私、伦理风险、社会公平、就业替代、军事安全、政治安全等诸方面，已成为国际治理需要统筹考虑并解决的重要现实问题。因此，应用环节的打通促进人工智能国际治理

① Ewen Callaway, “Major AlphaFold Upgrade Offers Boost for Drug Discovery,” *Nature*, 2024, pp.509-510.

② Amil Merchant, Simon Batzner, Samuel S. Schoenholz, Muratahan Aykol, Gowoon Cheon, and Ekin Dogus Cubuk, “Scaling Deep Learning for Materials Discovery,” *Nature*, 2023, pp.80-85.

的安全议程明显加速前置。

（二）技术逻辑：特殊的技术形态带来新型安全挑战

人工智能通常被视为新兴技术，而新兴技术一般具有新颖性、快速增长性、连贯性、重大影响性、不确定与模糊性等五大特质。其中，重大影响性、不确定性与模糊性揭示的是技术应用的显著特点。前者指对社会影响重大，后者则意味不确定这种影响对社会是否必然有利，不排除会带来一些“意想不到”（Unintended）甚至是“不希望看到”（Undesirable）的结果。[①] 即便如此，越来越多的专家学者仍认为，仅从新兴技术视角理解人工智能远远不够，“人工智能不同于我们已经探索过的其他工具或领域，如核武器或空间问题等，它带来的问题往往前所未见”。[②] 尤其是与其他新兴技术与应用相比，人工智能独有的智能涌现性带来的问题可能超出人类已有的历史经验，“人工智能具有自我创造、超强学习和超级进化的特性”。[③] 这种技术应用特性基本没有治理经验可以借鉴，自然而然加剧了人们的安全关切，安全问题也因此成为人工智能安全治理的重心。

源生于人工智能技术特性的安全风险大致可以分为以下三类：一是升级型风险。但凡技术都具有“双刃剑”效应，人工智能技术也不例外。作为一项赋能技术，它会加剧一些既有安全风险，如人工智能技术的“越狱”（Jailbreak）网络攻击带来更加严峻的现实威胁。人工智能技术的快速发展增加了其系统的复杂性和不可解释性，使得系统中的安全漏洞更难被发现。“越狱”攻击者可通过精心设计的输入，绕过人工智能系统的安全限制与规则约束等防护机制获得操纵系统的能力，从而获取用户敏感数据、制造恶意虚假信息、输出不良有害内容、发动网络攻击和协助真实犯罪等。例如，微软 2024 年推

① Daniele Rotolo, Diana Hicks, and Ben R. Martin, “What Is an Emerging Technology?” *Research Policy*, Vol.44, 2015, pp.1827-1843，https://arxiv.org/pdf/1503.00673v2.

② 《人工智能在军事领域发展前景如何？专家解读》，央视新闻，2024 年 9 月 14 日。

③ 薛澜、赵静：《人工智能国际治理：基于技术特性与议题属性的分析》，《国际经济评论》，2024 年第 3 期，第 56 页。

出一种人工智能“越狱”技术 Skeleton Key，该技术可绕过多个人工智能系统中的保护机制，迫其生成违反道德伦理和社会公序良俗的内容。①

二是难控型风险。如人工智能的幻觉（Hallucination）问题。人工智能幻觉指，人工智能系统输出内容与实际情况偏离，输出内容看似流畅合理，实则缺乏可靠依据、没有意义甚至完全错误。这主要是因为在构建人工智能系统模型时，实际训练数据存在偏差且很难涵盖所有场景，训练模型性能的不断提升导致其过度拟合，自然语言的多义性和复杂性使得系统缺乏对语义的深度理解。幻觉问题在文本生成任务中更为普遍，具体表现为人工智能系统“一本正经地胡说八道”，如在新闻报道撰写中虚构不存在的事件细节从而极大影响用户使用体验等。② 2023 年，谷歌公司人工智能系统 Bard（Gemini 前身）在公开演示中回答称“詹姆斯·韦布空间望远镜（JWST）首次拍摄到太阳系以外行星的照片”，事实上该答案是错误的，人工智能因幻觉问题提供不准确信息也因此受到广泛关注。③ 幻觉问题不太可能随着人工智能技术的发展自行消失，只能通过持续技术改进逐步缓解，这将在金融投资建模、自动驾驶、医疗诊断等高风险场景中带来严重安全隐患。

三是不可控型风险。如人工智能的涌现（Emergence）问题。近年来，不断有实践案例表明，随着模型规模（如训练计算量或参数数量）的增加，人工智能系统会表现出涌现现象——当模型达到某个临界规模后，其能力会突然从接近随机的水平跃升至远高于随机的水平。研究表明，该现象与模型规模密切相关，并且在处理多样化任务时表现得尤为突出。这种能力在本质上无法解释、不可预测甚至不受控制。从积极角度看，这可能是最具创造性的智能体现，

① Chris Mckay, “Microsoft Reveals ‘Skeleton Key’: A Powerful New AI Jailbreak Technique,” June 28, 2024, https://www.maginative.com/article/microsoft-reveals-skeleton-key-a-powerful-new-ai-jailbreak-technique/.

② See Ziwei Ji et al., “Survey of Hallucination in Natural Language Generation,” *ACM Computing Surveys*, 2023, pp.1-38.

③ Carrie Mihalcik, “Google ChatGPT Rival Bard Flubs Fact About NASA’s Webb Space Telescope,” February 9, 2023, https://www.cnet.com/science/space/googles-chatgpt-rival-bard-called-out-for-nasa-webb-space-telescope-error/.

或许是未来科学进步的巨大动力；从消极角度看，这或许也是智能机器控制人类场景的现实基础。

人工智能技术的“失控感”让国际社会各方认为，人类对于人工智能技术未来的掌控与治理准备不足。这也是为什么在 ChatGPT 推出之后，国际社会一度出现“喊停”声音，避免技术失控带来的重大安全风险成为人工智能发展的优先事项。2023 年 5 月，超过 350 名人工智能领域的行业高管、专家和教授在非营利组织人工智能安全中心（CAIS）网站上签署一份公开信称：“减轻人工智能带来的灭绝风险，应该与流行病和核战争等其他社会规模的风险一起，成为全球优先事项。”[①]《前沿人工智能安全承诺》签署方亦承诺，设定并评估“不可容忍”的风险阈值，在风险过大时暂停开发或部署人工智能模型和系统，以确保技术的安全和可信等。2023 年 7 月 18 日，联合国安理会举行首次题为“人工智能给国际和平与安全带来的机遇与风险”高级别公开会议指出，世界各国正考虑如何减轻新兴人工智能技术的危险，这种危险或将重塑全球经济并改变国际安全格局。[②] 2025 年 1 月 15 日，联合国大会全体会议上，古特雷斯再次呼吁，要避免人工智能技术发展失控失衡，确保人人都有平等获得最新知识的权利，支持发展中国家利用人工智能实现可持续发展；应建立人工智能国际独立科学小组，并启动 2025 年的人工智能治理全球对话。[③]

（三）政治逻辑：“高政治”叙事强化安全政策环境

当今世界正处于百年未有之大变局，又恰逢新一轮科技革命爆发，技术与政治双重因素下，科技领域已成为大国竞争博弈的关键高地，技术议题不可避免地被“政治化”。人工智能技术被视为驱动第四次科技革命的核心技术与重

① 《美媒：科技行业领袖警告AI可能给人类带来“灭绝风险”》，中国新闻网，2023年6月1日，https://www.chinanews.com.cn/gj/2023/06-01/10017541.shtml。

② 《联合国秘书长呼吁强化人工智能全球治理》，新华网，2023年7月19日，https://www.news.cn/world/2023-07/19/c_1129757277.htm。

③ 《古特雷斯呼吁国际社会携手应对四大挑战》，新华网，2025年1月16日，https://www.news.cn/world/20250116/9abe2811dbfb4d2caff215ab0433784f/c.html。

要应用，世界主要国家均从维护国家安全、提升国家竞争力与重塑国际力量格局的定位，谋划人工智能发展战略及参与人工智能国际治理，这使得地缘政治因素高度内嵌于人工智能国际治理发展全过程，具体表现为基于国家安全乃至国际安全的“高政治”叙事、不断强化“安全偏好”的国际政策环境等。

一是中美人工智能领域战略较量频密。“当前，在中美战略竞争已成常态背景下，两国在科技领域的博弈愈发激烈。数字时代的新一轮科技革命导致权力内涵、结构和体系均产生重大变革，新兴数字技术已成为决定中美国际竞争力和国际地位的重要因素。”[①] 人工智能更被描绘成两国之间的竞争甚至是零和博弈；美国更是很早就明确必须保持先发甚至压倒性优势，并将中国作为“最大竞争对手甚至是国家安全挑战”。2019 年，时任美国总统特朗普发布第 13859 号行政令《保持美国在人工智能领域的领导地位》（Maintaining American Leadership in Artificial Intelligence），试图促进美国人工智能技术和创新并保持优势地位；同年，美国国会研究局（CRS）发布《人工智能和国家安全》（Artificial Intelligence and National Security）研究报告，认为中国“试图在 2030 年前夺取人工智能发展的全球领先地位”。[②] 近两年来，伴随新一轮人工智能热潮，美国政界、学界不断强化此类安全叙事，声称“中国将全面挑战美国在人工智能领域及全球范围内的相对优势和领导地位，并对美国国家安全构成严重威胁”。拜登执政后，对中国人工智能发展的遏压力度不断加强，持续收紧出口管制措施，干扰国际科研合作，联合盟友共同构筑人工智能领域的“小院高墙”，甚至试图打造所谓人工智能的“平行体系”。2025 年 1 月，拜登在离任前继续放出“大招”，发布人工智能管制新规，[③] 通过分级管理方式进一步收紧对中国的人工智能芯片和技术出口限制，目的是把先进算力

① 参见余南平、戢仕铭：《西方“技术联盟”组建的战略背景、目标与困境》，《现代国际关系》，2021 年第 1 期，第 48—49 页。

② 参见吴沈括、崔鑫铭：《人工智能与国家安全的美国视野——美国国会 2019 年“人工智能与国家安全研究报告”研究》，《中国信息安全》，2020 年第 1 期，第 100—103 页。

③ 《最后的疯狂？拜登离任前最强 AI 禁令引众怒》，凤凰网，2025 年 1 月 14 日，https://tech.ifeng.com/c/8g7WiWF47YZ。

留在美国及盟友境内，同时寻求更多办法限制中国获取先进人工智能芯片和技术。

特朗普二次执政后，对前任政府的新规如何实施虽还有待观察，但鉴于其与拜登在人工智能领域维护美国优势的看法相同，因此即便具体执行过程有所差别，也不会出现大幅转向。2025 年 1 月以来，美国对于中国 DeepSeek 的态度就是明显的风向标。1 月 28 号，新上任的白宫新闻秘书卡罗琳·莱维特（Karoline Leavitt）首次召开情况简报会就声称，特朗普觉得 DeepSeek 发布的人工智能模型给美国人工智能行业敲响了警钟；美国国家安全委员会正对“DeepSeek 可能造成的影响”进行调查，白宫将全力确保美国在人工智能领域的领导地位。[①] 据美国 Newsmax 新闻网等媒体报道，特朗普 2 月 7 日接受采访时表示，DeepSeek 不会对美国国家安全构成威胁，美国最终可从这家初创公司的人工智能创新中“受益”。但值得关注的是，此种表述只能表明特朗普暂未动用国家安全叙事，但并不意味其会改变维护“美国领先”施政的竞争性逻辑。

中美互动将给人工智能国际治理带来全局性影响。“美国和中国在人工智能领域的竞争可能会彻底改变力量平衡，并对全球治理产生重大影响，这可能会对数据治理、技术标准、道德伦理和地缘政治局势产生影响。”[②] 当二者的互动是以安全为主要叙事逻辑时，其影响必然外溢至国际治理进程。因为这种国际治理的优先诉求必然是更有利于维护自身竞争优势或国家安全的权力争夺，而非基于共促发展的合作意向。

二是中美之外相关国家的现实诉求强烈。广大发展中国家基于地缘政治考虑，更多从国家安全乃至国际安全角度对人工智能进行“高政治”叙事，进一步形成了“安全偏好”的国际治理政策环境，如广大发展中国家在联合国框架

① 《白宫：正评估DeepSeek对国家安全的影响》，凤凰网，2025年1月29日，https://tech.ifeng.com/c/8gWzQ6E6ret。

② See Asia Maqsood, Ahyousha Khan, and Muhammad Usama Siddiqi, “US-China Competition in Artificial Intelligence: Implications on Global Governance,” *Journal of Asian Development Studies*, Vol.12, December 30, 2023, pp.481-493.

下提出“智能鸿沟”问题。所谓“智能鸿沟”就是在智能化发展过程中，不同主体对智能技术掌握与应用能力上的差距，体现为资源占有不均、应用能力不同、发展机遇不等。此轮人工智能发展浪潮虽尚处发展初期，但催生的“智能鸿沟”已初现端倪。

目前，全球有实力研发和推广新型智能系统的企业主要集中在中美两国。尤其是以美国的微软、谷歌、Meta（原脸书）等科技巨头为代表，它们在数据、算力以及高端芯片方面的优势使其能技术先发、市场先占，这种“赢者统吃”“强者恒强”的格局很难被打破。发展中国家经过多年努力与发达国家间缩小的差距，很可能因此次技术“换道”而进一步拉大，国际社会将再次面对新的南北问题。因此，相关国家高度担忧可能出现的“数字殖民”挑战与威胁，部分国家更提出“主权 AI”概念。它具体包括：技术发展路线具有一定自主性，不完全依赖；重要基础设施有一定可控性，不受制于人；产品应用体现本国国情与文化，防范“数字殖民”与外来价值观侵蚀等。非人工智能强国对自身安全与共同安全的强烈危机感，也必然反映到人工智能的国际治理进程中。

2025 年 2 月 10 日—11 日，在法国巴黎举行的“人工智能行动峰会”（AI Action Summit）就集中折射出上述地缘政治博弈。一方面，此次会议覆盖面更加广泛，相较于前两次在英国、韩国举行的峰会，广大发展中国家的参与度明显提升，反映出强烈的参与国际进程、争取发展机会的现实诉求。但另一方面，美国副总统万斯讲话凸显“美国优先”的政策立场，声称“美国人工智能应成为‘金标准’”。考虑到此次会议上来自中国的 DeepSeek 成为各方热议焦点，所谓“金标准”的言论在外界看来，无疑是宣示美国在人工智能领域的主导权。与此同时，万斯以反对过度监管为由，对欧盟等国的人工智能监管政策进行抨击，凸显了美国与欧洲国家在人工智能治理问题上的分歧。会后，美国、英国拒签最后的声明文件，再次向外界传递其强硬的政策立场。此次巴黎会议凝聚共识、共同行动的初衷也因此被打折扣，给未来国际治理进程蒙上一层阴影。

基于不同利益诉求的竞争与博弈，既是对治理话语权的争夺，更是对人工

智能未来发展全球版图的争夺。鉴此，不受地缘政治影响的人工智能发展、基于国际合作的人工智能安全，在相当长一段时间内仍然只会是国际社会各方的美好愿景或努力方向，高度内嵌的地缘政治要素及其影响是人工智能发展与治理进程中必须直面的现实。

三、“安全偏好”对人工智能治理的趋势性影响

技术的发展与治理从来不是单一的技术问题，而是技术与政策双向互动与共同塑造的结果。当前，人工智能的治理充满着“安全偏好”，这种偏好的底层逻辑正是治理理念形成与实践推进的内驱性、规律性动力，影响并贯穿了治理的全过程。随着人工智能技术与应用潜力和变革性影响的释放，相关治理也呈现出一些明确的特性。

（一）安全关切的解决成为国际治理优先目标

从发展逻辑看，此轮人工智能技术与应用的井喷式发展在客观上会带来更加严峻的“科林格里奇困境”（Collingridge’s Dilemma）。一是信息困境，即一项技术的社会影响与发展后果无法在应用早期被准确预料，如果因为担心不良后果而过早实施控制，技术很可能难以爆发。二是控制困境，即出现问题甚至风险威胁时，技术应用已高度内嵌至社会体系结构，实施改变或有效控制将十分困难并成本高昂，甚至难以或不能改变。因此，能否提前预判、实时跟进以及敏捷应对发展过程中的安全关切，就显得格外重要。

从技术逻辑看，人工智能技术本身的难测性与不可控性需要国际协调方能进行合理治理。毕竟，人工智能技术与应用具有显著的边际递增特性。由于数据长期积累和算法自主迭代等技术逻辑，人工智能的发展速度和影响程度也可能会出现边际报酬递增特征。[①] 而技术的快速更迭与加速落地更加缩短了政策

① W. Brian Arthur, “Increasing Returns and the New World of Business,” *Harvard Business Review*, Vol.74, No.4, July 1996, pp.100-109.

反应与评估时间，若无法把握节奏，势必会在不久的将来带来更严重的控制困境。因此，在发展红利和不可控风险之间，安全成为压倒性因素及大多数人的选择。

从政治逻辑看，历史上每一次科技革命都导致了全球格局的重塑和大国兴衰的出现。人工智能的广泛应用和“绝妙前景”使得对自身优势的追求必然是相关国家头号优先事项。比如美国早已旗帜鲜明地亮出“必须赢得主导性优势”的立场，因此，寄希望于大国间协调来达到“安全的人工智能”并不现实。当前人工智能领域的“军备竞赛”已有加剧甚至失控风险，这客观上进一步加剧了安全形势的恶化。例如，美国提出所谓人工智能领域的类“曼哈顿计划”[①]——以5000亿美元高投入打开所谓“星际之门”[②]。在此情况下，很多事关人工智能安全的重要议题也只能纳入国际机制，对于安全目标的保障必然成为国际治理的首要关切。作为制度性沟通平台，国际机制能在更大范围内、以更高效率实现主权国家间的协商对话和矛盾消解，以国际组织的制度性防范技术快速迭代带来的治理问题发酵。[③]

（二）技术社群将在国际治理中发挥更重要作用

在新技术应用中，科技公司在内的技术社群既是研发与产业落地的推进者也是实践者，因此在治理中的作用更多地体现在发展而非安全。科技公司与技术社群长期奉行技术自由主义，常常会对抗地缘政治因素带来的发展制约，即便愿意更多参与国际规则制定，但客观上仍然被局限在技术层面。但是，人工智能的出现和飞速发展及其安全因素的突出显现使技术群体参与国际治理的合理性与必然性大大增加。

① “US Government Commission Pushes Manhattan Project-Style AI Initiative,” Reuters, November 19, 2024, https://www.usnews.com/news/top-news/articles/2024-11-19/us-government-commission-pushes-manhattan-project-style-ai-initiative.

② 《5000亿美元投资，特朗普宣布建设“星际之门”AI基础设施》，中国新闻网，2025年1月22日，https://www.chinanews.com.cn/gj/2025/01-22/10357870.shtml。

③ Robert O’Brien, *Contesting Global Governance:Multilateral Economic Institutionsand Global Social Movements*, Cambridge University Press, 2000, p.136.

人工智能技术作为一项前所未有的技术形态，最不可预知或不可控风险恰恰源于技术特性，无论是技术层面的应对，还是政策层面的防范，都需要更专业的技术社群参与，这为技术群体参与人工智能国际治理提供了天然合理性。同时，由于人工智能技术的“强垄断性”，无论是数据资源、算力、算法还是大模型都掌握在科技巨头手中，它们为争取更好发展空间与政策环境，也必然会参与到人工智能国际治理进程中以提升话语权与影响力，从而更好地将技术与产业优势转化为治理实效。这也为技术群体参与人工智能国际治理提供了必然性。事实上，技术群体业已深谙各方对人工智能技术的安全关切，从其应用产品与服务重在强调“安全性”就可见一斑。鉴此，未来包括科技公司在内的技术社群会更主动地将其掌握的资源转化为更大话语权与主导权。

（三）围绕治理平台的竞争将更趋激烈

当前人工智能国际治理尚处初期，机制建设仍不完善，远未形成体系化、机制化且具有广泛代表性与认可度的治理平台。一方面，既有治理机制难以直接将人工智能议题全面纳入。例如，网络空间国际治理的相应机制主要基于互联网技术与应用，而人工智能技术与应用与之有着较大差异性，从专业性角度难以匹配。另一方面，现有治理机制较为松散，相关机制主要呈现出以科技大国或区域组织为单一中心、其他行为体为四周的“伞状结构”。[①]

中美两个大国在政治逻辑驱动下围绕人工智能竞争与博弈激烈，主要体现在国际治理机制的设想上。中国坚持一贯立场，高度重视联合国框架下国际治理的作用，明确支持联合国在人工智能全球治理领域发挥应有作用，认为在联合国框架下协调国际人工智能的发展、安全与治理等重大问题是必由之路。[②]美国则无论是出于对国际公共事务意愿下降，还是出于对中国人工智能发展建盟围堵的考量，更倾向于建立一种不仅自己能发挥主导作用，还可将中国影响

① 参见梅立润：《技术置换权力：人工智能时代的国家治理权力结构变化》，《武汉大学学报（哲学社会科学版）》，2023年第1期，第44—54页。

② 《关于全球治理变革和建设的中国方案》，外交部，2023年9月13日，https://www.mfa.gov.cn/ziliao_674904/zcwj_674915/202309/t20230913_11142009.shtml。

力剔除在外的国际治理机制。美国关于人工智能竞赛的高敏感性叙事映射到现实中，直接导致“国家间信任不足，从而引发‘超规格’的治理手段和‘高警惕’的安全互动，进一步阻碍全球范围内形成成熟有效的人工智能治理模式”。[①] 事实上，美国一方面阻碍全球性治理机制的形成，另一方面集中力量打造由己主导的区域性治理范本，企图通过示范或扩散效应使其升级为事实性的国际治理机制。例如，2024 年发布的《首尔宣言》[②] 就旨在七国集团内部率先建起人工智能全球治理框架。

广大发展中国家关于缩小“智能鸿沟”、维护“主权 AI”的诉求也势必需要一个更具全球性、能解决智能时代新南北问题的平台。同样，包括科技巨头、技术社群在内的非国家行为体也需要一个更能提升参与度与话语权的平台。但是，这个平台不能仅仅是技术产业导向的，更需要能够与国家主体进行有效对话与协调。无论国家主体立场如何，人工智能的发展正在极大改变主权国家同非国家行为体之间的权力分布和互动模式。[③] 因此，人工智能国际治理机制的完善需要打造更加合适的平台，而围绕平台的竞争也将更加激烈且复杂。这不仅涉及国家间博弈，还涉及国家主体与非国家主体间的协调。

结语

中国作为人工智能大国，以负责任大国姿态积极倡导人工智能治理进程，在国际上提出了以《全球人工智能治理倡议》《人工智能全球治理上海宣言》[④] 为代表的系列主张，为推动人工智能技术造福全人类作出前瞻探索和中国贡献。

① 鲁传颖、张璐瑶：《人工智能的安全风险及治理模式探索》，《国家安全研究》，2022 年第 4 期，第 91 页。

② *Seoul Declaration for Safe, Innovative and Inclusive AI: AI Seoul Summit 2024*, Department for Science, Innovation and Technology, May 2024.

③ See Juho Lindman, Jukka Makinen, and Eero Kasanen, “Big Tech’s Power, Political Corporate Social Responsibility and Regulation,” *Journal of Information Technology*, Vol.38, No.2, June 2023, pp.144-159.

④ 《人工智能全球治理上海宣言（全文）》，中国政府网，2024 年 7 月 4 日，https://www.gov.cn/yaowen/liebiao/202407/content_6961358.htm。

2025 年 2 月 11 日，为期两天的人工智能行动峰会在法国巴黎落幕，包括法国、中国、印度、欧盟在内的多个国家和国际组织共同签署了《关于发展包容、可持续的人工智能造福人类与地球的声明》（Statement on Inclusive and Sustainable Artificial Intelligence for People and the Planet），提出需就人工智能治理进行包容性的多利益相关方对话与合作，进一步加深信任和加强安全合作。[①] 国际社会的持续努力已为未来人工智能国际治理奠定一定基础。但现实表明，挑战仍然是巨大的，推进治理进程、实现治理目标，各方还需付出更多行动。未来较为现实的阶段性推进重点包括以下几方面。

一是更好回应全球共同安全关切。从国际议程设置角度，考虑到人工智能技术与应用发展初期的“安全偏好”，无论是人工智能大国还是发展中国家，最好的合作基础是回应与应对共同的安全关切。这些共同安全包括但不限于：与国际社会各方开展合作，共同探索有效应对技术应用安全风险的最佳实践；打破美国主导下的所谓“中美人工智能竞赛”的狭隘国家安全叙事，消解国际社会各方的冲突担忧；采取有效方式，切实关照广大发展中国家因为“智能鸿沟”面临的安全困境，积极推进相关能力建设，争取更广泛的合作力量，营造有利的国际政策环境。

二是积极调动科技公司与技术社群力量。鉴于技术与产业界在人工智能发展中的话语权与影响力日趋重要，国际社会应进一步调动其积极性，并引导它们有效平衡市场利益与安全关切，既要避免技术自由主义，又要防止过度政治化与安全化对其作用的反向抑制。中国企业与科研机构在 2024 年 6 月推出的“智能向善全球伙伴计划”就是很好的尝试。作为一项惠及全球的公共产品，它旨在建立一个全球性的合作网络，汇聚包括南方国家在内的各方力量，共同推动人工智能技术的健康发展，以确保技术进步能够造福全人类，实现可持续

① “Statement on Inclusive and Sustainable Artificial Intelligence for People and the Planet,” February 11, 2025, https://www.elysee.fr/en/emmanuel-macron/2025/02/11/statement-on-inclusive-and-sustainable-artificial-intelligence-for-people-and-the-planet.

发展目标。[①]

三是多元推动人工智能国际治理平台建设。当前地缘政治竞争态势下，国家间合作意愿与投入力度下降，但人工智能国际治理是大势所趋。为长远计，应该加大对机制建设的投入力度，搭建多梯度平台。一方面，国际层面应切实支持联合国框架的主渠道作用。联合国框架在全球性议题方面发挥的协调作用与影响力仍不可替代，尤其是在统合各方利益诉求方面，包括为全球南方国家提供发声渠道与沟通机制方面均发挥着重要作用。[②] 另一方面，鼓励区域层面搭建新的治理平台。全球人工智能产供链正在形成，越来越多的地区国家加入其中，包括中东、非洲、东南亚等地区都在积极布局。随着未来技术与应用的进一步推进，这些国家或地区在人工智能方面发展潜力巨大，同时也意味着关于人工智能治理的诉求亦会随之增加。这些区域性平台的搭建会成为人工智能国际治理生态的重要构成。

总而言之，人工智能是一片广阔但充满不确定性的“蓝海”。一方面，此轮人工智能浪潮已然在一定程度打通应用环节，技术落地的同时，新型安全风险也得以显现，国际治理尤其是安全治理的现实需求不断增加。另一方面，此轮人工智能浪潮恰逢百年未有之大变局，特殊历史背景下的“高政治”叙事不断影响国际政策环境，原本应该基于国际合作的人工智能国际治理充满巨大现实挑战。鉴此，人工智能国际治理过程中的理念认知、行动协调、政策制定以及价值选择，将不断塑造人工智能的未来。

（责任编辑：石刚）

① 《人工智能重大应用场景白皮书研究暨“智能向善”全球伙伴计划启动会成功召开》，中国公共管理案例中心，2024年6月28日，http://case.sppm.tsinghua.edu.cn/info/1002/2125.htm。

② 参见周桂银：《全球南方崛起与全球治理体系变革：以国际规则和制度为例》，《国际观察》，2024年第2期，第97—129页。