

人工智能全球治理的多边困境与中美战略稳定^{*}

栾心蔚 余南平

[内容摘要] 人工智能全球治理的困难,并不只是新兴技术发展过快而规则供给滞后,更是全球公共风险叙事与国家技术能力竞争之间的深层张力。人工智能正在成为国家能力、产业体系和国家安全判断的一部分,这使传统意义上多边治理中本已存在的主权边界、相对收益和规则解释权问题更加突出。世界各国虽能在安全、可信、负责任等原则问题上达成共识,却难以接受可能改变自身战略竞争位置的强约束规则。而芯片、算力、云平台、基础模型和技术开发生态的高度集中,又使全球规则形成带有明显的技术依赖和话语权非对称。由此,在理想与现实的悖论张力扩大中,全球人工智能治理难以直接走向理想中的统一制度,更现实的基础与出发点在于维持中美两大技术能力主体之间最低限度的建设性战略稳定,以防止人工智能全球治理被完全吸纳进阵营化技术竞争之中。

[关键词] 人工智能全球治理 多边困境 新兴领域国家安全 中美建设性战略稳定

[作者简介] 栾心蔚,华东师范大学政治与国际关系学院博士生;余南平,华东师范大学政治与国际关系学院教授

[中图分类号] D822 **[文献标识码]** A **[文章编号]** 2095-5715(2026)05-0109-22

过去几十年里,多边治理曾经承载了关于全球化的一种基本想象:当问题超出单一国家边界时,各国可以通过协商形成规则,通过制度安排降低不确定性,

^{*} 本文系上海市哲学社会科学规划重大研究专项课题“中国国际政治经济学自主知识体系构建研究”(项目编号:2025DZX028)的阶段性成果,并得到上海市未来与前沿学科(新兴领域国家安全)项目建设的支持。

通过共同承诺把冲突转化为可管理的分歧。然而,这一想象从来没有真正摆脱现实政治的制约,无论是贸易、金融、环境,还是后来的数字治理,多边治理都反复遭遇着同一困境:治理议题一旦进入国家利益、发展空间和安全边界深处,便会暴露收益不对称、能力不均衡和风险承受差异的深层矛盾,抽象共识也因此难以转化为具体约束,多边治理最终在“原则泛化”与“规则虚化”的内在背离中呈现出一种有共识而约束不足的治理状态。

人工智能的出现进一步加剧了这一困境。随着人工智能嵌入产业体系、社会运行和国家安全判断,以往多边治理中的相对收益、主权边界和规则解释权问题,在人工智能领域被重新压缩进一个更不稳定、更不均衡也更具安全敏感性的结构之中。在高度不对称与不确定的技术体系中,人工智能治理很容易转化为涉及技术能力分配、安全边界重塑与未来竞争格局预设的复合型议题。其核心张力表现为任何试图确立“实质约束”的规则努力,无一例外都会被各方解读为对自身未来相对地位的挤压、对发展空间的锁定,以及对安全边界的再次切割。

当技术仍未定型、能力分布又不对称、安全判断不断向国家内部运行结构回收时,多边治理能在何种基础上维持有效性?当共识难以变成约束、合作难以摆脱能力结构和安全疑虑牵引,人工智能全球治理的实现究竟依赖什么?中美作为人工智能发展中的技术能力主体,其互动方式如何影响多边治理的制度空间与规则边界?在竞争长期存在的条件下,人工智能全球治理如何避免被安全化与阵营化逻辑持续压缩?这些问题,不是治理热潮中的附带议题,而是理解人工智能全球治理现实限度、制度可能与未来走向时必须面对的基本问题。

一、全球治理的多边困境现实

冷战结束后的自由主义胜利叙事曾经强化了这样的理解:多边贸易规则体现了对监管性国家权力边界的自由主义设定。但这一理解遮蔽了多边共识机制的谈判、设定和运行始终发生在特定国际权力结构之中这一事实。全球治理始终伴随着不同国家围绕何种规则可以被普遍化、何种利益可以被制度化、何种国

家—市场关系可以被解释为“正常”而展开的持续竞争。^① 多边主义在形式上能够提供一种组织原则,但作为一种带有特定规则和行为原则的制度形式,它无法自动消除这些差异。^② 原则共识在进入多边治理的具体规则和制度约束转化阶段时,往往遭遇利益协调、制度嵌入、权力结构等阻碍,由此造成约束不足的治理状态,这正是本文所讨论的多边困境。

这种困境在传统贸易治理中已经显露。货物贸易曾经是多边治理较为成功的领域,因其相对容易被规则捕捉:商品可以分类、关税可以计算、原产地可以判定、市场准入可以谈判。可当贸易议题逐渐越过边境,涉及到产业政策、国内监管和发展权安排时,原有多边协商所面对的利益协调压力也随之上升。世界贸易组织的共识决策始终受到大国权力、成员结构和议题分布的影响,^③多哈回合谈判的长期停滞正说明,随着规则收益、调整成本和发展空间的分配变得更加复杂,多边贸易治理已经难以延续原有的利益协调方式和节奏。

服务贸易进一步加大了这种困难。金融、审计、通信、法律、教育等服务领域均与监管体系、职业资格、行业标准和主权安排相关联。一个国家是否接受外部会计规则、开放关键服务行业或认可他国监管标准,牵涉到的是主权制度边界。服务贸易中的让步往往意味着监管权、准入权和标准解释权的再分配。多边治理因此面临更加突出的制度嵌入难题。

数字时代曾给全球治理带来新的想象。网络连接似乎可以穿越地理边界,平台技术似乎可以降低交易成本,数据流动似乎可以形成更密集的全球合作。然而,随着数字基础设施日益嵌入经济社会运行,数字空间并没有让国家退场,相反,国家仍然通过规则制定和基础设施控制塑造数字秩序,国家特别是大国仍

① Anne Orford, “How to Think About the Battle for the State at the WTO,” *German Law Journal*, Vol. 24, No. 1, 2023, pp. 45 ~ 71.

② John Gerard Ruggie, “Multilateralism: The Anatomy of an Institution,” *International Organization*, Vol. 46, No. 3, 1992, pp. 561 ~ 598.

③ Richard H. Steinberg, “In the Shadow of Law or Power? Consensus-Based Bargaining and Outcomes in the GATT/WTO,” *International Organization*, Vol. 56, No. 2, 2002, pp. 339 ~ 374; Morten Boas et al., *Multilateral Institutions: A Critical Introduction*, London: Pluto Press, 2003, p. 44.

然是网络治理结构和监管标准形成中的重要行为体。^① 这意味着,数字治理既关乎连接,也关乎控制;既关乎开放,也关乎边界;既关乎市场效率,也关乎国家安全和秩序。这些张力的形成与协调,始终受到数字基础设施、数据资源和平台生态分布的影响,进而不可避免地嵌入既有权力结构之中。

从货物贸易时期的利益协调,到服务贸易时期的制度嵌入,再到数字治理时期更加凸显的权力结构问题,全球治理的多边困境始终包含着对市场运行方式、国家权力边界和规则解释权的安排。也正是在这一意义上,人工智能把多边困境推向了更加敏感的境地。人工智能不仅嵌入产业体系与国家安全边界,更逐渐渗透到知识生产、决策机制与社会组织方式中,其所触及的不只是技术应用本身,更是国家能力结构与权力运行方式的深层基础。^② 伴随人工智能技术属性的政治化,利益协调、制度嵌入、权力结构等障碍更加集中地作用于同一治理对象。由此,人工智能全球治理不只是既有多边机制面对新技术议题的适应问题,更是多边治理在技术能力、国家安全与规则主导权高度交织的背景下面临的结构性考验。^③

二、人工智能加剧多边困境

人工智能的通用性、嵌入性和能力依赖性使其不仅停留于技术本身,还成为国家竞争力、社会组织方式和新兴领域国家安全判断的一部分。^④ 人工智能越具有通用性与嵌入性,越难被划归为单一行业或特定风险框架来应对;而国家对其能力依赖越深,越会在治理中触及最不愿让渡的主权与安全底线。世界各国讨

① Daniel W. Drezner, "The Global Governance of the Internet: Bringing the State Back In," *Political Science Quarterly*, Vol. 119, No. 3, 2004, pp. 477 ~ 498; 高奇琦:《国家参与全球治理的理论与指数化》,《社会科学》2015 年第 1 期,第 5 ~ 7 页;戚凯、周祉含:《全球数字治理:发展、困境与中国角色》,《国际问题研究》2022 年第 6 期,第 47 ~ 64 页。

② Michael C. Horowitz, "Artificial Intelligence, International Competition, and the Balance of Power," *Texas National Security Review*, Vol. 1, No. 3, 2018, pp. 36 ~ 57.

③ Jonas Tallberg et al., "The Global Governance of Artificial Intelligence: Next Steps for Empirical and Normative Research," *International Studies Review*, Vol. 25, No. 3, 2023.

④ 余南平:《通用人工智能时代的国际权力重塑》,《中国社会科学》2025 年第 4 期,第 41 ~ 59 页。

论人工智能治理时,表面上是在回应共同风险,实质上却同时面对数据能否流动、能力能否扩散、规则由谁解释、约束由谁承担等更深层问题。

正因如此,人工智能全球治理中的共同风险难以顺畅转化为统一的制度约束,而是经历了一场层层递进的“位移”:治理对象安全化推动国家安全边界内移,压缩了跨国制度嵌入的空间;安全边界内移又加剧合作基础分化,使利益协调受制于能力竞争与约束强度分歧;当共识受阻,技术资源的不均衡便会使规则制定沦为权力结构的外化与延伸,并致使规则生成呈现出非对称性。在安全、能力与规则张力被推入现实约束中,大而全的理想化治理架构难以为继,承认治理限度并转向有限治理,成为超越多边困境、维持全球治理效能的必然选择。

(一) 安全边界的内移:人工智能与国家安全再定义

人工智能之所以会放大多边治理困境,首先在于它不断重构安全边界形态,使得原本清晰、集中于特定领域的边界呈现出跨层级、跨领域的交织特征。传统技术治理中的安全风险尚可被放置在产品、行业或应用场景之中,但人工智能并不安于这种边界。它以数据为燃料、算力为支撑、模型为组织形态,通过平台、接口和智能体进入产业链、网络空间、公共服务、情报系统和战略决策,并逐渐成为国家能力的一部分。人工智能安全治理由此超出“防止技术出错”的简单范畴,而是涉及数据、系统、行动和决策等核心安全边界能否被共同规则约束。

这种安全边界的内移,首先体现在数据维度。数据是观察人工智能安全边界重构的重要窗口,因为它直接牵涉国家对关键运行过程的持续控制与干预。底层数据是一个国家金融体系、产业结构、人口行为、社会治理和公共服务的映射。数据的流向在很大程度上塑造了模型的理解、预测和优化能力。而它一旦进入外部模型,还可能被建模、被预测、被转化为外部系统理解和影响本国经济社会运行的能力。^① 目前,美国国家安全局为贯穿采集、标注、存储、传输与调用

^① 国家级数据基础设施被置于人工智能治理与国家安全的首要交汇处,数据隐私、网络防御、信息安全和信息战风险从根本上不能分割处理。参见 Margaret Hu et al., “National Security and Federalizing Data Privacy Infrastructure for AI Governance,” *Fordham Law Review*, Vol. 92, No. 5, 2024, p. 1829。

各环节的数据供应链搭建了系统性安全防护框架,以落实严格的安全机制。^① 这不仅是出于对数据本身泄露风险的担忧,也因为数据一旦进入模型训练与优化过程,其所承载的结构性信息、行为模式与系统逻辑便会被转化为模型能力的一部分,并在模型参数与推理机制中长期固化,即使事后删除或回收原始数据,也无法将其从模型能力中彻底剥离。这种难以逆转的数据吸收过程,要求安全防护覆盖数据进入模型前后的各个环节。在此意义上,数据跨境流动不再只是信息转移,而是可能改变国家对本国金融、产业、社会治理和公共服务系统的可控性。对数据链条的强化保护,实质上是在防止关键数据通过外部模型和平台进入难以回收、难以追踪、难以控制的安全状态。

安全边界的迁移并不止于数据层面。模型的智能体化进一步改变了新兴领域安全风险形态。^② 当以大语言模型为核心的人工智能应用逐步向智能体形态扩展时,模型越来越多地连接代码环境、数据库、浏览器、办公系统和外部工具,开始从“生成回答”转向“执行任务”,安全问题因此从“说什么”转向“做什么”。^③ 此时,安全边界会沿着权限、工具和组织流程不断向金融、政务等关键领域延伸。这类风险在情报和战略决策领域更为突出。大规模公开数据经过人工智能处理后,也可能揭示原本并未直接呈现的私人信息、群体关系和行为模式。^④

① National Security Agency, “AI Data Security: Best Practices for Securing Data Used to Train and Operate AI Systems,” May 2025, <https://www.fbi.gov/file-repository/cyber-alerts/ai-data-security-best-practices-for-securing-data-used-to-train-and-operate-ai-systems-052225.pdf>.

② 本文所称“大语言模型”,是指通过大规模文本数据训练,能够理解和生成自然语言内容的机器学习模型;“智能体”则是指以大语言模型等人工智能模型为推理和决策基础,能够感知外部环境、规划任务、调用工具,并在一定程度上自主采取行动以实现特定目标的人工智能系统。二者并非同一层次的概念:前者属于基础模型,后者是整合模型、记忆、工具接口和行动执行的系统形态。本文所说的“智能体化”,特指以大语言模型为核心的人工智能应用由内容生成向任务规划、工具调用和行动执行扩展的过程。参见 Anna Gutowska, “What are AI Agents?” IBM, <https://www.ibm.com/think/topics/ai-agents>。

③ 谷歌威胁情报团队披露,政府支持的攻击者和信息操纵行为体已经使用生成式人工智能进行侦察、漏洞研究、代码辅助、内容翻译和社交工程;OpenAI 也披露,自 2024 年 2 月开始公开威胁报告以来,已处置并披露 40 多个违反使用政策的网络,涉及诈骗、恶意网络活动和隐蔽影响行动。参见 Google Threat Intelligence Group, “Adversarial Misuse of Generative AI,” January 2025, <https://services.google.com/fh/files/misc/adversarial-misuse-generative-ai.pdf>; OpenAI, “Disrupting Malicious Uses of AI: October 2025,” October 2025, <https://openai.com/zh-Hans-CN/global-affairs/disrupting-malicious-uses-of-ai-october-2025/>。

④ Ross Bellaby, “The Ethical Problems of ‘Intelligence-AI’,” *International Affairs*, Vol. 100, No. 6, 2024, pp. 2525 ~ 2542.

当这种能力与国家强制力、情报秘密性和安全机构的行动需求结合时,数据收集、模型分析和政策行动之间的边界会更加模糊,并进一步转化为战略风险:模型失误、数据污染、提示注入或外部攻击有可能从技术故障转化为公共安全事件,人工智能目标识别、态势感知、危机预警等有可能改变国家理解威胁、配置资源和使用力量的方式。^①

安全边界的内移带来了深刻的治理后果。人工智能所引发的安全顾虑,使世界各国高度警惕国家安全边界被重新划定、安全控制权归属被重新调整的风险。越是有效的国际规则,越可能涉及数据流动、模型评估和关键应用限制等多个方面。这些安排越具体,越会进入国家发展能力、安全能力和治理能力的核心区域。在当前人工智能多边治理议程中,欧盟通过《人工智能法案》《数据治理法》对数据使用与共享提出制度性约束,在经合组织与七国集团框架下也持续推动可信数据流动与数据治理原则的协调。这些实践表面上推动了合作与互联互通,深层上必然会触及国家对基础性信息资源和智能能力生成过程的控制。正是这种安全边界的向内推移,使人工智能多边治理从一开始就具有强烈的主权敏感性:它讨论的并非外部风险的共同防范,而是数据、系统、行动和决策等国家核心安全边界能否被共同规则触及。而当各国越来越将防范他国人工智能技术影响上升为国家安全问题时,全球人工智能治理也就越容易走向碎片化。^②

(二) 风险共识的限度:能力竞争与约束分歧

国际社会对人工智能风险治理并非没有共同语言。近年来,围绕安全、可信、负责任、以人为本、包容性发展等原则,全球人工智能治理已经形成若干框架性共识。从2023年英国首届人工智能安全峰会到2026年印度人工智能影响力峰会,全球人工智能峰会已经连续举办四届,相关议程也从前沿人工智能安全评

^① Toni Erskine and Steven E. Miller, "AI and the Decision to Go to War: Future Risks and Opportunities," *Australian Journal of International Affairs*, Vol. 78, No. 2, 2024, pp. 135 ~ 147.

^② Filippo Lancieri et al., "AI Regulation: Competition, Arbitrage and Regulatory Capture," *Theoretical Inquiries in Law*, Vol. 26, No. 1, 2025, pp. 239 ~ 262.

估扩展到创新发展、包容治理、能力建设和全球南方关切等多支柱议题。^①与此同时,人工智能全球治理也开始从阶段性峰会宣示向多边对话和组织化平台延伸。2026 年 7 月,联合国授权设立的“人工智能治理问题全球对话”在日内瓦首次举行,为各国政府和相关利益攸关方提供了开放、透明、包容的多边对话平台;^②随后,《关于成立世界人工智能合作组织的协定》在上海签署,世界人工智能合作组织作为独立政府间国际组织成立,并在首次扩容后增至 37 个成员国。^③二者分别体现了联合国框架内的开放对话安排和新设政府间国际组织的尝试,说明人工智能全球治理正在获得一定的议程承载和组织支撑。

这些进展表明,世界各国并非不能在风险认知层面达成共识,人工智能全球治理也正获得更多的制度承载。然而,治理认知的共识并不等于规则约束,新的机制供给也不会自动生成治理能力。会议参与、文件签署与核心规则塑造之间仍存在实质性差异,相关机制虽然能够容纳更多行为体参与议题讨论,却未必意味着各方能够平等进入规则形成和执行安排的关键环节。例如,在首尔峰会中,中国未获邀参加仅对特定国家开放的首脑会议,因而被排除在政府间核心文件《首尔宣言》的签署范围之外,显示出参与峰会与进入核心文件塑造之间仍存在差距。巴黎峰会则更集中地展示了这一困境。该峰会形成《关于发展包容、可持续的人工智能造福人类与地球的声明》。从文本来看,该声明延续开放、包容、可持续发展和全球协调等框架性原则,不以强制性规则为目标。然而美国以过度监管会抑制创新为由拒绝签署宣言,英国以该声明缺乏实际指导而效力有限为

① Prime Minister's Office et al., "The Bletchley Declaration by Countries Attending the AI Safety Summit, 1-2 November 2023," November 2, 2023, <https://www.gov.uk/government/publications/ai-safety-summit-2023-the-bletchley-declaration/the-bletchley-declaration-by-countries-attending-the-ai-safety-summit-1-2-november-2023>; Department for Science, Innovation and Technology, "Seoul Declaration for Safe, Innovative and Inclusive AI: AI Seoul Summit 2024," May 21, 2024, <https://www.gov.uk/government/publications/seoul-declaration-for-safe-innovative-and-inclusive-ai-ai-seoul-summit-2024>; Permanent Mission of France to the United Nations in New York, "Statement on Inclusive and Sustainable Artificial Intelligence for People and the Planet," February 11, 2025, <https://onu.delegfrance.org/statement-on-inclusive-and-sustainable-artificial-intelligence-for-people-and>.

② UNESCO, "UN Global Dialogue Opens with Urgent Call for Safe and Inclusive AI that Benefits All," July 6, 2026, <https://www.unesco.org/en/articles/un-global-dialogue-opens-urgent-call-safe-and-inclusive-ai-benefits-all>.

③ 《〈世界人工智能合作组织〉扩容,第二批 8 国签约加入,成员增至 37 国》,中国国际科技交流中心,2026 年 8 月 5 日, https://www.ciste.org.cn/gjjsmy/dwjl/art/2026/art_6eebe955ef6840a78367b9e91bb9b76f.html。

由同样拒绝签署。在这样相对宽泛的原则文件上,主要技术领先国家仍未能形成完全一致的制度表达,严重削弱了共同声明作为制度协调基础的代表性。^①

因此,此类机制能否从议题聚合走向有效治理,关键不只在于数量是否增加,还要取决于主要技术领先主体是否愿意接受并参与塑造具有约束意义的规则安排。而在人工智能领域,技术能力本身又深度嵌入国际竞争格局之中,世界各国往往将其视为提升国家竞争力和权力地位的关键战略能力。这一逻辑在美国国家人工智能政策叙事中表现得尤为明显。美国已将人工智能置于国家竞争和战略优势的核心位置。《美国人工智能行动计划》以“赢得人工智能竞赛”为目标,将加速创新、建设人工智能基础设施和引领国际人工智能外交与安全作为主要支柱。^② 所谓人工智能“军备竞赛”的表述正转变为一系列具体举措,例如美国与欧盟联手限制中国获取高端 GPU 及其他先进人工智能芯片,又如美、欧、中三方展开国际竞争,竞相向那些愿意将此类高端芯片生产转移至其各自管辖区的企业提供补贴等。^③ 人工智能“军备竞赛”叙事也正是在这种战略思维和语境中获得动员能力:它既可以为扩大投入、加速部署和放松监管提供理由,也会使其他国家更难相信安全治理完全是出于公共风险考量。当先进芯片、算力基础设施和前沿模型被纳入出口管制与联盟协调框架之中,技术脱钩便可能推动全球技术基础设施走向碎片化,并使不同标准、供应链和技术生态之间形成相互竞争的平行体系。^④ 由此,人工智能治理便不再只是风险合作议题,而成为技术竞争和安全竞争的延伸。

① Dan Milmo, “US and UK Refuse to Sign Paris Summit Declaration on ‘Inclusive’ AI,” *The Guardian*, February 11, 2025, <https://www.theguardian.com/technology/2025/feb/11/us-uk-paris-ai-summit-artificial-intelligence-declaration>; Zoe Kleinman and Liv McMahon, “UK and US Refuse to Sign International AI Declaration,” *BBC*, February 11, 2025, <https://www.bbc.com/news/articles/c8edn0n58gwo>.

② The White House, “America’s AI Action Plan,” July 2025, <https://www.whitehouse.gov/wp-content/uploads/2025/07/Americas-AI-Action-Plan.pdf>.

③ Chris Miller, *Chip War: The Fight for the World’s Most Critical Technology*, New York: Scribner, 2022; Foo Yun Chee, “EU Takes on United States, Asia with Chip Subsidy Plan,” *Reuters*, April 18, 2023, <https://www.reuters.com/technology/eu-agrees-chips-subsidies-plan-eu-industry-chief-says-2023-04-18/>.

④ Björn Fägersten and Tim Rühlig, “Infrastructure Development and Geoeconomic Competition: A Framework for Analysis,” in H. Borchert and J. Strobl, eds., *Storms Ahead: The Future Geoeconomic World Order*, Vienna: Raiffeisen Bank International, 2021, pp. 156 ~ 171.

即便是以拥有规则和规范权力著称的行为体,其人工智能治理也并非单纯追求抽象公共利益。在规则扩散受制于技术生态、制度协调又受到监管主权利限制的条件下,人工智能治理不太可能实现全球协调。^①因为在竞争环境下,规则本身就可能成为能力竞争的一部分,虽然模型失控、网络滥用、虚假信息和关键基础设施脆弱性构成共同风险,但由人工智能带来的产业升级、战略优势、军事能力提升、信息空间塑造以及规则影响力扩展,会在国家之间形成差异化分布,人工智能安全治理很容易被理解为“谁先受约束,谁先失去优势”。^②共同风险要求各国克制,相对收益则推动各国加速。对此,经典国际关系理论早已指出,在无政府结构中,国家并不只关心合作能否带来绝对收益,也会关心合作收益如何在各方之间相对分配。^③人工智能治理正是这一逻辑在新兴技术领域中的集中体现。因此,任何治理安排都不可避免地会被纳入战略竞争考量之中:一方面,各方承认人工智能可能带来的系统性风险;另一方面,又担心过强的约束会削弱自身在技术、产业和规则制定中的优势地位。由此,人工智能治理的共识困境并不主要表现为“是否承认风险”,而是表现为“是否愿意接受可能改变自身竞争位置的强约束规则”。

因此,在技术主权优先于全球共治的背景下,人工智能全球治理在相当长时期内可能呈现一种向“最低公约数”妥协的有限治理状态。所谓有限治理,是指统一、强制且普遍适用的全球制度难以形成,各国只能围绕少数共同风险,通过原则性共识、局部合作和软性约束维持基本协调的一种全球治理状态。其有限性具体表现为治理范围有所收缩、参与程度存在差异、规则约束强度较低,执行机制也缺乏统一性。在这种状态下,原则性文件、伦理宣言、自愿承诺、技术评估合作和安全研究网络仍会继续发展,相应的国际峰会、国际论坛还会召开,因为

① Marco Almada and Anca Radu, “The Brussels Side-Effect: How the AI Act can Reduce the Global Reach of EU Policy,” *German Law Journal*, Vol. 25, No. 5, 2024, pp. 646 ~ 663; Daniel Mügge, “EU Sovereignty: For Whom, to What End, and to Whose Benefit?” *Journal of European Public Policy*, Vol. 31, No. 8, 2024, pp. 2200 ~ 2225.

② 有关人工智能塑造国际权力的论述,参见余南平:《通用人工智能时代的国际权力重塑》,《中国社会科学》2025 年第 4 期,第 41 ~ 59 页。

③ Joseph M. Grieco, “Anarchy and the Limits of Cooperation: A Realist Critique of the Newest Liberal Institutionalism,” *International Organization*, Vol. 42, No. 3, 1988, pp. 485 ~ 507.

它们有助于维持最低限度的沟通和风险认知。但应该看到,任何具有实质性约束力的人工智能治理规则,都会在落地过程中遭遇明显阻力:规则文本趋于原则化,执行机制趋于软化,关键议题可以被有意留白,高敏感领域则更容易转向小范围协调、阵营化安排或单边政策工具,最终形成有共识而弱约束、有合作而难统合的有限治理格局。人工智能多边治理由此只能在有限合作、规则竞争与技术分化之间艰难推进。

(三) 规则形成的非对称性:技术依赖与话语权失衡

规则能否被提出、被接受并被制度化,不仅取决于各方对风险的共同认知,事实上更取决于规则背后依托何种技术体系、产业结构和权力关系。人工智能规则并不是在抽象的多边会议空间中自然生成的,而是在芯片、算力、云平台、基础模型、开发生态和安全联盟所构成的技术结构中被塑造的。谁掌握关键技术体系,谁就更容易把自身关于风险、安全和责任的理解嵌入国际规则之中;谁处于技术依赖位置,谁就更容易在规则形成中成为跟随者、接受者甚至被规训者。人工智能多边治理的困境,正是在这种话语权非对称性中逐渐显现。

首先,非对称性的现实来自全球人工智能发展能力的高度不均衡。技术依赖本身会预先塑造各方参与治理的能力、位置和选择空间,技术能力有限的国家往往依赖外国控制的基础设施,从而更容易受到地缘政治变化和外部影响。^① 目前全球人工智能发展水平极其不均衡,斯坦福大学《2026 年人工智能指数报告》显示,2025 年美国机构产出了 59 个重要人工智能模型,中国为 35 个,全球前沿模型生产仍主要集中在美国和中国;该报告同时指出,美国拥有 5427 个数据中心,数量是其他任何国家的十倍以上,并且几乎所有领先人工智能芯片都由台积电制造,使全球人工智能硬件供应链高度依赖单一关键节点。^② 极度的不平衡与非对称性说明,人工智能技术能力事实上已经高度集中于少数国家、企业和基础

^① Jieli Li, "Governing High-Risk Technologies in a Fragmented World: Geopolitical Tensions, Regulatory Gaps, and Institutional Barriers to Global Cooperation," *Fudan Journal of the Humanities and Social Sciences*, Vol. 19, 2026, pp. 113 ~ 137.

^② Stanford Institute for Human-Centered Artificial Intelligence, "Artificial Intelligence Index Report 2026," 2026, <https://hai.stanford.edu/ai-index/2026-ai-index-report>.

设施节点之中。除中美这样具备完整的技术产业链和大规模基础设施能力的国家外,全球多数国家并不具备全栈式人工智能自主能力。由此,即便在政治上主张治理权力的自主,在技术发展上也不得不接入少数领先国家和大型企业主导的基础设施,人工智能技术依赖形成了一个“极化”的悖论:人工智能越是向全球扩散,多数国家越离不开少数国家提供的基础设施,结果是技术越普及、核心能力反而越集中。对能力较弱国家而言,人工智能时代并不必然意味着平等进入未来,而更可能意味着在既有技术生态中被动适配,并因关键能力缺失而面临被进一步边缘化的风险。

其次,当上述技术能力不均衡进一步转化为对外部技术体系的持续依赖时,规则的非对称性就会持续强化。美国在芯片设计、算力供给、云基础设施、基础模型、开发框架以及资本与企业生态等多个关键环节形成了高度整合的技术体系,这种全栈式供给能力使他国人工智能发展从起点便嵌入美国设定的技术环境。围绕先进芯片、模型权重和高性能计算基础设施,美国已将出口管制、投资审查、云服务监管和联盟协调纳入人工智能竞争工具箱。美国商务部也多次更新人工智能扩散框架,围绕先进人工智能芯片和模型权重逐步建立分层准入与安全要求,并推动保护性措施的全球对齐。对此,《美国人工智能行动计划》将面向盟友和伙伴的人工智能确立为政策重点。美国商务部随后推动“美国人工智能出口计划”,要求美国科技企业组成产业联盟,向国际伙伴提供全栈式人工智能技术包;美国国务院推动的《硅谷和平宣言》倡议,也以人工智能和供应链安全为核心,在盟友和可信伙伴之间协调半导体、关键矿产、先进制造、数据基础设施、算力和人工智能模型等关键环节。^①而在技术供给条件控制下,谁能获得算力,谁能使用模型,谁能接入云平台,谁能进入安全合作网络,本身就成为规则形

^① The White House, “America’s AI Action Plan,” July 2025, <https://www.whitehouse.gov/wp-content/uploads/2025/07/Americas-AI-Action-Plan.pdf>; The White House, “Promoting the Export of the American AI Technology Stack,” July 23, 2025, <https://www.whitehouse.gov/presidential-actions/2025/07/promoting-the-export-of-the-american-ai-technology-stack/>; International Trade Administration, U. S. Department of Commerce, “Department of Commerce Begins Inaugural Call for Proposals for American AI Exports Program,” April 1, 2026, <https://www.trade.gov/press-release/department-commerce-begins-inaugural-call-proposals-american-ai-exports-program>; U. S. Department of State, “Pax Silica,” 2026, <https://www.state.gov/pax-silica>.

成之前的权力分配,技术依赖者只能在既有供给体系中争取更大的调整空间。

技术体系还具有较强的生态嵌入性和可扩展性,使得美国无须主动推动规则或者强化监管,便可以依托其广泛技术体系的深度嵌入与不可替代性,形成一种使他国在实践中趋于遵循其技术标准与运行逻辑的结构性影响力。以英伟达的 CUDA 技术生态为例,它已从软件工具演变为 GPU 计算的开发生态,并通过框架适配、应用优化和开发者社区形成持续的技术路径锁定。^① 对许多国家和企业而言,使用美国主导的芯片、云平台 and 模型服务,并不只是一种技术上的选择,也意味着接受与之配套的接口规范、开发环境、安全要求和商业规则。芯片、模型、云服务、开发框架和资本网络共同构成一种结构性吸附力,使其他国家在发展人工智能时不得不考虑与美国技术体系的兼容性和接入条件。由此,即便它们参与人工智能治理规则讨论,也往往处于被动适配而非共同塑造的位置。

最后,地缘政治竞争进一步推动话语权非对称向规则排他性转化。在大国战略竞争背景下,规则制定越来越发挥识别合作对象、限定技术扩散范围和划定竞争边界的制度功能。2025 年 5 月,美国商务部明确将“可信外国”作为技术合作对象,同时强调防止相关技术落入“对手”之手。^② 在这一过程中,“安全”“可信”“负责任”等概念除了回应了人工智能发展本身所带来的真实技术风险外,“安全”概念本身也成为区分合作对象、组织规则协调网络和限制技术扩散的政治语言。美国围绕人工智能技术栈出口、可信伙伴网络和供应链安全形成的一系列制度安排,通过将技术获取条件和基础设施接入资格与政治信任及伙伴关系相联系,在强化人工智能安全治理的同时,也重新界定了人工智能治理的参与边界。由此,人工智能规则越来越难以在开放、普遍的多边协商中形成,而是更可能首先在特定技术体系、价值叙事和安全伙伴网络内部成形。进一步看,在典

^① Stephen Nellis, Max A. Cherney and Krystal Hu, “Behind the Plot to Break Nvidia’s Grip on AI by Targeting Software,” Reuters, March 25, 2024, <https://www.reuters.com/technology/behind-plot-break-nvidias-grip-ai-by-targeting-software-2024-03-25/>.

^② Bureau of Industry and Security, U. S. Department of Commerce, “Department of Commerce Announces Rescission of Biden-Era Artificial Intelligence Diffusion Rule, Strengthens Chip-Related Export Controls,” May 13, 2025, <https://www.bis.gov/press-release/department-commerce-announces-rescission-biden-era-artificial-intelligence-diffusion-rule-strengthens>.

型的集体行动困境下,美国、中国与欧盟等关键主体之间无法形成合作,会显著削弱全球治理机制形成的可能,并加剧规则碎片化。^①

全球人工智能治理所面对的,是一个由技术能力、技术生态、技术权力所预先塑造的非对称的话语权场域:前沿模型的稀缺性、高门槛以及顶尖算力的难以复制,都使得外部对美国人工智能技术产生依赖,而这使美国更容易把技术优势转化为议程设置权力,技术生态控制使其规则倡议更容易获得扩散通道;对华竞争又推动规则更多在特定技术体系和伙伴网络内部形成,并使既有的话语权非对称带有更强的排斥性和阵营化倾向。

由此,人工智能全球治理既呈现出对全球共同安全协调的高度需求,又在实践中不断被技术依赖与大国战略竞争所重塑,其结果只能是在非对称结构中反复生成、调整并固化的话语权失衡,进而表现为在规则形成、解释与执行中的非对称性。这种结构性约束可能会不断削弱人工智能全球治理的实际效力,并使局部治理失灵演变为全面失序。因此,“有限治理”成为防止治理缺位与风险失控的现实选择。它承认规则覆盖、主体参与和执行能力的限度,将合作集中于仍可协调的共同风险,以较低的合作门槛维持必要的治理联系,进而构成人工智能全球治理在长期竞争中保持必要公共性与基本稳定的重要支撑。

三、人工智能全球治理的基本保障

在多边困境将人工智能全球治理推向有限治理的现实设定下,防止治理失效、秩序失序的关键,就在于建立起能够支撑有限治理的底线框架与战略依托。人工智能同时牵动算力基础设施、模型能力扩散、数据控制等多个相互交织的风险领域,这些能力要素同时具有军民两用属性、快速扩散特征与高度不确定的外溢效应,使人工智能的发展与应用成为潜在的战略不稳定来源。与“核战略稳

^① Olajide Olugbade, “In Search of a Global Governance Mechanism for Artificial Intelligence (AI): A Collective Action Perspective”, *Global Public Policy and Governance*, Vol. 5, 2025, pp. 139 ~ 161.

定”依赖“相互确保摧毁”的威慑逻辑不同,^①人工智能的战略稳定并不存在对等毁灭的清晰阈值,其稳定机制更依赖于持续沟通、规则互认和危机管理——这正是“建设性战略稳定”的核心内涵。换言之,核使用本身构成相对明确的危机底线,而人工智能并不具备这种不言自明的边界,其使用限度、安全阈值和不可接受后果并不天然存在,而是在不完全信息条件下通过反复博弈形成的。在这一背景下,人工智能全球治理所需要的是在持续竞争中控制风险、管控分歧并保留合作空间的建设性战略稳定。在全球人工智能能力分布高度不均的条件下,这种稳定能否形成并得到维持,主要取决于关键能力主体之间的互动。就当前的全球能力结构而言,中美两国构成最主要的键能力主体,双方关系也成为人工智能全球治理能否维持开放多边空间的关键变量。本文认为,中美以建设性战略稳定为战略框架,以有限共治为实践路径,构成人工智能全球治理的基本保障。需要澄清的是,建设性战略稳定侧重于战略关系维度,其“建设性”区别于传统依赖核威慑的“冷和平”,并不要求原则的深度趋同,而是强调在竞争常态下实现冲突的可管理性与合作空间的包容性;而有限共治则是这一战略稳定在治理实践维度的具体抓手。中美两国通过有限共治维持最低限度的原则互认、风险沟通和危机管控,是防止全球治理滑向阵营化、碎片化和泛国家安全化的现实路径。

(一) 关键能力主体:人工智能全球治理绕不开中美

人工智能全球治理虽以“全球公共风险”为正当性基础,但高度不对称的技术供给体系,使“全球规则”在形成之初便带有特定技术体系和价值框架的路径依赖。技术能力的集中不仅意味着发展优势的集中,也意味着规则解释权与执行空间在实践中的收缩。因此,能够在关键技术能力上占据主导地位的国家,既是规则制定的重要参与者,也是风险外溢与治理边界形成的核心节点,其行为选择将直接影响全球风险的生成方式与规则约束的可达范围。如果缺乏主要能力主体的实质参与,多边平台、区域安排或行业标准即使能够提供合法性和议题空间,也容易停留在原则表达或局部适配层面。因此,人工智能全球治理不仅是规

^① Thomas C. Schelling, *The Strategy of Conflict*, Cambridge, MA: Harvard University Press, 1960, pp. 35 ~ 54.

范设计问题,更是关键能力主体之间的关系管理问题。

所谓关键能力主体,是指能够形成体系性牵引,进而影响人工智能规则生成、风险认知与治理边界的行为体。这种牵引力来自算力、模型、数据、云平台、基础设施和技术生态等关键技术环节的相互支撑,体现为一国的技术选择能够影响其他主体的发展路径和治理立场,而不仅仅取决于其总体资源规模。从这一标准看,当前全球人工智能格局并不是一个多主体均衡竞争的结构,而是呈现出明显的能力集中于中美两国的趋势。无论是前沿模型、算力基础设施、资本投入,还是产业生态与应用扩散,中美都已形成体系性优势,并与其他国家之间形成明显的断层。

前沿模型是观察这一能力断层的直观切口。中国凭借开源化、低成本和场景化路线,已经在部分能力维度上进入与美国顶级模型竞争、抗衡的阶段。2025 年初以来,中美的前沿模型曾多次交替领先:DeepSeek-R1 曾在 2025 年 2 月短暂接近美国顶尖模型水平;^①2026 年 6 月,中国智谱 AI 研发的 GLM-5.2 模型进一步向美国顶级人工智能模型逼近,其中 GLM-5.2 在部分编码基准上超过美国的 GPT-5.5,并以显著更低的调用成本接近 Anthropic Claude Opus 4.8 等美国顶级闭源模型,在漏洞检测等网络安全任务上已可比肩 Anthropic 顶级模型 Mythos。^②前沿模型的性能提升直观呈现出中美之间激烈的技术竞争,但模型排名的短期变动尚不足以说明两国在全球人工智能能力结构中的地位。中美的主导地位还建立在算力、资本和能源等基础条件的高度集中之上,两国合计控制全球约 90% 的计算能力,并吸引全球 70%—80% 的人工智能投资。^③此外,国际能源署《能源

① Stanford Institute for Human-Centered Artificial Intelligence, “Artificial Intelligence Index Report 2026,” 2026, <https://hai.stanford.edu/ai-index/2026-ai-index-report>.

② Z. ai, “GLM-5.2: Built for Long-Horizon Tasks,” June 16, 2026, <https://z.ai/blog/glm-5.2>; Carl Franzen, “Z. ai’s Open-Weights GLM-5.2 Beats GPT-5.5 on Multiple Long-Horizon Coding Benchmarks for 1/6th the Cost,” VentureBeat, June 16, 2026, <https://venturebeat.com/technology/z-ais-open-weights-glm-5-2-beats-gpt-5-5-on-multiple-long-horizon-coding-benchmarks-for-1-6th-the-cost>; Robert McMillan, “China has Matched Anthropic in Cybersecurity, Resetting AI Race,” The Wall Street Journal, June 28, 2026, <https://www.wsj.com/tech/ai/chinese-ai-anthropic-mythos-cybersecurity-574b02c2>.

③ Rina Chandran, “The Great AI Divide: Navigating U.S. and Chinese Dominance,” June 9, 2026, <https://restofworld.org/2026/ai-divide-america-china-world/>.

与人工智能》报告还指出,2024年至2030年,全球数据中心用电需求将快速增长,其中中国和美国将贡献近80%。^①由此可见,中美之间的人工智能竞争不仅表现为前沿模型的性能竞逐,也建立在算力、资本和能源基础设施的持续扩张之上。两国因此共同塑造着全球人工智能的技术发展方向与竞争格局。

对于前沿人工智能而言,算力是模型训练、推理部署、产品迭代和服务扩张的基础条件;资本厚度则是支撑长期研发、数据中心建设、芯片采购和人才竞争的组织能力。人工智能竞赛的门槛由此大幅提高。多数国家即使能在局部场景、特定产业或监管制度上形成特色,也难以在算力和资本上与中美处于同一层级。中美的能力集中意味着,全球人工智能治理中的算力规则、能源约束、基础设施安全、云平台责任和数据中心监管都无法绕开两国。

由此可见,人工智能全球能力结构更接近一种中美“双核心、双驱动”的格局:一方面,美国依托先进半导体、资本密集型基础设施、专有创新体系和企业生态,仍然处于前沿模型和计算能力的中心位置;另一方面,中国在外部先进芯片被限制的压力下,通过开源扩散、工程效率、应用落地和产业嵌入,形成相对独立的人工智能生态。^②虽然欧洲具有监管制度优势,日本、韩国、荷兰和中国台湾地区在半导体供应链中具有关键节点能力,印度、中东国家以及部分全球南方国家也在数据、能源或应用场景方面拥有一定筹码,并能够在市场准入、规范扩散、议题设置等层面发挥治理作用,但这些优势大多仍是局部性、环节性的,尚不足以支撑完整的人工智能技术生态与体系能力。换言之,人工智能全球竞赛并不是传统经济强国之间处于同一层级的横向竞争,而是以中美为核心展开的结构性竞争:美国有能力凭借全栈技术优势构建人工智能价值链,中国则依托自身的数据规模、应用场景和工程化能力,具备突破并形成完整技术生态的现实可能,其他国家和地区更多是在不同环节中寻求补位、嵌入或依附。

^① International Energy Agency, “Energy and AI,” April 10, 2025, <https://iea.blob.core.windows.net/assets/de9deal3-b07d-42c5-a398-d1b3ae17d866/EnergyandAI.pdf>.

^② LGT Private Banking, “Who has the Edge in the AI Race - China or the USA?” April 13, 2026, <https://www.lgt.com/global-en/market-assessments/insights/investment-strategies/who-has-the-edge-in-the-ai-race-340062>; Kyle Chan et al., “Are the US and China Really in an AI ‘Race’?” Brookings, June 23, 2026, <https://www.brookings.edu/wp-content/uploads/2026/06/The-Beijing-Brief-20260623-US-China-AI.pdf>.

这与既往货物贸易、服务贸易体系下的价值链分工存在显著差异。在人工智能时代,技术能力不再只是价值链中的某一环节,而是同时构成基础设施控制、产业生态组织、规则影响力扩散和安全风险生成的基础条件。无论是模型安全评估、算力与数据监管、开源扩散还是接口标准设定,相关规则都必须通过国家监管与头部企业技术生态落实到具体应用环节,才能产生实际约束力。中美两国在能力规模、技术生态和制度影响力上的集中,使其成为全球人工智能能力结构、风险结构和规则结构的核心塑造者。由此,既有多边协调机制固然可以提供议题聚合和能力建设空间,但难以脱离中美技术生态独立形成强约束能力。人工智能全球治理的关键在于,中美两个关键技术能力主体之间能否形成最低限度的互动约束与风险管控框架。

(二)有限共治:中美建设性战略稳定的最低实现路径

中美作为人工智能竞赛的关键能力主体,技术赋能与权力竞争的叠加使两国战略竞争具有结构性和长期性,并与本轮技术革命的长周期相伴。人工智能革命的战略性、权力性和结果不可预测性,使技术能力强国担忧其未来权力位置被改写,进而持续加大战略应对与竞争投入。因此,人工智能全球治理的现实目标在于防止中美竞争失去边界。

对于人工智能这样兼具通用技术、战略能力和社会基础设施属性的新兴技术而言,大国竞争一旦完全脱离沟通机制,技术风险、产业竞争和安全疑虑便可能相互叠加,进而削弱全球性规则的实际约束力。在前述有限治理状态下,建设性战略稳定为中美管控竞争提供了战略框架,有限共治则构成这一框架在治理实践中的现实路径。有限共治是以中美两个关键能力主体之间的有限协调为基础,确保双方在最可能引发误判、外溢和危机升级的领域中,仍能保留必要的风险沟通、规则互认和危机管理机制,进而形成底线性稳定。这种底线性稳定能够防止竞争彻底消解风险治理,确保大国博弈不滑向不可沟通、不可预期与不可管理的失控状态,从而为人工智能全球治理保留基本的公共空间与合作可能。作为维持中美建设性战略稳定的手段,两国的有限共治应包含以下几个方面:

首先,有限共治需要稳定的政府间沟通机制。人工智能正在进入国家安全、

经济运行、社会治理和国际竞争的核心领域,其风险往往容易被安全化解读,也更容易在大国互疑中被放大。而若中美之间缺乏基本沟通渠道,技术与产业层面的事件与变化便可能被迅速解释为战略意图,治理分歧也可能进一步转化为安全对抗。2026年5月特朗普访华期间,中美元首围绕人工智能进行了建设性交流,并同意就人工智能问题举行两国政府间对话。^①这一安排说明,即便在技术竞争加剧的背景下,中美仍然存在围绕人工智能风险保持沟通的现实需求。这种沟通可以采取多种形式,包括但不限于高级别战略对话、主管部门常态化磋商、政府主导的政策与技术工作组交流,以及突发事件下的紧急联络与信息通报机制等,而此类对话的意义并不在于立即消除分歧,而在于防止双方在缺乏解释机制的情况下误判对方意图,并为未来治理规则的协调保留必要入口。

其次,有限共治需要在高风险领域形成最低限度的规则互认。围绕军事人工智能、关键基础设施安全、前沿模型风险等可能影响国际安全的高风险议题,中美不必共享同一套治理哲学,也可以承认某些风险具有不可接受的共同后果。2026年新加坡香格里拉峰会对话期间,有关军事人工智能治理的讨论表明,尽管在规则形成路径和制度平台选择上仍存在分歧,中美都意识到军事人工智能规则的重要性。^②军事领域的人工智能化本身也会对中美战略稳定产生负面冲击,中美应优先在军事人工智能等关联危机升级与误判风险的领域形成底线性共识,包括确立自主武器系统的人类控制原则,明确关键决策环节的可追溯性与可解释性要求,并对人工智能在预警、情报处理和辅助分析中的使用边界设置最低限度的行为约束,避免竞争突破危机管理的基本边界。^③

再次,有限共治需要防止人工智能治理机制完全阵营化。当前人工智能治理已经出现向技术联盟、安全伙伴网络和价值共同体内部收缩的趋势。2025年12月,美国发起“硅和平”倡议,强调通过国际合作建立可靠供应链和可信技术网

① 《外交部:中美同意开展人工智能政府间对话》,新华网,2026年5月19日, <https://www.xinhuanet.com/world/20260519/883ac1ee99c74a8fa2441da4d4b40e96/c.html>。

② Meia Nouwens, “Military AI Governance Under Strain: The US-China Dialogue,” International Institute for Strategic Studies, June 2026, <https://www.iiss.org/online-analysis/online-analysis/2026/06/military-ai-governance-under-strain-the-uschina-dialogue/>。

③ 丁伊、江天骅:《美国军备人工智能化对中美战略稳定的冲击》,《国际展望》2025年第2期,第88~109页。

络,但其合作框架主要围绕“可信伙伴”展开。伙伴身份由此逐渐成为进入相关技术网络的重要条件,甚至可能演变为参与相关合作的排他性条件。^① 如果这种趋势继续扩大,全球人工智能治理就可能从“共同规则”转向“阵营规则”,不同国家将在不同技术体系、标准体系和安全承诺之间被迫选择。对全球南方国家和中等技术水平国家而言,这不仅会加剧技术依赖,也会压缩其发展空间,削弱其规则参与能力。因此,即便中美分别拥有不同技术生态,也仍应维护联合国、二十国集团以及人工智能安全全球峰会等相对开放的多边平台,防止人工智能治理完全陷入阵营对立,为不同能力层级、不同发展阶段和不同制度背景的国家保留参与规则讨论和技术选择的自主空间。

最后,有限共治还意味着承认人工智能治理的阶段性目标。应该看到,由于人工智能技术还在加速演进之中,技术不确定性使得人工智能全球治理在短期内难以建立覆盖所有模型、算力、数据和应用场景的完整规则体系。2025 年发布的《国际人工智能安全报告》将现阶段任务限定于梳理通用人工智能的能力、风险与应对方法,并明确区分已有共识、存在分歧和证据不足的领域,此后,该报告采取滚动评估方式,通过专题更新和年度报告持续修正风险判断,为阶段性治理提供了参考。^② 因此,中美可以根据技术发展和风险证据的变化,为不同阶段设定可评估、可执行的治理目标,并据此逐步调整治理对象、政策工具和规则强度。倘若中美能够围绕各阶段的技术发展,就竞争边界、风险沟通和底线设置形成稳定预期,全球治理便可望在完整制度尚未形成时维持基本秩序,并为后续调整治理目标和规则强度保留空间。

只要中美之间仍然存在最低限度的沟通和互认,那么人工智能全球治理就仍有可能在碎片化的结构中维持基本运作,并为更多国家参与规则协商保留必

① 美国国务院拟向 35 个《人工智能机遇声明》签署国发出信函,警告若同时加入中国主导的人工智能合作机制可能导致其被排除在美国主导的合作体系之外。该信函是否发送及最终内容尚未确定,但表明美国存在利用多边机制推动伙伴国作出阵营选择的政策倾向。参见 Michael Martina, “U. S. to Tell Partners They Must Pick Sides in AI Race with China,” Reuters, August 14, 2026, <https://www.reuters.com/world/china/us-tell-partners-they-must-pick-sides-ai-race-with-china-2026-08-14/>。

② Yoshua Bengio et al., “International AI Safety Report 2025,” January 29, 2025, <https://internationalaisafetyreport.org/publication/international-ai-safety-report-2025>; Yoshua Bengio et al., “International AI Safety Report 2026,” February 3, 2026, <https://internationalaisafetyreport.org/publication/international-ai-safety-report-2026>。

要空间。而一旦中美建设性战略稳定关系破裂,有限共治将无以为继,两国的技术竞争、安全疑虑与规则分歧将进一步强化,人工智能全球治理则可能在大国安全竞争的持续牵引下,逐渐丧失作为公共风险治理机制的独立空间,再多的原则文件、伦理共识和多边倡议也会迅速失去实际约束力。同时,人工智能武器化的风险也会伴随技术升级而加剧,并给人工智能技术发展本身及未来世界秩序投下风险阴影。由此,中美有限共治虽然无法消除竞争,却能够为约束竞争风险、维持多边合作和推动全球治理逐步发展提供基本条件。

四、结 语

虽然围绕人工智能的风险讨论广泛展开,国际社会也并非缺乏关于安全、可信、负责任等原则性共识,但风险共识并不自动转化为强约束规则。必须看到,不同于传统贸易或一般数字治理议题,人工智能既具有跨国风险属性,又深度嵌入未来竞争格局;它既需要共同治理,又难以摆脱能力差异与技术依赖的结构性影响。世界各国真正的分歧,在于是否愿意接受可能削弱自身技术积累、产业优势和战略空间的制度约束。人工智能一旦成为数据、模型、系统和决策过程的组织方式,安全边界便由外部风险管理延伸至国家内部运行结构,治理议题也随之触及并试探利益博弈、安全防线与制度控制权的交叉敏感区,风险共识也由此显现出其原生性的内在限度。当共识难以稳定转化为一致的规则约束时,技术能力的不均衡便会更直接地影响议程设置和标准制定,使规则解释权和执行空间呈现出显著的不对称性。由此,本文所指的多边困境三层面——利益协调、制度嵌入与权力结构,在人工智能治理中相互牵引,使共识与合作难以进一步汇聚为统一且具有强约束力的治理安排。

在此现实下,有限治理虽不是理想制度的自然延伸,但成为各方维持最低限度合作与避免竞争彻底失序的现实取向。作为一种缺乏统一强制机制支撑的底线秩序,有限治理在很大程度上依赖关键技术供给者的自我约束与底线沟通。正是在这一意义上,人工智能全球治理难以绕开全球人工智能能力结构、风险结

构和规则结构的共同核心塑造者中国与美国及两者之间的互动方式。处于关键技术供给体系之外的国家,即便在形式上参与全球治理,也往往难以摆脱被动适配规则的地位;而中美之间的互动方式,则会直接影响人工智能全球治理是保留开放性的多边空间,还是滑向排他性的阵营对峙。因此,中美之间的建设性战略稳定是人工智能有限治理保持实际效力与开放空间的内在前提,也为人工智能领域分歧的管控、共识的积累与合作的维持保留必要的时间与空间。更长远地看,人工智能正逐步成为影响国家能力差距、国际权力分配和多边制度效力的基础变量。技术能力的集中将改变国际公共产品的供给方式,也会使形式上的主权平等与实际上的治理能力差距更加突出,未来国际秩序可能由此呈现出权力进一步集中、国家分层更加明显和制度竞争持续加深的趋势。在这一秩序演变过程中,前沿技术突变、重大安全事故、人工智能军事化升级或新一轮技术封锁,都可能削弱中美建设性战略稳定的可预期性。若双方无法落实有限共治,局部摩擦可能引发治理合作的整体收缩,并进一步削弱国际秩序的公共性与稳定性。也正是在这一意义上,推动中美关系在竞争中保持建设性战略稳定、在分歧中寻求有限共识,不仅关乎两国自身发展,更关乎人工智能时代全球公共安全与人类共同未来,也是构建开放包容、合作共赢国际治理格局的重要基石。

【责任编辑:卫蔚】