人工智能国际治理：基于技术特性与议题属性的分析*

薛 澜 赵 静

当人工智能为全球经济社会发展增添了新的不确定性时，人工智能国际治理便走到了一个历史关口。通用大模型商业应用引发各界对人工智能全球性议题的关注以及对相关发展红利和潜在风险的广泛讨论，在地缘政治环境日益复杂的背景下，人工智能发展也被赋予国家间竞争的涵义。人工智能国际治理不能被简单视为普通的新兴技术全球治理，其具有的复合技术特性和多重议题属性亟须关注。人工智能的技术特性和议题属性带来了国际治理难点：各国均不具有实践经验与主导能力，国内议题与全球议题的传导性加强，新旧技术治理范式间存在摩擦。人工智能国际治理实践也在“利益相关方治理”的基础上，增加了“主权国家战略博弈”的内涵，并呈现出各方主体“治理敏捷性”的实践特色与治理趋势。为构建具有包容性的人工智能国际治理机制，本文提出原则先行、分类敏捷、中间连接以及技术自治等治理思路，以启发国际社会各方参与主体的思考与行动。

关键词：人工智能 国际治理 新兴技术 全球治理体系 治理敏捷性

随着人工智能技术的高速发展和在公私领域的广泛应用，隐私保护、技术性失业、新型武器、数字贸易争端等问题引起关注，全球经济社会发展的不确定性也因人工智能而上升。一方面，人工智能发展速度之快和应用范围之广令其成为全球经济繁荣的主要驱动力，有助于发达经济体和新兴经济体的发展。[1]另一方

薛澜系清华大学苏世民书院院长、人工智能国际治理研究院院长、公共管理学院教授；赵静（通讯作者）系清华大学公共管理学院副教授、产业发展与环境治理研究中心副主任，Email: jingzhao09@tsinghua.edu.cn。

*本文系国家社会科学基金重大专项项目、中国科协国际前沿科技治理对话项目、清华大学自主科研计划“算法全球治理与分类治理的理论探索”（项目编号：2021THZWJC12）的阶段性研究成果。感谢审稿人和编辑部的意见，当然文责自负。

[1] Manyika J. and Spence M., “The Coming AI Economic Revolution”, *Foreign Affairs*, 102(6): 70-86, 2023.

面，人工智能的潜在风险和引发的治理问题备受全球关注。从个体隐私、伦理风险、社会公平、就业替代，到技术竞争、军事安全、地缘政治，人工智能引发的全球性问题已波及“地球村”的方方面面。人工智能引起国际社会的普遍关注，在国际层面规范技术的发展与应用已经成为影响各国重大政治经济利益和人类社会发展进步的关键议题。

近年来，随着人工智能的应用进一步扩大，技术释放的巨大潜力和带来的现实风险随之显现。在驱动经济增长和提升民生福祉方面，人工智能的强大能力令世界惊叹。根据麦肯锡全球研究所对人工智能经济潜力的预测，技术应用带来的生产率提升可每年为全球经济增加15万亿美元。[1]促进人工智能技术的创新、发展和扩散是各国的普遍共识和积极愿景。但与此同时，技术的风险和潜在危害也已进入国际社会的视野。技术的不确定性和影响的不可逆转性令其损害上限和危害规模无从得知，而这一新技术在群体间、产业间与国家间也产生了非对称的经济社会影响。这都是各国当前重要的国内政策议程和参与国际事务讨论的重点。特别是，人工智能还具有民用和军用的双重用途，防止军事领域滥用人工智能和避免技术军备竞赛也是重要的国际监管事项。

然而，从国际治理的维度看，人工智能可能带来的风险并不能被简单视为普通的新兴技术风险。人工智能技术发展的同时，地缘政治环境日益复杂，这使得技术发展被赋予国家间竞争的涵义，也使得国际治理更加重要、迫切和具有挑战性。以中国和美国为例，两国都是技术先驱并在资金上具有优势，都在争夺人工智能治理的领导地位，因而技术创新和芯片资源成为影响两国产业实力和两国关系的关键因素。此外，人工智能引发的全球性问题还与传统国际议题相结合，成为融合全球负外部性、个体伦理道德、国家安全风险、贸易与产业竞争等因素的多维治理话题，甚至可以说几乎覆盖当前所有全球性话题。这意味着，人工智能国际治理既不同于冷战前以安全与和平为目标、以国家权力为主导的“维持国家共生关系的治理事务”，也不同于21世纪以来围绕全球公共物品供给的责任分担和资金承诺进行讨论的“问题导向的治理话题”，这就需要从议题属性角度重新审视人工智能国际治理。

国际社会普遍认为，人工智能国际治理的目标是平衡各国经济、社会和文化

[1] McKinsey Company, “The Economic Potential of Generative AI”, June 2023, https://www.mckinsey.com/capabilities/mckinsey-digital/our-insights/the-economic-potential-of-generative-ai-the-next-productivity-frontier[2024-04-19].

诉求，在确保风险可控的前提下释放人工智能的潜在价值，从而赋能全球可持续发展。蕴含于技术之中的复杂问题使得人工智能国际治理的体系与规则构建迅速成为重要的全球治理议题。过去十年，人工智能议题已经迅速进入国际治理议程，从起初的非国家主体倡议、政府和国际性机构参与，演变到重要国际组织介入，形成了多主体、多角度、多层级的全球治理机制复合体。[1]尤其是随着2023年以来联合国积极参与、推动人工智能全球监管机构的建设和治理机制的构建，人工智能国际治理正式进入多方博弈的新阶段。当前松散的机制复合体在形成有效合力和达成共识上还存在诸多障碍，在形成合适的国际治理机制方面仍然任重道远。特别是在一些关键议题上，如消除社会对技术的恐惧，国际治理滞后已经制约了人工智能发展。

随着人工智能国际治理进入多方博弈的新阶段，国际社会如何在治理机制构建初期形成引导性的理念来避免人工智能治理失序？倘若人工智能国际治理无法延续传统技术治理模式而迫切需要新型治理机制，在国际社会联合起来设计国际治理框架之前，应当关注哪些要点？这些都是未来一段时间全球治理学界和实践界的重要关切。基于当前就如何规制人工智能众说纷纭的情况，本文跳出现有框架，探讨人工智能本身的技术特性和相关议题属性，分析当前人工智能国际治理面临复杂情况的根本原因。同时，在梳理人工智能理论探讨与治理实践的基础上，给出构建人工智能国际治理体系的新思路。

人工智能的议题属性与国际治理挑战

（一）人工智能的技术特性与议题属性

人工智能治理属于典型的新兴技术治理。在特定时空背景下，任何新兴技术都具有新颖性、不确定性、模糊性、相对快速的成长性、连贯性以及对经济社会的显著影响。[2]新兴技术兼具收益和风险的二重性，相关治理目标不仅包括构建

[1] Cihon P., Maas M. M. and Kemp L., “Fragmentation and the Future: Investigating Architectures for International AI Governance”, *Global Policy*, 11(5): 545–556, 2020; Alter K. J. and Meunier S., “The Politics of International Regime Complexity”, *Perspectives on Politics*, 7(1): 13–24, 2009; Raustiala K. and Victor D. G., “The Regime Complex for Plant Genetic Resources”, *International Organization*, 58(2): 277–309, 2004. 机制复合体主要指某一全球议题领域并行存在的多种治理机制，其特征是治理机制和治理主体多元化，彼此之间没有层级关系，不同机制在遵循原则、规制、程序等方面存在差异和冲突。

[2] Rotolo D., Hicks D. and Martin B. R., “What Is an Emerging Technology?” *Research Policy*, 44(10): 1827–1843, 2015.

鼓励技术转移、知识共享的规则，而且包括预防和应对技术创新产生的风险以及潜在的社会影响。[1]然而，新兴技术治理并非易事。一方面，当技术应用的负面影响显现时，新兴技术往往已成为整个经济和社会结构的重要组成部分，并形成了既得利益群体，以致对其治理变得极其困难、昂贵和耗时。[2]由于技术红利、社会福祉、伦理风险以及潜在社会影响，技术治理考验着人类社会的智慧。另一方面，技术创新对经济的驱动力，往往令其成为全球经贸博弈和各国产业竞争的焦点，具有通用技术特征的新兴技术更是如此。新兴技术发展历来是关系大国兴衰的重要因素之一。

类似于生物基因、信息科技等新兴技术，人工智能国际治理也会遇到挑战与困难。不同的是，人工智能应用涉及的各类全球性问题，如数据主权、自主武器、劳动市场结构、深度伪造，已远远超出传统技术治理涉及的风险与安全等单一议题范畴。各国都高度关注全球人工智能政策及战略。学者们指出，当前人工智能治理无法延续传统技术治理思路。[3]本文将具体原因归结为人工智能的复合技术特性和多重议题属性。

第一，复合技术特性。人工智能是一类技术的集合，具有各种应用类型和风险状况，且未形成固定的技术概念。[4]纵观技术发展史，人工智能并不是第一个具有突出特性的技术，[5]却是人类社会截至目前遇到的第一种将多类技术特性结合起来的技术。[6]

总体而言，人工智能具有如下典型的技术特征。一是技术进步的超快迭代速度。在技术创新规律方面，摩尔定律是预测计算机处理能力的经典标准。人工智能模型尤其是刚刚出现的通用人工智能模型的迭代速度远远超出这一定律，成为

[1] Kuhlmann S., Stegmaier P. and Konrad K., “The Tentative Governance of Emerging Science and Technology–A Conceptual Introduction”, *Research Policy*, 48(5): 1091–1097, 2019.

[2] Genus A. and Stirling A., “Collingridge and the Dilemma of Control: Towards Responsible and Accountable Innovation”, *Research Policy*, 47(1): 61–69, 2018.

[3] Flournoy M. A., “AI Is Already at War”, *Foreign Affairs*, 102(6): 56–69, 2023; Bremmer I. and Suleyman M., “The AI Power Paradox: Can States Learn to Govern Artificial Intelligence–Before It’s Too Late?” *Foreign Affairs*, 102(5): 26–43, 2023.

[4] Stone P., Brooks R., Brynjolfsson E. et al., “Artificial Intelligence and Life in 2030”, Stanford University, 2016, https://ai100. stanford. edu/2016–report[2024–05–07]; Russell S. J. and Norvig P., *Artificial Intelligence: A Modern Approach*, Boston: Pearson, 2020.

[5] 如生物技术、信息技术因具有较强的通用性而对经济社会具有明显的影响。

[6] Bremmer I. and Suleyman M., “The AI Power Paradox: Can States Learn to Govern Artificial Intelligence–Before It’s Too Late?” *Foreign Affairs*, 102(5): 26–43, 2023.

继集成电路、信息技术之后的技术创新代表。二是技术风险的不确定性与广泛性。不同于传统制造业技术的可改进性和风险个体化，而类似重组脱氧核糖核酸（DNA）、转基因技术，人工智能的技术风险具有不可计算、不受控制、易接触、传导广泛等特点，这都令其潜在损害没有上限，社会影响巨大且不可逆。三是技术扩散的便捷性。人工智能得益于开源知识和期刊论文，具有技术扩散的便捷性，并且影响其技术迭代与竞争力的因素是相关资源投入，如数据、算力、芯片、能源、人力资本。诸多技术在商业需求的驱动下实现了相对快速的跨国转移和技术应用。四是技术的智能涌现性和不可预期性。与一般的信息技术相比，人工智能具有自我创造、超强学习和超级进化的特性，这一“技术潜力”令人类社会对技术风险产生极大的担忧和恐惧。

第二，多重议题属性。人工智能引发各类全球性问题，尤其是人工智能“发展的流动性”和“风险的广泛性”导致国际社会将各类传统议题均转化为对人工智能国际治理的关切，呈现出典型的多重议题属性特征。

人工智能发展具有流动性，也就是指，随着技术创新与各类应用场景的开发，各类供需被快速催生，产业可以很快实现全球流动。人工智能技术发展促进跨国企业生产、研发以及多元应用场景的全球化，连接全球数字经济活动的各类需求与供给。在人工智能研究和开发中，数学基础、模型构建技能，叠加商业利益与创新，共同推动技术强劲发展，令各行各业都受到技术发展的影响。在全球化背景下，从发展角度来看，人工智能国际治理不可避免地涉及贸易、产业等传统全球经济治理议题，在一定程度上体现了“利益属性”。[1]并且，技术创新发展和知识红利释放还涉及技术合作共享、标准制定、数据利用等方面，使得人工智能国际治理还具有“共有产品属性”。

人工智能产生的风险具有广泛性。技术应用的范围从私域到公域，[2]涉及各行各业，导致几乎所有的个体、组织和国家都将直面技术风险，这将带来深远的社会影响。由此，人工智能的国际治理维度出现了类似气候变化、环境能源治理的特征，即因个别主体治理不到位而产生“外部性属性”。解决外部性问题则需要在国际层面形成具有包容性的人工智能治理体系。此外，人工智能国际治理涉

[1] 韩永辉、张帆、彭嘉成：“秩序重构：人工智能冲击下的全球经济治理”，《世界经济与政治》，2023年第1期；部彦君：“人工智能发展下的权力扩散态势解析与挑战应对”，《科学学研究》，2024年3月网络首发。

[2] van Noordt C. and Misuraca G.,“Artificial Intelligence for the Public Sector: Results of Landscaping the Use of AI in Government across the European Union”, *Government Information Quarterly*, 39(2): 101714, 2022.

及民族与地理数据、语言文化、国家安全、恐怖主义等问题，蒙上了地缘政治与国际关系的色彩。[1]从这一角度看，是国际社会对可能发生大规模经济社会变革的担忧，而不是人工智能技术的新颖性，使其成为亟须关注的问题。[2]

人工智能带来的安全风险也使得“维持国家共生”和“人类主宰命运”被列入国际治理议程。人工智能霸权成为军事优势的代名词，而霸权竞争将日益激烈，从而影响地缘政治走向。人工智能与军事技术的结合为国家间军事竞争带来更大的不确定性，如人工智能可按“机器速度”实施自动化网络攻击、合成化学武器、加速情报获取与分析，甚至在战争中误导自动化武器。这使大国面临困难选择。从国家安全战略看，各国难以接受人工智能军事化应用落后带来风险，但在缺乏应对军事打击的措施和国际安全盟约的情况下，各国间的技术竞赛也同样带来较大风险。[3]基辛格（Henry A. Kissinger）认为，第二次世界大战后全球安全很大程度上归功于形成防止核武器扩散的国际治理机制，而国际社会未来的安全取决于是否能够建立防止人工智能技术军备竞赛的机制。[4]由于人工智能在民用和军用之间的界限较为模糊，各国对人工智能军事应用的警觉将不可避免地影响民用领域的技术发展。人们在人工智能上的终极梦魇就是技术失控。人工智能的“觉醒”或将导致其摆脱人类控制，甚至成为人类的主人。如同当年转基因、克隆羊、试管婴儿等技术的出现引发科学界热议那样，防止技术失控将成为人工智能国际治理中的优先事项，相应的预防原则、预案准备、制度建设亟待商讨。

（二）人工智能国际治理面临的挑战

国际治理机制随着全球性问题的变化而演进，人类社会在不同历史时期获取了“看似切实可行”的不同解决方案。具体而言，第二次世界大战后的国际治理机制是为实现民族国家共生而形成，其核心是通过协议和谈判维持和平稳定的国际环境，由具有霸权的国家主导和传播治理理念，并构建国际层面的、制度化的组织与协调机制，如国际货币基金组织、联合国。[5]伴随世界发展与全球化进程

[1] 余南平：“新一代通用人工智能对国际关系的影响探究”，《国际问题研究》，2023 年第 4 期；戚凯：“ChatGPT 与数字时代的国际竞争”，《国际论坛》，2023 年第 4 期。

[2] Tallberg J., Erman E., Furendal M., Geith J., Klamberg M. and Lundgren M., “The Global Governance of Artificial Intelligence: Next Steps for Empirical and Normative Research”, *International Studies Review*, 25(3): 1–25, 2023.

[3] Flournoy M. A., “AI Is Already at War”, *Foreign Affairs*, 102(6): 56–69, 2023.

[4] 于潇清：“清华教授回忆十月底在美拜访基辛格情景：献策中美，忧心 AI”，2023 年 11 月 30 日，https://www.thepaper.cn/newsDetail_forward_25484784[2024-04-19]。

[5] 薛澜、关婷：“多元国家治理模式下的全球治理： 理想与现实”，《政治学研究》，2021 年第 3 期。

推进，国际事务日益增多，国内与国际事务的界限并不分明，这便改变了国际治理的议题结构和参与主体。全球治理从以权力为中心的传统范式，转变为以议题为中心的新范式，[1]一系列新的全球问题解决机制和各类国际组织日益兴起。进入21世纪，全球化加速，围绕各类严峻挑战，形成了多层次、多主体、多形态的国际治理体系和机制。与此同时，涉及部分议题的全球治理失序现象频繁出现，诸多国际事务陷入治理僵局和失序状态，传统国际组织难以就重大事件做出决策，看似越来越复杂的全球治理体系显得应对乏力、运行低效。[2]

纵观当前的全球性话题，如经济与贸易、环境与气候变化、基因与生物多样性、武器与恐怖主义，均呈现出差异化的治理结构和不同的治理难题。例如，在环境与气候变化领域，以《京都议定书》的签署为起始，形成了多种治理机制并存的局面，包括高度制度化的国际规则、高度碎片化的非正式机制以及位于谱系中间的机制复合体等，以解决涉及复杂科学问题、具有跨边界与跨部门影响、未来趋势不可预测的治理难题。即便如此，国际层面的气候变化谈判仍然陷入僵局，多边环境条约难以执行。再如国际贸易领域，全球市场自由化和公平贸易是治理目标，世界贸易组织框架下的国际贸易协定是实现治理目标的有效途径，然而，因跨越多个行业并涉及标准制定，在多国加入的情况下，贸易谈判成本不断增加，主权国家影响力下降，跨国企业和非政府组织开始负责制定全球性行业标准。

基于人工智能的复合技术特性和多重议题属性，与其他全球性议题相比，其国际治理存在以下三个突出的难点。

第一，国际规则制定的主导者和新兴参与者在全球治理经验方面站在同一起跑线上。一般而言，主导国通过输出自身治理经验影响国际规则制定。人工智能是新兴技术，在如何促进和规范人工智能发展方面，各国都缺乏治理经验。一方面，在安全与风险维度，国际规则制定的主导者和新兴参与者都站在同一起跑线上。人工智能技术创新迭代的超快速度使其治理经验的可复制性和推广性减弱，而且其相对的低成本、易扩散特性使得在技术上占据绝对主导地位的国家较难出现。当然，个别国家仍然在支撑人工智能发展的算力维度（如高端芯片）上处于

[1] 薛澜、俞晗之：“迈向公共管理范式的全球治理——基于‘问题—主体—机制’ 框架的分析”，《中国社会科学》，2015年第11期。

[2] Xue L., “The Shifting Global Order: A Dangerous Transition or an Era of Opportunity?” *Governance*, 25 (4): 535–538, 2012.

绝对领先的位置。[1]例如，斯坦福大学发布的2023年人工智能年度报告指出，人工智能领域的研究与生产在2010－2020年呈现出越来越明显的地理分散性趋势。[2]此外，人工智能的潜在风险极高，尚未发生重大灾难和事故之前，人类始终无法清晰认知其可能产生的风险与伦理问题。

另一方面，受经济利益驱使，各国存在激烈的技术竞争和对全球产业主导权的争夺。传统的规则制定国较难通过扩散政策理念引领国际治理。以数据跨境流动为例，即便欧盟出台的《通用数据保护条例》（GDPR）影响了众多国家的数据跨境流动规则，但全球仍然形成了自由流动、有条件流动、数据本地化等各类规则。[3]当前，在“充分保护认定”规则基础上形成的数据流动圈日益增多，但这些数据联盟在解决海量数据跨境流动需求方面表现欠佳。基于此，私营部门和非政府部门开始扮演数据中转站的角色，Gaia-X、Catena-X等欧洲共同数据空间和Dawex、The FSM等私营数据中介平台开始兴起。因此，在技术浪潮驱动下，治理规则受不同参与者影响，而治理的参与者也会随技术迭代而发生变化。在所有参与者都不具备治理经验且竞争充分的情况下，由谁提供公共物品和输出治理经验则成为影响治理机制形成的关键因素。

第二，国内议题与国际议题的传导性加强，导致形成具有包容性的规则体系并不容易。全球治理需建立在各国国内治理的基础上，尤其是面对具有外部性特征的全球议题时，各国首先需要做好国内治理。跨国公司的全球发展和以数据训练与算法迭代为特征的技术创新模式，使得人工智能引发的国内与国际事务以及背后的风险与治理问题紧密缠绕，并逐渐呈现互相传导的趋势。一方面，人工智能技术的易扩散性和风险广泛性意味着国际治理十分重要，不能存在监管漏洞，而各国监管体系的简单拼凑与对接事实上无益于减少全球性风险。例如，各国人工智能治理水平和相关规则的差异为非政府部门的“套利”行为提供了土壤，使

[1] National Science and Technology Council, “The National Artificial Intelligence R&D Strategic Plan”, May 2023, https://www. whitehouse. gov/wp-content/uploads/2023/05/National-Artificial-Intelligence-Research-and-Development-Strategic-Plan-2023-Update.pdf[2024-04-19]；薛澜、魏少军、李燕、贺俊、罗长远、余振、杨荣珍：“美国《芯片与科学法》及其影响分析”，《国际经济评论》，2022年第6期。

[2] Maslej N., Fattorini L., Brynjolfsson E., Etchemendy J., Ligett K., Lyons T., Manyika J., Ngo H., Niebles J., Parli V., Shoham Y., Wald R., Clark J. and Perrault R., “The AI Index 2023 Annual Report”, AI Index Steering Committee, Institute for Human-Centered AI, Stanford University, April 2023, https://aiindex. stanford. edu/wpcontent/uploads/2023/04/HAI_AI-Index-Report_2023.pdf[2024-04-19].

[3] Casalini F. and Javier L. G., “Trade and Cross-Border Data Flows”, OECD Trade Policy Papers, OECD Publishing, No. 220, 2019.

得研发、应用出现了从强规制国家向弱规制国家转移的现象。如果人工智能技术大国过度监管，则可能促使优秀的创新者和积极的投资者转移至监管力度较弱的国家。

另一方面，全球需要一套规制体系来预防各类风险，形成基本的安全共识。虽然当前各国在人工智能需要受到人类控制这一安全底线上具有共识，但遗憾的是，在人工智能应用的具体风险上，各国却因地域文化、意识形态、价值观念上的差异而形成不同的治理思路。这些差异可能是阻碍部分领域形成全球共识的关键原因。以人脸识别技术为例，各国对技术应用的限制存在较大差别。有些国家民众高度重视个人隐私保护，故难以接受人工智能的部分应用；有些国家民众具有较高的技术接受度，较为认可在合理风险范围内获取便捷的技术应用。再如，针对人工智能在智慧司法判决和预测性警务中的应用，也难以达成治理共识。因此，如何弥合认知差异、共同应对风险，成为人工智能国际治理的重要议题。同时，人工智能发展将从全球治理框架中获益。例如当跨国公司全球生产时，相对一致的标准、规则和监管体系可以极大减少跨国企业的合规成本，并增进民生福祉。因而，构建具有包容性的国际规则将有助于发挥制度效力，但构建过程也极具挑战性。

第三，新旧全球治理范式共存，并在多重议题中出现碰撞、冲突和制度摩擦。鉴于人工智能概念宽泛且涉及多重议题，以主权国家为中心的传统治理方式和利益相关方发挥主导作用的治理方式将同时存在。人工智能国际治理呈现出机制复合体、多中心治理、治理碎片化等特点。具体而言，当人工智能引发的议题具有“民族、国家属性”和“利益属性”时，主权国家和国际组织将积极参与其中，传统全球治理机制将发挥作用，如数据主权、自动化武器等议题；当具有“共有产品属性”和“外部性属性”时，多方治理主体和多元治理机制则扮演更重要的角色，如数据跨境流动、技术扩散等议题。当前，利益相关方很难清晰区分人工智能国际治理的“对象和问题”，也较难识别全球性议题背后的元问题。各方对技术的理解和专业知识积累需要较为漫长的过程。

在全球治理过程中，这些尚未被明确界定的全球性问题或将导致新旧全球治理范式间出现碰撞和冲突。尤其是当霸权国家试图提升在人工智能议题上的国际话语权时，意识形态先行的全球治理策略和国际合作将带来更大的分裂，导致技术治理被转化为关联意识形态、文化理念和地缘政治的议题。而这恰恰与多方治理主体倡导的包容性合作、平等参与等理念产生冲突。以算法的全球治理为例，

算法概念的模糊性以及与数据的复杂关系影响了算法全球治理议题的内涵，导致利益相关方对问题的理解存在差异。例如，当前由算法引发的全球治理议题中，针对算法是私人产品、集体产品还是公共产品，尚未形成共识。对问题较难定性使得相关对话和讨论并没有处于同一个层面。[1]主权国家关注的数据主权涉及算法的私人产品属性，而多方主体关注算法的集体产品或公共产品属性。因此，主权国家主导的规则体系和利益相关方倡导的治理理念之间的差异成为当前人工智能国际治理的首要难题，由此带来的制度摩擦减弱了治理效果，人工智能国际治理的机制复合体面临协调困境与合力不足的情况。[2]

人工智能国际治理的实践与特点

（一）人工智能国际治理进入博弈新阶段

自2016年人工智能成为全球性话题以来，据不完全统计，跨国家的国际性技术社群、非政府组织、企业联盟、国际组织等各类主体共提出近200项关于科技与伦理的治理建议，这些建议促进了人工智能应用领域的国际治理合作。相较而言，主权国家起初更多关注国内监管和产业布局，而较少参与该领域的全球性事务，各国甚至在产业竞争中不断将国内产业发展和监管问题，与全球贸易、科技伦理等议题挂钩，在一定程度上阻碍了治理共识的形成。究其原因在于，当技术竞争尚不明朗时，以隐私伦理、社会公平、国家安全等为名指责竞争对手的治理情况是阻碍他国技术发展的优先策略选择。但随着技术应用范围扩大和深度拓展，人工智能的经济效益、外部性及其治理影响日益引发关注。各国政府在人工智能领域的国际参与程度和战略部署开始明显改善，表现出试图影响全球层面技术标准、应用规范、治理准则的意图。尤其是在以生成式人工智能为代表的技术迭代和聊天机器人ChatGPT商业模式浮现后，人工智能国际治理的重要性和紧迫性迅速提升，各国政府加速了全球布局。

总体而言，当前的人工智能国际治理大致经历了两个主要阶段。

第一阶段，以利益相关方为主提出治理原则和倡议，构建包容性的国际发展环境。2019年之前，企业、非政府组织、技术社群、标准化组织等非国家主体是

[1] 贾开、赵静、周可迪：“算法全球治理：理论界定、议题框架与改革路径”，《中国行政管理》，2022年第6期。

[2] 陈颖、薛澜：“全球跨境数据流动治理的演进与趋势”，《国际经济合作》，2024年第2期。

人工智能国际治理领域的主要行动者。[1]这些主体从不同视角出发，围绕人工智能治理提出了丰富多样的伦理原则或倡议。其中，较有代表性的包括未来生命研究所提出的“阿西洛马人工智能原则”（2017年）、电气与电子工程师协会（IEEE）提出的“人工智能设计伦理准则”（2017年）、美国计算机协会（USACM）提出的“算法透明度和责任原则”（2017年）等。这些松散的论坛和私人部门的讨论是萌芽状态下人工智能国际治理的主要方式，旨在探索颠覆性技术的治理实践。[2]

当然，这一时期各国政府虽未在国际层面发挥主要作用，但在人工智能技术、产业甚至军事应用方面，加强了国内战略部署。例如，2016年美国颁布两个国家级政策框架——《国家人工智能研究与发展战略规划》《为人工智能的未来做好准备》，日本出台《第五期（2016－2020年度）科学技术基本计划》以促进人工智能发展。与美国和日本强调技术创新和保持国家竞争力不同，英国和法国分别在2016年和2017年发布人工智能相关战略，均关注人工智能的变革性影响，并强调潜在的伦理风险和法律问题。可以看出，无论是利益相关方提出的全球性倡议和原则，还是主权国家颁布的国内战略和计划，都在一定程度上为人工智能发展营造了包容性的国际环境。

第二阶段，主权国家、国际组织和机构参与程度加深，迈向“主权国家战略博弈”格局。一方面，各国政府逐渐参与人工智能国际治理，并以促进技术发展和国际科技合作为重点。尤其是，经过一段时间的技术创新竞赛后，一些国家意识到人工智能技术无法由一国垄断，转而重点开展国际技术合作与研发。以美国为例，其与多个国家在人工智能研发领域开展国际合作。例如，2021年与印度启动人工智能合作，2022年与澳大利亚开展联合资助以共同推进公平和可信赖人工智能研发。在已有跨国合作的基础上，2023年5月美国政府发布的《人工智能研发战略计划》明确提出，“为人工智能研发国际合作确立有原则和可协调的路径”，以多样化、多层次机制推动人工智能国际合作。

另一方面，国际组织和机构开始关注人工智能引发的多重风险，加强国际科技合作，制定标准规范，积极开展治理行动。例如，2019年经济合作与发展组织（OECD）提出人工智能治理首个政府间政策指导方针——《OECD关于人工智能

[1] Green J. F. and Auld G., “Unbundling the Regime Complex: The Effects of Private Authority”, *Transnational Environmental Law*, 6(2): 259–284, 2017.

[2] Morin J. F., Dobson H., Peacock C. et al., “How Informality Can Address Emerging Issues: Making the Most of the G7”, *Global Policy*, 10(2): 267–273, 2019.

的政府间政策指导方针》，旨在确保人工智能的系统设计符合公正、安全、公平和值得信赖的国际标准。2020年，七国集团（G7）发起具有一定排他性的“全球人工智能伙伴关系”（GPAI），以促进发达国家之间以及与具有相近价值观的国家在人工智能方面的国际研究合作，并于2023年共同呼吁制定和采用可信赖的人工智能国际技术标准，设立名为“广岛人工智能进程”的部长级论坛。与发达国家相比，发展中国家及相关国际组织虽然活跃度不高，但2023年以来也呈现出积极参与的态势，例如，金砖国家宣布成立人工智能研究工作组，希望加强信息交流和技术合作，形成具有广泛共识基础的人工智能治理框架和标准规范；非洲联盟发展机构与非洲联盟新兴技术高级别委员会共同发布《非洲人工智能战略》，试图监督和确保非洲国家在人工智能领域向共同目标发展。

联合国及其框架下的国际组织频繁讨论人工智能的国际治理机构与规则问题，并围绕具体议题和机构设立开展了一系列行动。例如，联合国教科文组织在2021年发布了人工智能治理的首个规范性全球框架——《人工智能伦理问题建议书》，并推动成员予以落实。2023年7月，古特雷斯（António Guterres）在联合国安全理事会召开的“人工智能与安全问题高级别公开会议”上呼吁成立类似于国际原子能机构的国际人工智能监管机构。2023年10月，联合国人工智能高级别顾问委员会成立，旨在为人工智能全球治理提供建议，协调和推动各利益相关方的治理行动，并于2023年12月发布报告《以人为本的人工智能治理》，以推动人工智能治理的国际合作。2022年互联网治理论坛提出其成为国际性协同平台的愿景，并于2023年5月成立“人工智能政策网络”，聚焦治理的互操作性、种族和性别包容性、人工智能环境影响三大议题。

2023年以来，随着各国政府、国际组织和机构积极涉足人工智能议题领域以及人工智能行业领袖签署公开信、呼吁警惕人工智能风险，[1]全球学术界和行业界都十分关注人工智能议题。这一系列行动和事件的发生，标志着人工智能国际治理正在进入主权国家战略博弈的关键阶段。例如，美国近期加强了与盟友之间

[1] 2023年5月，上百名专家联名发起警惕人工智能的公开信，建议将人工智能风险等级与流行病、核武器并列。签署人包括OpenAI首席执行官（CEO）萨姆·奥尔特曼（Sam Altman），谷歌DeepMind的CEO戴米斯·哈萨比斯（Demis Hassabis），美国Anthropic的CEO达里奥·阿莫代伊（Dario Amodei）。此外，微软和谷歌的多名高管也在名单之中。2023年3月，出于对伦理和社会责任的担忧，特斯拉CEO埃隆·马斯克（Elon R. Musk）、图灵奖得主书亚·本吉奥（Yoshua Bengio）、苹果联合创始人斯蒂夫·沃兹尼亚克（Stephen G. Wozniak）等一千多名人工智能专家和行业高管联合签署了公开信，呼吁人工智能实验室立即暂停训练比GPT-4更强大的人工智能系统，至少暂停6个月。

的人工智能治理磋商，支持印度成为全球人工智能伙伴关系主席国。中国在“一带一路”国际合作高峰论坛期间发布了《全球人工智能治理倡议》。在英国举行的首届全球人工智能安全峰会上，与会国签署了《布莱切利宣言》，同意通过国际合作形成人工智能监管方法。此外，随着乌克兰危机持续，全球各国对于自动化武器的担忧不断加剧，搭建供军事大国讨论人工智能军事化应用的对话平台显得尤为迫切。各国也增加了人工智能军事运用和国防安全方面的行动。例如，2023年10月30日，美国发布的关于人工智能的行政命令指出，最强大的人工智能系统开发者必须与美国政府分享安全测试结果和相关数据。[1]

（二）当前人工智能国际治理的实践特点

在ChatGPT发布和文生视频大模型Sora问世引起一系列连锁反应后，世界各国以及相关的国际组织、私营部门在人工智能国际治理上表现出前所未有的积极性和紧迫感。以往的新兴技术，往往在发生重大灾害或事故后国际社会才有所行动，而人工智能领域的全球治理可谓行动迅捷。回顾这些治理实践可以发现，当前人工智能国际治理呈现一些实践特色和发展趋势，相关分析将有助于未来的人工智能治理。

第一，软性规则引领治理行动。不同于传统国际技术治理中以硬性规则为主，软性规则逐渐成为人工智能国际治理的主要趋势。传统的硬性规则如国际技术标准、接口协议、合同范本，服务于“控制性”治理需求，如核心技术厂商形成企业联盟来促进标准制定或形成接口协议，以巩固垄断地位、保证国际贸易中商品和服务的质量或性能一致性，从而实现其技术标准推广、技术壁垒构建、专利产业化、产业霸权、市场锁定等目标，进一步提高相关国家的技术领导地位。在人工智能领域，围绕特定议题或产业形成的原则共识、发展指南、倡议协议、最佳实践、标准指南等软性规则或者软法悄然出现，如OECD发布的《关于隐私保护与个人数据跨境流动指南》、联合国教科文组织发布的《人工智能伦理建议书》。这些软性规则通常服务于发展需求，强调透明、公平和问责等原则，[2]适用于难以达成共识的国际谈判以及讨论政治敏感度高的议题等场景，以凝聚共

[1] “Fact Sheet: Biden-Harris Administration Announces Key AI Actions following President Biden’s Landmark Executive Order”, October 4, 2022, https://www.whitehouse.gov/briefing-room/statements-releases/2023/11/27/fact-sheet-president-biden-announces-new-actions-to-strengthen-americas-supply-chains-lower-costs-for-families-and-secure-key-sectors/[2024-04-19].

[2] Erman E. and Furendal M., “The Global Governance of Artificial Intelligence: Some Normative Concerns”, *Moral Philos Politics*, 9(2): 267-291, 2022.

识、开展合作、交换利益。软性规则在人工智能国际治理领域发挥了积极作用。[1]

第二，呈现多元监管思路。虽然各国近期才在人工智能国际治理领域有所行动，但针对人工智能的国内监管实践却已持续一段时间。各国都在尝试创新本国的监管方案与治理经验，全球呈现出多种类型的监管思路和实践举措。总体而言，全球对人工智能监管的思路从原先秉持技术中立的避风港制度，逐渐向穿透式监管、过程监管、互动式监管方向迈进，已超越了互联网技术治理的逻辑。但在具体做法上，各经济体又显现出较大差异。例如，美国针对人工智能较少采用硬性监管，主要采取市场主体自我规制模式，充分发挥企业作用和鼓励市场自发的制度建设，即“从行业实践到行政指引再到立法确认”的美式监管思路。欧盟则以风险治理思路为核心，以保护人的基本权利为出发点，构建了风险分类分级的治理谱系，相关治理行动走在世界前列。[2]2024年3月，欧洲议会正式通过《人工智能法案》，这是欧盟围绕人工智能产业建立保障措施的首次全面立法尝试。作为全球第一个全面的人工智能立法，该法案在一定程度上采纳了各成员国的意见，然而，各成员国对该法案能否有效实施普遍表示担忧。英国虽然提出了更加宽松、自愿的人工智能监管方法，但截至目前还缺乏有效的实施手段。中国构建了负责任发展的人工智能治理体系，提出了敏捷治理这一新颖的原则，创新性使用了“算法备案”“算法检查”等新型治理工具，获得了国际同行的好评。即便部分国家质疑人工智能治理中存在政治表演，[3]但当前各国多元监管实践一定程度上可为人工智能多样化的治理场景提供一些解决方案，也将加速人工智能国际治理进程。

第三，私营部门是主要的治理参与方。区别于传统技术治理中技术开发者主要被监管的情况，人工智能领域的私营部门具有较大的市场影响力和市场权力。过去几年，私营部门虽受制于各国监管政策，但同时也积极参与国内监管和国际治理。例如OpenAI首席执行官萨姆·奥尔特曼在美国国会的听证会上主动呼吁

[1] Wahlgren P.,“How to Regulate AI?” in Colonna L. and Greenstein S. eds., *Nordic Yearbook of Law and Informatics 2020–2021: Law in the Era of Artificial Intelligence*, Stockholm University: The Swedish Law and Informatics Institute, 2022.

[2] Laux J., Wachter S. and Mittelstadt B.,“Trustworthy Artificial Intelligence and the European Union AI Act: On the Conflation of Trustworthiness and Acceptability of Risk”, *Regulation & Governance*, 18(1): 3–32, 2023.

[3] Bareis J. and Katzenbach C.,“Talking AI into Being: The Narratives and Imaginaries of National AI Strategies and Their Performative Politics”, *Science, Technology, & Human Values*, 47(5): 855–881, 2022.

对人工智能加强监管，这一行动打破了以往“猫和老鼠”型的监管模式，而由被监管者主动要求监管并配合监管。面对数据跨境流动的制度障碍，私营部门发起成立的国际数据中介，相较于国家间协议和标准合同条款，更为直接有效。典型代表是由西门子（Siemens）、思爱普（SAP）、源讯（Atos）等22家公司成立的欧洲共同数据空间Gaia-X。此外，数字巨头正在积极优化内部治理，完善技术开发标准与规范，加强数据与算法的可靠性，出台审计流程与问责提案，并将治理理念沿产业链向行业传播。据Statista统计，2015－2020年21家全球数字企业发布了44个有关人工智能伦理的倡议，仅在2020年全球23个新的倡议中就有15个来自私营企业。与以往科技公司坚持包容性自规制不同，随着一些人工智能技术引发的社会风险持续发酵，数字企业对规制人工智能高风险应用形成了共同认知，这无疑有助于全行业接受社会监督和向负责任人工智能方向发展。[1]

第四，科技专家扮演重要角色。计算机、信息技术等学科领域的诸多科学家积极参加各类人工智能峰会和全球性论坛，提供了各自国家的人工智能技术场景、监管新规以及文化理念，成为人工智能国际治理的重要参与力量。[2]类似于气候变化、生物基因领域呈现的知识与技术门槛，人工智能领域的科学家群体在国际治理舞台上占据不可或缺的位置。一方面，鉴于人工智能应用产生较为广泛的影响，专家解读与知识传播在应对风险方面发挥了更为重要的作用。虽然各界存在批评人工智能专家群体过分信赖技术和“技术治理技术”这一思维倾向的声音，但不可否认，各国技术专家在国际公共舞台上发声、解释和讨论，有助于推进对人工智能技术与相关概念的理解，消除社会层面的技术恐慌，并将人工智能国际治理议题拉回关于技术中立的讨论，促使相关讨论具有多样性、代表性和公平性。另一方面，人工智能领域的科学共同体参与治理，也可在一定程度上解决部分国家参与全球治理缺乏人力资源的难题。在人工智能国际治理体系构建初期，既掌握技术原理和产业情况、又熟悉全球治理机制与国际规则的复合型人才相对匮乏。科学共同体不仅掌握技术与行业前沿，而且立场相对中立，因而在一定程度上可以解决技术权力不平衡问题，帮助在人工智能国际治理中处于弱势的部分国家，同时科学共同体还可以承担科学审查和同行评议等职责，以提升人工智能国际治理的科学性和公平性。

[1] de Laat P. B., “Companies Committed to Responsible AI: From Principles towards Implementation and Regulation ?” *Philosophy & Technology*, 34: 1135-1193, 2021.

[2] Ulnicane I., Knight W. and Leach T. et al., “Framing Governance for a Contested Emerging Technology: Insights from AI Policy”, *Policy and Society*, 40(2): 158-177, 2021.

形成具有议题适配性的国际治理机制：一些启示

处于各方博弈阶段的人工智能国际治理，事实上才刚进入国际治理体制形成的初期。然而，在全球治理过程中，最初各方主体如何构建国际治理机制，对该领域的长期发展和治理效果至关重要。在不存在全球政府的情况下，针对纷繁复杂的全球性事务，采用何种国际治理形态，是更加碎片化还是集中化，抑或形成机制复合体，一直是全球治理学界讨论的重要问题。[1]集中化的国际治理体系具有协调力和约束力强、参与便捷等特征，但也存在构建缓慢、路径依赖、脆弱易俘获等缺陷。分散化的国际治理体系虽然可能降低效率和抑制参与，但有利于在应对新兴议题过程中及时调试和促进创新。[2]即便机制复合体被普遍认为是结合了集中与分散特性的中间治理形态，治理体系的复杂化也往往导致混乱和无序。[3]人工智能国际治理体系的构建同样面临上述选择困境。尤其是在当下这一十字路口，任何国际治理机制的失败都将影响国际社会对人工智能治理的信心，甚至会将这一情绪传导至经济领域，对人类社会发展产生深远影响。对此，必须慎之又慎。

在思考如何构建人工智能国际治理机制时，从既有全球治理经验中寻找答案是必要且有价值的。更为重要的是，由于人工智能发展带来的全球性问题以及背后问题的复杂性，亟需一套新的治理方法。[4]当下可能难以找到新的治理方法和治理模式，但可以深入思考构建新的治理机制需要哪些新的思路，这或许具有启发性。

第一，原则先行。当前只有国际社会就如何治理人工智能达成一致，才可能构建行之有效的治理体系与机制。其中，技术治理、军事应用和智能潜力是最重

[1] Biermann F. and Kim R., *Architectures of Earth System Governance. Institutional Complexity and Structural Transformation*, Cambridge: Cambridge University Press, 2020.

[2] Morin J. F., Dobson H., Peacock C. et al., “How Informality Can Address Emerging Issues: Making the Most of the G7”, *Global Policy*, 10(2): 267–273, 2019.

[3] Alter K. J. and Raustiala K., “The Rise of International Regime Complexity”, *Annual Review of Law and Social Science*, 14: 329–349, 2018; Eilstrup-Sangiovanni M. and Westerwinter O., “The Global Governance Complexity Cube: Varieties of Institutional Complexity in Global Governance”, *The Review of International Organizations*, 17(2): 233–262, 2022.

[4] Taeihagh A., “Governance of Artificial Intelligence”, *Policy and Society*, 40(2): 137–157, 2021; Tallberg J., Erman E., Furendal M., Geith J., Klamberg M. and Lundgren M., “The Global Governance of Artificial Intelligence: Next Steps for Empirical and Normative Research”, *International Studies Review*, 25(3):1–25, 2023.

要的关注点。首先，人工智能国际治理要求达成促进发展和规避风险的全球共识。这意味着构建国际治理机制和解决全球性问题的前提是，识别、预防和充分讨论人工智能可能带来的风险和社会影响，同时不阻碍技术进步与创新发展。其次，借鉴管理核武器、化学武器的思路，针对人工智能的军事化应用问题，各国需要尽快构建讨论途径与平台，形成一些原则共识，从而避免各国由产业竞争走向军备竞赛和地缘政治冲突。同时，引导科技向善也可在一定程度上防范和制约全球恐怖主义。最后，要依赖科学界和产业界针对人工智能的“智能潜力”做出预判，秉持审慎的态度，形成预防优先于发展的原则共识，共同打造安全、可控、可信的人工智能。

第二，分类敏捷。人工智能的快速变化和对世界秩序的影响表明，其国际治理体系、机制和形态应当具有适应性、敏捷性、包容性和前瞻性。其中，分类敏捷是重要的治理体系构建途径。[1]一方面，考虑到人工智能作为一般通用技术的广泛性、多样性以及相关国际治理涉及多重议题的属性，任何宽泛的人工智能国际治理方案都可能欠缺落地性，并难以得到利益相关方的支持。只有对治理议题进行分类，针对不同议题属性构建多层次的治理图谱和政策工具箱，才能形成正确的对话基础，找到更为适宜的治理主体和治理机制。另一方面，鉴于人工智能的技术特性，当下无法预测其演变、用途、风险、回报，因而难以较早明确治理对象和内容，相关治理只能在监管互动中前行。在此过程中，应当根据技术发展及其带来的全球性问题调整治理对象和内容，这要求敏捷的治理思路渗透方方面面，包括发展与治理的节奏掌控、治理工具的灵活组合以及治理形态的变化。

第三，中间连接。一个行之有效的全球治理机制复合体能够有效发挥多元主体之间的协作功能。然而，国际组织、行业标准组织、非政府组织以及一些国家政府都提出了人工智能治理倡议或者制定相关规则，这些治理机制的重叠与冲突使得对该议题的讨论和应对方案变得更加复杂。虽然成立国际层面的委员会和评估机构有助于接触私营部门、汇集专业知识、形成专业权威以及促进各方达成共识，然而截至目前，尚未从其他全球治理领域看到良好的效果，甚至有学者认为当前国际体系正在经历条约泛滥危机。[2]在人工智能领域，科研工作者和产业界人士是一个庞大群体，且呈现高度的地域和文化多样性。与人类基因治理、化学

[1] 薛澜、赵静：“走向敏捷治理：新兴产业发展与监管模式探究”，《中国行政管理》，2019年第8期；贾开、赵静：“技术嵌入、价值倾向与算法分类治理”，《经济社会体制比较》，2023年第4期。

[2] Morin J. F., Dobson H., Peacock C. et al., “How Informality Can Address Emerging Issues: Making the Most of the G7”, *Global Policy*, 10(2): 267–273, 2019.

武器等议题不同，人工智能领域的科学家与产业界人士并非界限分明，相较而言更容易形成具有包容性的知识群体，并积极纳入学术期刊、行业联盟、专业组织等力量，从而形成人工智能治理的伙伴关系。作为衔接不同机制与层次的“中间连接体”，这些专业人员可在主权国家、国际组织、非政府组织之间发挥粘合剂作用，从而增强机制复合体的合力。同时需要认识到，主权国家的角色不可或缺，需要构建人工智能强国之间的对话渠道、确保多方主体在明确界定的议题范围内展开交流与沟通、共同维护人类社会的共同利益，避免相关讨论被意识形态渗透。

第四，技术自治。人工智能具有失控风险，同时也可能具备超出人类想象的更多技术能力。对于“技术治理技术”的思路，应当秉持“不排斥且可积极尝试”的立场，充分发挥人工智能赋能治理的作用。目前，人工智能已广泛应用于各类监管实践，扮演着自动化监管者的角色，这在金融、环境、质量监测等领域已较为普遍。谷歌、微软、国际商业机器公司（IBM）、阿里巴巴等跨国企业和数字巨头都开发了新颖的算法审查工具来加强对技术风险的监控，第三方技术安全审查企业也纷纷成立。今后，人工智能技术与应用或许能在构建国际治理机制这一软命题上发挥更大的作用。

当前，推动人工智能的国际治理刻不容缓，构建适宜的全球治理机制和组织体系仍是复杂而艰巨的任务。为了在促进发展的同时有效规避风险，有必要针对安全底线达成共识。例如，超越国家利益或意识形态，以人类命运共同体和人类社会共同利益为风险识别的基准，加强针对高风险技术研发和应用的限制。并且，在发展的维度上坚持可持续公平，包括给予所有国家参与国际治理的机会和表达诉求的渠道，促进发展中国家完善教育和创新体系以实现人工智能技术落地应用。在此基础上，各国加强对话，增进共识，共同推进人工智能国际治理框架的形成。

（责任编辑：邱静）