Oct. 8
2024

全球治理学科动态
2024 年第 5 期（总第 35 期）

人工智能与全球治理

中国社会科学院（CASS）
世界经济与政治研究所（IWEP）

本期执笔

张馨丹 张尊月 张朔宁

中国社会科学院世界经济与政治研究所
全球治理研究团队成果发布
Global Governance Perspectives

专题序言

人工智能是人类发展的新领域，既带来重大机遇，也伴随风险挑战，需要国际社会共同应对。近期，联合国未来峰会通过《全球数字契约》，全球人工智能治理迈出重要一步。中国亦先后提出《全球人工智能治理倡议》、《人工智能全球治理上海宣言》。本期《全球治理学科动态》以“人工智能与全球治理”为主题，探讨了人工智能的治理模式及其可能障碍，分析了国际合作应对人工智能风险的现实路径，并考察了人工智能治理不同模式中的结果和程序合法性问题。这些近期文献揭示了全球人工智能治理的必要性与复杂性，为此治理领域的框架设计和规则制定提供了有益启示。

本期目录

1. Michael Veale, Kira Matus, and Robert Gorwa, “AI and Global Governance: Modalities, Rationales, Tensions,” *Annual Review of Law and Social Science*, Vol. 19, No. 1, 2023.
2. Huw Roberts, Emmie Hine, Mariarosaria Taddeo, and Luciano Floridi, “Global AI Governance: Barriers and Pathways Forward,” *International Affairs*, Vol. 100, No. 3, 2024.
3. Lewis Ho et al., “International Institutions for Advanced AI,” Working Paper, July 2023.
4. Eva Erman and Markus Furendal, “Artificial Intelligence and the Political Legitimacy of Global Governance,” *Political Studies*, Vol. 72, No. 2, 2024.

本期审校

陈兆源、杨嘉豪

No. 1

Michael Veale, Kira Matus, and Robert Gorwa

Annual Review of Law and Social Science

Vol. 19, No. 1, 2023.

“AI and Global Governance: Modalities, Rationales, Tensions”

《人工智能与全球治理：模式、动因与张力》

人工智能（AI）在全球范围内的快速发展，正在深刻改变着社会、经济和政治格局。然而，围绕其治理的讨论仍然充满了不确定性和争议。本文旨在探讨 AI 在全球治理中的作用，分析不同的治理模式及其背后的动因，并解释这些模式所面临的张力，探究治理方式背后的逻辑。

理解全球 AI 治理的一种方式是关注**规则**，作者以 AI 在实际应用中所达到的效果为分类标准。AI 的发展治理包括 AI 系统的研发和维护的政策要求，以及模型发布前的审查（如仇恨言论和恐怖主义内容的检测）。此外，AI 治理还关注部署 AI 的过程和结果在政治、社会和经济方面的影响。AI 基础设施治理则涉及提供 AI 模型基础建设的上游公司产品（如硬件、操作系统、检测技术、云计算和网络建设等）的知识产权政策。

AI 全球治理的主要方式有以下部分：

1）**伦理原则与理事会**：近年来，许多 AI 相关的道德标准文件、理事会及多成员组织由科技公司主导设立。其中一些组织是公司内部的（如微软的 Aether Committee、IBM 的 AI Ethics Board 等），而另一些则由多个科技巨头联合成立（如 Partnership on AI）。这些组织旨在推动行业协作，形

成跨公司的 AI 治理体系，然而，实际效果有限。同时，有人担忧 AI 伦理原则可能成为公司获取监管话语权的工具。例如，谷歌的 AI 伦理委员会被认为试图赢得政治圈的青睐。

2）**行业治理**：虽然行业自治发展不均衡，但它对 AI 的整体发展有着重要影响。许多问题源于少数公司控制着 AI 领域的关键资源，并通过资助研究和技术工具影响 AI 治理进程。行业通过“制造共识”，在学术会议上推动 AI 技术的公平扩展。全球 AI 治理正集中于高利润的通用系统领域，当前挑战在于扩大其积极用途并抑制负面影响。未来，平台或将成为强大的治理者，欧盟正在《人工智能法案》中考虑对通用 AI 供应商施加监管。

3）**合同和执照**：合同条款被用于限制 AI 及其输出的用途。这一方法受开源软件知识产权制度的启发，旨在通过行为使用限制（如 RAIL 许可）来规范 AI 的应用，防止不当使用。然而，执行这些合同条款在全球范围内面临挑战，尤其是涉及全球范围内的用户时。

4）**标准**：自我监管的工程标准不仅确保了技术功能，还涉及隐私、言论自由等价值观。AI 系统的标准化正在逐渐采用类似方法，不仅关注功能性，还考虑法律合规和伦理问题。政府在 AI 标准的制定中采取了混合或协同方式，以促进合规并加强治理。

5）**国际协议**：由跨国组织制定的一系列 AI 原则，如经济合作与发展组织（OECD）和联合国教科文组织（UNESCO）的建议，虽然具有国际影响力，但与行业文件相比并无显著差异。欧洲委员会在推动全球 AI 治理法律框架方面发挥了重要作用，当前正加紧起草 AI 治理公约。然而，这一

过程受到包括美国在内的多方政治压力，且一些关键讨论是秘密进行的。

6）**具有域外效力的国内立法：**一些国家通过国内立法影响全球 AI 治理，例如欧盟的《通用数据保护条例》（GDPR）。此类立法往往具有域外效力，对全球企业产生深远影响。与 AI 密切相关的法规还包括数据保护、知识产权和竞争法等，它们限制了 AI 系统的开发与应用，并产生跨境影响。

文章随后对推动全球 AI 治理的动因进行了阐述。AI 带来了巨大的经济利益，例如提高生产效率、推动创新和优化资源分配等，但同时也对传统行业构成了威胁，特别是在自动化、智能化程度较高的领域，低技能工作者面临被 AI 取代的风险，导致就业市场不稳定。这迫使各国和企业寻找应对方案。此外，AI 依赖海量数据进行训练与优化，这带来了数据隐私问题。未经授权的数据使用以及算法偏见可能导致社会信任下降。同时，AI 技术的高能源消耗也引发了环境问题，电力和计算设备的使用导致了大量的碳排放，对环境造成了巨大压力。因此，AI 的发展不仅局限于某个国家或地区，其影响是跨国界的。通过全球治理机制，制定统一的治理框架和标准，有助于确保技术的安全和公平应用。

编译：张馨丹（北京外国语大学）

No. 2

Huw Roberts, Emmie Hine, Mariarosaria Taddeo, and Luciano Floridi

International Affairs

Vol. 100, No. 3, 2024.

“Global AI Governance: Barriers and Pathways Forward”

《全球人工智能治理：障碍与前进路径》

本文的研究问题主要聚焦于全球 AI 治理所面临的障碍及其改进措施。随着先进 AI 技术的商业化，如 OpenAI 的 ChatGPT，这些技术不仅带来了诸多益处，同时也带来了许多风险。比如，它们可能对国家安全构成威胁，或通过不平等的市场权力分配导致就业机会流失，从而影响社会经济的发展。这些跨越国界的风险引发了全球对更强有力的 AI 治理机制的呼吁。本文的核心问题是，如何在当前地缘政治和制度现实的背景下，推动建立有效的国际 AI 治理机制。

本文认为，尽管 AI 治理的重要性日益凸显，但由于其复杂性以及各国政策优先级的差异，国际合作面临诸多挑战。本文指出，现有的国际 AI 治理框架是一个弱“机制复合体（regime complex）”，由多个相互关联但独立运作的机构组成。为了实现有效的治理，需要加强这些机构之间的协调，并提升其能力，以支持相互促进的政策变革，最终实现跨多个政策领域的催化性变化。

本文首先回顾了当前全球 AI 治理的主要进展。例如，自 2014 年以来，联合国在《特定常规武器公约》框架下，讨论了致命自主武器系统的治理问题。2019 年，经济合作与发展组织（OECD）成员国通过了一套 AI 伦理原则，随后 G20 领导人也承诺遵守这些原则。2021 年，联合国教科文组织（UNESCO）的 193 个成员国通过了《人工智能伦理建议》。随后，本文探讨了国际关系中的一阶和二

阶合作问题如何适用于 AI 治理。一阶合作问题主要涉及国家间的地缘政治竞争与利益冲突，如国家安全和经济利益竞争。二阶合作问题则集中于国际制度的功能失调及各国政策优先级的不一致。例如，各国在实施 AI 伦理原则时的具体做法差异巨大。通过对现有国际制度的分析，本文还探讨了强化现有“机制复合体”的可行性，分析了各机构在制定和推广 AI 治理标准与政策方面的作用及其局限。这些机构包括现有的国际标准组织（如 ISO 和 IEC）、政府间组织（如 OECD 和 UNESCO）以及私营部门的治理机构（如全球 AI 合作伙伴关系和前沿模型论坛）。

研究结果显示，尽管存在显著的地缘政治障碍和制度功能失调，但通过加强现有国际机构之间的协调合作与能力建设，仍有可能逐步推动全球 AI 治理。国际标准组织在涉及伦理和价值观相关的治理问题上作用有限，因为这些组织主要关注技术层面的规范，伦理和社会影响等方面缺乏足够的专业知识与权威。虽然政府间组织在制定伦理原则方面取得了一定进展，但在落实和推广这些原则时面临巨大挑战。各国受到自身政策优先级和地缘政治利益的影响，导致执行效果不一致。私营部门的治理机制依赖于企业自愿参与，缺乏强制性和约束力，导致不同企业对治理标准的接受度和执行力不一，进而使得行业标准难以全面落实。

最后，本文强调，应通过强化现有国际 AI 治理机构之间的协调与合作，提升其治理能力，支持政策的相互促进。本文还建议利用现有专家机构，如 OECD 等组织的技术专长，提供权威信息，支持基于证据的国际合作。此外，本文主张在多个治理层次（国家、次国家、私营部门）上开展合作，以促进一致性政策和标准的发展。这些措施不仅有助于缓解当前的合作难题，还能为未来更为系统的全球 AI 治理奠定基础，从而在技术进步与风险控制之间实现平衡。

编译：张馨丹（北京外国语大学）

No. 3

Lewis Ho et al.

Working Paper, July 2023.

“International Institutions for Advanced AI”

《针对先进人工智能的国际制度》

人工智能的快速发展表明，全球人工智能治理需要国际制度发挥更大作用。强大的人工智能系统在带来诸多益处的同时，也伴随了风险。一方面，人工智能能够显著提高经济生产力，帮助解决重要的社会和技术挑战；另一方面，它也可能导致劳动力流失、透明性缺乏、结果偏见、利益分配不均，甚至威胁国家安全。能否充分发挥人工智能的优势并有效管理其风险，既涉及国内层面的决策，也涉及国际层面的协调。许多风险和挑战跨越国界，影响全球社会与经济。因此，越来越多的政策制定者、技术专家和人工智能治理专家呼吁加强以国际制度为核心的全球人工智能治理。

本文在强调国际人工智能治理必要性的基础上，提出了人工智能治理制度应具备的两大类职能：一是科学技术的研究、开发和推广；二是国际规则的制定与执行。作者进一步提出了四种不同模式的全球人工智能治理制度，并分析了各模式面临的困难。这些制度的参与者包括政府、非政府组织、私营部门和学术界等民间社会的利益相关方。

前沿人工智能委员会（Commission on Frontier AI）：旨在促进专家对先进人工智能带来的机遇与风险达成共识，提升公众对人工智能前景和问题的认识。这有助于科学地说明人工智能的使用和风险，并为政策制定者提供专业意见。该

模式的挑战在于：及时理解未来风险的科学挑战，政治化和成员结构对委员会公正性和合法性的影响。

先进人工智能治理组织（Advanced AI Governance Organization）：负责制定国际治理规范和标准，支持其实施，并监督治理体系的遵守情况。该模式的挑战在于：制定标准的速度和全面性，如何确保参与方的激励机制，以及明确治理范围的难度。

前沿人工智能协作（Frontier AI Collaborative）：作为国际公私合作伙伴，负责开发和推广尖端人工智能技术，帮助落后地区从中受益，并推动人工智能技术的全球普及，以实现安全与治理目标。该模式的挑战在于：如何克服利用人工智能技术惠及弱势群体的障碍，以及防止双重用途技术的传播。

人工智能安全项目（AI Safety Project）：通过汇集杰出的研究人员、工程师和计算人员，并为其提供计算资源和先进的人工智能模型，研究如何从技术层面缓解人工智能的潜在风险，进一步推动人工智能安全研究。该模式的挑战在于：可能将安全研究与前沿开发者隔离开来，以及安全问题和模型访问权限的管理。

本文的主要贡献在于，讨论了国际范围内进行人工智能治理的必要性，并对国际治理机构可能需要的职能进行了初步分类，为全球人工智能治理的初步研究做出了贡献。

编译：张尊月（中国社会科学院大学）

No. 4

Eva Erman and Markus Furendal

Political Studies

Vol. 72, No. 2, 2024.

“Artificial Intelligence and the Political Legitimacy of Global Governance”

《人工智能与全球治理的政治合法性》

当今世界，随着人工智能（AI）技术的快速发展，AI 治理的概念变得愈发重要。然而，现有的 AI 治理定义、研究和实践往往忽视了过程民主的重要性，因此缺乏足够的政治合法性。为此，本文为全球 AI 治理的政治合法性理论化提出一个一般性分析框架，并展示如何应用该框架来评估实际的全球 AI 治理案例。

首先，作者回顾了有关 AI 治理的现有文献，发现多数研究将“AI 善治”定义为“能够规避风险和产生有益结果的机制与结构”。然而，作者认为，这种结果导向的 AI 治理模式存在以下风险：第一，针对 AI 治理的政策与研究会过于依赖先前关于 AI 伦理的假设；第二，如果 AI 治理过于依赖 AI 伦理所确定的问题清单，其可能忽视不追求狭义目标的组织和机构；第三，结果导向的 AI 治理忽视了程序合法性对于 AI 善治的重要意义。因此，作者主张，具有合法性的 AI 治理必须同时兼备可取的程序和结果。作者假定，

1）具有合法性的 AI 治理必须在最小意义上是民主的；

2）政治合法性是良好治理最重要的程序属性之一。

AI 治理包含“通过 AI 的治理”（governance by AI）和“对 AI 的治理”

(governance of AI) 两种形式，前者指的是将 AI 技术作为现有治理机制的一部分，而后者则是指监管 AI 发展的治理机制本身。由于前者需要后者的权力下放，“通过 AI 的治理”若要满足民主标准就必须通过民主过程得到授权。基于此，作者进一步区分了 AI 治理中的授权实体（authorized entity）和托管实体（mandated entities），并由此建立了合法性链条：公众赋予授权实体为 AI 发展制定目标的权力，授权实体则将部分权力下放给托管实体，并对其负责，同时通过撤销权保证监管透明度。

本文还讨论了五种具有合法性的全球 AI 治理模式，

1）授权实体对 AI 治理制定普通法，例如欧盟的通用数据保护条例（GDPR）；

2）授权实体对 AI 治理制定特殊法，例如欧盟正在讨论的“AI 法”；

3）托管实体可以对 AI 治理制定具备适用性的一般政策，例如联合国安理会；

4）托管实体可以对 AI 治理制定特殊政策，例如北约在 2021 年推出的“AI 战略”；

5）托管实体可以利用 AI 科技促进其它领域的治理，如利用 AI 软件跟踪、分析犯罪率。

作者强调，授权实体自身不能合法地参与“利用 AI 的治理”而必须下放权力至托管实体，因为 AI 系统本身不能治理人类。此外，作者承认很多合法的现存 AI 治理实体并不符合上述五种框架，因此本文提出的分析框架并不完备。

作者指出，将民主决策外包给 AI 系统或 AI 领域私人行为体试图制定软性法律的治理都不具备政治合法性。本文对 AI 治理的贡献有二：首先，填补了当前有关 AI 治理的规范性研究中针对程序合法性关注不足的空缺；其次，提出了 AI 治理合法性标准的一般性分析框架，为区分现有 AI 治理种类发挥积极作用。

编译：张朔宁（达特茅斯学院）

全球治理研究团队

任　琳　熊爱宗　韩　冰　吴国鼎
陈兆源　黄宇韬　李　冲　韩永辉
宋　锦　田　旭　沈　陈　彭　博

研究助理团队

兰馨彤　张尊月　苏山岳　孟思宇
杨嘉豪　邢珏璠

中国社会科学院世界经济与政治研究所
全球治理研究团队成果发布
Global Governance Perspectives