专题文论·人工智能与数字治理

人工智能与全球治理：模式、理据和紧张关系

［英］迈克尔·维尔 基拉·马图斯
［德］罗伯特·戈尔瓦 张涛 译

【编者按】近年来，人工智能成为一个引人注目但又充满争议的议题。全球的行动者都在参与构建围绕人工智能的治理机制。但被治理的“对象”究竟是什么，如何治理，由谁治理，以及为什么治理等问题尚不明确。《俄罗斯学刊》编辑部选译该文，介绍国外学者在参考人工智能、计算治理以及更广泛的监管和治理方面文献的基础上对这些问题的阐释。原文刊登于《法律与社会科学年刊》（*Annual Review of Law and Social Science*）2023 年第 19 卷，经作者授权在《俄罗斯学刊》以中文发表。

【中图分类号】F490.2 【文献标识码】A
【文章编号】2095-1094(2024)03-0049-0024
【关键词】人工智能监管 算法监管 全球治理

【作者简介】迈克尔·维尔（Michael Veale），英国伦敦大学学院法学院副教授；基拉·马图斯（Kira Matus），中国香港科技大学公共政策学部教授（国籍不详）；罗伯特·戈尔瓦（Robert Gorwa）德国柏林社会科学研究中心研究员。
【译者简介】张涛，黑龙江大学信息管理学院教授。

【基金项目】国家社会科学基金一般项目“数智环境下情报分析算法风险治理路径研究”（项目批准号：22BTQ064）阶段性成果。

一、引言

人工智能（AI）议题引起全球高度关注。一些人认为人工智能是经济发展的希望，另一些人则视其为一种商业威胁，还有一些人认为它是一些突出的社会和环境问题的起因或催化剂。上述议题引发了有关全球治理的呼吁，而全球治理本身已成为一个颇具争议的话题。批评者认为，全球治理被工业界把持或完全不起作用，而支持者和密切参与者则认为自己正在建立引导未来的机制。本文列举并审视了各种全球治理倡议及不同的人工智能框架，将它们视为监管竞争的关键点。

本文首先确定了人工智能和全球治理的含义。然后，我们对新兴的人工智能全球治理的不同模式进行了批判性评估，如伦理委员会、行业治理、合同和许可、标准、国际协议以及具有外部影响的国内立法。最后，我们评估了支撑这些模式的特定理据和紧张关系，并关注到推动这些不同模式的利益和观念。

（一）我们在谈论人工智能哪些方面的内容?

人工智能已成为一个宽泛的、包罗万象的术语，越来越多地被炒作、误导和混淆所困扰。近年来，特别是在有关治理的讨论中，人工智能作为一个新术语，指的是以前被称为数据挖掘、大数据或机器学习的技术，而在某些人看来，其几乎等同于任何现代软件。这让那些寻求人工智能具体定义或强调模式识别和统计之外方法（如符号推理）的人感到困扰。本文关注的不是人工智能是什么（或应该是什么），而是当我们谈论人工智能的全球治理时，需要治理哪些技术和实践。

出于上述目的，人工智能一词通常被理解为一种实践，一门“应用科学和工程学科”[①]，旨在将人类认为智能的品质赋予特定的软件。因此，对于作为一种实践的人工智能治理，我们也必须从更广泛的视角来看待。它应该涵盖这一实践所使用的工具和流程，包括其可用性以及社会和物质影响。我们更愿意谈论人工智能模型、开发工具包、框架、数据集或其他人工智能产品，而不是谈论如何对“人工智能”进行治理，因为后者可能包含误导性的含义。这与即使是人工智能技术从业人员也无法就“算法”作为人工产物的定位达成一致的情况十分类似，

① Bryson，J.J. *The Artificial Intelligence of the Ethics of Artificial Intelligence: An Introductory Overview for Law and Regulation*.Oxford，UK: Oxford Univ. Press，2020，pp.2-25.

更不用说那些其工作对这些技术的开发和指导非常重要且作用更广泛的学科了[①]。人工智能还包括人工智能从业者及其所在的组织、人工智能技术的使用目的，以及围绕其使用的社会、经济和政治结构。它还应包括人工智能产物本身的特点，这些特点在不同背景下的表现形式，以及如何影响与之接触的人、环境和机构。本文专注于现有和明显新兴的、而非假设的人工智能技术。学术界、工业界以及慈善机构支持的某些智库对于人工通用智能或生存风险的高度推测性讨论高度关注，但这些内容不在本文的讨论范围内。

我们所讨论的某些方面的全球治理是人工智能特有的，而有些方面则是人工智能所不具备的。这与围绕计算治理或社会分类和划分的更久远问题重叠[②]。我们专注于将这些相关领域的讨论集中在人工智能上，同时强调人工智能全球治理的一些最重要方面可能确实是计算技术全球治理（或缺乏计算技术全球治理），人工智能加速了这一进程，并使其比以往任何时候都更加突出。

（二）什么是人工智能治理？

人工智能治理以及人工智能全球治理意味着什么还远未明朗，我们首先概述具有明显全球相关性的各种人工智能治理模式，包括公共、私人、非正式、正式和混合政策工具，从行业标准到涉及多个领域的国际制定的原则和法律。然后，我们会考虑他们不得不应对的一些紧张关系，这些紧张关系在未来可能会更加明显和重要。

理解人工智能全球治理的一种方式是着眼于监管的概念目标。我们可以根据它们试图塑造的人工智能实践的各个方面来区分规则：人工智能开发、使用和基础设施。

人工智能开发的治理包括在系统设计和维护过程中尝试应用各种要求，以实现一系列政策目标。这些目标可能包括更广泛的概念，如安全性或网络安全，具体的统计目标，如狭义的非歧视定义等。这些要求也可能试图提供透明度和监督机制，例如在 AI 系统部署前要求对其进行审计，或在其销售前要求在数据库中列出。这些开发要求可能是国家层面的而非全球性层面的，但就其国际市场的活

① Seaver，N. "Algorithms as Culture: Some Tactics for the Ethnography of Algorithmic Systems." November 9，2017.

② Bowker，G.C.，Star，S.L. *Sorting Things Out: Classification and Its Consequences*. Cambridge，MA: MIT Press，1999.

动而言，可能产生更广泛的影响[①]。将政策目标设计到技术中，就会引出这些目标从何而来的问题。例如，在国际上销售的招聘系统被发现嵌入了美国的非歧视规范，如五分之四规则，这可能无法解决其他司法管辖区的问题[②]。旨在检测仇恨言论或恐怖主义内容的系统必须处理这些在法律上具有国家特定定义的内容。此外，一些与人工智能相关的问题本质上是全球性的，例如，当具有特定功能（如文本或图像生成）的模型在国际范围内发布时，或者当训练过程涉及跨国重要碳排放以及提取材料及劳动力（包括数据标注的心理影响）的供应链时，这些都会不同程度地影响到价值链中的某些司法管辖区[③]。

人工智能使用的治理涉及部署具有政治、社会和经济后果软件的手段和目的。在实践中，这一类别可以涵盖广泛的范围：国际条约中关于自动决策的法律，如《个人数据自动化处理中的个人保护公约》（也称《108 号公约》）；涉及人工智能在医疗设备、警务和情报等国内部门中使用的跨国制度；规范通用人工智能固有的跨国使用制度，如内容审查、国际冲突、人道主义和国际警务。

人工智能基础设施的治理涉及试图超越上述开发与使用二分法的政策努力。这一视角促使人们关注苹果、谷歌、亚马逊和微软等纵向一体化公司对跨国技术堆栈的影响，这些公司销售硬件、操作系统、传感技术、云计算、网络基础设施，甚至专门训练的模型，所有这些都可能成为某些类型人工智能系统开发和使用的前提。这一领域尚不成熟，但正日益引起竞争监管机构的关注，有关数字主权的问题也正以各种形式进入国际谈判。人工智能系统还可被视为组织内部决策和控制的重塑点，这将全球政治经济问题与国内组织经历的实际选择、约束和影响联系起来[④]。

本文从全球治理的突出模式的简要概述开始，这些模式涉及上述一个或多个领域。在随后的部分中，笔者对这些模式凸显的紧张关系进行批判性讨论，从理论目标和实际应用两个方面来看，笔者认为人工智能的全球治理仍然是一个颇具

① Newman，A. "Watching the Watchers: Transgovernmental Implementation of Data Privacy Policy in Europe." *J. Comp. Policy Anal*，2011，no.2.

② Sánchez-Monedero， J.， Dencik， L.， Edwards， L. "What Does It Mean to 'Solve' the Problem of Discrimination in Hiring? Social， Technical and Legal Perspectives from the UK on Automated Hiring Systems." In *Proceedings of the 2020 Conference on Fairness，Accountability，and Transparency*，2020.

③ Matus，K.J.M.，Veale，M. "Certification Systems for Machine Learning: Lessons from Sustainability." *Regul. Gov*，2022，no.1.

④ Balayn，A.，Gürses，S. "Beyond Debiasing: Regulating AI and Its Inequalities." 2021. https://perma.cc/4UAV-3UFB

争议的概念，目前主要由行业利益驱动，因此值得仔细拆解和审查。最后笔者提出了一些政策制定者和研究人员在参与这些工作时应该考虑的几点问题。

二、人工智能的全球治理模式

任何文章想要列出一份当前详细、全面的人工智能领域全球治理倡议的清单都是徒劳的。制度环境瞬息万变，几乎每天都有新的倡议出现，以至于这样的清单还没开始制定就已经过时了。尽管如此，笔者还是可以确定并举例说明一些广泛的、理想类型的类别，以显示它们在多个问题领域的发展情况。正如笔者所展示的，这些治理模式都具有高度的政治性，并且越来越成为行业、各国政府、国际组织和公民社会之间日益重要的政治竞争场所。笔者（大致）按照制度复杂性递增顺序来组织这些内容。

（一）伦理准则和委员会

近年来涌现了大量人工智能伦理文件、人工智能伦理委员会和多方利益相关者机构。其中许多都是由大型科技公司或与之有密切联系的组织建立的，有些是公司内部机构，其产品的影响力赋予其全球重要性，如微软公司的人工智能道德（Aether）委员会、负责任的人工智能办公室和人工智能工程战略，国际商业机器公司（IBM）的人工智能伦理委员会，以及谷歌昙花一现的外部伦理委员会，即先进技术外部咨询委员会（ATEAC）。其他倡议由技术公司资助，但至少名义上是外部的。其中最典型的例子或许是亚马逊、苹果、谷歌、脸书（Facebook）、IBM 和微软于 2016 年成立的人工智能合作伙伴关系（PAI）。其他由行业资助的论坛则处于外围，如世界经济论坛或由财政支持的研究实体，如未来生命研究所或人类未来研究所。

这些实体通常在名义上寻求协调开发人工智能系统的行业参与者之间的行动，并旨在公司内部和公司之间塑造未来负责任的人工智能治理议程。然而，在实践中，它们的努力喜忧参半，更严重的是深陷争议。例如，PAI 最初吸引了一系列民间社会成员，但自成立以来，人们对其影响力的热情已经逐渐减弱。著名的非政府组织（NGO）“立即访问”（Access Now）于 2020 年退出，其在退出时表示，“PAI 没有影响或改变成员公司的态度，也没有鼓励它们系统地回应民间社

会或接受其咨询”[①]。行业伦理委员会制定的原则因表达含糊且实际上毫无意义而受到批评——没有执行力度或展示合规的机制。一些人认为，如果这些自愿性努力确实有效的话，那么其在处理商业模式和公司文化的核心伦理问题时应是游刃有余的[②]，或者侧重于工程师和设计选择，并有效地被归入企业逻辑和激励体系[③]。一项对数十个人工智能伦理框架的审查指出，这些伦理原则在如何解释伦理原则，重点涉及哪些问题、领域或行动者，以及它们应如何实施方面存在“实质性分歧”，同时在伦理框架方面，全球多数国家没有做出努力[④]。

最令人担忧的是，有观点认为，人工智能伦理原则主要被企业用于阻止监管行动[⑤]，其中“伦理”成为新的“行业自律”[⑥]，并且可能为面向公众的公关活动提供助力。例如，昙花一现的谷歌人工智能伦理委员会的成员似乎受到了更广泛政策策略的影响，该策略旨在赢得美国政界的支持，并回应共和党内对谷歌等公司偏袒民主党价值观和观点的担忧。对于那些认为伦理哲学作为一个领域对人工智能系统讨论有很多有益贡献的人来说，针对工具化伦理的强烈反对是令人遗憾的[⑦]。

（二）行业治理

当前，伦理准则以外的行业自律参差不齐，但仍然具有影响力。这在很大程度上取决于企业在人工智能领域设置的瓶颈。少数企业控制着促进人工智能系统的培训资源，包括物理资源（如连接廉价电力的大型图形处理单元集群）、认知资源（接触尖端研究人员群体的机会、向大学教授支付高薪的能力）和信息资源（访问数据集、支付大量标签费用的资源以及用于收集和实验的实时系统）。行业参与者赞助了大量关于技术工具的研究，尤其是通过高知名度的学术会议，一些

① "Access Now Resigns from the Partnership on AI." *Press Rel.*, October 13, 2020. https://www.accessnow.org/access-now-resignation-partnership-on-ai/

② Munn, L. " The Uselessness of AI Ethics. "2022. https://doi.org/10.1007/s43681-022-00209-w

③ Green, B. "The Contestation of Tech Ethics: A Sociotechnical Approach to Technology Ethics in Practice." *J. Soc. Comput.*, 2021, no.3.

④ Jobin, A., Ienca, M., Vayena, E. "The Global Landscape of AI Ethics Guidelines." *Nat. Mach. Intell.*, 2019, no.9.

⑤ Nemitz, P. "Constitutional Democracy and Technology in the Age of Artificial Intelligence."*Philos. Trans. R. Soc.* A. 2018, no.2133.

⑥ Wagner, B. *Ethics as An Escape from Regulation: From "Ethics-Washing" to Ethics-Shopping? In Being Profiled: Cogitas Ergo Sum.*, Amst. Univ. Press, 2018, pp. 84-88.

⑦ Bietti, E. "From Ethics Washing to Ethics Bashing: A Moral Philosophy View on Tech Ethics." *J. Soc. Comput.*, 2021, no.3.

学者将其描述为围绕“公平”等政治概念“制造共识”[①]以及推广浅薄的、非语境化的、以工程师和实验室为中心的方法，这些方法可以顺利而低成本地推广[②]。这种行业治理可被视为一种跨国政策创业形式，包括“专家话语的管理和沟通，而不是数据、证据或研究成果”[③]。

全球人工智能治理迅速趋向的一个具体方向是利润丰厚且公众关注的通用系统领域。为通用分析形式设计的人工智能系统是典型的双用途技术，既支持相对良性的目的，如生成库存图片或模板文本，也支持有争议的目的，如犯罪活动或破坏民主。一个重要的问题出现了，即如何在促进良好目的的同时限制或防止不希望出现的目的？

市场上许多性能最好的人工智能系统都是通过基于云计算的新型平台式商业模式以服务形式销售的[④]。这些模式在文本、图像分析或生成等领域被作为通用的基础能力进行销售，用户可以根据具体应用进行集成和定制。尽管用户经常带来自己的数据集来微调模型以适应他们的使用案例，但他们很少能够完全复制最终模型，通常只被允许通过应用程序编程接口将它们作为问答系统使用。

这种人工智能能力的分布形式使平台有能力充当重要的治理决策者。只要理想的模型是专有的，它们的使用就可以被限定在某些用途上。例如，谷歌仅允许其媒体行业的白名单客户使用其名人识别面部分类系统[⑤]。一些评论家表示担心，如果模型是开源的，治理就会变得困难或不可能，例如生成的文本没有隐写水印，无法辨别其是否为人造文本[⑥]。此外，为支持没有明显危害的产出（有明显危害的产出，如生成儿童性虐待图像和叙述）所需的标签资金，即使是单个模型也需要数十万美元，使用的外包劳动力每小时工资不到 2 美元，并提供有限的心

① Young，M.，Katell，M.，Krafft，P.M. "Confronting Power and Corporate Capture at the FAccT Conference." In *2022 ACM Conference on Fairness，Accountability，and Transparency*，2022.

② Gansky，B.，McDonald，S. "CounterFAccTual: How FAccT Undermines Its Organizing Principles." In *2022 ACM Conference on Fairness，Accountability，and Transparency*，2022.

③ Stone，D. "Transnational Policy Entrepreneurs and the Cultivation of Influence: Individuals，Organizations and Their Networks." *Globalizations*，2019，no.7.

④ Cobbe，J.，Singh，J. "Artificial Intelligence as a Service: Legal Responsibilities，Liabilities，And Policy Challenges." *Comput. Law Secur. Rev.*，2021.

⑤ BSR. "Google Celebrity Recognition API Human Rights Assessment." 2019. https://www.bsr.org/reports/BSR-Google-CR-API-HRIA-Executive-Summary.pdf

⑥ Aaronson，S. "My AI Safety Lecture for UT Effective Altruism." November 28，2022. https://scottaaronson.blog/?p=6823

理支持或咨询[①]。

目前尚不清楚平台在未来提供人工智能方面有多么不可或缺。人工智能模型的多边平台正在兴起，例如人工智能公司（Hugging Face），它为分发和使用他人制造的系统提供了基础设施。在当前的数字环境中，选定的实体是极其强大的监管者。应用商店是在许多移动设备上安装某些类型软件的唯一途径，它们通过技术限制和合同限制来规范软件的内容和数据使用等问题[②]。这些规则的一致性和质量一直备受关注，如优图（YouTube）按级别对其用户进行不同的监管[③]，而 Facebook 曾多次违反苹果公司的规则，其行为很可能会导致不太知名的软件被完全禁止[④]。

未来，人工智能全球治理似乎有可能与平台治理高度融合。如果重要的中介机构仍然存在，它们将既是强大的管理者，也是立法者监管人工智能系统的目标[⑤]。这将反映出此前信息技术法律领域的无数经验，即从社交媒体到加密货币等领域的中介机构成为“监管切入点”[⑥]或“瓶颈点”[⑦]。目前，欧盟部长理事会已经考虑在人工智能法案草案中针对通用人工智能系统的提供商规定一些义务，这进一步表明，他们可以通过真正地禁止特定用途，并在检测到“市场滥用”时采取行动来避免某些法律义务[⑧]。

（三）合同与许可

另一种新兴的、重要的、私有的人工智能系统跨国治理形式是使用合同条款

① Perrigo，B. "OpenAI Used Kenyan Workers on Less Than $2 Per Hour to Make Chatgpt Less Toxic." January 18，2023. https://time.com/6247678/openai-chatgpt-kenya-workers/

② Cowls，J.，Morley，J. *App Store Governance: The Implications and Limitations of Duopolistic Dominance.* In the 2021 Yearbook of the Digital Ethics Lab，2022，pp. 75-92；Marsden，C.T.，Brown，I. "App Stores，Antitrust and Their Links to Net Neutrality: A Review of the European Policy and Academic Debate Leading to the EU Digital Markets Act." *Internet Policy Rev.*，2023，no.1；Van Hoboken，J.，Fathaigh，R.Ó. "Smartphone Platforms as Privacy Regulators." *Comput. Law Secur. Rev.*，2021.

③ Caplan，R.，Gillespie，T. "Tiered Governance and Demonetization: The Shifting Terms of Labor and Compensation in the Platform Economy." *Soc. Media Soc.*，2020，no.2.

④ Carman，A. "What Would Happen if Apple Fully Banned Facebook from the App Store?" 2019. https://www.theverge.com/2019/2/1/18205291/apple-facebook-developer-bancertificate-app-store

⑤ Cobbe，J.，Veale，M.，Singh，J. "Understanding Accountability in Algorithmic Supply Chains. " In *Proceedings of the 2023 ACM Conference on Fairness，Accountability，and Transparency*，2023.

⑥ Finck，M. *Blockchain Regulation and Governance in Europe*. Cambridge，UK: Cambridge Univ. Press，2018.

⑦ Goldsmith，J.L.，Wu，T. *Who Controls the Internet? Illusions of a Borderless World*. New York: Oxford Univ. Press，2006；Tusikov，N. *Chokepoints: Global Private Regulation on the Internet*. Berkeley: Univ. Calif. Press，2016.

⑧ Counc. Eur. Union. "Proposal for a Regulation of the European Parliament and of the Council Laying Down Harmonised Rules on Artificial Intelligence (Artificial Intelligence Act) and Amending Certain Union Legislative Acts (Preparation for COREPER). "2022. https://perma.cc/564K-RCKR

来限制人工智能的使用及其产出。这种机制似乎受到了计算机治理史上重要创新的启发，即围绕开源软件出现的知识产权制度。

针对许多人工智能开发的专有的、由企业控制的本质，以及至少在一定程度上由于对数字主权和国家对某些重要人工智能系统和功能的依赖与日俱增的担忧，近年来人们对创建开源通用人工智能系统的兴趣有所增长。为了促进开源人工智能系统的查询、分发和开发，出现了一些平台，如 Hugging Face 等。然而，这些系统的易于访问性带来了一些政治问题，因为很少有制衡措施来防止它们被用于有争议的任务。对此，一些模型设计者转向合同法，保留模型的知识产权（即不将其发布到公共领域），而只提供有条件的使用许可[①]。长期以来，此类许可一直是软件和数字知识产权（IP）治理的核心部分，著作权许可系列，如通用性公开许可证（GPL），要求任何基于许可代码的衍生作品都必须按照相同或等效的许可条款进行发布，而知识共享许可则允许在一系列条件下重复使用内容，如署名、不篡改或用于非商业目的[②]。

在人工智能领域，负责任的人工智能许可（RAIL）倡议声称要超越这些倡议，增加“行为使用限制”。知识产权持有者将这种许可应用于稳定扩散模型，该模型被试图制作成类似于美国人工智能研究公司（OpenAI）专有的 DALL-E 2 图像生成模型。该许可证禁止用户使用模型对他人进行诽谤、提供医疗建议、用于执法或类似目的、用于意图歧视或具有广义的歧视效果的目的、用于某些完全自动化的决策（这是基于数据保护法的规定），以及以可能引起身体或心理伤害的方式剥削个人——这些规则似乎是直接取自欧洲委员会的《人工智能法案》草案[③]。尽管这些合同条款并不完美，而且其中一些条款措辞含糊不清（例如，从法律中提取相关定义部分，或未明确规定诽谤或一般非法行为的管辖权），但有效落实这些条款的最大障碍在于只有版权持有者才能执行版权许可。尽管软件公司可能相当了解该领域中数量有限的竞争软件公司，并能够收集足够的信息进行有效执行，但如果问题是围绕技术的使用而不是技术的开发或所谓的知识产权盗

① Contractor，D.，McDuff，D.，Haines，J.K.，Lee，J.，Hines，C.，et al. "Behavioral Use Licensing for Responsible AI." In *2022 ACM Conference on Fairness，Accountability，and Transparency*，2022.

② Guadamuz，A. "Viral Contracts or Unenforceable Documents? Contractual Validity of Copyleft Licenses. " *Eur. Intellect. Prop. Rev.*，2004，no.8.

③ Rombach，R.，Esser，P. "CreativeML Open RAIL-M. Hugging Face. "2022. https://huggingface. co/spaces/CompVis/stable-diffusion-license

窃，那么成功防止社会滥用似乎不太可能。与此相关的是，以这种方式授权开发人工智能系统的开源或公益团队很可能没有能力发现和监管通用人工智能系统的大量用户，也没有必要的法律资源来实现大规模的治理。我们或许可以设想另一个制度层面的执行机制——许多开源软件版权的基金会所创造的“以社区为导向”的执行原则[①]，但据我们所知，目前还没有此类提议出现。

一种在法律上更加微妙的尝试可见于一些平台试图对模型的输出而非模型本身进行授权，以此来管理其全球使用。OpenAI 是一家以生成新颖文本和图像模型而闻名的公司，它实施了一项内容政策，要求必须将生成的媒体内容披露为人造的，并且禁止某些主题，例如描述“非法活动”、涉及政治人物或宣扬“重大阴谋论”的主题。如果违反上述准则，那么声称在法律允许范围内拥有生成媒体内容所有权的 OpenAI 将保留撤销向用户提供许可的权利。然而，基于知识产权的有效执法程度可能受到两个因素的限制，即司法管辖权和制作系统用于生成媒体的提示所需的有限创造性技能或劳动力[②]。精心设计的提示可能会提供一些保护，但这似乎是一种有限的治理模式，因为提示的复杂性与潜在危害之间并没有可靠的联系。

（四）标准

计算机的全球治理历来在很大程度上依赖于自我监管组织创建的工程标准[③]。这些标准对于网络技术一直至关重要，因为网络技术要求各组件遵循相同的规则来实现功能。互联网工程任务组、万维网联盟和电气与电子工程师协会（IEEE）等机构管理着重要标准，如 TCP/IP、HTML 和 802.11（WiFi）等。然而，这些组织的产出具有政治维度，不仅仅促进功能性，还产生了与人权相关的、富于价值的设计理念和成果，例如与隐私和自由表达有关的理论和成果[④]。统一网络标准

① Ballhausen，M. "Copyright Enforcement. " In *Open Source Law，Policy and Practice*，2022.

② Guadamuz，A. "Do Androids Dream of Electric Copyright? Comparative Analysis of Originality in Artificial Intelligence Generated Works. "*Intellect.Prop. Q.*，2017，no.2；Guadamuz，A. "DALL•E Goes Commercial，But What About Copyright? "July 25， 2022. https://www. technollama. co. uk/dall% c2%b7e-goes-commercial-but-what-about-copyright

③ Harcourt，A.，Christou，G.，Simpson，S. "Global Standard-Setting in Internet Governance. "In *Global Standard Setting in Internet Governance*，2020.

④ Braman，S. "The Framing Years: Policy Fundamentals in the Internet Design Process，1969 - 1979." *Inf. Soc.*，2011，no.5；Cath，C. "The Technology We Choose to Create: Human Rights Advocacy in the Internet Engineering Task Force. " *Telecommun. Policy*，2021，no.6.

在国际上的普及使其发挥了重要的全球治理作用[①]，实质性的政策决定和价值观可以有效地利用其功能的必要性并在技术层面上得到体现。

标准流程可以增强资源丰富的现有企业的能力。在网络领域，大型企业采用的标准可能会迫使其他参与者也做出相同的变更[②]。参与任何标准制定，机构都需要投入大量人力资源，而那些拥有相关经验的人更有能力对其施加影响。成本会进一步限制问责制。许多与计算相关的标准都是通过专有标准机构制定的，在这些机构的标准制定过程中，其仅对付费成员开放，最终产品也是专有的，通常需要花费数百美元才能获得一整套相互关联的规则。

尽管如此，网络标准的成功使一些组织自愿寻求以同样的方法来治理人工智能系统。最早出现的标准之一是 IEEE P70xx 系列，这些标准包括已发布的透明度标准（7001-2021）、在设计中考虑道德问题的流程标准（7000-2021），以及关于偏见和“伦理驱动的推动”标准（7003TM，7008TM）。国际标准化组织（ISO）也有一系列标准，其中大部分正通过 2017 年成立的 ISO/IEC JTC 1/SC 42 委员会制定中。这些组织通常采用订阅模式，保留所发布标准的版权，并通过直接授权访问或经典型的私营国家标准机构，例如美国国家标准协会（ANSI）、英国标准协会（BSI）或德国标准化协会（DIN）授权本地化和翻译许可来赚钱。与大多数标准机构不同的是，国家测量机构通常是公共实体，它们也一直致力于人工智能标准化工作，其中包括美国国家标准与技术研究院，它建立了一个人工智能风险管理框架[③]，以及英国国家物理实验室，它是英国人工智能标准中心的合作伙伴。

并非所有标准的功能都是所必需的，有些标准可能只是向能力较低的组织传递知识，使其了解如何按照当前的最佳做法构建系统，类似于简单的知识输出，而不是协调机制。另一个重要角色是传递信号，计算标准可以向其他市场参与者传递最佳实践信号，例如向保险精算师发出有关保险目的的信号[④]。当立法者考

① DeNardis，L. *The Global War for Internet Governance*. New Haven，CT: Yale Univ. Press，2014.

② Cohen，J. E. *Between Truth and Power: The Legal Constructions of Informational Capitalism*. Oxford，UK: Oxford Univ. Press，2019；Ten Oever，N. " 'This Is Not How We Imagined It': Technological Affordances，Economic Drivers，and the Internet Architecture Imaginary." *New Media Soc.*，2021，no.2.

③ NIST (Natl. Inst. Stand. Technol.). *AI Risk Management Framework (AI RMF 1.0)*. Rep.，2023.

④ Shackelford，S.J.，Proia，A.A.，Martell，B.，Craig，A.N. "Toward a Global Cybersecurity Standard of Care: Exploring the Implications of the 2014 NIST Cybersecurity Framework on Shaping Reasonable National and International Cybersecurity Practices." *Tex. Int. Law J.*，2015，no.2.

虑是否需要更严格的监管时，它们可以被用作一种立法者的信号[①]；或当法院在评估侵权过失时，它们也可以作为一种信号[②]。

在人工智能标准制定方面，各国政府似乎主要采取了两类方法，按照标准和认证文献的分类，可分为混合型和共生型[③]。混合型方法旨在激励私营标准，以允许遵守欧盟《人工智能法案》中提出的法律文书的规定（下文将进一步讨论）。共生型方法认为私营认证体系可以加强其他治理系统的权威性和合法性，欧盟数据保护和网络安全法中的可选行业认证机制就是一个例子[④]。随着标准化激励措施超越功能性和网络一致性，人工智能标准的监管轨迹似乎更类似于可持续产品或产品安全等非计算标准化领域，而不是典型的网络标准。

（五）国际协议

在国际行业自律的同时，围绕人工智能问题出现了一系列政府间标准和组织，其中包括经济合作与发展组织（OECD）的《人工智能建议书》（2019）、联合国教科文组织的《人工智能伦理问题建议书》（2021）以及二十国集团的《人工智能原则》（2019）。在大多数情况下，这些标准和原则可以说与行业文件或行业利益并无明显相悖。例如，原则综述中并未提及数字竞争、权力或对技术的控制[⑤]。尽管数字权力在全球政策中具有重要意义，但它似乎是技术公司不愿意讨论的问题。一系列国家已经建立了一个人工智能全球伙伴关系（关于这一点，不要与上面讨论的由私营部门领导的 PAI 相混淆），以加强在经济合作与发展组织基础上的工作，尽管目前其影响尚不明确。

欧洲委员会（CoE）在人工智能治理方面发挥了主导作用，因为它审查了过去的数字和数据相关法律文书。《欧洲人权公约》是欧洲人权法院以及许多国际监督和分析法律体系的基础，它还牵头制定了《布达佩斯网络犯罪公约》《个人数据自动化处理中的个人保护公约》，后者是国际数据保护标准的基础，以及

① Marsden，C.T. *Internet Co-Regulation: European Law，Regulatory Governance and Legitimacy in Cyberspace*. Cambridge，UK: Cambridge Univ. Press，2011.

② Schepel，H. *The Constitution of Private Governance: Product Standards in the Regulation of Integrating Markets*. Oxford，UK: Hart，2005.

③ Cashore，B.，Matus，K.J.，Norris，R. "Pathways to Impact: Synergies with Other Approaches." In *Toward Sustainability: The Roles and Limitations of Certification*，2012.

④ Kamara，I. "Co-Regulation in EU Personal Data Protection: The Case of Technical Standards and the Privacy by Design Standardisation Mandate." 2017. https://ejlt.org/index.php/ejlt/article/view/545/725

⑤ Cowls，J.，Floridi，L. "A Unified Framework of Five Principles for AI in Society." 2019. https://doi.org/10.1162/99608f92.8cd550d1

《人权和人类尊严保护公约》（《奥维耶多公约》），该公约试图在更复杂的生物技术创新背景下保护人类尊严。欧洲委员会条约的影响力却是全球性的，已有 67 个国家批准或加入《布达佩斯网络犯罪公约》，其中包括许多欧洲以外的国家，如加拿大、美国、澳大利亚、尼日利亚和日本。

2019 年 9 月，欧洲委员会成立了人工智能临时委员会（CAHAI），以审查人工智能系统开发、设计和应用制定横向和跨界法律框架的可行性。2022 年 6 月，部长委员会指示 CAHAI 的继任机构人工智能委员会在就条约的可能要素开展工作后“迅速着手制定这一类文件”，CAHAI 的提案与上述原则明显不同，它建议围绕有效合规性和独立的国家监管机构制定条款，并建议禁止某些类型的人工智能系统，如那些进行生物识别分类或社会评分的系统。在撰写本报告时，人工智能委员会应美国的要求，以及由于受到来自民间社会机构的批评（他们声称没有做出此类最终决定），同意在第二次全体会议上，由一个仅包括潜在条约缔约方的封闭起草小组起草公约的建议草案。媒体报道强调，这是迫于美国压力进行的，因为欧洲委员会在外交方面对于吸引美国成为签署国非常感兴趣，美国不希望其关于缩小公约范围的谈判立场，如将私营机构全部排除在外的立场已经被公众所知[①]。与此同时，欧盟委员会设法获得了代表欧盟成员国进行谈判的权利，因其担心“公约可能会影响现有的和可预见未来的共同联盟规则或改变其范围”[②]。

（六）国内监管与域外监管的融合

人工智能治理的最具体形式是由各国政府或由各国政府组成的超国家协会（如欧盟）制定规则。与其他领域一样，在数字政策领域，这些法规往往会产生跨国影响，因为它们不仅会影响寻求在相关市场提供产品的国际企业，而且在某些条件下还会激励企业协调其规则以降低跨司法辖区的成本，从而可能会采用新的国家法规作为事实上的国际标准[③]。

① Bertuzzi，L. "US Obtains Exclusion of Ngos From Drafting AI Treaty." January 17，2023. https://www.euractiv.com/section/digital/news/us-obtains-exclusion-of-ngos-from-drafting-ai-treaty/

② Counc. Eur. Union. *Council Decision (EU) 2022/2349 of 21 November 2022 Authorising the Opening of Negotiations on Behalf of the European Union for A Council of Europe Convention on Artificial Intelligence，Human Rights，Democracy and the Rule of Law*. Counc. Decis.，Counc. Eur. Union，Brussels，2022.

③ Bradford，A. *The Brussels Effect: How the European Union Rules the World*. New York: Oxford Univ. Press，2020；Kalyanpur，N.，Newman，A.L. "Mobilizing Market Power: Jurisdictional Expansion as Economic Statecraft." *Int. Organ.*，2019，no.1.

欧盟及其成员国已迅速成为全球人工智能监管的重要参与者，自 2016 年以来，除了应用现有的通用工具外，还通过并提出了各种对机器学习有影响的新监管工具。最近，欧盟委员会提出《人工智能法案》草案，主要目的是为人工智能系统在教育评估、招聘或警务和司法等某些高风险应用领域的特性制定统一的最高标准[①]。这些系统的供应商将负责自我认证其是否满足某些基本要求，这些要求将在欧洲私营标准化机构、欧洲标准化委员会（CEN）、欧洲电工标准化委员会（CENELEC）及其国家成员制定的专有标准中详细说明。任何希望向欧洲市场销售高风险系统或远程交付高风险人工智能系统的供应商，都必须根据欧洲标准对其系统进行认证。因此，希望在全球市场部署模型的供应商很可能必须关注这些标准，这些标准包括对偏见、人为监督和网络安全等问题的考虑，并附带某些公共文件要求。

除《人工智能法案》草案外，许多其他欧洲监管文件也对人工智能的使用、开发或基础设施的某些方面进行监管，并具有不同的全球维度。拟议的《平台工作指令》是另一个例子，因为它倡导对算法管理中使用的系统透明度进行监督（尽管其地域范围限于平台工作实际执行的地方，这意味着如果国际平台选择不使用欧盟境内的工人，那么治外法权的影响是有限的）。《数字服务法案》和《平台对企业的公平条例》都包含一些与在提供特定服务的重大在线平台背景下人工智能推荐系统的透明度相关的条款[②]。由于推荐系统往往跨越国界，因此这种透明度将具有全球性影响力。围绕部署推荐系统的平台对它们推荐的内容以及推荐方式应承担的责任范围的辩论，也产生了重要的法律问题，这些问题正在由法院探讨。这些问题具有特别重要的全球意义，因为它们涉及人工智能的使用能在多大程度上被广泛的中介责任保护所覆盖，这可能会排除其他形式的治理。

除了这些针对人工智能或平台的监管形式外，政府政策制定的许多既定领域也对人工智能系统产生影响。数据保护法就是一个例子。世界各地的数据保护法在形式上通常相似，欧盟内外的许多政策法规都包含与《通用数据保护条例》第

① Veale，M.，Zuiderveen Borgesius F. "Demystifying the Draft EU Artificial Intelligence Act." *Comput. Law Rev. Int.*，2021，no.4.

② Cobbe，J.，Singh，J. "Regulating Recommending: Motivations，Considerations，and Principles. "2019. https://ejlt.org/index.php/ejlt/article/view/686/982

22 条类似的条款[①]，该条款规定了仅在自动处理基础上采取的重大决定和措施所必需的保障措施和法律依据。第 15 条规定可查阅一些重要的特征分析系统的处理逻辑，同样，该条款已在许多司法管辖区实施。这些规定代表了一些共享的最低信息基线。在某些情况下，也代表了在没有人类参与的情况下对人工智能的重要使用的透明度，但这些规定的起草方式使其难以适用于许多现实情况[②]。数据保护法中对人工智能开发而非使用有影响的法定限制虽然较少，但确实存在。通常开发者无法确立处理个人数据的法律义务，例如来自泄露或黑客攻击的数据，用这些数据训练机器学习系统是非法的。只要人工智能系统，尤其是大型语言模型或其他生成工具，可被视为个人数据集，就可能会导致其移动和销售受限，从而产生跨境影响[③]。

知识产权法与上述讨论的国内制度和其他治理制度相互作用。版权法可禁止在未经适当授权的情况下使用某些个人和非个人数据。与此类授权最相关的全球形式是文本和数据挖掘豁免，其于 2009 年首次在日本实施（现为《日本版权法》第 30-4 条、第 47-5 条），随后于 2014 年在英国实施（《1988 年版权、外观设计和专利法》第 29A 条），后来在欧盟的《数字单一市场指令》中有所体现[④]。随着从网络上抓取和使用文本和多媒体（尤其是来自专业的艺术家、作家和编码员的文本和多媒体）用于大规模生成模型的训练的增多，这些豁免正变得越来越重要。此类模型的单个用户是否对其所使用模型的输出享有权利，或者这些输出是否不符合版权等保护条件，也会影响上述全球契约治理的范围。此外，商业秘密法的国际轮廓也限制了监管机构对人工智能系统和数据采集的审查，以及对部署机构可以强制执行的透明度[⑤]。

在某些情况下，如果某些人工智能基础设施可能会产生更广泛的反竞争或损

① Binns，R.，Veale，M. "Is That Your Final Decision? Multi-Stage Profiling，Selective Effects，and Article 22 of the GDPR. "*Int. Data Priv. Law*，2021，no.4；Demetzou，K.，Zanfir-Fortuna，G.，Vale，S.B. "The Thin Red Line: Refocusing Data Protection Law on ADM，A Global Perspective with Lessons from Case-Law." *Comput. Law Secur. Rev.*，2023.

② Custers，B.，Heijne，A-S. "The Right of Access in Automated Decision-Making: The Scope of Article 15(1) (h) GDPR in Theory and Practice."*Comput. Law Secur. Rev.*，2022.

③ Veale，M.，Binns，R.，Edwards，L. "Algorithms That Remember: Model Inversion Attacks and Data Protection Law." *Philos. Trans.*，2018.

④ Caspers，M.，Guibault，L.，McNeice，K.，Piperidis，S.，Pouli，K.，Eskevich，M. "D3.3+ Baseline Report of Policies and Barriers of TDM in Europe (Extended Version)." 2017. https://perma.cc/W6MF-4WLM

⑤ Radauer，A.，Bader，M.，Aplin，T.，Konopka，U.，Searle，N.，et al. *Study on the Legal Protection of Trade Secrets in the Context of the Data Economy: final report*，2022.

害性影响，那么竞争法可能会阻止跨服务的数据合并，并可能阻止这些基础设施的统一。竞争管理机构正在探索人工智能与其现有权力之间的相互作用，同时他们和立法者也在寻求新的事前手段，如《数字市场法》或拟议中的英国数字市场部门。然而，迄今为止，鲜见直接围绕人工智能进行执法，因为监管机构通常考虑的是底层数据基础设施，而不是人工智能本身。

三、全球人工智能治理的理据和紧张关系

虽然人工智能的开发、使用和基础设施的全球治理目前正处于快速活动状态，但值得退后一步，揭示全球治理工作的潜在驱动因素和影响。从理论和实证角度来看，更清晰地理解全球监管体系的原理和局限性，可以阐明这些努力的潜力，但也可能会显现出紧张关系和不足之处。

（一）跨境治理问题

全球努力背后首要或许也是最明显的一个理由是地理因素，具体来说，是这些系统的规模和跨境性质及其影响。连接人工智能技术的供应和实施的价值链，以及它们所依赖的基础设施都具有跨境要素。同时，通过一套规则来治理它们有时在功能上是有利的（就像互联网协议一样），并且在经济上对希望受单一规则体系治理的企业来说也是有利的。

跨境全球治理也源于对跨境影响的担忧。人工智能价值链类似于商品供应链。这意味着，支持数据收集和培训的基础设施可能会对一些国家的低薪工人造成伤害，这些伤害远远超出使用这些数据的软件公司的劳动保护规定。平台经济中的内容审核已经证明了相关危害[①]。近年来，出现了关于数据标签对边缘化工人造成伤害的跨境诉讼。由于此类标签实践具有高度的非物质化、信息化和全球流动性，因此有观点认为，必须在全球层面处理由此产生的治理问题。

当然，全球治理并不是应对跨境影响的唯一途径。这些问题可以通过从正式的（非全球性）贸易协定到供应链合同要求的治理工具来解决。然而，这些更具地域局限性的方法也存在潜在的弊端，可能会将应对措施推向全球。在这种情况

① Gray，M.L.，Suri，S. *Ghost Work: How to Stop Silicon Valley from Building a New Global Underclass*. Boston: Houghton Mifflin Harcourt，2019；Roberts，S.T. *Behind the Screen: Content Moderation in the Shadows of Social Media*. New Haven，CT: Yale Univ. Press，2019.

下，关于逐底竞争的成熟论点为我们提供了一个有益的视角。这一论点是在环境监管学术背景下出现的，它认为如果缺乏集中的跨境治理制度，企业就会受到激励，将其对环境危害最大的活动转移到监管严格程度低、成本低和/或执法力度低的地区[①]。为了保持竞争力，监管更严格的地区或国家将被迫降低标准，从而导致监管标准整体呈下降趋势。一些研究指出，企业事实上没有将业务迁往标准较低的地区的原因有很多，其中包括成本、企业声誉等。然而，有研究表明，较贫穷的工业化程度较低的国家面临着这种监管标准降低的竞争压力，而收入较高、工业化程度较高、监管严格的国家则没有这种压力[②]。换句话说，即使这会对当地产生负面影响，缺乏集中的全球监管也意味着那些制度性较弱的国家将会以较宽松的监管为基础进行竞争。

虽然最初的研究主要集中在环境监管方面，但现在的研究已经扩展到劳务和其他监管领域，研究发现，劳务监管的减少会导致外国直接投资的增加[③]。对于人工智能体系而言，令人担忧的是，一些地方将成为更多有害活动（即低工资劳动）或开发有问题的软件的避风港，而这些在其他地方将受到监管，甚至是禁止。鉴于算法和数据可以轻易移动，这可能会破坏对能够用于已受提议监管的系统进行控制的尝试，例如在心理操纵领域的使用、作为进攻性物理或网络武器或作为监视系统。这也可能阻碍基础设施治理发展的尝试，因为人工智能价值链中的公司可能会寻找当地环境或劳动力政策能够为数据服务器或标签服务创造更具经济竞争力条件的司法管辖区[④]。

但就目前看来，针对建立全球治理体系、保护防止有害发展的做法或在其他地方限制人工智能产品开发等的潜在合法理由的努力似乎还不够，并没有在当前的实践中更多地体现。在这些潜在的框架中，全球治理是一项由国家主导的活动。例如，尽管一些私营监管系统在可持续商品认证方面发挥了作用，但它们不

① Esty，D.C. *Greening the GATT: Trade，Environment，and the Future*. Washington，DC: Inst. Int. Econ.，1994；Konisky，D.M. "Regulatory Competition and Environmental Enforcement: Is There a Race to the Bottom? " *Am. J. Political Sci.*，2007，no.4；Porter，G. "Trade Competition and Pollution Standards: ' Race to the Bottom' or ' Stuck at the Bottom'? " *J. Environ. Dev.*，1999，no.2.

② Porter，G. "Trade Competition and Pollution Standards: ' Race to the Bottom' or ' Stuck at the Bottom'? " *J. Environ. Dev.*，1999，no.2.

③ Olney，W.W. "A Race to the Bottom? Employment Protection and Foreign Direct Investment." *J. Int. Econ.*，2013，no.2.

④ Ensmenger，N. "The Environmental History of Computing."*Technol. Cult.*，2018，no.5.

能相互取代。事实上，私人主导的治理体系往往会相互竞争[①]，而国家主导的方法既能减少负面竞争影响，同时又能维护政治合法性——目前的人工智能全球治理工作恰恰缺乏政治合法性[②]。

（二）游说治理

治理机制本身也可以在促进行为者获得优先的政策选择和结果方面发挥作用，无论是在特定体制内还是考虑到跨体制的互动。我们既要探讨如何防止采用更严格的治理方式（如法律），也要探讨如何发出战略信号以扩大影响力。

1. 预防监管

如果说当前的尝试与上述理由并不特别吻合的话，那么一种更为愤世嫉俗的看法是，这些尝试恰恰是在试图阻止前文所述的国家驱动的治理类型。对于全球治理学者和跨境监管研究者来说，私营企业主导的自愿治理尝试与现有或未来国家主导的约束性监管之间的互动性质是一个长期争论的问题。无论是企业责任计划[③]、产品认证计划[④]，还是由企业主导的更广泛形式的“监管标准制定”[⑤]，公司做出的各种自愿承诺都可以战略性地部署，以减少政策制定者和公众对更强大、更健全规则的需求。它们可以使企业处于有利地位，在宣布各种承诺（或建立行业自我监管机构，寻求协调自愿标准、实践和规范）之后，这些承诺可以在方便的时候被简单收回或永远不会得到有效履行[⑥]。

技术人员主导的治理努力可以将一个政策领域描绘为充满不确定性，需要的知识看似触手可及，但永远都不会出现。计算机科学家可能会从著名的停机问题的角度来思考这个问题——一项可能永远不会结束的尝试。将人工智能描述为一个高度技术性的领域，利用了公共部门对科技公司作为“未来使者”的迷恋，这

① Lambin，E.F.，Thorlakson，T. "Sustainability Standards: Interactions between Private Actors，Civil Society，and Governments." *Annu. Rev. Environ. Resour.*，2018；Steer. Comm. State-of-Knowledge Assess. Stand. Certif. *Toward Sustainability: The Roles and Limitations of Certification*. Washington，DC: Resolve Inc，2012.

② Erman，E.，Furendal，M. *Artificial Intelligence and the Political Legitimacy of Global Governance*. Political Stud. In press，2022.

③ Auld，G.，Gulbrandsen，L.H.，McDermott，C.L. "Certification Schemes and the Impacts on Forests and Forestry." *Annu. Rev. Environ. Resour.*，2008.

④ Bartley，T. "Transnational Private Regulation in Practice: The Limits of Forest and Labor Standards Certification in Indonesia." *Bus. Politics*，2010，no.3.

⑤ Abbott，K.W.，Snidal，D. "The Governance Triangle: Regulatory Standards Institutions and the Shadow of the State." *In the Politics of Global Regulation*，2009.

⑥ Malhotra，N.，Monin，B.，Tomz，M. "Does Private Regulation Preempt Public Regulation?" *Am. Political Sci. Rev.*，2019，no.1.

些公司值得在认识论上被给予尊重[①]，同时它们所积累的人工智能专业知识也使它们处于规则制定的核心。立法者可能会因为误解而制定出损害行业的规则，这种论点在政治上有很强的影响力。然而，与人工智能治理有关的知识并不是通过更多的专业知识就能解决的，它最好被归类为“超常规的”[②]，这对科学在问题解决中所扮演的经典角色提出了挑战。人工智能治理真正涉及的问题往往不是技术性的，而是深层次的规范性和分配性问题，即哪些参与者在社会中做出决策，谁来承担风险，以及程序上和实质上的正义应该是什么样的。在事实不确定、价值观有争议、决策风险高的情况下，不太可能达成科学共识。因此，从技术或认识论不确定性的角度来阐述问题就成了一种拖延策略。尽管有些人梦想着将价值观标准化、编码化，并整合到透明的算法中，就像打上技术补丁一样。当新的政治主张和运动对我们的社会如此重要，并需要我们在更广泛的背景下不断反思算法实践时，它认为我们可以而且应该将政治锁定到位[③]。

由于地方和区域的政治偏好、文化或经济状况不同，人工智能治理问题的一些核心价值观也会有所差异。将在国际上有分歧的问题推到全球范围内需要达成共识，不同的国家监管可能是一种合适的替代方法，至少在中短期内是这样。辅助性政治——在较低水平上实现某些足够的结果，从根本上与许多科技公司选择的平台商业模式不相容，这些模式将畅通无阻（或不受阻碍）、低摩擦的规模扩张作为深层基础[④]。

2. 信号传递

一些研究表明，通过表现出社会责任感，企业可以更接近政策制定者[⑤]。在负责任的人工智能研究领域，上述治理方法的自愿创建、资助和支持已经与企业捕获交织在一起。分类法、逻辑以及思考、检验、评估和构建框架的方式，很可能会因为行业研究在公开学术论坛上讨论这些话题的比例，而被纳入国家主导的

① Zuboff，S. "Big Other: Surveillance Capitalism and the Prospects of an Information Civilization." *J. Inf. Technol.*，2015，no.1.

② Funtowicz，S.O.，Ravetz，J.R. "Science for the Post-Normal Age."*Futures*，1993，no.7;G20. *G20 Ministerial Statement on Trade And Digital Economy*，2019. https://www.mofa.go.jp/mofaj/files/000486596.pdf

③ Amoore，L. *Cloud Ethics: Algorithms and the Attributes of Ourselves and Others*. Durham，NC: Duke Univ. Press，2020.

④ Pfotenhauer，S.，Laurent，B.，Papageorgiou，K.，Stilgoe，J. "The Politics of Scaling." *Soc. Stud. Sci.*，2022，no.1.

⑤ Werner，T. "Gaining Access by Doing Good: The Effect of Sociopolitical Reputation on Firm Participation in Public Policy Making." *Manag. Sci.*，2015，no.8.

政策讨论。这种兴趣和专业知识的表征似乎导致了这样的情况：公共部门的“专家小组”主要由可能受到这类规定约束的实体所组成[①]。这并不是说这些公司中的许多人不是该领域的专家——恰恰是研究团队（尤其是来自学术界）被迅速雇佣和积聚到行业实验室，并在传统的学术场域发表论文，才使得在讨论任何与人工智能治理相关问题的会议上，行业代表的人数越来越多。一方面对技术持有批判性观点，另一方面又从行业研究职位中获益，近年来这种紧张关系变得更加明显，因为许多人由于在谷歌等大型技术公司组织发表研究成果而被解雇[②]。

（三）治理框架

建立新的制度和规范是一种创造性的行为，它可以朝着多个方向发展。它需要对未来进行推理，考虑技术和经济或社会关系将如何发展。那些设定舞台的人拥有相当大的权力，我们现在要考虑这种影响力。

1. 让人工智能成为现实

长期以来，科学和技术研究学者一直对有关未来的文本如何塑造未来感兴趣。设想未来技术的治理工作涉及创造期望，而这些期望反过来又会赋予某些行为者超越其他行为者的更大的权力[③]。因此，参与制定治理模式既能造就成功的科学家，也能造就成功的政策制定者[④]。在人工智能治理中，国家战略文件被明确强调为具有执行维度，即试图“将人工智能付诸实践”[⑤]。全球治理也可以从类似视角出发，既将某些愿景及其附带的商业模式纳入政策讨论，又进一步将注意力吸引到人工智能的核心公司上，因为它们正在实现人类技术的未来。

2. 国家的雄心

在公众关注度和监管压力不断增加的环境下，行业及其主导联盟越来越积极主动地引导在线内容的监管工作[⑥]。该倡议造成了两极分化，一些人认为它是棘

① Veale，M. "A Critical Take on the Policy Recommendations of the EU High-Level Expert Group on Artificial Intelligence." *Eur. J. Risk Regul.*，2020，no.1.

② Whittaker，M. "The Steep Cost of Capture. " *Interactions*，2021，no.6.

③ Van Lente，H.，Rip，A. *Expectations in Technological Developments: An Example of Prospective Structures to be Filled in by Agency.* In Getting New Technologies Together ，2012，pp. 203-230.

④ Voß J-P. "Performative Policy Studies: Realizing ' Transition Management'." *Innovation*，2014，no.4.

⑤ Bareis，J.，Katzenbach，C. "Talking AI into Being: The Narratives and Imaginaries of National AI Strategies and Their Performative Politics." *Sci. Technol. Hum. Values*，2022，no.5.

⑥ Caplan，R. "Networked Platform Governance: The Construction of the Democratic Platform." *Int. J. Commun.*，2023；Gorwa，R. "The Platform Governance Triangle: Conceptualising the Informal Regulation of Online Content." 2019. https://doi.org/10.14763/2019.2.1407

手的跨境言论治理的创新方法，而批评者则更倾向于认为它是一个存在重大缺陷的机构，甚至可能主要是一种公关行为[①]。无论其直接影响如何，监督委员会一直是有利于 Meta/Facebook 的更广泛合法化战略的一部分[②]：它允许公司使用“宪法隐喻”，并借助法律和国家建设的语言，进一步将自己塑造成可接受的、值得信赖的规则制定者[③]。

在人工智能领域，类似 PAI 这样的努力也有一些类似的特点。通过创建一个表面上由多方利益相关者参与、但实际上主要由行业主导和控制的政策论坛（其创始的 10 人董事会由 7 位行业代表、1 位学者、1 位资助者和 1 位公民社会代表组成），强大的平台型跨国公司不仅可以寻求影响人工智能全球治理的议程，确保其符合他们的广泛利益，还可以公开（向政策制定者）展示自己积极参与负责任且对社会有益的人工智能创新的深层问题。在这样做的过程中，这种尝试可以以具有影响力的方式框定政策领域和参与者，例如在劳动力方面，考虑工作的自动化而不是全球数据标注和处理流水线中工人面临的真实的问题[④]，或者是“人工智能应对气候变化”的一般概念，而不是信息和通信技术的生产、使用和处置过程所带来的日益严重的环境问题[⑤]。

3. 不恰当的人工智能中心主义

将治理问题视为人工智能系统的特征本身就是一种选择，而且这种选择并不总是十分恰当的。即使是占据研究文献的经典算法问题，在很多方面也根本不是以模型为中心的。在透明度等问题上，要将人工智能系统的不透明性与其使用的不可预知性、用户的知识和培训、文档或代码的可用性、对其输出的监督，或仅

① Arun，C. "Facebook's faces." *Harvard Law Rev. Forum*，2022；Klonick，K. "The Facebook Oversight Board: Creating an Independent Institution to Adjudicate Online Free Expression." *Yale Law J.*，2019，no.8；Pallero，J. "What the Facebook Oversight Board Means for Human Rights，and Where We Go From Here. " 2020. https://apo.org.au/node/307040

② Dvoskin，B. "Expertise and Participation in the Facebook Oversight Board: from Reason to Will." *Telecommun. Policy*，2022，no.5.

③ Cowls，J.，Darius，P.，Santistevan，D.，Schramm，M. "Constitutional Metaphors: Facebook's 'Supreme Court' and the Legitimation of Platform Governance." *New Media Soc.*，2022，no.5.

④ Posada，J. "Embedded Reproduction in Platform Data Work." *Inf. Commun. Soc.*，2022，no.6；Van Doorn，N.，Badger，A. "Platform Capitalism's Hidden Abode: Producing Data Assets in the Gig Economy." *Antipode*，2020，no.5.

⑤ Belkhir，L.，Elmeligi，A. "Assessing ICT Global Emissions Footprint: Trends to 2040 & Recommendations." *J. Clean. Prod.*，2018；Dobbe，R.，Whittaker，M. "AI and Climate Change: How They're Connected，and What We Can Do About It. " October 17，2019. https://medium.com/@AINowInstitute/ai-and-climate-change-howtheyre-connected-and-what-we-can-do-about-it-6aa8d0f5b32c

仅是其输入数据的人类可解释性区分开来，并非易事[①]。尽管有黑盒的隐喻，但许多黑盒本身却深陷于阴影之中，例如，商业模式或公共部门承包商的不透明造成的阴影。即使是白色的盒子，在黑暗中看起来也是黑色的。歧视问题不能轻易地与数据收集和再训练的动态过程分开，或者与人工智能系统输出相关的决策或措施的性质分开[②]。与特定任务相关的故障模式问题无法轻易地与这些任务的问题框架分开，其中一些任务可能根本无法预测或自动化[③]。问责问题既不能脱离更为漫长的社会技术发展过程[④]，也不能脱离对积极、批判性受众的需求。

使用人工智能领域中的权力所引起的担忧可能源自于与人工智能治理的辩论和框架无关的技术领域。推荐系统的设计空间与封闭的、整体式的社交媒体平台紧密交织在一起[⑤]。人脸识别技术的使用与公共场所摄像头的安装和重新利用密切相关。在加密的信息渠道中，机器学习对视听媒体进行不规则的算法扫描和标记（所谓的客户端扫描）的建议是通过锁定的终端设备实现的，而这些终端设备的检查或安装替代软件的途径有限[⑥]。这表明全球技术治理的脱节，也凸显了一个重大风险，即人工智能的全球治理是潜在监管者所选择的，也是被夸大的渠道，目的是将更深层次、更具结构性的问题排除在外。将问题框定为技术问题往往会导致将解决方案框定为技术问题[⑦]，而技术公司则是这一结果的主要受益者。

4. 非完全全球化

对于全球治理来说，一个至关重要的问题是谁能参与到决策中，以及谁能拥有治理的话语权。如果从历史角度来看待全球治理，将其与国际关系的更广泛模式以及国家与其他寻求“全球治理”的新兴行动者之间的各种互动联系起来，那

① Kemper，J.，Kolkman，D. "Transparent to Whom? No Algorithmic Accountability Without a Critical Audience. " *Inf. Commun. Soc.*，2018，no.14；Vaughan，J.W.，Wallach，H. *A Human-Centered Agenda for Intelligible Machine Learning.* In *Machines We Trust: Perspectives on Dependable AI*，Cambridge，MA: MIT Press，2021，pp.123-138.

② Lum，K.，Isaac，W. "To Predict and Serve?" *Significance*，2016，no.5.

③ Birhane，A. *Cheap AI.* In *Fake AI.* Manchester，UK: Meatspace，2021；Passi，S.，Barocas，S. "Problem Formulation and Fairness." In *Proceedings of the Conference on Fairness，Accountability，and Transparency*，2019.

④ Cobbe，J.，Lee，M.S.A.，Singh，J. "Reviewable Automated Decision-Making: A Framework for Accountable Algorithmic Systems." *In Proceedings of the 2021 ACM Conference on Fairness，Accountability，and Transparency*，2021.

⑤ Keller，D. "The Future of Platform Power: Making Middleware Work." *J. Democr.*，2021，no.3.

⑥ Abelson，H.，Anderson，R.，Bellovin，S.M.，Benaloh，J.，Blaze，M.，et al. *Bugs in Our Pockets: The Risks of Clientside Scanning.* Work. Pap.，Columbia Univ.，2021；Gorwa，R.，Binns，R.，Katzenbach，C. "Algorithmic Content Moderation: Technical and Political Challenges in the Automation of Platform Governance." 2020. https://doi.org/10.1177/2053951719897945

⑦ Green，B. *The Smart Enough City.* Cambridge，MA: MIT Press，2019.

么可以说，排除某些群体一直是全球治理项目的核心问题[①]。有人可能会说，“当代全球治理起源于一系列旨在通过管理欧洲各国势力之间的冲突来帮助确保白人全球主导地位的机构”[②]。因此，我们必须密切关注参与人工智能治理新机构的行动者，以及在这些机构中形成的代表和排除模式。

由政府、国际组织和行业牵头，为全球“善用人工智能”或“负责任的人工智能”制定政策倡议的许多努力，都陷入了“参与悖论”，涉及经济较不发达的利益相关者表面上的参与，却没有提供相应的资源和结构性改革来使他们能够有意义地参与其中[③]。全球多数国家的人工智能基础设施、开发和使用似乎都以物质提取主义和商业开发的系统逻辑为特征，尽管表面上有各种全球道德准则和多方利益相关者治理来塑造它们。这就导致了当前的发展态势，即在多元化背景下，政治行动者并没有接受去政治化的、被宣称为双赢的全球治理中的包容，而是在新兴的人工智能治理体系中发挥了一种对抗性的“挑战功能”，积极寻求通过一系列创造性技巧来抵抗和颠覆它。

四、结论：人工智能全球治理中的权力

如果说某种全球人工智能治理体系——由一套相互重叠的跨国私人标准和最佳实践制定倡议以及国际组织颁布的规范性宣言和原则组成，并由多个司法管辖区的新法律框架和现有法律框架共同支撑——正在慢慢形成，那么，我们最好从全球治理学术研究中关于其他行业相关体制的起源、影响和缺陷的为期二十多年的辩论中吸取教训，包括那些与数字政策问题不直接相关的行业。由于全球人工智能发展格局的特点是行业控制和影响力水平非常高，私人权利的问题至关重要。这些公司以及它们创建或参与的治理举措，如何实现与监管工作互动？它们推动非正式跨境治理形式的努力，如何与它们在全球背景下对公众影响力、游说和利益代表的更广泛政治策略相互作用？

由于人工智能强调的社会、经济和环境问题并不仅仅是技术本身的结果，而

① Avant，D.D.，Finnemore，M.，Sell，S.K. *Who Governs the Globe?* Cambridge，UK: Cambridge Univ. Press，2010.

② Murphy，C.N. "The Last Two Centuries of Global Governance." *Glob. Gov.*，2015，no.2.

③ Png，M-T. "At The Tensions of South and North: Critical Roles of Global South Stakeholders in AI Governance."In *2022 ACM Conference on Fairness，Accountability，and Transparency*，2022.

是与许多更广泛且截然不同的系统相互交织在一起，因此，如何构建问题的框架至关重要。从人工智能的角度来看会产生什么后果？全球对话通过人工智能的视角看待歧视、问责、能源使用和隐私等问题时，无论是全球范围的还是其他形式的，哪些潜在的治理形式被边缘化了？

当国家、公司、利益集团和其他参与者就新兴技术进行讨论时，他们正在为什么样的未来做准备，并为实现这一愿望做出怎样的贡献？在未来，对于那些不在场的行动者、社区、地区和利益集团来说，有什么暗示的角色或后果？他们之所以不在场，是因为他们被体制性地排除在外，还是因为他们缺乏资源、能力或联系而无法参与？如果他们能够设定讨论的框架和基调，这场对话会有何不同？综合来看，在考虑人工智能治理举措的影响时，我们应当牢记国际政治经济学批判学者提出的关键问题[①]——谁是受益者？谁真正受益？

（免责声明：作者不知晓任何可能被认为会影响本文客观性的隶属关系、成员资格、资助或金融控股情况。）

（责任编辑 李淑华）

① Sell，S.K. "Ahead of Her Time? Susan Strange and Global Governance." In *Susan Strange and The Future of Global Political Economy: Power，Control and Transformation*，London: Routledge，2016，pp.21-32.