

全球人工智能治理探究

——基于委托—代理理论视角

蔡翠红 张璐瑶

内容提要 在风险与机遇并存的背景下,如何整合国际组织、国家、科技企业、市民社会等多元主体,发挥其协同作用是当前人工智能治理的关键问题。基于委托—代理理论,国家在人工智能治理中发挥着“双职中枢”作用:既作为治理体系中的核心委托方以减轻信息不对称和代理成本,又作为跨层次互动的协调中枢,推动多元主体在单元、互动和系统三个层面上实现有效合作。这一作用在实践中面临代理成本复杂化、多方目标协调困难和治理韧性不足等挑战。美国、英国和中国三国基于自身制度环境发展出差异化策略:美国采取分布式监管降低单点成本,英国强调设计安全前置风险管控,中国注重统筹规划与稳健实施。这些不同实践路径下的共同目标意味着国家在促进多主体协调、减少信息不对称及增强系统韧性方面的关键作用不可替代,委托—代理框架为构建更具包容性和有效性的全

* 蔡翠红:复旦大学美国研究中心教授。(邮编:200433);张璐瑶:复旦大学国际关系与公共事务学院博士生。(邮编:200433)

** 本文系国家社科基金重点项目“总体国家安全观视阈下的数字主权研究”(项目编号:23AZZ002)的阶段性成果。感谢《国际政治研究》匿名审读专家的意见和建议,文责自负。

球人工智能治理提供了理论指导与实践参考。

关键词 非传统安全 人工智能 治理 委托—代理 多主体
协调

随着人工智能技术的加速发展,其应用范围和影响深度不断扩大,全球人工智能治理已成为国际社会的紧迫议题。2023 年 10 月,中国发布《全球人工智能治理倡议》,提出各国政府、国际组织、企业、科研机构、民间机构和公民个人等各主体需要秉持共商共建共享的理念,协力共同促进人工智能治理。^① 2024 年 9 月,联合国未来峰会通过《全球数字契约》,明确在全球范围内共同设计、使用并治理人工智能的承诺,强调需要协同国家(特别是发展中国家)和其他各利益攸关方的努力,使智能技术造福人类。^② 这些倡议均指向一个核心共识:有效的人工智能治理需要多主体之间的协同合作。

多主体协同虽已成为全球人工智能治理的共识,但其具体实现机制仍面临严峻挑战。这些挑战主要体现在单元、互动和系统三个层面:单元性问题源于人工智能赋能的广泛性和普惠性,致使各主体在治理体系中均具备独特角色和能力边界,这就需要明确国际组织、国家、科技企业、智能平台及公民个体在治理中的具体定位。互动性问题关注不同单元间的关系结构及互动规则,反映各主体在治理过程中的利益博弈与协作机制,核心在于构建有效的沟通协调渠道,使各主体能在竞争与合作中共同形成可持续的治理生态。系统性问题则聚焦于单元和互动关系构成的整体治理架构及其运行效能,探讨人工智能治理如何在多层次互动中实现系统协同与整体优化。这三个层面的问题相互交织,共同构成人工智能全球治理的核心难题。

在此背景下,本文提出的核心问题是:在全球人工智能治理框架下,如何

① 中央网络安全和信息化委员会办公室:《全球人工智能治理倡议》,2023 年 10 月 18 日, https://www.cac.gov.cn/2023-10/18/c_1699291032884978.htm, 2025-03-06。

② United Nations, “Pact for the Future, Global Digital Compact and Declaration on Future Generations,” September 2024, pp. 39-49, <https://www.un.org/en/summit-of-the-future/pact-for-the-future>, 2025-03-05.

构建有效的多主体协同机制？本研究基于全球主要国家和地区的人工智能治理文件、国际组织发布的治理框架、典型人工智能企业的治理实践案例，采用多层次分析框架进行研究，进而从单元层面梳理国际组织、国家政府、智能平台和公民个体在人工智能治理中的角色定位与责任边界，厘清各主体的能力特征与权责配置；从互动层面分析不同主体间的关系模式、权力分布与协调机制，识别哪种行为体能够在复杂的互动系统中发挥提纲挈领的作用；从系统层面在复杂的多主体关系网络中识别核心节点与结构特征，探索多主体治理的制度设计与优化路径。

一、人工智能治理中的委托—代理关系网络

在网络空间等数字领域，对行为体的认定有基本的“四分法”：即由于权力的流散，治理活动开始由拥有体系性权力的国际组织、拥有工具性或结构性权力的国家、拥有元权力的科技企业，以及拥有倡议权力的市民社会等四类主体承担。^① 这种“四分法”在人工智能领域也基本适用，但人工智能的生命周期涵盖从研发到部署、从设计到应用的多个阶段，不同阶段涉及的治理内容与责任各不相同。因此，在“四分法”框架基础上，有必要根据人工智能的技术特征对主体的能力类型和功能进行细分和调整。

首先，国际组织在人工智能治理中具备规范性治理能力，主要体现为制定和推动全球技术标准与伦理规范。人工智能技术发展具有全球性影响，单一国家或区域难以独立应对其伦理和安全挑战。因此，国际组织需通过协调各国行动、引导公共利益，推动全球治理标准化进程。例如，联合国通过的《人工智能伦理建议书》和经合组织(OECD)发布的《人工智能原则》(The OECD AI Principles)是国际组织发挥规范性治理作用的典型案例。^② 此外，国际组织还可以发挥倡导性作用，识别国家战略中忽视的领域。联合国发布了《治理人工

① 蔡翠红：《国家—市场—社会互动中网络空间的全球治理》，《世界经济与政治》2013年第9期，第90—112、158—159页。

② 耿召：《数字空间国际规则制定中的功能性组织角色：以经合组织为例》，《国际论坛》2023年第4期，第24—47、155—156页。

智能,助力造福人类:最终报告》,倡导利用人工智能促进全球公平、赋能可持续发展目标,鼓励国家和企业以发展的视角看待智能技术。在这一过程中,国际组织充当了技术发展与社会需求间的桥梁,进一步保证人工智能进步符合全球社会长远利益。

其次,国家拥有对人工智能的战略性治理能力,主要体现在政策制定、法律框架和技术战略设计方面。国家既是人工智能技术发展的推动者,也是监管者。在全球竞争日益激烈的背景下,国家通过制定国家人工智能战略、推动研发创新及规范技术应用,在保证人工智能技术与国家战略相协调、相一致的同时把握战略机遇期,推动国家以技术优势塑造全方位的战略优势。^① 例如,中国政府通过发布《新一代人工智能发展规划》,明确了未来人工智能发展的目标和路径,既要推动技术创新,又要应对技术应用带来的社会问题和安全挑战。国家的战略性治理能力不仅体现在技术发展规划上,还体现在确保技术符合伦理要求与国际规则的过程中。

再次,与国际组织和国家不同,科技企业在人工智能治理中扮演着更为多样化的角色,根据不同科技企业的业务和能力可以进行进一步细化:一类是在设计阶段的人工智能研发企业,这些企业由于对技术,特别是对算法和模型的掌控而拥有了安全性治理能力。安全性治理能力要求企业在技术研发阶段构建安全、可靠且透明的人工智能系统,尤其在应对“黑箱”问题和技术滥用风险时,研发企业肩负重要责任。^② 例如,美国开放人工智能研究中心(OpenAI)的“超级对齐”(Superalignment)团队在人工智能设计安全性领域开展了大量工作,包括开发可扩展训练方法、验证生成模型以及对整个对齐流程进行压力测试。^③ 美国开放人工智能研究中心的实践表明,安全性治理能力不仅要求研发企业具备足够的技术掌控力,还需预见技术应用的潜在风险;另一类是技术应用和部署阶段的智能平台,它们塑造用户在智能时代的活动特征与生态,因而

① 彭红梅、刘忠:《人工智能对国家安全的影响:基于美国四家智库的观点分析》,《情报杂志》2020年第7期,第36—41、28页。

② Qasim NAUMAN, “16 Top AI Firms Make New Safety Commitments at Seoul Summit,” <https://techxplore.com/news/2024-05-ai-firms-safety-commitments-seoul.html>, 2025-02-27.

③ “Introducing Superalignment,” July 5, 2023, <https://openai.com/index/introducing-superalignment/>, 2025-02-27.

更易掌握内容性治理能力。这意味着智能平台企业需保证人工智能技术应用符合社会规范,同时管理用户生成内容,防止虚假信息传播、歧视性算法应用及其他潜在恶意操纵。^① 典型案例包括阿里巴巴、脸书(Facebook)等平台企业,它们在提供人工智能服务的同时需要管理技术应用与用户行为。在此过程中,平台企业通过算法调控、内容审核等方式确保技术应用符合道德与法律要求,从而实现内容性治理。^② 当然,部分科技巨头(big tech)同时兼具研发与平台功能,如谷歌和字节等,它们拥有综合性治理能力。这些科技巨头不仅需进行技术安全性治理,还须负责平台运营中的内容管理。^③ 因此,它们在人工智能治理中发挥关键作用,往往同时承担技术创新、风险预防与用户管理三重责任。

市民社会在人工智能治理生态中发挥着不可替代的监督与纠偏功能,拥有与网络空间治理中类似的倡议性与反馈性治理能力。这种能力主要通过两种途径实现:一是利用公众舆论塑造社会共识与期待,二是提供直接的用户反馈影响企业决策。这些机制共同作用,促使企业与政府承担相应的治理责任,在人工智能规制体系中构建起自下而上的约束力量。在技术治理与伦理困境日益复杂化的当下,市民社会通过持续倡导算法透明性、决策问责制和技术应用公平性等核心原则,为人工智能治理注入了公共价值导向,有效推动了治理过程的开放性与公正性,成为在人工智能带来的“生存性危机”阴影下,制衡技术潜在风险的重要社会力量。

综合分析以上四类人工智能治理主体的能力结构与功能定位,可以清晰地发现它们之间存在明显的专业分工与能力互补关系——国际组织的规范性能力、国家的战略性能力、研发企业的安全性能力、平台企业的内容性能力,以及市民社会的倡议性能力各有侧重,形成了一个相互依存的治理生态系统。

① 杨云霞、陈鑫:《霸权国家互联网平台巨头话语权垄断及我国应对》,《世界社会主义研究》2021年第11期,第84—93、116页。

② Karen Huang and P. M. Krafft, “Performing Platform Governance: Facebook and the Stage Management of Data Relations,” *Science and Engineering Ethics*, Vol.30, No.2, 2024, p. 13.

③ Raquel Jorge-Ricart, “Big Tech Companies and States: Policy or Politics?” Elcano Royal Institute, March 2, 2020, <https://www.realinstitutoelcano.org/en/blog/big-tech-companies-and-states-policy-or-politics/>, 2025-03-05.

这进一步印证了单一主体独立塑造人工智能这一全方位颠覆性技术治理生态的内在局限性:如果治理仅依托于技术研发企业,不仅难以实现技术全生命周期的有效监管,更无法确保技术发展轨迹与广泛的伦理共识和国家战略需求相协调;若仅由国家主导,则可能缺乏足够的技术理解和创新活力,从而使得政策与产业难以相向而行;而纯粹的国际组织框架则往往面临执行力不足的困境。

在此背景下,可以合理假定,人工智能治理的总体目标必然是综合性的,需要整合各方优势以应对复杂的治理挑战。具备专业性治理能力的单一主体,当面对超出其核心能力范围的任务和项目时,自然会倾向于通过机构设置、用户协议、商业合作等多种形式的契约关系,将特定职能委托给拥有更专业资源和能力的主体,从而实现协同治理的目标。正是这种基于能力差异与互补的“委托”与“被委托”关系,构成了人工智能治理生态中最为基础的互动逻辑,由此形成以“委托—代理”(principal-agent)为主要特征的治理主体互动模式。“委托—代理”理论产生于经济学的研究背景下,遵循的基本假设是随着社会分工的细化和专业化程度的提高,委托人越来越难以独立完成所有任务,需要依靠专业的代理人来完成特定任务。^① 在人工智能治理中,各个主体都可以根据具体情境和目标,扮演委托人或代理人的角色,形成复杂的委托—代理关系网络。

在全球人工智能治理的框架下,四类主体之间形成了专业且相互关联的互动结构:

国际组织和平台企业在人工智能治理网络中普遍扮演“代理方”角色,这一定位源于它们在各自领域所具备的独特优势:国际组织在全球公共利益协调方面拥有制度性权威和跨国动员能力,平台企业则在数字生态系统构建和技术应用层面具备不可替代的专业能力和市场影响力,这些优势使它们成为其他主体实现治理目标的理想代理方。

作为主要代理方之一,国际组织主要接受国家的委托,通过调用成员国赋

^① 赵蜀蓉、陈绍刚、王少卓:《委托代理理论及其在行政管理中的应用研究述评》,《中国行政管理》2014年第12期,第119—122页。

予的权力和资源,以全球协作机制和规范化工具推动国际规则的制定与执行。^① 在人工智能全球治理中发挥着“软法”建构者和规范传播者的作用。例如,为了实现人工智能发展的安全、向善、符合全球人民的公共利益,中国不仅积极向联合国提交了《中国关于规范人工智能军事应用》和《中国关于加强人工智能伦理治理》等立场文件,而且以实际行动支持联合国决议,推动联合国在全球人工智能治理中发挥更加切实的作用。^②

而与国际组织面临的相对单一的委托关系不同,平台企业在人工智能治理中处于多方委托的交叉点,其代理职能的复杂性远超其他主体:国家基于内容监管和意识形态安全的战略考量对平台提出合规要求;研发企业出于技术安全和商业利益的保障将应用管理权委托给平台;普通用户则基于个人信息保护和 cultural 需求满足向平台授权使用其个人信息……这种多维度的委托结构一方面拓展了平台企业的职能范围和自主空间,使其在人工智能治理中拥有更大的决策权;另一方面,也使平台企业面临利益平衡的严峻挑战,在协调国家监管要求、技术商业价值和用户权益保护等多重目标时容易产生利益偏向和责任模糊,进而引发“道德风险”。^③

与之相对,除了在特定情境下对国家的部分治理需求提供有限代理之外,人工智能研发企业和市民社会更多地扮演人工智能治理中委托方角色。这两类主体在治理生态中具有显著的能动性,能够通过设计契约关系、实施技术控制和建立激励机制等多种手段主动影响代理人的行为,在整个“委托—代理”系统中发挥着战略引导与价值塑造的关键作用。^④

研发企业与平台企业之间的“委托—代理”关系尤为典型,体现了技术从创造到应用的价值实现路径。在这一关系中,研发企业作为技术原创者将应

① 张雪:《国家与国际组织互动关系的“委托—代理”解释框架》,《理论与改革》2021年第1期,第119—130、155—156页。

② 鲁传颖:《全球人工智能治理的目标、挑战与中国方案》,《当代世界》2024年第5期,第25—31页。

③ “道德风险”指的是契约关系的双方由于信息的不对称而拥有更多权力和资源优势的一方愿意去违背委托方目标攫取更多的私利和自主性的行为。参见 Hector Chade and Jeroen Swinkels, “The Moral Hazard Problem with High Stakes,” *Journal of Economic Theory*, Vol.187, 2020, 105032.

④ Michael C. Jensen and William H. Meckling, “Theory of the Firm: Managerial Behavior, Agency Costs and Ownership Structure,” *Journal of Financial Economics*, Vol.3, No.4, 1976, pp. 305-360.

用权授予平台企业,但同时也设定了明确的技术使用边界和安全标准,其核心关注点是保障技术成果的安全性与技术产品商业化后市场价值的最大化。这种委托关系往往通过正式的技术许可协议和战略合作伙伴关系来规范,构成了人工智能产业链中的重要价值传递机制。例如,英伟达(NVIDIA)作为人工智能计算技术的领先企业,向百度授权使用其图形处理器(GPU)和深度学习技术,百度则负责将这些底层技术整合到自动驾驶平台和云计算产品中进行商业化应用,双方通过这种分工合作共享技术创新的市场价值。^①

作为人工智能治理的另一主要委托方,市民社会更关注数据隐私保护、算法公平、内容审查等公共价值议题。不同于企业间基于商业契约的委托关系,市民社会对平台企业和国家的委托更多体现为一种社会契约,通过倡议、反馈等方式对国家和平台企业的政策施加影响,以此维护其所代表的公共利益。市民社会在人工智能治理中的能动性体现在对数据和内容治理的需求表达上,尤其在数据安全和算法透明度方面能够对平台企业形成有效约束。例如,Zoom 会议软件在 2023 年 3 月的隐私条款更新中曾增加一项内容,授权该公司自由获取用户的语音、视频等数据,将其投入机器学习的系统中。这一条款引发了大批量的用户抗议,致使其最终不得不撤销这一诉求,重新声明其不会在未经同意的情况下使用用户内容来训练人工智能。^② 用户和市民社会作为委托方的能动性作用不仅仅是对技术和平台行为的反向约束,也是塑造治理框架的一种路径,使平台企业在履行代理职责时,能够更好地满足公众的核心权益。

在人工智能治理复杂的“委托—代理”系统中,国家具有独特的核心地位,是系统的“双职中枢”。首先,“双职”特征体现了国家职能结构的双重性,即国家兼具重要的委托方和代理方两种角色。作为委托方,国家通过制定政策和提供资源,对人工智能技术的发展方向和治理框架进行引导。国家在发布的人工智能战略中,通常明确了优先发展领域,包括数据、算法治理,产业与行业

^① Ken Brown, “NVIDIA, Baidu Announce Partnership to Accelerate AI,” NVIDIA Newsroom, <http://nvidianews.nvidia.com/news/nvidia-baidu-announce-partnership-to-accelerate-ai>, 2025-03-05.

^② Smita Hashim, “How Zoom’s Terms of Service and Practices Apply to AI Features,” Zoom, February 7, 2024, <https://www.zoom.com/en/blog/zooms-term-service-ai/>, 2025-02-27.

能力的提升,以及治理机制建设、立法互操作性和安全管理等内容,这体现了国家对人工智能发展的战略性支持。^①同时,国家作为代理方,承担着推动公共利益与保障公共安全的职责,通过立法监管、行政执法和司法裁决等机制在保障数据隐私、算法透明度和内容合规等方面对企业进行监管,以回应社会需求。这种“双职”角色彰显了国家在人工智能治理中的多重利益协调能力,也意味着国家需要在技术创新与风险防控、产业效益与社会责任、国家安全与公民权益等多重目标之间寻求最优平衡。

其次,“中枢”指的是国家在整个人工智能治理系统中的节点定位,意味着除了研发与平台企业之间及用户—平台之间直接的商业性关系以外,其他治理链条均需经由国家这一核心关节点进行连接与传导:在国际治理框架下,国家作为全球规范的执行者,通过采纳和落实国际规则,将这些全球治理规范与本国的利益结合,以推动人工智能伦理标准、数据保护框架等的本土化实施。这一职能使国家成为连接国内外规范的枢纽,在本国政策与全球治理需求之间寻求平衡。^②此外,在国内治理方面,国家通过政策引导和法规实施,将市民社会的利益诉求传递到企业层面,促使企业在技术和内容治理方面更好地服务于公众利益。^③在这一过程中,国家能够作为社会需求与企业行为之间的桥梁,为维护公众利益提供制度保障。

最后,国家作为“双职中枢”并不是职能结构和节点定位的简单叠加,而是二者的有机整合,使其能够作为多层次、多维度治理网络的协调者,通过有效整合国际规则、企业行为准则和社会需求,形成一个系统化的政策框架,以保证人工智能多主体协作治理的合法性、有效性和可持续性。

基于对“委托—代理”关系网络的系统分析,本文提出以下研究假设:人工智能治理网络的有效构建和多主体协同模式的成功运行,在很大程度上依赖于国家在委托—代理系统中的核心地位和双重职能的有效发挥。然而,在具

① James S. Denford, et al., “A Cluster Analysis of National AI Strategies,” Brookings, December 13, 2023, <https://www.brookings.edu/articles/a-cluster-analysis-of-national-ai-strategies/>, 2025-03-05.

② [美]玛莎·芬尼莫尔:《国际社会中的国家利益》,上海人民出版社2012年版,第40—60页。

③ 周辉:《网络平台治理的理想类型与善治:以政府与平台企业间关系为视角》,《法学杂志》2020年第9期,第24—36页。

体治理实践中,国家在协调复杂的委托—代理关系时面临多方面挑战。

二、国家“双职中枢”角色的实践困境

在人工智能治理的复杂“委托—代理”系统中,国家的“双职中枢”角色虽然在理论上印证了其合理性,但在实践中面临诸多挑战。尽管在经济学、管理学和政治学的框架内,学者们已经对“委托—代理”关系中的问题进行了较为充分和系统的讨论,识别了由于“授权”“信息”和“契约”而建立起的成本、道德风险和逆向选择问题。^①然而,当问题聚焦于人工智能治理时,国家角色显现出新的复杂性。

首先,这种复杂性源于主体的特殊性。不同于传统市场关系中的委托—代理关系,国家和政府具备制度设计和技术干预的能力,可以通过多种手段来解决信息不对称问题。^②这一特点使得国家不仅依赖经济上的激励机制,还可以通过法制和技术手段实现对代理方的有效控制。正是这种多维度的能力结构,使国家在人工智能治理中超越了单纯的理性经济考虑,成为具有战略性资源配置能力的核心主体。

其次,人工智能治理中的客体,即人工智能技术本身的特殊性也决定了国家在治理中需要采取特殊的应对策略。委托—代理关系中的激励不一致问题在多个技术领域都有体现。例如,在能源领域,供需双方在节能投资上就存在这一问题:代理方往往因高昂的前期成本而缺乏投资动力。^③这种困境在需要前期高投入的技术领域尤为普遍。然而,人工智能技术具有不确定性、涌现性和战略变革性,这意味着国家在处理委托—代理关系时需要超越传统的管理工具,不仅要解决传统的激励问题,还需要在治理框架内应对技术快速迭代带

① Michael C. Jensen and William H. Meckling, “Theory of the Firm: Managerial Behavior, Agency Costs and Ownership Structure,” *Journal of Financial Economics*, Vol.3, No.4, 1976, pp. 305-360.

② 王金秀:《“政府式”委托代理理论模型的构建》,《管理世界》2002年第1期,第139—140页。

③ Adam B. Jaffe and Robert N. Stavins, “The Energy-Efficiency Gap What Does It Mean?” *Energy Policy*, Vol.22, No.10, 1994, pp. 804-810.

来的新挑战,并在不确定性中预见并引导技术发展方向。^①这使得人工智能治理成为一项极具挑战性的国家任务。

最后,人工智能治理的特殊性还体现在其效用评估标准上。传统基于理性经济关系的委托—代理关系效用衡量方式难以完全适用于人工智能治理场景,即国家需要闭环性思考“代理失灵”和“效用失败”的情况。作为能够广泛、持续产生风险的特殊技术,人工智能技术一旦出现治理失误,其后果可能是深远且难以逆转的。^②因此,国家不仅要关注代理方的合规性与执行效果,还需引入多层次的反馈和问责体系来降低信息不对称带来的偏差,并以此在复杂的利益协调中维持整体效用的平衡。

综合来看,国家在人工智能治理中“双职中枢”角色面临着三个关键问题:代理成本的复杂化、多方目标的协调问题及代理人效用失败后的救济机制,以下将针对这三个关键问题进行深入分析。

(一) “双职”结构与代理成本的复杂化

在人工智能治理中,国家的“双职”角色不仅显著放大了传统委托—代理关系中的代理成本,更因技术的特殊属性产生了新型成本结构:作为委托方和代理方的双重角色,国家不仅承担监督和控制责任,还需响应公众、产业界和国际社会的多元需求。在委托—代理关系中,代理成本是最具主导性的问题,通常被分为监督成本(Monitoring Cost)、约束成本(Bonding Cost)和因机会成本而产生的剩余损失(Residual Cost)。^③而在人工智能治理框架下,这些成本均呈现出新的技术特征,同时还出现了以往治理领域中不突出的时间成本问题。

首先,作为委托方,国家在人工智能治理中的监督成本会更加高昂。这一

^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,第28—34页。

^② Yoshua Bengio, “Reasoning through Arguments against Taking AI Safety Seriously,” July 9, 2024, <https://yoshuabengio.org/2024/07/09/reasoning-through-arguments-against-taking-ai-safety-seriously/>, 2025-02-27.

^③ Michael C. Jensen and William H. Meckling, “Theory of the Firm: Managerial Behavior, Agency Costs and Ownership Structure,” *Journal of Financial Economics*, Vol.3, No.4, 1976, pp. 305-360.

成本增长直接源于人工智能技术的专业性和发展速度。因此,国家作为委托方需投入大量资源进行精细化的监督,推动代理方行为的安全性和伦理合规性。^① 然而,人工智能领域技术壁垒的存在产生了结构性的信息鸿沟——技术研发企业和智能平台掌握关键技术细节及其发展路径,而作为监管主体的国家却处于信息劣势地位。^② 这种信息差距增大了国家在监督过程中维持治理透明度的难度,使其不仅需要高效的监督机制,还要面对代理方行为不可预测性带来的潜在风险。此外,行为透明度的缺失进一步加剧了国家在治理中的成本。代理方在技术开发和应用中的行为通常难以被国家完全掌握:一方面,科技企业在追求市场竞争优势过程中,往往选择性披露技术细节,甚至在隐私保护和责任归属等问题上可能规避部分伦理责任;^③ 另一方面,国际组织在推动全球人工智能治理规范时,可能过度强调普遍性原则而忽视特定国家的技术主权需求。为应对代理方行为的不可预测性,国家还需在治理框架中引入灵活调整机制,进一步加大了监督资源的需求。^④ 这种特殊的监督成本是传统技术领域治理所未曾面临的挑战。

其次,作为代理方,国家在人工智能中的约束成本呈现出明显的技术依赖性。在履行代理职责时,国家不仅需向公民和社会各界证明其在人工智能治理中的公信力,还需在技术快速演进过程中维持政策的连续性和适应性,使得其治理行为能够有效代表公共利益并保障社会的安全和福祉。在实践层面,这种约束表现为三方面的资源投入:第一,国家在治理框架中往往主动设定高标准的伦理规范和操作流程,并不断更新法律和政策以应对人工智能技术的快速发展,进而帮助国家在治理中履行对公民的承诺,也构成了其作为代理方

① 庞金友:《AI 治理:人工智能时代的秩序困境与治理原则》,《人民论坛·学术前沿》2018 年第 10 期,第 6—17 页。

② Daniel Byman, “Democratic Governments Are Failing to Leverage Technology Companies,” *Lawfare*, September 23, 2023, <https://www.lawfaremedia.org/article/democratic-governments-are-failing-leverage-technology-companies>, 2025-03-06.

③ Olga Solovyeva, “Why Tech Companies Can No Longer Ignore Their Role in Shaping Politics and Society,” *Global Voices Advox*, April 19, 2023, <https://advox.globalvoices.org/2023/04/19/why-tech-companies-can-no-longer-ignore-their-role-in-shaping-politics-and-society/>, 2025-02-28.

④ Colin van Noordt and Gianluca Misuraca, “Artificial Intelligence for the Public Sector: Results of Landscaping the Use of AI in Government across the European Union,” *Government Information Quarterly*, Vol.39, No.3, 2022, 101714.

的自我约束。中国发布的《新一代人工智能伦理规范》和美国的《人工智能权利法案蓝图》等都属于这一方面的相关措施；第二，为回应产业界对技术安全性、创新空间和公平竞争环境的需求，国家往往投入大量资源进行产业建设。^①例如，美国众议院拨款委员会自2023年6月开始重点推进立法，将人工智能纳入越来越多的项目中，推动政策与产业发展的正向互动；^②第三，国家的约束成本还包括在国际治理框架中维护信誉的投入。由于国家处于联通国际治理和国内治理的中枢位置，因而，要保证其治理政策和行为符合国际准则和标准，以维持国家在全球人工智能治理中的地位和信誉。然而，更为关键的问题在于，技术资源和信息的不对称迫使国家在很多情况下依赖前序环节中代理方的判断，而这种依赖并未得到相应的代理能力支持。^③例如，在评估大型语言模型长期的社会影响时，由于多数国家的国家级测评部门还未建立，国家往往需要依赖技术开发者提供的评估数据和风险分析，这种依赖不仅增加了政策决策的不确定性，还使国家在技术治理路径选择上处于被动地位，难以有效掌控治理的方向。

此外，剩余损失则源于国家在“双职”角色下不可消除的机会成本。即使国家构建了全面的监督框架和约束机制，代理方与国家之间的战略差异仍可能导致治理目标的偏离。这一问题在人工智能领域尤为突出，因为该技术具有多路径发展可能性和广泛的应用场景。例如，技术研发企业可能基于商业逻辑优先发展特定技术方向（如强化消费者行为预测的算法），而非国家战略优先的方向（如支持公共服务的人工智能应用）；国际组织可能推动以西方价值观为基础的全球人工智能伦理框架，而非兼顾不同发展阶段国家需求的多元治理模式。这种战略目标偏差反映了国家作为委托方在人工智能技术演进路径上的掌控局限，也体现了国家在代理角色中协调不同利益诉求的结构性

① 贾怡炜、戚湧、孙明汉：《新一代人工智能产业政策及治理体系现代化研究》，《中国科技论坛》2023年第12期，第51—60页。

② Peter Kasperowicz, “Congress Pushes Aggressive Use of AI in the Federal Government, Says AI ‘under-Utilized’ in Agencies,” Fox News, <https://www.foxnews.com/politics/congress-pushes-aggressive-ai-federal-government-under-utilized-agencies>, 2025-03-01.

③ Raquel Jorge Ricart, “Big Tech Companies and States: Policy or Politics?” Elcano Royal Institute, March 2, 2020, <https://www.realinstitutoelcano.org/en/blog/big-tech-companies-and-states-policy-or-politics/>, 2025-03-05.

挑战。

最后,人工智能治理对时间节点的敏感性带来了额外的时间成本,这是该领域中特有的新型代理成本。与其他技术领域不同,人工智能发展呈现出明显的“能力跃升”和“应用扩散”特征,使得治理活动面临严格的时间窗口约束。如果国家无法在这些关键的技术创新节点(如多模态大模型的突破)或技术扩散节点(如开源模型大规模商业化初期)上及时完成委托任务的分配和监督,便可能错过最佳治理时机,导致治理效果的削弱甚至失效。这种延误现象有可能引发“克林格里奇困境”(Collingridge Dilemma),即国家在快速发展的技术进程中未能及时响应而导致治理控制力下降的局面。^① 因此,时间成本在人工智能治理中成为必须关注的重要代理成本问题,不仅是监督效率的体现,也关乎治理的有效性和及时性。

(二) “中枢”定位与协调性难题

在人工智能治理中,国家作为“中枢”不仅需协调多方利益诉求,还需在国内外治理体系之间实现政策和资源的平衡。人工智能既是关乎国家竞争力的战略资源,又是影响社会各领域的普惠技术,这就要求国家将治理目标与国家安全、社会福祉、产业发展等多重战略目标相结合。然而,这一整合过程面临极高的协调难度,主要表现在目标协调、利益平衡、层次整合和资源优化四个关键方面。

第一,多方目标的差异性使国家在治理中面临适应性调节的压力。不同代理方基于其组织性质、核心利益和发展战略,形成了各自独特的治理期待:国际组织普遍强调技术的普惠性和全球秩序;研发企业追求技术创新与市场回报;平台企业关注用户体验与内容生态;市民社会则优先考虑隐私保护与算法公正。这种目标多元性直接考验着国家治理框架的适应能力和包容性。在此情境下,如果国家无法建立足够灵活的多层次治理机制,就会陷入“单一标

^① Audley Genus and Andy Stirling, “Collingridge and the Dilemma of Control: Towards Responsible and Accountable Innovation,” *Research Policy*, Vol.47, No.1, 2018, pp. 61-69.

准难以普适”的困境,无法有效回应多元治理需求。^①

第二,多方代理人的内在冲突性进一步加剧了国家的利益平衡的难度。在资源分配、数据使用和技术标准的设定方面,各方代理人可能存在根本性分歧。例如,科技企业和用户对数据使用的透明度需求可能相互矛盾,企业希望利用数据驱动创新,而市民社会则可能要求更严格的数据保护措施。这种差异性带来的冲突直接加剧了国家的协调难度。资源分配的失衡可能导致某些代理方退出合作或消极履行职责,政策执行的不一致性则可能削弱整体治理框架的权威性,利益协调的失败甚至可能引发社会信任危机。因此,国家需要超越传统的自上而下的监管模式,构建更具包容性的多元协商机制,通过制度设计创造各方代理人利益的交集点,以实现人工智能治理中的利益平衡与共识建构。

第三,跨层次的双层博弈使国家的协调任务更加复杂。作为国内外人工智能治理的枢纽,国家需在国际合作与国内需求之间找到平衡,以协调跨层次的治理目标。因此,国家本身就面临着“双层博弈”的困难,要寻找国际利益和国内利益团体之间的交集以推动政策的执行。^② 例如,在国际治理中,国家可能需遵循国际准则,如数据共享、跨国监管和人工智能伦理标准,但国内治理中的具体需求和利益诉求可能要求国家在这些方面保持灵活。国家既要保护国内利益,又需在国际场合中展现出合作姿态,以试图在全球治理体系中占据主动地位。因此,国家在协调国内外政策时需投入额外的资源,以平衡不同层次治理中的多方需求。

第四,资源调配压力进一步加剧了这一协调难题。人工智能治理涉及技术研发、标准制定、风险评估、监管执法等多个环节,每个环节都需要大量的专业人才、技术设施和制度资源作为支撑。在资源总量有限的情况下,国家需在不同治理主体、不同治理领域之间进行战略性配置,以实现治理效能的最大化。如果资源分配失衡,不仅可能降低治理效率,还可能导致治理主体之间产

① Thomas A. Hemphill, “Regulating Nanomaterials: A Case for Hybrid Governance,” *Bulletin of Science, Technology & Society*, Vol.36, No.4, 2016, pp. 219-228.

② Robert D. Putnam, “Diplomacy and Domestic Politics: The Logic of Two-Level Games,” *International Organization*, Vol.42, No.3, 1988, pp. 427-460.

生不信任,进而引发公众对国家治理能力的质疑。为此,国家在资源调配过程中需设计有效的分配机制,以支持人工智能技术的有序发展,同时保障社会利益和公共价值的实现。

(三) “双职中枢”角色与治理韧性问题

从人工智能治理的整体系统来看,国家作为“双职中枢”须具备高水平的治理韧性,以应对代理方可能出现的“代理失灵”及其带来的连锁反应。代理方的代理失灵指的是代理方因技术偏差、行为失误或其他主观或客观原因未能达成国家设定的治理目标。^① 由于人工智能技术在数据隐私、算法公正性和社会安全等敏感领域具有不可逆的影响,国家在此背景下不仅要维持治理的稳定性,还应具备有效的救济机制,以保障治理系统的持续性和恢复力。

第一,治理系统的脆弱性是效用失败带来的直接挑战。人工智能技术的复杂性和快速迭代性使得代理方在治理任务中承担高度的技术和伦理责任,但一旦代理方出现失误,治理系统的薄弱环节将被放大。特别是在人工智能研发企业在人工智能设计安全等关键场景的任务中,代理方的失误可能引发有害算法等问题不可控的风险扩散,影响到其他治理领域的稳定性。^② 在这一背景下,国家须具备应急响应和风险控制能力,是其能够在代理方效用失效时能够迅速调整治理手段,防止治理系统出现系统性失效。这要求国家在治理体系中设立灵活的风险预警和应急机制,从而在关键节点上保有应对突发事件的主动权。

第二,社会信任的维系是国家在治理韧性中的核心任务。代理方在数据隐私、算法公平性等领域未能实现治理目标,容易引发公众对国家治理的信任危机,进而影响到社会对人工智能治理的支持度。^③ 信任危机的加剧不仅削弱

① 尹润瀚、钟裕民:《慈善组织运行中委托代理失灵及其矫正之策》,《南京社会科学》2023 年第 6 期,第 73—80 页。

② 康骁:《行政法如何应对生成式人工智能:基于算法、训练数据和内容的考察》,《云南社会科学》2024 年第 4 期,第 80—88 页。

③ 李思艺:《迈入技术与信任相融合的数字治理时代:加拿大数字政府建设的启示》,《情报理论与实践》2022 年第 1 期,第 205—212 页。

了公众对国家和代理方的支持,还可能使公众质疑国家在人工智能治理中的权威性和公正性。国家作为代理方需承担赢得和维持社会信任的责任,这包括确保代理方在数据安全和伦理合规方面的透明度,并在出现失误时快速而透明地采取补救措施,以恢复公众对治理体系的信任度。为此,国家需在治理体系中建立清晰的问责机制,保证在出现问题时,公众能够得到及时的解释和行动响应,以维持治理的公信力。

第三,创新与风险控制的平衡是实现治理韧性的关键。人工智能的涌现性和不确定性要求国家在治理过程中既要鼓励创新,又需要控制风险。代理方在推动技术突破的同时也面临潜在的风险,国家需通过动态调整机制实现对创新的支持,同时在潜在的意外影响中保持治理的稳定性。^① 国家在此背景下需调整和创新治理框架,推动在代理方的创新活动和公共利益间建立平衡,以便在出现风险时能够及时调整,降低风险带来的负面影响。

第四,建立系统性救济机制是国家治理韧性的长期保障。鉴于人工智能技术可能产生不可逆的影响,一旦代理方未能履行其职责,国家须具备针对性的救济方案,以将风险控制在最低范围内。系统性救济机制不仅包括事后补救措施,还涉及在事前和事中建立监测与干预机制,以减少代理方效用失效对治理目标的冲击,使国家在快速变化的技术背景中拥有稳健的治理基础,进而保证人工智能治理体系的可持续性。

三、策略对比:美、英、中三国在人工智能治理中的不同选择

国家在人工智能治理系统中的关键性角色和现实性挑战,共同提出了从国家层面设计破局思路的必要性。在前文关于国家“双职中枢”角色及其面临挑战的理论分析基础上,本节将通过对美国、英国和中国三个代表性国家人工

^① The White House, “Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence,” October 30, 2023, <https://www.whitehouse.gov/briefing-room/presidential-actions/2023/10/30/executive-order-on-the-safe-secure-and-trustworthy-development-and-use-of-artificial-intelligence/>, 2025-03-01.

智能治理实践的考察,验证该理论框架的解释力,并探究不同国家在应对共同挑战时采取的差异化策略。这三个国家虽然在政治制度、技术发展阶段和战略目标上存在差异,但在人工智能治理的委托—代理关系网络中,均需承担“双职中枢”角色,面临相似的代理成本、协调性和韧性挑战。然而,由于各国特有的制度背景、治理传统和战略优先级,它们在具体治理实践中形成了各具特色的应对策略。美国的分布式监管、英国的设计安全与灵活策略、中国的统筹与稳健策略,代表了应对相同理论挑战的三种典型路径,为理解“双职中枢”角色在不同国家语境中的多样化实现形式提供了典型案例。

(一) 美国的人工智能治理:分布式监管

在应对委托—代理关系问题上,美国的人工智能治理框架采取独特的分布式监管方法,以实现削减代理成本、增加协调性和韧性的目标。分布式监管的核心在于多部门分工、信息透明及风险控制,目的是在技术发展的同时减少治理障碍。其方法包括广泛的部门协作、风险导向的政策指导和灵活的监管方式,适应了技术变化和主体协同带来的治理挑战。

首先,美国的分布式监管模式从三个维度构建了系统性解决方案,这些解决方案共同作用,显著降低了国家在“双职中枢”角色中面临的代理成本:第一,在应对人工智能治理中的信息不对称问题时,美国通过强调透明性和制定多层次的信息共享机制来增加治理可控性。2019 年发布的《保持美国在人工智能领域的领导地位的行政命令》(EO 13859)和 2023 年的《关于安全、可靠和可信开发及使用人工智能的行政命令》(EO 14110)要求联邦机构加强对算法使用的透明度,从而推动科技企业等代理人在数据隐私、算法偏见等方面的操作能够接受监督。^①《关于安全、可靠和可信开发及使用人工智能的行政命令》要求 50 余个联邦机构完成 100 多项行动,包括任命首席人工智能官员,并要求各机构建立人工智能使用策略,以帮助政府在关键领域减少信息不对称。此

^① The White House, “Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence,” October 30, 2023, <https://www.whitehouse.gov/briefing-room/presidential-actions/2023/10/30/executive-order-on-the-safe-secure-and-trustworthy-development-and-use-of-artificial-intelligence/>, 2025-03-01.

外,美国科学技术政策办公室(OSTP)发布的《国家人工智能研发战略计划》要求各联邦机构采用透明度和开放数据的政策,并加强各部门之间的技术信息交流。^① 这个分布式的治理框架在增强政府对代理方了解的同时,还支持人工智能开发过程中的跨部门信息共享和合作,缓解了由于技术复杂性带来的信息不对称问题。第二,美国的分布式结构允许各联邦机构在其领域内制定适合自身职能的人工智能政策,从而降低了约束成本和机会成本。例如,美国联邦贸易委员会(FTC)主要负责人工智能技术在消费者保护和反垄断方面的监管。^② 而美国食品药品监督管理局(FDA)则专注于医疗健康领域中的人工智能应用。^③ 这种专业化分工不仅提高了监管效率,还避免了单一机构承担全面监管责任可能导致的资源过度消耗和知识盲点,直接缓解了国家作为代理方的约束成本过高问题。第三,为应对国家在人工智能委托—代理系统中的能力不对称性,美国司法系统作为行政措施的补充,有助于分散治理成本。2024 年底,美国司法部表示,已向审理谷歌垄断案的法官提出可能要求谷歌母公司字母表(Alphabet)剥离部分业务,包括 Chrome 浏览器和安卓操作系统,以遏制其市场主导地位,特别是在人工智能相关业务上的潜在垄断优势。这类司法干预不仅削弱了垄断可能带来的负面影响,还为美国人工智能治理的行政系统提供了治理能力的补充,试图解决技术资源和信息的不对称迫使国家依赖代理方判断的挑战。^④

其次,面对“中枢”定位所带来的多方目标协调难题,美国对内通过跨部门协作和风险为基础的政策框架实现了代理方的多元目标对接,而对外则凭借其盟伴体系承接和塑造全球性利益。

① Select Committee on Artificial Intelligence, “The National Artificial Intelligence Research and Development Strategic Plan,” May 2023, <https://www.nitrd.gov/pubs/National-Artificial-Intelligence-Research-and-Development-Strategic-Plan-2023-Update.pdf>, 2025-03-06.

② Federal Trade Commission, “FTC Announces Crackdown on Deceptive AI Claims and Schemes,” September 25, 2024, <https://www.ftc.gov/news-events/news/press-releases/2024/09/ftc-announces-crackdown-deceptive-ai-claims-schemes>, 2025-03-01.

③ Robert M. Califf, “Harnessing the Potential of Artificial Intelligence,” US Food & Drug Administration, 2024, <http://fda.gov/news-events/fda-voices/harnessing-potential-artificial-intelligence>, 2025-03-06.

④ 庞昕熠:《美司法部为反垄断提议拆分谷歌部分业务》,《经济参考报》2024 年 10 月 11 日,第 4 版, <http://jjckb.xinhuanet.com/20241011/1fd87aa4308b459f968ab381f401e885/c.html>, 2025-03-01.

在国内层面,美国通过建立包容多元主体的协商机制,缓解了“多方目标的差异性”带来的协调压力。2020年,美国推动搭建国家人工智能计划的机构系统,利用国家人工智能咨询委员会(NAIAC)等多方参与的机制,吸纳了学术界、产业界和民间社会等领域的代表,直接向总统和国会提供关于人工智能发展的政策建议。这种多元参与的机制使得美国的人工智能政策不仅具有国家主导性,还能够反映多方利益的平衡,从而进一步减少治理冲突,实现高效协调。^① 在国际层面,美国利用其强大的盟友网络,通过美欧贸易与技术委员会(TTC)、美日印澳四边机制(QUAD)以及奥库斯(AUKUS)同盟等平台,将国内治理需求与国际规则制定有机结合从而实现国内外政策的协调一致,^②即通过塑造国际规则来减少国内治理成本。

最后,针对“双职中枢”角色面临的治理韧性挑战,美国设计了一套同样遵循分布式逻辑的风险控制与快速响应机制,通过分散风险评估和责任分配,增强了整体治理体系的恢复力。一方面,美国建立了多层次的安全评估和风险预警系统,直接应对理论框架中提出的“治理系统脆弱性”问题。例如,美国国家标准与技术研究院(NIST)发布的《人工智能风险管理框架》(AI-RMF)聚焦于技术全生命周期的设计安全与韧性建设,为各行业提供了系统性的风险评估工具,构建了从风险识别到风险应对的完整流程;^③另一方面,美国通过建立分布式的审核和评估机制,强化了“社会信任的维系”能力。美国人工智能专责委员会负责定期为联邦部门制定并发布人工智能战略计划,明确指导方针和评估标准,为各机构的人工智能活动设定目标和优先事项,从而确保美国人工智能治理体系的整体一致性与响应力。^④ 这种分权与协调并重的设计,不仅提高了治理的透明度和可问责性,还在风险事件发生时提供了明确的责任划

① 张璐瑶:《敏捷治理与人工智能治理机制创新》,上海国际问题研究院硕士论文,2023年。

② 蔡翠红、李煜华:《美日印澳四边机制网络安全合作探析》,《现代国际关系》2024年第5期,第24—42、140页。

③ Elham Tabassi, “Artificial Intelligence Risk Management Framework (AI RMF 1.0),” NIST AI 100-1, National Institute of Standards and Technology (U. S.), Gaithersburg, MD, January 26, 2023, NIST AI 100-1.

④ National Artificial Intelligence Initiative Office(US), “Charter of the Select Committee on Artificial Intelligence,” January 5, 2021, <https://trumpwhitehouse.archives.gov/wp-content/uploads/2021/01/Charter-Select-Committee-on-AI-Jan-2021-posted.pdf>, 2025-03-01.

分和应对路径,直接回应了理论框架中关于“代理方在出现失误时快速而透明地采取补救措施”的要求。

(二) 英国的人工智能治理:设计安全与灵活策略

英国在人工智能治理中采取了以“设计安全、灵活应对”为核心的治理框架,直接针对国家“双职中枢”角色面临的信息不对称、代理成本高昂及多方协调等挑战。这一框架体现了英国在行政传统和全球治理影响力背景下对“双职中枢”实现路径的重新定义。

第一,英国高度重视人工智能治理中的设计安全,在控制代理成本方面采取了多层次的透明和安全措施,这种“源头治理”策略直接改变了传统委托—代理关系中的信息与权力结构。英国的设计安全机制主要采取两方面措施:一方面,为提高治理的透明性和降低信息不对称风险,英国推行了算法透明性记录标准(Algorithmic Transparency Recording Standard)。该标准要求公共部门在应用算法时公开隐私、准确性和公平性等关键信息,使得政府能够更好地监督代理方的算法操作。^① 透明性标准的建设通过提高信息的可及性,使国家能够以较低成本监督代理方行为;另一方面,英国通过“设计即安全”,将安全要求内置于技术发展的早期阶段,从而从源头上降低监督成本。^② 例如,英国引入了人工智能的“沙盒”机制作为技术创新的安全试验环境,使代理方能够在受控条件下测试新技术的效果。^③ 这一机制不仅降低了国家在监督上的资源消耗,还鼓励代理人在符合安全标准的情况下实现创新,解决了“人工智能技术的专业性和发展速度”带来的高昂监督成本问题,将被动监督转变为主动引导。

① Department for Science, Innovation & Technology(UK), “Algorithmic Transparency Recording Standard - Guidance for Public Sector Bodies,” January 5, 2023, <https://www.gov.uk/government/publications/guidance-for-organisations-using-the-algorithmic-transparency-recording-standard/algorithmic-transparency-recording-standard-guidance-for-public-sector-bodies>, 2025-03-01.

② UK Government Security, “Secure by Design Principles,” <https://www.security.gov.uk/policy-and-guidance/secure-by-design/principles/>, 2025-03-01.

③ Ryan Morrison, “Government Backs UK AI Regulatory Sandbox,” *Tech Monitor*, March 16, 2023, <https://www.techmonitor.ai/digital-economy/ai-and-automation/government-backs-ai-regulatory-sandbox>, 2025-03-01.

第二,面对“中枢”定位带来的多方目标挑战,英国通过建立系统性治理框架和平衡激励机制,促使代理方目标与国家治理方向的一致性。首先,英国在国内采取灵活的分层治理框架,其科学、创新和技术部(DSIT)设立了人工智能治理的五项原则(安全性、透明性、公平性、问责性和救济性),为各部门的人工智能治理提供统一的准则。^①但在具体实施上允许各代理方在自己的领域中根据需求灵活调整,既减少了目标冲突,也为各机构在不同治理需求间建立了通用的标准和共识。其次,英国通过建立跨部门协作机制,试图解决和应对“多方代理人的内在冲突性”问题。英国通过建设数字监管合作论坛(Digital Regulation Cooperation Forum, DRCF)使其关键监管机构(如金融行为管理局 FCA 和通信办公室 Ofcom)得以共享信息、互通政策。^②作为跨机构的协作平台,数字监管合作论坛使英国在政策调整和快速应对技术变化时更加灵活,从而推动了跨部门的协调性和一致性,减少因政策不一引发的治理冲突。最后,在国际合作中,英国依托其主办的全球人工智能安全峰会,推动了全球人工智能安全标准和伦理准则的协调。作为英国的“王牌平台”,全球人工智能安全峰会汇集了来自各国的政府、企业、学术和社会组织的代表,围绕人工智能伦理、安全和发展等主题开展深入讨论。^③该峰会为英国与全球伙伴提供了独特的协商平台,使英国能够主动推动全球性的人工智能安全标准,并将其作为全球治理中的重要议题。

第三,针对“双职中枢”角色面临的治理韧性挑战,英国在优先从国家和政府层面关注人工智能安全的韧性建设,进而保证代理方在失效后能够迅速恢复并保持治理的连贯性。首先,英国通过人工智能安全研究所(AI Safety Institute)等专门机构为治理体系提供全面的风险评估和响应机制。人工智能安全研究所承担了风险预评估和故障救济的职能,在代理方失效后能迅速采取

① Department for Science, Innovation & Technology(UK), “Implementing the UK’s AI Regulatory Principles: Initial Guidance for Regulators,” <https://www.gov.uk/government/publications/implementing-the-uks-ai-regulatory-principles-initial-guidance-for-regulators/102aa401-60f6-46e8-95dc-b48150eba7dd>, 2025-03-01.

② Competition and Markets Authority(UK), “The Digital Regulation Cooperation Forum,” April 4, 2023, <https://www.gov.uk/government/collections/the-digital-regulation-cooperation-forum>, 2025-03-01.

③ GOV. UK, “About the AI Safety Summit 2023,” <https://www.gov.uk/government/topical-events/ai-safety-summit-2023/about>, 2025-03-01.

补救措施,以防止治理失效带来的系统性风险。^①前置性安全设计不仅为国家在治理失效时提供了恢复力,也为代理方的行为设立了清晰的安全基准,降低了治理失效的可能性。其次,英国采取了逐步推进的迭代性监管方式,通过收集和分析实践数据,不断调整政策,以实现快速响应和适应性救济,在治理体系失效时能迅速恢复秩序。最后,为进一步保障治理韧性,英国采用了非立法先行的灵活监管方式,即在法律框架确立前先进行政策试行,数字监管合作论坛的算法研究项目分析了算法如何影响竞争、隐私和消费者权益,从而帮助各机构制定一致的政策。在促使代理方的行为符合国家治理框架的同时,为未来立法提供了稳健的支持。

(三) 中国的人工智能治理:统筹与稳健策略

中国的人工智能治理框架通过中央主导和地方试点结合、标准化监管与多利益攸关方的协同合作,展现出“统筹与稳健”相结合的治理思路。中国在代理成本控制、目标协调和救济韧性建设等方面的政策,既保障了技术创新的灵活性,又兼顾了国家的安全与社会需求。

首先,针对理论框架中提出的监督成本和信息不对称挑战,中国构建了一套自上而下与自下而上相结合的系统化标准体系,通过整体规划和局部试验的有机统一,实现了代理成本的结构化降低。一方面,中国通过建立全面的监管标准体系,从制度层面减少了信息不对称问题。在“新一代人工智能”治理框架下,中国形成了涵盖技术开发、应用部署和市场监管的全面政策体系,《算法推荐管理规定》《深度合成管理规定》和《生成式人工智能服务管理办法》等政策规定不仅要求技术企业在数据使用、算法操作方面实现高透明度,还要求高风险技术需接受安全审查。这些标准为治理过程提供了清晰的合规要求,使国家能够有效监督代理方行为,降低了因信息不对称带来的监管难度;另一方面,中国的“试点型”监管策略通过地方试点增强了治理机制的灵活性和稳健性。以深圳、上海等地的人工智能试点项目为例,这些地区在算法透明性和

^① “The AI Safety Institute (AISID),” <https://www.aisi.gov.uk/>, 2025-03-01.

数据隐私方面进行测试,为政策全国推广提供了先行经验,减少了政策调整带来的资源投入,优化了国家的代理成本。^①

其次,面对“中枢”定位带来的多方目标协调挑战,中国在人工智能治理中的多方目标协调通过中央统筹和多层协同的方式进行,推动治理一致性并与代理方的自主创新需求相适应。在国内治理层面,中国通过明确的部门分工和层级责任机制,缓解了“多方目标差异性”带来的协调压力。党中央和国务院负责顶层战略规划,科技部和网信办等部门分别承担技术引导和内容监管职责,地方政府则负责创新试点和落地实施。这种清晰的责任分工进一步保证了政策从制定到执行的高效传导。在具体策略上,中国通过负面清单制度和分级管理体系,对不同风险等级的人工智能应用采取差异化监管。这一机制使国家可以优先集中资源监管高风险技术,同时赋予低风险技术更多创新空间。通过风险导向的监管方式,国家能够在技术创新与社会安全之间找到平衡,保证各代理方目标与国家战略的协同一致。^② 例如,针对敏感数据的人工智能应用,中国要求代理方进行严格的信息安全和数据隐私审查,而在其他创新领域则提供了较宽松的环境,在提升科技企业创新活力的同时保持了治理的灵活性。此外,中国在国际互动中,更多基于人类命运共同体的指导思想,不排斥任何组织和平台的治理进程,并积极参与联合国主导的各类人工智能治理进程。例如,2024年7月,联合国通过《中国提出的加强人工智能能力建设国际合作决议》,而中国的两名专家也在联合国人工智能高级别咨询机构中积极开展工作,推动发布了《治理人工智能,助力造福人类》等重要文件。^③

最后,针对“双职中枢”角色面临的治理韧性挑战,中国采取了以前瞻预防为主、多层次救济为辅的整体性风险管控策略,增强了治理体系应对风险的稳

① 崔亚东、杨华:《人工智能的地方立法研究》,《政府法制研究》2021年第5期。

② 贾开、薛澜:《人工智能伦理问题与安全风险治理的全球比较与中国实践》,《公共管理评论》2021年第1期,第122—134页。

③ Cuihong Cai, “China Shows Deep Learning in Shaping Global AI Governance,” East Asia Forum, September 13, 2024, <https://eastasiaforum.org/2024/09/13/chinas-network-connection-issues-in-global-ai-governance/>, 2025-03-01.

定性和恢复力。中国更关注人工智能带来的社会、个人和产业影响,并强调以人为本和未雨绸缪。中国的治理策略不仅关注技术本身的创新,更加注重其对社会、个人和产业的深远影响,强调以人为本的原则。在这方面,中国制定了一系列针对人工智能的伦理标准和治理框架,例如,《新一代人工智能伦理规范》和《数据安全法》不仅规定了企业在技术开发中的伦理责任,还要求在实施新技术前进行充分的风险评估。此外,中国高度重视社会信任维系,通过多元化的公众参与机制增强了治理的社会基础。针对人工智能可能带来的隐私侵犯、算法歧视等问题,中国建立了专门的投诉反馈渠道和多层次的监督机制,确保受影响的个人和组织能够表达诉求并获得响应。这种机制设计直接回应了“委托—代理”理论框架中关于“社会信任的维系是国家在治理韧性中的核心任务”的要求,通过增强公众参与促进了治理的开放性和包容性。^①最后,中国特别强调以人为本的价值导向,将技术发展置于更广泛的社会发展和人类福祉框架中考量。这种价值引导不仅体现在政策文件中,也渗透到具体的技术标准和应用规范中,确保技术发展的方向与社会需求和公共利益保持一致。

(四) 策略对比:不同国家“双职中枢”角色的实现路径

通过对美国、英国和中国三国人工智能治理实践的考察,可以发现:尽管作为“双职中枢”的国家在人工智能治理中同样面临一系列挑战,然而,各国基于自身制度环境和战略优先级,形成了各具特色的应对策略。

首先,在代理成本控制层面,美国注重通过分散治理责任的方式,利用广泛的跨部门协作和信息共享机制,降低治理过程中的代理成本,体现了其联邦制度特征和权力制衡理念。英国则以设计安全为核心,通过在设计初期介入标准设立,将透明性前置,使治理更具韧性和可控性,反映了其完善的行政传统。中国的治理特征则在基于“以人为本”和稳定的原则上推动治理

^① 中央网络安全和信息化委员会办公室:《中央网信办部署开展“清朗·整治违规开展互联网新闻信息服务”专项行动》,2024年10月3日,https://www.cac.gov.cn/2024-10/03/c_1729643985936583.htm, 2025-03-01。

框架的统筹应用,以政策一致性和分层监督确保各主体行为合规,将治理透明度与国家整体战略目标结合,形成整体规划与分层实施相结合的稳健监管模式。

其次,在应对协调性挑战方面,三国均建立了多层次的协调机制,但在参与主体的权重分配和决策流程上呈现出明显差异,反映了各自治理理念中对“多元参与”与“政府引导”的不同侧重。美国更强调市场主导与多元参与,并借助广泛的国内外多边协作来进行协调,赋予其全球标准影响力,充分体现了美国市场经济和全球盟友体系的双重优势。英国则强调灵活的治理框架,以“五项原则”协调国内目标,同时依托全球人工智能安全峰会在国际合作中强化其主导地位,更重视专业化协调平台和原则性指导的平衡作用。相比之下,中国的协调逻辑更注重国内治理的统筹一致,通过中央统筹规划和地方试点相结合的方法,确保多层次治理目标的有序对接与协调,体现了系统性规划和层级责任的统筹。

最后,在应对治理韧性挑战方面,美国偏向分布式问责和快速响应的模式,重视灵活性和治理系统的分散化,使其能够应对技术变化带来的不确定性。英国的策略则偏重于在韧性建设中逐步推进,注重在治理失效后的快速恢复与调整,以保持治理体系的连贯性和适应性。而中国的治理韧性建立在稳健的救济体系之上,通过系统化的伦理规范和层次清晰的治理机制保障治理的延续性,确保技术和社会影响之间的平衡。

三国策略共同呈现了多层次治理的多样性和适应性:美国的分散协作提升了政策灵活性,英国的安全设计为治理的稳定提供了基础,而中国的稳健统筹则确保了整体政策的连续性和有效性。通过比较三国的人工智能治理的不同举措,可以发现,在多重委托关系中,人工智能的全球治理可吸收多样化的机制和逻辑,进而破除委托—代理关系中的成本和道德困境,实现技术发展与公共利益的平衡与协同。

结 语

全球人工智能治理的多主体协调问题使得各国在技术监管、利益协调、韧性建设等方面面临复杂挑战。本文借助“委托—代理”理论框架,揭示了国家的双重角色定位和中枢节点功能,为复杂治理现象提供了结构性解释,并识别了国家在人工智能治理中面临的三大核心挑战——代理成本高、协调复杂和韧性不足,这些挑战构成了各国治理实践的共同着力点。

面对上述挑战,美国、英国和中国的治理路径虽各有侧重,却展现了共通性,即以创新和安全为核心,在兼顾多元利益的同时,构建具备持续适应性的治理机制。这三种策略分别从分布式治理、设计安全和统筹稳健的角度,形成了相互映衬又各具特色的治理模式。

通过对三国治理实践的比较分析,可以发现,“双职中枢”并非一个刚性的角色定位,而是一个具有弹性的理论框架,能够适应不同国家的制度环境和战略需求。国家在履行“双职中枢”角色时,可以根据自身禀赋和发展阶段,选择不同的实现路径:或分散权力降低单点成本,或前置设计减少安全风险,或系统规划强化整体协同,进而有效应对智能时代中的风险与挑战。