人工智能全球扩散中的“安全悖论”与规制

鲁传颖[1]　杨理伟[2]

【内容摘要】 以 DeepSeek、Qwen、LLaMA 为代表的开源大模型降低了技术准入门槛，在缩小数字鸿沟与建立主权人工智能方面呈现显著正向效应。然而，人工智能技术赋权效应与安全威胁呈现非线性增长态势，形成技术普惠与风险扩散的“安全悖论”，表现为风险内生与价值外生的技术本体矛盾、通用性与战略性的应用属性冲突以及技术迭代与治理响应的时滞困境。在此三重矛盾的复合作用下，各国陷入发展与安全的两难困境，片面追求开源模型的技术采纳将导致风险敞口的持续扩大，过度强调安全控制则可能丧失技术发展先机，最终形成技术扩散与安全威胁的循环。因此，后发国家亟须在认知“安全悖论”的基础上构建适应性治理框架，国际社会需要加强人工智能安全的全球治理，加大对后发国家的能力建设支持，以此实现人工智能技术扩散效益与安全治理的动态均衡。

【关 键 词】 人工智能　开源模型　技术扩散　安全悖论　风险治理

【作　　者】 1 鲁传颖，同济大学政治与国际关系学院副院长、教授，同济大学网络空间国际治理研究基地常务副主任；

2 杨理伟，同济大学政治与国际关系学院博士后。（上海 200092）

人工智能正在加速步入“普及”时代。随着技术开源浪潮的兴起，人工智能不再是少数科技强国的专属利器，而是以低成本、低门槛的方式加速扩散并广泛渗透至全球。各国政府、企业乃至个人用户纷纷拥抱开源人工智能技术，试图借助其强大的算力与智能能力，重塑医疗、教育、金融、治理等多个领域的运行范式。特别是在“全球南方”国家，人工智能更被寄予打破发展瓶颈、实现跨越式进步的厚望。以人工智能赋能的智慧农业、智能医疗、数字政务等新业态，正成为这些国家追赶技术前沿、缩小数字鸿沟的重要路径。然而，这一技术乐观主义的叙事背后，却潜藏着深层的安全风险焦虑。人工智能技术扩散的进程呈现出一种“安全悖论”，即技术赋权效

应与安全威胁呈现正相关态势。开源技术的低门槛特性在实现技术普惠的同时，也催生了新的脆弱性——缺乏充分安全治理能力的行为主体，在拥抱技术赋权的过程中，不可避免地暴露于复杂的内生风险之中。更为严峻的是，人工智能的广泛扩散可能引发新的风险社会形态。当人工智能深度嵌入关键基础设施、社会治理机制和经济运行体系时，技术系统的任何异常都可能触发跨域连锁反应，形成超越传统风险管理范畴的系统性危机。这种风险的本质不再是局部的、可控的技术故障，而是具有全球性、网络性特征的生存性风险。

当前关于人工智能风险的研究主要从技术伦理、算法治理、数据安全等技术本体层面，探讨算法偏见、隐私泄露、决策透明度等内生性风险问题。[①]同时，部分研究亦开始关注人工智能在就业、司法、教育等领域的社会化应用风险。[②]然而，既有研究多聚焦技术内生风险，未能深入揭示技术扩散过程中风险与价值实现之间的内在悖论。基于此，本文旨在从技术扩散视角切入，通过阐释人工智能扩散的“安全悖论”，揭示技术扩散过程中安全与发展之间的张力，为理解人工智能风险传导机制及构建适应性治理框架提供理论支撑。

扩散形态：人工智能扩散的“安全悖论”

人工智能扩散的“安全悖论”是指，技术扩散是人工智能价值实现的必要路径，但扩散本身即构成风险失控的充分条件。这一悖论的本质在于，不扩散则价值无法实现，扩散则风险必然放大。与以往通用技术不同，人工智能是首个在价值实现机制与风险放大机制上高度重合且相互强化的技术形态。这种内在矛盾在技术演化中呈现自我强化、不可逆转与难以调和的特征。人工智能扩散的“安全悖论”形态根植于技术本体、应用诉求与治理结构的三重矛盾中，共同塑造了技术扩散过程中价值与安全对立的复杂关系。

（一）风险内生与价值外生的本体矛盾

从技术本体的底层逻辑看，人工智能是风险内生、价值外在的技术形态，这一特征构成了其“安全悖论”的第一重根源。一方面，人工智能的价值实现高度依赖于外部激活。作为典型的数字技术，其技术潜能只有通过向多元应用场景扩散、被大规模用户持续使用，才能转化为现实的商业、社会乃至政治效用。另一方面，扩散本身却触发了内生风险的层级跃迁，原本在实验室环境中尚可追溯与局部控制的技术缺陷，在规模部署后演化为不可预测的系统性涌现风险，并进一步从一国之内的单域风险转变为全球范围内的跨域安全挑战。

人工智能内生安全（AI Safety）深嵌于其核心技术架构之中，集中体现为算法黑箱、结果涌现与模型幻觉等内在属性，并随每一次实际部署而被必然携带。这种风险并非技术发展初期的阶段性缺陷，而是当前技术范式的固有特性，且会在全球扩散过程中被持续放大。首先是算法黑箱问题。以深度学习为代表的现代人工智能系统，其决策逻辑依赖海量参数的非线性组合与复杂交互，内部推理过程对人类观察者高度不透明，难以解释与验证，从而显著削弱了人类对系统行为的认知能力与控制能力，放大了不可控风险。[③]其次是涌现能力的不可预测性。模型在规模扩展过程中可能自发展现出训练阶段并未显式设计的高阶能力，例如多步推理、代码生成和语义类比等。这类能力并非经由明确编程实现，而是从大规模数据学习中自发涌现，在带来性能跃升的同时，也引入了难以评估和提前识别的安全不确定性。[④]例如，大语言模型在特定诱导条件下可能绕过既有安全约束，生成恶意内容或执行“越狱攻击”（jailbreak attacks）以规避内置的安全护栏。[⑤]

最后是模型幻觉问题。由于大语言模型基于概率预测生成输出，其结果并不始终对应客观事实，错误信息往往以高度连贯、流畅自然的语言呈现，普通用户难以识别真伪，从而在现实应用中显著放大信息误导与决策失误的风险。[⑥]

与风险内生相对的是人工智能的价值外在属性。作为数字技术，人工智能并不具备自足性的价值生成机制，其经济与社会效益依赖于用户规模、使用场景与社会需求等外部条件的共同激活。正因如此，技术扩散并非可选路径，而是价值实现的必要前提。这一机制在现实中驱动科技企业不断加快人工智能的市场化与全球化进程，以抢占市场先机、积累用户数据、形成规模效应，并摊薄高昂的研发与算力投入成本。以 OpenAI 的 ChatGPT 为例，其在 2022 年 11 月发布后短时间内用户迅速突破 1 亿，刷新互联网产品增长纪录。这种爆发式扩张体现的，正是先占领市场、后完善治理的商业逻辑。

风险内生与价值外生的技术本体矛盾，在技术扩散中显现为扩散与风险之间的正反馈循环。一方面，技术扩散成为风险传导与放大的必要载体。人工智能系统在跨越地理与文化边界进行全球部署时，其内在的算法偏见、决策不透明性与潜在安全漏洞会通过技术网络被成倍放大并叠加。另一方面，扩散过程本身又不断催生新的风险形态。在不同社会语境与治理能力条件下，同一算法架构在发达国家可能表现为创新工具，而在治理能力薄弱的地区却易演变为社会操控、信息操纵甚至政治干预的工具，由此产生显著的“风险外溢”效应。更为关键的是，技术本体层面的矛盾在全球技术竞争结构中被进一步强化，并呈现出典型的“囚徒困境”式集体行动困局。若所有国家协同限制扩散速度以完善治理，全球整体风险可控；但任何单一行为体若单方面限制技术扩散，将面临技术竞争中的相对劣势。如欧盟虽然在《人工智能法案》中设置了严格的安全规制要求，但面对美中两国在人工智能领域的快速发展，欧盟委员会仍不得不提出一揽子精简、调整数字与科技监管的改革建议，以提升欧洲竞争力。[⑦]这一困境也迫使人工智能技术扩散中安全优先的理性选择在现实中往往让位于发展优先的战略考量，进而在全球范围内诱发治理标准下调与风险累积的“竞相逐底”现象。

（二）通用性与战略性的内在属性冲突

从技术应用的属性来看，人工智能是历史上首个同时兼具通用性与战略性的技术形态，这一双重属性构成了其“安全悖论”的第二重根源。一方面，人工智能作为典型的通用目的技术，其价值实现依赖于跨领域、大规模的开放扩散。正如电力、互联网等典型通用技术一样，只有通过广泛普及、多场景复用与规模效应积累，才能转化为全社会的生产力跃升与福祉增进。另一方面，人工智能同时具备显著的战略属性，其核心能力直接关乎国家权力投射、军事竞争优势与地缘政治博弈，因而必然受到类似核技术的严格管控与扩散限制。正是这种通用性与战略性的并存，使任何扩散策略都陷入两难困境。追求技术普惠必然削弱战略可控性，而强化战略管控则势必压缩通用技术的普惠潜能，二者共同构成技术扩散价值与扩散风险之间难以调和的矛盾。

人工智能的通用性是指其具备跨领域迁移、多场景复用与广泛适配的技术特征，其应用价值的实现高度依赖于全球范围内的开放扩散与规模化部署。人工智能技术呈现出显著的“一次开发、多场景复用”属性，单一算法框架可被迅速移植至金融决策、医疗诊断、交通管理、能源调度等异质性领域，在复制过程中不断降低边际成本，形成典型的规模效应。作为数字时代的关键基础设施，人工智能更被赋予了弥合数字鸿沟、促进全球发展均衡的期待。这种价值导向驱动技术供给应该开放共享、降低准入门槛，以实现全球范围内的技术民主化与发展包容性。

人工智能具有战略性属性，即其作为国家权力工具与战略竞争焦点的本质特征。人工智能已成为大国战略竞争的核心技术，其发展水平直接决定国家在全球权力结构中的相对地位。斯坦福大学人工智能研究所发布的《2025 人工智能指数报告》显示，2024 年美国机构共发布 40 个具有重大影响的基础模型，远超中国的 15 个及欧洲的 3 个。[⑧]技术供给的高度集中超越了单纯的经济竞争范畴，演化为涉及技术标准、制度主导权和国际治理能力的新型权力博弈。[⑨]与此同时，人工智能技术的开发往往兼具民用与军事双重目标。例如，用于自动驾驶的感知算法可转化为无人作战系统，用于自然语言处理的模型可转用于情报分析与信息战，用于医疗影像识别的技术可应用于军事目标识别。这种两用性使得任何民用技术扩散都潜藏军事能力扩散的可能，从而触发战略管控的必然诉求。

应用属性的通用性与战略性矛盾，最终表现为技术扩散逻辑的内在冲突与风险结构的层级跃迁。一方面，开放扩散与选择性管控之间的矛盾，使得任何单一国家若单方面强化战略安全限制，均可能在全球人工智能竞争中丧失市场规模、数据积累与技术迭代优势，从而陷入技术落后与战略被动的双重风险。另一方面，扩散过程本身又推动了风险性质的根本转变。技术从实验室阶段到国内部署，其风险尚可通过行政、法律与伦理框架相对有效地约束；但一旦扩散至全球，伴随应用场景的碎片化、使用主体多元化与治理能力差异化，原有风险会逐步引发多重战略风险。更为关键的是，这些风险在全球技术发展不均衡的结构中呈现出显著的不对称分布。技术先进国同时占据价值获取与规则制定的有利地位，而发展中国家则往往承担数据供给者、技术使用者与治理规则接受者的多重被动角色，面临技术红利不足与风险承受的局面。[⑩]

（三）技术迭代与治理响应的时滞困境

从技术治理的角度来看，人工智能是一种技术迭代不断加速而治理响应滞后的技术形态，这一特征构成了其“安全悖论”的第三重根源。一方面，人工智能的技术迭代表现出显著的内生加速特征。作为典型的数据驱动型技术，其性能提升高度依赖算力扩张、数据积累与算法优化所形成的正反馈循环。每一次技术迭代不仅带来能力提升，也为后续突破奠定更高的能力基座，从而形成自我强化的加速机制。以大语言模型为例，OpenAI 从 GPT-1 到 GPT-4 的发布间隔从最长的 21 个月缩短至 12 个月，但模型能力却呈现指数级跃迁。[⑪] 从技术机理看，人工智能技术迭代的内生加速源于其技术范式本身，并在总体上呈现出可预期的指数型增长轨迹。从 2019 年通用语言模型 GPT-2 的 15 亿参数，到 2020 年 GPT-3 的 1750 亿参数，再到万亿级参数规模的大模型，模型体量在短期内实现数量级跃升，相应训练算力亦呈几何级增长。[⑫]

与技术迭代的内生加速相对的，是治理响应的制度性滞后。治理作为社会制度安排，受制于认知形成、规则协商与制度固化的固有周期，其总体演进呈现出相对线性的特征。通常，法律与监管框架面临严格的程序性约束。规则制定通常经历问题界定、公众咨询、多方利益协调、专业评估与民主审议等环节，难以跨越既定程序。例如，美国《算法问责法案》自 2019 年提出以来历经多年仍未完成立法，而同期人工智能技术已发生多轮代际跃迁。欧盟《人工智能法案》从提出草案到正式生效历时近三年，其间大语言模型迅速从实验室技术演变为全球性通用应用，其风险形态与治理需求已明显超出最初的制度设想。此外，人工智能风险具有显著的社会建构属性。新型技术风险往往需要在实际部署与应用过程中逐步显现，从技术发布到风险识别再到社会共识形成，依赖于持续的公共讨论、文化适应与价值调适。当监管者围绕某一代技术进行制度回应时，下一代技术及其新型风险往往已进入扩散阶段。技术迭代的内生加速与治理响应的外部滞后相互叠加，最终表现为治理失效与风险敞口的持续扩大。

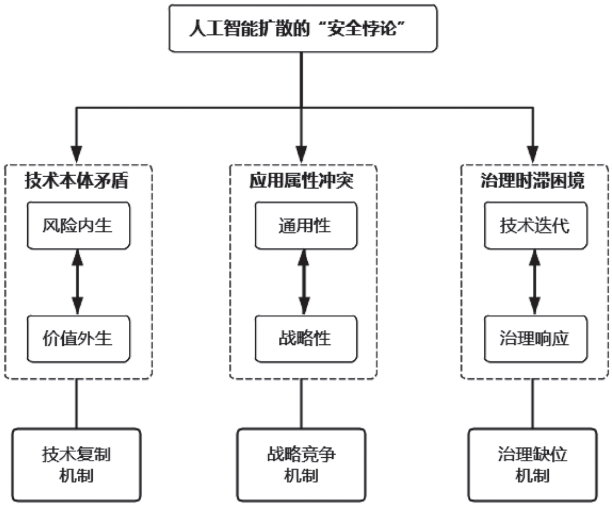

图 1　人工智能扩散的“安全悖论”形态及其传导机制

扩散链路：“安全悖论”的传导机制

人工智能扩散的“安全悖论”并非抽象的理论推演，而是在具体的技术复制路径、应用扩张实践与多元行为体互动中被不断生产和强化。人工智能所具有的开放性、适应性与跨界性，使其安全风险得以在技术扩散过程中跨越原有边界，在不同场景与制度环境中经历复制、外溢、渗透与激化的动态演化。正是在这一过程中，“安全悖论”通过多个关键传导链路将抽象悖论转化为现实后果。

（一）技术复制机制：风险—价值错配的复制化扩散

人工智能的开源化、模块化、低门槛化，使其成为人类技术史上首个价值外生释放与风险内生复制高度同构的技术形态。在这一技术架构下，每一次技术扩散既是价值潜能的释放，也是内生风险的完整复制，二者在技术层面不可分离。

一方面，开源路径是人工智能价值外生性的关键激活机制。依托开源模型、共享框架与社区协作，人工智能显著降低了技术获取与再创新门槛，使多元行为体得以参与能力扩展与场景探索。截至 2024 年底，Hugging Face 等主流开源社区已累计提供超过一万种预训练与微调模型，涵盖通用大模型、垂直领域小模型等多种技术形态，大幅降低了创新门槛。GitHub 统计发现，2024 年生成式人工智能相关开源项目数量同比增长 98%，贡献总量增长 59%。[13]这一扩散逻辑通过开源降低门槛、吸引开发者、积累应用场景，使人工智能的潜在能力得以迅速转化为经济和社会效用。

另一方面，开源路径也是风险内生性的完美复制装置。算法黑箱、数据偏见、系统脆弱性等风险要素，原本即内嵌于模型结构之中，随着模型能力的无差别扩散被原样复制并快速外溢。潜在的恶意行为体能够在毫无技术壁垒的情况下，低成本地获取、定制乃至变异这些高阶工具，实现如自动生成虚假信息、高级网络攻击和敏感技术研发等非预期用途。由此，开源平台尽管本意在于促进创新与协作，却在技术扩散路径中不可避免地引发了系统性安全隐患和潜在威胁。更为重要的是，开源生态的动态演化特性进一步推动了风险的“再生产”与“增殖”。作为一个开放且持续演化的技术生态，模型的自我复制、下游修正及多源协同往往超越初始开发者的控制边界，既有的安全约束与伦理嵌入在扩散过程中被不断稀释甚至解除。例如，原本内嵌内容过滤机制以

防止有害生成的开源模型，可能因易于重构而被绕过或解除相应限制，进而被滥用于仇恨言论、欺诈行为等不当场景。[14]在此背景下，开源模型的风险“再生产”不仅侵蚀现有的安全措施，更导致风险类型和攻击方式的持续演进与多样化。

由此，开源路径集中体现了人工智能“安全悖论”的两难困局：保留模型能力与修改自由，意味着同时保留算法黑箱等本体性风险；而通过加密、审查或封闭化削弱风险，又必然损害开源生态所依赖的可修改性与创新性。此外，开源生态的去中心化、跨国界及社区自治属性，使责任归属与风险追溯愈发模糊，治理真空与“公地悲剧”问题随之凸显。[15] 在缺乏强制性治理框架与统一监管标准的情境下，伦理准则和行业自律往往流于软约束。

（二）战略竞争机制：竞争驱动的扩散加速与风险外溢

人工智能的全球扩散并非自然的市场演化，而是由大国战略竞争驱动的加速过程。各国在技术领先、国家安全、经济竞争力的战略焦虑下，通过产业扶持、人才争夺与标准输出等政策工具，推动人工智能技术在国内的规模化部署与跨境流动。这种由竞赛逻辑主导的扩散路径，在释放通用技术红利的同时，也同步激活了战略依赖、军民转化与跨境风险外溢等多重战略风险。

先发国家在战略竞争导向下推进技术扩散，客观上重构了风险的跨境传导链条。以美国为例，2025 年 7 月 23 日，特朗普政府正式发布《赢得人工智能竞赛：美国人工智能行动计划》，明确提出通过向“所有愿意加入美国人工智能联盟的国家出口其包含硬件、模型、软件、应用程序和标准在内的完整的人工智能技术栈来满足全球对人工智能的需求”，并强调“开源和开放权重模型可能在全球范围内的某些业务领域和学术研究中成为全球标准。因此，它们也具有地缘战略价值”。[16] 这一扩散策略在降低技术门槛的同时，也引导后发国家在缩小技术差距的过程中对既有技术体系形成深度依赖。从底层算力、开发框架到核心模型能力，应用生态被嵌入由先发国家主导的技术栈之中。

（三）治理缺位机制：治理失配下的风险渗透与激化

人工智能的快速扩散与治理体系的建构之间形成了显著的速度差与结构差。应用场景扩张、技术迭代加速与跨域传导同步推高治理的边际成本，而传统治理体系在知识储备、信息掌控与权威配置上的先天局限，使其难以适应去中心化、高度技术化的数字安全情境。在这一错配之下，治理真空与监管盲区不断生成，风险借由扩散过程中的场景渗透、跨主体转移与责任弥散持续放大，最终由局部技术隐患演化为系统性安全威胁。

首先，技术供给方与技术消费方之间的不对称，使风险在扩散过程中呈现明显的外部化趋势。以发达国家及大型科技企业为代表的技术供给方，能够通过技术架构设计、服务条款设定与数据处理规则，将潜在风险因素嵌入技术系统之中，并在跨境扩散过程中转嫁给技术消费方。以后发国家为代表的技术消费方在获得人工智能能力的同时，被动承担了算法偏见、数据泄露、系统漏洞等多重风险。[17] 2024 年，牛津大学发布的《人工智能治理的哪些方面应国际化？》白皮书指出，大型跨国科技企业可能通过严格的本地合规措施控制其人工智能产品的相关风险，但当该产品输出到监管能力相对薄弱的市场时，原本受控的风险因素可能在新的应用环境中被激活并放大，形成“风险外部化”现象。[18]

其次，人工智能时代安全治理能力的结构性约束进一步加剧了风险扩散的势头。在人工智能时代，传统的官僚性等级治理体系在扁平化和私人化特征突出的数字安全情境中表现出明显的适应性“能力赤字”。在等级化的官僚体制中，责任划分机制固化，应急响应方式僵化，专业人才配备滞后，均难以满足人工智能安全治理跨领域、快速变化的现实需求。政府也不能完全掌握存

在安全风险的关键信息基础设施、数据和算法，大型科技企业作为技术研发与创新的核心驱动力，凭借对关键算法、海量数据资源以及技术平台的深度垄断，已然成为影响安全治理的主导力量。面对去中心化技术扩散的不可逆转趋势，国家的安全治理能力受制于体制和资源瓶颈，政府不得不愈发依赖市场主体和技术社群在关键信息基础设施与核心技术环节的安全能力补位。这一转变反而造成政府权威的弱化和安全治理能力的碎片化。

最后，风险认知不足与规制工具的滞后共同塑造了风险疏漏的“制度洼地”。在应用优先的发展导向下，后发国家在推动人工智能广泛落地的过程中，往往将经济绩效与产业突破置于安全规制之前，对算法安全、伦理约束与数据合规问题重视不足。表面的快速普及掩盖了深层次的风险积累，模型投毒、算法歧视、数据泄露等问题在应用扩散过程中不断叠加并跨域传导。根据《全球人工智能安全指数》对 40 个国家的调研数据显示，虽然有 18 个国家配备了与人工智能安全治理相关的政策工具，但仅有 8 个国家在国家层面建立了系统化的法律、政策与技术治理框架。[19] 在具体操作中，企业与科研机构虽然设立了伦理审查机制，但受限于专业隔离与沟通不足，伦理考量往往游离于技术研发与部署的核心流程之外。[20] 安全议题被简化为技术参数调节或工程补救，缺乏对系统性社会风险与外部效应的整体考量，监管实践因而呈现碎片化、事后补救的治理常态。在人工智能深入公共治理、关键基础设施与社会管理等敏感领域时，这类制度性缺位尤为突出，潜在安全隐患更易被边缘化并固化为监管“盲区”。

风险外溢：“安全悖论”的多重影响

人工智能扩散的“安全悖论”，并不仅仅局限于单一领域或特定主体，而是在技术生态、国家社会、国际安全层面引发了多重影响。这种多重影响既表现为监管缺失下的技术生态恶化，也体现为社会治理能力滞后引发的风险乱象，并最终在国际安全层面形成“水桶效应”，即个别国家的安全短板可能成为全球安全体系的薄弱环节，影响国际社会的整体安全。

（一）技术生态恶化

技术复制机制通过开源路径实现了人工智能能力扩散与风险复制的同构，其在技术生态层面的直接后果是技术发展的无序化、技术依赖的结构性固化以及技术滥用的产业化。在后发国家和地区，由于监管框架缺失、技术筛选能力不足、规则制定权薄弱，这三重后果表现得尤为显著，导致技术生态从促进创新的良性循环转向风险累积的恶性循环。

首先，对人工智能研发与应用主体的规则约束不力，由此导致技术发展的无序化。在缺乏统一技术标准，如算法透明度、数据合规性、安全审计规范和强制性约束机制的情境下，科技企业常优先披露技术优势或专利成果，而对潜在风险如算法偏差、数据泄露保持沉默。这种制度环境为大量低质量、高风险的人工智能技术创造了生存空间，使其得以规避必要的安全审查程序快速扩散。中国软件评测中心联合多家机构发布的《Top 开源大模型安全测评报告（2024）》显示，在接受测评的 20 款开源大模型中，当面临内容分割重组、角色伪装、直接问答、语言切换、混合攻击等多维检测方式时，绝大多数模型在公共安全、道德伦理、不良信息传播和网络安全等领域均表现出防护能力不足。[21] 监管机制的滞后性和有效性不足，实质上为技术生态注入了大量风险因子，侵蚀其可持续发展的基础。

其次，监管缺失助长了技术生态的外部依赖性与“算法殖民”。后发国家在自主研发能力和

规则制定权方面的双重劣势，使其在面对外部技术输入时缺乏有效的筛选和风险管控机制。当本土监管框架未能及时建立或与国际标准脱节时，大量未经本地化适配的外来人工智能模型得以直接渗透本土市场，从而催生“算法殖民”的新型依附关系，即核心技术标准、数据流向规则乃至模型价值偏好均由外部主体主导，本土技术生态系统沦为全球科技巨头的技术输出目标市场。[22] 联合国贸易和发展会议发布的《2025 年技术与创新报告》揭示了这一不平衡格局的量化特征：全球 40% 的私人研发投资集中于 100 家公司，其中绝大多数总部设在美国。[23] 技术发展的不平衡，直接加剧了技术生态的不稳定性和外部风险传导的敏感性。

此外，监管缺失催生的“负向创新”亦不容忽视。当制度缺位使得违法作恶成本显著低于合规成本时，人工智能技术便成为非法产业链技术升级的重要工具。深度伪造技术的门槛降低与生成式人工智能的普及，使得传统网络犯罪呈现出技术复杂化与产业链条化的新特征。一方面，诈骗手段从简单的文本欺骗升级为多模态深度伪造，涵盖语音克隆、视频换脸、实时交互等多元化形式。如 2025 年 6 月，网络犯罪分子利用深度伪造技术模仿美国国务卿马尔科·鲁比奥（Marco Rubio）的声音和写作风格，通过伪造账户成功接触了至少五名政府官员。[24] 另一方面，犯罪组织结构日趋专业化，形成了从算法优化、工具开发、数据采集到诈骗实施的完整产业生态。根据德勤的最新报告，与深度伪造相关的网络攻击损失预计将从 2023 年的 123 亿美元飙升至 2027 年的 400 亿美元，复合年增长率达 32%。[25]这种“负向创新”的扩散效应不仅直接危害个体权益与社会秩序，更通过技术路径的扩散，恶化全球技术生态的信誉基础与安全信心。

（二）国家社会风险乱象

在人工智能技术向后发国家的无序扩散中，由于治理体系的缺位与滞后，原本潜藏于技术系统中的风险可能向社会各领域无节制地渗透和放大，造成国家层面的风险乱象。

首先，治理缺位加剧了社会信任机制的系统性危机。信任作为社会运行的基础性资源，在人工智能时代面临“算法黑箱”、决策不可解释性的多重冲击。生成式人工智能技术的低门槛特征，使其在缺乏有效内容溯源与问责机制的情况下，极易异化为虚假信息批量生产的技术工具，从而为认知操纵、舆论引导和社会分化提供技术支持。深度伪造技术的恶意滥用、虚假信息的智能化生成与病毒式传播、算法操纵下的“信息茧房”效应等，在治理缺位的环境中得以野蛮发展，直接冲击社会信任的基础。[26]

其次，治理缺位加剧了社会分配领域的结构性不公。在缺乏算法透明度审查、偏见审计及有效救济途径的制度框架下，人工智能被广泛部署于就业筛选、信贷审批、司法裁量等关键领域，技术“理性”的外壳掩盖并放大了其训练数据中固有的社会偏见和刻板印象，使得已处于社会不利地位的群体进一步被边缘化。例如，智能招聘算法容易作出歧视女性和有色人种的决策；智能金融算法可能作出降低弱势群体的信贷得分、拒绝向特定族群提供贷款服务，以及倾向于向低收入群体定向推送高息贷款产品等歧视性决策。[27] 更严重的是，算法偏见的社会影响具有累积性和自我强化的特征，被算法系统边缘化的群体往往缺乏足够的社会资源和技术能力来挑战这些不公正的自动化决策，从而在社会分层体系中陷入劣势地位的持续循环。

再次，技术替代冲击在社会保障体系缺位的情况下，正在引发大规模的结构性失业与社会不稳定风险。人工智能驱动的自动化浪潮对后发国家劳动密集型产业的冲击尤为剧烈。高盛发布的《人工智能对经济增长的潜在影响》报告指出，当前人类所执行工作任务中的 1/4 具备被人工智能系统完全替代的技术可行性，全球范围内可能受到直接影响的就业岗位数量超过 3 亿个。[28]

而后发国家的职业再培训体系、失业救济网络及产业转型引导政策普遍准备不足。历史经验表明，大规模的结构性失业往往带来社会矛盾的激化和政治极化的加剧，对国家治理稳定性构成严重威胁。

此外，后发国家在技术扩散中数据主权的流失与隐私保护的系统性失效，对国家安全构成了深层威胁。在全球数字经济中，后发国家往往扮演数据资源的供给方而非规则的制定方。在缺乏强制性数据本地化存储、跨境流动审查及平台主体责任法规的情境下，海量涉及国民身份、关键基础设施运营的高敏感数据，通过不受约束的人工智能应用与服务流向境外。[29] 尼日利亚竞争和消费者保护委员会 2024 年调查发现，Meta 在未获得用户明确同意的情况下，擅自实施跨境数据传输和存储行为，并对用户强制执行具有明显剥削性质的隐私政策条款，严重侵犯了尼日利亚用户的合法权益。[30] 这种“数据殖民”的技术实践不仅削弱了后发国家的经济竞争力，更使后发国家在面对外部网络攻击、情报渗透与地缘政治胁迫时处于极度被动的脆弱地位。

（三）国际安全中的“水桶效应”

人工智能技术的全球扩散在国际安全领域呈现出典型的“水桶效应”，即全球人工智能安全的整体水平并非取决于技术最先进国家的防护能力，而是受制于治理能力最薄弱地区的安全底线。在高度互联的数字时代，后发国家在人工智能治理方面的短板与漏洞，不仅威胁其自身安全，更通过网络空间的互联互通性、技术供应链的全球化布局以及安全威胁的跨境传播机制，冲击全球安全架构。

首先，网络空间的互联互通性使得后发国家的局部安全漏洞成为全球性威胁的入口。在人工智能技术生态系统中，安全风险不再遵循传统的地理约束和主权边界，而是通过数字基础设施的连通性实现跨境传播，任何单一节点的脆弱性都可能通过网络效应放大为整体的系统性风险。以僵尸网络攻击为例，网络安全公司 Akamai 的数据显示，全球范围内大规模分布式拒绝服务攻击的主要源头集中分布在网络安全治理体系相对薄弱的中东、非洲以及亚太部分地区。[31]这些地区和国家的人工智能防护系统缺陷被恶意行为体利用，成为攻击全球关键基础设施的“跳板”。

其次，全球化供应链的深度融合放大了局部治理失效的连锁风险。现代人工智能产业呈现出高度分工的全球化特征，从芯片制造、算法开发到应用部署，各环节分布在不同国家和地区，形成“你中有我、我中有你”的生产格局。在这种高度整合的供应链体系中，任何单一节点的薄弱环节都可能因外部恶意行为、内部安全管理不足、组件更新与管理机制不安全而被阻断，进而通过供应链的关联效应向整个系统传导风险。2025 年 8 月曝光的 Nx 供应链投毒事件为这种风险传导机制提供了最新的实证案例。在该事件中，攻击者先窃取主流开源构建工具 Nx 的发布令牌，随后在多个新版本更新包内植入恶意脚本。代码一旦在开发者本地运行，会立即调用已安装的云端模型命令行工具，全盘扫描钱包密钥、云凭据和源代码，并将结果回传至外部服务器。Nx 在全球约七成大型企业中应用广泛，有毒版本仅上线五个小时就波及数万名开发者，并进一步通过集成管道向企业生产环境扩散，造成大规模密钥泄露和潜在的国家关键基础设施风险。[32] 这种“蝴蝶效应”式的风险扩散充分体现了全球化背景下安全风险的非线性传播。

此外，当前国际人工智能安全合作机制的不平衡发展，使得“水桶效应”的负面影响进一步加剧。现有的国际人工智能安全合作主要集中在少数技术发达的大国之间，广大后发国家往往被排除在核心治理机制之外，或仅能以被动接受者的身份参与。由联合国秘书长组建的高级别人工智能咨询机构发布的《为人类治理人工智能》最终报告显示，在联合国 193 个会员国中，仅有 7

个国家实质性参与了近期国际社会提出的七项重要人工智能治理倡议，而多达 118 个会员国完全缺席相关治理进程，这些被边缘化的国家主要来自“全球南方”地区。[33] 这种“合作鸿沟”使得后发国家难以获得必要的技术支持、能力建设援助以及制度经验分享，不仅无法有效提升全球人工智能治理体系中最薄弱环节的治理能力，反而通过治理标准的分化，加剧国家间安全防护能力的差距，进而维持甚至扩大了其作为全球“安全短板”的地位。[34]

治理规约：“安全悖论”的规制机制

人工智能扩散的“安全悖论”本质上反映了技术扩散与风险治理之间的失衡。当下，全球范围内的人工智能的安全治理普遍面临“治理赤字”与“能力鸿沟”的双重挑战，亟须在技术快速演进、规则多元分化的现实下，探索更具包容性、韧性与适应性的新型治理框架，推动法律、伦理与技术标准的多维衔接，强化能力建设和赋权型国际合作，构建化解“安全悖论”的规制机制，从而实现人工智能扩散过程中的风险可控与利益共赢。

（一）后发国家人工智能安全治理的体系构建

在全球人工智能技术加速扩散的背景下，后发国家既需抓住技术带来的发展机遇，又必须审慎应对随之而来的安全治理挑战。这些挑战不仅体现为技术能力的不足，更深层次地根植于全球治理体系的权力不对称、规范话语权削弱及数字主权维护的多重博弈中。因此，后发国家应将人工智能安全治理纳入国家安全战略，秉持“自主、协同、前瞻”的原则，统筹本土实际与国际格局，形成具有韧性与适应性的国家安全屏障。[35]

第一，建立本土化的分级分类治理体系与动态风险响应机制。2023 年中国出台的《生成式人工智能服务管理暂行办法》率先确立了备案加分类分级的监管框架，规定生成式人工智能服务中“具有舆论属性或社会动员能力”的应用需履行安全评估与算法备案，一般应用则实施事中事后监管，体现了风险导向的差异化治理逻辑。[36] 2024 年北京人工智能数据训练基地监管沙盒成为全国首个人工智能监管沙盒项目，该项目通过设立限定性条件，以真实用户为对象进行深度测试，降低监管不确定性。[37] 这一制度设计的引入有效缓解了时滞困境。当技术扩散速率超过传统立法周期时，通过设定不同风险等级的合规宽限期，在安全底线与创新空间之间建立缓冲带。后发国家可借鉴此经验，根据本国技术发展阶段与治理能力，构建风险评估矩阵与分级响应清单的适应性治理体系。

第二，争取全球人工智能规则制定中的话语权，提升议价能力。技术扩散中的隐性权力结构往往通过 API 接口控制、开源协议条款和云服务依赖等方式实现。例如，OpenAI 的 API 使用协议禁止用户将其模型用于开发竞争性人工智能系统，Meta 的 LLaMA 虽标榜开源，但其商业许可限制月活用户超 7 亿的企业使用。[38]这些条款构成了技术霸权的契约化。后发国家企业即便获得模型权重，仍受制于数据回流条款、算法黑箱审计权与跨境数据处理的不对等约束。因此，后发国家应通过区域联盟强化集体议价能力。2024 年非洲联盟第 37 届首脑会议通过的《数字贸易议定书》，允许成员国要求计算设施本地化，对源代码、算法等自愿转让或获取，对人工智能等新兴技术授权未来设立“技术工作组”制定实施细则。[39] 这一机制通过集体标准设定对冲了单一国家在技术谈判中的弱势地位，值得“全球南方”国家借鉴。

第三，突破核心技术“卡脖子”环节，掌握安全治理的技术底座。安全治理能力的根本，最

终取决于自主可控的技术能力。当前，人工智能领域的技术瓶颈主要集中在可信人工智能、隐私计算与安全检测等关键层面。后发国家必须将资源集中于这些“卡脖子”环节，力求实现技术突破，从而掌握安全治理的技术主动权。例如，大力发展可信人工智能技术，特别是模型的可解释性、鲁棒性和公平性研究，能够有效防范对抗性攻击、数据投毒和模型偏见等风险；积极布局隐私计算技术，如联邦学习、同态加密和安全多方计算，可以在不暴露原始数据的前提下实现数据价值的共享与利用，有效化解数据开发与隐私保护之间的内在冲突；同时，将大模型行为监测、内容安全审计平台等纳入国家“安全新基建”范畴，构建国家级的人工智能安全检测基础设施，将有助于减少对外部安全工具和服务的依赖，构筑坚实的技术底座。

第四，构建多元主体协同的责任分担机制。随着人工智能技术的全球扩散，治理主体愈加分散，治理模式亟须从传统的“政府单中心”转向多主体协同。以欧盟《人工智能法案》为例，其创设的“监管金字塔”模式，依据人工智能系统可能带来的风险水平，将其划分为不可接受风险、高风险、有限风险和低风险四个层级，并为不同层级的系统及其各个主体配置了差异化的法律责任与合规义务。[40] 中国 2022 年施行的《互联网信息服务算法推荐管理规定》也明确了企业的“算法安全主体责任”，要求企业建立算法安全评估、数据安全管理及应急处置机制，并向监管部门报送年度算法安全报告。[41] 然而，不容忽视的是中小企业往往面临较高的合规成本，可能因监管过载导致创新被抑制。因此，在强化企业主体责任的同时，建立配套的支撑体系至关重要。后发国家可探索建立由政府资助或引导的第三方合规服务体系，通过认证一批专业的第三方机构，为广大中小企业提供标准化、低成本的风险评估、算法审计和合规咨询服务，这不仅能有效降低全社会的合规门槛，更能促进技术创新与安全治理的双赢。

（二）全球人工智能安全治理的协同规制

人工智能技术的全球性流动与风险外溢已在国际体系引发“治理赤字”和“安全困境”的叠加效应。作为全人类共享的科技创新，人工智能在全球协同机制尚未健全的背景下，易沦为国家间战略竞争、“技术民族主义”乃至地缘政治角逐的工具，加剧全球安全风险的循环与放大。因此，亟须超越单边路径，构建一个具有韧性、适应性和包容性的协同治理框架。

第一，推进跨文化兼容的伦理标准与技术互操作体系。解决全球人工智能伦理和价值冲突的关键在于形成可操作的技术解决方案，而非仅限于意识形态的宣示。在此背景下，伦理协议的分层嵌入机制提供了有效的解决路径。在模型架构层嵌入“价值观中性”的基础安全协议，在应用层允许根据地域文化定制化调整，IBM 的 watsonx.governance 平台实现了这一设计，企业可在统一的安全基座上叠加符合当地法规的合规模块，既保证基础安全，又尊重文化多样性。[42] 此外，“跨文化算法审计”机制可以为事后监督提供有效手段。例如，联合国教科文组织启动的“全球人工智能伦理观察站”项目，联合不同文化背景的专家团队，对主流人工智能系统在多元文化场景下的偏见表现进行审计。[43] 这类审计结果倒逼企业改进训练数据和奖励模型，为技术普惠与价值兼容提供了实践通道。

第二，建立常态化的风险信息共享与协同响应机制。人工智能技术扩散带来的风险具有突发性和跨境性特征，需要在全球范围内形成常态化的风险信息共享和协同响应机制。2024 年 11 月，美国、英国、新加坡等 10 余国组建了“人工智能安全研究所国际网络”（International Network of AI Safety Institutes），旨在联合开展对前沿人工智能模型的测试评估、共享风险研究成果、推动评估标准的国际互认，并计划扩容至更多的发展中国家。[44] 后发国家可主动参与此类机制，同时应

探索利用“区块链存证”与“联邦审计”的技术方案，通过区块链技术记录人工智能系统的关键决策日志，在争议发生时可以由多个国家的监管机构联合调取审计数据，既保证问责透明，又避免数据跨境引发的主权争议。

第三，构建具有约束力的全球治理机制与履约保障体系。当前全球人工智能治理体系主要依赖软法，缺乏有效的硬约束。为应对这一挑战，建议借鉴《蒙特利尔议定书》在臭氧层保护方面的成功经验，推动制定《全球人工智能安全公约》，明确各国在高风险人工智能系统监管、开源模型安全审查和算法透明度等方面的最低标准，并设定达标时间表。此外，应建立“人工智能安全合规指数”，定期评估各国履约情况；设立“全球人工智能安全基金”，支持发展中国家建设人工智能安全实验室、培训监管人才和采购安全检测工具；同时，公约应引入“技术转让义务”条款，要求掌握核心安全技术的国家以优惠条件向发展中国家转让可信人工智能、隐私计算等防御性技术，避免治理能力鸿沟固化为新的“数字殖民”。

（三）能力建设与技术赋权的国际支持框架

全球人工智能安全治理的根本挑战之一，在于各国间普遍存在的能力鸿沟与赋权不对称。这一差距已成为数字时代结构性不平等的具体体现。当非洲国家连基础算力设施都依赖外部援建时，要求其同步建立人工智能安全治理体系无异于缘木求鱼。因此，构建以赋权为导向、兼顾包容性的国际能力建设支持框架，是实现全球人工智能安全治理体系可持续发展的前提。

第一，搭建开放式的全球人工智能安全知识共享网络。知识赤字是能力鸿沟的首要根源。2025 年，联合国教科文组织在 G20 峰会上宣布将培训超过 15000 名公务员掌握人工智能应用，鉴于人工智能等新技术对司法系统的影响，还将培训 5000 名法官和检察官，同时启动泛非孵化器，支持 1500 名研究人员参与人工智能的发展。[45] 中国可发挥在人工智能技术普及方面的优势，通过“数字丝绸之路”倡议输出经验，例如将《生成式人工智能服务管理暂行办法》的核心逻辑制作成多语种政策工具包，协助共建“一带一路”国家建立适配本国国情的监管框架。

第二，建立多元化的技术援助与资源共享机制。印度的“数字公共基础设施”（DPI）模式为创新的技术援助提供了借鉴。印度通过构建开放且可互操作的技术底座，如统一身份认证系统、支付和数据交换系统降低了后发国家数字服务开发的成本。2024 年，印度将 DPI 理念扩展至人工智能治理领域，推出“India AI Mission”，包括机器学习、合成数据生成、算法偏见缓解、隐私增强工具、可解释人工智能、人工智能治理测试、人工智能伦理认证和算法审计工具。[46] 这种“授人以渔”的赋权模式比单纯的资金援助更具可持续性。由此，国际社会应推动建立“全球人工智能安全技术共享池”，通过各国自愿贡献非敏感的安全技术，如开源的对抗样本防御算法和隐私计算协议，形成全球公共产品。同时，多边开发银行应设立人工智能安全专项贷款，支持发展中国家采购算力设施和建设安全检测平台。

第三，构建尊重自主权的赋权型合作关系。能力建设的目标是培育发展中国家的内生治理能力，而非建立新的依附关系。非洲联盟 2024 年通过的《非洲大陆人工智能战略》明确拒绝“一刀切”的外部方案移植，强调人工智能发展与应用必须根植于非洲的实际，优先解决数字基础设施短缺、数据主权保护和本土语言模型开发等核心关切。[47] 卢旺达与加纳率先制定国家人工智能战略，探索在全球技术标准与本地发展需求之间实现平衡，以推动可持续发展并维护数字主权。这些实践表明，赋权型合作的关键在于尊重受援国的议程设定权，确保技术转让不附加“价值观捆绑”，让能力留存而非依附关系成为合作的最终成果。

结语

“安全悖论”作为人工智能扩散的内在张力与结构性隐患，其本质在于科技赋能下风险转移与不确定性扩大的态势。这一悖论不仅体现为安全防线的脆弱性与风险链条的延展化，更映射出治理理念、制度框架与伦理规范的适应性滞后，由此导致系统性安全困境的产生。面对“安全悖论”不断加剧的现实，任何对技术自主、利益最大化的片面追求都应受到理性规制和社会共识的约束，确保科技发展服务于公共安全与伦理底线。

基于此，亟须系统把握人工智能扩散背景下“安全悖论”的理论逻辑与演化机理，整合政策、技术与伦理等多维治理资源，构建动态、协同与前瞻性的综合安全治理体系。唯有如此，才能在保障技术安全可控的前提下，充分激发人工智能技术的创新潜力和社会价值。未来，面对人工智能扩散进程中“安全悖论”的长期化与复杂化趋势，各利益攸关方须保持高度的审慎意识与责任担当，强化信息共享、风险共治与价值协商。只有将安全规制体系前置并嵌入技术创新与应用的全过程，充分调动全球治理合力，方能有效规避“安全悖论”演变为现实危机，促进人工智能发展的可控、可信、可靠，最终实现智能技术对人类社会文明的正向引领与深远促进。

注释：

① Manoj K. Kamila and Sahil S. Jasrotia, “Ethical Issues in the Development of Artificial Intelligence: Recognizing the Risks,” *International Journal of Ethics and Systems*, vol.41, no.1, 2025, pp.45-63; Joshua Hatherley, “A Moving Target in AI-assisted Decision-making: Dataset Shift, Model Updating, and the Problem of Update Opacity,” *Ethics and Information Technology*, vol. 27, no. 2, 2025, p. 8.

② Gabbiadini Alessandro, et al., “Artificial Intelligence in the Eyes of Society: Assessing Social Risk and Social Value Perception in a Novel Classification,” *Human Behavior and Emerging Technologies*, no.1, 2024, pp.1-10; 丁元竹：《人工智能技术的潜在社会风险及治理体制机制研究》，《行政管理改革》2025 年第 7 期。

③ Jenna Burrell, “How the Machine ‘Thinks’: Understanding Opacity in Machine Learning Algorithms,” *Big Data & Society*, vol 3, no. 1, 2016, pp.1-12.

④ 王峰：《人工智能意识“涌现论”的概念误区与未来视野》，《华东师范大学学报》（哲学社会科学版）2024 年第 2 期。

⑤ 李南等：《面向大语言模型的越狱攻击综述》，《计算机研究与发展》2024 年第 5 期。

⑥ Adam Tauman Kalai, et al., “Why Language Models Hallucinate,” *OpenAI & Georgia Tech*, 2025-09-14, https://cdn.openai.com/pdf/d04913be-3f6f-4d2b-b283-ff432ef4aaa5/why-language-models-hallucinate.pdf.

⑦ 新华网：《欧盟计划放宽数字监管法规》，2025 年 11 月 20 日，https://www.news.cn/world/20251120/3103485540ba4b8b99323c09db368f23/c.html。

⑧ Stanford HAI, *Artificial Intelligence Index Report 2025*, 2025-04-03, https://hai.stanford.edu/ai-index/2025-ai-index-report.

⑨ 沈逸、宫云牧：《技术权威与数字时代的国际技术竞争——以移动通信技术领域为例》，《国际安全研究》2023 年第 6 期。

⑩㉖ 黄日涵、姚浩龙：《被重塑的世界？ ChatGPT 崛起下人工智能与国家安全新特征》，《国际安全研究》2023 年第 4 期。

⑪ OpenAI, *GPT-4 Technical Report*, 2023, https://arxiv.org/pdf/2303.08774v3.

⑫ 澎湃·机器之心：《1750 亿参数，史上最大 AI 模型 GPT-3 上线：不仅会写文章、答题，还懂数学》，2020 年 5 月 30 日，https://m.thepaper.cn/newsDetail_forward_7634184。

⑬ GitHub, *Octoverse 2024*, 2024-10-29, https://github.blog/news-insights/octoverse/octoverse-2024/.

⑭ Elizabeth Seger, et al., “Open-Sourcing Highly Capable Foundation Models: An Evaluation of Risks, Benefits, and Alternative Methods for Pursuing Open-Source Objectives,” in *Centre for the Governance of AI*, 2023, p. 2, https://arxiv.org/abs/2311.09227.

⑮ 张凌寒、何佳欣：《开源人工智能负责任创新的法律保障》，《法治社会》2025 年第 3 期。

⑯ The White House, *Winning the Race: America's AI Action Plan*, 2025-08-05, https://zw.usembassy.gov/winning-the-race-americas-ai-action-plan/.

⑰ Shakir Mohamed, et al., “Decolonial AI: Decolonial Theory as Sociotechnical Foresight in Artificial Intelligence,” *Philosophy &*

Technology, vol. 33, 2020, pp. 659-684.

⑱ University of Oxford, "What Should Be Internationalised in AI Governance?" 2024-11-13, https://www.oxfordmartin.ox.ac.uk/publications/what-should-be-internationalised-in-ai-governance.

⑲ Center for Long-term Artificial Intelligence（CLAI）, et al., *Global Index for AI Safety*, Feb 2025, https://agile-index.ai/Global-Index-For-AI-Safety-Report-EN.pdf.

⑳ Thilo Hagendorff, "The Ethics of AI Ethics: An Evaluation of Guidelines," *Minds and Machines*, vol. 30, no. 1, 2020, pp. 99-120.

㉑ 中国软件评测中心：《Top 开源大模型安全测评报告（2024）》，2024 年 12 月，https://www.vzkoo.com/document/20241225a9b5ebc53b7f78f63cba75bf.html。

㉒ Chinmayi Arun, "Transnational AI and Corporate Imperialism," *Carnegie Endowment for International Peace*, 2024-10-08, https://carnegieendowment.org/research/2024/10/transnational-ai-and-corporate-imperialism?lang=en.

㉓ 联合国贸易和发展会议：《2025 技术和创新报告：包容性人工智能促进发展》，2025 年 5 月 19 日，https://unctad.org/system/files/official-document/tir2025overview_ch.pdf。

㉔ 上观新闻：《美国外交电报曝光：有人利用 AI 技术深度造假，国务卿鲁比奥也被冒充》，2025 年 7 月 9 日，https://finance.sina.com.cn/jjxw/2025-07-09/doc-infewcvu1391640.shtml。

㉕ Deloitte Center for Financial Services, "Generative AI is Expected to Magnify the Risk of Deepfakes and Other Fraud in Banking," 2024-05-29, https://www.deloitte.com/us/en/insights/industry/financial-services/deepfake-banking-fraud-risk-on-the-rise.html.

㉗ 洪永森、史九领：《人工智能的政治经济学分析》，《学术月刊》2024 年第 1 期。

㉘ Goldman Sachs, *The Potentially Large Effects of Artificial Intelligence on Economic Growth*, 2023-03-26, https://www.gspublishing.com/content/research/en/reports/2023/03/27/d64e052b-0f6e-45d7-967b-d7be35fabd16.html.

㉙ 张春宇：《非洲人工智能发展的现实图景与多重挑战》，《西亚非洲》2025 年第 5 期。

㉚ Chinedu Asadu, "Nigeria Fines Meta $220 Million for Violating Data Protection and Consumer Rights Laws," *"Associated Press*（*AP*）, 2024-07-19, https://apnews.com/article/nigeria-meta-fine-facebook-whatsapp-9c79447e348dcaaa1b8c59898e60c7fa.

㉛ Akamai：《2023 年 DDoS 趋势回顾与 2024 年实用行动策略》，2024 年 1 月 9 日，https://www.akamai.com/zh/blog/security/a-retrospective-on-ddos-trends-in-2023。

㉜ Dan Goodin, "Nx NPM Packages Poisoned in AI-Assisted Supply-Chain Attack," *The Register*, 2025-08-27, https://www.theregister.com/2025/08/27/nx_npm_supply_chain_attack/.

㉝ UN. Advisory Body on Artificial Intelligence, *Governing AI for Humanity : Final Report*, September 2024, https://digitallibrary.un.org/record/4062495?v=pdf.

㉞ 余南平：《通用人工智能时代的国际权力重塑》，《中国社会科学》2025 年第 4 期。

㉟ 郎平：《强化人工智能安全治理》，《前线》2024 年第 5 期。

㊱《生成式人工智能服务管理暂行办法》，https://www.cac.gov.cn/2023-07/13/c_1690898327029107.htm。

㊲ 张苑、瞿晶晶、张晅昱：《人工智能监管沙盒：概念界定、实践经验及对我国的启示》，《世界科技研究与发展》2025 年增刊第 S1 期。

㊳ CSDN：《Llama-Chinese 模型版权指南：商用许可与开源协议解读》，2025 年 11 月 9 日，https://blog.csdn.net/gitblog_01185/article/details/154556377。

㊴ 孙志娜：《非洲数字贸易规则塑造及中国因应》，《西亚非洲》2025 年第 4 期。

㊵ European Commission, *EU AI Act*, 2024-08-01, https://artificialintelligenceact.eu/the-act/.

㊶《互联网信息服务算法推荐管理规定》，https://www.gov.cn/zhengce/2022-11/26/content_5728941.htm。

㊷ IBM Corporation, "Scale trusted AI with watsonx.governance," 2025-06-25, https://www.ibm.com/products/watsonx-governance.

㊸ UNESCO, "Global AI Ethics and Governance Observatory," https://www.unesco.org/ethics-ai/en.

㊹ 中国科学院科技战略咨询研究院：《美国启动人工智能安全研究所国际网络》，2025 年 3 月 18 日，http://casisd.cas.cn/zkcg/ydkb/kjzcyzxkb/2025/zczxkb202501/202503/t20250318_7560406.html。

㊺ UNESCO, "AI in Africa: UNESCO Unveils New Solutions for Its Development at the G20," 2025-09-30, https://www.unesco.org/en/articles/ai-africa-unesco-unveils-new-solutions-its-development-g20.

㊻ "Last Chance to Apply: IndiaAI Mission's Safe & Trusted AI EoI Deadline Extended," 2024-12-20, https://indiaai.gov.in/article/last-chance-to-apply-indiaai-mission-s-safe-trusted-ai-eoi-deadline-extended.

㊼ African Union Commission, *Continental Artificial Intelligence Strategy*, 2024-08-09, https://au.int/en/documents/20240809/continental-artificial-intelligence-strategy.

编辑　杜运泉