生成式人工智能治理国际研究评述：焦点议题与立场分野

张树华　张夏添

〔摘要〕 生成式人工智能作为极具影响力的新兴技术，正深刻影响和形塑着人类的政治、经济和社会发展等诸多领域，并在全球各界掀起广泛关注与普遍探讨。当前，关于生成式人工智能的全球研究焦点已从单纯的技术规制升维至文明形态重构，这种立场分野本质上是数字文明时代生产关系的全球性调适，既折射出国际权力结构的深刻变迁，更预示着人机共生社会新型治理范式的孕育需求。国内外学界基于“风险–成因–对策”三重进路建构了生成式人工智能的治理方案，这就需要对既有研究作归纳比较，进一步寻找未来生成式人工智能的治理之道，以平衡定性与定量的落差、关注理念与道路的撕裂、链接实践与理论的时差、填补设计与落实的断层、弥合域内与域外的隔离。

〔关键词〕 生成式人工智能；人工智能；治理

〔中图分类号〕D630　〔文献标识码〕A　〔文章编号〕1000－4769 (2025) 03－0025－12

当今世界百年未有之大变局加速演进，全球创新版图和经济结构经历深刻调整。生成式人工智能作为引领新一轮科技革命和产业变革的战略性技术，具有溢出带动性很强的“头雁”效应，将对全球经济社会发展和人类文明进步产生深远影响，正在重塑知识生产机制，重构权力运行逻辑，重绘文明演进轨迹。生成式人工智能治理研究是把握人类文明智能化转型历史方位的关键性课题。其治理体系构建不仅关乎技术风险防控效能，更涉及数字时代文明秩序的主导权配置。当前国际社会面临根本性挑战，关键在于如何确立智能技术爆炸性增长与人类文明可持续发展之间的动态平衡关系。这种平衡既需要突破传统治理框架的路径依赖，又必须维系技术伦理的文明基线，其复杂程度远超历次技术革命带来的治理命题。以比较视野开展系统性研究，须以人类命运共同体理念为坐标，在技术创新空间与文明守护责任之间建立战略均衡，为智能时代生成式人工智能治理提供超越零和博弈的元规则框架。

一、生成式人工智能的研究热点与研究进路

面对纷繁复杂的国内外形势和新一轮科技革命与产业变革，中国共产党第二十届中央委员会第三次全体会议通过的《中共中央关于进一步全面深化改革 推进中国式现代化的决定》指出，要贯彻新发展理念，健全因地制宜发展新质生产力体制机制，推动技术革命性突破、生产要素创新性配置、产业深度

〔作者简介〕张树华，中国社会科学院政治学研究所研究员、博士生导师；
张夏添，中国社会科学院大学政府管理学院博士研究生，北京　102488。

〔基金项目〕中国社会科学院学科建设登峰战略“中外政治与国家治理”（DF2023YS37）；国家社会科学基金重点项目“健全全过程人民民主制度体系的理论与实践创新研究”（24ZD070）

转型升级，发展以高技术、高效能、高质量为特征的生产力，完善推动人工智能等战略性产业发展政策和治理体系，引导新兴产业健康有序发展。①生成式人工智能是人工智能技术的最新突破，是新质生产力的典型代表与核心所在，正在快速迭代与发展中不断催生全球新产业、新模式和新动能。

生成式人工智能（AI-Generated Content，AIGC）是指通过人工智能算法对数据进行收集、分析、加工、创造而智能生成的多媒体内容，包括音频、文本、图像、视频及多模态内容。它是人工智能领域理论建构、技术设计、应用场景的最新发展，以高度灵敏、深度交互、快速迭代、加速融合的技术特点深度嵌入并形塑诸多领域，成为全球科技、经济、文化、政治等的新型基础设施。2025年1月，中国深度求索公司自主研发的生成式人工智能DeepSeek以其高智能、快响应、低成本、全开源的技术应用特点震动全球，生成式人工智能的技术、应用、开发、治理进入新一轮博弈。作为极具突破性的新兴技术②，生成式人工智能技术带来的生产力解放或将进一步推动人类社会的繁荣发展③，但技术的不确定性与不稳定性也将对社会产生显著影响。④自标志性产品ChatGPT发布以来，生成式人工智能的发展与治理已成为各国乃至全球重大政治经济利益和人类发展进步的关键议题。

本文以“生成式人工智能”“治理”“人工智能治理”“生成式人工智能治理”等为关键词，使用CiteSpace软件在CNKI数据库中选取核心期刊、CSSCI期刊进行文献检索及分析，并以AI-Generated Content、AIGC、Generated AI为关键词，在Web of Science数据库中选取2021年以来生成式人工智能相关研究。通过逐层筛选和逐步聚焦的分析思路，在治理视域下梳理、归纳、总结当前国内外对于生成式人工智能治理的已有讨论。通过对现有研究的解构与再建，探索未来生成式人工智能的治理之策。

（一）生成式人工智能的研究热点与发展趋势

当前，学界就生成式人工智能的探讨基本上已覆盖了哲学、经济学、法学、教育学、文学、理学、工学、军事学、管理学等学科门类，主要集中在计算机科学与技术、公共管理、新闻与传播、行政法及地方法制、教育理论与教育管理、图书情报与数字图书馆、知识产权法等学科领域，既探讨了生成式人工智能对数字政府及电子政务⑤、知识生产与学科建设、智能传播与大众传播⑥等的推动作用，也对此技术带来的算法与数据使用⑦、人机伦理争议⑧、网络空间安全⑨等冲击作了分析。作为一项全球性重要议题，生成式人工智能的快速发展也引起了海外学者的普遍关注。以“AI-Generated Content”为关键词于Web of Science数据库中检索2021年至2024年8月的相关研究，共获得510篇有效文献，主要集中在计算机科学（151篇）、编程（108篇）两大领域，并对社会发展、大众传播、公共管理等学科和领域有所探讨。其中，对深度学习、大语言模型等新兴技术的影响和未来发展格外关注。可见，尽管各个学科侧重不同，但都从不同维度关注了生成式人工智能的风险挑战。

本文所指的生成式人工智能全球治理，主要指向全球代表性国家及学者对于生成式人工智能的治理，但也涉及技术的“国际治理”。以“生成式人工智能”为主题和关键词在CNKI数据库中筛选核心期刊及CSSCI论文，经过CiteSpace筛选过滤重复文献并梳理关键词聚类节点后，本文基于386份有效文件梳理了2021—2024年度学界对生成式人工智能的研究情况。

---

① 《中共中央关于进一步全面深化改革 推进中国式现代化的决定》，2024年7月21日，https://www.gov.cn/zhengce/202407/content_6963770.htm，2024年11月2日。

② Allison Stanger, et al, “Terra Incognita: The Governance of Artificial Intelligence in Global Perspective,” *Annual Review of Political Science*, vol. 27, 2024, pp. 445-465.

③ James Manyika and Michael Spence, “The Coming AI Economic Revolution: Can Artificial Intelligence Reverse the Productivity Slowdown?” *Foreign Affairs*, vol. 102, no. 6, 2023, pp. 70-86.

④ Daniele Rotolo, Diana Hicks and Ben R. Martin, “What is an Emerging Technology?” *Research Policy*, vol. 44, no. 10, 2015, pp. 1827-1843.

⑤ 何哲等：《ChatGPT等新一代人工智能技术的社会影响及其治理》，《电子政务》2023年第4期。

⑥ 彭兰：《从ChatGPT透视智能传播与人机关系的全景及前景》，《新闻大学》2023年第4期。

⑦ 张红春、章知连：《从算法黑箱到算法透明：政府算法治理的转轨逻辑与路径》，《贵州大学学报》（社会科学版）2022年第4期。

⑧ 曾一果：《人工智能迷思与数字技术伦理的现实建构》，《新闻与写作》2023年第4期。

⑨ 郁建兴、刘宇轩、吴超：《人工智能大模型的变革与治理》，《中国行政管理》2023年第4期。

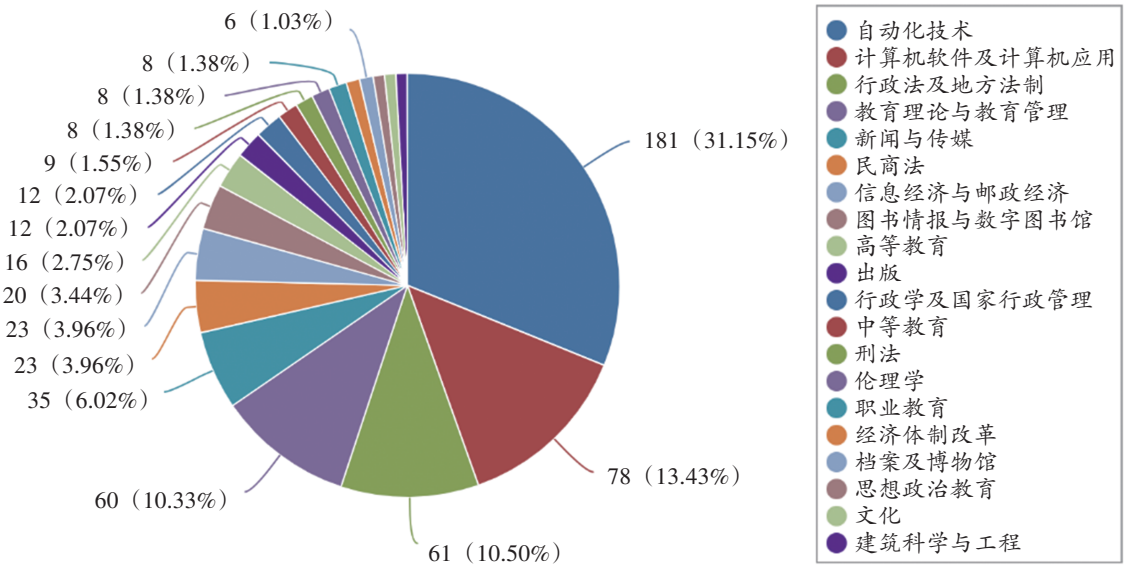

图1　2024年度CSSCI以“生成式人工智能”为主题的文献发表情况

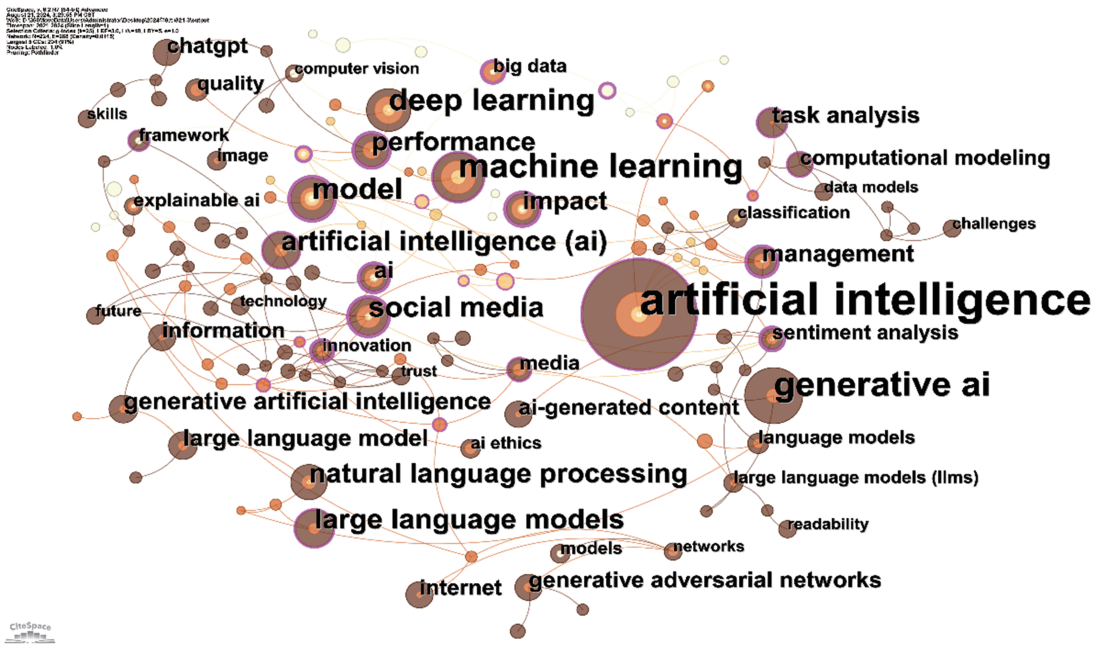

图2　2021—2024年度WOS以“AI-Generated Content”为主题的文献发表情况

通过对关键词聚类的进一步整合，既有研究呈现三个特点。一是当前对生成式人工智能的研究具有多主体、分散式的特征，以“人工智能”为核心，关注基于技术发展带来的机遇和挑战，呼吁合理使用新技术。二是着重探讨风险规制、风险治理、敏捷治理、数据安全、国家安全等议题，对风险和治理的关注是目前学界研究生成式人工智能的重点。三是尽管对社会治理、技术风险、技术革命、主体性等有所讨论，但相对而言并未形成热点。

生成式人工智能治理的重要性与必要性已备受关注。当前，从中国的《生成式人工智能服务管理暂行办法》①、欧盟的《人工智能法》到英国的《人工智能白皮书》，全球范围内提出的人工智能治理准

①《生成式人工智能服务管理暂行办法》，2023年7月10日，https://www.gov.cn/zhengce/zhengceku/202307/content_6891752.htm，2024年2月13日。

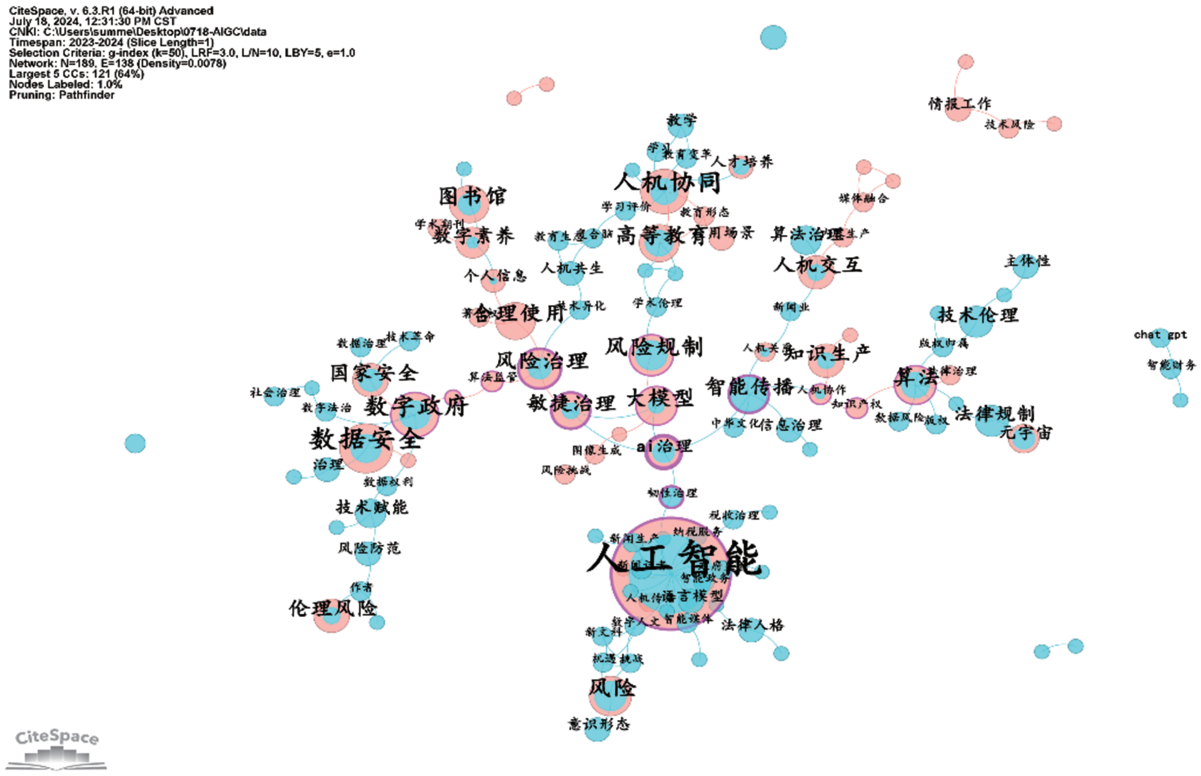

图3　生成式人工智能的研究热点

则已达160多份。新技术的出现激发和推动着新的治理议题的发展。随着生成式人工智能技术的成熟、应用和推广，其风险与治理已备受学界热议。以“治理”为主题和关键词，在CNKI数据库中筛选核心期刊及CSSCI论文，经过CiteSpace筛选过滤重复文献并梳理关键词聚类节点后，本文基于1751份有效文件，梳理了2015至2024年关于治理理论、治理策略的研究进路。可见，随着大数据、算法等技术快速发展，对生成式人工智能及相关人工智能技术的关注已成为当前治理研究领域的重要议题。

（二）生成式人工智能治理的研究进路与核心议题

目前，生成式人工智能治理的研究进路可划分为三个方面。首先，梳理生成式人工智能相关风险，并进行类型学概括和总结，如“责任性风险”“失控性风险”“侵权性风险”“歧视性风险”“社会性风险”。其次，探析生成式人工智能风险的生成原因。有学者以ChatGPT为例，指出生成式人工智能风险的生成源自数据库建立的真假双重性与匿名性、人机交互模式下用户隐私的消匿性、算法技术的单向性、技术沉溺的诱导性以及工具理性逻辑的价值判断真空性。①最后，探析生成式人工智能风险的治理向度，并提出相应治理策略。尽管不同学者所提出的治理模式设计不尽相同，但都将研究焦点集中在治理理念、治理主体、治理策略和治理系统四个方面。

一是在治理理念上如何协调安全与发展的关系。当前生成式人工智能的治理逻辑仍较为离散和模糊，主管部门的“技术风险焦虑”造成了某种程度上的治理资源“浪费”和失衡。②这就需要明确生成式人工智能的技术特性和社会可信之间的内在关联，形成对“技术迭代–场景变革–风险涌现”治理逻辑的认知与共识，确定不同风险等级的划分标准，建立具有开放性的风险认知体系。

二是在治理主体上如何兼顾主流与支流的平衡。从政府、企业、个人到第三方组织，如何合理地协调、统筹不同主体之间的权责分配关系是生成式人工智能治理的关键。③不同治理主体的博弈影响着协

① 张诗濠、李赟、李韬：《ChatGPT类生成式人工智能的风险及其治理》，《贵州社会科学》2023年第11期。
② 邓建鹏、马文洁：《人工智能“逐案设法”治理模式的优化》，《南京社会科学》2024年第6期。
③ 郭海玲、卫金金、刘仲山：《生成式人工智能虚假信息协同共治研究》，《情报杂志》2024年第9期。

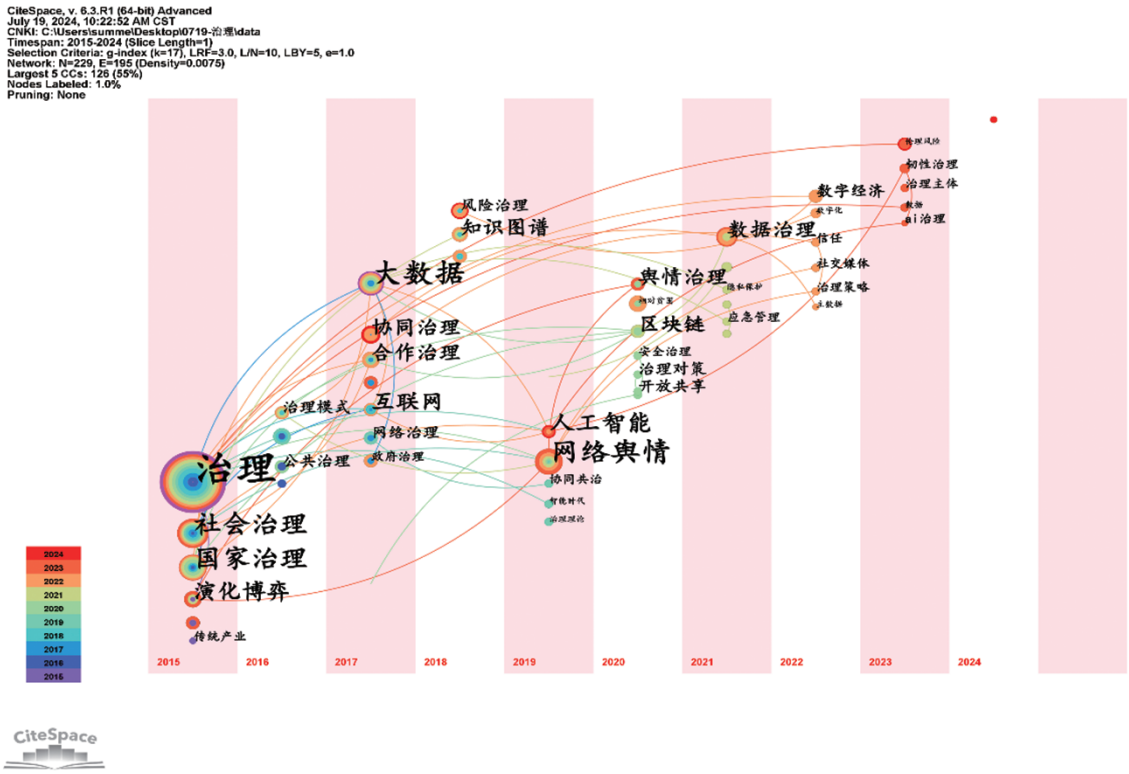

图4　2015—2024年治理领域的研究趋势和研究热点

同治理的稳定性[①]，在不确定环境下造成了治理主体的疲软和失措[②]，人工智能产业政策如何演进仍需进一步探析。[③]

三是在治理策略上如何统筹柔性工具与刚性工具。从架构性风险看，生成式人工智能治理需要实现突破审慎探索的有限实践。在关于技术可控与技术失控的争议张力下，基于生成式人工智能多元系统治理的共识已基本形成，对“柔性工具”与“刚性工具”的有机组合与调配仍待探索。[④]

四是在治理系统上如何协同地区与国别间的合作。人工智能时代的社会秩序乃至国际秩序发生着深刻变化，其重构过程与技术治理密切相关。[⑤]同时，组织协同的秩序成为生成式人工智能迭代更新中的新议程，对既往发展模式造成全新冲击。生成式人工智能日益成为大国之间博弈的重要端口，在国内政策掣肘和国际冲突加剧的对抗态势下[⑥]，生成式人工智能治理合作已成为技术向善发展的重中之重。

二、生成式人工智能的风险与挑战

生成式人工智能具有弥散性、裂变性、融合性，其技术应用则具有扩张性、敏捷性和智能性，它在推动传播、教育、金融、军事等领域智慧化发展时，也带来了数据安全、经济安全、文化安全、国家安全等风险挑战，引起了学者们的关注与担忧。当前，关于生成式人工智能风险的探讨主要集中在算法风险、数据风险、技术创新风险、产业可持续发展风险、生命伦理与人机情感风险、智能鸿沟与信息贫

① 王治莹等：《考虑政府监管机制的虚假信息治理三方演化博弈分析》，《安全与环境学报》2023年第12期。
② 郭文斌：《基于大数据分析的互联网信息时代公共安全治理方法——评〈面向国家公共安全的互联网信息行为及治理研究〉》，《中国安全科学学报》2023年第11期。
③ 朱齐宇等：《不确定环境下我国人工智能产业政策的演化形成研究》，《现代情报》2025年第2期。
④ 姜李丹、薛澜：《我国新一代人工智能治理的时代挑战与范式变革》，《公共管理学报》2022年第2期。
⑤ 戚凯、崔莹佳、田燕飞：《美欧英人工智能竞逐及其前景》，《现代国际关系》2024年第5期。
⑥ 赵申洪：《全球人工智能治理的困境与出路》，《现代国际关系》2024年第4期。

困风险[①]、知识产权与产品责任风险、国家主权风险、国际协作风险等。基于风险的递进性、复合性特征，我们可以将上述风险分为技术层风险、内容层风险、应用层风险和社会层风险，这四类风险彼此交叉、耦合、衍变和互动。

技术层风险。生成式人工智能技术包含数据、算法、算力三大要素，目前衍生于这三大要素的风险已初露峥嵘。一是在数据要素上，生成式人工智能在收集、筛选、标注、加工、再造海量数据的过程中，数据已不单单是生产资料[②]，也产生了隐私泄露、数据造假和“数据投毒”，在实践过程中产生网络信息生态风险[③]；因不同应用所抓取数据库的不同，生成式人工智能的再创造具有明显偏差。二是在算法要素上，算法黑箱、算法歧视、算法偏见、算法后门等技术的“异化”催生了价值失衡。[④]随着技术资源日益成为重要的生产要素，生成式人工智能技术及其应用的工具理性已经超越价值理性，智能算法已成为权力斗争的附庸。三是在算力要素上，算力基础设施存在国别、地区，甚至软硬件的不均衡发展与巨大差异，更因日益增长的算力需求而成为高消耗、高污染的行业，影响范围已辐射至环保等多个领域。

内容层风险。基于生成式人工智能技术而产生的内容导致虚假信息大量涌现，并由此造成意识形态风险迅速扩张。以大语言模型为代表的生成式人工智能在海量数据、先进算法、强大算力的加持下建构了前所未有的大型语义网络，在运行中以强大的信息编码能力塑造了“作为意识形态的表达符号和编码工具”的意识形态话语[⑤]，并作为最新技术力量发挥着对意识形态话语的引导、凝聚、动员、约束、辩护乃至异化等功能，造成了“黑镜”中的对垒。[⑥]一是缘起于技术路径，人工智能技术与政治的深度互嵌已衍生出科技势能向政治势能转变的趋势[⑦]，掌握核心技术的国家其影响力、渗透力、辐射力显著增强[⑧]，在生成式人工智能技术所重塑的意识形态话语国际传播格局中，划分了大国政治与意识形态博弈的新边界。二是诞生于应用场景，话语主体的裂变式繁殖冲击与消解着主流意识形态的权威性，实体意识形态话语权在向网络虚拟空间转化的过程中，“信息茧房”进一步离散了话语内容[⑨]，窄化乃至偏化了认知与价值。三是并发于技术融合、人机交互过程中产生了新型意识形态战场，基于沉浸式交互场域的意识形态柔性塑造，赋予了“数字利维坦”精确又隐秘开展意识形态场域斗争的新焦点。[⑩]

应用层风险。生成式人工智能应用的不均等性导致数字鸿沟与信息贫困快速扩大，不同地区及国家间的发展不平等加剧。近三十年来，数字鸿沟经历了从最初的接入鸿沟到网络使用的素养鸿沟、再到智能时代以数据为核心的智能鸿沟三个阶段，在新兴技术变革之下已演变为物理接入、技术接入到使用接入的巨大差异。[⑪]目前，生成式人工智能大模型等的应用快速扩大了数字鸿沟，从国家内部到国家之间都存在人工智能发展失衡的问题。《世界互联网发展报告2022》指出，发达国家与发展中国家之间仍存在较深的“数据鸿沟”，在互联网接入、普及、使用、合作等方面差异显著。[⑫]以联合国最近公布的全球80亿人口规模进行计算，全球仍有近30亿人未接入互联网。这种近乎对半分的“鸿沟”已随着人工智能技术的竞争、创新、发展而急剧扩大，数字不平等与社会不平等高度融合，信息贫困中环境贫困与自身贫困复杂交织[⑬]，“数字穷人”的生存空间将不断被挤压，为全球人工智能领域发展的平等化、健

① 张夏添：《生成式人工智能技术与国际传播新格局》，《世界社会主义研究》2024年第9期。
② 张文魁：《数据治理的底层逻辑与基础构架》，《新视野》2023年第6期。
③ 侯东德、张丽萍：《生成式人工智能背景下网络信息生态风险的法律规制》，《社会科学研究》2023年第6期。
④ 邓伯军：《人工智能的算法权力及其意识形态批判》，《当代世界与社会主义》2023年第5期。
⑤ 冯冉、陈锡喜：《系统观念视域下新时代社会主义意识形态话语创新的三个着力点》，《湖北社会科学》2023年第7期。
⑥ 向征：《“黑镜”中的对垒：生成式人工智能背景下网络意识形态风险与防范》，《社会科学战线》2024年第4期。
⑦ 欧阳林洁、张永红：《生成式人工智能应用的意识形态风险：命题由来、生成机制与治理进路》，《学术探索》2023年第11期。
⑧ 杨章文：《ChatGPT类生成式人工智能的意识形态属性及其风险规制》，《内蒙古社会科学》2024年第1期。
⑨ 王珺：《新技术变革下意识形态治理研究————理论检审、现实叩问与治理出路》，《马克思主义研究》2023年第2期。
⑩ 王海威：《人工智能诱发隐性意识形态话语风险的逻辑机理及化解策略》，《马克思主义研究》2024年第4期。
⑪ 张笑、孙典：《再谈“数字鸿沟”：新兴技术关注度与社会公平感知》，《科学学研究》2024年第10期。
⑫ 参见中国网络空间研究院：《世界互联网发展报告2022》，北京：电子工业出版社，2022年。
⑬ 张小倩、张月琴、杨峰：《国内外信息贫困研究进展：内涵阐释、研究设计及内容综述》，《图书馆论坛》2018年第8期。

康化、和谐化带来严峻挑战。

社会层风险。总览人类社会，生成式人工智能革命下人的“异化”与科技伦理困境引人担忧。生成式人工智能是一种尚未成熟且迅速发展的革命性、颠覆性新技术，作为整个社会的基本技术支撑正发展成为一种全新的异己力量。一方面，技术的解放力量在使事物工具化后成为技术解放的桎梏，使人趋于工具化。智能科技对人类的宰制使人沦为智能社会系统中的“奴隶”与“附庸”，技术应用的加速使人愈发被捆绑于空间异化、时间异化、自我异化和社会异化之中。另一方面，智能机器人在模糊人机界限的同时对人的主体地位带来冲击①，何以为人的问题愈发突出。人机关系的改变触及了人的本质问题，“人”的概念的模糊消解着哲学层面的自洽，颠覆着“人”所构建的社会。

此外，不同于仅仅观照技术视域下多学科“嫁接式”的风险构想，有学者已关注到人与技术互动时公众对生成式人工智能风险的看法与态度，其结合LDA模型、情感分析、社会网络分析和T-TOE分析模型，基于Reddit平台中以ChatGPT为主题的公众评论文本，对公众对于生成式人工智能的观察角度进行梳理。②分析发现，公众对生成式人工智能的关注，主要涉及学校教育、人机互动、反思讨论、技术变革、运行机制、运用领域等六个方面，尤其关注技术变革和人机互动。尽管整体上公众对生成式人工智能保持乐观态度，但在人机交互等领域也持消极态度。

三、生成式人工智能治理的全球实践

新一轮全球智能浪潮以生成式人工智能的主流化为基本点和增长点，各国已就生成式人工智能的治理提出各具特色的应对办法，对于未来生成式人工智能的安全与发展具有一定的借鉴意义。当前国际层面的生成式人工智能研究主要涉及概念之争、影响分析、应用尝试、未来发展四个议题③，部分学者立足于国际比较视野，分析和总结了欧盟、美国、中国等地区和国家的生成式人工智能治理之策。

（一）生成式人工智能治理的欧盟方案

《欧盟人工智能法案》是走在国际前列的，在治理体系、治理范围、治理机制、治理工具等的设计和选取上有着创新性，已有学者对欧盟生成式人工智能治理的策略展开讨论。部分学者基于法案内容探讨了人工智能立法的必要性及可行性④，提出人工智能立法何以促进技术向善和安全治理。⑤具体来看，有学者从立法、监管和执行三个层面分析了欧盟数据治理的主体，并从个人数据、公共数据、人工智能数据、企业数据和网络数据等治理客体剖析了其治理体系，寻求完善中国数据治理体系之策。⑥有学者指出，受地缘政治影响，欧盟的发展策略在“去风险”框架下侧重于强化抵御外部风险的能力和保证自我权益的效力，通过制度性竞争和扩散性规则抢占人工智能治理“把关人”的关键地位。⑦也有学者指出，《欧盟人工智能法案》采取了基于风险分类的治理路径，在明确禁止不可接受的人工智能技术风险基础上，对高风险人工智能技术进行了相应规制，并根据技术特点划定了风险等级和治理措施。⑧

（二）生成式人工智能治理的美国方案

美国生成式人工智能的技术开发和产业发展处于全球领先地位，其生成式人工智能治理策略呈现总体温和、发展导向、立法迟滞的特点。2023年1月，美国商务部国家标准与技术研究院发布《人工智能

---

① 孙伟平：《人工智能与人的“新异化”》，《中国社会科学》2020年第12期。

② 陈升、刘子俊、张楠：《数字时代生成式人工智能影响及治理政策导向》，《科学学研究》2024年第1期。

③ 朱禹、叶继元：《人工智能生成内容（AIGC）研究综述：国际进展与热点议题》，《信息与管理研究》2024年第4期。

④ 胡铭、洪涛：《我国人工智能立法的模式选择与制度展开——兼论领域融贯型立法模式》，《西安交通大学学报》（社会科学版）2024年第4期。

⑤ 张玉洁：《重述“技术公益主义”：人工智能立法的一般理论与实践》，《法治研究》2024年第4期。

⑥ 方晓娟、陈媛媛：《欧盟数据治理政策法规体系研究》，《数字图书馆论坛》2024年第6期。

⑦ 宋黎磊、陈悦：《“去风险”视域下欧盟人工智能战略的推进及影响——以〈人工智能法案〉为例》，《战略决策研究》2024年第4期。

⑧ 沃尔夫冈·多伊普勒：《〈欧盟人工智能法案〉的背景、主要内容与评价——兼论该法案对劳动法的影响》，《环球法律评论》2024年第3期。

风险管理框架》，以风险划分为基础针对人工智能领域设计了治理工具。2023年10月，美国正式颁布总统行政命令《安全、可靠和值得信赖的人工智能开发和使用》，对生成式人工智能领域的数据安全等作出明确规定。2024年5月，美国参议院人工智能工作组发布《推动美国在AI领域的创新：参议院AI政策路线图》，旨在促进和保持美国在人工智能领域的优势地位。在人工智能领域的政策制定方面，尤其是在中美两国服务大众和保障公平上，有学者在对比中美两国的人工智能国家科技战略决策模型后指出，中国更侧重多元主体协作下稳定的技术发展，而美国更偏向制衡式决策，且在国际视野和站位中具有较为突出的领先优势。①也有学者指出，美国在技术上的“上半场优势”与中国在发展上的“下半场优势”，正在技术、产品、服务、应用和资本等方面展开较量，并将随着用户规模的转移和生态链的博弈而产生不同的竞争优势。②

（三）生成式人工智能治理的中国方案

中国在生成式人工智能治理领域的实践探索与理论研究已走在国际前列。早在2017年，中国就发布了《新一代人工智能发展规划》，2019年国家新一代人工智能治理专业委员会提出以发展“负责任的人工智能”为核心的八项治理原则。2023年，《生成式人工智能服务管理暂行办法》等关注生成式人工智能健康发展的文件陆续出台，从技术发展、风险治理、服务规范、监督检查和法律责任等方面，为生成式人工智能的可持续发展提供了政策保障。中国的生成式人工智能治理方案明确了监管技术、服务和产业的相关规则、标准与要求，在应对监管和治理的复杂性时，充分保障政策执行的一致性。《中华人民共和国网络安全法》《中华人民共和国科学技术进步法》《生成式人工智能服务管理暂行办法》等构成了生成式人工智能治理的基础，在分类分级监管策略下进行风险评估和有效干预。有学者指出，中国未来的生成式人工智能治理应加快标准化转型，推进可持续发展，应对全球竞争格局③，进一步构建包容、审慎的规范导向，实现发展与稳定的动态平衡。④

生成式人工智能的技术复合性、应用流动性和风险广泛性，对私域和公域都带来深远影响，传统治理方式难以应对新的挑战与要求，具体问题如下：

一是治理结构僵化。⑤传统的科层制治理结构难以承载生成式人工智能发展带来的开放性、灵活性和不确定性。因为，在生成式人工智能技术开发的前端和过程中，相应科技伦理的认定与规范尚不完善，事前监管较为薄弱。⑥与此同时，生成式人工智能的应用存在着法律主体模糊不清、智力成果难认证等复杂问题，因尚未形成系统性法律法规，现阶段存在事后监管缺失等风险。⑦

二是治理方法滞后。既有法律体系的建构是建基于人类行为的因果关系，难以直接适用于由数据和算法构成的智能环境。作为新兴技术的典型代表，生成式人工智能具有“一管就死、一放就乱”的产业发展特征，若不及时跟进治理策略，将对新兴技术的健康发展造成影响。

三是治理范围狭隘。作为人类社会首个融合多类技术特性的新兴技术，生成式人工智能带来了远超新兴技术本身的规范难题，技术的快速应用与扩散已带来数据主权、自主武器、深度伪造等新议题。

四是治理合作错位。从生成式人工智能的国际治理看，世界各国尚不具有绝对的实践经验与主导能力，国内议题与全球议题间的传导性加强，新旧技术治理范式间存在摩擦，难以形成包容、开放和可对话的全球性治理体系。当前，生成式人工智能的国际治理形成了以软性规则引领治理、以多元主体引导监管、以私营部门为参与主体、以科技专家为治理顾问的四大特征。⑧但技术先发国家与技术后发国家

---

① 岳昆、房超：《中美人工智能国家科技战略决策模式比较研究——基于多元决策视角》，《中国科技论坛》2023年第1期。

② 方兴东、钟祥铭、黄浩宇：《Sora冲击波后中美AI差距研判——新一轮智能浪潮中美“半场优势”分析模型与趋势》，《西北师大学报》（社会科学版）2024年第5期。

③ 汝鹏等：《智能引领未来：生成式人工智能的社会影响与标准化治理》，《电子政务》2025年第1期。

④ 王一臻、何伏刚：《生成式人工智能风险治理与模式完善研究》，《科学决策》2024年第11期。

⑤ 贾开、蒋余浩：《人工智能治理的三个基本问题：技术逻辑、风险挑战与公共政策选择》，《中国行政管理》2017年第10期。

⑥ 宋应登、霍竹、邓益志：《中国科技伦理治理的问题挑战及对策建议》，《科学学研究》2024年第8期。

⑦ 龙柯宇：《生成式人工智能应用失范的法律规制研究——以ChatGPT和社交机器人为视角》，《东方法学》2023年第4期。

⑧ 薛澜、赵静：《人工智能国际治理：基于技术特性与议题属性的分析》，《国际经济评论》2024年第3期。

均不具备全球人工智能治理经验，生成式人工智能领域的研究与生产已呈现明显的地理分散性，其国际治理也增加了“主权国家之间战略博弈”等因素。目前，生成式人工智能在军用和民用上的界限是模糊的，国际社会也尚未形成防止人工智能技术“军备竞赛”的机制，这些都将深刻影响着人类命运共同体的未来发展。

四、生成式人工智能治理的理论建构

“治理”在学术话语、政治话语及生活话语中的用法存在差异，跨界使用这一概念则加剧了概念的模糊性[①]，治理理论也在“治道”和“治术”上陷入了不同流派的理论争议。[②]随着技术迭代，生成式人工智能已不再仅仅是物质转化为组织产出的活动，其所带来的社会冲击具有基础性、广泛性、综合性和颠覆性，是对人类社会建构及其运行状态的根本变革。

基于不同的价值判断，乃至于不同的文化背景、治理传统、组织结构和决策方式，不同国家和地区的生成式人工智能治理路径有所不同，有学者将生成式人工智能治理的逻辑导向分为四种类型。[③]

一是侧重工具主义。工具主义强调技术的高水平发展，以期通过技术自动调试和解决治理中存在的问题。需要注意的是，工具主义所强调的技术，并非绝对“中立”。伴随先进技术发展的技术霸权、技术垄断已“脱钩断链”，很难在技术的自由发展中实现对人类社会的普惠。

二是关注过程主义。过程主义主张对新兴技术应用实施过程控制，并以此实现对社会秩序的干预，逐渐实现预期治理目标。然而，这通常会形成技术强国掌握对治理的绝对话语的现实窘境。

三是重视权变主义。权变主义强调根据具体情况适时调整治理策略，没有一成不变或普世万能的治理模式。目前，欧盟、美国、英国等地区和国家均实施了具有不同侧重的治理方案，强调因势利导，结合本国特点，对生成式人工智能进行阶段性、实验性治理。这些治理方案虽具备一定的灵活性，但也容易导致政策的随机性与不稳定性，在各自为政的“碎片化”治理中难以达成全球共识与普遍合作。

四是依托建构主义。建构主义认为人类的关系结构是由社会共识决定的，霍布斯文化、康德文化和洛克文化塑造了敌人关系、朋友关系和竞争关系。若以建构主义视域下的关系角色定义来解构不同主体之间的相互关系，则难免有削足适履之嫌。因为在治理过程中，三种关系角色往往处于相互切换和不断融合之中。

事实上，上述四种逻辑导向并不是割裂的、离散的，而是以不同程度和不同组合融汇于治理实践中，既彼此促进，又博弈制衡。总的来看，学界关于生成式人工智能治理的理论建构大致可分为敏捷治理、韧性治理、回应型治理和险基治理四种类型。

（一）敏捷治理

2018年，世界经济论坛提出敏捷治理（Aigle Governance）这一概念，并以此对第四次工业革命中的政策制定问题展开讨论。敏捷治理是一种具有灵活性、适应性、流动性、柔韧性的行动策略，遵循“以人为本”的治理理念，旨在探索第四次工业革命中适应技术变革的政策产生、政策审议、政策制定和政策实施，具备参与广泛度、时间灵敏度等特征。[④]在人工智能技术创新背景下，2019年，中国发布《新一代人工智能治理原则——发展负责任的人工智能》，提出了人工智能治理的框架和行动指南，将敏捷治理纳入八项基本原则。2023年，中央网信办发布的《全球人工智能治理倡议》指出，敏捷治理是应对人工智能风险的必要之策。当前，敏捷治理已经成为生成式人工智能治理的典型范式。

在敏捷治理视域下，治理原则确立基于抽象的法律指导，治理关系建构依托互动的监管关系，治理工具选择依据策略的轻重缓急。从政策过程看，走向敏捷治理具有多目标间平衡、动态过程优化、工具灵活转化的三条优化路径。敏捷治理具有政府主导的组织形态、弹性化的结构形式以及穿透式治理的流

① 杨雪冬、季智璇：《政治话语中的词汇共用与概念共享——以“治理”为例》，《南京大学学报》（哲学·人文科学·社会科学）2021年第1期。

② 柳亦博：《治理理论的“视差”：术道分离与术道合一》，《探索与争鸣》2021年第11期。

③ 胡键：《全球数字治理：理论问题、价值目标和治理工具》，《国际经贸探索》2024年第7期。

④ 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。

程模式三大特征。[①]有学者指出，敏捷治理以全面性治理格局、适应性治理机制、灵活性治理工具实现了治理范式更新。[②]有学者认为敏捷治理模式需要采取多元合作互动、预防与应对并举的治理模式。[③]在具体实践中，敏捷治理的重要载体是“分类”理念，即基于不同类型的人工智能风险而匹配差异化的治理工具。按照“问题界定–分类框架–政策工具箱”的分析思路，有学者在敏捷治理范畴内提出分类治理的框架及与之适配的政策工具箱。在技术维度下基于数据特异性的强弱，分别就稳定性、正确性、效率性的功能目标和隐私性、公平性、透明性的价值目标设计了数据库建设、定期披露要求等工具；在业态维度下基于系统自主性的强弱，针对事前、事中、事后设计了准入要求、保存记录、责任分配等工具。[④]

（二）韧性治理

为应对生成式人工智能日益复杂且不确定性极高的风险与挑战，关注公共安全的“韧性治理”被引入生成式人工智能的治理框架中。韧性治理以治理为研究焦点，强调治理体系和治理能力的适应性、整体性、开放性、包容性，以稳健的制度系统为核心框架、灵活的政策系统为行为指导、弹性的组织系统为行动载体，通过健全完整、多样嵌套及学习演进的治理工具，保障治理系统在常态与非常态间的有序运行与衔接转换。[⑤]

有学者基于“韧性–脆弱性–韧性治理”和“TOE”分析框架，在基础理论与技术实践的双向支撑中提出生成式人工智能的“韧性治理”。TOE（（techonology-organization-environment），即“技术–组织–环境”的分析框架，是一种紧密围绕技术应用多元场景的通用性、综合性、灵活性、可操作性的分析工具，从组态视角强调某种结果的产生是多要素相互影响并协同发挥作用的非线性、非单一性、非因果对称的机制效应。在生成式人工智能技术迭代发展的背景下，“韧性治理”与TOE框架的互嵌与合意，体现在以系统性、演进式、开放性、多元化的治理路径应对生成式人工智能的风险与挑战。

生成式人工智能的韧性治理具备三个特征。一是以“技术”治理“技术”，充分发挥生成式人工智能的正向效能而促进相关技术的自适改进、自我更新、功能触达和融合赋能。二是以“组织协作”引领“组织决策”，强调多元主体的优势互补，在组织化协同的基础上精准施策，以短期、中期、长期的组织化协作动态应对生成式人工智能的衍生风险。三是以“环境营建”应对“环境挑战”，在保障技术创新发展的基础上建立健全保障机制，建构更具韧性的治理环境。[⑥]

（三）回应型治理

生成式人工智能的回应型治理思路缘于“回应性监管”理论。回应性监管理论认为只有融合政府监管与非政府监管的混合模式，对监管治理权进行合理分配，并选择“金字塔”型监管工具来实现最佳监管效果。[⑦]回应性监管以“回应”为代表特征，以“塑造”为价值内核，以“协同”为监管手段，以“关系型”为理论基石。在回应性监管概念下，监管主体间的权力分配与让渡形式主要包括三方主义（tripartism）、强化型自我监管（enforced self-regulation）以及部分行业监管（partial industry intervention）。相较于传统的“命令–控制”型监管，回应性监管通过引入第三方公共利益集团、向监管对象分配一定的监管权、加强某部分监管对象的竞争力，来实现监管制度的平衡性、市场环境的公平性、监管对象的活跃性。

近些年来，回应性监管理论被广泛应用于社会安全多个领域，并随着生成式人工智能的快速崛起与发展而进入学者的视野。在生成式人工智能的数据安全领域，有学者以“回应型治理”统筹生成式人工

① 何宇华、李霞：《生成式人工智能虚假信息治理的新挑战及应对策略——基于敏捷治理的视角》，《治理研究》2024年第4期。

② 张凌寒、于琳：《从传统治理到敏捷治理：生成式人工智能的治理范式革新》，《电子政务》2023年第9期。

③ 赵梓羽：《生成式人工智能数据安全风险及其应对》，《情报资料工作》2024年第2期。

④ 薛澜、贾开、赵静：《人工智能敏捷治理实践：分类监管思路与政策工具箱构建》，《中国行政管理》2024年第3期。

⑤ 容志、宫紫星：《理解韧性治理的一个整合性理论框架——基于制度、政策与组织维度的分析》，《探索》2023年第5期。

⑥ 许源源、陈智：《新一代人工智能技术社会化应用的脆弱性风险及其韧性治理研究——以ChatGPT为例》，《电子政务》2023年第9期。

⑦ 杨炳霖：《回应性监管理论述评：精髓与问题》，《中国行政管理》2017年第4期。

智能的数据主权、意识形态安全、网络安全及国家安全等方面的系统性风险，要求秉持算法透明原则，建构风险管控机制、多层次数据保障体系及以标准、法律、技术为框架的治理体系。①有学者认为，回应型治理是以人工智能应用的影响结果为治理对象，从风险管理视角对人工智能治理提出流程性或实质性要求。如欧盟2021年发布的《人工智能法》（提案）就对人工智能应用风险进行分级分类，并区分了禁止性应用和高风险应用。在界定其内涵的同时，针对高风险应用提出了程序性要求（认证、救济、责任、备案、登记等），实质性要求（质量管理、信息披露、风险管理等）。

（四）险基治理

在传统风险与新兴风险并发、系统性风险与局部性风险交织的当下，随着生成式人工智能的爆发，全球风险社会加速到来，“风险”日益成为监管与治理的源点，以风险为基型监管治理——有学者将其简称为险基监管（risk-based regulation）——正日渐兴起。②险基监管以承认国家干预能力和干预责任的有限性为前提，从以风险为对象跨越至以风险为工具，依风险而干预而非仅仅确保绝对安全，提倡以监管的理性化而实现监管优化，致力于精准有效地实施监管策略。从构成上看，险基监管包括事先设定风险控制的监管目标和可接受的风险容量（risk appetite），合理评估风险发生的概率及预期后果，根据不同的风险级别对监管对象展开不同强度和不同频率的监管干预。③险基监管随着实践发展而逐步演进，发展出了真正的回应性险基监管（really responsive risk-based regulation），提倡从监管对象的行为与态度、监管工具的策略、监管体制的绩效、监管制度的动态变化等维度提升监管干预的针对性。在根据风险级别制定监管策略的理念下，险基监管模型也逐渐覆盖高风险和低风险对象。有学者提出了良好监管干预设计矩阵（Good Regulatory Intervention Design，GRID），倡导根据监管对象风险的动态变化而采取具有灵活弹性的监管策略。④风险评级是险基监管理论的决策基础。随着针对多主体的风险信息披露机制在提升监管效能方面的广泛应用，该机制已发展为该监管框架的重要补充工具。

险基监管以“风险”为理论的建构点和生发点，以对风险的预防、判断、分级、约束、监督、治理为行动指针。尽管其生发于“监管”领域，但衍生于此的“风险”及其相关工具也被广泛应用于治理范畴中。有学者基于38份政策文本的扎根分析，梳理了生成式人工智能数据的采集、存储、标注、运输、输出和销毁六个阶段的风险，提出了生成式人工智能数据风险的治理路径。⑤在以“人工智能”“风险”“治理”为关键词检索中国政府出台的生成式人工智能相关规范性政策后，有学者对生成式人工智能的风险进行类型学分析，深入探讨了风险类型与治理工具的匹配性，以期能对风险进行有效治理。⑥

五、生成式人工智能治理的未来展望

生成式人工智能仍在高速发展，生成式人工智能治理将在动态发展中不断调试、演进和革新。2024年6月，习近平总书记在致世界智能产业博览会的贺信中指出，“中国愿同世界各国一道，把握数字化、网络化、智能化发展机遇，深化人工智能发展和治理国际合作，为推动人工智能健康发展、促进世界经济增长、增进各国人民福祉而努力”。⑦2024年7月，李强总理在出席世界人工智能大会暨人工智能全球治理高级别会议时指出，人工智能发展迫切需要世界各国深入探讨，凝聚共识，共抓机遇，共克挑

① 斜晓东：《论生成式人工智能的数据安全风险及回应型治理》，《东方法学》2023年第5期。

② 刘鹏、张崙楠、王力：《基于风险的政府监管：理论发展与实践应用》，《中国行政管理》2024年第3期。

③ Henry Rothstein, Oliver Borraz and Michael Huber, “Risk and the Limits of Governance: Exploring Varied Patterns of Risk-Based Governance Across Europe,” *Regulation & Governance*, vol. 7, no. 2, 2013, pp. 215-235.

④ Julia Black and Robert Baldwin, “When Risk-Based Regulation Aims Low: A Strategic Framework,” *Regulation & Governance*, vol. 6, no. 2, 2012, pp. 131-148.

⑤ 徐伟、何野：《生成式人工智能数据安全风险的治理体系及优化路径 ——基于38份政策文本的扎根分析》，《电子政务》2024年第10期。

⑥ 陈少威、吴剑霞：《人工智能治理风险和工具的识别与匹配研究——基于政策文本的分析》，《中国行政管理》2022年第9期。

⑦ 《共创共享 携手并进——习近平主席致2024世界智能产业博览会贺信引发业界强烈共鸣》，2024年6月21日，https://www.gov.cn/yaowen/liebiao/202406/content_6958719.htm，2024年10月1日。

战。①如何有效治理生成式人工智能风险，如何推动技术应用向善发展，事关人类福祉。在未来，对生成式人工智能治理的研究可从以下方面着力：

一是平衡定性与定量的关系。当前生成式人工智能风险研究呈现显著的认识论分野——在风险本体论的建构过程中，质性分析的范式主导性与量化验证的方法论缺位形成结构性矛盾。现有风险分类框架多囿于宏观叙事的先验性推定（如风险类型学的主观赋义），在风险存在性的实证检验、风险层级的可计算建模、损害后果的指标化测度以及治理效能的循证评估等关键维度，存在认识论层面的方法论失衡。这要求建构全要素数据追踪体系，通过技术生命周期的观测性学习与反事实因果推断，实现风险表征从质性推演向数理实证的范式跃迁。

二是关注理念与道路的撕裂。在技术地缘政治重构与文明发展梯度并存的当代语境下，技术演进路径的伦理锚定、安全边际的认知共识、治理协同的集体行动框架等核心议题，正遭遇深层的价值共识分裂与制度路径分歧。当前的治理实践呈现出双重悖论：一方面，多元主体通过制度创新试图建构普适性技术治理公域；另一方面，在大国竞合的技术主权化趋势中，治理体系正加速形成策略性区隔与制度性壁垒。这种差异正导致学术研究面临价值理性抉择——在国家安全范式与技术利他主义的矛盾性中，治理优先序的差异化配置已成为不可回避的事实。

三是链接实践与理论的时差。技术风险具象化与制度建构的异步性、社会实践先行与理论阐释的滞后性，在此轮智能革命中呈现显著张力。生成式人工智能的技术奇异性及其应用场景的超复杂性，已突破传统治理范式的解释边界——无论中国抑或海外，产业实践正以指数级速度催生多维技术模态，形成治理认知的“解释赤字”。既有的规制工具不仅面临动态适配困境，更存在与技术迭代周期同步失焦的范式危机。更深层的挑战在于，传统治理的认知图式与价值坐标，在技术异化催生的社会形态跃迁中，正遭遇价值理性与工具理性的结构性错位。

四是填补设计与落实的断层。既有研究虽在多中心治理框架下提出了多元主体协同、政策工具组合等理论模型，但治理实践的本质远非抽象的理论建构。作为嵌入国家历史发展轨迹与全球竞合格局的复杂系统，其实际运作始终面临制度惯性、利益协调与技术异化等多重现实约束。当前研究尚未有效回应“理想模型-现实情境”的转化难题——精巧设计的治理方案往往在落地过程中遭遇实践变形，而执行阶段涌现的新矛盾又会反向解构既有理论预设。这种双向互动关系持续要求学术界深化对治理动态性的认知。

五是弥合域内与域外的隔离。当前生成式人工智能治理研究呈现显著不均衡，相较于对单一国家治理框架的密集探讨，跨国协同机制的理论建构略显滞后。生成式人工智能技术固有的流动性本质与风险外溢特征，已然塑造了诸多超越主权疆界的全球性公共议题。然而，在地缘技术竞争态势下，关键行动者之间尚未建立系统性、制度性、多层次的对话机制，这种治理对话的碎片化状态与技术的统合性风险形成了鲜明对比。

总的来看，生成式人工智能的技术、产业、生态发展在高歌猛进之下，仍具备一定程度的不确定性与模糊性，以期毕其功于一役的治理策略已不现实。现有讨论以及就生成式人工智能治理提出多维度多角度多向度的探析，在未来或可在此五类视域下继续着力，以构建统筹发展与安全、活力与秩序、效率与公平的治理模式，保障和促进生成式人工智能的向善发展。

（责任编辑：陈　果）

---

①《李强出席2024世界人工智能大会暨人工智能全球治理高级别会议开幕式并致辞》，2024年7月4日，https://www.gov.cn/yaowen/liebiao/202407/content_6961222.htm，2024年10月2日。