生成式人工智能治理研究报告

（2026 年）

中国人民大学人工智能治理研究院
生成式人工智能研究团队

程絮森	信息学院
杨建梁	信息资源管理学院
阮神裕	法学院
殷佳敏	商学院
郑敏睿	公共管理学院
塔娜	新闻学院
王裕平	新闻学院

摘　要

本报告围绕生成式人工智能的发展应用与治理问题展开系统研究，立足大模型、多模态模型、智能体和具身智能加速演进并向经济社会各领域渗透的时代背景，分析生成式人工智能对生产方式、产业组织、法律责任、人机关系、医疗服务、公共治理和新闻传播的影响。当前，生成式人工智能已由内容生成工具发展为连接知识生产、任务执行和组织协同的重要基础能力，治理对象也由模型本体延伸至训练数据、生成内容、应用场景、责任链条和社会影响。系统梳理其发展图景、典型风险与治理路径，有助于推动我国形成促进创新、保障安全、权责清晰、以人为本的治理体系。

本报告从八个方面梳理生成式人工智能发展的主要进展、应用图景、突出问题和治理路径，分别是智能经济发展、中外治理进展比较、侵权责任、人智协作、智慧医疗、社会空间分析与公共治理、双向人机对齐以及新闻传播治理。

在生成式人工智能与智能经济发展方面，生成式人工智能正在将自然语言、多模态理解、代码生成、工具调用和智能体执行能力嵌入研发、生产、流通与消费过程，推动生产流程、创新方式、商业入口和产业组织加快重构。当前，高质量行业数据供给不足、算力成本上升、模型可靠性欠缺以及智能体责任边界模糊，仍然制约高价值场景的深度应用。未来应加强可信基础设施建设，完善数据、版权和责任规则，推动场景牵引式应用，并提升中小企业与劳动者的人工智能应用能力。

在中外生成式人工智能治理进展比较方面，2025—2026 年治理重点已由模型准入扩展至模型、数据、内容、场景、评测和责任链条的复合治理。我国政策具有地方扩散速度快、场景牵引特征强、能力供给工具丰富等特点；欧盟、美国、英国、新加坡和日本分别形成法律义务、创新基础设施、公共评测、软法工具以及企业指南与经济安全相结合的治理路径。我国下一阶段需要补强模型持续评测、事故报告、训练数据与版权合规、内容凭证、重点场景分级治理以及跨区域和国际规则互操作机制。

在生成式人工智能侵权责任方面，信息交互型人工智能可能引发著作权、不正当竞争和人格权侵害，决策交互型人工智能可能在招聘、信贷和消费等场景中形成偏见、歧视和不合理差别待遇。现行法律原则上可以通过过错责任回应相关侵权，但合理注意义务、通知后的必要措施以及多主体责任划分仍需细化。报告提出，应建立覆盖模型训练、系统部署、服务运营和内容传播的分层注意义务，完善“通知—必要措施”机制，并依据风险创设、控制、获益和防范能力识别各类主体责任。

在生成式人工智能人智协作方面，人类的创造力、价值判断和情感感知正与生成式人工智能的计算、数据处理和规则化执行能力加速融合，形成能力协同、价值协同和动态协同相结合的发展形态。人智协作已进入工业、医疗、教育、金融和公共服务等领域，但仍面临模型黑箱与幻觉、意图理解不足、权责体系不完善、复合型人才短缺和数字鸿沟等问题。未来

应加强核心技术研发，完善全流程监管和责任追溯机制，强化人才培养，并推动算力、数据和技能资源向基层地区与中小企业下沉。

在生成式人工智能智慧医疗方面，应用场景已经覆盖居民健康助手、临床诊疗辅助、医疗文书、医学影像、药物研发和公共卫生治理。医疗生成式人工智能正由技术展示进入政策牵引、产业落地和治理塑形并行推进的新阶段，评价重点也将转向真实临床环境中的安全性、有效性和可持续性。针对医疗幻觉、临床证据不足、数据安全和算法偏倚等问题，未来应加强本地验证、真实世界监测和不良事件报告，推动全生命周期监管，并促进相关应用向基层医疗、家庭健康管理和连续照护拓展。

在生成式人工智能社会空间分析与公共治理方面，地理大模型、空间语言模型和合成空间数据正在推动空间分析向预测、推演和辅助决策拓展，并在国土空间规划、城市治理、应急管理和公共服务设施布局中展现应用潜力。与此同时，“空间幻觉”、算法黑箱和历史数据偏见可能削弱公共决策的准确性、可解释性与公平性。未来应构建权责边界清晰的空间智能制度体系，发展监管沙盒、第三方审计、空间一致性检查和风险预警工具，保留重大公共决策中的人类审核与干预，并促进空间智能公平普惠应用。

在生成式人工智能时代双向人机对齐方面，报告将对齐理解为人工智能适配人类意图、语境和价值，以及人类认识和调适人工智能能力与局限的动态互动过程。提示词兼具意图表达、语境建构和规范嵌入功能，是推动意图、语境、风格、情感和价值对齐的关键媒介。双向对齐受到文化差异、语言模糊性、模型能力边界以及效率与安全等张力影响。治理应覆盖提示输入、模型响应和多轮互动全过程，完善风险识别与交互审计机制，提升公众提示素养，并防范过度信任、情感依附和思维外包。

在生成式人工智能新闻传播治理方面，大模型、智能体、AI 搜索和多模态生成正在重塑新闻采集、生产、分发与消费机制。生成式人工智能能够提高新闻生产效率和跨语种传播能力，也可能带来虚假信息、来源错配、版权争议、流量收益失衡、合成内容泛滥和公众信任流失等风险。未来应坚持人类编辑责任，建立新闻机构内部人工智能分级使用制度，对生成、辅助、改写、翻译和情景演绎内容实行分类标识，强化 AI 新闻中介平台的来源、授权和补偿责任，建设内容凭证、数字水印等真实性基础设施，并提升公众 AI 新闻素养。

展望未来，生成式人工智能将进一步转向智能体执行、流程嵌入和跨系统协作，治理重点也将由一次性准入扩展至训练、部署、更新、调用和退出全过程。持续评测、事故报告、场景分级、人工复核和责任追溯将成为重要制度安排，可信数据、公共算力、评测平台、内容凭证和合规工具将构成规模化应用的基础条件。未来治理应贯穿以人为本的价值导向，保障人类监督、专业责任、公平可及和公众信任，推动生成式人工智能安全、可信、普惠发展。

关键词：　生成式人工智能　智能经济　复合治理　侵权责任　人智协作　智慧医疗　空间智能　双向人机对齐　新闻传播治理

目　录

第一章 生成式 AI 与智能经济发展前景

程絮森[1,2] 黄天乐[1]

1.中国人民大学信息学院 北京 100872

2.中国人民大学人工智能治理研究院 北京 100872

生成式人工智能（Generative AI，以下简称生成式 AI）成为智能经济发展的关键变量。与上一阶段以感知识别、推荐匹配、流程自动化为主的人工智能应用不同，生成式 AI 的核心影响不只是“生成内容”，而是把自然语言、多模态理解、代码生成、工具调用和智能体执行能力嵌入研发、生产、流通、消费和治理过程。2026 年政府工作报告首次提出打造智能经济新形态，使智能经济从技术应用议题上升为经济形态重构议题。随着“人工智能+”行动全面部署、大模型备案和应用登记持续推进、智能体与行业模型加速落地，生成式 AI 已经从单点工具进入规模化应用阶段。本章重点讨论生成式 AI 如何推动智能经济新形态发展、重点场景中已经出现哪些真实进展、当前制约何在，以及如何通过治理安排促进其可信、可控、可持续地赋能智能经济。

一、生成式 AI 推动智能经济新形态从政策部署走向规模落地

（一）政府工作报告首提智能经济新形态，明确发展方向

2026 年政府工作报告在“加紧培育壮大新动能”部分提出打造智能经济新形态，并把深化拓展“人工智能+”、促进新一代智能终端和智能体推广、推动重点行业领域人工智能商业化规模化应用、培育智能原生新业态新模式、支持开源社区建设、实施超大规模智算集群和算电协同等新基建工程、建设高质量数据集、完善人工智能治理等内容放在同一政策框架中部署[1]。智能经济不再只是数字经济中的一个应用方向，而是围绕人工智能技术、智能终端、数据要素、算力基础设施和产业场景形成的新型经济形态。

从政策演进看，智能经济新形态承接了数字经济、平台经济和“人工智能+”行动的既有基础，但不再停留于“人工智能赋能某个环节”。它更强调人工智能作为核心引擎，数据作为关键要素，算法作为核心生产力，算力作为新型基础设施，通过人机协同、跨界融合和共创分享重塑产业形态、就业结构和生活方式[2]。因此，讨论生成式 AI 与智

[1] 新华社：《快讯：政府工作报告提出，促进新一代智能终端和智能体加快推广》，教育部转载，2026 年 3 月 5 日，https://www.moe.gov.cn/jyb_xwfb/moe_1946/2026/202603/t20260305_1430052.html

[2] 经济参考报：《政府工作报告首提“打造智能经济新形态”释放什么信号？》，陆家嘴金融网转载，2026 年 3 月 10 日，https://www.ljzfin.com/info/86885.jspx。

能经济发展前景，关键是说明生成式 AI 如何成为智能经济新形态中最具扩散力和重构力的技术载体。

智能经济新形态的发展前景主要体现在生产方式智能化、产业形态智能原生、基础设施一体化和治理要求前置化四个方面。研发设计、生产排程、供应链协同、质量控制和客户服务将逐步由模型和智能体参与组织，智能终端、智能体、具身智能、行业大模型和智能服务平台也将形成新的产品形态和商业模式。

（二）战略牵引强化，智能经济从政策概念转向发展目标

2025 年 8 月，国务院发布《关于深入实施“人工智能+”行动的意见》，明确提出推动人工智能与经济社会各行业各领域广泛深度融合，加快形成人机协同、跨界融合、共创分享的智能经济和智能社会新形态，并围绕科技、产业、消费、民生、治理、全球合作等六大领域部署重点行动。该文件还提出，到 2027 年新一代智能终端、智能体等应用普及率超过 70%，到 2030 年超过 90%，到 2035 年全面步入智能经济和智能社会发展新阶段[3]。这意味着，生成式 AI 不再只是互联网平台和内容产业的技术热点，而被纳入新质生产力培育、产业体系升级和社会治理能力提升的中长期框架。

从智能经济的形成机制看，生成式 AI 的特殊性在于其直接作用于“知识生产—任务执行—组织协同”链条，能够以自然语言作为通用接口，把企业文档、业务系统、专业模型和外部工具连接为面向具体任务的智能工作流。

（三）应用数据快速增长，生成式 AI 进入常态化供给阶段

备案与登记数据反映出我国生成式 AI 服务供给正在迅速扩容。截至 2024 年 12 月 31 日，共 302 款生成式人工智能服务完成备案，2024 年共有 105 款调用已备案模型能力的应用或功能完成登记；截至 2025 年 12 月 31 日，累计已有 748 款生成式人工智能服务完成备案，435 款生成式人工智能应用或功能完成登记；2026 年 1 月至 2 月又新增 48 款服务备案、46 款应用或功能登记，截至 2026 年 2 月 28 日累计达到 796 款备案服务和 481 款登记应用或功能[4]。这一组数据说明，生成式 AI 的市场供给已从少数通用聊天工具，扩展到办公、教育、营销、研发、客服、工业、医疗、金融等多场景服务矩阵。

使用侧的增长同样显著。2024 年初我国日均 Token 消耗量约为 1000 亿，截至 2025 年 6 月底已突破 30 万亿，1 年半时间增长 300 多倍；同期我国已建设高质量数据集超

[3] 国务院：《关于深入实施“人工智能+”行动的意见》，2025 年 8 月 26 日， 。

[4] 国家互联网信息办公室：《国家互联网信息办公室关于发布 2025 年生成式人工智能服务已备案信息的公告》，2026 年 1 月 9 日，https://www.cac.gov.cn/2026-01/09/c_1769688009588554.htm；国家互联网信息办公室：《关于发布生成式人工智能服务已备案信息的公告（2026 年 1 月至 2 月）》，2026 年 3 月 17 日，https://www.cac.gov.cn/2026-03/17/c_1775482074695536.htm；国家互联网信息办公室：《国家互联网信息办公室关于发布 2024 年生成式人工智能服务已备案信息的公告》，2025 年 1 月 8 日，https://www.cac.gov.cn/2025-01/08/c_1738034725920930.htm。

过 3.5 万个，总体量超过 400PB，各地高质量数据集累计交易额近 40 亿元，数据交易机构挂牌的高质量数据集总规模达到 246PB[5]。Token 消耗量可以被视为大模型应用活跃度的重要侧面，高质量数据集则构成行业模型和专业智能体落地的底座。二者同步增长，说明生成式 AI 应用已经从“用户尝鲜”走向“产业调用”，并开始形成模型调用、数据供给、场景反馈相互促进的循环。

（四）智能体和开源生态降低门槛，推动生成式 AI 成为通用生产工具

2025 年以来，生成式 AI 能力演进的突出方向是从“回答问题”走向“完成任务”。智能经济的关键矛盾正在从“有没有模型”转向“能不能把模型嵌入业务流程”。智能体之所以重要，正在于其可以围绕具体目标进行任务拆解、工具调用、结果校验和跨系统执行，使生成式 AI 从知识问答工具转化为业务流程中的“行动单元”。

开源模型生态则进一步降低了生成式 AI 扩散门槛。DeepSeek-R1 于 2025 年 1 月发布，模型采用 MoE 架构，总参数 671B、每个 Token 激活参数 37B，并提供多个蒸馏版本，降低了企业、开发者和研究机构部署推理模型的门槛[6]。阿里通义千问开源生态也在 2025 年快速扩张，截至 2025 年 9 月已开源 300 余个模型，覆盖语言、代码、图像、语音、视频等多模态能力，全球下载量突破 6 亿次，衍生模型超过 17 万个[7]。开源模型和低成本推理共同推动智能经济从“大厂专有能力”转向“中小企业可调用能力”，也使行业模型、私有化部署和本地化智能体具备更强可行性。

二、生成式 AI 重塑智能经济的增长机制

（一）重塑生产方式，从局部提效转向流程再造

生成式 AI 首先改变的是企业内部知识工作的组织方式。在办公、客服、软件开发、营销、法务、财务、人力资源等环节，生成式 AI 能够完成信息检索、文本起草、代码生成、材料汇总、表格分析和方案比较等任务，把分散在不同岗位、系统和文档中的知识重新组织为可调用的流程能力。企业竞争力不再只取决于是否购买大模型服务，而取决于能否围绕订单处理、研发管理、供应链调度、质量控制和客户服务等高频流程重构数据和责任链条。

[5] 国家数据局：《国家数据局：截至今年 6 月底 日均 Token 消耗量已经突破 30 万亿》，2025 年 8 月 15 日，https://www.nda.gov.cn/sjj/swdt/mtsy/0815/20250815194322139344581_pc.html。

[6] DeepSeek-AI, “DeepSeek-R1,” GitHub repository, accessed May 2, 2026, https://github.com/deepseek-ai/DeepSeek-R1。

[7] 《通义千问 Qwen 荣获“领先科技奖”》，2025 年 11 月 7 日，https://developer.aliyun.com/article/1687908。

（二）重塑创新方式，从经验驱动研发走向人机协同创新

生成式 AI 正在改变研发创新的起点和节奏。在科学研究中，大模型可用于文献综述、假设生成、实验方案设计和数据解释；在工业研发中，多模态模型和生成式设计工具可参与产品外观、结构方案、工艺参数和仿真结果的迭代；在软件产业中，代码生成模型已把“自然语言—代码—测试—修复”连接为连续过程。智能经济因此呈现“创新民主化”和“专业化增强”并存的趋势。低门槛模型工具让中小企业、个人开发者和垂直行业专家获得生成、分析和自动化能力；高价值场景则仍要求专业数据、工艺知识、验证体系和安全责任。

（三）重塑商业模式，从平台撮合转向智能体代理服务

生成式 AI 还将深刻影响数字商业的入口结构。过去，用户通过搜索引擎、电商平台、内容平台和应用商店主动查找信息，平台通过流量分发和广告匹配组织交易。随着智能体能力增强，用户可能越来越多地把需求直接交给 AI，由 AI 完成比价、筛选、咨询、下单、售后和跨平台任务执行。智能经济的前端入口因此可能从“人浏览平台”转向“智能体代理人完成任务”。智能体可以降低用户决策成本，提高服务匹配效率，推动个性化消费、智能导购、企业采购、差旅安排、金融咨询等场景升级；同时也可能改变平台流量分配和市场竞争秩序，并因错误推荐、隐性诱导、数据抓取和未经授权交易引发责任争议。

（四）重塑产业组织，模型、数据、算力、场景形成新型协同体系

生成式 AI 推动智能经济发展的深层机制，是把模型、数据、算力、应用场景和治理规则重新组合为新型产业基础设施。模型提供通用认知和生成能力，数据决定行业知识密度，算力决定训练和推理成本，场景提供反馈闭环，治理规则则决定哪些能力能够被稳定、可信、合规地部署。国务院“人工智能+”行动把提升模型基础能力、加强数据供给创新、强化智能算力统筹、促进开源生态繁荣、提升安全能力水平等列为基础支撑，正是因为智能经济不是单一技术突破的结果，而是多要素协同的结果[8]。未来智能经济竞争将不只是模型参数规模竞争，更是数据治理、场景组织、工程落地和可信治理能力的综合竞争，生成式 AI 推动智能经济的价值创造机制如表 1 所示。

表 1 生成式 AI 推动智能经济的价值创造机制

价值创造机制	主要作用对象	经济价值形成方式	适用边界与治理重点
知识生成	文档、知识库、设计方案、代码与多模态内容	将分散知识转化为可检索、可复用、可迭代的方案草稿，降低	适合结构化程度较高的知识工作，需要保留人工复核、引用核

[8] 人民网：《国务院发布关于深入实施“人工智能+”行动的意见：强化 8 项基础支撑能力》，2025 年 8 月 26 日，https://finance.people.com.cn/n1/2025/0826/c1004-40550587.html。

价值创造机制	主要作用对象	经济价值形成方式	适用边界与治理重点
		知识获取和初稿生成成本。	验和版权审查。
流程重组	研发、采购、客服、营销、财务、人力等业务流程	以自然语言接口连接业务系统和工具链，把分散步骤组织为连续任务，提高跨部门协同效率。	需要明确权限边界、审批节点和异常回退机制，避免模型越权执行。
决策辅助	经营分析、供需匹配、风险识别、资源配置	汇总多源信息并生成备选方案，帮助管理者更快识别趋势、比较方案和形成判断。	应定位为辅助判断而非自动决策，关键决策应保留可解释依据和责任主体。
人机协同	一线员工、专业岗位、智能体和行业模型	让员工通过对话式交互调用专业工具，使模型能力嵌入具体岗位和团队协作。	需要开展岗位再设计和能力培训，防止对模型输出形成机械依赖。
服务个性化	教育、医疗、金融、消费、政务等面向用户服务	基于用户需求和上下文生成差异化内容与服务建议，提高服务响应速度和匹配度。	对个人信息保护、公平性、内容安全和服务边界要求更高。

三、生成式 AI 赋能智能经济的重点场景与典型案例

（一）智能制造，从工业知识问答走向生产现场协同

制造业是生成式 AI 进入实体经济的关键场景。2025 年，京东物流“亚洲一号”智能产业园内的“狼族”机器人通过大模型全域感知能力，实现从“被动响应”向“主动预测”的决策升级，可在复杂环境中进行路径规划、动态避让和群体协同，相关机器人队伍已在全球超过500个仓库应用；京东工业打造的工品查智能体可处理大量物料信息，通过多智能体编排自动识别商品属性、补全技术参数、覆盖行业标准并完成打标分类，较传统方式提效超过 10 倍[9]。生成式 AI 在制造和供应链场景中的价值不只是写报告或答问题，而是把“物料知识、空间感知、业务规则、执行系统”连接起来，参与仓储、采购和工业品治理等核心流程。

具身智能也开始从展示走向真实工况。2025 年 7 月，网易灵动在世界人工智能大会发布面向露天矿山挖掘机装车场景的具身智能模型“灵掘”。该模型采用多模态数据驱动的端到端技术路径，训练数据来自真实矿山作业场景；在内蒙古霍林河北露天煤矿环

[9] 《人民日报》：《AI 重构供应链 京东推动人工智能深度应用》，2025 年 9 月 25 日，https://paper.people.com.cn/rmrb/pad/content/202509/25/content_30106966.html。

境中，“灵掘”单机装车效率已达到人工 80%，近 70%作业时间无需人为干预，并适配极寒、高粉尘等复杂环境[10]。生成式 AI 正在从纯软件场景向物理世界延伸，但也提示治理重点必须同步延伸到设备安全、现场责任、异常接管、数据闭环和事故追溯。

（二）智能服务，知识密集型服务进入“专业增强”阶段

金融、法律、咨询、医疗辅助、教育培训等知识密集型服务，是生成式 AI 最容易形成规模化影响的领域。这些业务高度依赖文本、规则、案例和专业判断，既适合大模型进行知识检索、文档生成和风险提示，也要求专业人员对结果进行审核。与传统信息系统不同，生成式 AI 可以直接理解自然语言需求，并通过智能体调用数据库、表格、合规规则和外部工具，形成面向业务目标的辅助决策。

但智能服务不能被理解为“机器替代专家”。在金融投研、医疗问诊、法律审查等场景中，生成式 AI 的优势是快速整理信息、发现关联、生成初步意见和提示风险；其短板则是幻觉、过度自信和责任承担能力不足。因此，智能服务的可持续发展依赖“人机协同责任链”，模型负责生成和辅助，专业人员负责判断和签署，机构负责场景准入、日志留存、质量评测和用户告知。

（三）智能消费，从内容生成走向个性化服务和智能导购

消费场景是生成式 AI 扩散最快、用户感知最强的领域。2025 年以来，文本生成、图像生成、视频生成、数字人直播、智能客服、个性化学习、智能导购等应用持续增长。通义千问、DeepSeek、豆包、Kimi 等模型和应用降低了公众使用门槛，企业也开始把生成式 AI 接入营销策划、商品素材生成、广告投放和用户运营。通义千问开源模型覆盖语言、编程、图像、语音、视频等多模态能力，下载量和衍生模型数量快速扩张，这为大量消费端应用提供了低成本模型底座[11]。

智能消费的深层变化，是消费服务从“标准化推荐”转向“交互式生成”。用户可以用自然语言表达偏好、预算、场景和约束，由生成式 AI 生成个性化方案。旅游、教育、家装、健身、保险、汽车等低频高决策成本消费，都可能因智能体介入而改变用户决策路径。与此同时，虚假宣传、AI 生成内容版权、未成年人保护、诱导式推荐和深度合成诈骗等风险也会随之放大。2025 年 9 月 1 日起施行的《人工智能生成合成内容标识办法》要求对 AI 生成合成内容进行显式和隐式标识，为智能消费场景中的内容可信和平台责任提供了重要制度工具[12]。

---

[10] 光明网：《全球首个工程机械具身智能模型“灵掘”发布》，2025 年 7 月 28 日，https://tech.gmw.cn/2025-07/28/content_38178517.htm。

[11] 《阿里云：2025 年通义大模型全球下载量 6 亿次，衍生模型 17 万个》，2025 年 9 月 24 日，https://finance.sina.com.cn/tech/roll/2025-09-24/doc-infrrepi6431872.shtml。

[12] 中国政府网：《四部门联合发布〈人工智能生成合成内容标识办法〉》，2025 年 3 月 15 日，https://www.gov.cn/lianbo/bumen/202503/content_7014283.htm；中国政府网：《〈人工智能生成合成内容标识办法〉助力辨别虚假信息 推进从生成到传播全链条治理》，2025 年 3 月 19 日，https://www.gov.cn/zhengce/202503/content_7014404.htm。

（四）智能研发与开源生态，低成本模型推动创新扩散

智能经济的另一类典型场景，是生成式 AI 作为研发基础设施进入开发者和企业创新过程。DeepSeek-R1 的开放权重和蒸馏模型，使许多企业能够以较低成本获得推理能力，在代码生成、知识库问答、行业助手、自动化分析等场景中进行私有化部署或二次开发[13]。通义千问开源生态的大规模下载和衍生模型，则说明开源模型正在成为智能经济中的“能力公共品”，开发者可以基于通用模型快速构建垂直应用，企业也可以在保证数据安全的前提下进行定制化训练和部署。

开源生态具有双重效应。一方面，它降低了创新门槛，推动中小企业和行业机构参与生成式 AI 应用开发；另一方面，模型能力开放也可能带来滥用、越权调用、恶意自动化和安全评测不足等问题。因此，发展开源生态还要配套模型许可、能力分级、应用备案、安全评测、漏洞披露和行业自律机制。

四、生成式 AI 发展智能经济面临的关键制约

（一）高质量数据供给不足，行业知识难以充分模型化

生成式 AI 进入智能经济深水区后，瓶颈不再只是通用语料和算力规模，而是行业数据、专业知识和业务反馈能否被合法、准确、安全地转化为模型能力。制造、医疗、金融、能源、交通等行业积累了大量高价值数据，但往往存在格式不统一、质量参差、权属不清、跨部门流通困难、隐私和商业秘密保护要求高等问题。若缺乏高质量数据集和可信数据空间，行业模型容易停留在“会说行业术语”的层面，难以真正理解工艺流程、风险约束和业务规则。数据治理还涉及版权和个人信息保护。生成式 AI 训练和检索增强过程中可能使用公开网页、企业文档、用户交互数据和专业数据库，如果授权基础、使用边界和收益分配机制不清，企业会因合规不确定性降低应用意愿。

（二）算力成本与能源约束上升，中小主体应用能力分化

生成式 AI 的训练和推理都依赖高强度算力。随着多模态模型、推理模型和智能体应用普及，推理调用频率显著上升，企业面临的不只是训练成本，还包括持续服务成本、响应延迟、模型更新和安全监测成本。对大型平台而言，算力投入可以通过云服务和生态规模摊薄；对中小企业而言，模型调用成本、私有化部署成本和人才成本可能成为应用门槛。若缺乏公共算力、行业共性模型、低成本开发平台和面向中小企业的服务包，生成式 AI 可能强化企业之间的数字鸿沟。

[13] TechCrunch, “DeepSeek claims its ‘reasoning’ model beats OpenAI’s o1 on certain benchmarks,” January 27, 2025, https://techcrunch.com/2025/01/27/deepseek-claims-its-reasoning-model-beats-openais-o1-on-certain-benchmarks/。

（三）模型可靠性不足，制约高价值场景深度落地

生成式 AI 在开放语境中具有很强表达能力，但在产业场景中，表达流畅并不等于结果可靠。模型幻觉、事实错误、偏见输出、提示注入、工具调用错误、对异常工况不敏感等问题，都会影响其在金融、医疗、制造、政务、交通等高价值场景中的应用。尤其是在智能体模式下，模型不仅生成建议，还可能调用外部工具、修改数据、触发交易或控制设备，错误输出的后果会从信息层面扩展到现实层面。因此，生成式 AI 赋能智能经济必须建立“可评测、可审计、可追溯、可干预”的可靠性体系。低风险场景可以通过提示、标识和用户确认降低风险；中高风险场景则需要引入红队测试、基准评测、人工复核、权限控制、日志留存和异常熔断等机制。

（四）智能体应用带来新的责任、竞争和安全问题

智能体是生成式 AI 进入智能经济的关键形态，也是治理难度最高的形态。传统软件按照固定规则运行，责任主体相对清晰；智能体则可能根据自然语言目标动态规划路径、调用多个工具、与其他系统交互，并在不同平台之间完成连续任务。一旦出现错误下单、泄露数据、误操作设备、生成侵权内容或规避平台规则等问题，责任应由模型提供者、智能体开发者、部署企业、平台方还是用户承担，仍需要更清晰的制度安排。智能体还可能重塑市场竞争，改变搜索、电商、广告、内容平台和企业软件的流量分配方式，并引发数据权益、平台公平接入和网络安全问题。

五、面向生成式 AI 驱动智能经济的治理建议

（一）建设可信生成式 AI 基础设施，降低产业应用公共成本

推动生成式 AI 赋能智能经济，首先要把可信基础设施建设作为公共能力来推进。建议围绕公共算力、行业高质量数据集、模型评测平台、内容标识与溯源、智能体安全测试、合规工具包等方向形成基础支撑。对中小企业和科研机构，可以通过算力券、模型服务券、行业共性数据集和低代码智能体平台降低应用成本；对高风险行业，则应建设统一的模型安全评测和场景准入标准。可信基础设施还应服务于监管和产业双方，使监管部门能够基于统一标准识别模型能力、应用风险和事件责任，也使企业获得可操作的合规路径。

（二）以场景牵引推动“生成式 AI+产业”深度落地

生成式 AI 应用不能只追求“模型上新”和“概念展示”，而应围绕真实产业问题开展场景牵引。建议优先选择制造研发、工业质检、供应链协同、医疗辅助、金融风控、政务服务、教育培训、科研创新等价值高、数据基础较好、责任边界可设计的场景开展试点。每个试点都应同步形成行业模型、数据规范、评测指标、责任流程和安全预案，

避免“只做应用、不建规则”。场景牵引还应强调复用，推动成功应用沉淀为行业知识库、流程模板、评测集和合规指引，把生成式 AI 从示范项目转化为普惠能力。

（三）完善数据、版权与责任规则，稳定企业应用预期

智能经济需要稳定的制度预期。建议进一步明确生成式 AI 训练数据、检索增强数据和用户交互数据的合规使用边界，推动公共数据、行业数据和企业数据在可授权、可计量、可追溯条件下流通。对版权问题，可探索训练使用、生成输出、相似性判断和收益分配的分类规则，既保护权利人合法权益，也避免过度不确定性抑制模型创新。责任规则方面，应区分模型提供者、应用开发者、部署机构和用户的不同控制能力与收益程度。对于智能体自动执行、专业服务辅助和物理设备控制等场景，应建立更严格的授权确认、人工接管和责任追溯机制。

（四）提升劳动者和中小企业生成式 AI 应用能力，防止智能经济能力分化

生成式 AI 既会创造新岗位，也会改变既有岗位能力结构。智能经济的治理目标不应只是防范替代风险，还应帮助劳动者和中小企业获得新工具。建议面向重点行业建立生成式 AI 应用培训体系，将提示工程、模型校验、数据安全、内容标识、智能体使用和人机协同纳入职业技能培训。对中小企业，可由行业协会、产业园区和公共服务平台提供模板化智能体、行业知识库、合规问答和案例库，降低其从“不会用”到“用得好”的门槛。同时，应建立生成式 AI 对就业和组织结构影响的动态监测机制。对于入门级岗位、流程性知识工作和外包服务等受影响较大的领域，需要提前开展岗位转型和再培训；对于模型测试、智能体运维、行业模型评测等新岗位，应加强职业标准和人才培养。

（五）推动开放生态与安全治理协同发展

开放生态是生成式 AI 赋能智能经济的重要优势。我国应继续支持开源模型、开放评测、行业联盟和开发者生态，鼓励企业、高校、科研机构和中小开发者围绕垂直场景形成模型共创机制。但开放并不意味着无边界。建议建立开源模型能力分级、风险提示、许可证合规、漏洞披露和滥用举报机制，对高风险能力、敏感数据和关键基础设施场景实施更严格管理。国际层面，应围绕开源模型安全、内容标识、智能体责任、数据跨境合规等议题提出可操作方案，推动形成兼顾创新、安全和普惠的生成式 AI 治理框架。

第二章 中外生成式 AI 治理进展比较

杨建梁[1,2] 徐子杭[1] 郑庆雯[1] 刘思思[1]

1.中国人民大学信息资源管理学院 北京 100872

2.中国人民大学人工智能治理研究院 北京 100872

2025 年以来，生成式 AI 治理的重点已由模型准入和服务规范，逐步扩展至模型、数据、内容、场景、评测和责任链条的复合治理。本报告基于北大法宝人工智能相关政策文本，经人工筛选形成我国人工智能治理核心政策样本，并结合政策工具分析、主题演化分析和国际比较，对 2025—2026 年中外生成式 AI 治理进展进行系统研判。研究发现，我国政策呈现地方扩散速度快、场景牵引特征强、能力供给工具丰富的特点，治理主题正从早期基础能力建设和场景试验，转向大模型、数据语料、内容标识、行业应用、评测标准和安全责任的联动治理。国际上，欧盟侧重法律义务，美国突出创新基础设施，英国强调评测证据，新加坡注重软法工具，日本将企业指南与经济安全供给相结合。比较来看，我国已形成以场景扩散和能力供给为主要特征的治理路径，但仍需在模型透明、训练数据与版权合规、第三方评测、事故报告、内容凭证、跨区域执行一致性和国际规则互操作等方面补强。报告建议，下一阶段应把“人工智能+”行动形成的场景优势转化为可评测、可验证、可追责的治理能力，推动生成式 AI 治理从政策动员走向制度化、工具化和闭环化。

一、总体形势：生成式 AI 治理进入复合治理阶段

2025 年以来，生成式 AI 治理的核心问题正在发生结构性变化。早期政策讨论主要集中于大模型备案、生成式 AI 服务规范和算法安全边界，但随着大模型快速进入政务、教育、医疗、工业、金融、交通、文化内容生产和公共服务等领域，治理对象已经从模型本体扩展到训练数据、生成内容、应用场景、算力基础设施、模型评测和责任链条。生成式 AI 治理不再是单一模型监管问题，而是模型、数据、内容、场景和能力体系相互耦合的复合治理问题。

中外政策均反映了这一趋势。欧盟以《人工智能法案》为制度框架，并通过通用人工智能模型义务，将通用模型纳入透明度、版权政策和系统性风险管理框架。[1][2]美国在

[1] European Parliament and Council. Regulation (EU) 2024/1689 laying down harmonised rules on artificial intelligence [EB/OL]. (2024-06-13)[2026-05-03]. https://op.europa.eu/en/publication-detail/-/publication/dc8116a1-3fe6-11ef-865a-01aa75ed71a1/language-en.

[2] European Commission. General-Purpose AI Code of Practice; General-purpose AI obligations under the AI Act[EB/OL]. (2025)[2026-05-03]. https://digital-strategy.ec.europa.eu/en/policies/contents-code-gpai.

2025 年以后突出人工智能行动计划、基础设施建设和国际技术竞争，同时以 NIST 生成式人工智能风险管理框架保留技术治理工具。[3][4][5]英国依托人工智能安全研究所持续开展前沿模型评测，把网络安全、生物化学、自主性和越狱风险转化为公共证据。[6]新加坡通过生成式人工智能治理框架和 Project Moonshot 测试工具，将软法原则转化为企业可执行的评测流程。[7][8]日本则将企业人工智能指南与云服务、算力供给和经济安全相结合。[9][10]

与上述路径相比，我国政策的突出特征是地方扩散速度快、应用场景牵引强、能力供给工具丰富。2025 年以来，地方政策明显从“鼓励应用”进一步延伸到政务服务、教育、医疗、工业、数据语料、算力平台、内容标识和行业合规等具体治理环节。与此同时，我国仍需在通用模型透明、训练数据与版权披露、第三方评测、事故报告、跨区域执行一致性和国际规则互操作方面进一步补强。

本报告的基本判断是：2025—2026 年，生成式 AI 治理的国际竞争焦点已经从“是否监管模型”转向“如何将模型、数据、内容、场景和基础设施纳入同一政策工具组合”。我国下一阶段不宜简单停留在“加强监管”或“促进发展”的二分思路，而应转向提升治理能力密度，即在关键场景中同时具备规则、标准、评测、工具、责任和反馈机制，使生成式 AI 能够被识别、被评估、被纠偏、被追踪和被问责。

二、政策数据与分析方法

本报告政策文本源自北大法宝法律法规数据库，以“人工智能”为检索关键词获取中央和地方相关政策，并经人工筛选保留直接涉及模型监管、算法备案、数据安全、内容标识、场景治理、评测标准、伦理责任和算力能力供给等治理议题的政策文件。经筛选，本报告使用人工智能治理核心政策 351 条，其中 2025 年至 2026 年 4 月政策 166 条；进一步聚焦生成式 AI 治理，得到生成式 AI 治理核心政策 173 条，其中 2025 年至 2026

[3] The White House. Executive Order 14179: Removing Barriers to American Leadership in Artificial Intelligence[EB/OL]. (2025-01-23)[2026-05-03]. https://www.govinfo.gov/content/pkg/FR-2025-01-31/pdf/2025-02172.pdf.
[4] The White House. Winning the AI Race: America's AI Action Plan[EB/OL]. (2025-07-23)[2026-05-03]. https://www.whitehouse.gov/articles/2025/07/white-house-unveils-americas-ai-action-plan/.
[5] National Institute of Standards and Technology. Artificial Intelligence Risk Management Framework: Generative Artificial Intelligence Profile[EB/OL]. (2024-07-26)[2026-05-03]. https://www.nist.gov/publications/artificial-intelligence-risk-management-framework-generative-artificial-intelligence.
[6] UK AI Security Institute. Frontier AI Trends Report[EB/OL]. (2025)[2026-05-03]. https://www.aisi.gov.uk/frontier-ai-trends-report.
[7] Infocomm Media Development Authority. Model AI Governance Framework for Generative AI[EB/OL]. (2024-05-30)[2026-05-03]. https://www.imda.gov.sg/resources/press-releases-factsheets-and-speeches/factsheets/2024/gen-ai-and-digital-foss-ai-governance-playbook.
[8] AI Verify Foundation. Project Moonshot[EB/OL]. [2026-05-03]. https://aiverifyfoundation.sg/project-moonshot/.
[9] Ministry of Economy, Trade and Industry. AI Guidelines for Business Ver. 1.0[EB/OL]. (2024-04-19)[2026-05-03]. https://www.meti.go.jp/english/press/2024/0419_002.html.
[10] Ministry of Economy, Trade and Industry. Approval of Plans for Ensuring a Stable Supply of Cloud Programs[EB/OL]. (2024-04-19)[2026-05-03]. https://www.meti.go.jp/english/press/2024/0419_001.html.

年 4 月政策 127 条。近期生成式 AI 治理核心政策中，地方政策 110 条、中央政策 17 条，表明生成式 AI 治理已从中央规则制定扩展到地方场景落地和行业应用治理。[11]

政策筛选遵循“人工智能对象+治理信号”双重口径。人工智能对象包括生成式 AI、大模型、AIGC、通用人工智能、智能体、算法推荐、深度合成等；治理信号包括治理、监管、合规、伦理、可信、安全、风险、备案、审查、评估、评测、标准、个人信息保护、数据安全、网络安全、内容标识、水印、版权、知识产权、责任、执法、应急等。对仅属于产业扶持、项目申报、奖补兑现、企业培育或一般示范应用的文件，除非文本中具有明确治理对象和治理工具，否则不纳入治理核心政策样本。边界样本由研究团队结合政策标题、发文主体、政策目的和正文条款进行人工复核，重点剔除仅以人工智能为产业方向，但未提出治理规则、治理工具或治理责任的文本。

在分析方法上，报告使用三类方法。第一，使用政策工具框架，将政策文本中的治理工具分为监管约束型、激励财政型、能力供给型、需求应用型和信息协调型。第二，围绕报告主线构建七类治理主题，包括模型监管与算法备案、数据语料与隐私安全、内容治理与标识水印、场景应用与行业治理、评测标准与合规工具、安全责任与监管执法、算力平台与能力供给。第三，按 2017—2019 年、2020—2022 年、2023—2026 年三个阶段构建主题演化桑基图，观察不同阶段治理主题之间的延续、分化和转向关系。主题和工具均采用多标签识别口径：同一政策可同时归入多个主题或工具类别；关键词识别主要用于形成可比统计，边界样本再结合政策语义进行人工校正。因此，本报告的量化结果用于揭示政策结构和变化趋势，不替代逐项政策法律解释。

国外政策部分采用代表性官方政策样本，重点覆盖欧盟、美国、英国、新加坡和日本。国外样本用于比较不同司法辖区的治理路径和政策工具组合，不与我国政策数量进行简单相加或直接数量比较。

三、我国政策主题演化：从基础能力和场景试验走向联动治理

主题分析显示，我国人工智能治理政策的主线并非单向地从“产业促进”转向“风险监管”，而是在不同阶段形成了不同的政策组合。2017—2019 年，政策主题主要集中在智能化应用、基础能力建设、产业平台和场景试验。该阶段治理色彩相对间接，更多体现为通过智能化改造、平台建设和示范应用为后续治理积累场景基础。

2020—2022 年，算法伦理、标准规范、评测工具和行业治理开始上升。该阶段的政策不再仅强调人工智能作为产业技术的应用价值，而开始关注算法、数据、伦理、标准、

[11] 北大法宝法律法规数据库. 人工智能相关政策文本检索数据[DB/OL]. 检索关键词：人工智能；检索范围：中央与地方政策文件；经人工筛选形成本报告政策分析样本. 检索时间：2026-04-30.

规范和责任等治理要素。随着生成式 AI 和大模型技术快速发展，治理对象逐渐从一般智能化应用扩展到模型能力、生成内容、数据安全和行业场景风险。

2023—2026 年，主题明显转向模型、数据、内容、场景、安全责任和算力能力的联动治理。近期政策不仅关注大模型和生成式 AI 服务，也关注数据语料、内容标识、水印、评测标准、政务服务、教育医疗、工业制造、交通运输和城市治理等具体场景。这表明我国生成式 AI 治理已进入“中央规则定边界、地方场景促落地”的扩散阶段。

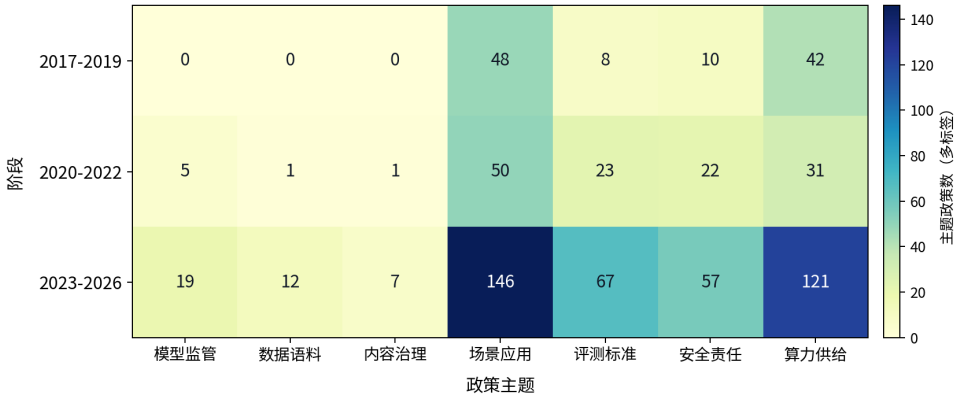

图 1 我国人工智能治理政策主题的阶段分布

图 1 显示，2017—2019 年政策主题主要集中于场景应用与行业治理、算力平台与能力供给；2020—2022 年评测标准与合规工具、模型监管与算法备案开始出现并增强；2023—2026 年场景应用、能力供给、评测标准、安全责任和模型监管同步上升，显示治理对象已经由场景扩散逐步走向复合治理。

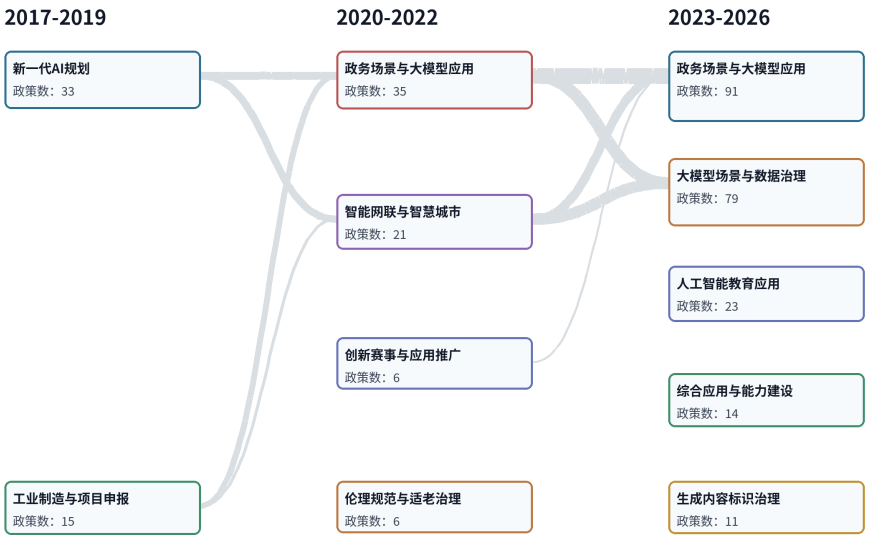

图 2 我国人工智能治理主题演化桑基图

如图 2 所示，阶段主题并非使用同一套固定标签，而是在每一阶段分别提取主要主题，并通过相邻阶段主题词相似度观察主题流动。2017—2019 年的“新一代 AI 规划”“工业制造与项目申报”等主题，主要流向 2020—2022 年的“智能网联与智慧城市”“工业制造与项目申报”等主题；2020—2022 年的产业应用、智能网联和工业制造主题，又进一步流向 2023—2026 年的“政务场景与大模型应用”“大模型场景与数据治理”“人工智能教育应用”和“生成内容标识治理”。这说明，政策主题并不是简单重复，而是经历了从规划引导和产业应用，到场景扩散，再到大模型、数据、内容和行业治理并行展开的演化过程。

这一主题演化具有政策含义。第一，风险治理正在进入具体应用环节。生成式 AI 风险并不只发生在模型训练阶段，也发生在政务问答、教育评价、医疗辅助、金融咨询、广告生成、文化内容生产和公共传播等场景。第二，治理能力越来越依赖公共基础设施。没有公共评测工具、合规指南、数据语料规范、算力服务平台和场景分级制度，原则性规则难以转化为可执行治理。第三，地方政策正在成为生成式 AI 治理落地的重要载体，但也可能带来执行尺度不一致、场景风险识别不足和治理工具碎片化等问题。

进一步看，主题演化并不是简单的“监管增强”过程，而是治理对象的不断外溢。早期政策中的场景应用和能力供给，主要服务于智能化改造和产业升级；中期政策开始引入伦理、标准、规范和评测；近期政策则把模型能力、数据语料、内容标识、应用场景和安全责任同时纳入治理对象。这说明生成式 AI 治理具有强烈的链条属性：模型风险会通过数据来源、内容生成、平台分发和行业应用被放大，也会通过评测、标准、标识和责任机制被抑制。

从政策文本看，近期地方政策中“人工智能+政务服务”“人工智能+教育”“人工智能+医疗健康”“人工智能+制造”“人工智能+城市治理”等表述明显增多。这一变化首先是国家“人工智能+”行动在地方实施层面的政策传导，而不宜简单理解为地方政策自发转向应用导向。国务院《关于深入实施“人工智能+”行动的意见》明确提出，以科技、产业、消费、民生、治理、全球合作等领域为重点，推动人工智能与经济社会各行业各领域广泛深度融合，并对教育教学、医疗健康、城市治理等场景作出部署。[12]因此，地方政策中的“人工智能+”表述，实质上是在把国家层面的场景化部署转化为本地任务清单、应用工程和治理要求。需要强调的是，应用扩散本身并不必然等于治理增强；只有当场景应用同步配置风险评估、责任边界、标准规范、数据合规、人工复核和反馈机制时，才构成治理意义上的场景治理。这类政策的价值不只是推广应用，更重要的是把抽象治理要求转化为场景流程：例如政务场景需要明确辅助决策边界和人工复核要求，

[12] 国务院. 国务院关于深入实施“人工智能+”行动的意见：国发〔2025〕11 号[EB/OL]. (2025-08-26)[2026-05-03]. https://www.gov.cn/zhengce/zhengceku/202508/content_7037862.htm.

教育场景需要关注学生数据、评价公平和未成年人保护，医疗场景需要处理临床责任和专业审查，工业场景需要关注生产安全、知识产权和供应链安全。因此，场景治理具有较强政策基础，也应成为下一阶段制度建设的重点。

如表 1 所示，地区政策主题分布也显示出不同地方政策重心的差异。近期政策中，广西、上海、江苏、北京、山东、河南、浙江、湖北等地出现较多治理核心政策，但主题组合并不完全相同。广西、上海、北京更突出场景应用和算力能力供给，江苏、浙江、湖北在安全责任、评测标准和行业场景治理方面相对突出。鉴于地区样本数量受政策发布频率、数据库收录情况和地方专项政策结构影响，上述分布不能直接等同于地方治理能力高低，其主要意义在于提示地方政策正在围绕产业基础、公共服务需求和数字政府建设重点形成不同侧重。

表 1 近 3 年地区 AI 治理政策主题分布

地区	近期治理核心政策数	主要主题
广西	14	场景应用与行业治理、算力平台与能力供给、评测标准与合规工具
上海	11	场景应用与行业治理、算力平台与能力供给、评测标准与合规工具
江苏	8	场景应用与行业治理、算力平台与能力供给、安全责任与监管执法
北京	7	场景应用与行业治理、算力平台与能力供给、评测标准与合规工具
山东	7	场景应用与行业治理、算力平台与能力供给、评测标准与合规工具
河南	7	场景应用与行业治理、算力平台与能力供给、评测标准与合规工具
浙江	6	算力平台与能力供给、场景应用与行业治理、安全责任与监管执法
湖北	5	场景应用与行业治理、算力平台与能力供给、安全责任与监管执法

四、我国政策工具组合：场景牵引、能力供给与监管约束并行

政策工具分析显示，我国生成式 AI 治理不是单一监管逻辑，而是场景牵引、能力供给、信息协调和监管约束并行展开。本报告依据政策工具理论，将政策文本中的工具划分为五类，并以关键词词典识别政策中是否包含相应工具（见表 2）。该方法不替代法律解释，但可以用于观察政策工具组合的结构特征。

表 2 我国生成式 AI 治理政策工具

政策工具类型	识别逻辑	典型关键词	治理含义
监管约束型	政策是否设置禁止、备案、审查、评估、责任等要求	备案、审查、安全评估、算法备案、个人信息保护、内容标识、水印、处罚	划定风险边界和责任要求
激励财政型	政策是否通过财政资金或市场激励引导主体行为	奖励、补贴、专项资金、政府采购、算力券、服务券	降低合规和应用成本
能力供给型	政策是否建设基础设施、数据、平台、人才和技术能力	算力、数据集、语料、平台、实验室、人才、标准、开源	提供治理和应用能力基础
需求应用型	政策是否通过场景开放和示范应用牵引治理落地	政务、医疗、教育、文旅、制造、交通、金融、示范、试点	将治理要求嵌入具体场景
信息协调型	政策是否通过标准、指南、清单、协同机制降低不确定性	指南、指引、评估、标准规范、清单、目录、专家委员会	形成规则预期和协同机制

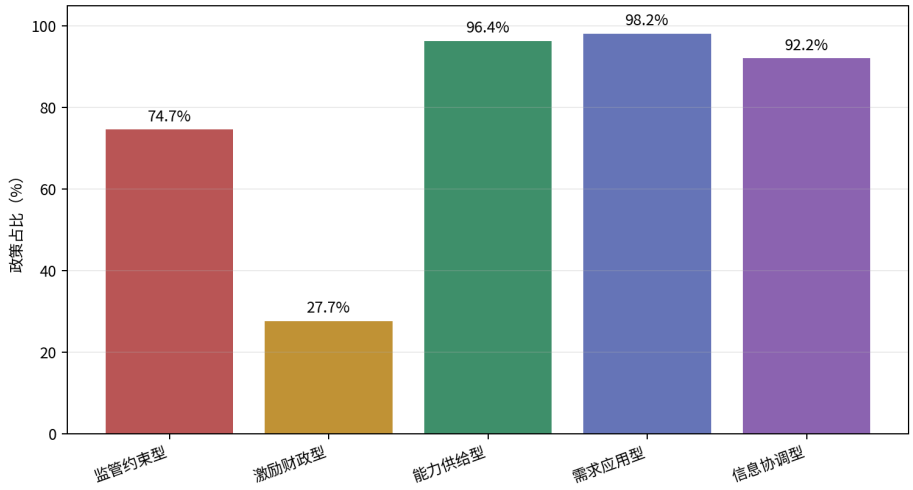

图 3 近期人工智能治理核心政策的政策工具占比

如图 3 所示，在 166 条近期人工智能治理核心政策中，需求应用型、能力供给型和信息协调型工具命中率较高，分别为 98.2%、96.4%和 92.2%；监管约束型工具命中率为 74.7%，说明备案、审查、责任、安全边界等治理要求已经较多嵌入政策文本；激励财

政型工具命中率为 27.7%，相对较低。这一结构表明，近期政策并不是单纯依靠财政激励推动应用，而是更多通过场景任务、能力平台、标准指南和监管约束组合推进治理。图 3 采用多标签统计口径。同一份政策文本如果同时包含备案审查、算力供给、场景示范、财政支持等工具，会被同时计入多个政策工具类别。因此，图中各类工具占比反映的是“含有该类工具的政策占样本政策的比例”，各项相加可以超过 100%。

图 4 政策主题与政策工具关联矩阵

如图4所示，矩阵值表示同时命中某一政策主题和某一政策工具的政策数量。其中，场景应用主题与需求应用、信息协调、能力供给工具共同命中的政策数较高，分别为 112 条、109 条和 108 条；算力供给主题与能力供给、需求应用、信息协调工具共同命中也较为集中，分别为 91 条、91 条和 90 条。结果表明，近期政策正在从单一目标政策转向复合工具组合，场景任务、能力平台、信息协调和监管约束之间的联动关系更加突出。

从中央与地方分工看，相对而言，中央政策更强调规则、备案、标准、安全边界和内容治理，地方政策更强调应用场景、算力平台、数据语料、公共服务和产业生态。二者共同形成我国生成式 AI 治理的基本结构：中央政策更多发挥底线设定和制度框架作用，地方政策更多承担场景转化和应用流程嵌入功能。但这种分工并非绝对，中央也会部署重点场景，地方也会出台标准、评测和监管细则。

这种结构的优势在于，能够迅速推动生成式 AI 进入真实应用场景，并通过算力券、语料券、公共数据、开放平台和试点示范降低应用门槛。其潜在风险也比较清楚：如果

地方政策过度强调场景扩散和能力供给，而模型透明、训练数据合规、内容标识、事故报告和第三方评测不足，就可能形成“应用扩散快于治理能力”的结构性张力。

因此，我国下一阶段的政策重点不应从“促进应用”简单转向“抑制应用”，而应从政策发布密度转向治理能力密度。所谓治理能力密度，是指在关键场景中同步配置规则、标准、评测、工具、责任和反馈机制，使生成式 AI 应用能够被识别、被评估、被纠偏、被追踪和被问责。

政策工具组合还揭示了一个需要警惕的问题：能力供给和场景应用工具过强，而评测、透明和事故反馈工具相对不足，容易形成“前端部署积极、后端治理滞后”的局面。对于生成式 AI 而言，风险往往不是在单一环节集中爆发，而是在模型更新、数据接入、插件调用、用户交互和场景迁移过程中动态生成。如果缺乏持续评测和反馈机制，地方场景越多，治理压力越大。因此，政策工具组合需要从“供给—应用”双轮驱动，升级为“供给—应用—评测—反馈—问责”闭环。

从政策实施角度看，地方政府应避免把生成式 AI 治理简单理解为建设平台、发放算力券或发布场景清单。平台和场景只是治理落地的载体，真正决定治理效果的是场景准入条件、模型调用边界、数据使用规则、人工复核机制、风险监测指标和责任分配机制。对于高风险场景，尤其是政务、医疗、金融、教育和公共传播，应当要求应用单位在部署前完成场景风险评估，在部署中保留日志和人工干预通道，在部署后建立投诉、纠错和事故报告机制。

五、国际比较：主要经济体形成五类治理路径

国外政策样本显示，主要经济体并未沿着同一条路径推进生成式 AI 治理，而是形成了五类具有代表性的政策组合。

欧盟代表法律义务型治理。《人工智能法案》以风险分级为基本框架，对高风险人工智能系统和通用人工智能模型设置合规义务。2025 年以来，通用人工智能模型义务和 GPAI Code 成为欧盟生成式 AI 治理重点，涉及技术文档、版权政策、训练内容摘要、系统性风险识别、风险缓解和严重事件报告。[13][14]欧盟路径的优势在于法律义务清晰，尤其是在通用模型透明度、版权、系统性风险和监管执行方面形成制度化框架；短板是合规成本较高，可能增加中小企业创新负担。对我国而言，欧盟经验的启示不是照搬强监管，而是补强模型透明、训练数据摘要、版权合规和系统性风险文档。

[13] European Parliament and Council. Regulation (EU) 2024/1689 laying down harmonised rules on artificial intelligence [EB/OL]. (2024-06-13)[2026-05-03]. https://op.europa.eu/en/publication-detail/-/publication/dc8116a1-3fe6-11ef-865a-01aa75ed71a1/language-en.

[14] European Commission. General-Purpose AI Code of Practice; General-purpose AI obligations under the AI Act[EB/OL]. (2025)[2026-05-03]. https://digital-strategy.ec.europa.eu/en/policies/contents-code-gpai.

美国代表创新基础设施型治理。2025 年，美国联邦人工智能政策突出创新领导、基础设施和国际竞争。第 14179 号行政令强调移除被视为阻碍人工智能创新的联邦政策壁垒，并要求制定人工智能行动计划；美国人工智能行动计划提出 90 多项联邦政策行动，聚焦加速创新、建设人工智能基础设施、领导国际外交与安全。[15][16]与此同时，NIST 生成式人工智能风险管理框架为组织提供识别、测量、管理生成式 AI 风险的技术工具。[17]美国路线并非没有治理，而是将治理嵌入国家竞争战略。

英国代表评测证据型治理。AI Security Institute 的核心能力是持续测试前沿模型，评估网络安全、生物化学、自主性、越狱、防护有效性等风险，并将评测结果转化为公共证据。[18]英国路径的价值在于，它把治理从规则文本推进到事实生产：监管者、企业和社会可以围绕模型能力、风险边界和防护效果形成共同证据基础。对我国而言，英国经验提示我们，模型治理不仅要有备案和标准，还要建设持续性、公开性、可比较的模型评测体系。

新加坡代表软法框架与测试工具型治理。其生成式人工智能治理框架强调系统、平衡和多方参与，Project Moonshot 则将治理要求转化为大语言模型测试工具，帮助企业开展基准测试、红队测试和报告生成。[19][20]新加坡路径的优势是把抽象原则转化为可执行工具，降低企业合规成本，增强国际互操作。对我国地方政策而言，启示在于：地方和行业政策不仅要提出安全、可信、合规要求，还应配套低门槛、可复用的测试工具和合规流程。

日本代表企业指南与经济安全供给型治理。日本通过面向企业的人工智能指南整合人工智能研发、利用和治理要求，面向广泛人工智能业务经营者提供行为框架；同时从经济安全角度支持云程序和算力供给，强调生成式 AI 对产业活动和社会生活的影响，以及国内计算资源和服务韧性的重要性。[21][22]日本经验说明，生成式 AI 治理并不止于模型规则，也延伸到关键基础设施和计算资源稳定供应。

---

[15] The White House. Executive Order 14179: Removing Barriers to American Leadership in Artificial Intelligence[EB/OL]. (2025-01-23)[2026-05-03]. https://www.govinfo.gov/content/pkg/FR-2025-01-31/pdf/2025-02172.pdf.

[16] The White House. Winning the AI Race: America's AI Action Plan[EB/OL]. (2025-07-23)[2026-05-03]. https://www.whitehouse.gov/articles/2025/07/white-house-unveils-americas-ai-action-plan/.

[17] National Institute of Standards and Technology. Artificial Intelligence Risk Management Framework: Generative Artificial Intelligence Profile[EB/OL]. (2024-07-26)[2026-05-03]. https://www.nist.gov/publications/artificial-intelligence-risk-management-framework-generative-artificial-intelligence.

[18] UK AI Security Institute. Frontier AI Trends Report[EB/OL]. (2025)[2026-05-03]. https://www.aisi.gov.uk/frontier-ai-trends-report.

[19] Infocomm Media Development Authority. Model AI Governance Framework for Generative AI[EB/OL]. (2024-05-30)[2026-05-03]. https://www.imda.gov.sg/resources/press-releases-factsheets-and-speeches/factsheets/2024/gen-ai-and-digital-foss-ai-governance-playbook.

[20] AI Verify Foundation. Project Moonshot[EB/OL]. [2026-05-03]. https://aiverifyfoundation.sg/project-moonshot/.

[21] Ministry of Economy, Trade and Industry. AI Guidelines for Business Ver. 1.0[EB/OL]. (2024-04-19)[2026-05-03]. https://www.meti.go.jp/english/press/2024/0419_002.html.

[22] Ministry of Economy, Trade and Industry. Approval of Plans for Ensuring a Stable Supply of Cloud Programs[EB/OL]. (2024-04-19)[2026-05-03]. https://www.meti.go.jp/english/press/2024/0419_001.html.

上述国际比较表明，生成式 AI 治理并不存在唯一最优模式。欧盟偏重法律义务，美国偏重基础设施和创新竞争，英国偏重评测证据，新加坡偏重软法工具，日本偏重企业指南和经济安全。结合图 3 和图 4 可以看到，我国近期政策中能力供给、需求应用和信息协调工具使用较多，且与场景应用、算力能力、评测标准等主题高度关联，因此更接近“场景扩散与能力供给型”路径。这一模式的优势是组织动员能力强、场景资源丰富、地方试点速度快，能够在较短时间内推动技术进入真实经济社会场景；不足是不同地区和行业的规则颗粒度、评测能力和责任机制可能不一致。下一阶段，我国需要在保持场景扩散优势的同时，把欧盟的透明义务、英国的评测证据、新加坡的工具化合规和日本的基础设施韧性吸收进自身治理体系。

表 3 主要司法辖区的生成式 AI 治理路径及对我国启示

司法辖区	主导路径	主要治理对象	核心政策工具	对我国启示
欧盟	法律义务型	高风险人工智能、通用模型、系统性风险	法律义务、透明文档、版权政策、系统性风险管理	补强模型透明、训练数据摘要和版权合规
美国	创新基础设施型	创新生态、基础设施、国际竞争	行动计划、基础设施、标准框架、采购规则、出口管制	将治理嵌入基础设施、标准和国际竞争
英国	评测证据型	前沿模型能力和安全风险	国家评测机构、红队、越狱测试、能力趋势报告	建设持续性、公开性、可比较的模型评测体系
新加坡	软法工具型	企业部署、大语言模型应用、可信生态	软法框架、测试工具、AI Verify、Project Moonshot	将治理原则转化为低成本工具链
日本	企业指南与经济安全型	人工智能业务经营者、算力和云服务	企业指南、算力补贴、经济安全法工具	把生成式 AI 治理与计算资源韧性结合
我国	场景扩散与能力供给型	模型、数据、内容、场景、行业应用	备案、标准、场景政策、算力平台、数据语料、地方试点	补强透明披露、第三方评测、事故报告和跨区域一致性

六、重点议题：模型、数据、内容、场景和评测的联动治理

第一，模型监管的重点正在从准入备案转向持续评测和系统性风险管理。生成式 AI 模型上线前的备案、评估和安全承诺是必要的，但不足以应对模型能力快速迭代带来的风险。模型在部署后可能因插件、工具调用、微调、检索增强、场景适配和用户交互而产生新风险。因此，应在既有备案制度基础上强化持续评测、版本更新报告、重大能力变化评估和严重事件报告。

持续评测应当覆盖三个层面：一是基础能力评测，包括模型准确性、鲁棒性、可解释性、幻觉倾向和对抗攻击抵抗能力；二是安全风险评测，包括违法有害内容生成、隐私泄露、网络攻击辅助、生物化学风险和自动化代理风险；三是场景适配评测，包括特定行业中的专业可靠性、责任边界、人工复核要求和误用成本。只有把模型评测与场景评测结合起来，才能避免“通用模型测试合格、具体场景使用失控”的问题。

第二，数据治理正在从一般数据安全扩展到训练数据、语料质量和版权合规。生成式 AI 依赖大规模训练数据和高质量语料，训练数据来源、授权、隐私保护、版权合规和数据质量将直接影响模型风险。地方政策已经开始出现语料券、高质量数据集和公共数据开放工具，但仍需在训练数据来源说明、版权争议处理、数据脱敏、数据质量评估和公共语料治理方面形成更明确的规则。

第三，内容治理正在从违法内容处置扩展到内容标识、水印和传播责任。生成式 AI 内容风险不仅包括违法有害信息，还包括虚假信息、深度伪造、误导性广告、版权侵权、未成年人保护和公共传播风险。我国已围绕深度合成、生成合成内容标识等形成政策基础，下一步应推动显式标识、隐式水印、内容凭证、平台留痕和跨平台识别机制协调。

内容标识制度的关键不是简单要求“加标”，而是建立可验证、可追踪、可互认的内容治理基础设施。显式标识有助于普通用户识别生成内容，隐式水印有助于平台和监管部门追踪内容来源，内容凭证有助于在新闻、广告、教育、司法、金融等高信任场景中保持证据链。未来应推动内容标识规则与平台审核、版权保护、未成年人保护和公共事件应急机制衔接，避免标识制度停留在形式合规。

第四，场景治理正在成为我国生成式 AI 治理最具特色的政策方向。量化结果显示，近期治理核心政策中地方政策占比较高，且大量进入政务、教育、医疗、工业、交通和城市治理场景。这说明“人工智能+”行动正在通过地方政策转化为具体场景任务，并推动治理要求进入应用流程。问题在于，不同场景风险差异显著：政务场景涉及行政责任和公共信任，教育场景涉及未成年人和评价公平，医疗场景涉及生命健康和专业责任，金融场景涉及消费者保护和系统性风险。因此，场景治理不能停留在推广应用，而应形成场景分级、风险清单、评测要求和责任边界。

第五，评测标准和合规工具是连接规则与执行的关键环节。没有评测工具，规则难以落地；没有标准和指南，企业无法形成稳定预期；没有第三方评测和公共工具，中小企业和地方部门难以承担合规成本。英国和新加坡的经验说明，国家级评测机构、测试工具和合规流程对于生成式 AI 治理至关重要。我国应将模型评测、红队测试、数据合规检查、内容标识验证和场景风险评估作为公共治理能力建设重点。

七、趋势判断与政策建议

综合中外政策分析，可以形成四个趋势判断。第一，生成式 AI 治理将从模型监管转向模型、数据、内容、场景联动治理。第二，治理能力将越来越依赖技术工具和公共基础设施。第三，地方场景将成为我国生成式 AI 治理的重要优势和风险来源。第四，国际规则竞争将围绕透明义务、模型评测、数据版权、内容标识和基础设施安全展开。

据此，提出六方面建议。其中，持续评测与事故报告、训练数据和版权合规、重点场景分级治理应作为近期优先补齐的三项制度短板；内容标识、公共工具和跨区域互操作则应作为支撑性制度同步推进。

（1）建立模型持续评测和事故报告机制。在现有备案和安全评估基础上，针对大模型能力变化、重大版本更新、工具调用、插件接入和高风险场景部署，建立持续评测、重大变化报告和严重事件报告制度。重点应从“一次性上线合规”转向“全生命周期风险管理”，要求模型提供者和重点应用单位对版本更新、能力跃迁、外部工具调用和高风险场景迁移进行记录、评估和报告。对政务、医疗、金融、教育、公共传播等高影响场景，应建立严重事件报告和事后复盘机制，使模型风险能够被及时识别、纠偏和追责。

（2）完善训练数据、语料和版权合规规则。推动训练数据来源说明、公共语料治理、版权争议处理、数据脱敏、数据质量评估和高质量数据集标准建设，将地方语料券、数据集政策与国家数据治理规则衔接起来。对于公共数据、行业数据和网络公开数据，应区分数据来源、授权状态、敏感程度和使用边界，形成可操作的数据合规清单。地方在建设语料库、发放语料券和推动高质量数据集时，应同步嵌入版权审查、个人信息保护和数据质量评估要求，避免数据供给扩张快于数据治理能力提升。

（3）推进生成内容标识和内容凭证基础设施。围绕显式标识、隐式水印、内容凭证、平台留痕和跨平台识别，建立统一技术规范和验证机制，避免标识规则碎片化。内容标识不应停留在形式合规，而应与平台审核、版权保护、广告治理、未成年人保护和公共事件应急处置衔接。建议优先在新闻传播、网络视听、广告营销、教育评价、司法证据等高信任场景开展内容凭证试点，逐步形成跨平台、可验证、可追踪的生成内容治理基础设施。

（4）建立行业场景分级治理制度。针对政务、教育、医疗、金融、交通、文化内容等重点场景，制定风险清单、准入条件、评测指标、责任边界和退出机制，推动场景应用与场景治理同步部署。对低风险场景，可以通过指南、标准和测试工具降低合规成本；对高风险场景，应设置更严格的准入、评测、人工复核和责任追踪要求。地方推进“人工智能+”行动时，应把场景开放、应用示范和治理要求同时纳入任务清单，避免场景扩散快于风险识别和责任配置。

（5）建设公共评测和合规工具体系。借鉴英国和新加坡经验，推动模型评测、红队测试、内容安全测试、数据合规检查、场景风险评估工具公共化，降低企业和地方部门合规成本。公共评测体系应覆盖基础能力、安全风险和场景适配三个层面，既评价模型准确性、鲁棒性、幻觉倾向和对抗攻击抵抗能力，也评价违法有害内容生成、隐私泄露、网络攻击辅助、生物化学风险和自动化代理风险。对中小企业和地方部门而言，低成本、可复用的合规工具比原则性要求更能提升治理落地能力。

（6）加强跨区域执行一致性和国际互操作。我国地方政策扩散速度快，但应避免同类场景执行标准不一致的问题。建议建立跨区域政策工具清单、评测结果互认机制和重点场景治理指南，同时加强与欧盟、美国、英国、新加坡、日本在模型评测、内容标识、数据版权和安全标准上的互操作。下一阶段的关键不是单纯增加政策数量，而是把政策动员能力和应用场景优势转化为可执行的治理闭环，在模型评测、数据合规、内容标识、场景分级和事故反馈上形成统一制度框架，提升我国在生成式 AI 治理中的制度供给能力、规则参与能力和跨场景治理能力。

总体而言，2025—2026 年生成式 AI 治理已经进入复合治理阶段。后续政策应坚持发展和安全并重，把“人工智能+”行动形成的场景优势，转化为可评测、可验证、可追责的治理能。

第三章 生成式 AI 侵权责任

阮神裕[1, 2]

1.中国人民大学法学院 北京 100872

2.中国人民大学人工智能治理研究院 北京 100872

一、生成式 AI 的侵权风险

依据人际交互的方式，生成式 AI 可以分为两类：一是向用户提供信息的人工智能，即信息交互型人工智能，如豆包、DeeSeek、ChatGPT 等向社会公众开放的问答式人工智能，用户从中获取何种信息，取决于用户所提的问题。二是向用户提供决策的人工智能，即决策交互型人工智能，如在授信、招聘、招生等场景中使用生成式 AI 进行自动化决策。两种类型的生成式 AI 的侵权风险各不相同。

（一）信息交互型人工智能的侵权风险

信息交互型人工智能所造成的侵权风险，集中表现为“内容生成—内容呈现—内容传播”过程中的权利侵害。信息交互型人工智能可能侵害的民事权益包括：

一是对著作权的侵权风险。生成式 AI 所生成的内容可能再现、改写、变形或替代他人作品，构成著作权侵权。例如，在“上海 A 文化发展有限公司与杭州 B 智能科技有限公司著作权侵权及不正当竞争纠纷案（杭州奥特曼案）”中，杭州 B 智能公司运营的生成式 AI 平台，通过提供基础模型与 LoRA 模型训练功能，使用户能够上传奥特曼图片并训练生成模型，从而生成与奥特曼形象实质相似的图片，并在平台中传播相关模型与图片，被法院认定为侵害著作权。[1]又如在“上海新创华文化发展有限公司诉广州年光公司网络侵权责任纠纷案”中，被告广州年光网络科技有限公司经营 Tab 网站，该网站具有 AI 对话和 AI 生成绘画功能。原告发现，在 Tab 网站 AI 绘画模块输入“生成奥特曼”“奥特曼拼接长发”“生成插画风格的奥特曼”等提示词后，系统可生成与案涉奥特曼形象相同或相似的图片，供用户查看和下载。法院经比对认为，部分生成图片在多个关键特征上与案涉奥特曼形象具有极高相似度，构成实质性相似。被告在应诉后采取了关键词屏蔽措施，但庭审中输入“迪迦”等与奥特曼相关的其他关键词时，Tab 仍能生成与案涉奥特曼形象高度相似的图片。法院判决被告平台提供的生成式 AI 服务构成对著作权的侵害。[2]

[1] 浙江省杭州市中级人民法院（2024）浙 01 民终 10332 号民事判决书。
[2] 广州互联网法院（2024）粤 0192 民初 113 号民事判决书。

又如在“腾讯科技（北京）有限公司等与北京百度网讯科技有限公司等著作权侵权及不正当竞争纠纷上诉案”中，百度公司运营“度加”创作工具，通过“AI 成片”功能，根据用户输入文本自动匹配并调用素材库中的视频片段（3－7 秒），拼接生成视频内容。相关素材片段与百家号原视频在时长、URL 等方面存在差异，法院据此认定百度对原视频进行了剪辑、切条并另行存储。百度公司运营的“度加”创作工具被认定为侵害著作权。[3]再如，在“福州市某品电子商务有限公司、姚某渊等人侵犯著作权案”中，被告福州市某品公司与姚某渊等人共谋，在未经著作权人许可的情况下，通过导入他人美术作品并使用生成式 AI“图生图”功能生成高度相似图片，从中选取最接近原图者，用于制作拼图并在电商平台销售牟利。被告福州市某品公司与姚某渊等人的行为被判决构成侵害著作权。[4]

二是构成不正当竞争行为。生成式 AI 具有极强的信息内容生产能力，可能破坏当前一些平台的信息生产机制，构成不正当竞争行为。例如，在“A 科技（上海）有限公司与合肥 B 信息技术有限公司、合肥 C 信息科技有限公司、杭州 D 互娱科技有限公司著作权侵权及不正当竞争纠纷案（小红书种草案）”中，A 公司经营小红书平台，长期积累了以真实消费体验和生活方式分享为核心的“种草”内容生态。B 公司、C 公司通过 AI 写作工具提供“小某书种草文案”“小某书旅游攻略”等一键生成服务，并以“符合小某书调性”等表述进行宣传；D 公司则提供该工具下载。法院认定 B 公司、C 公司以特定平台场景为应用层，提供具有明确指向性和诱导性的生成式 AI 服务，未尽合理注意义务，损害了 A 公司基于真实可靠内容生态形成的竞争优势和商业利益，也影响消费者和其他经营者利益，构成不正当竞争。[5]在“阿里巴巴集团控股有限公司等诉李某某不正当竞争纠纷案”中，李某某发布题为《阿里数字控股有限公司是真的吗》的文章。该文由生成式 AI 生成，内容称“阿里数字控股有限公司”系阿里巴巴集团子公司、是阿里数字化转型的重要布局，并配有含“阿里巴巴 Alibaba.com”字样及品牌图形的图片。法院查明，文中所称公司实际为阿里数字控股（深圳）有限公司，系案外人注册设立，与两原告无任何关联或合作关系，阿里巴巴集团也未发布相关信息。法院判决被告自媒体发布的信息不实且未显著标识，被告传播失实信息，违反诚实信用原则，构成不正当竞争行为。[6]这个案件尽管也涉及到生成式 AI，但是生成式 AI 仅仅是传播不实行为的工具，不正当竞争行为与生成式 AI 本身没有实质性关联。

三是对人格权的侵权风险。生成式 AI 的生成内容可能捏造事实、丑化形象、克隆声音、重构人格，进而侵害名誉权、肖像权、声音权、隐私权和个人信息权益。例如在“程某诉孙某网络侵权责任纠纷案（AI 恶搞案）”中，原告程某与被告孙某同为摄影交

[3] 湖南省长沙市中级人民法院（2024）湘 01 民终 18114 号民事判决书。
[4] 北京市通州区人民法院（2025）京 0112 刑初 558 号刑事判决书。
[5] 浙江省杭州市中级人民法院（2025）浙 01 民终 3998 号民事判决书。
[6] 浙江省杭州市滨江区人民法院（2024）浙 0108 民初 10311 号民事判决书。

流微信群成员，被告孙某未经原告程某同意，使用 AI 软件将原告程某用作微信头像的肖像照片生成衣着暴露的动漫风格图片，并发送至该摄影交流微信群内，后原告程某多次制止，被告孙某仍继续使用涉案 AI 软件将原告程某的上述微信头像肖像照片 AI 生成衣着暴露且身体畸形的动漫风格图片并微信私信原告。被告孙某的行为被判决侵害了原告程某的肖像权、名誉权和人格尊严。[7]

除此之外，基于生成式 AI 的“数字复活”也对死者人格利益的保护形成了一定的挑战。所谓“数字复活”，通常是指利用逝者生前的照片、视频、录音、文字记录、社交媒体内容等数据，通过生成式 AI 合成其声音、影像、表情、语言风格乃至互动人格，使逝者以“数字人”“AI 视频”“虚拟对话对象”等形式重新出现。2026 年有媒体报道“399 元 AI 复活亲人”的服务，指出部分商家以“思念营销”吸引消费者，尤其容易触动老年人、失独家庭等群体的情感需求。[8] 2025 年 10 月，已故茶学家张天福被 AI“复活”为茶企代言一事引发广泛争议。媒体报道称，相关视频通过 AI 模拟张天福的形象和声音，并以其口吻推广茶企品牌；张天福遗孀明确反对，认为视频未经其同意，是对逝者的“丑化”和“侮辱”，并表示将通过法律途径维权。[9]

（二）决策交互型人工智能的侵权风险

决策交互型人工智能所造成的侵权风险，主要是自动化决策中的偏见、歧视或者不合理差别待遇等风险。实践中存在如下现象：

一是消费领域的“大数据杀熟”。2025 年央视网报道直播购物“杀熟”现象时提到，辽宁阜新李女士因经常在某拆卡直播间消费，发现老用户购卡单价比新用户高 40 元；江苏消费者朱女士在同一电商平台同一家商铺先后两次购买同一款床上用品，第二次价格比第一次高 20%，商品页面并未提示价格差异原因。经核实，商家利用算法根据用户购买历史、浏览频率等数据，对老用户制定更高价格策略，最终同意退还差价并承诺整改价格算法。[10]

二是招聘领域的 AI 筛选和 AI 面试。2026 年春招季，多家媒体报道 AI 已深度介入就业全流程，用人单位使用 AI 筛选简历系统、AI 面试官等工具，系统可以对简历与岗位匹配度打分，并给出是否进一步了解的建议。AI 招聘可能把行业内既有历史偏见固化到模型中，企业和求职者都不易察觉；如果筛选机制中设置“空窗期不超过 3 个月”“年龄不超过 37 岁”等条件，AI 会严格执行，而人类 HR 在遇到优秀候选人时本可酌情放行。[11]

---

[7] 北京互联网法院（2024）京 0491 民初 21163 号民事判决书。
[8] 参见 https://news.bjd.com.cn/2026/04/02/11665427.shtml
[9] 参见 https://www.stdaily.com/web/gdxw/2025-10/17/content_416908.html
[10] 参见 https://news.cctv.cn/2025/04/29/ARTI4G3JYYI5zPUIyUF9Vyug250429.shtml
[11] 参见 https://finance.sina.com.cn/jjxw/2026-04-15/doc-inhupqxe4263848.shtml

三是金融信贷中的智能风控。2026 年，21 世纪经济报道专访央行原副行长李东荣时，李东荣提出快递骑手、网约车司机等新型从业者就业和收入高度不稳定，传统风控模型难以覆盖；在利用 DeepSeek 等技术降低金融服务成本时，必须避免因数据画像过于“精准”，反而将最需要金融服务的人“精准地排除在外”。[12]

二、生成式 AI 的侵权责任制度及其不足

（一）归责原则

生成式 AI 侵权原则上应当适用过错责任。依据《民法典》第 1165 条和第 1166 条的规定，在法律没有特别规定生成式 AI 侵权责任适用无过错责任的情况下，生成式 AI 的侵权责任应当适用过错责任。所谓过错责任，是指侵权责任的成立以行为人存在过错为前提的情形。也就是说，受害人在主张生成式 AI 服务提供者承担侵权责任时，应当提出证据证明生成式 AI 服务提供者存在过错。

理论上有一些学者主张生成式 AI 服务提供者应当承担无过错责任。例如有的学者主张生成式 AI 构成“产品”，可以适用产品责任，而《民法典》第 1202 条所规定的产品责任是无过错责任，因此生成式 AI 侵权应当适用无过错责任。[13]有的观点提出，尽管生成式 AI 侵权原则上可以适用产品责任，但是那些纯粹提供信息的情形，如只是提供一段文字、一张图片或者一个音视频的情形，则应当例外地不适用产品责任。[14]又如，有的学者主张生成式 AI 侵权责任的归责原则应当直接采取无过错责任，不需要通过适用产品责任的方式间接适用无过错责任。也就是说，生成式 AI 生成难以预见的致害内容时，逻辑上最契合的归责原则是无过错责任原则。[15]

本报告认为，生成式 AI 侵权责任是否直接采取无过错责任的归责原则，属于“立法论”视角，须以现行法作出明文规定为必要。从“解释论”视角来看，在现行法还未作出调整之前，生成式 AI 侵权责任不能直接适用无过错责任。

一种可能存在的路径是通过适用产品责任的方式，间接适用无过错责任。但是这一路径面临以下两个问题：一是生成式 AI 在提供信息或决策时，是否可以适用产品责任，或者在何种情况下可以适用产品责任，仍然存在争议。二是即便适用产品责任，生成式 AI 不存在“制造缺陷”，只可能存在“设计缺陷”，而“设计缺陷”的认定通常采取风险-效用衡量法，这一方法实质上需要检讨生成式 AI 提供者是否在合理成本范围内没有采

[12] 参见 https://www.21jingji.com/article/20260306/herald/34a67c667ca27d1e6b0f8696c179bdc1.html
[13] 参见李雅男：《生成式人工智能的法律定位与侵权归责路径》，载《比较法研究》2025 年第 3 期，第 71 页；窦海阳：《人工智能产品侵权归责何以脱离产品责任？》，载《中外法学》2025 年第 2 期，第 1-2 页；曹建峰：《人工智能提供者的产品责任》，载《法律科学（西北政法大学学报）》2026 年第 1 期，第 51 页。
[14] 参见曹建峰：《人工智能提供者的产品责任》，载《法律科学（西北政法大学学报）》2026 年第 1 期，第 50-51 页。
[15] 参见徐伟：《生成式人工智能服务提供者侵权归责原则之辨》，载《法制与社会发展》2024 年第 3 期，第 197-198 页。

取更优的设计方案，实质上也是检讨了提供者的过错。就此而言，将生成式 AI 界定为产品，适用产品责任，最终仍然无法逃避检讨提供者是否存在过错的问题。[16]

从司法实践来看，生成式 AI 侵权案件适用的不是产品责任，而是《民法典》第 1165 条所规定的过错责任。例如，在“梁某诉杭州深度求索人工智能基础技术研究有限公司网络侵权责任纠纷案（AI 幻觉第一案）”中，法院明确将生成式 AI 界定为服务，而非产品，不适用产品责任。[17]在“杭州奥特曼案”[18]和“广州奥特曼案”[19]中，法院明确适用的是过错责任，并且检讨了生成式 AI 服务提供者的过错。

（二）保护范围

生成式 AI 侵权责任的保护范围，是指生成式 AI 所生成的内容，可能侵害何种民事权益。《民法典》第 1165 条第 1 款规定：“行为人因过错侵害他人民事权益造成损害的，应当承担侵权责任”。其中“侵害他人民事权益”发挥着“合法”或者“不合法”之二元符码的判断功能，只有一个行为侵害他人民事权益，该行为人才有可能承担侵权责任。所谓“民事权益”，包括物权、债权、知识产权等财产权，生命权、身体权、健康权、姓名权、肖像权、名誉权、隐私权和个人信息等人格权，以及因婚姻家庭关系所产生的人身权利等等。

承上文所述，生成式 AI 对民事权益的侵害，依据人际交互的方式不同可以作如下区分：

第一，信息交互型人工智能在内容生成、呈现和传播过程中，可能侵害财产权和人格权。（1）生成式 AI 的生成内容可能侵害著作权，也可能构成不正当竞争。（2）生成式 AI 的生成内容也可能侵害名誉权、肖像权、声音权、隐私权和个人信息权益。

尽管信息交互型人工智能具有上述侵权风险，但是现行法中的著作权、人格权等相关规定，足以回应生成式 AI 的侵权风险。也就是说，生成式 AI 尽管改变了信息生产的方式，但是何种信息构成对著作权的侵害，何种信息构成对名誉权、肖像权、声音权、隐私权和个人信息权益的侵害，则仍取决于著作权和各项具体人格权的法律构造，因而不会对侵权责任法的保护范围造成冲击。例如，生成式 AI 大幅降低了伪造肖像的成本，但是伪造肖像的行为构成《民法典》第 1019 条第 1 款中的“利用信息技术伪造肖像”的行为，因此仍然落入肖像权的保护范围。[20]

第二，决策交互型人工智能在嵌入私人机构或者公共机构的决策流程时，可能实施歧视、预测或者不合理差别待遇。目前司法实践中还没有出现生成式 AI 实施歧视行为

[16] 参见林洹民：《人工智能侵权的过错责任》，载《法学研究》2025 年第 5 期，第 115-116 页。
[17] 杭州互联网法院（2025）浙 0192 民初 18143 号民事判决书。
[18] 杭州市中级人民法院（2024）浙 01 民终 10332 号民事判决书。
[19] 广州互联网法院（2024）粤 0192 民初 113 号民事判决书。
[20] 参见阮神裕：《论人工智能时代肖像可识别性的判断标准》，载《法学家》2026 年第 1 期，第 15—30 页。

的案件，因此还不清楚各级人民法院将如何回应此类案型。但在理论上，生成式 AI 实施歧视，或者作出不合理的预测等行为，究竟侵害何种民事权益，可能面临争议。例如，AI 招聘可能把行业内既有历史偏见固化到模型中，企业和求职者都不易觉察，如我国曾经出现过“地域歧视”的现象，倘若相关数据被用于 AI 训练，则该 AI 招聘也可能形成“地域歧视”的结果。但是，被歧视的主体何种民事权益遭受侵害，现行法中并未直接规定。当前司法实践主要是通过《民法典》第 990 条第 2 款的“一般人格权”，即“除前款规定的人格权外，自然人享有基于人身自由、人格尊严产生的其他人格权益”进行回应。也就是说，生成式 AI 的歧视行为可能侵害的是“一般人格权”或“人格尊严”。又如，AI 预测也可能构成侵权，例如利用 AI 预测某人将来是否可能犯罪、是否可能成为诚实有效率的雇员、是否会按时偿还贷款等。有的学者指出，AI 预测可能损害人的自主性或者人格尊严。例如，将某人标记为“高犯罪风险”“高违约风险”“不适合岗位”“高辍学风险”，如果该标签被传播或被机构据以采取不利措施，就可能产生类似名誉损害、人格贬损或机会剥夺的效果。[21]

由此可见，决策交互型人工智能可能对被决策者造成的不利影响，主要是通过一般人格权或“人格尊严”进行保护。不过，人格尊严的法律构造存在一定的模糊性，某种行为是否侵害人格尊严，通常存在较大的利益衡量空间。因此，可以想象决策交互型人工智能的广泛部署将会引发人格尊严的范围和类型发生进一步的扩张。这需要等待更多司法实践的总结。

（三）过错认定

理论和实务中的主流观点均认为生成式 AI 侵权适用的是《民法典》第 1165 条第 1 款的过错责任。如此一来，如何判断生成式 AI 服务提供者的过错，就成为具体个案中最为重要的问题

判断生成式 AI 服务提供者是否存在过错，主要考察其是否尽到合理注意义务。王利明教授认为，对被告是否存在过错的检讨，并不是从“生成结果造成侵权”直接推定服务提供者有过错，而是将问题转化为：被告作为生成式 AI 服务提供者是否违反了客观注意义务。其基本标准是“同质行业理性人标准”，即考察在损害发生时，同行业一般服务提供者能否发现该生成内容可能侵权，或者被告是否已经采取符合当时技术水平的必要预防措施；如果即便采取合理措施仍无法避免损害，则不宜认定过错。[22]

例如，在“杭州奥特曼案”中，由于涉案奥特曼图片和 LoRA 模型出现在平台广场、IP 作品等较为显著的位置，属于平台可以较明显感知的侵权信息；叠加奥特曼 LoRA 模

[21] See Hideyuki Matsumi & Daniel J. Solove, *The Prediction Society: AI and the Problems of Forecasting the Future*, 2025 University of Illinois Law Review 1 (2025), pp.5-6.
[22] 参见王利明：《生成式人工智能侵权的归责原则与过错认定》，载《中国法律评论》2025 年第 4 期，第 23 页。

型后可以稳定生成奥特曼形象，且模型可被其他用户反复调用，“布莱泽奥特曼”LoRA 模型使用次数已达 1000 次，侵权内容扩散风险较高，故平台应当预见到著作权侵权发生的可能性，故此法院判决被告平台存在过错。[23]

理论上还有一个争议，即生成式 AI 侵权是否适用《民法典》第 1195 条第 1 款中的“通知-删除规则”或“避风港规则”。所谓通知删除规则，是指网络用户利用网络服务实施侵权行为的，权利人有权通知网络服务提供者采取删除、屏蔽、断开链接等必要措施。网络服务提供者接到通知后，应当及时将该通知转送相关网络用户，并根据构成侵权的初步证据和服务类型采取必要措施；未及时采取必要措施的，对损害的扩大部分与该网络用户承担连带责任。肯定说认为，生成式 AI 可以适用通知删除规则，其主要理由是我国法上的通知规则不是免责条款，而是归责条款。也就是说，通知不是为了让服务提供者免责，而是权利人用来证明服务提供者在通知后具有过错的一种方式。网络侵权一般采过错责任，受害人要证明服务提供者有过错；发送合格通知正是证明其“知道”或应当采取措施的方式之一。因此，通知规则可以作为认定生成式 AI 服务提供者通知后过错的规则直接适用，而不是只能通过类推适用进入生成式 AI 场景。[24]

也有观点认为，生成式 AI 侵权不能直接适用《民法典》第 1195 条的通知规则，但可以类推适用。生成式 AI 侵权适用通知规则的根据，不在于其完全符合《民法典》第 1195 条的文义构造，而在于其与传统网络侵权共享相同的制度理由，即通过通知机制具体化服务提供者的注意义务，在权利保护与技术发展之间建立动态平衡。 因此，生成式 AI 场景中所谓“通知删除”，更准确地说应当转化为“通知—必要措施”：必要措施可以包括屏蔽、过滤、阻止类似内容再次生成、风险提示、限制诱导侵权用户使用服务等，而不只是传统意义上的删除链接。[25]

否定说则认为，生成式 AI 侵权不能适用《民法典》第 1195 条的通知删除规则。生成式 AI 服务提供者输出内容的行为属于“作为”，不符合网络侵权规则所调整的关系；对生成式 AI 侵权适用通知规则与知道规则，既无助于预防和制止同一侵权行为，也不利于鼓励生成式 AI 技术创新，因此不能将网络侵权规则适用或类推适用于生成式 AI 侵权责任。[26]

生成式 AI 生成侵权内容时，能否适用通知删除规则，目前仍停留在理论讨论层面。“杭州奥特曼案”和“广州奥特曼案”中虽然均涉及生成式 AI 输出侵权内容的情况，但是法院没有直接回应是否适应通知删除规则的问题。但是可以想象，随着生成式 AI

[23] 参见杭州市中级人民法院(2024)浙 01 民终 10332 号民事判决书。
[24] 参见徐伟：《论通知规则在生成式人工智能侵权中的适用》，载《现代法学》2024 年第 3 期，第 145 页。
[25] 参见王利明、包丁裕睿：《论“通知”规则在生成式人工智能作品侵权中的类推适用》，载《比较法研究》2025 年第 4 期，第 12 页。
[26] 参见程啸：《论生成式人工智能服务侵权不应适用网络侵权规则》，载《比较法研究》2026 年第 1 期，第 125—137 页。

的广泛部署，这一问题肯定会在实践中出现，届时各个法院可能呈现不同的判决，因此应当在司法解释中进行明确规定。

（四）责任主体

所谓责任主体，是指因过错侵害他人民事权益而造成损害，因此应当承担侵权责任的主体。通常而言，生成式 AI 侵权责任的责任主体是“生成式 AI 服务提供者”，问题在于：服务提供者是一个广义的概念，包括基础模型开发者、应用或服务部署者以及人工智能价值链上的其他主体。如此一来，当生成式 AI 输出侵权内容时，谁应当承担侵权责任，或者共同承担侵权责任，就会成为法院不得不面临的问题。

判断生成式 AI 侵权责任主体，不能仅以“谁提供了服务”作形式判断，而应当围绕侵权风险的形成、控制与实现过程，考察不同主体在模型训练、系统部署、内容生成、结果分发等环节中的实际控制能力与过错形态。换言之，责任主体的识别应从“服务链条”转向“风险控制链条”：谁对侵权风险具有预见可能性、避免可能性和控制可能性，谁就可能成为侵权责任的承担者。

首先，基础模型开发者并非当然承担责任。基础模型开发者主要负责模型架构设计、数据训练、参数优化与安全对齐，其过错通常表现为训练数据来源违法、训练语料中明显包含大量侵权内容却未采取必要清洗措施，或者明知模型在特定类型内容上存在高频侵权风险而未进行安全校准、过滤与风险提示。如果侵权输出直接源于基础模型本身的系统性缺陷，例如模型持续生成特定权利人的肖像、声音、作品片段或者侮辱诽谤性信息，则基础模型开发者对侵权结果具有较强的风险创设与风险控制地位，不能仅以其未直接面向终端用户提供服务为由完全免责。但如果基础模型已经采取合理的数据治理、模型评测、内容安全机制，而侵权结果主要是后续部署者改变模型能力边界、关闭安全措施或者诱导性调用所致，则基础模型开发者不宜被机械地纳入责任范围。例如，在“杭州奥特曼案”中，被告平台所使用的大模型，实际上是一个开源大模型。尽管该基础模型具有生成奥特曼图片的能力，但是法院没有简单判决基础模型开发者也要承担侵权责任。[27]

其次，应用或服务部署者通常是法院识别责任主体时更直接的对象。原因在于，部署者直接面向用户提供具体服务，决定模型的使用场景、交互界面、提示词模板、内容过滤规则、输出审核机制以及商业化运营方式。生成式 AI 的侵权输出并不是抽象模型能力的自然外溢，而往往是在特定应用场景中被组织、引导和分发的结果。例如，图片生成平台允许用户输入真实人物姓名生成高度相似的形象，语音合成平台允许上传少量音频即可复制他人声音，聊天机器人在医疗、金融、法律等高风险场景中输出误导性内

[27] 参见杭州市中级人民法院(2024)浙 01 民终 10332 号民事判决书。

容，均与部署者所设定的服务边界和安全措施密切相关。因此，部署者负有根据服务类型、用户规模、应用场景和可能侵害的权益类型配置相应治理措施的义务。其未尽到合理审核、提示、限制、记录和处置义务的，应当成为主要责任主体。 再次，人工智能价值链上的其他主体也可能在特定情形下承担相应责任。例如，数据提供者明知其提供的数据集中包含未经授权的作品、肖像、声音或者个人信息，仍将其用于模型训练或商业化开发，可能构成共同侵权或者帮助侵权。插件开发者、模型微调者、接口集成者通过外挂工具、参数调整或者能力扩展显著改变模型输出能力，使其更容易生成侵权内容，也可能因扩大风险而承担责任。平台经营者如果只是提供一般性的算力、云服务或者分发渠道，原则上不宜当然承担侵权责任；但其若对特定侵权应用具有实质性参与、收益分享、推荐分发或者明知侵权风险仍放任扩散，则可能因未尽合理注意义务而承担相应责任。

由此可见，生成式 AI 侵权中的责任主体并非单一、固定，而应当结合具体技术结构和商业模式进行分层判断。基础模型开发者、应用部署者、数据提供者、插件开发者、平台经营者之间既可能单独承担责任，也可能因共同过错构成共同侵权。法院在个案中不宜简单套用“谁运营平台谁负责”或者“谁开发模型谁负责”的单线逻辑，而应当追问侵权风险由谁创设、由谁控制、由谁获益、由谁最有能力防止。只有以风险控制能力与过错内容为核心，才能在避免责任过度扩张的同时，防止各主体借助复杂技术链条相互推诿责任。

三、生成式 AI 侵权责任的对策建议

第一，明确生成式 AI 侵权责任的基本归责路径，避免以无过错责任简单替代过错认定。现阶段，在法律尚未对生成式 AI 侵权作出特别规定的情况下，应当坚持以《民法典》第 1165 条第 1 款为基础，原则上适用过错责任。生成式 AI 具有高度复杂性和不确定性，但不能仅因其生成结果造成损害，就当然推定服务提供者存在过错。司法裁判和未来规则设计应当把重点放在服务提供者是否违反合理注意义务上，即考察其在训练数据治理、模型安全评测、内容过滤、风险提示、用户管理、日志留存和投诉处置等方面是否达到同类行业理性经营者所应达到的注意水平。对于产品责任路径，应当保持审慎。即便在某些场景中可以将生成式 AI 系统理解为具有产品属性，也不宜由此当然适用无过错责任。尤其是在信息生成、内容问答、图像生成等场景中，损害往往并非来自传统意义上的制造缺陷，而是来自训练数据、模型设计、应用场景和用户提示词共同作用下的输出风险。此时，所谓设计缺陷的判断实质上仍然离不开风险—效用衡量和合理替代方案检讨，最终仍要回到服务提供者是否采取合理预防措施的问题。因此，制度上

应当以过错责任为基础，通过细化注意义务来降低受害人的证明负担，而不是简单引入无过错责任。

第二，建立生成式 AI 服务提供者的分层注意义务和合规治理标准。生成式 AI 侵权的预防，不能仅依靠事后赔偿，而应当前移到模型训练、系统部署、服务运营和内容传播的全过程。对于信息交互型人工智能，应当重点要求服务提供者建立训练数据来源审查机制、权利人投诉处理机制、敏感提示词和高风险内容过滤机制、侵权内容复现阻断机制以及生成内容标识机制。对于可能生成著作权作品、知名角色形象、真实人物肖像、声音或者虚假事实的服务，应当根据风险等级配置更高强度的审核和限制措施。对于决策交互型人工智能，应当要求部署者建立算法影响评估、偏见测试、决策解释、人工复核和申诉救济机制，防止自动化决策在招聘、授信、招生、保险、平台治理等领域固化既有歧视或者形成不合理差别待遇。尤其是在涉及重大人身权益、财产权益和机会利益的场景中，不能让模型结论直接替代人的判断，而应当保留有实质意义的人工干预。未来可以通过司法解释、部门规章或者行业标准，将不同场景下的注意义务具体化，使法院在个案中过错认定有相对明确的判断基准，也使企业能够据此建立可验证、可追溯、可审计的合规体系。

第三，完善“通知—必要措施”机制，将其作为认定服务提供者事后过错的重要规则。生成式 AI 侵权是否直接适用《民法典》第 1195 条的通知删除规则，理论上存在争议，但从制度功能看，通知机制对于生成式 AI 场景仍然具有重要意义。生成式 AI 服务提供者未必能够事前识别所有侵权风险，尤其是著作权、肖像权、声音权、名誉权等权益的侵害，往往需要权利人提供具体线索和初步证据。因而，制度上应当承认合格通知可以使服务提供者获得侵权风险的具体认知，并据此产生采取必要措施的义务。不过，生成式 AI 场景中的必要措施不能简单等同于传统网络侵权中的删除、屏蔽、断开链接，而应当结合服务类型进行扩展。例如，平台应当可以采取删除已生成内容、屏蔽特定提示词、限制特定模型调用、下架侵权 LoRA 模型、阻止相似内容再次生成、暂停高风险用户权限、增加风险提示、保留生成日志等措施。对于权利人通知后仍持续生成同类侵权内容的，服务提供者应当对损害扩大部分承担相应责任；对于已经采取合理技术措施但仍无法完全避免个别侵权输出的，则不宜过度加重其责任。由此，通知机制既可以缓解权利人的维权困难，也可以避免将服务提供者置于无限审查义务之下。

第四，构建以风险控制能力为核心的责任主体识别规则，防止技术链条复杂化导致责任落空。生成式 AI 侵权往往涉及基础模型开发者、应用部署者、数据提供者、模型微调者、插件开发者、云服务商、分发平台和终端用户等多个主体。未来制度不应采取“谁提供服务谁负责”或者“谁开发模型谁负责”的单一标准，而应当围绕风险由谁创设、由谁控制、由谁获益、由谁最有能力防止来分层认定责任。基础模型开发者主要对

训练数据合规、模型安全能力和系统性输出风险承担注意义务；应用部署者直接面向用户提供服务，决定具体使用场景、交互方式、内容过滤和商业推广，通常应当承担更直接的治理义务；模型微调者、插件开发者、数据提供者如果通过数据输入、参数调整或者功能扩展显著提高侵权风险，也应当在其控制范围内承担责任；仅提供一般云服务、算力服务或者中立分发渠道的主体，原则上不宜当然承担责任，但其明知侵权风险并实质参与、推荐、获益或者放任扩散的，应当承担相应责任。通过这种分层责任结构，可以避免基础模型开发者、部署者和平台经营者之间相互推诿，也可以防止责任范围无序扩张。最终目标应当是在权利保护、技术创新和产业发展之间形成可预期的责任分配机制。

第四章 生成式 AI 人智协作图景与展望

程絮森[1,2] 曾昂[1]

1.中国人民大学信息学院 北京 100872

2.中国人民大学人工智能治理研究院 北京 100872

一、生成式 AI 人智协作的核心内涵与发展背景

在人工智能技术迭代迈入系统性进化的关键阶段，生成式 AI 已从单纯的技术工具，逐步与人类智能深度融合，形成“人智协作”的全新范式。不同于传统人机协作中“人主导、机辅助”的单向互动模式，生成式 AI 人智协作实现人类创造力、价值判断、情感感知与生成式 AI 高效计算、海量数据处理、规则化执行能力的有机结合，推动生产方式、生活场景与治理模式的全方位变革。当前，全球生成式 AI 技术持续突破、应用场景不断深化，人智协作已从概念探索走向规模化落地，成为驱动新质生产力发展的关键引擎，其发展背景可从政策引导、技术突破、产业需求三个维度进行解析。

（一）生成式 AI 人智协作的核心内涵

生成式 AI 人智协作的核心是“以人为中心”的协同体系，其核心内涵包含三个层面：一是能力协同，人类聚焦创意构思、战略决策、伦理判断等高阶智能活动，生成式 AI 承担重复性、事务性、计算性工作；二是价值协同，生成式 AI 的技术迭代与应用落地始终围绕人类需求展开，避免技术异化，凸显“技术服务于人”的核心导向；三是动态协同，人智双方通过持续交互实现能力互补升级，人类通过反馈优化生成式 AI 模型，生成式 AI 通过数据支撑辅助人类优化决策。与传统人机协作相比，生成式 AI 人智协作打破了“工具与使用者”的边界，形成了智能伙伴式的共生关系，其核心差异在于生成式 AI 具备了自主生成、自主推理、自主适配的能力，能够主动适配人类需求，而非被动执行指令。

从本质上看，生成式 AI 人智协作并非替代人类，而是通过技术赋能实现人类智能的延伸与放大。人智协同的核心价值在于“1+1>2”的协同效应，人类的主观能动性与生成式 AI 的客观高效性相结合，能够突破单一主体的能力边界，解决传统单一智能无法应对的复杂问题，尤其在创意创作、复杂决策、精准服务等领域展现出独特优势[1]。例如，在内容创作领域，人类提供核心创意与情感表达，生成式 AI 辅助完成素材整理、

[1] Lu, T., & Zhang, Y. (2025). 1+1>2? Information, Humans, and Machines. Information Systems Research, 36(1), 394～418. https://doi.org/10.1287/isre.2023.0305

语言润色、格式优化，既保留人类的创作个性，又提升创作效率；在医疗诊断领域，生成式 AI 辅助医生完成影像识别、病例分析，人类聚焦病情判断与治疗方案制定，有效降低误诊率、提升诊疗效率，这正是人智协作核心内涵的生动体现。

（二）生成式 AI 人智协作的发展背景

政策层面，全球各国纷纷将生成式 AI 发展与治理纳入国家战略，为生成式 AI 人智协作提供了明确的政策导向。我国高度重视人工智能发展，将“人工智能+”提升至国家战略高度，2025 年 8 月，国务院发布《关于深入实施“人工智能+”行动的意见》，系统提出科技、产业、消费、民生、治理、全球合作等六大重点领域，明确推动生成式 AI 与人类协同发展，强化前瞻规划与安全可控。2025 年 10 月，党的二十届四中全会将全面实施“人工智能+”行动写入“十五五”规划建议，并强调加强人工智能等新兴领域国家安全能力建设，标志着生成式 AI 人智协作已完全融入国家中长期发展战略。2025 年 12 月，中央经济工作会议在部署 2026 年经济工作时，进一步强调深化拓展“人工智能+”，完善人工智能治理，为生成式 AI 人智协作的规模化落地提供了政策保障。2026 年，“智能体”首次被写入《政府工作报告》，明确了其国家战略地位。国务院及各部委相继出台行动意见，设定了到 2027 年应用普及率超 70%的量化目标，并配套“人工智能券”、开放公共数据等举措，形成了全方位支撑体系。国际层面，美国人工智能政策发生战略转向，欧盟人工智能立法迈向落地阶段，均在严监管与促创新间寻求平衡，推动生成式 AI 人智协作的规范化发展。

技术层面，生成式 AI 技术的突破性跃进为人大智协作提供了坚实支撑。2025-2026 期间，生成式 AI 技术实现了多维度里程碑式突破，从具备感知与理解能力的工具，加速演变为能够自主决策与执行的智能体。基础模型持续迭代，根据中国信通院“方升”测试数据，截至 2025 年 12 月，头部语言大模型综合能力提升约 30%，多模态理解能力增长超过 50%，推理、编程等能力实现了“又好又快”的发展[2]。具身智能进入实训期，端到端 VLA 架构与世界模型加速探索，显著提升了机器人在未见环境中的泛化性与认知边界，推动生成式 AI 从“软件智能”向“实体智能”跨越。AI 智能体（Agent）快速发展，AI 智能体具备了规划和使用工具的能力。它以大语言模型为大脑，融合记忆、规划、工具与行动四大核心组件，能够自主完成从意图理解到任务交付的全过程。中国企业级 AI 智能体市场正呈现爆发式增长态势，2025 年，该市场规模已达到 212 亿元，预计到 2026 年将增至 449 亿元，而到 2029 年，这一数字有望突破 3320 亿元，2024-2029 年的年复合增长率（CAGR）高达 107%。[3]这些技术突破，为生成式 AI 与人类的深度协同提供了技术可能，推动人智协作从简单工具辅助向复杂场景协同演进。

[2] 中国信息通讯研究院，人工智能产业发展研究报告（2025 年），2026，https://www.caict.ac.cn/kxyj/qwfb/bps/202602/P020260202487301304903.pdf

[3] 中国通信工业协会数据中心委员会，中国指挥与控制学会情报与智能认知专业委员会，科智咨询，AI 智能体赋能行

产业层面，千行百业的数字化转型需求，成为生成式 AI 人智协作落地的核心驱动力。随着我国人工智能产业进入规模化、深层次融合应用的新阶段，企业对效率提升、成本降低、创新升级的需求日益迫切，而生成式 AI 人智协作能够有效破解传统产业发展中的痛点难点。据中国信通院测算，2024 年我国人工智能核心产业规模已突破 9000 亿元，同比增长 24%，2025 年有望达 1.2 万亿元，截至 2025 年底，我国人工智能企业数量超过 6000 家，全球占比达 16%，形成了从基础底座、模型框架到行业应用的完整产业体系[4]。我国大模型调用需求快速增长，2026 年 3 月，中国日均 Token 调用量突破 140 万亿，相比 2024 年初的 1000 亿，两年增长超千倍。同月，中国 AI 大模型周 Token 调用量连续三周超越美国，成为全球 AI 应用活跃度最高的国家之一。全球最大 AI 模型 API 聚合平台 OpenRouter 最新数据显示，3 月 16 日至 22 日，全球 AI 大模型总 Token 调用量为 20.4 万亿，仅中国就达 7.359 万亿，占全球的 36%。。[5]无论是工业领域的生产流程优化、农业领域的精准种植，还是服务业领域的个性化服务，都对生成式 AI 人智协作提出了迫切需求，推动人智协作场景不断丰富、应用不断深化。

二、生成式 AI 人智协作的发展图景

2025-2026 年 4 月，生成式 AI 人智协作进入规模化落地的关键时期，呈现出“技术持续突破、场景全面渗透、生态日趋完善”的发展态势。从技术支撑来看，基础模型、具身智能、智能体等核心技术的迭代，为人智协作提供了更强大的能力支撑；从应用场景来看，人智协作已渗透到工业、农业、医疗、教育、办公等多个领域；从产业生态来看，“开源大模型+海量应用场景”的模式逐步形成，推动人智协作向普惠化方向发展。

（一）技术支撑：核心技术迭代赋能人智协作升级

基础模型的跨越式发展，为人智协作提供了核心能力支撑。国产大模型发展迅速，以 DeepSeek、通义千问等为代表的国产大模型，通过开源战略引领全球生态，其中阿里通义系列模型全球下载量已突破 6 亿次，极大地降低了广大开发者和中小企业的创新门槛。基础模型的多模态能力持续提升，能够实现文本、图像、声音、视频等多维度信息的融合处理，打破了不同信息形态的壁垒，为人智协作提供了更丰富的交互方式。

具身智能的快速发展，推动人智协作从“虚拟场景”向“物理场景”延伸。具身智能处于从实验室验证向规模化商用过渡的关键期。在工业领域，基于具身智能的无人驾驶挖掘机依托端到端具身智能模型，在真实严苛环境中已达到接近人工的装车效率；在

业决策：趋势与实践白皮书（2026），2026，https://www.cidc.org.cn/news/20260428/1233729920836829184.html

[4] 中国信息通讯研究院，人工智能产业发展研究报告（2025 年），2026，https://www.caict.ac.cn/kxyj/qwfb/bps/202602/P020260202487301304903.pdf

[5] 人民邮电报，日均 Token 调用量达 140 万亿，2026，https://www.news.cn/tech/20260326/8026bd54df3f489a8c79d4438cac96b3/c.html

物流领域，京东物流通过大模型实现了人类调度与机器人执行的高效协同[6]。这些应用表明，具身智能的发展正在打破生成式 AI 与物理世界的壁垒，使人智协作能够深入到生产、物流等实体场景，提升协同效率与质量。

AI 智能体的普及应用，重构了人智协作的模式。例如，在自动化代码开发场景中，Agent A 根据需求生成代码，Agent B 负责检查 bug/安全漏洞/规范，Agent C 监督整体流程、记录审计日志。在营销方案设计场景中，Agent A 拆解需求，多个 Agent 并行生成不同方案，再由 Agent 进行评估决策。AI 智能体的发展，使人智协作从“一对一”的简单协同，升级为“多对多”的复杂协同，进一步拓展了人智协作的应用边界。

智算基础设施的升级，为生成式 AI 人智协作提供了算力支撑。随着生成式 AI 模型规模的扩大与应用场景的增多，对算力的需求持续攀升，吉瓦级（GW）智算集群建设加码。高质量数据集建设正向智能生成、多元专业、合规治理方向深化，为生成式 AI 人智协作提供了高质量的数据支撑。

（二）多领域应用：人智协作场景持续渗透落地

医疗领域的生成式 AI 人智协作，有效提升了诊疗效率与质量，缓解了医疗资源紧张的问题。在影像诊断领域，生成式 AI 能够快速识别 CT、MRI 等影像中的病变区域，辅助医生完成诊断。在临床治疗中，生成式 AI 能够根据患者的病历数据、基因信息，辅助医生制定个性化治疗方案，人类医生负责方案审核与调整，实现了“精准医疗”与人智协作的结合。此外，生成式 AI 还能辅助医生完成病历书写、药物调配等事务性工作，让医生能够将更多精力投入到诊疗工作中。

教育领域的生成式 AI 人智协作，正在重构教育教学模式，实现“个性化教学”与“精准育人”。在课堂教学中，生成式 AI 能够根据学生的学习情况、知识掌握程度，生成个性化的学习方案，辅助教师完成教学任务，人类教师则聚焦学生的情感引导、思维培养与个性化辅导。

办公领域的生成式 AI 人智协作，成为提升办公效率的核心手段，推动办公模式从“传统办公”向“智能办公”转型，为人智协作提供了更便捷的工具支撑。在文档处理中，生成式 AI 能够辅助人类完成文档撰写、编辑、翻译、总结等工作。在会议管理中，生成式 AI 能够自动记录会议纪要、提取会议重点、安排会议任务，辅助人类完成会议组织与跟进工作。在客户服务中，生成式 AI 智能客服能够处理常见的客户咨询、投诉等问题，人类客服则聚焦复杂问题的解决与客户情感沟通，实现了生成式 AI+人类的协同模式。

[6] 中国信息通讯研究院，人工智能治理研究报告（2025 年），2026，https://www.caict.ac.cn/kxyj/qwfb/ztbg/202602/P020260213607027089305.pdf

（三）产业生态：人智协作生态日趋完善普惠

开源生态成为生成式 AI 人智协作发展的重要支撑，推动技术普惠化。开源成为生成式 AI 产业的标配，社区协同演进推动技术普惠，国产大模型通过开源战略，降低了开发者与中小企业的创新门槛。阿里通义系列模型、DeepSeek 等国产大模型开源后，衍生出大量的应用场景与二次开发产品，形成了开源大模型赋能海量应用场景的生态模式，推动生成式 AI 人智协作从少数企业的尖端科技，转变为千行百业可便捷调用的普惠性创新基础设施。

表 1 国产大模型开源情况

模型名称	模型厂商	开源/闭源	模型能力特点
通义千问 Qwen3.6 系列	阿里巴巴	开源（Qwen3.6-2B/7B/14B/27B/72B）闭源（Qwen3.6-Max/Plus）	中文能力顶尖，全模态支持，长文本处理，稠密架构高效能
文心一言 ERNIE 6.0	百度	闭源	知识增强技术，企业合规套件完善，多模态融合，政务场景适配
豆包 Doubao-Seed-2.0-pro	字节跳动	闭源	中文对话流畅，多轮记忆能力强，工具调用能力领先
GLM-5.1	智谱华章	开源（GLM-5-1.8B/9B/40B）闭源（GLM-5.1-Pro/Ultra）	逻辑推理与编程能力突出，Agent 智能体能力领先，国产算力适配好
DeepSeek V4 系列	深度求索	开源	推理成本低，显存占用少，百万上下文窗口，Agent 编码评测开源第一
盘古大模型 5.5	华为	闭源	昇腾芯片协同优化，推理成本降低，工业质检、金融风控等高精度场景适配
混元 Hy3 preview	腾讯	开源（Hy3 preview）闭源（混元大模型 Plus）	重建基础设施后首款模型，具身任务优化多模态模型 HY-Embodied-0.5-X 同步开源
Kimi K2.6	月之暗面	开源	编程能力全球领先，SWE-Bench Pro 评测 58.6% 登顶，超越 GPT-5.4 和 Claude Opus 4.6
星辰大模型	央企联合	部分开源	国内首个全模态、全尺寸、全国产的央企大模型体系，Token 经营新范式
MiniMax Abab5	MiniMax	开源	轻量级高效能，消费级硬件可运行，中文对话自然度高

基准测试体系不断完善，为人智协作的规范化发展提供保障。随着生成式 AI 人智协作的规模化落地，对生成式 AI 能力的评估与规范提出了更高要求，中国信通院等权威机构推出了“方升”测试等基准测试体系，涵盖大模型综合能力、智能体性能、具身智能适配性等多个维度，为企业选择生成式 AI 产品、优化人智协作模式提供了参考依据。

产业链协同日趋紧密，形成了“基础层-模型层-应用层”的完整产业链。基础层方面，智算集群、国产芯片、高质量数据集等基础设施不断完善，为生成式 AI 人智协作提供了算力、数据支撑；模型层方面，通用基础大模型、垂直行业大模型协同发展，形成了完整的产品矩阵；应用层方面，各类应用场景不断丰富，覆盖工业、医疗、教育、办公、农业等多个领域，形成了一批典型的人智协作应用案例。

全球合作不断深化，推动生成式 AI 人智协作的全球化发展。人工智能已成为全球重点多边机制的核心议题，全球治理模式正从单边碎片化走向多边协商，更具包容性的标准生态正在形成。我国积极参与全球生成式 AI 治理与合作，推动国产大模型与开源生态走向全球，同时引进国际先进技术与经验，促进我国生成式 AI 人智协作水平的提升。国际合作的深化，也有助于破解全球生成式 AI 人智协作面临的共性问题，如技术标准不统一、治理体系不完善等，推动人智协作的规范化、全球化发展。

三、生成式 AI 人智协作的问题、展望与对策建议

当前生成式 AI 人智协作虽实现规模化落地，但仍面临多维度突出问题，结合技术迭代趋势与产业发展需求，明确未来发展方向并提出针对性对策，方能推动其高质量发展。

（一）当前面临的突出问题

技术层面，核心瓶颈尚未突破。生成式 AI 模型“黑箱效应”显著，可解释性不足，且部分模型存在“幻觉”与策略欺骗行为，降低协作可靠性。具身智能面临高质量数据短缺、跨场景泛化难、软硬协同不稳定等工程化难题，难以大规模商用。人智协同适配性不足，生成式 AI 难以精准理解人类主观意图与情感需求，且个性化适配、快速学习能力欠缺，难以实现理想的“人机共生”状态。

治理层面，体系不完善且权责模糊。我国尚未出台专门针对生成式 AI 人智协作的法律法规，难以应对决策失误追责、著作权归属、数据安全保护等新型法律纠纷。人类与生成式 AI、研发企业、应用企业之间的权责划分不清晰，出现问题易相互推诿，多智能体协作场景的追责难度更突出。监管体系滞后于技术发展，以事后监管为主，事前预防与事中管控不足，且行业标准不统一，导致不同企业产品兼容性差，影响规模化发展。

社会层面，认知偏差与人才、数字鸿沟制约发展。部分人群存在“技术恐惧”，担心生成式 AI 替代工作，另有部分人群过度依赖生成式 AI，忽视人类主观能动性。生成式 AI 人智协作复合型人才缺口巨大，现有人才供给难以满足快速发展的产业需求，从事 AI 芯片、算法设计等关键技术研发的顶尖人才依然匮乏，具备 AI 技术与行业背景的复合型人才比较稀缺[7]。数字鸿沟突出，基层地区、老年群体、中小企业因缺乏技术、资金与技能，难以享受人智协作带来的便利。

（二）未来发展展望

未来 3-5 年，生成式 AI 人智协作将进入高质量发展阶段，实现从“规模化落地”向“精细化协同”跨越。技术层面，可解释 AI 技术将破解“黑箱效应”，主流生成式 AI 模型可解释性大幅提升。具身智能将突破工程化瓶颈，实现工业、医疗等领域大规模商用；多智能体将向协同化、自主进化方向发展，多智能体协作系统将在企业端全面普及。

应用层面，场景将实现全领域深化覆盖。工业领域向研发、生产、调度、运维全流程协同升级，人智协作大量普及。医疗领域延伸至疾病全周期管理；教育领域构建个性化终身学习体系，成为主流教学模式；同时逐步渗透到航天、海洋、环保等新兴领域，拓展应用边界。

生态层面，协同共赢格局逐步形成。法律法规体系将不断完善，我国将形成健全的生成式 AI 人智协作法治体系，全球协同治理机制逐步建立；监管体系升级为“事前预防、事中管控、事后追责”全流程模式，实现精准监管；人才培养体系不断完善，人才缺口基本填补；数字鸿沟显著缩小，生成式 AI 人智协作实现普惠化发展，覆盖千行百业。

（三）对策建议

技术层面，加大核心技术研发力度。政府设立专项资金，支持可解释 AI、具身智能等核心技术研发，推动产学研协同创新；由权威机构牵头制定统一技术标准，提升产品兼容性与协作水平；加强数据安全与隐私保护技术研发，推动隐私计算应用，建立数据分级分类管理机制，保障数据合规安全。

政策层面，完善治理体系与政策引导。加快推进生成式 AI 人智协作相关立法，明确各方权责，填补法律空白；创新监管模式，建立事前备案、事中监测、事后追责的全流程监管体系，利用生成式 AI 技术提升监管效率；加大政策扶持，对中小企业、基层地区给予资金与技术支持，完善基层智算基础设施，推动技术普惠。

[7] 郑振宇，加快培养高水平人工智能人才，光明日报，2025，https://epaper.gmw.cn/gmrb/html/2025-09/26/nw.D110000gmrb_20250926_2-06.htm

企业层面，强化主体责任与创新落地。坚守合规底线，建立健全内部管理制度，加强数据安全与模型安全管理，完善责任追溯机制；加大研发投入，深化场景应用，打造个性化解决方案，加强用户反馈优化；完善人才培养与引进机制，通过校企合作、员工培训、高端人才引进，缓解人才缺口。

社会层面，引导正确认知与缩小发展差距。通过多种形式普及生成式 AI 人智协作知识，消除认知偏差，宣传典型应用案例；优化高校专业设置，开展校企联合培养与技能培训，扩大复合型人才供给；政府、企业、社会组织协同发力，推出低成本产品，开展技能培训，推动算力与数据资源下沉，缩小数字鸿沟。

综上，生成式 AI 人智协作作为驱动新质生产力发展的关键引擎，需正视当前问题，依托技术创新、治理完善、社会协同，破解发展瓶颈，推动其向更高效、更精准、更普惠的方向发展，充分释放人机协同的巨大价值。

第五章 生成式 AI 智慧医疗应用图景与展望

殷佳敏[1, 2]

1.中国人民大学商学院 北京 100872

2.中国人民大学人工智能治理研究院 北京 100872

一、引言

生成式 AI 智慧医疗是“人工智能+”行动在民生福祉领域的重要组成部分。2025 年 8 月，国务院发布《关于深入实施“人工智能+”行动的意见》[1]，提出以科技、产业、消费、民生、治理、全球合作等领域为重点，推动人工智能与经济社会各行业各领域广泛深度融合，并在民生服务中提出探索推广人人可享的高水平居民健康助手，有序推动人工智能在辅助诊疗、健康管理等方面应用。2025 年 11 月，国家卫生健康委、国家发展改革委、工业和信息化部、国家中医药局、国家疾控局联合印发《关于促进和规范“人工智能+医疗卫生”应用发展的实施意见》[2]，明确以新一代人工智能深度赋能卫生健康行业高质量发展，坚持政府引导、多方参与、创新驱动、安全可控，推动预防、诊疗、康复、健康管理等全链条连续智能服务。

与此同时，世界卫生组织在关于健康领域大规模多模态模型的伦理与治理指南中指出[3]，大规模多模态模型能够接受一种或多种数据输入并生成多样化输出，预期将广泛应用于医疗服务、药物研发和公共卫生。美国 FDA[4]则在 2025 年发布关于 AI 赋能医疗器械软件功能的生命周期管理与上市申请建议草案，强调以全生命周期方法评估安全性和有效性。

近年来，以大模型、多模态模型、智能体和生成式交互界面为代表的生成式 AI 加速进入医疗卫生领域。与上一轮以影像识别、辅助筛查和规则型临床决策支持为主的“医疗 AI”不同，本轮生成式 AI 智慧医疗的核心变化在于：一是模型能力从单点识别走向多模态理解、推理和生成；二是应用边界从医生工作站延伸至基层诊疗、药物研发和公共卫生；三是人机关系从工具辅助转向共同完成任务的协同关系，甚至在部分场景中呈现出智能体自主规划、跨系统调用和持续学习的趋势。

[1]参见 https://www.gov.cn/zhengce/zhengceku/202508/content_7037862.htm
[2]参见 https://www.gov.cn/zhengce/zhengceku/202511/content_7047018.htm
[3]参见 https://www.who.int/publications/i/item/9789240084759
[4]参见 https://www.fda.gov/media/184856/download

总体来看，生成式 AI 智慧医疗呈现出三个鲜明特征。第一，应用从演示验证走向深度嵌入，越来越多产品嵌入电子病历、临床文书、医学影像、患者服务和药物研发流程。第二，治理从原则导向走向场景化、可操作规范，监管部门开始围绕基层诊疗、医学影像、药物研发和公共卫生等场景提出更具操作性的要求。第三，风险从单一模型错误扩展为系统性责任与合规风险，包括幻觉与误导、隐私泄露、数据偏倚等。

二、生成式 AI 智慧医疗应用图景

（一）居民健康助手与患者服务

居民健康助手是生成式 AI 智慧医疗最贴近公众的一类应用。相较传统互联网问诊或健康百科，生成式模型能够通过自然语言理解居民症状、生活方式、既往病史和健康目标，生成个性化健康建议、就诊指引、复诊提醒、用药注意事项和慢病管理方案。我国《卫生健康行业人工智能应用场景参考指引》[5]已将院后健康管理、智能病历辅助生成、处方前置审核、智能随访等纳入典型应用场景，并提出通过智能服务提高患者依从性、减轻医护人员负担、优化医疗资源配置。

2025 年以来，患者服务类应用的一个重要趋势是从单次咨询转向全流程陪伴。国家卫生健康委等五部门的《实施意见》[6]提出，到 2027 年，基层诊疗智能辅助、临床专科专病诊疗智能辅助决策和患者就诊智能服务将在医疗卫生机构广泛应用。到 2030 年，基层诊疗智能辅助应用基本实现全覆盖，二级以上医院普遍开展医学影像智能辅助诊断、临床诊疗智能辅助决策等人工智能技术应用。这意味着居民健康助手不再只是商业互联网平台的消费级产品，而将逐步进入基层医疗卫生机构、家庭医生签约服务、慢病随访和公共卫生管理体系。

典型案例是社区健康服务场景中的数字医生。据人民日报报道，北京市海淀区花园路社区卫生服务中心已出现面向居民的“数字医生”交互屏，能够围绕居民主诉进行问答，并提供个性化健康管理建议[7]；报道同时指出，辅助诊疗、处方审核等智能应用正在成为基层医生的帮手。这类场景的治理重点在于明确其健康咨询、就医指引、风险提示的辅助定位，避免其被公众误解为可替代执业医师的诊断结论。对于未成年人、老年人、慢病患者和心理脆弱群体，还应强化风险提示、转诊触发和人工复核机制。

（二）临床诊疗辅助

临床诊疗辅助是生成式 AI 智慧医疗最受关注、也最具高风险属性的场景。传统 AI 医疗器械多集中于医学影像辅助检测和辅助诊断，例如肺结节、眼底病变、骨折、脑卒

[5]参见 https://www.nhc.gov.cn/guihuaxxs/c100133/202411/3dee425b8dc34f739d63483c4e5c334c.shtml
[6]参见 https://www.gov.cn/zhengce/zhengceku/202511/content_7047018.htm
[7] 参见 https://www.peopleapp.com/column/30050872591-500007231613

中等单病种识别。生成式 AI 进入后，模型能够整合病历文本、影像、检验检查、用药记录、指南知识和医患对话，生成鉴别诊断、诊疗建议、报告草稿和风险预警。其价值不只在于提高单个诊断环节的效率，更在于帮助医生组织复杂信息、形成诊疗假设、解释不确定性并支持多学科协作。

国际上，Google 的 AMIE 研究代表了诊断对话智能体的重要进展。2025 年发表在 Nature 的研究介绍了 Articulate Medical Intelligence Explorer，这是一种面向临床病史采集和诊断对话优化的大语言模型系统，研究通过模拟诊断对话环境和自动反馈机制提升模型能力[8]。这类研究表明，医疗生成式 AI 正在从回答医学考试题转向参与临床推理过程，但其研究环境与真实临床场景之间仍存在差距，包括患者叙事不完整、责任主体不清和本地化指南差异等。

与此同时，开源医疗基础模型也在加速发展。Google 于 2025 年发布 MedGemma[9]，定位为面向健康 AI 开发的开放模型，可用于医学文本、医学图像理解、报告生成和视觉问答等任务；2026 年 1 月又更新 MedGemma 1.5 并推出 MedASR[10]，强化医学影像支持和医疗语音转文本能力。开源模型降低了科研机构、医院和中小企业进入医疗 AI 开发的门槛，但也放大了模型数据来源、性能验证和责任追溯问题。尤其在临床诊疗场景中，开源模型不可未经验证直接用于高风险场景，必须经过本地数据验证、预期用途声明和持续监测。

（三）医疗文书与临床工作流

医疗文书生成也是目前生成式 AI 医疗商业化主要方向之一。目前，医生将大量时间用于病历书写、会诊记录、出院小结、转诊信等。生成式 AI 可通过环境语音识别、自然语言处理和结构化病历生成，将医患对话自动转化为可编辑的临床文书，降低行政负担。2025 年 3 月，Microsoft 发布 Dragon Copilot[11]，将 Dragon Medical One 的语音听写、DAX 的环境聆听能力、医疗场景优化的生成式 AI 和安全防护结合，定位为临床工作流 AI 助手，用于文书、信息调取和任务自动化。

Abridge 与 Mayo Clinic 的合作则体现了环境文书系统进入大型医疗机构的趋势。2025 年 1 月，Abridge 宣布与 Mayo Clinic 达成企业级协议，计划首先面向约 2000 名临床医生部署 AI 临床文书平台，覆盖多专科和多护理场景[12]。这类案例说明，生成式 AI 医疗应用的商业价值往往并非来自替代医生诊断，而是来自文书、编码、质控等低风险

[8] 参见 https://www.nature.com/articles/s41586-025-08869-4
[9]参见 https://deepmind.google/models/gemma/medgemma/
[10]参见 https://research.google/blog/next-generation-medical-image-interpretation-with-medgemma-15-and-medical-speech-to-text-with-medasr/
[11]参见 https://www.microsoft.com/en-us/microsoft-cloud/blog/healthcare/2025/03/03/meet-microsoft-dragon-copilot-your-new-ai-assistant-for-clinical-workflow/
[12]参见 https://www.abridge.com/press-release/mayo-clinic-announcement

但高频工作流。对医院而言，文书类 AI 也可能成为未来临床智能体的入口：一旦系统掌握医患对话、病历结构和临床上下文，就可能进一步扩展至检查建议、医嘱草拟、随访安排和质控提醒。

但文书类应用同样存在治理难点。第一，环境聆听涉及高度敏感的医患对话，必须明确患者知情同意、录音留存、数据使用和模型训练边界。第二，AI 生成文书可能遗漏关键阴性症状、误写病史或过度结构化表达，医生签署前必须进行人工审核。第三，若 AI 文书被直接用于医保支付、司法证据或医疗纠纷认定，错误文本的责任归属将更加复杂。

（四）医学影像、病理和多模态数据

医学影像仍是 AI 医疗最成熟的落地领域。美国 FDA 的 AI 赋能医疗器械清单旨在识别已获准上市的 AI 医疗器械，为创新者、医生和患者提供透明度。2026 年 2 月，基于 FDA 资料，Reuters 报道称美国获授权的 AI 医疗器械数量已超过千项，其中影像相关产品占据重要比例。同时，部分 AI 设备也引发召回和上市后安全监测讨论[13]。这说明影像 AI 虽然成熟度较高，但仍需要在真实世界中持续评估性能漂移、设备差异、患者亚群差异和临床工作流适配。

（五）药物研发与临床试验

生成式 AI 在医药研发中的作用正在从发现候选分子延伸至靶点识别、蛋白结构预测、分子生成、药效毒性预测等。2025 年，Insilico Medicine 及合作者在 Nature Medicine[14] 发表生成式 AI 发现的 TNIK 抑制剂 Rentosertib 用于特发性肺纤维化的 2a 期临床试验结果。该案例的重要意义不在于证明 AI 药物研发已经全面成熟，而在于显示 AI 生成候选药物开始进入以人体试验结果检验其价值的阶段。监管层面，FDA 已发布关于使用 AI 支持药物和生物制品监管决策的指南草案，提出基于风险的可信度评估框架，用于评估 AI 模型在特定使用情境中的可信性[15]。

（六）公共卫生与行业治理

公共卫生是生成式 AI 智慧医疗的重要增量场景。传统公共卫生系统依赖上报、统计和专家研判，存在信息滞后、跨部门协同难、基层负担重等问题。生成式 AI 可用于传染病舆情与病例信息摘要、应急预案生成、健康科普内容生成、疫苗接种提醒和公共卫生政策模拟。我国《实施意见》也把公共卫生列为重点应用方向之一，并要求夯实基础设施、丰富医疗数据供给、优化算力算法、加强中试基地建设和标准支撑，同时强化

[13] 参见 https://www.fda.gov/medical-devices/software-medical-device-samd/artificial-intelligence-enabled-medical-devices
[14] 参见 https://www.nature.com/articles/s41591-025-03743-2
[15] 参见 https://www.fda.gov/news-events/press-announcements/fda-proposes-framework-advance-credibility-ai-models-used-drug-and-biological-product-submissions

行业监管、预警机制、数据安全和个人隐私保护[16]。这意味着生成式 AI 智慧医疗的治理不应仅停留在单个模型或单个产品，而应覆盖数据、算法算力、审计和行业监管的全链条。

三、重点问题与风险挑战

（一）医疗幻觉与临床证据不足

生成式 AI 的生成机制决定了其可能在不确定或信息不足时产生看似可信但事实错误的回答。普通场景中的幻觉可能造成信息误导，医疗场景中的幻觉则可能影响诊疗判断、用药理解和就医时机。OpenAI 在 2025 年发布 HealthBench，构建包含 5000 个多轮健康场景对话的开放评测集，目标是根据医生专家认为重要的标准评估模型在真实健康场景中的表现。HealthBench 等评测的出现说明行业已经认识到，仅凭医学考试题得分不能证明模型适合临床使用，评估必须覆盖多轮对话、拒答能力、安全性、患者理解、临床上下文和不确定性表达[17]。

医疗生成式 AI 的证据问题还体现在外部有效性不足。许多研究在模拟病例、病例报告或结构化数据上表现良好，但真实世界中的患者叙述更模糊，检查结果更不完整，医生工作流更复杂。模型在一个国家、医院或人群中表现良好，并不意味着可迁移至其他地区和群体。尤其在基层医疗、老年患者、罕见病、儿童、孕产妇和多病共存场景中，模型错误可能更难被及时发现。因此，生成式 AI 医疗应用必须从模型准确率转向临床有效性和安全性，逐步引入前瞻性验证和应用持续监测。

（二）数据安全、隐私保护与可信数据供给矛盾

医疗数据具有高度敏感性和强情境属性。生成式 AI 模型需要大量高质量数据训练和微调，但医疗数据共享又受到隐私、伦理、数据产权、网络安全和跨境流动限制。具体而言，医疗生成式 AI 不仅处理结构化数据，也会处理医患对话、自由文本、音频、影像和多模态数据。此外，环境文书和健康助手可能持续收集患者对话与生活行为，智能体可能跨系统调取挂号、检查、药品、医保和可穿戴数据。如果缺乏数据脱敏和审计机制，生成式 AI 可能形成对个人健康画像的过度聚合。目前，我国《实施意见》提出，到 2027 年建立一批卫生健康行业高质量数据集和可信数据空间，这正是回应数据不能用、不敢用、不会用以及模型缺乏高质量本土医学数据之间矛盾的重要方向[18]。

[16] 参见 https://www.gov.cn/zhengce/zhengceku/202511/content_7047018.htm
[17] 参见 https://openai.com/zh-Hans-CN/index/healthbench/
[18] 参见 https://www.gov.cn/zhengce/zhengceku/202511/content_7047018.htm

（三）算法偏倚

医疗 AI 模型的性能高度依赖训练数据。如果训练数据主要来自大型三甲医院病患，模型可能在基层机构、欠发达地区、少数群体、儿童、老年人和罕见病患者中表现下降。生成式 AI 智慧医疗若要服务分级诊疗和基层能力提升，就必须避免技术越先进，越集中于高等级医院的反向不平等。基层应用不是简单下放模型，而是要建设面向常见病、多发病、慢病管理和公共卫生服务的可解释系统，并与上级医院远程会诊、转诊和质量控制机制联动。

四、典型案例分析

（一）临床文书案例：Dragon Copilot 与 Abridge

Dragon Copilot 和 Abridge 的共同点在于，它们没有把替代医生诊断作为首要卖点，而是从临床文书、对话摘要、信息检索和流程自动化切入[19]。Microsoft 将 Dragon Copilot 定位为统一的临床工作流 AI 助手，结合语音听写、环境聆听、生成式 AI 和医疗安全防护。Abridge 与 Mayo Clinic 的企业级协议则表明大型医疗系统愿意在可控场景中扩大生成式 AI 部署，以缓解医生文书负担并改善患者沟通。

（二）诊断智能体案例：AMIE 与 HealthBench

AMIE 的价值在于展示了生成式 AI 可通过对话进行病史采集和鉴别诊断，但其更大的治理意义是提出了如何评估医疗对话 AI 的问题[20]。HealthBench 则从另一侧回应这一问题，通过医生专家标准构建现实健康场景评测。这说明医疗生成式 AI 竞争的核心不只是模型参数，而是能否在真实交互中正确识别风险、表达不确定性、建议就医、遵循指南并避免越权诊疗。

（三）药物研发案例：Rentosertib

Rentosertib 案例表明[21]，生成式 AI 已经能够参与靶点发现和候选小分子设计，并将成果推进至人体临床试验。其治理意义在于，AI 药物研发不应只关注发现速度，还应关注模型可信度、实验可重复性和候选分子失败率。

[19] 参见 https://www.microsoft.com/en-us/health-solutions/clinical-workflow/dragon-copilot
[20] 参见 https://www.abridge.com/press-release/mayo-clinic-announcement
[21] 参见 https://www.nature.com/articles/s41591-025-03743-2

五、未来趋势研判

（一）从模型能力竞争走向临床证据竞争

当前及下一阶段生成式 AI 竞争将从谁的模型更强转向谁能证明在真实临床环境中安全、有效、可持续。医疗机构采购生成式 AI 系统时，将更关注本地验证结果、医生采纳率、错误率、患者安全事件、文书节省时间、临床结局改善和合规成本。第三方评测、注册研究、真实世界监测和行业基准将成为市场准入的重要条件。HealthBench、AMIE 等国际进展提示，医疗 AI 评测正在从静态题库走向多轮真实场景[22]。

（二）从单点产品监管走向全生命周期治理

医疗生成式 AI 具有持续更新、跨系统集成和场景迁移特征，因此传统一次审批、长期使用的监管方式面临挑战。FDA 强调 AI 医疗器械全生命周期管理，欧盟《人工智能法》要求高风险 AI 系统满足数据治理、透明度、人类监督和风险管理等要求[23]。我国也已形成 AI 医疗器械指导原则体系，并正在通过“人工智能+医疗卫生”政策推动应用基础、标准支撑和安全监管同步建设。当前及下一阶段监管重点将包括模型更新审批、性能漂移监测、不良事件上报和算法偏倚审计。

（三）从医院中心走向基层、家庭和连续照护

我国政策已明确到 2030 年基层诊疗智能辅助应用基本实现全覆盖[24]。因此，生成式 AI 智慧医疗将逐步从三甲医院创新项目走向社区卫生服务中心、乡镇卫生院、养老机构和家庭健康管理等应用场景。基层场景的关键不是部署最强模型，而是建立标准化知识库、常见病路径和远程协作机制。生成式 AI 若能帮助基层医生提高问诊完整性、用药安全、慢病随访和风险识别，将对医疗资源均衡具有重要意义。

六、结语

医疗是高风险、高伦理敏感、高责任约束的特殊领域。近两年来，生成式 AI 智慧医疗从技术展示进入政策牵引、产业落地和治理塑形并行的新阶段。其应用图景已覆盖居民健康助手、临床决策支持、医学影像、医疗文书、药物研发和公共卫生等多个环节。生成式 AI 智慧医疗不能简单套用通用大模型的发展逻辑，而必须在临床证据、数据安全、责任界分和人类监督基础上稳步推进。

[22]参见 https://openai.com/zh-Hans-CN/index/healthbench/
[23] 参见 https://eur-lex.europa.eu/eli/reg/2024/1689/oj
[24] 参见 https://www.gov.cn/zhengce/zhengceku/202511/content_7047018.htm

第六章 生成式 AI 在社会空间分析与公共治理中的应用与挑战

郑敏睿[1, 2]

1.中国人民大学公共管理学院 北京 100872

2.中国人民大学人工智能治理研究院 北京 100872

一、社会空间分析发展新形势

（一）“人工智能+”行动推动空间智能迈入规模化落地新阶段

2025 年 8 月国务院发布《关于深入实施“人工智能+”行动的意见》，10 月党的二十届四中全会将全面实施“人工智能+”行动写入“十五五”规划建议，标志着人工智能及其治理已完全融入国家中长期发展战略。作为人工智能与地理学、城市科学交叉融合的前沿领域，通用空间智能正从技术探索和试点示范，全面进入规模化、深层次融合应用的新阶段。

从产业规模看，国内地理大模型企业数量已突破 120 家，形成了从通用地理基础大模型到城市规划、自然资源、交通物流等垂直行业大模型的完整产品矩阵。应用深度从早期的地图导航、遥感解译等简单工具，向重塑公共治理核心生产流程演进。例如，基于人工智能的城市应急指挥系统使灾害响应效率提升 60%，公共服务设施智能布局方案使资源覆盖率平均提高 25%。这标志着空间智能不再仅是地理信息产业的增效器，更是驱动数字公共治理变革的重构者。

这一规模落地新阶段呈现出两个鲜明特征：一方面，空间智能与实体经济的结合点正从数据处理向自主决策突破，能够基于多源空间数据自动生成政策评估报告和规划方案；另一方面，开源地理大模型生态快速发展，极大降低了广大开发者和地方政府的创新门槛，推动空间智能从少数科研机构的尖端科技，转变为千行百业可便捷调用的普惠性融合创新基础设施。

（二）生成式 AI 技术突破催生空间分析范式革命

2025 年，人工智能技术实现了多维度、里程碑式的突破性跃进，AI 正从具备感知与理解能力的技术工具，加速演变为能够自主决策与执行的智能实体。这一变革在空间

分析领域表现得尤为深刻，传统基于线性假设和统计拟合的方法已难以应对当代复杂地理问题，生成式 AI 正在重塑我们理解、模拟和规划地理空间的方式。

传统空间分析方法存在三大核心局限：一是微观空间数据的稀疏性，问卷调查和抽样调查只能获得离散的点数据，难以形成连续的空间分布图景；二是复杂系统建模能力不足，难以捕捉城市与社会系统内在的非线性演化规律；三是非结构化数据空间化转换困难，大量蕴含空间信息的政策文本、社交媒体资源无法有效转化为可分析的空间数据。

生成式 AI 通过三大技术突破解决了上述问题：一是空间语言模型实现了自然语言与地理空间的双向映射，能够将“北京天安门广场南侧”等自然语言描述转化为机器可理解的空间坐标和拓扑关系；二是涌现出复杂空间推理能力，能够从简单的空间事实出发进行多步逻辑推演和反事实推理；三是合成空间数据技术实现了空间知识的“无中生有”，能够填补观测数据空白，生成反事实场景用于政策模拟。

二、生成式 AI 在社会空间分析的发展

（一）地理大模型与空间映射

地理大模型的出现标志着空间分析进入了的新时代。与传统针对特定任务设计的空间模型不同，地理大模型通过在海量多模态地理数据（包括卫星影像、地图数据、地理文本、传感器数据等）上进行预训练，获得了对世界的通用表征能力。空间语言模型作为其核心组件，能够将自然语言中的空间描述（如“北京天安门广场南侧”、“长江中下游平原”）转化为机器可理解的空间坐标和拓扑关系，实现了语言空间与地理空间的双向映射。

Mai 等人[1]（2025）针对当前人工智能 + 地理”领域“技术快速迭代但缺乏系统理论框架”、“通用大模型直接迁移地理任务效果不佳”、“技术发展与伦理治理脱节”三大核心问题，首次全面系统地定义了下一代“人工智能 + 地理”的概念、技术体系和发展路线图。他们明确指出地理基础大模型不是“通用大模型 + 地理插件”，而是原生融入地理空间独特属性的专用基础模型，必须内置空间异质性、拓扑关系、地理过程规律三大核心地理约束。

（二）复杂空间推理能力的涌现

生成式 AI 最具革命性的贡献在于其涌现出的复杂空间推理能力。传统空间分析主要停留在位置提取和空间关联分析层面，即从数据中识别地理实体的位置并计算其空间

[1] Mai, G., Xie, Y., Jia, X., et al. (2025). Towards the next generation of geospatial artificial intelligence. International Journal of Applied Earth Observation and Geoinformation, 136:104368.

相关性。而生成式模型能够从简单的空间事实出发，进行多步逻辑推演，甚至进行反事实推理和因果推断。

这种能力使得空间分析从描述性研究走向了预测性和处方性研究。例如，当评估一条新建高速铁路的经济影响时，生成式模型不仅能够计算空间可达性的变化，还能综合考虑产业结构、人口流动、资源禀赋等多种因素，生成分区域、分行业的详细影响评估报告，并提出针对性的政策建议。

（三）认识论转变

生成式 AI 带来了空间分析认识论的根本性转变。传统定量地理学遵循“观测数据—统计模型—经验知识”的研究范式，强调对既有观测数据的挖掘和拟合，知识的边界被数据的边界所限制。而生成式 AI 引入了合成数据的概念，实现了空间知识的“无中生有”。

合成空间数据不是对真实数据的简单复制，而是基于真实数据的统计规律和地理约束生成的、具有相同分布特征但不存在于现实世界的数据。它能够填补观测数据的空白，解决数据稀疏性问题；能够生成反事实场景，用于政策模拟和风险评估；还能够保护数据隐私，避免敏感地理信息的泄露。Hochmair 等人[2]（2025）指出生成式 AI 已成为地理空间科学最具变革性的技术，但现有研究存在“重应用轻方法”、“重单一模型轻系统对比”、“重技术实现轻质量评估" 三大缺陷，缺乏对 AI 驱动地理空间数据生成全流程的系统性梳理。他们认为合成数据的广泛应用将使空间分析从数据驱动转向知识驱动，极大地拓展了地理学研究的可能性空间。但是需要对合成数据进行真实性、有用性、合规性的三维评估。生成式空间分析与传统空间分析在核心特征上有显著差异，详见表 1。

表 1 传统空间分析 vs. 生成式空间分析范式对比

维度	传统空间分析	生成式空间分析
输入数据	结构化空间数据（矢量、栅格）、有限非结构化数据	多模态数据（文本、图像、音频、视频）、合成数据
模型核心机制	统计拟合、规则推理、判别式学习	生成式学习、自监督预训练、涌现式推理
输出形态	统计结果、预测值、静态地图	合成数据、动态模拟、自然语言报告、交互式可视化
核心能力	描述、解释、预测	描述、解释、预测、生成、设计、推演
知识边界	受限于观测数据	超越观测数据，可生成反事实知识

[2] Hochmair, H. H., Juhász, L., & Li, H. (2025). Advancing AI-Driven Geospatial Analysis and Data Generation: Methods, Applications and Future Directions. ISPRS International Journal of Geo-Information, 14(2), 56. https://doi.org/10.3390/ijgi14020056.

三、热点议题

（一）社会地理与社会绩效的空间评估

长期以来，微观社会数据的空间分布不均一直是制约研究精度的关键因素。问卷调查往往只能在有限的采样点收集数据，而人口普查数据又存在更新周期长、粒度粗的问题。生成式模型能够基于有限的采样数据和辅助地理数据（如土地利用、夜间灯光、POI分布），生成高分辨率、连续分布的合成社会空间数据。例如，Sukthong[3]（2025）提出了 GAN-VAE 混合生成模型，结合 GAN 和 VAE 的优势，解决小样本条件下的空间数据补全和预测问题，为地理空间情报提供更可靠的技术支撑。这种方法不仅能够填补数据空白，还能够生成不同时间节点的合成数据，用于分析社会空间结构的动态演变。在公共卫生领域，生成式模型可以模拟传染病的空间传播过程，预测不同防控措施下的疫情发展趋势，为公共卫生决策提供科学依据。

社会绩效评价的多维推演是生成式 AI 的另一重要应用方向。城市社会绩效包括资源公平性、宜居性、社会福祉、包容性等多个维度，这些指标往往难以直接量化，且相互之间存在复杂的非线性关系。传统评价方法通常采用加权求和的方式，主观性强且难以反映系统的整体性。生成式多模态大模型能够整合来自不同来源的异构数据（包括统计数据、遥感影像、社交媒体文本、市民投诉记录等），对城市社会绩效进行全方位、多维度的动态评估。例如，模型可以分析社交媒体上的文本情感，结合空气质量、噪音水平、公共服务设施分布等数据，生成城市不同区域的宜居性指数地图。更重要的是，模型能够进行“what-if”情景模拟，推演不同政策干预下社会绩效的变化趋势。例如，模拟在某个区域新建一所学校或医院后，周边居民的教育和医疗可及性将如何改善，以及这种改善会对房价、人口流动产生什么影响。

（二）数字公共治理中的空间决策与优化

政策文本的空间化解析为数字公共治理提供了新的技术手段。公共政策本质上是一种空间干预行为，几乎所有政策都具有空间维度。然而，传统政策分析主要关注文本内容和制度设计，对政策的空间影响缺乏系统评估。生成式大语言模型能够自动解析长文本政策文件，提取其中的空间要素（如政策覆盖区域、重点发展区域、限制开发区域）和政策措施，然后结合地理数据生成政策影响力的空间辐射圈。例如，模型可以分析国家“十四五”规划中关于区域协调发展的内容，自动识别出京津冀、长三角、粤港澳大湾区等重点发展区域，以及交通、能源、生态等领域的重大项目布局，然后生成政策影

[3] Napat Sukthong. (2025). Enhancing spatial data analysis with generative AI: a novel approach for geospatial intelligence, Proc. SPIE 13731, Seventh International Conference on Image, Video Processing, and Artificial Intelligence (IVPAI 2025), 137311E (29 August 2025); https://doi.org/10.1117/12.3075499

响力的空间分布图，直观展示不同区域从政策中受益的程度。这种方法不仅能够提高政策分析的效率，还能够发现政策文本中隐含的空间矛盾和冲突，为政策优化提供建议。

公共服务设施的生成式布局是生成式 AI 在空间规划领域的突破性应用。公共服务设施（如学校、医院、公园、消防站）的布局直接关系到居民的生活质量和社会公平。传统布局方法通常基于可达性分析和优化算法，但这些方法难以同时考虑多种约束条件和复杂的社会需求。基于扩散模型和强化学习的生成式空间设计方法能够在满足多种约束条件（如人口分布、服务半径、建设成本、生态保护）的前提下，自动生成多种可行的设施布局方案，并对每个方案进行量化评估。例如，模型可以根据一个城市的人口分布和现有医疗资源情况，生成多个新建医院的选址方案，每个方案都包含具体的位置、规模和服务范围，并计算出每个方案的平均就医时间、服务人口覆盖率和建设成本。决策者可以根据这些信息选择最优方案，或者对不同方案进行组合优化。这种方法不仅能够提高规划的科学性和效率，还能够实现公共资源的公平分配。

（三）地球系统与城市计算的跨模态生成

Text-to-Map（文本到地图）技术实现了空间可视化的革命性突破。传统地图制作需要专业的 GIS 软件和技能，制作周期长，难以满足快速变化的需求。生成式 AI 使得任何人都可以通过自然语言指令快速生成定制化的专题地图。用户只需输入“绘制 2025 年北京市各区县 GDP 分布图”、“展示过去十年中国人口老龄化的空间变化趋势”或“标出我家附近 3 公里范围内的所有公园和健身房”，模型就能够自动检索相关数据，选择合适的地图投影和符号系统，生成美观、准确的专题地图。更高级的 Text-to-Map 系统还支持交互式编辑，用户可以通过自然语言调整地图的样式、范围和内容，实现 "所想即所得" 的空间可视化。这种技术极大地降低了空间分析的门槛，使得地理信息能够更广泛地服务于公众和决策者。

遥感与空间矢量数据的合成为地球系统科学和城市计算提供了新的数据来源。高分辨率遥感影像是监测地球表面变化的重要手段，但受限于卫星重访周期和天气条件，往往难以获得连续、清晰的观测数据。生成式模型能够基于有限的观测数据生成缺失的影像，或者将低分辨率影像提升为高分辨率影像。例如，GAN 和变分自编码器（VAE）可以根据多云天气下的遥感影像生成无云的合成影像，或者根据历史影像预测未来的地表覆盖变化。在城市计算领域，生成式模型可以根据城市规划图和建筑设计图生成三维城市模型，用于城市风环境、热环境和交通流的模拟。Lu 等人[4]（2025）通过对 Web of Science 核心合集 2014-2024 年发表的 2386 篇相关论文进行时空和语义分析，揭示了“人工智能 + 地理”在城市计算领域的发展规律和研究热点。研究发现，2021 年以来

[4] Lu, F., Cheng, S., & Wang, P. (2025). GeoAI enabled urban computing: status and challenges. Annals of GIS, 31 (4), 537–555. https://doi.org/10.1080/19475683.2025.2552152.

该领域论文发表量呈爆发式增长，研究重点已从单一任务模型转向多模态基础大模型。并且在智能交通系统、公共安全、环境监测、可持续城市发展的四大核心领域有了相关突破性发展。

四、挑战与风险

（一）“空间幻觉”难题

生成式模型最突出的问题是“空间幻觉”，即模型常常生成看似合理但实际上违背地理常识和拓扑约束的空间关系。Janowicz 等人[5]（2026）在对 ChatGPT 的地理推理能力进行全面评估后发现，模型在回答关于距离、方向、边界和地形的问题时经常出现错误，甚至会虚构不存在的地理实体和空间关系。

空间幻觉产生的根本原因在于生成式模型缺乏对地理空间的物理理解。模型学习到的是文本和图像中空间描述的统计规律，而不是真实世界的物理定律和地理约束。例如，模型可能知道“长江发源于青藏高原，注入东海”，但它并不真正理解河流的流向是由地形决定的，因此可能会生成“长江自东向西流”这样的错误描述。此外，训练数据中的错误和不一致也会被模型学习并放大，导致更严重的空间幻觉。

解决空间幻觉难题需要从多个方面入手：一是构建高质量的地理知识图谱，将明确的地理常识和拓扑约束注入模型；二是开发空间感知的生成式模型架构，使模型能够更好地理解和表示空间关系；三是建立严格的空间验证机制，对模型生成的结果进行地理一致性检查。

（二）算法黑盒与空间可解释性

生成式模型的“黑盒”特性是其在公共政策制定和科学研究中应用的主要障碍。在量化研究中，研究人员需要理解模型的决策过程，以确保结果的科学性和可靠性；在公共政策制定中，决策者需要向公众解释政策的依据，以获得公众的理解和支持。然而，当前的大语言模型和扩散模型包含数十亿甚至上万亿个参数，其内部工作机制极其复杂，难以进行直观的解释。

空间可解释性问题比一般的 AI 可解释性问题更加复杂，因为空间数据具有独特的拓扑结构和自相关性。传统的模型解释方法（如特征重要性分析、注意力可视化）在应用于空间数据时往往效果不佳，难以揭示模型是如何利用空间信息进行决策的。例如，当模型预测某个区域的犯罪率较高时，我们无法确定它是基于该区域的人口密度、经济水平、历史犯罪记录，还是其他空间因素。

[5] Janowicz, K., Mai, G., Zhu, R., Gao, S., Wang, Z., Hu, Y., & Bennett, L. (2026). Geography According to ChatGP T -- How Generative AI Represents and Reasons about Geography.

提高生成式空间模型的可解释性需要跨学科的合作。计算机科学家需要开发专门针对空间数据的可解释 AI 技术；地理学家需要提供关于空间过程和地理规律的先验知识；社会科学家需要研究算法决策对社会的影响。只有这样，我们才能打开算法黑盒，使生成式 AI 的决策过程更加透明和可信。

（三）伦理风险与算法偏见

生成式 AI 在空间分析中的应用带来了一系列伦理风险，其中最突出的是算法偏见及其导致的空间不平等。生成式模型是在历史数据上训练的，而历史数据中往往包含着各种社会偏见和不平等。这些偏见会被模型学习并在生成结果中放大，形成数字红线，即算法在空间维度上对特定群体的系统性歧视。例如，国内某主流地理大模型在公共服务设施布局任务中，明显偏向高收入社区：在相同人口规模下，高收入社区获得的公共服务设施数量是低收入社区的 1.7 倍。此外，合成空间数据的滥用也可能带来严重的伦理问题。恶意用户可以利用生成式模型生成虚假的地图和地理信息，用于误导公众、实施诈骗或进行军事欺骗。在隐私保护方面，虽然合成数据可以在一定程度上保护个人隐私，但如果模型生成的数据过于逼真，仍然可能被用于识别特定个人或群体，详见表 2。

表 2 生成式 AI 在空间分析中的风险

风险维度	具体风险点	影响程度	发生概率	应对策略
技术风险	空间幻觉与地理错误	高	高	注入地理知识图谱；建立空间验证机制
	模型可解释性不足	高	中	开发空间可解释 AI 技术；引入人类专家审核
	模型泛化能力差	中	中	使用多样化训练数据；进行跨区域验证
数据风险	训练数据质量低	中	高	建立高质量地理数据集；进行数据清洗和标注
	数据更新不及时	中	中	建立动态数据更新机制；融合实时传感器数据
	数据隐私泄露	高	低	使用差分隐私技术；生成合成数据替代真实数据
伦理风险	算法偏见与空间不平等	高	中	进行算法公平性审计；建立多元参与的决策机制
	虚假地理信息传播	高	中	开发虚假信息检测技术；加强公众地理素养教育
	技术滥用与恶意使用	高	低	制定相关法律法规；建立技术使用伦理规范

五、未来展望

（一）构建边界清晰的空间智能权责框架

坚持统筹发展和安全，建立健全涵盖高中低位阶的法律法规、监管政策、技术标准等制度体系。一是明确数据产权制度边界，界定地理数据的所有权、使用权和收益权，探索建立基于贡献度的空间数据权益分配模式。二是合理设定责任范畴，针对空间智能系统研发、部署、使用环节主体多元、因果链条复杂的特性，明确归责原则，设定与技术能力相适应的多方主体法律责任。三是探索推进场景应用规则，面向国土空间规划、城市治理、应急管理等具体应用场景，结合行业特性需求细化上位法规则，制定专项法规和标准。

（二）发展敏捷动态的空间智能监管工具

一是推行创新场景“沙盒监管”试点。对有望推动空间智能取得突破性进展的创新应用，在有条件的区域试点“监管沙盒”，明确申请流程、测试期限、风险监测、责任豁免及争议解决等核心机制。二是强化空间智能安全能力建设。构建国家级地理大模型测试验证平台，大力支持第三方审计、评估、认证机构发展，提供模型测试验证、空间一致性检查、算法公平性审计等服务。三是构建多主体协同联动的风险预警与应对机制。依托监管平台实时归集监管数据，推动企业定期报送风险阶段性评估报告，建立完善内部监测与预警机制。

（三）塑造人机协同的空间决策模式

一是科学划定人机交互边界。坚持技术发展以增进人类福祉为根本宗旨，明确空间智能的辅助性工具地位，确保人类在空间决策中的最终控制权。在涉及公共利益和重大安全的决策场景中，必须保留人类专家的审核和干预环节。二是建立动态伦理监测机制。优化覆盖全生命周期的社会伦理影响评估机制，动态监测空间智能对公众心理健康、行为模式、社会家庭结构、就业结构等方面的影响。三是深化透明披露要求。制定空间智能透明度标准，提升训练数据集的代表性与标注规范性，开发空间可解释 AI 技术。空间智能系统提供决策建议时，应当以显著提示方式向用户披露 AI 身份和决策依据。

（四）推进公平普惠的空间智能治理

一是弥合数字鸿沟确保公平受益。实施全面的全民数字素养与空间智能通识教育计划，将其融入国民教育体系与社区公共服务。通过基础设施投资、公共 AI 服务供给、技能培训等方式，确保不同地区、不同群体公平地获取和受益于空间智能技术。二是加强反垄断保障公平竞争市场环境。关注大型平台利用空间智能生态系统构建封闭壁垒、限制互操作性等行为，维护开放、公平的市场环境。推动关键空间智能基础设施的开放

与共享，降低中小企业创新门槛。三是建立健全就业替代社会保障体系。设置专项就业缓冲基金，完善灵活就业社保体系，探索全民基本收入等前瞻性的全民保障机制，帮助受空间智能技术影响的劳动者实现职业转型。

（五）重点研究方向

生成式 AI 与空间分析的融合不是简单的技术叠加，而是一场深刻的学科革命。它要求计算科学与社会科学进行深度交汇，打破传统学科之间的壁垒。计算机科学家需要学习地理学的空间思维和理论方法，以开发更加符合地理规律的 AI 模型；地理学家和公共管理学者需要掌握生成式 AI 的技术原理和应用方法，以更好地利用这一工具解决复杂的地理和社会问题。面向未来，生成式空间分析领域有许多值得深入研究的前沿方向：

一是基于大模型的小样本空间推演。当前生成式模型需要大量的训练数据才能取得良好的效果，但在许多地理应用场景中，标注数据非常稀缺。研究如何利用大模型的通用知识和迁移学习能力，在小样本甚至零样本条件下进行准确的空间推理和预测，具有重要的理论和实践价值。

二是可解释的社会空间生成方法。如前所述，生成式模型的“黑盒”特性是其在社会科学研究和公共政策制定中应用的主要障碍。研究如何提高生成式空间模型的可解释性，使模型的决策过程更加透明和可信，是未来研究的重点方向。

三是多智能体生成式空间模拟。城市和社会系统是由大量相互作用的个体组成的复杂系统。研究如何利用生成式 AI 和多智能体系统技术，模拟个体的行为决策及其在空间上的涌现效应，能够帮助我们更好地理解城市演化和社会变迁的规律。

四是负责任的生成式空间 AI。研究生成式 AI 在空间分析中应用的伦理、法律和社会影响，建立负责任的 AI 开发和使用框架，确保技术的发展符合人类的共同利益，是一个亟待加强的研究领域。

第七章 生成式 AI 时代双向人机对齐

塔娜[1, 2]

1.中国人民大学新闻学院 北京 100872

2.中国人民大学人工智能治理研究院 北京 100872

一、引言：对齐问题成为生成式 AI 的核心议题

进入 2025 年以来，生成式 AI（Generative AI）在全球范围内加速渗透至内容生产、决策支持、教育服务与情感交互等多个关键领域。随着大语言模型能力的持续增强，人机互动的形态正在发生根本性转变：人工智能不再只是被动执行命令的工具，而逐渐成为参与意义建构的互动主体。在这一转变过程中，“对齐（alignment）问题”迅速从技术议题上升为社会议题，成为制约生成式 AI 进一步发展的关键瓶颈[1]。

传统意义上的对齐，强调人工智能对人类意图的准确理解与执行，属于单向适配逻辑。然而，在当前复杂多变的应用情境中，仅依赖单向对齐已难以支撑高质量的人机互动。一方面，模型难以完整理解用户隐含的语境、价值与情感需求；另一方面，用户也往往对模型能力边界、响应机制及潜在偏差缺乏清晰认知。这种双向失配不仅影响任务完成效果，也削弱了用户对系统的信任基础。

在此背景下，“双向人机对齐”成为理解当代人机关系的重要理论框架。所谓双向对齐，是指人工智能对人类意图、语境与价值的适配过程，以及人类对人工智能能力、行为模式及限制条件的认知调适之间形成的动态互动过程。这一过程并非一次性完成，而是在人机持续交互中不断修正、协商与重构。

本研究报告以“提示词（prompt）”为核心切入点，探讨其在双向人机对齐中的关键作用。报告聚焦三个核心问题：第一，双向人机对齐的结构与机制如何构成；第二，提示词如何作为语言媒介驱动对齐过程；第三，对齐如何在不同情境中表现出差异化形态。在此基础上，进一步分析对齐失效情境，并提出对应的治理路径与对策建议。

[1] 李霄骞，杨婧文，塔娜. 提示词设计如何塑造人机关系：从双向对齐到情境化互动[J/OL]. 新媒体与网络, 2025.

二、双向人机对齐的理论框架：从控制接口到关系机制

（一）从单向适配到双向协商

在生成式 AI 早期发展阶段，人机关系主要体现为从输入到输出的线性结构，用户通过指令（即提示词）驱动模型生成结果，对齐的核心在于提高模型对提示词的理解准确性。然而，随着模型能力的提升与应用场景的复杂化，这种单向逻辑逐渐暴露出局限性。用户的表达往往不完整、含糊甚至存在矛盾，而模型的输出也可能在语义、情境或价值层面出现偏差。

因此，对齐逐渐从技术校准转向互动协商。人机双方都参与到意义的建构之中，用户通过不断调整提示策略以适应模型特性，模型则通过语境推断与反馈机制不断逼近用户意图。对齐不再是静态目标，而是一种持续发生的过程性状态。

（二）提示词的关系性转向

提示词在这一过程中扮演着关键角色。提示词不仅是传递指令的技术接口，更是建构人机关系的语言媒介。它既规定了互动的起点，也在一定程度上预设了互动的走向。具体而言，提示词具有三重功能：其一是意图表达功能，通过结构化语言显化用户目标；其二是语境建构功能，通过补充背景信息限定生成范围；其三是规范嵌入功能，通过风格、情感与价值设定引导模型输出。这种多层功能使提示词成为连接人类认知与机器生成逻辑的桥梁。

（三）双向对齐的五维机制

在具体机制上，双向人机对齐可以从五个维度进行分析：

第一，意图对齐。指用户目标是否被模型准确识别并转化为输出结果。意图表达的清晰度直接影响对齐质量。

第二，语境对齐。强调人机双方对情境背景、角色关系与任务环境的理解一致性。缺乏语境信息往往导致生成偏差。

第三，风格对齐。涉及语言表达方式、语气与交互节奏的匹配。风格一致性有助于提升交互自然性与可接受度。

第四，情感对齐。指模型在回应中对用户情绪的识别与适配能力，在社交与支持性场景中尤为关键。

第五，价值对齐。是对齐体系中的最高层，涉及伦理规范、文化立场与社会价值的匹配。

这五个维度呈现出层级关系：价值对齐为整体提供规范边界，风格与情感对齐作为中介支撑互动体验，意图与语境对齐则直接影响任务执行效果。

（四）对齐失效的基本机制

对齐并非总能成功实现，其失败往往源于以下几类机制：提示词表达不充分导致意图偏差，语境信息缺失造成理解错误，模型在复杂情境下的推理能力不足，以及价值冲突引发的不一致输出。此外，恶意提示或提示注入攻击也可能直接破坏对齐结构，使模型偏离原有目标。

三、情境化对齐：双向对齐的实践形态

（一）任务导向情境：协作型对齐

在任务导向场景中，人机互动以效率与准确性为核心目标。提示词通常呈现出高度结构化特征，包括明确的任务分解、步骤指引与输出约束。在此情境下，对齐表现为一种协作关系。用户通过提示词设定任务框架，模型则在该框架内执行具体操作。随着多轮交互的展开，双方逐渐形成分工机制：人类负责目标设定与评估，人工智能承担信息处理与内容生成。这种协作关系的质量主要取决于对齐程度的高低。此外，可解释性提示（如要求模型展示推理过程）与置信度提示（如说明不确定性）也是提升对齐质量的重要手段，有助于增强用户对模型行为的理解与控制感。

（二）社交导向情境：关系型对齐

在社交与情感支持场景中，对齐的目标从任务完成转向关系建构。提示词在此不再仅仅用于信息传递，而更多承担情绪调节与互动引导功能。通过角色设定、语气调整与情感嵌入，用户可以引导模型呈现出特定的互动风格，如“倾听者”“导师”或“陪伴者”。在持续互动中，用户往往会将模型视为具有社会属性的对象，从而形成拟社会关系。这种关系的稳定性依赖于情感与风格层面的持续对齐。

（三）情境转换中的对齐张力

在现实应用中，任务与社交情境往往共同存在。例如，在教育场景中，人工智能既需要提供知识支持，又需维持鼓励与互动氛围。这种情境混合对对齐提出了更高要求。若模型在任务场景中过度拟人化，可能降低专业性；而在情感场景中过于机械，则会削弱用户体验。因此，不同情境之间的切换与平衡，成为对齐机制面临的重要挑战。

四、典型情境分析：对齐的实现与失效

在生成式 AI 的实际应用中，双向人机对齐并非抽象理论，而是在具体互动过程中不断被建构、修正与破坏的动态实践。从现实情境来看，对齐既可能在多轮协作中逐步收敛，也可能在关键节点迅速失效，其运行状态高度依赖提示词设计、用户认知以及任务环境的综合作用。

在高水平对齐的场景中，人机互动通常呈现出逐渐收敛的特征。以 AI 辅助写作为代表，用户往往从模糊需求出发，通过多轮提示逐步细化目标结构。在这一过程中，用户学习如何以更结构化、约束化的方式表达意图，而模型则通过上下文积累不断修正理解路径，使输出逐渐逼近预期。这一过程体现了典型的双向适配：不仅 AI 在“理解”人，人也在学习如何被 AI 理解。对齐由此不再是单次完成的结果，而是通过互动不断逼近预期目标的过程。

类似机制也存在于情感支持类应用中。当用户在提示中嵌入明确的情绪线索与情境描述时，模型可以生成相对稳定且具有共情特征的回应。在连续交互中，这种情绪一致性被逐步强化，使用户产生被理解的感受，并在心理层面形成一定程度的依附关系。值得警惕的是，这种情感对齐并非源于模型真实理解情绪，而是源于提示词成功构建了一个可被用户感知为情感一致的互动结构。

然而，对齐同样具有显著的不稳定性。在提示表达不充分或存在歧义时，模型往往基于不完整信息进行推断，导致输出偏离用户真实意图。这种意图对齐的偏差会迫使用户进入反复修正循环，显著增加交互成本。在语境层面，当提示缺乏必要背景信息时，即便语言表面流畅，内容仍可能与实际情境脱节，形成形式正确、实质错误的输出结果。此外，在价值敏感议题中，模型可能生成带有偏见或不符合社会规范的内容，反映出价值对齐的脆弱性。

更为复杂的是，对齐机制在特定条件下可能被主动操控。提示注入（prompt injection）通过嵌入误导性或恶意信息，诱导模型偏离原有任务目标。这类情境表明，对齐不仅可能失败，还可能被利用，成为攻击系统行为的入口，从而暴露出当前生成式 AI 在安全性与鲁棒性方面的结构性不足。

在上述多种情境基础上，可以归纳出具有解释力的如下关键发现。首先，双向人机对齐本质上是一种持续发生的动态过程，而非一次性完成的匹配结果。对齐依赖于多轮互动中的不断修正与反馈，具有明显的过程性特征。其次，对齐呈现出结构上的非对称性。尽管理论上强调双向适配，但在实践中，用户往往承担更多调适成本，需要不断优化提示策略以适应模型特性。第三，对齐具有显著的情境依赖性。在任务导向场景中，对齐主要表现为效率与准确性的提升，而在社交情境中，则转化为信任、亲密感与持续

互动的生成机制。第四，对齐存在感知与实际的偏差。用户往往依据语言流畅性等表面线索判断模型能力，从而高估其理解水平，这种偏差可能带来认知风险。第五，对齐具有明显的脆弱性与可操控性，提示词的不确定性或恶意输入都可能迅速破坏既有对齐状态。第六，对齐过程正在呈现出反向塑造效应，模型输出不仅回应用户需求，还可能在长期互动中影响用户的认知方式与判断结构。换言之，双向人机对齐并非稳定结构，而是一种高度动态、情境敏感且易受干扰的互动过程。

五、双向人机对齐的演化机制：过程、结构与影响

以上情境揭示了双向人机对齐如何表现，进一步的问题则在于这种对齐为何呈现出动态性、不稳定性与反向影响等特征。这就需要将经验性发现上升为机制性解释，从过程、结构与社会影响三个层面加以理解。

从过程层面看，双向人机对齐可以被理解为一种循环式协商机制。在人机互动中，提示词并非一次性输入，而是持续迭代的表达装置。用户通过提示表达意图，模型据此生成回应，用户再根据回应调整表达方式，进入下一轮互动。这一过程构成一个不断循环的结构：表达—解释—反馈—修正—再表达。在这一循环中，对齐并非逐步线性提升，而是可能出现反复震荡与阶段性收敛。对齐的质量取决于这一循环是否能够有效减少信息不确定性，并逐步建立共享语境。这一机制解释了为何多轮交互通常优于单轮输入，也揭示了提示词在对齐中的核心地位。

从结构层面看，双向人机对齐内在地包含多重张力关系，这些张力决定了对齐难以达到完全稳定状态。其一是控制与自主之间的张力。用户试图通过提示词控制模型输出，但模型自身的生成逻辑与概率机制使其具有一定自主性，两者之间始终存在偏差空间。其二是效率与安全之间的张力。高度结构化的提示可以提升效率，但也可能限制模型灵活性；而增强安全约束则可能降低响应的开放性与创造性。其三是拟人化与去人格化之间的张力。在社交情境中，拟人化有助于提升体验，但过度拟人化又可能导致用户误判模型能力边界。这些结构性张力意味着，对齐并不是可以彻底优化的问题，而更像是一种需要持续调节的平衡状态。

再次，从认知与社会影响层面看，双向人机对齐正在形成一种共塑机制。传统观念中，人机关系被理解为人类主导技术工具的单向关系，但在生成式 AI 环境中，模型输出开始反过来影响用户的认知结构与决策方式。当用户在反复互动中逐渐适应模型表达逻辑，其思维方式也可能发生变化，例如更倾向于结构化表达、依赖模型建议进行判断等。与此同时，用户通过提示词不断调整模型行为，实际上也在参与模型使用层面的训练过程。这种双向塑造关系表明，对齐不仅是技术适配问题，更是一种深层的认知互动机制。

进一步来看，对齐还存在明确的边界条件。首先，由于文化差异与价值多元，不同用户之间对“合理输出”的理解存在显著差异，这使得价值对齐难以形成统一标准。其次，语言表达本身具有模糊性与不完备性，提示词不可能完全覆盖用户意图，从而使对齐始终存在不确定空间。最后，模型本身的训练数据与算法机制也会对对齐结果产生限制，使其无法在所有情境下实现稳定表现。

因此，可以将双向人机对齐理解为一种有限稳定系统，它能够在一定范围内通过互动实现相对一致，但始终受到表达、认知与技术条件的多重约束。这种系统既具有适应性，也具有脆弱性；既能够促进高效协作，也可能引发认知偏差与行为依赖。

综合来看，双向人机对齐不应单纯被建构为技术问题，而应被视为一种需要在人机动态互动中持续优化的关系过程。理解这一点，对于后续构建更加稳健、透明且可治理的人机互动体系具有重要意义。

六、对策建议：提示词驱动的双向人机对齐治理机制与路径

基于前文对双向人机对齐机制及其情境化表现的系统分析可以发现，生成式 AI 的治理问题已从传统以模型性能与内容控制为核心的技术议题，转向以交互过程、关系结构与价值协商为核心的综合性问题。在这一转变中，提示词不再只是驱动模型生成的输入工具，而成为连接人类认知、机器响应与社会规范的关键媒介。因此，围绕提示词展开治理设计，是实现稳定、可持续人机对齐的重要路径。

从技术层面来看，当前对齐机制仍以结果导向为主，即通过输出约束来降低错误或风险。然而，双向人机对齐本质上是一个持续循环的互动过程[2]，其运行依赖于“表达—解释—反馈—修正”的动态机制。其中，提示词的不确定性往往成为对齐失效的重要来源。当提示表达模糊或信息缺失时，模型容易产生偏差，并在多轮交互中不断放大误解[3]。因此，技术治理应从单一的输出控制转向全过程嵌入。一方面，需要在提示输入阶段强化安全与稳健机制，通过提示注入防御、语境一致性检测与异常语义识别等手段，降低输入不确定性对模型行为的干扰；另一方面，应提升模型在互动中的自我调节能力，使其能够识别不完整信息并主动发起澄清，通过展示推理路径与不确定性信息增强可解释性，从而在多轮交互中逐步逼近用户真实意图。随着任务情境与社交情境的相互交融，模型还需具备一定程度的情境识别与策略切换能力，以缓解不同互动目标之间的对齐张力。

[2] SHEN H, KNEAREM T, GHOSH R, et al. Towards Bidirectional Human-AI Alignment: A Systematic Review for Clarifications, Framework, and Future Directions[EB/OL]. arXiv, 2024.
[3] YANG C, SHI Y, MA Q, et al. What Prompts Don’t Say: Understanding and Managing Underspecification in LLM Prompts[EB/OL]. arXiv, 2025.

从制度层面来看，提示词作为人机互动的核心接口，其治理已不再只是技术问题，而需要纳入多层级制度框架之中。宏观层面，应在现有生成式 AI 监管体系中明确提示词的关键地位，将其视为影响生成结果的重要环节，并推动形成围绕提示设计与内容生成的责任划分机制。同时，应强化模型运行的透明性与可审计性，使生成逻辑在一定程度上对监管与公众可解释。在平台层面，有必要建立覆盖提示输入与生成输出的双重治理机制，对潜在高风险提示进行识别与约束，并通过动态监测手段对多轮交互中的风险演化进行持续评估。此外，通过提供生成依据与逻辑说明，可以增强用户对模型行为的理解与信任。在行业层面，应推动提示词设计规范与人机交互伦理标准的建设，尤其是在价值多元与文化差异背景下，通过规范化路径缓解价值对齐的不确定性与冲突问题。

从用户层面来看，双向人机对齐的实现不仅依赖模型能力，也取决于用户的表达策略与认知水平。现实互动中，用户往往承担更多的调适成本，需要不断优化提示策略以适应模型特性。因此，提升“提示素养”成为治理体系的重要组成部分。这一能力不仅体现在能够清晰表达目标、语境与约束条件，还包括在多轮交互中根据反馈动态调整提示策略，并对生成结果进行批判性评估的能力。同时，有必要强化用户对人工智能能力边界的认知，避免因语言流畅性而高估模型理解能力，从而引发过度信任或依赖风险。在此基础上，应引导用户形成协作导向的使用观，将生成式 AI 视为辅助工具而非决策主体，以降低长期互动中反向塑造效应的潜在影响，即模型输出对人类认知方式与判断结构的反向影响。

从社会层面来看，提示词驱动的人机互动正在重构传统社会关系结构，其影响已超越技术范畴而进入认知与文化层面[4]。首先，在社交与情感支持情境中，提示词通过角色设定与情绪引导，可能促使用户对人工智能产生拟社会关系甚至情感依附。因此，有必要在产品设计与制度规范中强化人工智能的身份标识，通过明确其非人属性与交互边界，防止过度拟人化带来的认知偏差。其次，提示词作为价值嵌入的重要媒介，其设计过程可能影响生成内容的立场与倾向，从而在潜移默化中参与社会价值的传播与建构。在多元文化与价值差异并存的背景下，如国际传播等语境中，应通过多元价值校验与跨文化适配机制降低偏见风险。随着人机互动的持续深化，用户的表达方式与思维结构可能在长期互动中发生变化，形成一种人机共塑的认知机制，对于这一趋势，应考虑在教育与社会层面强化批判性思维培养，以防止对技术的过度依赖及思维外包现象。

总体来看，围绕提示词展开的双向人机对齐治理路径，同样体现出从结果控制向过程治理、从技术规制向关系调节的转型趋势。提示词作为连接人类与人工智能的关键节点，其设计与使用方式不仅影响单次交互效果，更在长期互动中塑造人机关系的结构与

[4] RAPP A, CURTI L, BOLDI A. The Human Side of Human-Chatbot Interaction: A Systematic Literature Review of Ten Years of Research on Text-Based Chatbots[J]. International Journal of Human-Computer Studies, 2021, 151: 102630.

边界。通过在技术、制度、用户与社会层面形成协同机制，可以在提升生成质量的同时，实现安全、可信且可持续的人机互动体系。

七、结语

生成式 AI 时代的人机关系，正以提示词为核心媒介，在持续互动中不断生成与重构。双向人机对齐并非一次性完成的技术匹配，而是在不确定性条件下，通过多轮交互不断协商与修正的动态过程[5]。在这一过程中，提示词既是意图表达与语境建构的工具，也是价值嵌入与关系调节的重要载体。

从更广阔的视角来看，生成式 AI 的治理问题，本质上是如何在不确定性中实现可控协作与关系稳定的问题。只有在提示词设计优化、模型能力提升与制度规范完善之间形成协同机制，才能在技术发展与社会需求之间建立相对稳定的平衡结构。由此，人机关系有望从功能性对齐逐步走向结构性共生，在提升效率与创造力的同时，维持必要的安全边界与价值秩序，从而推动生成式 AI 实现长期、可持续发展。

[5] TREIMAN L S, HO C, KOOL W. The Consequences of AI Training on Human Decision-Making[J]. Proceedings of the National Academy of Sciences, 2024, 121(33): e2408731121.

第八章 生成式 AI 在新闻传播领域的治理

王裕平[1, 2]

1.中国人民大学新闻学院 北京 100872

2.中国人民大学人工智能治理研究院 北京 100872

2025 年以来，生成式人工智能对新闻传播领域的影响已从“工具层面的效率提升”进入“生态层面的结构重塑”。大模型、智能体、多模态生成、AI 搜索与个性化推荐共同改变了新闻的采集、生产、分发、消费和反馈机制。新闻机构不再只是使用 AI 辅助写稿、翻译、摘要、剪辑和数据处理，而是与平台型 AI 系统、搜索引擎、内容聚合器和智能终端共同构成新的公共信息基础设施。

新闻传播业的特殊性在于，它既是 AI 技术的应用场景，也是社会事实、公共讨论和价值判断的生产场域。生成式 AI 一方面可以降低内容生产成本、提升多语种传播和数据新闻能力，帮助主流媒体扩大触达面；另一方面也可能放大虚假信息、版权争议、新闻来源错配和公众信任流失等问题。未来新闻传播治理不能只把 AI 看作“内容生成工具”，而应把它作为“新闻中介系统”和“公共信息基础设施”来治理。

一、新闻传播领域生成式 AI 应用的新形势

第一，新闻生产正在从“人机辅助”走向“流程嵌入”。在新闻机构内部，生成式 AI 已广泛进入选题策划、资料检索、录音转写、外文翻译、标题生成、稿件摘要、视频剪辑、图像生成和用户互动等环节。中央广播电视总台于 2025 年发布人工智能发展白皮书，提出构建人工智能媒体应用平台、打造“央视听媒体大模型 2.0”，反映出主流媒体正从单点工具试用转向组织级智能化基础设施建设。

第二，新闻分发入口正在从“搜索与社交平台”转向“AI 答案与智能体”。Reuters Institute《2025 数字新闻报告》显示，AI 聊天机器人和界面开始成为新闻入口，全球受访者中每周使用 AI 聊天机器人获取新闻的比例约为 7%，25 岁以下人群达到 15%；同时，受众对 AI 生成新闻仍保持谨慎，更能接受“人类主导、AI 辅助”的新闻生产方式。这意味着，AI 系统不只是把新闻“再分发”，而是在用户提问、模型检索、答案合成和来源呈现过程中，对新闻事实进行二次组织。新闻从“文章—链接—阅读”的模式，逐渐转向“问题—答案—少量引用”的模式。

第三，AI 搜索和摘要正在改变新闻机构的流量与收益结构。Pew Research Center 对 2025 年 3 月美国 Google 搜索行为的研究发现，当搜索结果页面出现 AI 摘要时，用户点击传统搜索结果链接的比例为 8%，没有 AI 摘要时为 15%；用户点击 AI 摘要中引用来源链接的比例仅为 1%。该研究还显示，约 18%的 Google 搜索生成了 AI 摘要。对新闻机构而言，这种“答案优先”的信息获取方式可能削弱网站访问量、广告曝光和订阅转化，使平台与媒体之间围绕内容使用、流量分配和收益补偿的关系更加紧张。

第四，合成内容规模化生成使新闻真实性治理更为紧迫。AI 生成文本、图片、音频和视频已经显著降低虚假信息生产成本，形成“数字泔水”、仿冒账号、伪造新闻现场、AI 主播搬运和深度伪造等新型内容风险。2025 年 10 月，欧洲广播联盟和 BBC 牵头的研究评估了 18 个国家、14 种语言中四类 AI 助手的 3000 余条新闻相关回答，发现 45%的回答至少存在一个重大问题，31%存在严重来源问题，20%存在重大准确性问题。这说明，AI 新闻风险不仅来自“恶意伪造”，也来自模型在摘要、归因、更新和语境判断中的系统性失真。

二、关键进展与典型案例

一是新闻机构与 AI 平台的授权合作加速。2025 年 2 月，OpenAI 与 Guardian Media Group 宣布内容合作，使 ChatGPT 用户可以访问 Guardian 新闻内容，并以归因和链接方式呈现。2025 年 4 月，OpenAI 与《华盛顿邮报》达成合作，ChatGPT 将在相关问题回答中展示该报报道的摘要、引用和原文链接。 这类合作表明，高质量新闻正在成为 AI 搜索、问答和实时信息服务的重要训练与检索资源。其积极意义在于改善来源质量、增加归因透明度、探索授权收益；其潜在问题在于，少数大型媒体可能更容易进入 AI 分发体系，中小媒体和地方新闻机构面临被边缘化风险。

二是版权诉讼推动新闻数据权益边界重塑。2025 年 2 月，美国法院在 Thomson Reuters 诉 Ross Intelligence 案中支持 Thomson Reuters，认定 Ross 使用 Westlaw 内容训练竞争性法律搜索工具不属于合理使用，该案被视为 AI 版权争议中的重要判例之一。2025 年 4 月，数字媒体出版商 Ziff Davis 起诉 OpenAI，指控其未经许可使用旗下 ZDNet、PCMag、CNET、IGN、Lifehacker 等内容训练 AI 系统。 2025 年 12 月，《纽约时报》又起诉 Perplexity AI，称其未经授权复制、展示和分发大量文章，并存在幻觉内容错误归因问题。 这些案件说明，新闻版权争议已从“训练阶段能否使用作品”扩展到“AI 答案是否替代新闻产品”“引用是否构成合规归因”“幻觉是否损害媒体商誉”等更复杂问题。

三是媒体内部 AI 使用标准逐步形成，但透明度不足仍突出。AP、Guardian、WIRED 等媒体均强调人类编辑责任、事实核验和 AI 使用披露。Guardian 在 2026 年 3 月更新的

生成式 AI 原则中强调，AI 只应在有助于原创新闻时用于编辑工作，并要求透明、许可、合理补偿和高级编辑批准。但实际执行仍存在缺口。2025 年一项针对美国 1500 家在线报纸、18.6 万篇文章的预印本研究发现，约 9%的新发布文章可能部分或完全由 AI 生成，且 AI 使用披露非常少；该研究虽需谨慎看待检测误差，但提示新闻业 AI 透明度规范尚未充分落地。

四是中国 AI 内容治理进入“标识+专项整治+场景治理”阶段。2025 年 3 月，国家网信办等四部门发布《人工智能生成合成内容标识办法》，自 2025 年 9 月 1 日起施行，要求规范 AI 生成合成文本、图片、音频、视频等内容标识。2025 年 4 月，中央网信办部署“清朗 • 整治 AI 技术滥用”专项行动，聚焦违规 AI 产品、AI 换脸拟声、谣言、不实信息、色情低俗内容和假冒他人等问题。 到 2026 年 4 月，中央网信办又部署“清朗 • 整治 AI 应用乱象”专项行动，将未履行大模型备案登记、安全审核能力不足、训练语料安全、AI 数据投毒、生成合成内容标识不到位、“数字泔水”和侵害未成年人权益等纳入重点整治。 对新闻传播领域而言，这意味着 AI 治理已从原则倡导转向可执行的内容标识、平台责任和应用合规。

五是国际治理对“公共利益信息”的透明义务趋严。欧盟《人工智能法》第 50 条要求，AI 生成或操纵的内容应清晰标记并可检测；部署者发布用于告知公众公共利益事项的 AI 生成或操纵文本，应披露其人工生成或操纵属性，但经过人类审查或编辑控制且有人承担编辑责任的情形可适用例外。欧盟委员会还于 2025 年 11 月启动 AI 生成内容标记和标签行为准则制定工作，以落实 AI 法透明度义务。 这一制度方向对新闻机构具有直接启示：AI 披露不能停留在笼统声明，而应区分辅助性编辑、实质性生成、深度伪造、公共利益文本和娱乐讽刺等不同情形。

三、主要影响与风险

首先，生成式 AI 提升了新闻生产效率，但也可能弱化新闻专业判断。AI 可快速处理海量材料、生成背景说明、协助跨语种传播，特别适合财经快讯、体育赛况、天气灾害、公告整理、会议纪要和数据新闻等结构化场景。但新闻价值判断、信源关系维护、现场观察、利益冲突识别和公共责任承担仍依赖记者。若机构把 AI 当成低成本替代品，容易导致新闻同质化、语义模板化、采访缺席和事实核验外包，最终损害原创报道能力。

其次，AI 答案引擎正在重塑议程设置机制。传统搜索至少向用户展示多个来源，而 AI 搜索倾向于合成单一答案。模型选择哪些来源、如何排序、是否引用新闻媒体、是否压缩异议观点，都可能影响公众对议题重要性和事实边界的理解。尤其在突发事件、公共卫生、灾害预警、选举政治和国际冲突中，AI 系统若引用过时信息或低质量来源，就可能形成新的“算法误导”。EBU/BBC 研究所揭示的大量来源错误和语境错误表明，新

闻中介型 AI 的治理重点不应只是“模型是否会编造”，还应包括“检索是否可靠、引用是否充分、答案是否保留不确定性”。

再次，版权和数据权益成为新闻业可持续发展的核心问题。新闻机构投入大量采访、编辑、核验和法律风险成本，形成可信内容资产；AI 平台则通过训练、检索和摘要将这些资产转化为用户答案。如果缺乏授权、归因和收益分配机制，新闻机构可能出现“内容越被使用、原站越无流量”的困境。授权合作可以提供一种路径，但也可能造成强者更强。未来更需要行业集体谈判、标准化授权接口、机器可读版权声明、来源收益追踪和公共利益新闻补偿机制。

第四，合成内容冲击新闻信任。深度伪造、AI 配音、虚拟主播、自动搬运号和低质 AI 号可以伪装成新闻生产主体，以低成本制造“看起来像新闻”的内容。新闻信任危机并不只来自假新闻本身，更来自公众无法判断哪些内容经过专业核验。若“AI 生成”“AI 辅助”“AI 改写”“AI 翻译”“AI 复原画面”“AI 情景演绎”之间没有清晰标识，受众可能把说明性合成误认为事实影像，也可能把真实报道误判为伪造，形成“真实性犬儒主义”。

第五，劳动结构和新闻教育面临重构。AI 会替代部分重复性编辑、编译、标题、快讯和初级资料整理工作，但也会创造 AI 事实核验、数据验证、模型审计、提示工程、交互新闻设计、信源数据库建设和算法监督报道等新岗位。新闻人才培养应从“会不会使用 AI”升级到“能不能监督 AI”，包括数据素养、统计推理、版权意识、算法审计、跨平台传播和公共伦理能力。

四、治理原则与路径建议

第一，确立“人类编辑责任不可转移”原则。新闻机构可以使用 AI，但不能把事实责任、价值判断和公共后果转交给模型。凡涉及公共利益、突发事件、重大政策、司法案件、医疗健康、金融风险、未成年人和个人名誉的内容，应坚持人工采编、人工核验、人工签发。AI 生成内容不得直接进入发布端；AI 辅助稿件应保留可追溯记录，包括使用工具、提示词范围、引用来源、人工修改和审核责任人。

第二，建立新闻机构内部 AI 分级使用制度。可将 AI 应用分为四类：低风险辅助类，如转写、检索、格式整理；中风险编辑类，如摘要、标题、翻译、资料卡片；高风险生成类，如新闻稿主体、评论草稿、采访问题和视频脚本；禁止或严格限制类，如虚构采访、伪造现场、无授权复刻声音肖像、生成未标识新闻图片。不同类别对应不同审批、披露、留痕和复核要求。AP 提出的“AI 可辅助但记者负责”的原则，可作为新闻机构制度设计的重要参照。

第三，落实 AI 生成合成内容标识制度。新闻报道中使用 AI 生成或显著修改的图片、视频、音频、图表和文字，应在用户首次接触处以清楚方式标识，不能只放在页面底部或笼统写入服务条款。对新闻图片和视频，应区分“画质修复”“示意图”“情景再现”“AI 生成”“AI 合成声音”“虚拟主播播报”等不同标签。中国《人工智能生成合成内容标识办法》和欧盟 AI 法第 50 条均体现出“显性披露+机器可读标识”的治理方向。

第四，强化 AI 新闻中介平台的来源责任。AI 搜索、聊天机器人、智能体和内容聚合平台在提供新闻答案时，应承担高于普通工具的公共信息责任：优先引用权威、原始和多元来源；清晰呈现出处、时间和不确定性；对突发事件设置更新提示；对公共安全、选举、医疗和金融类问题采取更严格的拒答、转介和人工审核机制；不得以 AI 摘要替代新闻产品而规避授权和补偿。平台还应开放外部审计接口，让研究机构评估其来源质量、引用偏差、幻觉率和对新闻网站流量的影响。

第五，推动版权授权和收益分配机制制度化。新闻机构与 AI 平台的合作不应只依赖个别商业谈判。建议建立新闻内容机器读取标准、授权登记平台、训练数据使用说明、引用收益追踪和集体授权机制。对公益性、地方性、调查性新闻，可探索公共基金、平台反哺和数据使用费制度，避免 AI 平台只吸收新闻内容价值而不承担新闻生产成本。

第六，建设新闻真实性技术基础设施。C2PA 内容凭证、数字水印、来源签名、媒体资产指纹库和可信时间戳可以提高内容溯源能力。Cloudflare 于 2025 年推出一键式 Content Credentials 方案，帮助出版者和媒体机构在图片传播中保留数字历史记录，这显示内容溯源正在从专业工具走向基础设施。但技术标识并非万能，元数据可能丢失，真实图片也可能被断章取义。因此，技术治理必须与人工核验、现场采访、语境说明和平台处置机制结合。

第七，提升公众 AI 新闻素养。治理不能只面向机构和平台，也应面向受众。公众需要理解：AI 答案不是新闻原文，AI 引用不等于事实核验，AI 图片不等于现场证据，AI 摘要可能遗漏关键背景。学校、媒体和平台应共同开展 AI 新闻素养教育，引导用户查看原始来源、比较多方报道、识别合成标识、警惕情绪化和高度个性化的信息推送。

五、未来展望

展望未来，生成式 AI 对新闻传播的影响将呈现三大趋势。其一，新闻产品将从“固定稿件”转向“动态交互”。用户可能通过智能体要求“解释这条新闻对我所在城市的影响”“用两分钟视频说明事件背景”“比较三家媒体观点差异”。新闻机构需要把报道、数据库、时间线、图表和问答系统整合为可交互知识服务。其二，新闻竞争将从“发布速度”转向“可信知识资产”。当 AI 可以快速改写和摘要公开信息，独家采访、现场核验、

专业数据库、权威解释和稳定声誉将更具价值。其三，治理将从“内容标识”走向“过程问责”。未来的关键问题不是简单标注“由 AI 生成”，而是说明 AI 在何处介入、依据哪些来源、由谁审核、错误如何纠正、权益如何补偿。

总体而言，生成式人工智能不是新闻业的外部变量，而正在成为新闻传播生态的内生结构。新闻机构应主动建设可信数据、可信流程和可信品牌；平台企业应承担新闻中介责任；监管部门应坚持包容审慎、分类分级和场景治理；公众也应提升对 AI 信息环境的辨识能力。只有在技术创新、专业伦理、版权秩序和公共责任之间形成新的平衡，生成式 AI 才能真正服务于高质量新闻传播和健康有序的公共舆论生态。