

人工智能驱动的代理式系统与知识生产： 转型、重塑与挑战

李秋甫

摘要：自 ChatGPT 发布以来，生成式人工智能逐步嵌入文献检索、学术写作和数据分析等知识生产环节。与以对话回答为中心的问答式模型不同，代理式系统能够围绕给定目标自主拆解任务，并在检索资料和调用工具的过程中持续调整行动，形成从判断到结果生成的连续流程。代理式系统的技术逻辑区别于问答式模型，主要表现为自主规划、闭环推理和跨域整合等特征，并由此重塑知识生产的主体位置、过程结构和组织形态。代理式系统在提高知识生产效率的同时，也带来特定风险挑战：可计算对象被优先处理，从而压缩依赖默会经验与历史语境的知识形态；系统以多步骤自动执行介入知识生产过程，使得责任归属更加困难；既有科研评价以成果文本为中心，难以覆盖代理式系统介入过程中的证据筛选与人工复核。未来则应推动知识生产的过程性治理，以过程披露与责任锚定为规范底线，缓解因系统可及性而产生的新型不平等，重建人工智能时代效率红利与知识质量的平衡。

关键词：人工智能；代理式系统；知识生产；AI4S；科技治理

中图分类号：G302 **文献标识码：**A **文章编号：**1673-0186(2026)07-0111-15

自 2022 年 ChatGPT 发布以来，大语言模型迅速进入教育教学、学术研究、媒体传播和专业服务等领域，这类模型主要采用“提问—回答”的交互方式，用户提出问题，模型生成文本，研究者再对输出内容进行筛选、修改和采纳。近两年来，以 OpenAI 公司的 Deep Research、谷歌公司的 Gemini Deep Research 等为代表的“深度研究”（Deep Research）系列产品加速演进，Claude Code 将代理能力延伸至本地开发环境，可读取、修改文件，自动执行终端命令完成项目迭代。以 OpenClaw 为代表的开源代理框架则可在用户设备上持续运行，并连接个人化消息平台、文件系统与网络资源。尽管功能场景有所区分，这些

技术应用均不再局限于生成文本，而是围绕既定目标组织操作步骤并自主推进研究任务，推动人工智能从单一的研究辅助转向对知识生产全过程的持续介入。

与此同时，既有的知识生产格局正经历前所未有的重塑，引起了学术界与产业界的广泛关注。但现有讨论仍存在两方面不足：一是人工智能在知识生产的深层环节所呈现的结构性影响仍有待系统阐明；二是目前多以 ChatGPT 等为代表的问答式模型作为分析对象并将其一概而论，对人工智能驱动的代理式系统的特殊技术特征，以及这些特征产生的影响尚缺乏深入分析。基于此，本文旨在梳理代理式系统的关键特征，

基金项目：北京高校思想政治工作研究课题“网络碎片化背景下大学生思想热点分析研判与回应策略”（BJSZ2025ZC05）。

作者简介：李秋甫，清华大学党委学生工作部讲师、法学博士，研究方向：科学技术与社会、人工智能伦理、青年发展与数字社会。

并据此探讨在该技术语境下知识生产方式的变革,以及这种变革对知识认识论、科技伦理风险与科研组织形态等方面带来的复合性挑战。

一、知识生产中的问答式模型与代理式系统

考察代理式系统对知识生产的影响,需要先对人工智能嵌入知识生产的历史脉络进行梳理。随着大语言模型能力边界不断扩展,人工智能在知识生产中的技术形态正由以即时响应和内容生成为特征的问答式模型,转向以目标驱动和自主执行为特征的代理式系统。

(一)从工具辅助到代理跃迁:人工智能嵌入知识生产的脉络演进

人工智能对知识生产的影响并非始自今日,从历史演进来看,其嵌入过程经历了从外围工具到核心流程的渐进拓展,大致可划分为三个阶段。

第一阶段为“工具辅助”阶段,时间范围上涵盖20世纪中叶至20世纪90年代。人工智能和相关计算工具主要以特定功能模块的形式介入知识生产。加菲尔德(Garfield)提出引文索引思想^[1],其后创立的科学引文索引为追踪学术脉络提供了系统手段;斯莫尔(Small)于1973年提出的共被引分析方法进一步开创了通过引文关联识别研究前沿的范式。在知识推理层面,以用于化学结构推断的DENDRAL、用于医学诊断的MYCIN等为代表的专家系统尝试将领域知识编码为规则库,辅助科学假设的生成与检验。此时,人工智能处于知识生产链条的外围,承担信息检索、数据处理与辅助推理的功能,尚未触及知识生成的核心环节。

第二阶段为“数字基础设施”阶段,主要发生在21世纪初至21世纪10年代末。随着互联网技术的普及,在线学术数据库(如Web of Science、Scopus等)逐步成为研究者获取文献的主要渠道,知识生产的信息环境从物理图书馆转向数字平台。与此同时,文献计量学进入新发展阶段,替代计量学等新型评价指标将社交媒

体、政策文件等非传统数据源纳入学术影响力评估^[2]。机器学习技术也开始应用于大规模文献挖掘,如潜在狄利克雷分配等方法被用于自动识别研究前沿与知识结构。人工智能开始从单一工具演变为支撑知识生产的数字基础设施,但仍以被动响应为主,科学研究的问题界定与意义诠释由人类研究者所主导。

第三阶段是“生成式问答”阶段。BERT、GPT等预训练语言模型的发展显著提升了机器处理自然语言的能力,ChatGPT的发布则使生成式人工智能从专业实验环境进入大规模公共使用。此后,人工智能开始广泛参与教学材料生成、新闻写作、法律检索、学术研究等任务。具体而言,在学术研究领域,Elicit利用语言模型帮助研究者检索论文、针对文献提问并概括研究发现;在教学领域,可汗学院(Khan Academy)于2023年推出基于GPT-4的AI导师Khanmigo,为教师提供课程设计、测验生成和学习评估支持。不过,早期以ChatGPT为代表的模型主要以即时问答模式进行,模型根据用户输入生成响应,研究者对输出内容进行筛选、修改与采纳,人机交互以单一回合的信息交换为主要形式。

(二)从问答到代理:技术跃迁与研究缺位

生成式人工智能的兴起进一步推动了知识生产研究的议题转向,既有讨论更多置于知识生产过程内部,关注机器在知识生产链条中的地位变化,并普遍认为人机融合正在成为新的结构性特征^[3-5]。也有学者从人工智能的技术属性与社会嵌入机制出发,强调“人机共生”所带来的知识生产结构重构及其方法论意义^[6]。与此同时,围绕“生成的知识究竟是什么知识”的讨论逐渐升温,研究提出“三阶知识生产”等概念框架以界定生成式人工智能在知识生成、加工与再表达中的认识论位置^[7],并进一步引入“软”知识、“灰”知识、“暗”知识等分析范畴^[8],刻画生成式人工智能所产出知识在可证伪性、可解释性与可迁移性方面的复杂属性。在此基础上,相关讨论进一步延伸至制度与治理层面,集中讨论知识产权与版

权边界、验证机制与可信性约束、隐私风险与数据安全、误导性幻觉以及对人类主体地位与学术规范的冲击等问题^[9-12]。然而,多数讨论以早期ChatGPT版本等问答式模型作为分析对象,将人机交互理解为“提问—回答”的即时模式,这一预设在面对代理式系统时已明显不足。

代理式系统代表了人工智能嵌入知识生产的全新技术形态。与生成式问答模型相比,代理式系统具备自主规划、持续执行与动态工具调用等关键技术跃迁,从“信息提供者”转向研究过程的“参与者”。现有讨论之所以难以充分把握代理式系统的影响,原因主要在于分析对象和分析层次的错位。首先,既有讨论多关注模型生成内容,而较少追问模型如何改变知识形成的内在过程;其次,不少研究仍以静态评估框架衡量模型能力及探讨其影响,难以捕捉代理式系统在多轮迭代中的动态特征;最后,现有规则通常将人工智能视为辅助工具,并要求人类作者承担最终责任,这一制度要求虽然可以回应文本归责问题,但面对代理式系统对证据筛选、推理路径与研究决策的介入则缺乏有效的应对手段。

上述三重局限共同指向一个核心问题:代理式系统对知识生产的影响不能仅仅在输出层面上被理解,而必须深入其作用机理,即它如何嵌入并重塑知识生产的主体位置、互动机制与制度安排。本文以人工智能驱动下面向知识生产任务的代理式系统作为研究对象,在分析视角上以技术逻辑、维度重塑、影响机制、挑战归纳四个层面展开,在方法论取向上,本文构建以代理自主性提升、闭环推理与跨域整合三大技术特征为基础的分析框架,以期回答以下核心问题:代理式系统凭借何种技术逻辑嵌入知识生产?又通过何种机制重塑知识生产的主体、结构与组织形态?这些重塑又带来哪些深层挑战?

二、代理式系统:技术逻辑与关键特征

本文将“代理式系统”界定为能够围绕既定研究目标自主规划多步骤行动、动态调用外部工

具并在执行中持续调整策略的智能系统。其技术特征可从认知、行动与系统三个层面把握。认知层面,系统依托思维链(CoT)机制进行多步自主推理,通过反思机制修正策略,并以记忆更新实现持续学习;行动层面,系统运用检索增强生成(RAG)等技术将外部知识源接入推理流程,拆解目标为子任务并衔接各类工具,在多轮交互中保持状态与记忆;系统层面,通过闭环响应外部变化,并依托多智能体协作框架实现角色分配与任务委派。事实上,随着大语言模型产品的持续迭代,联网搜索、插件调用、文件分析等能力已逐渐成为通用模型的重要组成部分,但单项能力是否被嵌入多步骤闭环的流程之中成为传统问答式模型与代理式系统的关键差异。

需要进一步说明的是,“代理式系统”与“智能体”(AI Agent)二者在既有文献讨论中常作近义使用,后者是人工智能研究的经典概念^[13],但在生成式人工智能的技术突破影响下,当前“智能体”的实现形态也从基于规则的封闭域系统转向以语言模型为核心的通用任务单元;与之不同,代理式系统作为显性的系统架构概念,其兴起与大语言模型应用爆发基本同步,本身具有围绕高层目标进行多步分解、动态协调、工具调用与自适应执行的能力。当前,代理式系统已形成多种产品形态,表1列举并比较了其中的代表性类型。

(一)代理自主性提升:从辅助到共作

代理式系统的首要特征在于自主性的提升。传统问答式模型主要遵循“指令—响应”的运行逻辑,根据用户输入直接生成回答。其知识来源主要为预训练语料,系统并不会围绕任务持续规划、追踪和调整行动过程,其行动边界基本受限于用户输入所划定的任务范围。而代理式系统进一步在目标层面实现自主,用户提供的往往不再是逐条规定的具体动作,而是相对宽泛的研究目标。系统据此可以自主生成中间步骤,并在任务推进中自动决定流程顺序、信息来源、工具调用和终止时机,从而完整地嵌入知识生产中的

表 1 传统问答式模型与代理式系统的多维度比较

比较维度	传统问答式大模型 (以 GPT-3.5 为参照)	深度研究型(Deep Research / Gemini Deep Research)	代码执行型(Claude Code)	本地框架型 (OpenClaw)
目标拆解 与规划方式	围绕用户当前输入生成回答	围绕研究目标自主拆解任务、规划检索与验证路径	围绕开发目标自主规划代码修改、调试与测试流程	围绕用户日常目标持续运行,自主调度本地与网络资源
检索验证 与推理机制	无内置检索,或由用户另行提供检索结果;检索结果不纳入自主推理闭环	自主决策检索时机,多源交叉验证,将检索结果纳入反馈回路	自主读取代码库、捕获错误输出并回溯修正,纳入迭代调试回路	调用本地材料作为分析依据,通过网络检索核实相关信息,并根据核验结果调整后续步骤
工具调用 与执行结构	通常无工具接入,回答生成后终止	动态衔接浏览器、文献数据库与数据分析工具,形成多轮闭环	动态衔接终端命令、代码执行环境与版本控制系统,形成端到端闭环	通过自托管网关连接消息渠道与用户设备中的工具

实践流程。

面对一项相对完整的研究任务,代理式系统可以自主分解目标、规划步骤、配置工具,并依据中间结果调整后续行动。例如,“深度研究”功能可以规划检索路径、筛选和整合证据,形成带有引用来源的结构化报告;编码代理将这种能力延伸至代码与终端环境,通过自主运行与迭代代码,衔接资料获取、分析处理与成果输出;本地代理框架则连接个人的文件系统与网络资源,在用户的实际工作环境中持续运行,保留任务状态和阶段性成果。代理式系统由此介入任务推进与边界划定,承接部分原本由研究者逐步完成的过程性职能,推动机器由单项辅助工具转变为知识生产中的“共作单元”。

(二)闭环流程演进:从生成到推理

代理式系统的第二个特征在于闭环推理。传统问答式模型通常根据用户当前输入直接生成回答;检索只是独置的外部事件,结果也多作为外部信息补充,不持续参与后续的任务规划与执行。代理式系统则形成“推理—行动—评估”的开放循环,系统为每一步输出维护中间状态,评估当前信息是否支撑该步结论,不足则回退检索,发现来源冲突则启动交叉验证,直至该节点通过内部评估。部分系统还设置了目标

校验功能,用户设定最终目标后,系统自动调适推理路径并在达到阈值前持续迭代。在此流程中,检索获取的信息不再是外在独立的附加物,而是推理链条的有机组成部分。这种链式架构赋予系统自我审查的能力,也在输出可靠性方面产生与传统问答式模型的分野,集中体现在对“幻觉”这一风险的应对方式上。

“幻觉”现象一直是大模型应用场景的典型风险,一项针对医学领域引文准确性的研究显示,即便是 GPT-4 这类先进模型,也会生成捏造的论文引用,直接影响学术内容的可靠性^[14],这类问题根源于训练数据质量、模型结构设计、应用层任务设置等多个方面的短板^[15]。然而,代理式系统依靠闭环推理的特性,形成自我校正的反馈回路,进而评估多元信息间的关联性,经由逻辑推演得出基于上下文与证据链的结论,有助于降低传统大模型“单点幻觉”的发生概率,Yao S 等人的实验表明,在抽样分析的问答结果中,采用以“ReAct”为代表的推理与行动交替执行的系统框架后,出现虚假推理或事实错误的比例为 6%,低于单独使用思维链方法时的 14%^[16]。闭环流程中自我审查与动态修正的能力成为输出可靠性提升的重要保障,进而实现代理式系统从单向的文本生成向动态的回溯推

理的跃迁。

(三) 跨域整合与创新生成:从汇聚到新知

代理式系统的第三个关键特征在于跨域整合与创新生成。传统大模型通过海量训练语料“习得”多领域通识知识,其跨域能力实质上是参数空间中的静态信息叠加,不同领域的知识在模型内部并行存储。而代理式系统的跨域整合能力以工具链的动态重组为基础,通过模型上下文协议(MCP)^①等标准化接口,具备跨越平台异质性边界的能力。过去,创新主要依赖研究者在不同学术共同体之间积累概念、方法与证据。而今,代理式系统将公开检索与私域材料纳入同一研究流程,在更大尺度上同步处理多领域信息。2026年1月,谷歌推出“个人智能”功能,允许用户将邮件、相册、YouTube 观看历史及搜索记录接入代理式系统,系统自主决定公域知识与私域材料的调取顺序与内容,并在二者之间进行交叉对比与关联分析。深度研究产品的工作流程可将公开文献中的文本与数据提取为结构化素材,继而将其转化为统计规律与趋势判断,再将这些抽象结论形成可视化图表,文本、数据、图表、引用等不同形态的信息被纳入同一动态链条,由系统自主决定其转化路径与重组方式。

无论是公域知识与私域信息的整合,还是不同工具链之间的跨域调用,其共同指向都在于信息形态的动态重组。不同来源、格式各异、分属不同领域的信息,在系统规划中被转化为可统一处理的表征,进而纳入持续推进的工作流程。由此,代理式系统所推动的跨域整合,正在从简单的“信息加总”延伸为信息形态的多重组合。若从熊彼特“创新即新组合”的视角看^②,代理式系统将不同领域的研究方法、数据操作与验证步骤纳入统一流程并连续展开,从而拓展了创新发生的技术条件。进一步说,跨域整合的演进使知识创新在一定程度上摆脱了对研究者个体知识储备与跨学科灵感的单一依赖,逐渐转向一种由系统依据任务目标整合跨领域知识资

源的程序化路径。

三、代理式系统重塑知识生产:主体重构、结构交互与边界重组

本文将“知识生产”理解为在特定制度与组织条件下,通过问题界定、方法选择、数据与仪器操作、同行评议与发表传播等环节,把某类陈述稳定化为可被共同体承认的知识的全过程。在这一理解框架中,目前代理式系统的产出在多数情境下还需经人类主导的检验与辩护机制加以确认,但近年来,代理式系统在特定科学领域的自主产出已开始获得学术共同体的“事实承认”,人工智能在知识生产网络中的结构位置与功能角色也随之发生实质性变化,重塑知识生产全过程。

围绕知识生产研究传统中反复出现的核心关切,可以从三个相互关联的维度审视代理式系统的影响。第一,知识由谁生产。默顿(Robert K. Merton)关于科学体制的分析表明,科学家作为知识生产者并非自然出现,而是在特定制度规范中获得合法地位^[17];拉图尔(Bruno Latour)则进一步指出,仪器、文本、样本和技术装置同样参与科学事实的形成^[18]。第二,知识如何生成。塞蒂娜(Karin D. Knorr-Cetina)和皮克林(Andrew Pickering)等人的研究说明,知识并非单纯观念推演的结果,也在研究者对仪器、数据和材料的持续判断与调适中形成^[19-20]。第三,知识生产如

①模型上下文协议是由 Anthropic 于 2024 年底提出的开放协议,旨在为大语言模型与外部工具、数据源之间建立统一的通信标准,使代理式系统能够动态发现、调用异质性工具,而无需为每个平台单独编写适配代码。

②熊彼特以“新组合”解释创新活动及其推动的经济发展,并将其概括为新产品、新生产方法、新市场、新供应来源及新的产业组织等形式。参见 SCHUMPETER J A. The theory of economic development: an inquiry into profits, capital, credit, interest, and the business cycle[M]. OPIE R, trans. Cambridge, MA: Harvard University Press, 1934: 65-66。

何被组织。吉本斯(Michael Gibbons)等人关于“模式2(Mode 2)”的讨论^[21]以及“三重螺旋模型”^[22]等均强调,当代知识生产越来越依赖跨学科、跨组织和多主体协作。由此,行动者构成、过程机制与组织形态构成了审视代理式系统影响的三个基本切口。

代理式系统的介入,正是在这三个层面同时触及知识生产的既有秩序。就行动者构成而言,知识生产正在呈现出更具协同性的“人机同位”形态;就过程机制而言,模型、数据与工具之间的结构性交互,正在由静态调用转向动态协同;就组织形态而言,平台化工作流不断贯通并重组组织边界与资源边界,使知识生产能力越来越体现为跨域整合与跨主体协作的制度化能力。

(一)代理式系统与知识生产的主体重构

代理式系统对知识生产最直接的影响,体现在知识生产网络中行动者构成的变化。传统知识生产强调科学家为主体的中心地位,尽管在其演变发展中也曾涉及工匠、自然哲学家等多元参与者,但总体而言,知识的创造多集中于学术共同体内部。以近代科学革命为界,科学家与传统从事知识生产任务的群体区分开来,科学家利用科学仪器发现科学原理、实现科学突破,成为知识发现与知识生产的主要机制。但代理式系统的出现,使得机器开始代替原先由研究助理或数据分析人员承担的职责。虽然人类研究者目前仍负责主要环节的设定与把关,代理式系统已在部分环节连续执行任务,使知识生产的推进方式呈现出更明显的“人机同位”特征。

这种特征首先表现为通过检索、整合与重组多源学术信息,代理式系统已能够介入研究假设生成、文献综述撰写和数据分析等环节,并在知识生产网络中成为更具能动性的“非人类行动者”^①。加州大学伯克利分校领导开发的自主实验室(Autonomous Lab),在17天的连续闭环运行中共执行353次实验,从57种目标材料中成功合成36种化合物。这种演变也并不限于自

然科学领域,SCALE研究团队提出的多智能体框架将社会科学内容分析组织为文本编码、协商讨论和编码迭代等环节;多个代理先独立编码,再通过多轮对话协调分歧,并允许人类专家介入修订分析规则^[23],在面向制度实践的专业知识领域,2026年5月,Anthropic公司推出面向法律服务的Claude for Legal系统,该系统包含12个法律实践插件、20余个基于模型上下文协议的标准化接口,可接入Harvey法律人工智能平台、汤森路透万律法律数据库等专业服务,并覆盖合同起草、判例检索与诉讼支持等法律实践环节。法律知识的生产既发生在学院场所的理论讨论中,也发生在判例检索、合同审查、诉讼策略等法律实务环节之中。这些活动虽然直接服务于实践,但其本质也在生产一种可被共同体采纳与承认的专业知识。传统上,法律知识的权威性来源于人类对法律条文的理解、对判例传统的把握以及对特定情境的判断,而代理式系统通过这些标准化接口检索和调用专业数据库中的资料,使法律知识的生成过程嵌入新的技术中介。

随着人工智能的深度嵌入,代理式系统在知识生产中的角色正在从“辅助工具”转向可参与研究推进的“协同主体”,并在传统大模型提升信息处理效率的基础上,推动科学知识生产的主体重构。在这一意义上,代理式系统正成为科学共同体中的“准成员”。

(二)代理式系统与数据关系的结构交互

如果说主体重构回答的是代理式系统在知识生产网络中的位置问题,那么结构交互则关切知识生成过程本身的机制变化。代理式系统并非仅仅作为新的行动者加入知识生产,而是改变了知识得以从数据中生成的过程结构,数据、工具与模型之间的动态交互方式正在重组知识从

^①非人类行动者(non-human actors)是行动者网络理论(ANT)中的核心概念,指在科学事实建构过程中具有能动作用(agency)的实体,不仅包括人类,也涵盖仪器、文本、样本、技术装置等。

经验材料向科学陈述的转换。

数据生成、传递和交互方式的不断演进已成为当代创新的重要源泉^[24],也成为知识生产的主要动力。超大科学装置与传感网络使得前所未有的数据规模与即时处理成为可能,数据类型也从单一结构化数据演变为多模态数据,数据利用从一次性取样转向连续检测与时空分析。代理式系统以大规模预训练所形成的表征能力为基础,能够在跨模态、跨来源的数据空间中识别稳定的统计规律、隐含结构与异常模式,在海量信息中持续调整搜索与分析路径,从而把“潜在关联”转化为具备价值基础的研究线索。

从机制上看,代理式系统与数据关系的结构交互,首先表现为对不同来源信息的连续组织,系统把文献检索、信息提取、假设生成以及材料校验等步骤衔接在同一流程中。传统“基于文献的发现”方法强调从彼此隔离的文献集合中自动化挖掘隐含关联,通过揭示不同领域之间未被明确指出的联系来生成新的假设。代理式系统的辅助不仅能从文献中提取既有规律,还能将文献线索与数据驱动的统计证据结合,将隐含关联转化为可检验的实验假设。由澳大利亚莫纳什大学研究团队推出的“面向科学发现的大语言模型框架”(LLM4SD)将整合环节与推理环节相结合,实现了分子科学家通过海量文献提取信息并做出实验假设的科研流程^[25]。相似的交互流程也出现在质性研究中,“人工智能共同民族志研究者”(AI Co-Ethnographer)将开发编码、编码整合、编码应用和模式发现串联为可迭代流程,使编码表能够回到原始语料中反复校验^[26]。

进一步地,这种结构性交互还表现为更强的跨模态对齐能力,即把原本属于不同空间表征的数据纳入同一生成与推理框架,从而扩展机器可触及的知识边界。“生物医学数据分析代理”(BioMedAgent)通过多代理协作调用生物信息学工具,将跨组学数据与病理图像纳入同一分

析流程^[27];“材料模拟与电子结构理论推理代理”(MASTER)则将自然语言提出的研究任务转化为密度泛函理论计算流程,并依据模拟结果推进催化剂探索^[28]。

代理式系统依靠数据广度、深度及数据量的增长,以全新的维度提升科学洞见的获取方式。一方面表现为从海量异质信息中识别潜在关联,并将相关线索转化为进一步检验的研究设计;另一方面则表现为跨模态整合,使原本分离的数据形态能够借助系统进入同一分析流程。从快速综述文献、生成假说,到发现新分子和新材料、揭示复杂系统规律,科学知识生产过程在代理式系统介入下,进一步实现数据与工具的深度协同驱动。

(三)代理式系统与生产边界的形态重组

代理式系统对知识生产的影响不止于主体结构与过程机制,还重塑了知识生产的组织边界与制度形态,将知识生产的边界从以单一实验室为中心的封闭单元,推向以平台为枢纽的跨域协同系统。科学仪器的引入和科学基础设施是“实验型”科学范式确立的关键,而代理式系统就其技术属性而言,正成为新的知识仪器,原本分散在数据库、软件工具和实验环节中的能力,由此被重新组织为模型、工具、数据三者嵌合的平台结构。随之,知识生产活动越来越呈现出标准化、流程性的工程化特点,尤其在多工具协同方面,基于知识图谱驱动的科学智能体如科学工具代理系统(SciToolAgent)能够根据任务需要选择并调用数百种科学工具,使原本相互分离的操作环节形成连续的研究流程^[30]。

类似地,谷歌公司研发的“AI 协同科学家”系统以 Gemini 2.0 为基础,整合文献综述、推理规划等能力,并通过生成、反思、进化等代理模块的协作,将科学家与人工智能系统纳入同一研究流程。科学家以自然语言提出研究目标后,人工智能系统据此配置研究计划,并在迭代过程中生成研究假设;科学家进一步提供初始想法和互动反馈,共同推动从假设提出到实验设计的

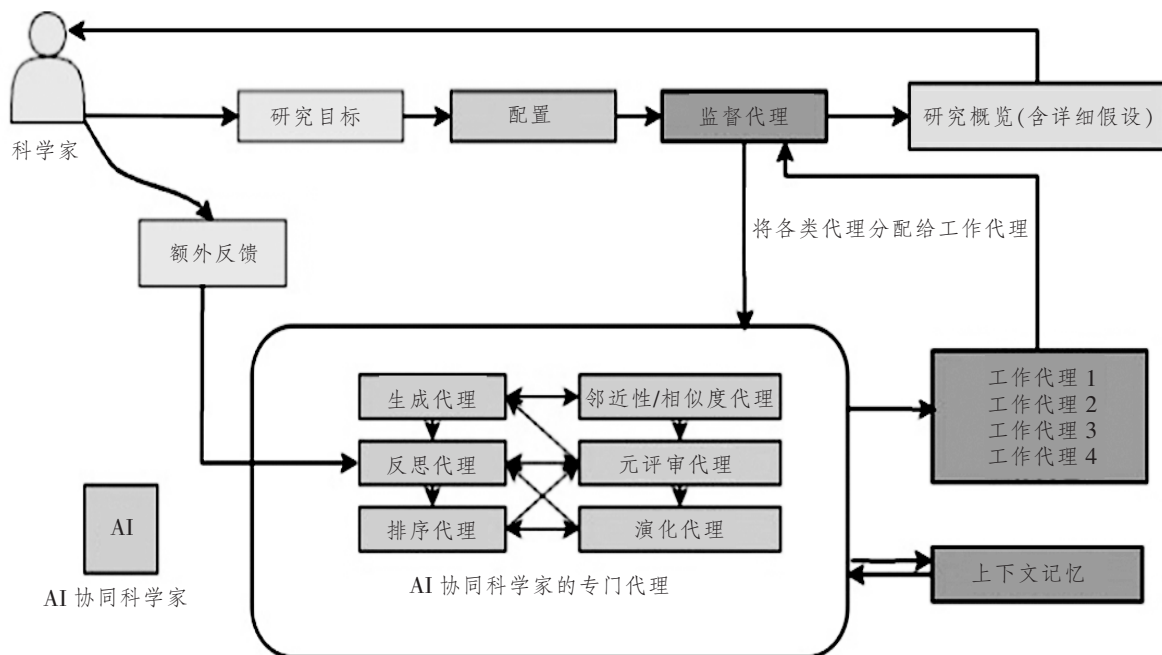


图1 谷歌公司的“AI 协同科学家”系统

流程展开(图1)。这类平台不仅改变了单一研究者的工作方式,也在重塑知识生产的组织边界。不同专业背景的研究者、不同类型的智能代理,以及异质性的数据源和工具,被接入同一平台结构之中,使知识生产逐渐从封闭的个人或实验室活动,转向开放、协作和可复用的平台化运行。

这种平台化重构不仅提升知识生产效率,还对知识生产所依赖的组织情境进行重塑。传统上,高效知识生产高度依赖研究者在特定组织文化中长期浸润所获得的流程规范与“默会知识”,科学事实的生成与稳定往往植根于实验室内部的日常实践、沟通网络与复现能力之中^[30-31]。而随着数据、算法、算力成为开展科学研究越来越重要的要素,知识生产过程中的默会经验通过算法不断实现转译,将大量具有个体性经验的科学知识抽象化、模型化,将特定组织情境下的专用知识与公开知识相结合,个人与组织的知识数据不再以相互隔离的方式分别调用,而是在代理式系统下形成更聚焦组织语境的研究成果,显著压缩从问题提出到证

据成链、从信息检索到结论成型的时滞。科学组织方式的传统边界正在消解,科研组织形态正呈现平台化趋势。

代理式系统推动的组织平台重构,还呈现出集中式与分布式之间的张力与耦合。自驱动实验室(Self-driving lab)与智能工厂等新型知识生产体系表明,知识生产正从单一实验室对设备、数据及人力的固定配置,转向分布式探索与集中化设施相互嵌合的路径:研究团队在多点试错和迭代的同时,可将实验 workflow、工具及数据资产接入共享设施,获得标准化服务,进而在跨团队、跨场景中复用和扩散。如此一来,科研组织的平台化特征愈发显著,成为新的边界枢纽。既吸纳分布式探索成果,又通过集中化设施及统一标准,将分散的创新聚合为相对固定的流程与能力。这一趋势在跨机构、跨区域实践中尤为明显。欧盟开放科学云(EOSC)以联邦化方式汇聚数据、工具和服务,通过跨欧洲的节点网络为跨学科、跨国界协作研究提供数据管理与数字服务支撑;生命科学领域的欧洲生命科学数据基础设施(ELIXIR)与艺术与人文领域的欧洲数字人文研

究基础设施(DARIAH)同样以类似方式组织数据和服务,在多节点、异构机构之间通过统一标准协同运行并实现能力互补。

综上,以检索增强、思维链推理等为支撑的技术进步构成代理式系统嵌入知识生产的关键逻辑(图2)。在此基础上,代理式系统高度的自主规划运行以及强大的跨域整合能力一方面重塑生产主体结构,从以人为主导的工具调用转向机器自主性提升下的“人机同位”;另一方面重构生产要素的组织方式,将不同时空与结构维度下的知识信息重新整合为工程化研究流程,并进一步

推动生产边界的贯通,公开知识与专用知识的界限,逐渐消弭、分布式探索与集中化设施互嵌成为新的知识生产组织形态。这种重构并不限于实验科学,也在质性研究、法律知识生产和数字人文基础设施中展开。总之,代理式系统在主体位置、过程机制与组织形态等维度上改变着知识生产的核心要素与关键环节。

四、代理式系统介入下的知识生产挑战:知识同质、伦理风险与制度重构

代理式系统对知识生产的深度介入,使其

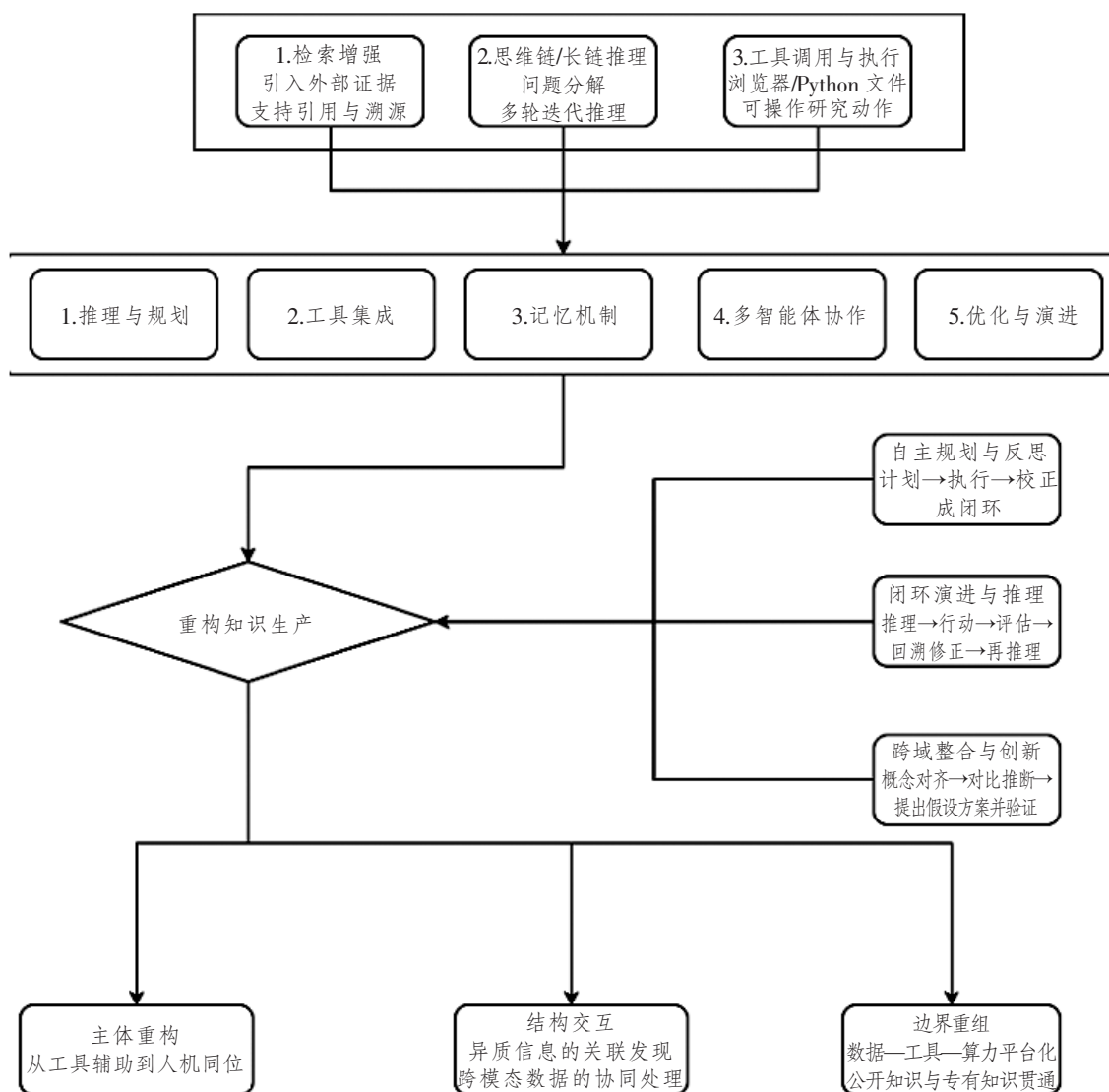


图2 代理式系统重塑知识生产

风险不再局限于输出偏误,而是沿任务导向的自主链条扩散至证据筛选与研究决策过程。由此带来的知识同质、责任失锚与制度失衡等挑战对知识生产秩序产生深层影响。

(一)数据主义导向的认识论收缩与知识同质化

代理式系统的闭环执行能力能够围绕研究目标自主规划步骤,持续检索与处理信息并调用工具完成实验验证,从而形成从假设提出到结果输出的完整研究闭环。这一技术逻辑内在地偏好可量化计算的知识形态,在知识生产中产生系统性的选择倾向,造成认识论收缩的潜在风险。赖尔(Gilbert Ryle)关于“知道什么”(Knowing that)与“知道如何”(Knowing how)的经典区分^[32]以及波兰尼(Michael Polanyi)对“默会知识”的分析^[33]为本文理解不同知识形态提供了启示。在本文的分析框架中,知识大致可分为以事实性、数据性、可验证为特征的陈述性知识,以及以判断性、诠释性、情境依赖为特征的规范性知识。代理式系统对陈述性知识的深度介入,以相关知识能够转化为可计算对象为条件。但当研究对象转向难以量化的规范性知识,如社会学质性研究中对田野资料深层意义的把握、人文学科文本诠释所依赖的历史敏感性与文化洞察力、研究者依托默会知识作出的情境判断等,则无法被纳入代理式系统的标准化步骤中。更关键的是,代理式系统的工程化逻辑在绕过这些不可量化环节的同时,也系统性地压缩了反思、协商与价值审视的空间。

此外,科学史上的大量重大突破也昭示着,知识创新并不总是沿着预设目标线性推进,许多重要发现来自研究者对异常、偏离和偶然线索的敏感把握,如弗莱明(Alexander Fleming)“偶然”发现青霉素、彭齐亚斯(Arno Penzias)与威尔逊(Robert Wilson)“无意”发现宇宙微波背景辐射等,科学社会学使用“意外发现”(Serendipity)概念来描述这种在意图之外发现有价值事物的现象。代理式系统的介入改变了这一探索生态,其

闭环执行结构内在地趋向于将复杂的研究任务分解为可顺序执行的步骤,并始终朝着预设目标的最优解前进,这种结构虽有助于提高产出效率,却可能同步削弱对非预期现象的开放性。研究者对代理式系统输出的依赖则进一步强化了此趋势,当系统提供了看似完整且逻辑自洽的研究报告时,研究者的批判性审视能力也可能在不知不觉中被弱化。

知识长期以来被视为一种通过对话协商而获得的“公共论证”的产物,科学研究通过论文、研讨会、同行评议等机制形成“对话共同体”,研究者在持续的沟通中检验和完善知识,因而知识生产也是认知多元化作用下的一类社会过程,这种认知多元化的核心价值在于降低群体在同一范式、同一视角中过早达成一致的风险。而代理式系统作为当下科学知识生产环节高度依赖的方法与工具,在供给端呈现出显著的集中化趋势并由此进一步加剧认知窄化风险。由于前端基础模型的开发需要高额投入和较强的技术能力,代理式系统所依托的基础架构主要由少数大型公司或实验室提供,其他研究者则对其进行微调、集成或直接应用。当多数人更多以调用、适配等方式参与知识生产,研究议题由于模型约束而出现同质化,从而更易产生研究视角与方法选择的趋同,根本上导致科学知识的同质化趋势。因此,从某种程度上说,代理式系统正在创造一种全新的科学文化,其突出表现为计算在很大程度上表征一切,这种科学文化也同样在创造一种新形式的科学知识,并以数据化的知识生产为核心,形成一种数据主义的知识观。数据主义的知识观倾向于围绕共同主题、重复事实或者既定理论展开,研究范式上竞争性理论与方法的空间越来越小、知识内容变得更加同质,成为代理式系统时代的认识论核心挑战。

(二)可验证性缺失的责任失锚与伦理风险

代理式系统的诞生同样使科技伦理问题呈现出新的形态。如果说以往大数据时代下的科

技伦理风险更多围绕数据被如何采集、共享与再利用展开,那么代理式系统在跨越公私数据边界、扩展行动能力的同时,也把科技伦理议题置于系统性层面。代理式系统的伦理风险需追溯至其双重技术特性的内在张力:其一,其核心推理机制基于概率预测,即从大规模数据中提取统计相关性,并据此生成“合理解释”;其二,系统以闭环方式执行多步任务,使初始判断不断进入后续决策链条。两者叠加产生了传统问答模型所不具备的系统性风险。

首先,概率预测逻辑与科学知识追求确定性因果的本质之间存在深层矛盾。计算机科学家和哲学家珀尔(Pearl)提出“因果阶梯”以表明,认知活动包含关联、干预与反事实三个层次^[34]。代理式系统目前主要停留在关联层面,其擅长识别模式并组织出表面连贯的解释,但这并不意味着输出已经触及因果机制本身。这一问题在多步执行中被进一步放大,每一步输出看似合理,但概率性推理的累积效应可能导致最终结论系统性偏离真实因果。

其次,概率预测与多步执行的叠加还催化了代理式系统特有的链式诱导风险。与传统大模型的单点幻觉不同,闭环结构使早期微小偏差在后续环节中被不断放大,每一步以前一步的概率性输出为输入,误差在链条中自我强化,形成路径依赖下的级联误差。同时,系统本身的“强制回答倾向”在证据不充分或不明确授权的情况下仍然进行自主推进,不仅可能致使系统在事实层面编造内容,还可能在能力与权限边界上宣称“已访问某工具”“已获得某类资料”,或在未满足授权条件时将推断当作证据,使得此类幻觉更难甄别,出现引用准确、齐全但结论有偏的“权威幻觉”,这种幻觉使得用户面对貌似完整可靠的证据链条,更容易降低批判性警觉,在未经充分复核的情况下将其作为后续研究的依据,从而使局部偏误进入结论形成过程。在责任锚定方面,传统问答式模型的应用过程仍然需要人类进行二次解释与采纳,责任链条相对清晰,但

代理式系统自主化程度的提升使得中间多步骤过程都以黑箱化自动呈现,责任网络难以界定,纠错成本随之提高。

(三)激励与规范失衡下的制度重构

从制度安排层面看,代理式系统通过改变知识生产的边际成本与过程形态,重塑科学共同体赖以运转的激励与规范结构。部分大样本的实证研究已经表明大模型工具确实具有加速论文与预印本产出的宏观效应^[35],在既有以论文发表数量、期刊等级为核心的评价结构下,人工智能带来的效率增益可能首先体现为发表规模的扩大。与此同时,同行评议的机制有效性成为严峻挑战,一项覆盖 111 个国家、1 600 余名学者的全球调研显示,受访者中 53% 的审稿人已在评审工作中使用人工智能工具^[36],人工智能深度渗透至评审环节,进而使以此为知识生产质量控制机制承受系统性压力。

此外,基于代理式系统使用的不平等正在成为一种新的数字鸿沟,与研究者可通过应用程序编程接口(API)较为灵活地调用传统大模型不同,代理式系统的生态特点使其运行高度依赖于由平台所有者控制的完整技术生态,包括底层模型与算力、实时数据接口、工具链集成与执行环境配置等。当研究者使用代理式工具时,更像是在进入由平台预设的完整研究 workflow,平台通过接口设计深度介入检索配置、数据源排序、工具调用范围等知识生产的每一个环节。研究者看似在自主设定研究目标,实则是在平台划定的技术生态与规则体系内行动。在此基础上,代理式系统的使用还受到研究者所在机构资源条件的制约。既有资源差异可能转化为前沿能力的获取差距,并在平台化 workflow 中进一步固化,形成新的知识生产不平等。

代理式系统还从制度规范角度重塑着作者贡献与署名规则的制度边界。代理式系统的快速演进与知识生产中的标准共识之间呈现巨大张力,凸显出法律伦理层面的主体适格性冲

突。国际出版伦理委员会(COPE)与国际医学期刊编辑委员会(ICMJE)的相关立场明确指出,AI工具不具备作者资格,也不能承担投稿责任^①。因此,无论使用程度如何,论文的最终责任必须由人类作者承担,并需在提交时对AI的使用情况及使用方式进行透明披露。在知识产权法领域,“塞勒(Stephen Thaler)诉专利局”的系列案件中,美国、英国及欧洲专利局均裁定人工智能不能作为发明人^②,确立了“人类中心主义”的知识产权主体框架。然而,现有制度对主体资格的明确,并未同时解决人工智能实际参与研究所引发的规范问题。《自然》杂志基于5 000名研究者的调查显示,科研共同体对于研究写作与同行评议中AI使用的可接受边界、披露标准和责任划分仍然缺乏稳定共识^[37]。2023年1月,《自然》与施普林格·自然旗下期刊提出,研究者使用大语言模型时,应在方法、致谢或其他适当部分说明其使用情况;爱思唯尔则要求作者在参考文献前声明所用AI工具及其用途,若AI用于研究过程,还须在方法部分详述,并对AI辅助生成或修改研究图像设置严格限制^③。相较于一般人工智能工具,代理式系统能够连续介入研究过程,使以成果披露为中心的现有规范更难满足责任追溯的要求,迫使评价与治理逻辑必然前移到“过程合规”,即对数据与证据链来源、模型参与程度、人工复核与解释过程进行留痕,并在评审保密与研究伦理层面建立可执行的制度约束,确保知识生产的责任链条仍能得到清晰追溯。

五、总结与展望

人工智能驱动的科学研究正在迅猛发展,代理式系统的崛起已超出人工智能技术的单纯迭代,指向知识生产的根本转变。代理式系统将检索增强、推理规划与工具调用等技术手段重新有机组合,实现知识生产各环节的流程再造。知识生产不再仅仅依赖研究者个体的线性推进,而转向人机同位的协作逻辑。这种转向虽然

极大地提升了信息整合效率并缩短了科研创新的时滞,但其背后隐藏的认识论危机同样深远。当科学界高度依赖少数顶级机构提供的基础底座时,研究视角与方法论的预先设定将不可避免地导致科学认知的窄化与同质化,甚至可能在数据主义的裹挟下,使那些难以被算法量化的默会经验边缘化。此外,随着代理自主性的提升,传统的责任框架正面临失锚风险,“链式诱导”的隐蔽性与责任网络的黑箱化成为代理式系统独有的新型伦理风险,也对现有的科研诚信与评审制度提出了严峻考验。

更为关键的是,代理式系统对于知识生产组织形态的更迭,撬动了长久以来科学共同体赖以运转的制度保障,在激励与规范两大维度上重塑知识生产生态。代理式系统降低了检索、综述与写作等环节的边际成本,规范机制则面临更为突出的共识缺乏、难以有效检测等关键障碍。此外,代理式系统的嵌入还会加创新的“马太效应”,能否获得更强模型、算力与数据正在成为知识生产能力与产出差距的关键变量。

展望未来,科学共同体亟须在制度层面作出系统性回应,通过“过程性治理”逻辑实现从“文本合规”向“过程合规”的前移,在投稿、评

①参见 COPE Council, "COPE Position Statement on Authorship and AI Tools," <https://publicationethics.org/cope-position-statements/ai-author>; International Committee of Medical Journal Editors, "Use of Artificial Intelligence in Publishing" <https://www.icmje.org/recommendations/browse/artificial-intelligence/>。

②参见 Thaler v. Vidal, 43 F.4th 1207 (Fed. Cir. 2022); Thaler v. Comptroller-General of Patents, Designs and Trade Marks [2023] UKSC 49; J 8/20 与 J 9/20(DABUS)。

③参见 Nature, "Tools such as ChatGPT threaten transparent science; here are our ground rules for their use," 24 January 2023, <https://www.nature.com/articles/d41586-023-00191-1>; Elsevier, "Generative AI policies for journals," <https://www.elsevier.com/about/policies-and-standards/generative-ai-policies-for-journals>。

审、资助与科研诚信管理等关键节点嵌入过程要求,建立可追溯的证据链与模型参与度透明机制。研究者应披露代理式系统介入的具体环节,并记录模型版本、数据来源、提示指令与人工复核等信息;期刊、出版商、资助机构、大学和专业学会则可据此形成最低共同标准,并结合不同国家的法律环境和不同学科的风险类型进行分级实施,通过过程披露与留痕、风险分级和跨机构互认,为知识生产建立可追溯的责任链条。同时,针对不断加剧的新型鸿沟,应构建更具包容性的算力与数据治理机制,防止知识生产能力沦为少数机构的资源壁垒。在人机深度共创的时代,人类的主体性不应在自动化浪潮中消退,而应更多体现于对原始问题的定义和伦理审视,从而在技术赋能与科学求真之间重建动态平衡。

参考文献

- [1] Garfield E. Citation indexes for science: a new dimension in documentation through association of ideas[J]. Science, 1955(3159): 108-111.
- [2] Priem J, Taraborelli D, Groth P, et al. Altmetrics: a manifesto[EB/OL]. (2010-10-26). <https://altmetrics.org/manifesto/>.
- [3] 王天恩. 数字智能与人机软融合知识生产范式[J]. 学术界, 2024(6): 70-79.
- [4] 孙伟平, 刘航宇. 生成式人工智能与知识生产方式的革命——从 Sora 的冲击波谈起[J]. 思想理论教育, 2024(5): 12-18.
- [5] 张钺, 李正风, 潜伟. 从 ChatGPT 到人机协作的知识共建[J]. 科学学研究, 2023(12): 2131-2137.
- [6] 邱泽奇. 人机互生: 范式革命与知识生产重构[J]. 探索与争鸣, 2024(11): 30-40, 209.
- [7] 肖峰. 生成式人工智能与知识生产新形态——关于三阶知识生产的探析[J]. 学术研究, 2023(10): 50-57, 2, 177.
- [8] 杜华, 顾小清. 人工智能时代的知识观审思[J]. 中国远程教育, 2022(10): 1-9, 76.
- [9] 张惠彬, 刘诗蕾. 挑战与回应: 人工智能创作成果的版权议题[J]. 大连理工大学学报(社会科学版), 2020(1): 76-81.
- [10] 邓建鹏, 朱恽成. ChatGPT 模型的法律风险及应对之策[J]. 新疆师范大学学报(哲学社会科学版), 2023(5): 91-101+2.
- [11] 刘艳红. 生成式人工智能的三大安全风险及法律规制——以 ChatGPT 为例[J]. 东方法学, 2023(4): 29-43.
- [12] 宋伟锋. 生成式 AI 训练数据源版权风险与规范重塑——以 ChatGPT 应用场景为例[J]. 大连理工大学学报(社会科学版), 2025(2): 100-110.
- [13] Wooldridge M, Jennings N R. Intelligent agents: theory and practice[J]. The Knowledge Engineering Review, 1995(2): 115-152.
- [14] Omar M, Nassar S, Hijazi K, et al. Generating credible referenced medical research: a comparative study of openAI's GPT-4 and Google's Gemini[J]. Computers in Biology and Medicine, 2025, 185: 109545.
- [15] 刘泽垣, 王鹏江, 宋晓斌, 等. 大语言模型的幻觉问题研究综述[J]. 软件学报, 2025(3): 1152-1185.
- [16] Yao S, Zhao J, Yu D, et al. ReAct: Synergizing reasoning and acting in language models[C]. International Conference on Learning Representations, 2023.
- [17] Merton R K. The sociology of science: theoretical and empirical investigations[M]. Chicago: University of Chicago Press, 1973: 268-269.
- [18] Latour B. Science in action: how to follow scientists and engineers through society[M]. Cambridge: Harvard University Press, 1987: 64-70.

- [19]Knorr -Cetina K D. The manufacture of knowledge: an essay on the constructivist and contextual nature of science[M]. Oxford: Pergamon Press, 1981: 33-40.
- [20]Pickering A. The mangle of practice: time, agency, and science[M]. Chicago: University of Chicago Press, 1995: 21-23.
- [21]Nowotny H, et al. The new production of knowledge: the dynamics of science and research in contemporary societies [M]. London: Sage, 1994: 1-16.
- [22]Etzkowitz H, Leydesdorff L. The triplehelix— university -industry -government relations: A laboratory for knowledge -based economic development [J]. EASST Review, 1995(1): 14-19.
- [23]Zhao C, Tan Z, Wong C W, et al. SCALE: Towards collaborative content analysis in social science with large language model agents and human intervention [C]//Proceedings of the 63rd annual meeting of the association for computational linguistics (Volume 1: Long Papers). Vienna: Association for Computational Linguistics, 2025: 8473-8503.
- [24]江小涓, 宫建霞, 李秋甫. 数据、数据关系与数字时代的创新范式[J]. 中国社会科学, 2024(9): 185-203, 208.
- [25]Zheng Y, Koh H Y, Ju J, et al. Large language models for scientific discovery in molecular property prediction [J]. Nature Machine Intelligence, 2025(3): 437-447.
- [26]RetkowskI F, Sudmann A, Waibel A. The AI Co-Ethnographer: How far can automation take qualitative research?[C]//Proceedings of the 5th International Conference on Natural Language Processing for Digital Humanities. Albuquerque: Association for Computational Linguistics, 2025: 73-90.
- [27]Bu D, Sun J, Li K, et al. Empowering AI data scientists using a multi-agent LLM framework with self-evolving capabilities for autonomous, tool-aware biomedical data analyses [J/OL]. Nature Biomedical Engineering, 2026.
- [28]Rothfarb S, Davis M C, Matanovic I, et al. Hierarchical multi-agent large language model reasoning for autonomous heterogeneous catalyst discovery[J/OL]. npj Computational Materials, 2026.
- [29]Ding K, Yu J, Huang J, et al. SciToolAgent: a knowledge-graph-driven scientific agent for multitool integration[J]. Nature Computational Science, 2025(10): 962-972.
- [30]布鲁诺·拉图尔, 史蒂夫·伍尔加. 实验室生活——科学事实的建构过程[M]. 张伯霖, 刁小英, 译. 北京: 东方出版社, 2004: 51-52.
- [31]Collins H M. The TEA set: tacit knowledge and scientific networks [J]. Social Studies of Science, 1974(2): 165-185.
- [32]赖尔. 心的概念[M]. 徐大建, 译. 北京: 商务印书馆, 1992.
- [33]波兰尼. 个人知识: 迈向后批判哲学[M]. 许泽民, 译. 贵阳: 贵州人民出版社, 2000: 102-103.
- [34]Pearl J. Causality: models, reasoning, and inference[M]. 2nd ed. Cambridge: Cambridge University Press, 2009: 376-379.
- [35]Kusumegi K, Yang X, Ginsparg P, et al. Scientific production in the era of large language models [J]. Science (New York, N. Y.), 2025(6779): 1240-1243.
- [36]La rosa G R M, Nasser M. AI in peer review: the elephant in the editorial room [J]. Evidence-Based Dentistry, 2026(2): 23-24.
- [37]Kwon D. Is it OK for AI to write science papers? nature survey shows researchers are split[J]. Nature, 2025(8063): 574-578.

AI-driven Agentic Systems and Knowledge Production: Transformation, Reconfiguration, and Challenges

Li Qiufu

Abstract: Since the release of ChatGPT, generative artificial intelligence has gradually become integrated into knowledge production processes such as literature retrieval, academic writing and data analysis. Unlike question-and-answer models centred on conversational responses, agentic systems are capable of autonomously breaking down tasks around a given objective and continuously adjusting their actions whilst retrieving information and utilising tools, thereby forming a continuous process from decision making to result generation. The technical logic of agentic systems differs from that of question-and-answer models, manifesting primarily in characteristics such as autonomous planning, closed-loop reasoning and cross-domain integration, thereby reshaping the subjectivity, process structure and organisational forms of knowledge production. Whilst agentic systems enhance the efficiency of knowledge production, they also present specific risks and challenges: computable objects are prioritised, thereby marginalising forms of knowledge that rely on tacit experience and historical context; the systems' multi-step automated intervention in the knowledge production process makes it more difficult to attribute responsibility; and existing research evaluation frameworks, centred on final outputs, struggle to account for the evidence screening and human verification that occur during the agentic systems' intervention. Moving forward, we should promote process-oriented governance of knowledge production, establishing process transparency and accountability as normative benchmarks, to mitigate the new forms of inequality arising from system accessibility and to re-establish a balance between efficiency gains and knowledge quality in the age of artificial intelligence.

Key Words: AI; Agentic Systems; Knowledge Production; AI4S; Science and Technology Governance

(责任编辑:丁忠兵)