

从第四范式到第五科研范式： 科研智能体驱动的知识生产方式变革

王爱清

宁德师范学院

摘要 经验科学、理论科学、计算科学与数据密集型科学，构成了理解现代科学研究方式变迁的常用框架。随着大语言模型、AI for Science、自主实验平台与多智能体科研系统的推进，人工智能已不止于处理数据或加速仿真，而是进入问题界定、文献理解、假设生成、实验设计、代码执行、数据分析、论文写作与同行评议等环节。本文在“四大科研范式”基础上提出并界定“第五科研范式”：以科研智能体为核心组织单元，以算力、科学数据、基础模型、自主实验平台和人机协同机制为支撑，通过自主化、闭环化、平台化与可审计化的方式组织科学发现。它与第四范式的差别不在于“数据更多”或“算法更强”，而在于研究主体、流程结构、验证机制和科研组织形态的同步重构。为避免概念泛化，本文提出三项最低判定标准：智能体须进入两个以上相互反馈的核心研究环节；系统须能依据中间证据修改行动计划，并与计算或实验验证相连接；全过程须具备可审计的人类监督与责任安排。文章分析第五科研范式的形成条件与运行机制，讨论其对科研组织、劳动分工、评价制度和国家创新体系的影响，并借助 A-Lab 自主合成论文的更正、AI Scientist 的第三方实测等案例说明，自动化程度的提高与知识可靠性的提高并不同步。因此，发展科研智能的关键不只在模型与算力投入，也在于科学数据体系、实验接口、可重复性基础设施与责任规则的同步建设。

关键词 第五科研范式；科研智能体；AI Scientist；AI for Science；知识生产；国家创新体系

DOI <https://doi.org/10.6914/tpss.080404> 文章编号 2664-1127.2026.0804.24-42

一、问题提出

（一）一个被更正的自动化实验

2023 年 11 月，劳伦斯伯克利国家实验室与加州大学伯克利分校的团队在《自然》报告了一套名为 A-Lab 的自主实验系统。它把材料计算数据库的稳定性预测、文献挖掘得到的合成经验与机器人固相合成串成一条流水线，在 17 天连续运行中，从 58 个目标化合物里做出了 41 个，实验规划、执行与物相判读基本没有人工介入^①。报道随即把它称作“机器人科学家”的雏形。

两年之后，结论被部分收回。伦敦大学学院的固体化学家 Palgrave 等人指出，被称为“新化合物”的

© The Author(s) 2026. CC BY-NC-ND 4.0. **Article History:** Published: 15 August 2026 (online), (print).

社会科学理论与实践 (Shèhuì kēxué lǐlùn yǔ shíjiàn) *Theory and Practice of Social Science*, ISSN 2664-1127 (Print). Published by Creative Publishing Co., Limited 开元出版有限公司. Email: cpcl@cpcl.hk.

^①Szymanski N. J., Rendy B., Fei Y., et al. An autonomous laboratory for the accelerated synthesis of inorganic materials[J]. *Nature*, 2023, 624: 86-91. <https://doi.org/10.1038/s41586-023-06734-w>

产物中,有相当比例只是已知无序相的有序化描述;更关键的是,用于确认物相的X射线衍射精修由算法自动完成,其证据强度不足以支撑“新材料”的判定。2026年2月,原作者刊出更正:论文中的“新”本意是“对预测平台而言是新的”,而非“对科学而言是新的”;经人工重新分析衍射数据,40个被报告为成功的案例中36个判定正确,另有4个无法确证^①。

这段插曲之所以适合作为本文的起点,是因为它同时呈现了两件事。其一,自动化科研系统确实已经能够把预测、执行与判读连成完整链条,这在十年前还只是设想。其二,链条上的每一次自动判断都可能把一个未被察觉的偏差一路传导到结论,而且传导速度与自动化程度成正比。前者是范式问题,后者是治理问题,二者必须一起讨论。

(二) 从第四范式到“第五”的命名之争

近代以来,人类科学大体经历了从观察实验到理论建模、从计算模拟到数据密集型发现的演进。2009年, Tony Hey、Stewart Tansley 与 Kristin Tolle 主编的《第四范式:数据密集型科学发现》整理了 Jim Gray 关于科研方式变迁的思考,把经验科学、理论科学、计算科学之后的数据密集型科学概括为第四科研范式^②。此后,“第四范式”成为理解大数据、数据库、机器学习与 e-Science 如何改变科学研究的常用概念。

生成式人工智能与科研智能体的兴起,把这一序列继续向前推。2022年,微软研究院在 AI4Science 的论述中提出,深度学习可能开启科学发现的“第五范式”^③;2024年以来, AI Scientist、自主科研系统与 AI co-scientist 加速涌现,使人工智能从“辅助科学家”走向“参与科学发现”^④;2025年,《PLOS Biology》刊文认为“人工科学智能”的时代正在到来, AI scientists 不仅辅助研究,还能生成假设、驱动发现并自主检验^⑤。

需要承认的是,这些提法并未形成统一命名。有的把“第五范式”落在深度学习这一方法上,有的落在“人工科学智能”这一主体上,也有研究者干脆拒绝使用“范式”一词,认为它在库恩的原意之外被过度借用。本文仍沿用“第五科研范式”,但把它当作一个需要设定判定条件的分析概念,而不是一个已经确立的事实描述——第六部分将专门说明滥用这一标签会带来什么代价。

从社会学的角度看,真正的问题既不是人工智能能否替代科学家,也不是 AI for Science 能否提高某个领域的研发效率,而是:当人工智能系统开始承担知识生产的核心环节,科学研究的组织方式是否已经改变?如果已经改变,其特征是什么?它对国家创新体系、科研组织与科技治理制度意味着什么?回答这些问题,需要越过技术应用综述,进入科学社会学、创新经济学与科技治理的交叉地带。

本文的判断是:科学研究正在从第四范式走向第五科研范式。第五科研范式不等同于“AI 驱动科学”这一泛称,也不只是第四范式中机器学习方法的自然延伸,它以科研智能体为核心组织单元,以算力、数据、模型与实验平台为基础设施,以人机协同和自主闭环为运行机制。用一句话概括这一变化:科学研究正从“人类科学家使用工具发现知识”,转向“人类科学家与科研智能体共同生产、验证并承担知识”。

围绕这一判断,全文拟回答四个问题:第一至第四科研范式的演进逻辑是什么;第五科研范式何以成立,与第四范式的实质区别在哪里;科研智能体如何改变科学发现的运行机制;这一变化对我国国家创新

①Szymanski N. J., Rendy B., Fei Y., et al. Author Correction: An autonomous laboratory for the accelerated synthesis of inorganic materials[J]. Nature, 2026, 650: 8100. <https://doi.org/10.1038/s41586-025-09992-y>

②Hey T., Tansley S., Tolle K. The Fourth Paradigm: Data-Intensive Scientific Discovery[M]. Redmond: Microsoft Research, 2009.; Gray J. Jim Gray on eScience: A Transformed Scientific Method[A]. In: Hey T., Tansley S., Tolle K. The Fourth Paradigm: Data-Intensive Scientific Discovery[M]. Redmond: Microsoft Research, 2009.

③Bishop C. AI4Science to empower the fifth paradigm of scientific discovery[EB/OL]. Microsoft Research Blog, 2022-07-07. <https://www.microsoft.com/en-us/research/blog/ai4science-to-empower-the-fifth-paradigm-of-scientific-discovery/>

④Lu C., Lu C., Lange R. T., Foerster J., Clune J., Ha D. The AI Scientist: Towards Fully Automated Open-Ended Scientific Discovery[EB/OL]. arXiv:2408.06292, 2024. <https://arxiv.org/abs/2408.06292>; Gottweis J. et al. Towards an AI co-scientist[EB/OL]. arXiv:2502.18864, 2025. <https://arxiv.org/abs/2502.18864>

⑤Miolane N. The fifth era of science: Artificial scientific intelligence[J]. PLOS Biology, 2025, 23(6): e3003230. <https://doi.org/10.1371/journal.pbio.3003230>

体系建设有何启示。

（三）分析层次与研究方法

对第五科研范式的分析不能停留在技术能力清单层面。把大模型生成文本、调用代码或控制仪器的能力依次罗列，很容易把局部工具进步误判为整体范式变迁。科学范式的变化至少牵动四个层次：认识论关注何种证据可被接受、机器生成的解释能否构成科学解释；方法论关注假设、实验、验证与复现如何连接；组织论关注人类研究者、智能体、数据平台与实验设施如何分工；制度论则关注责任、评价、产权、安全与公共资源配置如何调整。只有这四个层次出现相互强化的结构变化，“第五科研范式”这一概念才有使用价值。

方法上，本文结合历史比较、概念分析与制度分析。纵向以第一至第四科研范式为比较框架，考察知识来源、研究工具、核心基础设施与研究主体的变化；横向以第四范式与第五范式为比较对象，从驱动要素、行动主体、流程结构、验证机制与组织形态辨析边界；再以科研智能体参与知识生产的完整链条为分析单元，考察问题理解、假设生成、实验执行、结果解释与知识表达之间的反馈关系；最后把这一变化放回国家创新体系，讨论公共科研基础设施、大学与科研院所、科技企业、评价制度与风险治理之间的新关系。

这一方法强调过程而非单一结果。科研智能体是否生成了一篇形式完整的论文，并不足以证明它完成了科学发现；更重要的是追踪问题从何而来、数据是否可靠、假设如何形成、实验能否复现、结论能否被反驳，以及责任主体是否清晰。开篇的 A-Lab 案例正说明了这一点：仅看“17 天做出 41 个化合物”这一结果，与追踪其物相判定的证据链，会得出相当不同的评价。

（四）研究价值与文章结构

本文的理论意图，是把第五科研范式从技术趋势判断推进为知识生产方式分析。文章并不把“第五范式”简单定义为人工智能用于科研，而是把三项条件当作概念成立的门槛：科研智能体能否形成自主但受约束的行动链，能否与物理实验或计算验证相连接，能否留下可追溯的知识记录。这样界定，既能解释为什么一般的数据分析、模型预测或写作辅助仍属于第四范式，也能说明第五范式并不排斥实验、理论、计算与数据，而是对既有范式的重新编排。

实践意图则在于为我国科研智能基础设施建设提供组织与治理视角。科研智能的发展不只需要模型参数和算力，也需要高质量科学数据、标准化实验接口、领域知识工程、跨机构合作协议、审计记录与安全边界。只建通用模型而不管科学数据与实验平台，智能体容易停留在文本层面；只追求自动化效率而不管复现、责任与评价，则可能扩大低质量知识供给。科研智能政策因此需要从单点项目支持，转向“数据—模型—平台—人才—制度”的协同建设。

结构上，第二部分回顾范式的叠加演进并界定第五科研范式；第三部分分析其形成条件与知识生产闭环；第四部分讨论核心特征以及科研组织、劳动分工和评价方式的重构；第五部分考察其对国家创新体系的影响并提出我国路径；第六部分澄清理论边界，讨论可信性、伦理、安全与责任治理，最后给出结论与后续研究议程。

二、科研范式演进与第五科研范式的理论界定

（一）四大科研范式的演进逻辑

“科研范式”并非由某一学科统一规定的制度化分类，而是一种用于解释科学研究方式演进的分析框架。第一至第四科研范式的划分，通常用来说明科学发现如何随观察手段、理论工具、计算能力与数据基础设施的发展而阶段性变化。这里必须强调，不同范式之间并非替代关系，而是叠加、互补与重组的关系。今天的科学研究仍然离不开观察实验、理论建模、计算模拟和数据分析，改变的是这些要素之间的组合方

式。

第一范式是经验科学，核心方法是观察、实验、记录与归纳。无论早期天文学的星象观测、生物学的物种分类，还是医学中的临床经验积累，都依赖对自然现象的直接接触。经验科学贴近事实，能为知识生产提供原始证据，其局限是难以形成可推广的普遍解释。

第二范式是理论科学，核心方法是数学建模、逻辑推演与理论解释。牛顿力学、麦克斯韦电磁理论、相对论、量子力学以及许多社会科学模型，都体现了理论科学的力量：从有限观察中抽象出一般规律，并以形式化方式解释和预测复杂现象。其局限则在于，面对高度复杂、非线性和多尺度的系统时，理论模型往往需要很强的假设和简化条件。

第三范式是计算科学。数字计算机和高性能计算的发展，使科学家能够通过数值计算和仿真实验研究复杂系统。气候模拟、流体力学、核反应模拟、分子动力学、复杂经济系统仿真，都是它的典型应用。计算科学的意义在于，它让许多无法直接实验或无法解析求解的问题获得了可操作的研究路径，“模拟”由此成为与观察、理论并列的科学方法。

第四范式是数据密集型科学。进入 21 世纪，科学仪器、传感器、数据库、互联网平台和高通量实验持续产生海量数据，科学发现越来越依赖大规模数据的采集、存储、清洗、分析与可视化。基因组学、天文大数据、材料数据库、社会计算和数字人文，都带有这一特征。以 AlphaFold 为代表的蛋白质结构预测是标志性事件：在 2020 年 CASP14 评估中，AlphaFold2 在全局距离检验上的中位得分达到 92.4，首次在多数目标上接近实验解析精度^①。第四范式使数据本身成为发现规律的重要来源，也使机器学习、数据库技术和云计算平台成为科研基础设施的一部分。

四大范式的属性差异可以概括如表1所示。演进过程表明，科研方式的变化通常由三类因素共同推动：认识对象的复杂化、工具与基础设施的升级、知识组织方式的变化。经验科学重在事实获取，理论科学重在规律表达，计算科学重在复杂系统模拟，数据密集型科学重在从海量数据中发现结构。第五科研范式的出现，正是这些因素进一步叠加的结果：科学问题更复杂，数据和模型规模更大，实验与仿真成本更高，单个研究者已难以独立掌控完整的知识生产链条，而智能体系统开始进入这一链条并承担部分自主性工作。

表 1 第一至第五科研范式的比较

范式	主导方法	关键基础设施	研究主体与分工	典型例证
第一范式 经验科学	观察、实验、记录与归纳	观测仪器、实验室、标本与档案	个体研究者及学徒式团队	第谷的天文观测记录；林奈的物种分类
第二范式 理论科学	数学建模、逻辑推演与形式化解释	数学工具、学术共同体与出版体系	理论家主导，实验者承担验证	牛顿力学；麦克斯韦电磁理论；量子力学
第三范式 计算科学	数值计算与仿真实验	高性能计算机、算法库与科学软件	学科专家与计算专家分工协作	气候环流模式；分子动力学；核反应模拟
第四范式 数据密集型科学	大规模数据采集、清洗、关联与机器学习	科学数据库、云平台、高通量仪器	跨学科团队，数据工程角色出现	人类基因组计划；天文巡天数据库；AlphaFold 结构预测
第五范式 智能体驱动科研	目标约束下的任务分解、工具调用、执行与迭代	“算力—数据—模型—实验平台”复合体，及日志与权限体系	人类设定问题与承担责任，智能体承担过程执行	Coscientist 的自主化学实验；A-Lab 的自主合成；AI Scientist 的端到端论文生成

从科学哲学角度看，科研范式不是对工具的简单命名，而是一套关于重要问题、有效证据、合格方法与共同体规范的相对稳定安排。库恩关于科学革命的讨论提示，范式变化通常伴随问题选择、评价标准与

^①Jumper J., Evans R., Pritzel A., et al. Highly accurate protein structure prediction with AlphaFold[J]. Nature, 2021, 596: 583–589. <https://doi.org/10.1038/s41586-021-03819-2>; Tunyasuvunakool K., Adler J., Wu Z., et al. Highly accurate protein structure prediction for the human proteome[J]. Nature, 2021, 596: 590–596. <https://doi.org/10.1038/s41586-021-03828-1>

共同体实践的调整^①。因此，第五科研范式能否成立，不能只以模型性能是否提高为依据，还要看科学共同体是否开始围绕智能体重新配置研究流程、判断标准与协作制度。技术突破是必要条件，却远不是充分条件。

范式演进还具有明显的累积性。理论科学没有取消经验观察，计算科学没有取消理论建模，数据密集型科学也没有让实验和因果解释失去意义。每一种新范式都把此前的范式转化为新的输入条件：实验产生可计算的数据，理论为模型提供约束，仿真为假设提供低成本试验场，数据库为跨尺度发现提供材料。第五科研范式同样谈不上是“第五种孤立方法”，它的作用在于通过科研智能体把观察、理论、计算和数据组织成可持续运行的任务网络，其结构关系如图1所示。

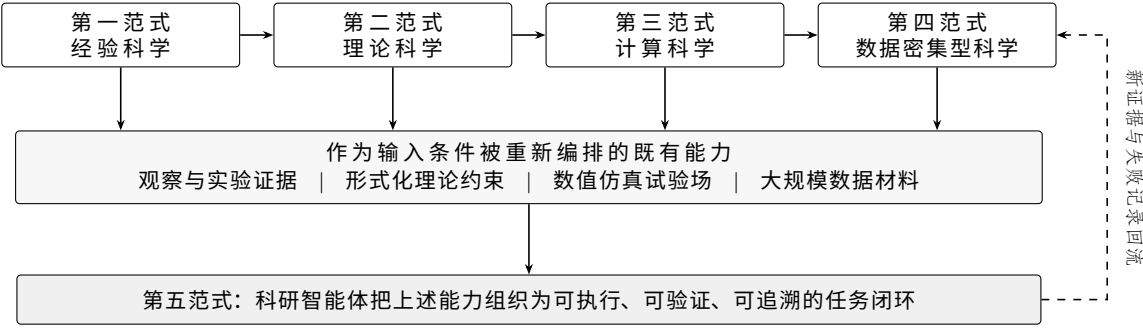


图1 科研范式的叠加演进与要素重组

这种累积关系意味着，未来最具竞争力的科研系统未必拥有最大的模型，而是能够对多种证据和方法作出有原则的组合。面对可由明确规则描述的问题，理论模型和数值计算仍然透明、稳定；面对高维搜索空间，数据和机器学习更有效率；面对涉及价值选择、研究方向与责任承担的问题，人类科学家的判断不可替代。第五范式的功能，在于依据任务特征调度这些能力，并记录调度依据与验证过程。

波兰尼关于默会知识的分析，进一步划出了智能化科研的边界^②。实验操作、异常识别、问题品位与理论直觉中，存在大量难以完全编码的经验。科研智能体从文本、数据和反馈中学到部分模式，却未必能自动获得实验室长期积累的情境判断。因此，第五范式的发展方向并不是把默会知识全部形式化，而是设计更有效的人机接口：让人类研究者能把关键判断嵌入任务约束，让智能体的搜索结果能被专家理解和修正，并把新的实践经验沉淀为可复用而不过度简化的组织知识。

（二）第五科研范式的概念界定

本文将第五科研范式界定为：以科研智能体为核心，以算力、科学数据、基础模型、自主实验平台和人机协同机制为支撑，通过问题理解、假设生成、实验设计、仿真计算、数据分析、知识表达与自我迭代等环节的智能化闭环，推动科学发现和知识生产的新型科研范式。

这一界定的第一个约束落在“科研智能体”上。它指的不是一般意义上的人工智能工具，而是面向科学研究任务构建的智能行动系统，能够在目标约束下进行文献阅读、问题分解、方案设计、工具调用、代码执行、实验管理、结果解释与文本生成。与传统软件工具相比，科研智能体的区别在于任务连续性、上下文理解能力和有限度的自主行动能力——它可以在没有逐条指令的情况下走完多个步骤。

与之相连的是基础设施问题。第五范式依托的从来不止一个模型，它需要由“算力—数据—模型—实验平台”组成的复合系统：大模型提供推理、生成与交互能力，科学数据库提供知识来源，高性能计算和云平台提供算力，自主实验室与机器人平台提供物理世界的验证能力。离开这一复合基础设施，科研智能体就只能停留在文本智能，无法抵达真实的科学发现。中国科学技术大学团队研制的“机器化学家”是一

①Kuhn T. S. The Structure of Scientific Revolutions[M]. Chicago: University of Chicago Press, 1962.

②Polanyi M. The Tacit Dimension[M]. Chicago: University of Chicago Press, 1966.

个可对照的例子：它把文献机读、机器人实验与理论计算三个模块连成一体，在高熵合金催化剂的筛选中把原本需要长期人工试错的过程压缩到可控范围内^①。

运行方式上，第五范式的特征是闭环。传统 AI for Science 往往聚焦某一环节，例如用机器学习预测蛋白质结构、筛选材料或分析图像；第五范式强调从问题提出到结果验证的连续循环：提出假设、设计实验、执行实验、分析结果、修正模型、生成新假设，并在循环中提高发现效率。闭环的意义与其说在于提速，不如说在于让反馈能够改变系统后续的搜索方向。

至于主体结构，第五范式并不意味着人类科学家退场，而是意味着其角色发生转移：从直接完成所有研究步骤，转向提出高价值问题、设定边界条件、进行价值判断、审查研究结果并承担最终责任；科研智能体则负责扩展搜索空间、提高试错效率、连接知识片段与自动化执行重复性任务。最后一点关乎组织形态：未来的科学竞争不只是单个科学家、单个实验室或单项设备的竞争，也是科研智能平台、科学数据生态、模型训练体系、实验自动化系统与治理能力的竞争，知识生产因而可能从“实验室中心”走向“平台中心”。

综合起来，第五科研范式与第四范式的分野，与其说落在从“大数据”到“大模型”的技术更替上，不如说落在从“数据驱动发现”到“智能体组织发现”的结构变化上。第四范式强调数据规模、数据处理与数据发现；第五范式强调智能体的自主性、流程的闭环性、知识生产的平台性以及人机协同的制度化。

（三）第四范式与第五范式的多维比较

两种范式的差异可以沿七个维度展开，如表2所示。这里挑出其中三点作进一步说明，因为它们最容易被混淆。

表 2 第四范式与第五范式的多维差异及判别要点

比较维度	第四范式（数据密集型）	第五范式（智能体驱动）	判别要点
稀缺资源	高质量大规模数据及其处理能力	可供智能体调用的任务环境：数据、模型、软件、算力、仪器接口与权限规则	数据是否被嵌入目标明确、反馈连续的行动链
人工智能的角色	分析工具，研究目标与流程顺序由人类预设	受约束的行动代理，可分解目标、选择工具并依据中间结果修改计划	系统能否在没有逐条指令时完成多步骤任务
流程结构	“数据进入—模型处理—结果输出”的线性链条	“提出—设计—执行—分析—更新”的循环	反馈是否成为系统能力，而非由研究者手工组织
验证对象	模型精度、统计显著性与外部数据表现	增加过程可追溯性、工具调用合规性、数据版本一致性与解释忠实度	研究记录是否从辅助材料上升为验证对象
组织单元	课题组为主，辅以数据库与计算中心	跨机构科研智能平台与智能体网络	团队能力更多来自资源独占还是资源编排
失败的处置	多数不进入正式记录	被结构化编码为可检索的负面结果	失败记录能否改变后续搜索空间
责任安排	作者与所属机构	需在人类研究者、机构、平台与模型提供者之间分层界定	是否存在可核对的授权与操作日志

第一点是流程结构。第四范式常表现为线性过程，环节之间的反馈多由研究者手工组织；第五范式则把反馈本身变成系统能力，使实验失败、数据异常和同行批评能够直接触发新的检索、计算或实验任务（图2）。线性流程当然也可以反复运行，区别在于谁来发起下一轮，以及发起的依据是否被记录下来。

^①Zhu Q., Zhang F., Huang Y., et al. An all-round AI-Chemist with a scientific mind[J]. National Science Review, 2022, 9(10): nwac190. <https://doi.org/10.1093/nsr/nwac190>

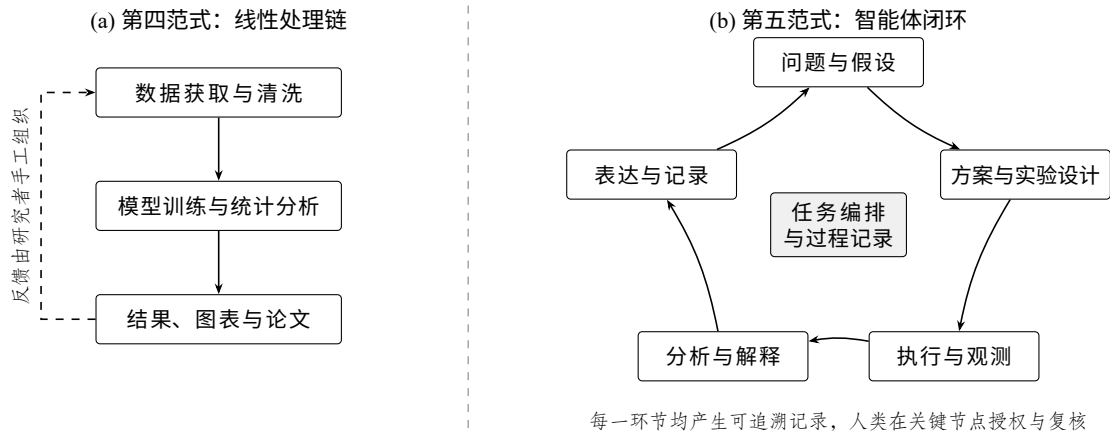


图2 第四范式的线性链条与第五范式的闭环链条

第二点是验证对象的扩展。第四范式重点检验模型精度、统计显著性和外部数据表现；第五范式在保留这些标准之外，还必须验证行动过程是否可追溯、工具调用是否合规、数据版本是否一致、自动生成的解释是否忠实于证据，以及人类究竟在哪些环节作出了实质判断。A-Lab 的更正正是在这一维度上发生的：其合成结果本身多数没有问题，出问题的是自动化物相判定这一中间环节的证据强度^①。验证对象一旦从最终结果扩展到完整过程，研究记录就从辅助材料上升为科研基础设施。

第三点是组织形态。第四范式推动了数据库、计算中心与开放平台的发展，但课题组仍是主要行动单元；第五范式可能形成跨机构的科研智能平台，多个智能体围绕文献、计算、实验、审校和管理分工协作。平台既能降低专业工具的使用门槛，也可能形成新的集中化权力——谁控制科学数据、基础模型和实验接口，谁就可能影响研究议题、资源配置与知识可见性。这使第五范式不仅是效率问题，也是科研公共性问题。

由此可以给出第五范式的最低判定标准，共三条：其一，科研智能体进入两个以上相互反馈的核心研究环节，而不仅提供写作或检索辅助；其二，系统能够根据中间证据调整行动计划，并与计算或实验验证相连接；其三，全过程具备可审计的人类监督与责任安排。三条同时满足，才宜把有关实践视为第五科研范式的典型形态。这一标准偏严，好处是能把大量“使用了 AI 的第四范式研究”排除在外。

三、第五科研范式的形成条件与知识生产机制

（一）形成条件：技术侧与制度侧

第五科研范式的出现有其条件，它是多重技术与制度条件同时成熟的结果。这些条件可以分为技术侧和制度侧两组，前者决定智能体“能做什么”，后者决定它“被允许做什么、做完之后算不算数”。

技术侧的第一项变化是科学问题复杂度的持续提高。生命科学中的蛋白质折叠、药物发现与基因调控，材料科学中的结构—性能关系，气候科学中的长期预测，能源科学中的聚变控制，社会科学中的复杂行为建模，都需要跨学科知识、海量数据和高强度计算。传统科研流程在这类问题面前容易暴露三个瓶颈：知识碎片化、试错成本高、研究周期长。

第二项变化来自科学数据的规模与类型。高通量测序、自动化显微成像、遥感观测、粒子探测、网络行为数据、文献数据库与开放科学平台不断扩展数据边界。第四范式已经把数据变成了发现的重要资源，

^①Szymanski N. J., Rendy B., Fei Y., et al. An autonomous laboratory for the accelerated synthesis of inorganic materials[J]. Nature, 2023, 624: 86–91. <https://doi.org/10.1038/s41586-023-06734-w>; Szymanski N. J., Rendy B., Fei Y., et al. Author Correction: An autonomous laboratory for the accelerated synthesis of inorganic materials[J]. Nature, 2026, 650: 8100. <https://doi.org/10.1038/s41586-025-09992-y>

但规模扩大同时带来数据理解、整合与可信性问题——数据越多，判断哪些数据可用的成本越高。科研智能体在跨库检索、语义标注、模式发现和异常检测上能够承担相当一部分工作。

第三项变化是基础模型的跨任务迁移能力。大语言模型、多模态模型和科学基础模型突破了传统机器学习只能处理窄任务的局限，可以理解自然语言、生成代码、调用工具、总结文献、设计流程，并在领域之间迁移知识。这为科研智能体提供了通用认知接口，使其能够把文献、数据、代码、实验设备与科研人员连接起来。

第四项变化最关键，也最容易被忽略：自主实验平台的成熟。没有它，科研智能体只能在文本和计算空间里活动。Coscientist 系统由大模型驱动，可以检索文献与仪器文档、编写控制代码并操作液体处理设备，完成了钯催化交叉偶联反应的自主优化^①；中国科学技术大学团队研制的“机器化学家”把文献机读、机器人实验与理论计算三个模块连成一体，用于高熵合金催化剂的筛选^②；A-Lab 则把材料数据库预测与固相合成机器人相接^③。这三个系统的共同点，是让“假设生成”与“现实验证”第一次在同一条自动化链路上闭合。

制度侧的条件同样不可省略。开放数据、开放代码、预印本平台、数字实验记录、科研协作平台与在线同行评议，为智能体创造了可读取、可调用、可追踪的环境；科研过程越数字化，智能体越容易嵌入其中；流程越标准化，自动化和可审计化越容易实现。与之配套的是研究对象和流程的“可接口化”：文献需要机器可读的元数据，数据需要明确语义和版本，计算工具需要稳定接口，实验设备需要能接收标准化指令，审批制度也需要把权限和风险转化为可执行规则。接口化并不等于把科研活动机械化，它只是为不同知识环节建立可验证的连接；如果文献、样本、代码、参数和仪器状态彼此割裂，再强的模型也只能在局部生成建议。

还有一项条件常被略过，即质量基础设施。传统质量控制主要依靠实验规范、统计检验、同行评议和机构伦理审查；智能体参与之后，还需要数据血缘追踪、模型版本管理、提示与工具调用日志、自动决策回放、权限分级和异常终止机制。这些机制决定了系统能否回答三个问题：结论由何种证据产生，哪一步由谁授权，错误能否定位和纠正。A-Lab 的更正之所以能够完成而不必撤稿，恰恰是因为原始衍射数据被完整保留，可以重新人工分析^④。可审计能力越弱，自动化规模越大，错误知识被快速复制的风险就越高。

最后，开放并不等于无条件使用。科学数据中可能包含隐私信息、商业秘密、敏感地理信息或双重用途知识，因而需要在开放共享与安全控制之间形成分层规则：基础文献和低风险数据可广泛开放，受限数据通过可信计算环境调用，高风险实验和敏感模型必须设置人工审批与使用记录。制度设计的目标不是消除风险，而是让风险与任务价值、使用者资格和控制能力相匹配。

（二）科研智能体驱动的知识生产机制

第五科研范式的运行机制，可以概括为“问题—知识—假设—实验—验证—表达—迭代”的智能闭环（参见图2）。

起点是问题理解。科研智能体通过读取文献、数据库、实验记录和研究计划，形成对研究领域的语义理解，不仅检索已有知识，还能识别研究空白、争议点和潜在问题。与传统文献检索不同，这种理解具有任务导向性和上下文连续性，可以围绕特定目标持续整合信息。

其次是假设生成。科学发现的关键在于提出有价值的假设，智能体可以通过跨文献关联、知识图谱推

①Boiko D. A., MacKnight R., Kline B., Gomes G. Autonomous chemical research with large language models[J]. Nature, 2023, 624: 570–578. <https://doi.org/10.1038/s41586-023-06792-0>

②Zhu Q., Zhang F., Huang Y., et al. An all-round AI-Chemist with a scientific mind[J]. National Science Review, 2022, 9(10): nwac190. <https://doi.org/10.1093/nsr/nwac190>

③Szymanski N. J., Rendy B., Fei Y., et al. An autonomous laboratory for the accelerated synthesis of inorganic materials[J]. Nature, 2023, 624: 86–91. <https://doi.org/10.1038/s41586-023-06734-w>

④Szymanski N. J., Rendy B., Fei Y., et al. Author Correction: An autonomous laboratory for the accelerated synthesis of inorganic materials[J]. Nature, 2026, 650: 8100. <https://doi.org/10.1038/s41586-025-09992-y>

理、模拟结果分析和模型生成能力提出候选假设。但生成假设不等于假设成立。此处恰恰是自动化最容易失手的地方：对 AI Scientist 系统的一项第三方评测发现，它常把已经确立的概念误判为“新颖”，例如把随机梯度下降中的微批处理当作创新点^①。判断假设是否具有科学意义、理论价值和实验可行性，仍然依赖人类科学家。

第三是实验设计。智能体可以依据假设生成实验方案、仿真方案或计算流程，涵盖变量选择、样本设计、参数设置、对照组安排、评价指标与风险控制；计算任务可直接生成代码并调用算力，实验任务则与自动化平台衔接，形成实验队列。

第四是执行与监控。智能体执行代码、调度仿真、监控实验过程、记录异常并动态调整流程。在这一环节，它从文本生成者变为行动执行者，其能力边界取决于工具调用权限、实验平台开放程度、模型可靠性与安全约束。前述评测中，约 42% 的自动实验因代码错误而失败，另有一部分产生了看似成立实则误导的结果^②——执行能力与执行可靠性之间存在明显落差。

第五是结果分析与解释。智能体能对实验数据进行统计分析、模式识别、可视化呈现和初步解释，并把结果与既有理论和文献相联系。这一环节效率提升明显，风险也最集中：错误解释、幻觉推理和过度拟合都可能在此产生，且往往披着规范的统计外衣。第五范式因此必须建立结果复核、可重复性验证和人类专家审查机制。

第六是知识表达。AI Scientist 类系统已经展示出自动生成论文、图表和评审意见的能力，其开发者称单篇成本可低至 15 美元以内^③。但成本下降与质量提升并不同步：独立评测指出，这类系统生成的稿件引文中位数仅为 5 条且多为过时文献，个别稿件甚至保留了“Conclusions Here”一类占位文字与凭空生成的结果^④。知识表达的自动化会显著降低科研文本的生产成本，也可能带来低质量论文激增、引用污染、评价失真和同行评议负担加重。第五科研范式不应把“自动生成论文”当作目标，其核心标准只能是“可验证的新知识生产”。

最后是迭代优化。智能体能够依据实验失败、评审意见、数据异常和人类反馈更新研究计划，形成连续迭代。就此而言，第五科研范式的价值不在于一次性给出答案，而在于构建可持续演化的科研系统。

（三）多智能体协作与任务编排

复杂科研任务往往难以由单一智能体可靠完成。一个模型既要理解理论，又要编写代码、调用仪器、分析数据并审查结论，容易出现能力错配与错误累积。多智能体系统提供了专业化分工的思路：文献智能体负责证据检索和引用核查，理论智能体负责概念建模和假设生成，计算智能体负责编程、仿真与统计分析，实验智能体负责仪器调度和过程监测，审校智能体检查数据一致性、方法合规性和论证漏洞，安全智能体则对高风险行动保留否决权。角色之间通过明确的输入输出协议连接，把复杂任务分解为相对可验证的模块（图3）。

多智能体协作的价值不只是并行提速，更在于形成内部质疑机制。若所有智能体共享同一模型、同一数据和同一目标函数，所谓协作可能只是把同一种偏差重复若干遍。有效的协作应当允许角色采用不同的证据来源、方法假设和评价标准：生成方案的智能体追求搜索空间扩展，审查方案的智能体优先寻找反例和约束；统计智能体检查量化结果，领域专家智能体检查机制解释；安全智能体对高风险行动拥有否决权。这种结构化分歧有助于减少单一路径依赖。

①Beel J., Kan M.-Y., Baumgart M. Evaluating Sakana's AI Scientist: Bold Claims, Mixed Results, and a Promising Future?[EB/OL]. arXiv:2502.14297, 2025. <https://arxiv.org/abs/2502.14297>

②Beel J., Kan M.-Y., Baumgart M. Evaluating Sakana's AI Scientist: Bold Claims, Mixed Results, and a Promising Future?[EB/OL]. arXiv:2502.14297, 2025. <https://arxiv.org/abs/2502.14297>

③Lu C., Lu C., Lange R. T., Foerster J., Clune J., Ha D. The AI Scientist: Towards Fully Automated Open-Ended Scientific Discovery[EB/OL]. arXiv:2408.06292, 2024. <https://arxiv.org/abs/2408.06292>

④Beel J., Kan M.-Y., Baumgart M. Evaluating Sakana's AI Scientist: Bold Claims, Mixed Results, and a Promising Future?[EB/OL]. arXiv:2502.14297, 2025. <https://arxiv.org/abs/2502.14297>

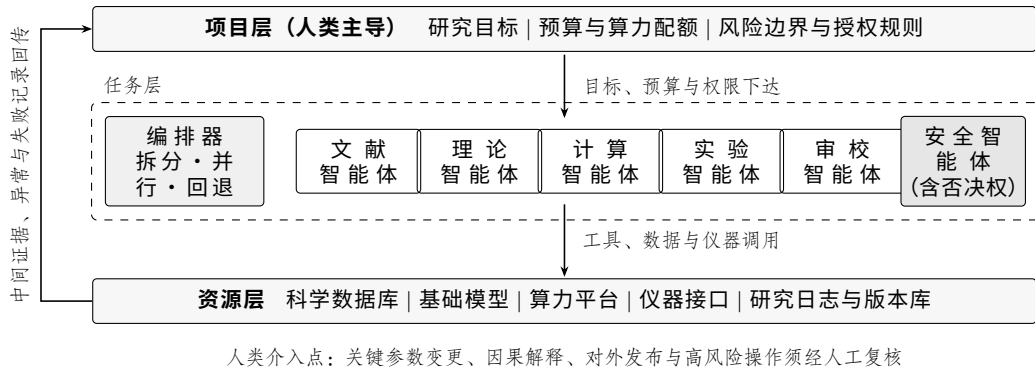


图3 多智能体分层编排结构与人类介入点

任务编排是多智能体科研系统的核心。编排器要决定任务如何拆分、何时并行、何时回退、哪些结果需要交叉验证、何时请求人类介入。过度集中于一个编排器会形成单点失误，完全分散又可能导致目标冲突和资源浪费。较稳健的设计是分层编排：项目层设定研究目标、预算与风险边界，任务层动态分配工具和数据，关键节点设置独立验证，异常时回到最近一次被确认的状态。

科研任务还有长周期特征。药物研发、材料实验和长期社会调查可能持续数月乃至数年，其间研究条件、数据版本和人员判断都会变化。智能体因此必须具备稳定的记忆与状态管理能力，能够区分原始证据、阶段性推断和已被否定的方案，避免把过时结论重新引入当前任务。长期记忆不能只是不断累积文本，还应包含证据等级、时间戳、版本关系、责任人和适用条件。

（四）人类监督、验证层级与失败管理

人机协同的重点不是要求人类在每一步签字，而是按风险和不确定性设计差异化监督。格式转换、代码单元测试、公开文献整理这类低风险、可逆且易复核的任务，可由智能体自动完成并保留日志；样本选择、因果解释、关键参数变更和对外发布，应由专业研究者复核；涉及生命安全、环境风险和双重用途的行动，则必须经机构审批。监督强度与潜在损害、可逆性和证据不确定性相匹配，是这一设计的基本原则。

验证同样应当分层。第一层是语法与运行验证，确认代码、公式和工具调用能够正确执行；第二层是数据与统计验证，检查样本、变量、缺失值、模型设定与稳健性；第三层是领域验证，判断结果是否符合专业知识并能解释异常；第四层是独立复现，由不同人员、模型或实验平台重新执行关键步骤；第五层是共同体验证，通过开放数据、同行评议和后续研究接受反驳。智能体可以支持每一层，却不能用较低层的运行成功替代较高层的科学成立——A-Lab 的合成流程在第一、二层运行无误，问题出在第三层的领域判断上。风险等级、监督方式与验证层级的对应关系，可参见表3。

表3 风险等级、人类监督方式与最低验证层级的对应

风险等级	任务示例	智能体自主度	人类监督方式	最低验证层级与记录要求
低	格式转换、代码单元测试、公开文献整理、描述性统计	可连续自动执行	事后抽查，保留完整日志	L1 运行验证；保存提示、工具清单与执行环境
中	样本选择、变量构造、模型设定变更、结果的因果解释	分步执行，关键节点暂停	专业研究者逐节点复核并署名	L2 数据与统计验证、L3 领域验证；保存数据快照与版本差异
高	新化合物合成、活体实验、涉敏数据调用、结论对外发布	仅在授权范围内执行，禁止自行扩权	机构审批加双人授权	L4 独立复现；保存授权记录、操作回放与异常处置过程
极高	双重用途研究、可自我改进的实验流程、高危试剂操作	默认不授予自主执行权	伦理与安全委员会前置审查	L4 与 L5 共同体验证；隔离环境运行，敏感步骤阻断并留痕

失败管理是科研闭环中被长期低估的一环。传统科研叙事习惯突出成功结果而略过失败过程，智能体则有条件系统记录无效假设、异常参数和负面结果。对失败作结构化编码，可以减少重复试错并帮助识别理论边界。但失败记录本身也受目标函数影响：如果系统只奖励快速获得显著结果，它可能倾向于规避高风险问题或淡化不利证据。科研智能平台的评价机制因此应当奖励真实记录、稳健验证和及时终止无效路径，而不是只奖励正向产出。

归结起来，人类科学家在第五范式中的核心作用有四项：提出值得研究的问题，设定可接受的风险，判断解释是否有意义，维护科学规范并承担公共责任。智能体扩大了可搜索的知识空间，却不能决定哪些知识值得追求。把价值判断明确保留在人类和科学共同体手中，是科研智能获得正当性的前提。

四、第五科研范式的核心特征与科研组织重构

（一）核心特征的三个层面

与此前四种范式相比，第五科研范式的特征可以归入三个层面：能力层面的自主性与闭环性，组织层面的协同性与平台性，制度层面的可审计性与制度依赖性。三组特征互为条件，任何一组单独出现都不足以支撑范式判断。

先看能力层面。自主性指科研智能体能在目标约束下自主完成多步骤任务，而不是被动等待逐条指令，这是它区别于普通 AI 工具的关键；但自主性必须与边界控制配套，在生物安全、网络安全、化学合成和双重用途技术领域尤其如此，不能允许智能体无限制探索。闭环性则指流程闭环而非单点智能化：只有当问题提出、假设生成、实验设计、执行验证、结果分析与知识表达形成连续反馈，科研智能才真正改变了知识生产方式。这两项特征是一体两面——没有闭环的自主只是失控，没有自主的闭环只是流水线。

组织层面的协同性，说的是人类科学家、科研智能体、实验平台与计算资源之间形成新型分工：人类负责价值判断、理论创造、伦理责任与战略选择，智能体负责扩展搜索空间、提高执行效率、组织复杂流程。平台性则意味着第五范式高度依赖基础设施，未来的科学竞争可能表现为科研智能平台之间的竞争——谁拥有更高质量的科学数据、更强的科学基础模型、更完善的自动化实验平台和更可信的治理体系，谁就可能获得更高的知识生产效率。

制度层面包含两项特征。可审计性要求科研过程记录、模型调用记录、数据来源记录、实验操作记录和结果复核机制齐备；没有可审计性，智能体可能制造大量难以追溯的错误知识，削弱共同体的信任基础。制度依赖性则提醒我们，第五范式不仅是技术范式，也是制度范式：它要求科研评价、知识产权、数据共享、学术伦理、人才培养和安全监管同步调整。制度若仍停留在论文数量评价和封闭实验室管理，第五范式很可能演变为低质量自动化产出，而不是科学发现的革命。

（二）科研组织从课题组中心走向平台协同

传统科研组织以首席研究者、课题组和固定实验室为中心，资源围绕人员编制、设备和经费展开。第五范式不会让课题组消失，但会使其与平台形成更紧密的依赖。一个团队能否快速调用跨学科文献、异构数据库、专业模型、远程仪器和自动化实验流程，可能比它是否独占全部设备更重要。组织能力因此从“拥有多少资源”部分转向“能否以可信方式编排资源”。

平台协同能降低中小团队进入复杂研究的门槛。过去只有大型机构才负担得起的计算、数据和仪器，可以通过共享平台按任务调用；标准化智能体还能把部分隐性操作转化为可复用流程。但平台化也可能削弱团队的独立性：如果关键模型、数据或实验接口由少数机构控制，研究者可能被迫接受平台预设的问题分类、评价指标和知识边界。科研平台既是公共基础设施，也可能成为新的学术资源分配者。

防止平台垄断，需要互操作和可迁移机制。研究任务、数据格式、模型接口和实验记录应尽量遵循开放标准，使团队能在不同平台之间转移流程并验证结果。由公共经费支持建设的科学数据和模型，应明确

公共访问、合理使用与长期保存责任；对商业平台，则需通过合同和监管防止数据锁定、结果不透明和服务中断影响科研连续性。

（三）科研劳动分工与专业角色变化

科研智能体将重新分配研究活动中的时间。文献整理、代码模板、常规统计、图表生成和格式审校可被大幅自动化，研究人员因而有机会把更多时间用于概念创造、实验判断和跨学科沟通。《自然》2023 年对 1,600 余名研究者的调查显示，超过半数受访者预期 AI 工具在未来十年会对本领域“非常重要”或“不可或缺”，与此同时，他们对研究方式被改变也表达了明确担忧^①。这种期待与担忧并存的态度值得重视：效率收益并不会自动转化为高质量研究。若评价制度仍以产出数量为主，节省下来的时间可能被用于生产更多同质化论文，反而加重研究人员的负担。技术变革必须与评价改革同步，否则前者只会放大后者的缺陷。

研究团队内部也会出现新的专业角色。除传统的理论研究者、实验研究者和数据分析人员外，还需要科研任务架构师、科学数据工程师、模型验证人员、智能实验工程师、研究审计人员与科研伦理专家。任务架构师负责把科学问题转化为可执行而不过度简化的任务，数据工程师维护数据质量和血缘，验证人员独立检查模型和实验结果，审计人员确保日志、权限和贡献记录完整。这些角色共同构成科研智能系统的可信运行基础，其中多数在现行岗位序列和职称体系中尚无对应位置。

青年科研人员的训练路径同样会改变。入门阶段的资料整理和基础计算，过去既是劳动，也是理解学科的过程。若这些任务全部交给智能体，研究者可能更快拿到结果，却失去形成问题意识和方法直觉的训练机会。教育设计不应简单取消基础训练，而应把智能体当作可解释的练习伙伴：要求学生检查证据、复现代码、比较不同方法，并说明为什么接受或拒绝机器给出的建议。经过这种训练，研究者才可能从工具使用者成长为负责任的科研组织者。

（四）知识公共性与平台权力

科学知识具有公共性，科研智能的基础资源却有集中化倾向。训练高质量科学模型需要大量数据、算力和专业工程投入，自主实验平台的建设成本同样不低，这使大型企业和少数科研机构更容易形成优势。缺乏公共投入的情况下，部分学科可能因商业回报不足而得不到模型和平台支持，边缘议题、小语种知识和地方经验也可能在训练数据中被系统性忽视。

国家因此需要把科研智能视为新型公共科研基础设施。公共性并不要求所有资源完全免费，而是要求核心能力具有可及性、连续性和透明性：平台应说明数据来源、模型适用范围和重要限制，为不同规模机构提供合理的访问层级，并保留研究者使用替代工具和导出研究记录的权利。对公共卫生、气候、农业和基础科学等具有显著社会价值的领域，公共平台应承担长期维护责任，避免完全依赖短期项目或商业市场。

平台的推荐和排序机制还会影响知识可见性。智能体在文献检索和问题生成中反复引用高频文献，可能强化主流观点并压缩非主流假设的空间。平台需要设计多样性和反事实搜索机制，主动呈现不同理论、负面结果和低频证据。科学创新常常来自对既有共识的有根据质疑，科研智能不应把“最常见的答案”误当作“最有价值的答案”。

（五）科研评价与贡献识别

智能体参与之后，论文署名已无法完整呈现知识生产过程。数据整理者、模型开发者、平台维护者、实验执行者和审计人员都可能作出关键贡献。国际学术出版界近年推行的贡献者角色分类，把作者贡献拆解为概念构思、方法设计、软件、验证、数据管理、资源提供等十余项，正是对这一问题的回应；在智能体参与的情境下，这类分类还需增补“任务编排”“模型验证”与“过程审计”等条目。智能体不能承担法律和伦理责任，但其使用范围、模型版本和关键输出应当透明披露。

^①Van Noorden R., Perkel J. M. AI and science: what 1,600 researchers think[J]. Nature, 2023, 621: 672–675. <https://doi.org/10.1038/d41586-023-02980-0>

评价对象也应从最终论文扩展到可复用的研究资产。高质量数据集、经过验证的代码、负面结果、实验协议、模型评测和开放工具都具有长期价值。若评价体系承认这些贡献,科研智能带来的效率才可能用于建设共同体资源;若仍只奖励论文数量,平台就会倾向于优化文本产出而非可靠发现。第五范式要求把复现率、证据质量、开放程度和社会价值纳入评价核心——这一点说起来简单,落实起来涉及职称、经费和学科评估的连锁调整,恐怕是全部变革中阻力最大的部分。

五、国家创新体系影响与我国发展路径

(一) 第五科研范式对国家创新体系的影响

国家创新能力不再只取决于科研人员数量、研发经费规模和论文产出,也越来越取决于科研智能基础设施的建设能力。这一影响至少体现在六个方面。

首先是科研组织方式。传统科研组织以课题组、实验室、项目和学科为中心;第五范式下,科研组织将更依赖跨学科平台、共享数据库、模型服务、自动化实验设施和智能体协作网络。团队能力不再是个人学术能力的简单加总,而是人、机、数、算、平台之间的组织能力。

其次是科研效率与创新速度。智能体可以显著降低文献综述、代码编写、实验设计、数据分析和论文初稿的成本,使研究者把更多精力投向高价值问题判断与理论创造。对药物发现、材料设计、能源技术和生命科学这类试错成本高、搜索空间巨大的领域,第五范式有可能大幅缩短研发周期。不过第三部分已经说明,效率提升与可靠性提升并非同一件事。

第三是国家科技竞争的分化形态。未来国家之间的科研能力差距,可能不只体现为人才和经费差距,也体现为算力供给、科学数据资源、科学基础模型、自主实验平台与科研智能治理能力的差距。拥有完整科研智能生态的国家,将在前沿科学和关键技术突破中占据更大主动权。

第四是科研评价体系承受的压力。当AI可以生成论文、代码和实验报告,以论文数量、引用次数和项目成果为核心的评价方式将更加脆弱。已有研究指出,人工智能正在广泛进入科学假设生成、实验设计、数据解释和知识发现过程^①。评价需要转向问题价值、数据质量、可重复性、真实贡献、开放资源和社会影响等维度,否则自动化论文生产只会加剧学术泡沫。

第五是知识产权与责任归属的复杂化。智能体参与发现后,成果归属如何界定?模型开发者、数据提供者、平台运营者、实验执行者与人类科学家之间如何分配权利和责任?当智能体给出错误结论或引发实验风险时,责任由谁承担?这些问题需要新的法律和治理框架,目前尚无成熟答案。

第六是科研安全与伦理治理的重要性上升。智能体可能被用于双重用途研究、敏感技术设计和高风险实验流程;缺乏权限管理、风险评估和审查机制,第五范式就会带来安全隐患。发展科研智能必须坚持促进创新与防范风险并重。

还有两点需要补充。其一,第五科研范式对不同学科和机构的影响并不均衡。数据标准化程度高、实验流程易自动化、评价指标相对明确的领域,可能较早形成智能闭环;依赖田野经验、历史语境、深度访谈和价值解释的领域,智能体更适合作为资料组织和比较分析工具。大型机构能率先建设平台,中小机构则可能通过共享服务获得能力,也可能因数据和人才不足而进一步边缘化。创新政策需要正视这种差异,避免把少数技术密集型领域的成功模式机械推广到所有学科。

其二,国家创新体系将面临“速度与方向”的双重问题。科研智能可以提高搜索和试错速度,但速度不等于方向正确。如果社会资源过度集中于容易量化、容易自动化且短期可见的任务,基础理论、长期观

^①Wang H., Fu T., Du Y., et al. Scientific discovery in the age of artificial intelligence[J]. Nature, 2023, 620: 47–60. <https://doi.org/10.1038/s41586-023-06221-2>; Stevens R., Taylor V., Nichols J., Maccabe A., Yelick K., Brown D. AI for Science: Report on the Department of Energy Town Halls on Artificial Intelligence for Science[R]. U.S. Department of Energy, 2020. <https://www.osti.gov/biblio/1604756>

测和高不确定性探索就可能受到挤压。政府与科学共同体需要保留对长期问题、公共问题和非共识问题的支持，使智能化效率服从于多元创新目标，而不是反过来由平台指标决定研究方向。

（二）我国发展第五科研范式的途径

面对第五科研范式，我国的机遇与挑战同样明显。机遇在于较完整的产业体系、快速发展的人工智能技术、丰富的科研场景和持续扩大的科技投入；挑战在于高质量科学数据、原创科学基础模型、自动化实验平台、跨学科人才和治理体系仍需加强。从支撑结构看，科研智能建设可以划分为基础设施、能力、组织与制度四层，各层之间存在自下而上的支撑关系与自上而下的约束关系（图4）。

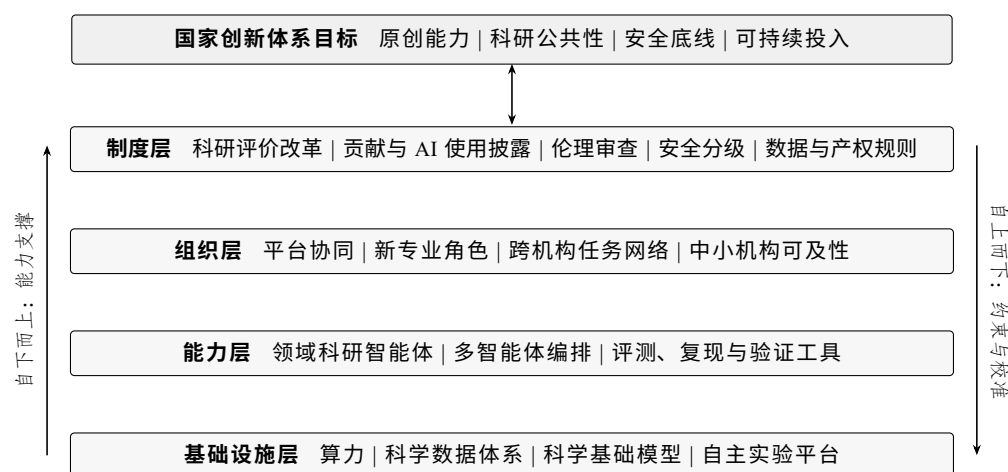


图4 科研智能的四层支撑结构与国家创新体系的接口

第一，建设国家级科研智能基础设施。应围绕生命科学、材料科学、能源科学、气候科学、先进制造和社会科学计算等重点领域，建设面向科学发现的算力、数据、模型与实验平台。科研智能不能只依赖通用大模型，还应发展面向学科知识、实验流程与科学推理的专业基础模型。

第二，构建高质量科学数据体系。我国自2018年施行《科学数据管理办法》以来，已陆续建成一批国家科学数据中心，但数据分散、格式不一、质量参差、共享动力不足的问题仍然存在。下一步的重点应从“建库”转向“可用”：推动数据标准化、可信采集、开放共享与安全使用，建立跨机构、跨学科、跨地区的协同机制，并把数据被复用的次数纳入建设单位的考核。

第三，发展自主可控的科研智能体。应支持高校、科研院所和科技企业联合开发面向重点学科的智能体，包括文献、实验设计、仿真、数据分析、论文审校与科研管理等类型，同时形成可解释、可审计、可追责的技术规范。

第四，建设自动化实验室与智能科研平台。第五科研范式的关键瓶颈，在于从“文本智能”走向“实验智能”。我国已有“机器化学家”这样的先行探索^①，下一步应在生物医药、化学合成、材料筛选、农业育种和工业测试等领域扩展智能实验平台，使AI生成的假设能够被快速、低成本地验证。

第五，改革科研评价制度。要防止科研智能变成“自动化论文生产机器”。评价体系应更重视原创问题、真实贡献、可重复性、数据和代码开放、平台建设、长期积累与社会价值。对AI参与生成的成果，应建立透明标注和贡献说明制度。2023年印发的《负责任研究行为规范指引》已划出几条红线：不得使用生成式人工智能直接生成申报材料，不得将其列为成果共同完成人，不得直接使用未经核实的由其生成的参考文献。这些禁止性规定是必要的起点，接下来要解决的是如何把原则要求转化为可核查的操作规程——披露采用什么格式、过程留痕保存多久、抽查由谁执行。

^①Zhu Q., Zhang F., Huang Y., et al. An all-round AI-Chemist with a scientific mind[J]. National Science Review, 2022, 9(10): nwac190. <https://doi.org/10.1093/nsr/nwac190>

第六，培养复合型科研智能人才。第五科研范式需要既懂学科知识，又懂人工智能、数据治理、实验自动化和科学伦理的人才。高校和科研机构应推动科学教育、工程教育与社会科学教育交叉融合，并为前述新增的科研角色设置相应的岗位序列与晋升通道。

第七，建立科研智能治理体系。应围绕数据安全、模型安全、实验安全、知识产权、责任归属、学术诚信和双重用途风险，建立分级分类治理机制。高风险领域设置人工审查、权限控制与过程记录要求；低风险领域则应减少不必要的审批，鼓励开放试验。

（三）分阶段建设科研智能体系

上述七条路径不可能同时推进，需要有先后。近期应优先补齐可直接复用的基础能力，中期推动数据、模型与实验平台的实质连接，长期则形成开放而有韧性的国家科研智能生态。各阶段的重点任务、主要风险与评估指标可参见表4。

表4 我国科研智能体系分阶段建设的重点任务与评估指标

阶段	时间跨度	重点任务	主要风险	评估指标
近期	约1—3年	开展重点学科科学数据普查；在文献检索、代码验证、数据分析等低风险环节建设领域智能体；统一研究日志、模型版本与贡献披露规范；在高校和院所建设小规模试点环境	以演示效果替代可靠性；重复建设通用模型；试点缺乏可比基线	数据资产清单完整度；低风险任务的错误率基线；日志与披露规范的采纳率
中期	约3—7年	在材料、生物医药、农业、能源、先进制造等领域建设跨机构智能实验网络；同步形成模型评测、数据治理、伦理审查与安全响应机制；以开放接口支持第三方开发专业智能体	平台锁定与数据垄断；实验安全事件；评价制度滞后于产能扩张	关键研究周期缩短幅度；跨机构与跨平台复现率；平台对中小机构的服务覆盖面
长期	7年以上	形成多中心、可互联验证的国家科研智能生态；基础算力与通用能力共享，专业数据、评价标准与实验规范保持多样性	单一技术路线形成系统性风险；研究方向被平台指标牵引	生态内可替代供应者数量；负面结果与失败记录的完整度；欠发达地区科研可及性的改善程度

需要强调的是，近期目标不应是追求“全自动科学家”的宣传效果，而是形成可靠、可复制的基础模块。评估指标的设计尤其重要：除模型准确率和任务完成率外，还应考察数据质量改善、失败记录完整度、人工复核工作量、跨平台复现率、科研人员时间配置、平台服务覆盖面以及安全事件响应能力。对公共投入建设的平台，更应考察其是否改善了中小机构和欠发达地区的科研可及性。只有把效率、质量、公平与安全纳入同一评价框架，科研智能建设才不至于滑向单纯追求规模和演示效果。

（四）社会科学在科研智能建设中的作用

第五科研范式不仅需要自然科学和工程技术，也需要社会科学参与。科学社会学可以研究智能体如何改变共同体规范、声望结构和知识权威；经济学可以分析平台投入、数据定价和创新激励；管理学可以研究人机团队、跨机构协作和项目治理；法学处理数据权利、知识产权和责任分配；伦理学则需要界定可接受风险与人类监督原则。缺少社会科学参与，技术制度可能在实践中放大不平等，或形成难以纠正的路径依赖。

社会科学研究本身也在被科研智能改变。大规模文本、政策文件、企业数据和行为记录可以被智能体协同处理，定性与定量方法之间的连接更为紧密。但社会科学的对象具有反身性，研究结论可能改变被研究者的行为，数据也常受制度语境和权力关系影响。因此，社会科学智能体不能只追求预测准确率，还需说明概念如何定义、样本如何形成、谁被排除在数据之外，以及结论可能产生何种社会后果。

我国可以建设面向社会科学的公共语料、政策知识库和可审计分析平台，为重大政策评估、社会风险

研判和公共治理研究提供支持。此类平台应特别重视个人信息保护、群体公平与解释责任，避免把概率性结果直接用于高影响决策。社会科学的加入，能够把第五科研范式从技术效率框架扩展为兼顾制度、文化与公共价值的创新框架。

六、理论边界、治理风险与结论展望

（一）需要警惕的三种误区

概念一旦流行，边界就容易失守。就“第五科研范式”而言，至少有三种误区需要预先排除。

技术决定论是第一种。第五科研范式不会由某个大模型自动带来，也不会仅靠算法进步自然实现，它依赖科学数据、实验平台、科研制度、伦理规范与人类科学家的共同作用。缺少制度和组织创新，技术能力可能转化为低质量产出甚至科研风险——本文开篇的案例已经说明，一条自动化程度很高的链条，仍可能在证据判定这一环节失守。

替代论是第二种。科研智能体不会简单替代科学家。科学研究不只是信息处理过程，也是问题选择、意义判断、理论想象和责任承担的过程。人工智能可以提高搜索和执行效率，但高价值科学问题的提出、研究方向的选择和结果的价值判断，仍要由人类科学共同体完成。

泛化论是第三种，也是最常见的一种。并非所有使用 AI 的科研活动都属于第五科研范式：用机器学习处理数据，可能仍属第四范式；用大模型辅助写作，更不足以构成第五范式。只有当科研智能体进入多环节闭环流程，并改变了知识生产的组织方式，才谈得上第五科研范式。第二部分给出的三条最低判定标准，正是为了抵御这种概念通胀。

澄清这三点，是为了让第五科研范式成为一个可以被检验、也可以被证伪的分析概念，而不是一个流行技术标签。

（二）可信性、可解释性与可重复性

可信性是科研智能区别于一般内容生成的首要要求。大语言模型能够生成语义连贯的解释，但语义连贯并不保证事实正确、引用真实或因果关系成立。一旦智能体获得工具调用和实验控制能力，错误就会从文本层扩展到数据、代码和物理操作层：一个早期的错误假设会影响后续检索范围、实验参数和结果解释，并在闭环中被不断强化。系统因此必须能够识别不确定性、主动寻找反证，并在证据不足时停止而非补全答案。

更隐蔽的风险发生在研究者一侧。Messori 与 Crockett 在《自然》上提出，科研中的 AI 常被想象为四种角色——提供答案的“先知”、替代人类被试的“替身”、处理海量数据的“量化者”、裁定证据的“仲裁者”——每一种想象都可能诱发相应的认知幻觉：把机器给出的关联当作对机制的理解，把工具能触及的范围当作问题的全部，把统计过程当作价值中立。当整个共同体依赖同一批模型和同一套数据时，这些幻觉会汇合成“认知单一化”，表面上产出更多，实际探索的空间反而收窄^①。这一批评之所以重要，是因为它指向的不是模型缺陷，而是使用者的判断结构。

可解释性需要区分层次。对研究者而言，需要理解模型为何选择某一假设、数据或方法；对审计者而言，需要追踪调用了哪些工具、使用了哪个版本和参数；对公众和管理者而言，则需要知道结果的适用范围、主要限制与潜在风险。并非所有模型内部机制都能完全透明，但关键科研决策必须形成可理解的理由链和证据链，否则所谓解释只是事后生成的合理化文本。

可重复性问题的第五范式中获得了新的技术变量。需要提醒的是，复现困难并非人工智能带来的新病症：2016 年《自然》对 1,500 余名研究者的调查显示，超过七成受访者曾尝试复现他人实验而未能成功，

^①Messori L., Crockett M. J. Artificial intelligence and illusions of understanding in scientific research[J]. Nature, 2024, 627: 49–58. <https://doi.org/10.1038/s41586-024-07146-0>

半数以上连自己的实验也无法复现^①。智能体的加入把这一旧问题推到了新的量级——同一任务在不同模型版本、不同提示、不同检索时间和不同平台权限下可能产生不同结果。传统论文只报告方法和参数已不足以支持复现，还需保存模型标识、提示结构、工具清单、数据快照、随机种子、执行环境与人工干预节点。对无法公开的敏感数据，也应通过可信计算环境或独立第三方验证提供复核渠道。

评测方式也要相应改变。验证科研智能体不能只使用标准答案式测试，真实科研问题往往没有预设答案；更值得考察的是系统能否提出可证伪假设、识别证据冲突、在实验失败后合理调整、区分相关与因果，并承认超出能力边界。评测应结合历史案例回放、开放问题挑战、跨平台复现与专家盲审，在不同任务和时间尺度上长期观察其稳定性。一次成功的演示，不能替代持续评测。

（三）责任、伦理、安全与治理原则

智能体参与知识生产后，责任链条容易被技术复杂性稀释：研究者可能把错误归因于模型，平台可能把风险归因于用户，模型开发者又强调自己只提供通用工具。治理制度必须坚持责任不可外包的原则——提出研究目标并授权高风险操作的人类研究者承担直接责任，所属机构承担管理和监督责任，平台与模型提供者对已知缺陷、权限设计和安全措施承担相应责任。完整日志是划分责任的证据，却不能替代事前审查。

知识产权需要在激励创新与促进共享之间取得平衡。智能体可能同时使用公共文献、机构数据、商业模型和自动化实验平台，成果难以归属于单一主体。合同和政策应明确训练数据、生成代码、实验数据、模型改进与最终发现的权利边界，保障数据提供者和基础设施维护者的合理回报；同时应防止对通用科学事实、公共数据和基本研究流程的过度占有，维护科学知识的公共领域。

学术诚信治理面临新的失范形式。除虚构引用、代写论文和隐瞒 AI 使用外，还可能出现自动化择优呈报有利结果、批量生成低质量假设、操纵同行评议、重复使用未经验证的数据，以及借智能体掩盖真实贡献等问题。期刊、基金机构和科研单位应要求披露 AI 参与的实质环节，而不是笼统标注“使用了 AI”。有效的披露至少应说明四项：所用模型及版本，用于哪些环节，人工核查的方式与范围，以及是否影响了数据、方法或结论。我国 2023 年出台的科技伦理审查制度已把涉及人工智能的研究活动纳入审查范围，接下来的问题是如何把审查要求落到具体的操作规程和记录格式上。

安全治理需要按领域和能力分级。普通文献整理与高风险生物实验不能适用同一套规则，治理对象也应同时覆盖输入数据、模型能力、可调用工具与可能产生的结果。高风险任务可采取身份认证、最小权限、双人授权、隔离环境、敏感步骤阻断与事后审计等措施；低风险任务则应减少不必要审批，支持开放创新。分级治理既避免“一禁了之”压制合理研究，也避免用创新名义绕开必要控制。

治理还须具备动态性。模型能力、科研场景和风险形态持续变化，固定规则容易滞后。机构可以建立由领域专家、数据与安全人员、伦理和法律专家共同参与的持续评估机制，定期更新任务分类、权限清单和应急预案；发生错误或安全事件后，应鼓励及时报告和跨机构共享教训，避免因惩罚过重而导致风险被隐瞒。第五范式的治理目标，是建立一个能够学习和纠错的制度闭环，与技术闭环相互制衡。

（四）结论与研究展望

科研范式的演进反映了人类认识世界方式的变化：经验科学依赖观察实验，理论科学依赖数学建模，计算科学依赖数值模拟，数据密集型科学依赖大规模数据发现。今天，科研智能体、AI Scientist、自主实验平台与科学基础模型的发展，正在把科学研究推入新的阶段。本文把这一阶段概括为第五科研范式，并给出了三条判定标准，以免它变成一个可以贴在任何 AI 应用上的标签。

第五科研范式的实质，是科研智能体驱动的知识生产方式变革。它不是第四范式的简单延伸，而是研究主体、研究流程、组织结构、验证机制与治理制度的系统重构；其核心机制是以人机协同为基础的智能

^①Baker M. 1,500 scientists lift the lid on reproducibility[J]. Nature, 2016, 533: 452–454. <https://doi.org/10.1038/533452a>

闭环，其核心基础设施是算力、数据、模型与实验平台，其核心目标是提高科学发现效率并扩展人类知识的边界。但本文的分析同时表明，自动化程度与知识可靠性并不同向变化：链条越长、速度越快，未被察觉的错误传播得越远。这正是把可审计性写入范式定义的理由。

对我国而言，第五科研范式既是前沿科技发展的机会，也是国家创新体系建设的新命题。谁能率先建立可信、开放、可控、可审计的科研智能体系，谁就可能在未来的科学发现和关键技术突破中占据主动。发展科研智能，不能只关注模型能力和工具应用，还应同步推进科学数据体系、科研智能平台、自动化实验基础设施、复合型人才培育与科研治理制度建设。

后续研究至少有三个方向值得展开：从科学哲学层面研究智能体参与知识生产后的证据、解释与因果关系；从创新体系层面研究科研智能平台如何重塑国家、企业、高校与科研机构之间的关系；从治理层面研究科研智能的伦理、安全、责任与评价机制。这三个方向都需要经验材料的支撑，而目前可用的案例仍然稀少。

最后回到开篇的例子。A-Lab 的故事之所以值得反复引用，不在于它证明了机器能否做科学，而在于它展示了一个更基本的事实：自动化系统给出的结论，最终仍要通过人的重新检验才能进入知识体系；而使这次检验成为可能的，是被完整保存下来的原始数据。第五科研范式能走多远，很大程度上取决于我们是否愿意为这类看似琐碎的记录付出代价。

引用本文 王爱清. 从第四范式到第五科研范式：科研智能体驱动的知识生产方式变革 [J]. 社会科学理论与实践, 2026, 8(4): 24–42. <https://doi.org/10.6914/tpss.080404>.

Cite This Article Aiqing WANG. (2026). From the Fourth to the Fifth Scientific Paradigm: Transformation of Knowledge Production Driven by Scientific Research Agents. *Theory and Practice of Social Science (Shèhuì kēxué lǐlùn yǔ shíjiàn)*, 8(4): 24–42. <https://doi.org/10.6914/tpss.080404>

From the Fourth to the Fifth Scientific Paradigm: Transformation of Knowledge Production Driven by Scientific Research Agents

Aiqing WANG

Ningde Normal University, Ningde 352100, China

Abstract Empirical, theoretical, computational and data-intensive science together form the conventional framework for understanding how modern research has changed. With the advance of large language models, AI for Science, autonomous laboratories and multi-agent research systems, artificial intelligence has moved beyond data processing and simulation acceleration into problem formulation, literature understanding, hypothesis generation, experimental design, code execution, data analysis, paper writing and peer review. Building on the four-paradigm framework, this paper defines a fifth scientific paradigm: one in which scientific research agents serve as the core organizing unit, supported by computing power, scientific data, foundation models, autonomous experimental platforms and human-AI collaboration, and in which discovery is organized in an autonomous, closed-loop, platform-based and auditable manner. What separates it from the fourth paradigm is not more data or stronger algorithms, but the simultaneous restructuring of research subjects, workflows, validation mechanisms and organizational forms. To prevent conceptual inflation, three minimum criteria are proposed: agents must enter at least two mutually reinforcing core research stages; the system must revise its plan on the basis of intermediate evidence and connect to computational or experimental verification; and the whole process must carry auditable human oversight and accountability. The paper analyses the enabling conditions and operating mechanisms of this paradigm, examines its consequences for research organization, division of labour, evaluation and national innovation systems, and draws

on the correction issued to the A-Lab autonomous synthesis paper and third-party evaluations of the AI Scientist to show that greater automation does not move in step with greater reliability. Developing scientific intelligence therefore requires not only models and computing power, but also scientific data systems, experimental interfaces, reproducibility infrastructure and rules of accountability.

Keywords Fifth scientific paradigm; scientific research agent; AI Scientist; AI for Science; knowledge production; national innovation system

〔责任编辑：黄欣 cpcl@cpcl.hk〕