科技评论

AI for Science 的前景、瓶颈和风险

李国杰

摘要 基于对各种人工智能(artificial intelligence, AI)大模型和相关文献提供的预判信息的综合分析,对未来 5 年科学智能(AI for Science, AI4S)的发展前景给出谨慎乐观的预判。在分析 The AI Scientist-v2 平台自动生成顶级会议论文后指出,实现科研闭环全自动化有本质性困难,验证不完备和科研目标漂移是制约 AI4S 发展的障碍。从计算理论的角度来看,AI4S 验证瓶颈的根源在于科学问题的半可判定性,可验证性的边界就是自动化的边界。根据可验证性,科学技术问题可分成 2 大类问题:R1 问题和 R2 问题。AI4S 的风险在于把本该按 R2 治理的问题当成 R1 问题来处理。本文指出,AI4S 可能比通用人工智能(artificial general intelligence, AGI)安全风险更紧迫,AI4S 的风险不是“AI 变聪明”,而是“人类失去退出权”。

关键词 AI for Science (AI4S);验证不完备;目标漂移;半可判定性;R1 问题;R2 问题;AI 治理

1 对未来 5 年科学智能发展前景的预判

过去 3 年中,科学智能(AI for Science, AI4S)已取得里程碑式成果,如 Google 公司 DeepMind 团队在生物、材料领域推出的 AlphaFold 3 和 GNoME 平台,国内深势科技团队推出的通用人工智能(artificial general intelligence, AGI)分子模拟平台 DPA-3 等。科研实践已经证明:人工智能(artificial intelligence, AI)在可形式化、可验证、高度重复的科学任务上,效率远超人类。这不是单点技术突破,而是科研范式的根本性转变,从“人类试错”到“AI 预测+验证+人类判断”。算力成本的指数级下降、科学大模型走向通专融合、全球 AI 投入增加等因素,将促使 AI4S 进入高速发展期。学术界对 AI4S 的发展前景预判有明显的分歧,有学者认为 21 世纪内 AI 会产生 1 万个爱因斯坦,也有学者认为 AI4S 只是个辅助工具,不可能具有原始创新能力。基于对各种 AI 大模型和相关文献提供的预判信息的综合分析,对未来 5 年 AI4S 的发展前景给出谨慎乐观的预判,供读者参考。

AI4S 最有可能取得重大进展的领域包括药物研发、材料科学、工程设计、地球科学和数学等。在药物研发中,AI 将显著加速靶点发现、虚拟筛选、先导化合物优化和早期安全性评估。在乐观情景下,2030 年前后可能出现首批以 AI 为核心参与发现和设计的小分子药物获批上市。AI 有望将部分药物发现阶段从传统的 5~6 年压缩到 2~3 年,发现阶段的研发成本可降低 40%~60%。完整药物研发仍将受到临床试验、生物复杂性和监管要求等刚性约束,通常仍需 8~12 年。到 2030 年,相当比例的新药研发项目会使用 AI 辅助工具,但获批新药中由 AI 主导发现和设计的比例可能仍有限,到 2035 年前后,这一比例有望提升到 20% 以上。

在材料科学中,DeepMind 的

中国科学院计算技术研究所,
北京 100190

收稿日期:2026-05-25;修回日期:2026-07-04
作者简介:李国杰,研究员,中国工程院院士,研究方向为计算机体系结构、并行算法、人工智能、大数据、计算机网络、信息技术发展战略等,电子信箱:lig@ict.ac.cn
引用格式:李国杰. AI for Science 的前景、瓶颈和风险[J]. 科技导报, 2026, 44(14): 17-22; doi:10.3981/j.issn.1000-7857.2026.05.00106

GNoME 平台已预测约 38 万种潜在稳定晶体材料, A-Lab 等自动化实验平台也展示了快速合成和验证新材料的能力。但从“预测稳定”到“可制造、可表征、可应用”仍存在巨大鸿沟。到 2030 年前后, AI 预测材料中可能有数百到上千种获得不同程度的实验验证, 但真正进入中试或产业化的将主要集中在电池电解质、固态离子导体、催化剂、热电材料和特殊功能材料等高价值领域。AI 有望把部分材料候选发现周期压缩到 2~3 年, 但包含验证、中试、可靠性测试和产业认证的完整周期通常仍需 5~10 年。2035 年前后, “按需设计材料”可能首先在航天合金、量子器件材料、半导体材料和小批量高价值功能材料中初步实现; 钢铁、水泥等大宗工业材料仍将主要依赖传统工艺演进和长期工程验证。

在工程设计特别是芯片设计中, AI 辅助电子设计自动化(electronic design automation, EDA)已从概念走向实践。AI 将在寄存器传输级(register transfer level, RTL)生成、验证、PPA(功耗、性能、面积)优化、布局布线和设计空间搜索等环节显著提高效率。未来 5 年, AI 有望成为芯片设计流程中的常规工具, 在部分模块和验证任务中实现数倍效率提升; 但对完整系统级芯片(system on chip, SoC)项目而言, 设计周期缩短 30%~50% 应视为乐观区间。2030 年前, 先进芯片设计仍将以“AI 辅助+人类架构师和验证团队决策”的混合模式为主。

AI 在天气预报、气候模拟、地震/海啸预警、碳循环和资源勘探中会有重大进展, 但科研效率的提高不如药物和材料领域那样好量化, 因为地球系统不可控、验证周期长、极端事件样本稀缺。AI 在形式化证明、猜想生成等方面会快速进步。由于 AI 生成数学证明的速度远远超过人类消化证明的速度, 数学正从“证明稀缺时代”转入“证明过剩时代”, 数学家的价值将从“第一个写出证明”转向选择问题、验证证明、消化证明、理解意义和判断重要性[1]。

对 AI4S 持悲观态度的学者认为, AI 是语法层级的工具, 不能产生规则和语义, 不能提出新的科学问题, 常常用哥德尔不完备定理为依据来论证 AI 的“不可能”, 这是一种误解。哥德尔定理限制的是“形式系统的证明能力”, 不是“探索或创造能力”。科学不完全是形式系统, 科学的真理性依赖于实验和观测, 不全是形式证明。实际上, AI 自动生成规则与语义的潜力远超大多数人的想象。语义不一定由人类预定义, 可以从 AI 系统的交互、学习和预测中生长出来。现代 AI 系统(尤其是具身智能、多智能体、元学习架构)具备修改自身假设空间、引入新变量、构建新表示的能力。

2 实现科研闭环全自动化有本质性困难

由日本 Sakana AI 公司联合牛津大学、加州大学伯克利分校共同研制的 AI 科学家(The AI Scientist-v2)平台生成的论文, 曾提交至第十三届国际表征学习大会(The Thirteenth International Conference on Learning Representations, ICLR 2025)的 ICBINB 研讨会, 接受双盲同行评审。该系统提交的 3 篇全 AI 生成论文中, 有 1 篇获得平均评审分数 6.33, 高于该研讨会的平均接受阈值, 但离主会论文接受门槛还有较大差距(按照预设实验协议, 该论文在评审后被撤回)。2026 年, 相关研究论文《Towards end-to-end automation of AI research》发表于《Nature》, 学术界认可这是科研流程自动化的重要里程碑[2]。

The AI Scientist 需要用户提供种子想法, 无法对其自身结果进行批判性评价, 还不是一个独立的科学主体, 但它提出了若干发人深省的问题: AI4S 能实现科研完全自动化吗? 人类如果以“前现代式”的态度完全相信 AI, 会不会像基辛格所言, 将出现“人类认知大逆转”? 人类目前是不是站在“黑暗启蒙”的悬崖边上, 应如何应对 AI4S 带来的潜在风险? [3]

科学活动并不是一个单一的技术过程, 至少包含 2 个本质不同的层面: 一是“执行层”, 即在既定问题、方法和评价标准下生成假设、设计实验、运行仿真、分析数据并检验结果; 二是“判断层”, 即决定哪些问题值得研究、哪些异常值得追踪、哪些结果具有科学意义, 以及何时需要改变研究目标和评价标准。人工智能正在迅速掌握执行层中的大量环节, 但对于判断层, 人工智能仍面临更深层的限制。

科研活动本质上是根植于开放世界的认知实践, 其对象、背

景、历史语境、社会需求和价值维度不能被任何封闭的形式系统完全表示。若要实现真正意义上的科研完全自动化,机器不仅要能够生成理论假设和检验结果,还必须能够自主评价科学意义,并在新的发现和失败面前自我修正评价准则。问题在于,这种“生成假设—评价意义—修正规则”的自指闭环无法被还原为一个封闭、稳定、可事前完备验证的算法过程。在图灵计算框架下,系统可以扩展搜索空间、改进模型和优化局部目标,但不能保证自身在开放语义空间中的评价准则始终正确、稳定且具有正当性。

因此,科研全自动化的障碍并不只是当前模型能力、算力或实验机器人水平不足,而是涉及计算系统与科学共同体之间的层级差异。现有主流 AI 系统本质上仍是外源目标驱动的优化体系,即使未来 AI 具备更强的目标更新和自我改进能力,其评价准则的正当性、科学意义的判定以及风险责任的承担,仍不能由系统内部完全闭合地解决。把人类科学共同体中的主体性意义建构、价值判断和制度性责任,直接等同于算法系统中的目标函数优化,就混淆了价值活动与功能运算的层级边界,因而是一种范畴错误。AI 可以使科研执行层高度自动化,并在若干封闭或半封闭领域实现端到端自动科研;但在开放世界意义上,科学研究难以实现完全机器自治。未来更可能出现的是“AI 主导执行过程、人类科学共同体保留最终判断权”的混合科研范式,而不是取消人类评价权的全自动科研体系。

3 制约 AI4S 发展的 2 大障碍

除了缺乏对科学发现和技术发明有价值的高质量数据、算力不足等大家熟知的瓶颈以外,制约 AI4S 发展的障碍还包括验证不完备和科研目标漂移。

3.1 AI4S 的“验证瓶颈”

AI 生成的可能途径太多,验证跟不上。这一问题被称为“AI4S 的验证瓶颈”或“假说爆炸危机”。AI4S 最大的制约因素不是算法或算力,而是人类验证能力的线性增长与 AI 生成能力指数增长之间的剪刀差。AI 产生海量“看似合理”的假说,绝大多数无法被验证,导致科研资源浪费在低优先级候选上。“AI 幻觉”也常常被误认为是“发现”。

以新材料发现为例,经过第一性原理计算(这也是 AI4S 的贡献),筛选出热力学稳定的候选材料有 38.1 万种,目前只有 736 种材料与独立实验结构匹配,匹配率仅为 0.19%。这一案例说明,提高验证率是发现有价值的新材料的关键环节。新药研制的瓶颈也体现在临床实验验证,不管用不用 AI 技术,批准一款新药正式上市还需要 10 年左右时间。

3.2 目标漂移是 AI4S 的主要障碍之一

AI4S 容易出现目标漂移,这是因为科学目标天然做不到完全量化,“理解、发现、解释”具有高维、模糊和依赖语境的特征,不存在完美单一指标。代理指标只与真实目标弱相关,密度泛函理论(density functional theory, DFT)计算确认形成能为负的候选材料不一定可合成;药物小分子与生物靶点的高亲和力不保证能形成新药。在强优化压力下,AI 大模型往往能在高维空间找到人类想不到的指标漏洞,快速获得“高分”,但偏离了原始的科学意图。新材料研究的真实目标是发现可实验合成、性能优异的新晶体,实际上的代理指标变成 DFT 形成能、结构相似度、对称性、晶格常数等,一味优化某个代理指标就会出现目标走样的所谓“Goodhart 现象”。在药物设计中,过度优化单一指标(如亲和力),可能生成有致命毒性的分子。

4 从可验证性的角度审视 AI4S

4.1 半可判定问题

从计算理论的角度看,AI4S 验证瓶颈的根源在于科学问题的半可判定性。所谓半可判定问题是指一个问题有解时能在有限步内找到解,无解或无最优解时,永远无法在有限时间内证明“不存在”。换句话说,“半可判定问题”是“正例可验证、负例不封闭”的问题。在实际生活中,最明显的半可判定问题是无人驾驶汽车的安全问题,出了安全事故一定能发现,但至今没有出安全事故,不能通过事前测试验证保证以后在任何环境下永远不出安全事故。原因是无法事前枚举今后可能遇到的所有场景,无法做完备测试验证。大量的 AI4S 科研任务,和无人驾驶一样,事前无法完备验证,只能边运行边验证。AI4S 不可完备验证不是技术能力问题,而是原

理上不可能,通过模型升级、算力提升、数据规模增大不可能突破半可判定性与不可完备验证的理论边界。

半可判定问题的解决思路,不是把“半可判定”问题整体上变成“完全可判定”的问题,而是在有界约束、有限搜索域、有限精度、有限时间内,实现工程上可终止、可验证、可收敛。第一层,人为有界截断无穷空间,将无穷搜索空间强行切成有限有界闭集;第二层,通过设计粗筛→中筛→精算→实验 4 级漏斗,分层多级做判定,把半可判定问题拆成多层局部可判定的子问题;第三层,放弃“全局最优”,改为“满意解收敛”;第四层,主动学习+实验闭环,用物理真值补全逻辑判定。

4.2 可验证性的边界

可验证性的边界就是自动化的边界,长远来讲,一切可验证的问题都将被 AI(机器自动化)接管,机器会吞噬所有“可验证语义空间”。现在,AI 正在扩展可验证边界,把“科研意义”压缩为“预测能力”。相当一类根植于开放世界的科学研究(如新材料发现、分子设计、复杂系统模拟等),其核心问题天然具有半可判定特征,因此不存在能够“全覆盖、无遗漏、事前保证”的完备验证体系。在这类问题中,系统的可验证性在原则上是局部的,而不可验证边界是结构性的永久存在。人工智能倾向于通过批量生成、快速计算与局部校验,同时扩展可验证的范围并指数级制造新的不可验证边界;而人类则通过制度规范、学术共识、方法论革新、实验实践、范式转换与意义判断,系统性地管理无法被完备证明对错的部分。换言之,AI 的核心价值是高效拓展“可被验证正确的疆域”,而人类的不可替代作用是治理随 AI 能力同步扩张的不可验证性空间,并决定科学探索的方向与价值边界。这一本质分工决定了 AI4S 时代的科研治理核心,是强化人类对不可验证性空间的治理能力,而非追求完全机器自治。

4.3 AI4S 涉及 2 大类问题:R1 问题和 R2 问题

针对科学研究的开放世界属性与 AI4S 特有的半可判定性,将 AI4S 涉及的科学技术问题分成 2 大类:R1 问题,指在当前公认科学范式内,其物理正确性、成功条件与安全性能够通过有限步骤、可重复、可形式化的方法,在进入下一个不可逆决策节点前被充分验证的问题;R2 问题,指在同一范式内,上述关键性质不存在有限步骤的完备封闭验证方法,只能依赖局部测试、统计证据、同行共识与持续事后纠偏,且验证结果具有范式依赖性、时间局限性与概率性。从逻辑复杂性的角度看,R1 类问题是可判定问题,R2 类问题是半可判定问题。

传统计算机科学问题基本上属于 R1 问题,而绝大多数 AI4S 问题属于 R2 问题。这一本质差异决定了传统软件工程验证范式无法直接迁移,必须建立以人类终审为核心的分层治理体系。R1 问题的核心特征是存在客观的验证标准,验证失败的成本可控,典型的场景包括已知分子的量子力学性质计算、确定性物理过程的数值模拟、判定材料是否违反已知化学成键规则等;R2 问题的核心特征是无封闭的验证边界,AI 生成能力会指数级放大其验证规模与风险,典型场景包括新材料的长期稳定性预测、新药的临床前毒副作用评估、复杂气候系统的长期演化模拟等。该分类以“不可逆决策节点”替代传统工程领域的“部署前/后”单一时间分界,承认科学验证的相对性与范式依赖性,为 AI4S 时代的人机分工与风险治理提供了可操作的底层框架。R1 问题可实现全流程自动化,大幅提升研发效率;而在所有 R2 问题的关键决策节点,人类的经验判断、风险评估与价值权衡具有不可替代的核心作用。

R1 科研的特征是问题定义与验证标准在求解之前已被封闭给定;R2 科研的特征是问题定义、求解过程与验证标准在系统运行过程中共同生成,因此不存在完备且事前可定义的验证体系。不能再用 R1 科研的规范来评价 R2 科研,否则会出现 2 种极端:一是“不可解释就不可信”,导致错误地否定 AI 生成的有价值成果;二是“看起来有效就接受”,导致高风险。我们必须引入新的验证机制:审计、回滚、门控、多主体评估体。

R1 科研解决已定义的问题,R2 科研要重新定义“什么是问题、什么是解释、什么是正确”。R1 问题的可接受性逻辑是:证明/测试/验收→通过→上线。R2 问题的可接受性逻辑是:局部证据→受限部署→持续监测→动态扩展或收缩。R1 是验证先于部署;R2 是治理伴

随部署。R2 系统的伦理核心不是“零风险”,而是“不失控”,不是“证明它绝对安全”,而是“证明它比现状更好,并且还能被管住”[4]。

AI4S 的核心挑战,往往不只是“能不能生成”,而是“生成出来的行为能否在行动前被充分验证”,这正是区分 R1/R2 问题的动机。验证的对象往往不是静止系统,而是系统—环境共同演化体,这会让很多问题从表面 R1 滑向实质 R2。AI4S 问题不是天生 R2,而是研究阶段像 R1,部署阶段滑向 R2。AI 是一种持续扩大 R2 问题域的技术。AI 之所以有用,正因为它进入了传统验证体系覆盖不足的区域。AI4S 的风险不在于它“本质上不可验证”,而在于人们很容易把本该按 R2 治理的问题,当成 R1 问题来开发、验收和部署。

5 AI4S 可能比通用人工智能安全风险更紧迫

很多人认为 AI4S 是科研领域的事,没有多大安全风险,这是认识上的误区。其实,AI4S 很可能先于 AGI 进入“不可逆阶段”。AGI 的不可逆性目前还是理论上的,是未来式的风险,在现实中仍然可以被“延缓、冻结、限制部署”。AI4S 与 AI+*X* 的不可逆性是现在进行时。AI4S 和 AI+*X* 一旦失控,会立刻发生 3 件事:(1) 科研流程改变。如果由模型决定“做不做实验”,人类只保留验证权,这就是不可逆的科研流程锁定。(2) 理论地位发生改变。从“以理论为决策依据”变成“以有效性为决策依据”。(3) 系统依赖形成。一旦材料、药物、能源、工程安全开始依赖 AI 模型输出,就无法回到“不用 AI 的世界”。真正危险的不是“AI 变聪明”,而是“人类失去退出权”。

AI4S 并不需要自我意识,只要 AI 在材料、能源、生物医药、工程物理等关键领域中,持续给出“可复现、可工程化、可规模化”的有效结果,它就会迅速成为科研和工程决策链条中的核心环节。这种风险并非来自智能失控,而是来自效率优势所带来的路径锁定。由于 AI 系统在原则上无法形式化验证其长期正确性,风险已经从可判定区间进入不可判定区间,这一跃迁往往以“工程成功”的形式悄然完成。一旦关键产业、安全工程或国家级科研体系开始依赖 AI 模型输出,退出成本将迅速高于继续使用的成本,治理选项会急剧收缩。中国在 AI4S 领域具备率先突破的制度与工程条件,“先使用、后解释”的技术文化盛行,导致 AI4S 更可能在中国率先实现系统性落地。这既是机遇,也意味着风险更早显现。若缺乏明确的分级治理与退出机制,AI4S 可能在未被识别为“风险技术”的情况下,完成向不可判定决策系统的迁移。因此,在人工智能治理框架中,应该将 AI4S 作为独立重点领域进行前瞻性风险评估和分类管理,防止在不自觉中形成难以调整的技术路径依赖。凡是涉及重大安全、生命健康、国家战略或长期制度安排的科学与工程决策,均不得完全由 AI4S 系统主导。任何情况下,均须明确保留人类决策权、责任主体和可回退路径。

为克服 AI4S 的验证瓶颈,可以在以下 6 方面做出努力:(1) 全面推广“生成即验证”范式,边生成边验证,从源头消除逻辑缺陷;(2) 集中力量建设 5~10 个全流程无人化材料合成、药物筛选与芯片设计等工程技术测试平台;(3) 建立基于 R1/R2 分类的分层验证标准,明确不同风险等级问题的验证强度、方法与决策权限;(4) 改革科研评价与激励机制,将验证工作纳入科研评价体系,设立专门的验证成果奖项;(5) 设立国家级验证基础设施基金,支持自动化实验平台等关键技术攻关与开放共享;(6) 推动建立国际 AI4S 验证标准与互认机制,建设全球共享的验证数据库,整合全球科研资源共同应对验证挑战。

参考文献(References)

[1] Klowden T, Tao T. Mathematical methods and human thought in the age of AI[EB/OL]. (2026-03-29) [2026-05-15]. https://doi.org/10.48550/arXiv.2603.26524.

[2] Lu C, Lu C, Lange R T, et al. Towards end-to-end automation of AI research[J]. Nature, 2026, 651(7681): 914-919.

[3] 亨利·基辛格, 埃里克·施密特, 克雷格·蒙迪. 人工智能时代与人类价值[M]. 北京: 中信出版集团, 2025.

[4] 李国杰. 基于可判定性理论的人工智能系统安全风险分类[J]. 计算机研究与发展, 2026, 63(3): 539-547.

Prospects, bottlenecks, and risks of AI for Science

LI Guojie

Institute of Computing Technology, Chinese Academy of Sciences, Beijing 100190, China

Abstract Based on a comprehensive analysis of various AI large models and predictive information provided by relevant literature, this paper presents a cautiously optimistic forecast for the development prospects of AI4S over the next five years. After analyzing the automatic generation of top-tier conference papers by The AI Scientist-v2 platform, it is pointed out that there are inherent difficulties in achieving full automation of the research loop, and incomplete verification and drift in research objectives are obstacles that constrain the development of AI4S. From the perspective of computational theory, the root cause of the verification bottleneck in AI4S lies in the semi-decidability of scientific problems, and the boundary of verifiability is the boundary of automation. According to verifiability, scientific and technological problems can be divided into two major categories: R1 problems and R2 problems. The risk of AI4S lies in treating problems that should be governed as R2 problems as R1 problems. This paper sharply points out that AI4S may pose more urgent security risk than AGI, and the risk of AI4S is not "AI becoming smart", but rather "humans losing the right to withdraw ".

Keywords AI for Science (AI4S); incomplete verification; goal drift; semi-decidability; R1 problem; R2 problem; AI governance

（责任编辑 魏来）