

AI4S: 重塑科学发现的新范式

胡志坚

科学发现的历程,伴随着人类认知世界方式不断革新的过程。从远古先民对自然现象的简单观察,到现代科学家借助精密仪器探索宇宙奥秘,每一次认知工具与研究方法的突破,都推动着科学发现范式的迭代演进。今天,人工智能技术与科学研究的深度融合催生了 AI4S(人工智能驱动的科学, AI for Science),这一全新的科学发现范式正突破传统研究的边界,重塑科学探索的逻辑框架,为人类解开自然之谜注入前所未有的强大动力。

21 世纪以前,科学发现范式的演变大致经历了三个具有里程碑意义的阶段。在牛顿之前的漫长岁月里,科学发现处于经验科学范式阶段,观察与简单实验是认知自然的主要方式。科学家们通过肉眼观测、直观感受收集数据,试图从纷繁复杂的自然现象中找出规律,为科学研究奠定了实证基础。1687 年,牛顿《自然哲学的数学原理》的出版标志着理论计算范式的诞生,数学工具成为解释和预测自然现象的核心载体。理论计算与实验观察共同构成了近现代科学发展的双轮,推动科学研究实现了跨越式进步。20 世纪后半叶,随着科学问题日益复杂,传统分析方法难以应对高维、非线性的复杂系统,而计算机技术的突破使得科学家们能够通过数值模拟逼近复杂问题的近似解,计算科学范式应运而生。这种研究模式广泛应用于气候科学、分子动力学、材料科学和天体物理学等许多科学领域。

以往的科学发现范式各自存在难以逾越的局限性。实验研究往往耗时费力,大型实验的开展需要巨额投入,且存在“孤岛现象”,难以实现跨实验的宏观综合分析。数学模型为了简化复杂系统,必须进行假设抽象,这天然导致模型与现实存在偏差,且许多复杂方程组无法获得解析解,限制了理论计算的应用范围。超级计算机模拟虽能弥补部分缺陷,但设备成本高昂、仿真周期漫长,且仿真本身受到数学模型、数据表示以及初始条件等的限制,难以全面探索参数空间。这些局限性在面对重大复杂科学问题时尤为突出,亟需新的范式突破。

AI4S 的出现为彻底克服传统科研范式的局限性带来了希望。21 世纪以来,随着实验设施与模拟技术的升级,数据量呈爆炸式增长。卫星遥感、基因测序仪、强大的望远镜、同步加速器、中子源和电子显微镜、超算中心,以及大型国际科学设施等,每年产生的数据量已经达到数 PB 级,从这些不断增长的海量数据中提取有意义的科学见解是科学界面临的重大挑战。数据密集型科学逐渐崭露头角。如今,科学家不再依赖直接观测,而是通过大型仪器捕获数据或模拟生成数据,经软件处理后再进行分析,这种“数据先行”的模式成为科学发现新范式的核心特征。随着人工智能和机器学习等数据分析技术的不断突破,AI4S 正在加快重塑这一新范式。AI4S 是将人工智能技术应用于科学大数据,将其转化为科学发现的“富矿”,从中挖掘有价值的科学洞察。这种变革正在深刻改变多个科学领域的研究方式,展现了巨大的前景。

AI4S 推动科学研究规模从双峰科学向中等规模科学演变。到 20 世纪末,世界上大部分科学呈现双峰模式:一端是预算不足 100 万美元的小科学,由单个研究人员或小团体开展,数据生命周期较短;另一端是预算超 10 亿美元的大科学项目,建设周期长达数十年,虽能产生海量数据,但技术冻结早、风险厌恶度高,缺乏敏捷性。中间状态的科学项目几乎没有。而 AI4S 与探测器和传感器技术快速发展的结合,催生了预算在 1000 万至 1 亿美元之间的中等规模科学。这类项目依托先进自动化仪器和最新 AI 工具,既能创建大规模数据集,又保持了研究的灵活性,成为当今科学突破的核心阵地。斯隆数字巡天、1000 个基因组项目等成功案例,证明了中等规模科学的“最佳点”价值,这里是最具创新性的研究所在,为科学发现释放了巨大的智力能量。

数据科学成为 AI4S 不可或缺的一门新兴科学。科学发现越来越依赖大规模数据集及其质量,计算机与算法贯穿于数据采集、转换、组织、分析的全流程。正如计算科学从应用于不同学科的计算机科学中出现一

样,数据科学是从 AI 研究中出现的。数据科学继承了自动数据处理、大数据、云计算、数据挖掘、深度学习等技术成果,专注于大数据的管理与利用,其工具不仅包括人工智能,还涵盖统计学、运筹学等多学科方法。以往被视为“灾难”的数据洪流,如今通过数据科学的技术手段转化为有价值的科学资源。值得注意的是, AI 本身的突破离不开高质量数据集的支撑,而可解释 AI 的发展更依赖数据科学的进步,二者相互促进。

数据分析方式的变革是 AI4S 范式的重要组成部分。数据分析方式发生了重大变化,经历从本地分析到云端处理的演变。这一演变与音乐行业的发展轨迹高度相似:从早期“黑胶唱片模式”的本地数据存储与分析,到“CD 模式”的介质传输与本地处理,再到“iTunes 模式”的在线数据库查询,最终发展为当前的“流模型”——数据与计算集中于云端,实现服务器端的数据选择、转换与 AI 驱动的模式识别。这种变革打破了数据获取与分析的时空限制,让全球科学家能够共享数据资源、协同开展研究,极大提升了科学发现的效率。

不同学科之间的收敛趋同和综合程度越来越高是 AI4S 范式的另一显著特征。尽管物理学、化学、生物学等学科的传统计算方法差异显著,但 AI4S 范式与 AI 技术能够普遍适用于所有科学领域。不同的科学领域正在呈现非常相似的数据密集型研究活动。这为构建通用的科学研究工具与实践体系提供了可能,可以避免各学科“重复发明轮子”的资源浪费。斯隆数字巡天平台的演进就是典型案例,该平台从天文学专用工具,逐步迭代为模块化、通用化的科学服务平台,适用于从生物学到社会学等多个领域,实现了跨学科的资源共享与方法互通。

AI4S 并非单一技术的应用,而是由数据、模型、算力、算法、工作流及协作网络等构成的复杂生态系统。其效能取决于各组件的协同发展与交互效率。构建鲁棒、高效、可扩展的生态系统,成为新范式能否主导未来科学发展的关键。尽管 AI4S 已在部分领域取得重大突破,甚至助力获得诺贝尔奖,但这一范式在 21 世纪才诞生,其生态系统远未成熟,面临着诸多亟待解决的挑战。

数据方面,实现科学数据的可发现、可获取、可互操作、可复用(FAIR 原则)是 AI4S 生态的核心,但要使数据从“出生”开始就符合 FAIR 标准仍面临诸多挑战。当前绝大多数公共领域数据(包括数据集、代码、算法、工作流、模型和软件等)仍缺乏规范描述,复用难度大,数据准备工作繁琐,距离形成闭环的 FAIR 标准还有较长距离。基准测试领域, AI 技术在科学问题中的适用性与有效性评估仍处于“黑箱”状态,缺乏统一、连贯的比较框架,难以区分经典科学方法与 AI 方法的本质差异,阻碍了技术的优化与推广。

不确定性量化与可解释性不足是 AI4S 面临的核心技术挑战。AI 通过挖掘数据相关性进行预测,但缺乏对复杂系统的分析理解,若无法量化预测中的不确定性,其在科学研究中的应用将受到严格限制。只有将不确定性量化与可解释性相结合, AI4S 才能成为传统科学范式的可行替代方案。同时,大规模科学合作需要跨地域、跨领域的数据流与计算资源协同,工作流的复杂性呈指数级增长,对全球网络集成系统的可靠性与可预测性提出了前所未有的要求。

大型语言模型在科学领域的应用价值仍需验证。基础模型的同质化增加了错误传播风险,涌现性则导致系统行为难以预判,需要在硬件与工程层面加大投入,开发下一代基础模型。此外, AI 驱动的科学发现自动化引发了科学目的评估问题,如何明确科学追求的价值导向,避免研究方向迷失在“AI 黑箱”中,成为亟待解决的伦理与认知挑战。

科学智能系统的反思能力与因果推理能力不足,也是制约 AI4S 发展的关键瓶颈。当前 AI 尚未掌握科学推理中的多样化认知过程,无法独立提出研究问题、制定研究策略,也难以在已知知识框架下定位新发现的价值。而因果推理作为科学研究的核心目标,机器学习仅能识别数据相关性,难以实现基于观察数据的因果推断,这一领域的突破仍需长期探索。

我们正生活在 AI4S 科学发现新范式快速演变的时代。尽管挑战重重,但 AI4S 带来的范式变革已不可逆转,前景可观。它不仅是研究工具的革新,更是科学方法和思维方式的重塑,让人类能够以更高效、更全面的方式探索自然规律。从基础科学研究到应用技术创新, AI4S 正在为各个领域注入新的活力,推动科学技术发展进入规模化、协同化、智能化的新时代。面向未来,构建成熟的 AI4S 生态系统需要全球科学界的共同努力。随着科技的不断进步与生态的日益完善, AI4S 必将成为主导未来科学发展的核心范式,助力人类解开更多宇宙奥秘,推动社会文明实现跨越式进步。

(胡志坚 中国科学学与科技政策研究会理事长,

宁波工程学院中国-中东欧国家创新合作研究中心(宁波欧亚研究中心)特聘研究员)