Discussion Note | 论鉴系列

总 No.2 期/2026 年 7 月

AI 与科学发现的未来：从基础机制到认识论、方法论与社会学维度的综述

执行摘要

自二十世纪中叶计算机科学诞生以来，科学发现的方法论经历了从纯粹的经验观察与理论推演，向计算模拟乃至数据驱动模式的深刻演进。近年来，伴随深度学习架构的突破、算力的指数级增长以及海量科学数据集的积累，人工智能（AI）在科学研究中的角色已经发生了根本性的本体论转变。如果说过去的 AI 仅仅是被视为用于数据分析和模式识别的边缘辅助工具，那么当今的 AI 则已跃升为科学发现的基础性机制与核心引擎。从解决困扰生物学界半个世纪的蛋白质折叠难题，到预测极端气候事件，再到自动生成科学假设与指导高通量实验，人工智能正在以前所未有的规模、广度与速度重塑整个科学探究的边界。

《代达罗斯》（Dædalus）2026 年冬/春季专刊以“AI 与科学：发现的未来是什么？”为题，汇集了涵盖生命科学、物理学、地球系统科学、计算机科学以及社会科学等多个领域的前沿洞见。与之呼应，斯坦福大学以人为本人工智能研究院（HAI）于近期召开的“AI+Science：加速发现”大会也指出，AI 正在像昔日的望远镜和显微镜一样，为科学探究开辟全新的视野。

然而，与传统观测工具不同的是，AI 不仅允许人类“看见”微观或宏观现象，它更具备在人类心智无法独立处理的庞大复杂数据集中提取、理解并利用高维模式的能力。随着这一进程的不断深入，科学界逐渐达成一种审慎的共识：面对高度复杂且充满噪音的自然世界，单纯依赖统计相关性的数据驱动模式已显露出其固有的局限性。未来的科学发现必须走向“知识中心化（Knowledge-Centric）”，将物理定律、因果推理、第一性原理以及非人类中心主义的科学哲学深度融合于系统设计之中。本综述旨在全面剖析“AI 驱动的科学（AI for Science，AI4S）”在当前及未来的发展轨迹，考察其在各核心学科的突破，并深入分析自动化科研工作流带来的认识论重构与社会学挑战。

在未来，人类科学家的角色将不可避免地发生蜕变：从繁重数据的解析者，升华为探究架构的规划者、因果机制的批判者以及伦理界限的守夜人。通过结合物理约束、可解释的概率语言模型以及高度透明的开源研究生态，人工智能最终将作为人类智能最强有力的伙伴，而非替代者，在这个加速发现的黄金时代，共同构筑一个更加深刻、包容、且充满智性光辉的新科学图景。

第一章 生命科学与医学中的多模态架构与节点生物学

在生命科学领域，人工智能已从单纯的一维序列分析过渡到对生物系统复杂三维结构、多组学交互以及系统动力学的精确建模。这一范式的转变首先体现在结构生物学中。通过 AlphaFold 系统，研究人员成功解决了预测蛋白质三维结构的巨大挑战。传统的计算方法在面对一个典型蛋白质约 10^{300}种可能构象时，往往因组合爆炸而陷入计算灾难。而 AlphaFold 通过学习生物学数据的深层“语法”，绕过了计算成本极高的从头模拟，实现了实验级别的精度，并预测了超过两亿种蛋白质的结构。在此基础上，AlphaFold 3 进一步揭示了蛋白质如何与 DNA、RNA、配体及其他蛋白质相互作用以执行复杂的生物学过程。与此同时，结合物理学启发的扩散模型（Diffusion Models），AI 不仅能解析已知生命物质的结构，更能通过引入特定的靶向约束条件，从随机氨基酸坐标出发，生成自然界中前所未有的、具有特定结合能力的蛋白质分子，从而在药物设计领域实现了数量级上的成功率提升。

伴随高通量单细胞测序（如 scRNA-seq）和空间转录组学技术的进步，生命科学的研究正从整体组织的平均表达量测量，向精确至单一细胞的多模态分析演进。人类细胞图谱（Human Cell Atlas, HCA）项目即是这一趋势的集中体现，其目标是绘制整个人体的三维、多尺度细胞图谱。在这一进程中，“虚拟细胞（Virtual Cells）”的概念应运而生。通过自监督学习，基础模型能够将海量单细胞图谱中的转录组学、表观基因组学及蛋白质组学数据压缩并映射到统一的潜空间中，捕捉细胞系统的深层关联。这种模型不仅限于静态的细胞分类，而是致力于解决“细胞预测问题（Cell-Prediction Problem）”。它们能够通过对比学习或交叉模态预测（例如从组织学图像推断基因表达模式），预测细胞在未见过的条件或治疗手段下的动态行为。

在节点生物学（Nodal Biology）的理论框架下，AI 的应用进一步展现了其在疾病机制研究中的巨大潜力。许多表面上看似不相关的单基

因遗传病（例如某种遗传性失明、家族性阿尔茨海默病与某种遗传性肾病），实际上在更深层次的生物学通路上交汇于同一个可成药的机制节点。例如，研究人员发现 TMED 货物受体系统是连接数十种不同疾病的关键节点；这些疾病均表现为突变产生的毒性蛋白在细胞内异常积累，最终导致细胞死亡。通过部署大规模的细胞扰动研究，并利用 AlphaFold 对该货物受体进行原子级的三维可视化，科学家成功设计出了能够精准嵌入靶点口袋的小分子药物，从而引导受体将其毒性货物运送至溶酶体进行降解。这种将人类科学直觉与 AI 高通量预测相结合的节点生物学范式，为系统性应对数以千计的罕见遗传病提供了高度可扩展的解决方案。

在全球健康平等的语境下，AI 在药物发现中的应用也引发了关于数据主权与包容性的深刻反思。长期以来，全球南方的医学研究主要依赖于全球北方生成的数据和工具，非洲的丰富遗传多样性在主流基因组数据集中代表性严重不足。通过引入 AI 主导的药物发现机制和药物基因组学研究，非洲的科研机构（如南非开普敦大学的 H3D 中心以及 Ersilia 开源计划）正致力于开发针对非洲本土流行病与被忽视热带疾病的治疗方案。例如，利用 AI 预测传统非洲草药中天然产物的活性，并简化其结构以降低合成难度。然而，要实现这一愿景，必须解决本地数据获取、基础设施建设以及研究话语权的问题，防止“数据殖民”。建立注重隐私保护的联邦学习（Federated Learning）框架，确保模型在具有生物地理学代表性的数据集上进行训练，是保证 AI 真正促进全球公共卫生公平的前提条件。

第二章 物理学、天文学与量子计算：情境与第一性原理的融合

与生命科学中多模态、高维但存在内在统计规律的数据环境不同，物理学研究的子领域在数据结构、基础设施以及证据标准上呈现出极度的异质性。正如学者 Tess Smidt 所指出的，“物理学是不同的（Physics Is Different）”；这种差异要求 AI 的应用必须深度契合特定领域的工艺（Craft）与数据文化。

在粒子物理学领域，大型强子对撞机（LHC）等实验装置会产生规模庞大且高度同质化的数据集。由于质子-质子碰撞的条件被严格控制，研究者每秒必须处理高达三千万次的对撞数据。在这种极端环境下，即便只有轻微的统计波动，在海量数据中也可能伪装成有意义的信号（因此粒子物理学确立了 5σ 即 99.99994%置信度的严格发现阈值）。为了应对这一挑战，机器学习技术被深度集成于数据筛选管线中。部署在现场可编程逻辑门阵列（FPGAs）上的超快速分类器能够在纳秒级延迟内过滤无效数据，而图神经网络（GNNs）则在密集的碰撞环境中重构粒子轨迹。这些模型的训练高度依赖于 GEANT4 等精确模拟软件生成的标注数据，且必须通过持续的校准与控制样本验证来抵御模型设定的系统性偏差。

相比之下，凝聚态物理学、材料科学与化学等原子尺度科学面临的则是高度碎片化、异质且充满近似的实验数据。无论是 X 射线衍射（XRD）、透射电子显微镜（TEM）还是角分辨光电子能谱（ARPES），不同探测手段提供的信息往往难以直接统一。尽管薛定谔方程在理论上可以描述所有量子系统，但在实际计算中，密度泛函理论（DFT）等近似方法不可或缺。在此背景下，AI 不仅被用于对实验光谱进行去噪和实时图像分析，更被用于构建物理引擎的替代模型（Surrogate Models）。例如，基于机器学习的原子间势函数能够以极高的计算效率逼近 DFT 的精度；而等变神经网络（Equivariant Neural Networks）则将物理学的对称性（如平移、旋转及置换不变性）直接编码于网络架

构之中，从而提升模型在分子动力学模拟中的可靠性与数据效率。

在天文学与宇宙学领域，AI 的应用正推动该学科向着超大规模模拟与跨模态预训练的方向演进。由于恒星与星系的质量、温度和密度等属性无法直接测量，只能通过观测其发射的光来反向推断，深度学习的黑盒特性在此遭遇了严峻的认识论挑战。然而，通过建立诸如 Aion-1 或 AstroCLIP 等涵盖数以亿计天体、融合多达三十九种观测模态（包括低分辨率图像、光谱和时间序列数据）的天文基础模型，天体物理学家能够以更高的分辨率重构宇宙结构的演化，甚至突破现有理论框架的束缚，探索标准模型之外的未知领域。

更为深刻的范式转移发生于量子物理与 AI 的交汇点。随着量子计算技术的发展，量子人工智能（Quantum AI）有望打破经典计算的信息论瓶颈。量子传感器能够直接感知处于叠加态和纠缠态的量子数据，若利用量子处理器对这些数据进行直接处理，所需训练样本的数量相较于经典机器学习方法将呈现指数级的下降。在超越标准量子极限（Standard Quantum Limit）的精密测量中（如搜寻轴子暗物质），通过空腔量子电动力学（cQED）系统构建的量子神经网络能够原生且高效地实现联想记忆与组合优化。如果神经网络能够成功逆向工程并模拟从蛋白质折叠到行星动力学的物理规律，这或许暗示着“信息”本身比物质和能量更具本体论优先性，物理学定律可能正是这种可学习的信息处理过程的涌现属性。

第三章 地球系统、气候模型与深地科学的约束优化

面对气候变化这一系统性、长期性的人类挑战，对物理地球的动态演变进行精准预测本质上是一个极度复杂的“约束优化”问题。在此领域，纯数据驱动方法不可避免地会遭遇“不可解前沿（Intractable Frontier）”——即长期的非平稳性气候漂移与短期的混沌变异性（如蝴蝶效应）之间的矛盾。

目前，AI 在某些“可解前沿（Tractable Frontier）”中已展示出显著的价值。例如在极端天气事件的短期预测中，基于 AI 的全球天气模型（如 FourCastNet）表现出了超越传统数值天气预报的计算速度与预测精度。FourCastNet 利用基于四十年卫星观测与物理分析数据的 ERA5 数据集进行训练。尽管其训练样本仅约五万个（与大语言模型的万亿级语料相比微乎其微），但该模型依然成功提前预测了飓风“李”和“贝丽尔”的轨迹。这种预测能力源于气象现象中鲜明的物理特征模式被模型迅速捕捉。更为关键的是，FourCastNet 采用了球面傅里叶神经算子（Spherical Fourier Neural Operator, SFNO）架构，该架构原生支持球面几何的等变性，从根本上避免了传统网格状神经网络在处理全球数据时于极点附近产生的失真和灾难性崩溃现象。这使得 AI 模型不仅在短期预测中表现优异，甚至能够向亚季节或长周期气候建模进行稳定外推。

然而，要支撑 AI 模型的持续、指数级扩张，必然面临能源供给与算力消耗的深刻矛盾。训练单个大型基础模型所消耗的电力可等同于数千个家庭的年用电量，且 AI 算力中心需要提供能够响应瞬间计算峰值、全天候且极其稳定的基荷电力（Firm Power），这绝非仅靠太阳能和风能等间歇性可再生能源就能解决。学者 Gege Wen 将这种困境比作古希腊神话中的“双重衔尾蛇（Double Ouroboros）”悖论：AI 的发展受制于能源基础设施的升级，而重构能源系统又必须高度依赖 AI 带来的计算革命。

打破这一僵局的关键在于地球的深层地下空间。无论是结合碳封存的天然气发电，还是深层地热能的开发与大规模储能，地下工程的实施环境都处于高度异质性且不可见的岩层之中。传统的储层数值模拟工具（建立于上世纪九十年代油气繁荣时期）计算耗时长且充满不确定性。为解决这一难题，利用 AI 构建地下物理模拟预测系统成为唯一可行的出路。例如，ccsnet.ai 平台将海量预训练的流体力学与热力学偏微分方程（PDEs）求解结果输入模型，使用户能够在数分之一秒内获得关于二氧化碳注入后地层压力变化及流体运移的精准预测。通过融合物理学守恒定律与现场实验数据，这些物理学启发的 AI 模型大幅缩短了能源基础设施的设计周期，使全天候无碳能源的稳定供给成为可能。

第四章 神经科学、计算模型与生物智能的启示

在解开自然界奥秘的诸多尝试中，理解人类大脑本身的结构与功能依然是科学界最具挑战性的任务之一。神经系统连接组学（Connectomics）致力于绘制出从感官输入到运动输出的完整突触级神经回路图谱。通过电子显微镜技术（如透射电子显微镜 TEM 与扫描电子显微镜 SEM）、关联光电子显微镜（CLEM）以及膨胀显微镜等前沿成像手段，研究人员正在获取从线虫（C. elegans）、果蝇（Drosophila）到人类大脑皮层的超大规模纳米级解析度数据。

然而，面对动辄 PB 级别（Petabyte）的神经影像数据，传统的手工追踪和标注已经完全失效。深度学习模型（如 3D U-Nets 与泛洪填充网络）的介入，将原始像素自动转化为三维神经元分支与突触结构，从而使连接组学跨越了计算瓶颈。不仅如此，多模态连接组学还将转录组学与解剖学数据融合，构建出反映大脑动态状态与分子多样性的“分子连接组”。在此基础上，AI 研究人员正试图构建大脑的“数字孪生（Digital Twins）”，这些因果预测模型不仅能模拟特定视觉或运动皮层在感官刺激下的神经放电活动，还能指导设计脑机接口控制协议，甚至有望通过闭环干预控制癫痫等神经系统疾病的发作。

尽管如此，目前最先进的硅基 AI 系统与数亿年进化而来的生物智能之间，依然存在着在数据效率、能源效率以及鲁棒性上不可逾越的鸿沟。如下表所示，生物大脑在资源消耗上展现出了令人惊叹的极简主义美学与涌现的系统可靠性。

正如学者 Surya Ganguli 所观察到的，大脑通过三磷酸腺苷（ATP）的代谢实现了类似于“智能电网”的预测性能源分配机制，在时间与空间上极其精确地满足神经元的能耗需求。生物计算的过程虽然在单一突触层面充满噪音且缓慢，但其通过高度并行的模块化架构实现了极低的能耗与高度的鲁棒性。这种基于物理底层逻辑的计算原理提示

我们，未来的 AI 设计或许需要跳出纯粹的数字逻辑抽象，回归至基于随机性、模拟物理状态的新型计算底座。

智能系统	典型功率消耗	时钟周期/信号延迟	数据需求特征	鲁棒性与纠错机制
生物大脑(人类)	~20 瓦特	毫秒级 (Slow & stochastic)	极高效率(Few-shot learning)	系统级涌现纠错，对噪音高度免疫
大型 AI 模型(LLMs)	数兆瓦(Megawatts)	纳秒级 (Fast & deterministic)	海量需求(Trillions of tokens)	易受对抗样本(Adversarial examples)攻击

第五章 迈向知识中心化的 AI 与认知科学的融合

大语言模型和基于 Transformer 的架构虽然在模式识别和序列生成任务中取得了卓越的成就，但其本质上依然是基于表面相关性的自回归系统。这种模式对应于心理学中所描述的系统 1（System 1）——即快速、直觉驱动的认知系统。然而，真正的科学发现需要的是系统 2（System 2）级别的认知：深思熟虑、逻辑演绎、反事实推理以及对第一性原理的严格遵循。

深度推理网络与领域对齐的潜空间

为了克服纯数据驱动方法的局限，计算机科学家 Carla Gomes 提出了知识中心化 AI（Knowledge-Centric AI）的范式。其核心在于通过可解释的神经网络架构将科学领域先验知识硬编码到模型的推断过程中。以“深度推理网络（Deep Reasoning Networks, DRNets）”为例，该框架的有效性可以通过解决重叠的数独游戏（Multi-MNIST-Sudoku）来直观理解。在这一任务中，视觉编码器首先将图像转化为包含符号概率的潜空间；随后，基于逻辑规则的推理模块（如同解数独的约束条件）通过连续可微的熵函数在潜空间中实施软约束；最后由生成式解码器重构并输出解开的图样。

这一抽象框架在科学领域的投射同样令人瞩目。在材料科学中，从 X 射线衍射（XRD）图谱中反推晶体结构相图（Crystal-Structure Phase Mapping）是一个典型的 NP 难问题。借助物理与对称性启发的 PSI-DRNets 模型，研究人员将布拉格定律（Bragg's Law）和热力学约束融入潜空间，成功解析了此前人类专家无法破解的 Bi-Cu-V 氧化物系统，发现了性能优异的太阳能燃料合金相。类似地，在生态学中，通过 DMVP-DRNets 处理 eBird 的公民科学数据，模型在考虑多物种共现与环境变量的严格生态约束下，实现了针对五百多种北美鸟类的大规模联合

物种分布建模。这证实了知识中心化的 AI 不仅能够在数据稀缺的条件下保持泛化能力，还能反哺 AI 底层算法的创新。

概率语言模型与模型综合架构

从认知科学的维度来看，麻省理工学院的 Joshua Tenenbaum 指出，“语言并不代表思维的全部”。人类的因果推断能力远远早于语言的习得。因此，试图仅通过扩大文本预训练规模来孕育通用科学智能，是一条充满脆性（Brittleness）的路径。Tenenbaum 团队提出了以贝叶斯推理和概率编程（Probabilistic Programming）为核心的架构，探索“概率思维语言（Probabilistic Language of Thought, PLoT）”。

在“模型综合架构（Model Synthesis Architecture, MSA）”中，人类心智并不携带一个涵盖万物、庞大且单一的世界模型，而是在面临具体情境时，利用语言作为内部关联的检索工具，从记忆中调取相关碎片，实时合成一个小巧、定制化的贝叶斯概率程序。例如，在推断一场复杂的拔河比赛结果时，MSA 架构能够将自然语言描述翻译为 PPL（概率编程语言）代码，加入包括选手状态、环境影响等因果变量，进而在该临时生成的模型中进行极其高效的贝叶斯推断。这种结合了 LLM 语言翻译能力与严格概率推理的神经符号（Neuro-symbolic）架构，为构建具有高度科学常识与反事实演绎能力的“博学者 AI（AI Polymath）”指明了方向。

第六章 自主科研工作流与“自动科学”的边界

科学发现的方法在过去几个世纪中基本保持不变：提出假设、设计实验、收集数据、分析结果。然而，随着智能体（Agentic AI）和自动化硬件的快速迭代，科学探索正迈入“闭环自动实验室（Autonomous Laboratories）”时代。在材料学与化学领域，例如由 Connor Coley 等人开发的自动化平台以及 A-Lab 项目，已经能够实现从理论配方到真实物理合成的全天候不间断操作。然而，将屏幕上的算法翻译为真实世界中机械臂对粘稠液体和易敏粉末的处理，中间存在着巨大的“执行鸿沟（Execution Gap）”。

在更广义的虚拟科学探索中，基于 LLM 的智能代理体系已经被整合到了研究的整个生命周期中。系统性研究指出，自动研究过程（Auto-Research）可被划分为四个核心阶段，如下表所示 ：

研究阶段	AI 扮演的核心角色与功能	当前表现与技术瓶颈
1. 创造与生成 (Creation)	理念生成、大规模文献检索（如通过 Elicit, SciSpace 等工具提取证据综合）、代码编写与虚拟实验调度。	在结构化数据和工具调用的辅助任务上表现优异；但面对需要深层次判断科学新颖性的任务时极度脆弱。
2. 撰写与重构 (Writing)	依据实验数据和图表，生成初稿、提炼逻辑论点并确保科学术语格式的连贯性。	输出流利且符合学术规范，但容易陷入自我循环和格式同质化。

3. 验证与同行评审 (Validation)	模拟多专家代理对论文提出反驳、寻找方法论漏洞，并提出修改建议。	容易遗漏深层逻辑谬误，并在缺乏确凿外部事实检验时妥协 。
4. 传播与影响 (Dissemination)	生成学术海报、社交媒体演示文稿以及交互式的问答代理系统。	极大地降低了知识传播的边际成本。

尽管全自动化系统能以低至 15 美元的惊人成本端到端地生成一篇科学论文，但这种高产出的背后隐藏着一个危险的“能力悬崖（Capability Cliff）”。研究表明，AI 生成的具有一定新颖性的科学理念在转化为具体执行代码时，其质量会发生断崖式的降级。更令人担忧的是，AI 由于其固有的“幻觉”特性，正在以前所未有的速度污染着人类积累了数百年的科学文献库。

2025 年的一项大规模评估揭示，随着生成式 AI 的普及，科学文献中出现了爆发式的虚假引用（Fabricated Citations）。研究人员发现，在 2025 年单年，就有至少 146,932 条由 AI 伪造的文献条目混入了科学文献数据库中。这些并不存在的研究大多首先出现在预印本平台上，但令人震惊的是，绝大部分伪造引用最终成功逃过了传统的人类同行评审机制，被正式出版的期刊文章所收录。数据显示，包含至少一条虚假引用的论文比例在急剧攀升：从 2023 年的 1/2828 篇，恶化至 2025 年的 1/458 篇，2026 年初，这一数字进一步恶化至 1/277 篇。

科学文献中AI伪造引用的激增趋势

数据表明，随着生成式AI的普及，2025年有超过14.6万条伪造引用进入科学文献，凸显了在自动化科研闭环中引入严格事实核查机制的紧迫性。

数据来源: The Economic Times

这种由于过度自动化而引发的认识论危机表明，彻底移除人类监管的“完全自主科学代理”在当前阶段完全不可靠。科学探索的核心仍应是以人类主导下的合作机制（Human-governed Collaboration）为主，通过设立诸如 REFORMS（针对机器学习科学的共识建议）与 CLAIM（医学成像 AI 清单）等详尽的方法论核查标准，来保障科学发现的严谨性、可重复性及透明性。

第七章 自主科学哲学：重新定义科学理解与机器到人类的教学

当AI系统发展至不仅能提供准确的数值预测，更能发现超出人类直觉的设计和规律（例如在量子光学的超高维纠缠生成实验或高灵敏度引力波探测器的设计中，AI寻找到了人类专家耗时数月仍无法完全解释的全新拓扑结构）时，我们不可避免地触及了科学哲学的根本问题：如果一个极其复杂的模型给出了正确的答案，但人类无法理解其内部逻辑，这还能算是获得了“科学理解”吗？

为了应对这一挑战，学者Mario Krenn和Heather Champion提出了“自主科学哲学（Philosophy of Autonomous Science, PAS）”的框架，并列举了定义人工科学时代核心精神的十个问题。其中最为迫切的任务是，如何将人类科学家的非量化驱动力——如理解力、新颖性、兴趣、惊奇与创造力——转化为非人类中心、可计算的目标函数？

例如，在衡量“科学惊奇（Scientific Surprise）”时，计算机科学通常借助贝叶斯惊奇（即先验与后验概率分布之间的Kullback-Leibler散度）来量化信息增益，或者使用香农信息量来衡量事件的意外程度。然而，这些纯粹信息论维度的指标，真的能捕捉到人类科学家在顿悟瞬间产生的带有“美学与优雅（Aesthetic or Elegance）”成分的认知情感吗？同样，关于“创造力（Creativity）”的定义，诸如早期的IBM Chef Watson实验通过最大化新颖性（利用贝叶斯惊奇）与有效性（依据化学风味组合的愉悦度数据）成功生成了新式菜谱。然而，将这种创造力近似方程转移至物理学和化学的前沿探索中，是否会导致AI代理在搜索广袤的组合空间时，忽略了那些需要跨越深层范式转换（Paradigm Shift）的非连续性突破？

在此基础上，科学理解理论的先驱Henk W. de Regt指出，真正的科学理解不在于逻辑演绎的完美，而在于科学家运用理论解决新问题的“可理解性（Intelligibility）”与泛化能力。为了防止人类在超级

AI 面前退化为仅作数据诠释的旁观者（正如 Ted Chiang 在其科幻短篇《捡拾桌上碎屑（Catching Crumbs from the Table）》中所预言的那样），科学家们提出应当大力开发“教师 AI（Teacher AI）”。这种系统专门致力于充当人类与超级智能之间的翻译官，它被赋予丰富的可视化工具、数学降维方法与引导式教学能力，使得高维度的外星般（Alien）的 AI 科学发现，能够被降解、提取并内化为人类能够掌握的新物理学定律与科学直觉。

第八章 科学社会学：文化、治理与人类的能动性

将人工智能嵌入科学架构，绝不仅仅是一次技术工具的升级，它深刻地牵动着科学界的权力分配、认知习惯以及制度伦理。正如社会学家Alondra Nelson所言，AI必须同时被视作社会科学的提问对象、研究客体与工具。我们必须重温诸如马克斯·韦伯关于“工具理性”的牢笼警告、W.E.B.杜波依斯对自动化加剧不平等的剖析，以及赫伯特·马尔库塞关于技术抑制批判性思维的诊断，以审视当今算法力量的扩张。

在科研的微观实践中，认知科学家M. J. Crockett和Lisa Messeri敏锐地捕捉到了AI工具带来的“理解的错觉（Illusions of Understanding）”。当大语言模型被奉为可以总结海量文献的“神谕（Oracle）”或处理复杂计算的“代行者（Surrogate）”时，研究人员极易产生认知卸载（Cognitive Offloading）的惰性。由于AI的输出往往极其流畅且看似自信，缺乏批判性思考的学者可能会放弃对底层数据的严格审视。长此以往，科学研究可能会不可避免地陷入“认知单作（Epistemic Monocultures）”的陷阱——即算法更倾向于在数据丰富、范式成熟的领域内进行安全迭代，反而窄化了科学探究的广角镜头，使得那些真正具有颠覆性但数据稀疏的科学问题被边缘化。这也从另一个侧面论证了，相较于不可靠且缺乏透明度的黑盒商业大模型，开源科学模型对于维护科学界透明和可追溯的认识论基础显得至关重要。

在更宏大的制度治理层面，2026年5月5日于斯坦福大学召开的“AI+Science：加速发现”大会确立了一个核心主旨：“AI改变了哪些问题是可解的，但它并不能告诉我们哪些问题才是真正重要的。唯有依靠人类的判断与价值观，才能决定科技发展的终极意义”。技术不应脱离公共善（Common Good）的约束。为此，学术机构正积极进行纠偏与战略投资。2026年4月14日，圣克拉拉大学（Santa Clara University）宣布接受NVIDIA执行副总裁Debora Shoquist提供的近

2500 万美元巨额捐赠，正式成立了“坎宁安·舒奎斯特应用 AI 与人类潜能中心（Cunningham Shoquist Center for Applied AI and Human Potential）”。作为该校第四个“杰出中心”，它根植于耶稣会的人本主义使命，明确将确保 AI 系统的公平性、安全性、透明度以及提升人类尊严作为科研部署的根本导向，并在医疗、机器人等诸多下游应用中贯彻伦理约束。这一中心的成立表明，在紧邻世界顶尖科技巨头的硅谷腹地，学术界正试图用“道德约束（Restraint）”和“以人为本”来平衡资本对算法速度的狂热追求。

结论：人机协同时代的科学重构

综上所述，人工智能与科学发现的融合已经超越了单纯的工具主义叙事，演变为一场触及科学本体论、方法论及社会伦理的深远革命。从破解生命的复杂折叠与基因表达序列，到跨越宇宙演化的时空尺度；从构建地球气候的物理神经网络算子，到深入微观原子的量子态操控；在打破纯粹数据驱动的局限、迈向知识中心化和概率推理基础的进程中，AI 正在不可逆转地拓展科学探索的可解边界。

然而，在这个从传统实验室向自动化、自主化乃至具备“多模态博学者”潜质的科研生态演进的过程中，人类面临的挑战同样前所未有。大规模幻觉与虚假文献的泛滥、黑盒模型带来的理解鸿沟，以及技术应用中可能加剧的不平等和认知单作危机，都在提醒我们：技术可以极大缩短“提出问题”到“获得答案”的距离，但它无法代替人类去追问“为什么”以及“这有何意义”。

在未来，人类科学家的角色将不可避免地发生蜕变：从繁重数据的解析者，升华为探究架构的规划者、因果机制的批判者以及伦理界限的守夜人。通过结合物理约束、可解释的概率语言模型以及高度透明的开源研究生态，人工智能最终将作为人类智能最强有力的伙伴，而非替代者，在这个加速发现的黄金时代，共同构筑一个更加深刻、包容、且充

满智性光辉的新科学图景。

联系我们：CEISD@Shanghaitech.edu.cn