AI预测知识的可靠性考察

——从科学实践哲学视角看

Examining the Reliability of AI Predictive Knowledge:
A Perspective from Philosophy of Science in Practice

宋奇峰 /SONG Qifeng

（中国社会科学院大学哲学院，北京，102400）
(School of Philosophy, University of Chinese Academy of Social Sciences, Beijing, 102400)

摘　要：人工智能驱动科学研究已成为当下的热点议题，随之而来的是关于AI预测能力及其预测结果可靠性的质疑。现有研究多从规范性视角探讨AI的可靠性，但此类视角无法揭示AI预测知识可靠性的来源。科学实践哲学将可靠性置于具体科学实践中众多科学产品相互缠结的动态过程中，为该问题的深入提供了富有启发性的理论路径。以AlphaFold为例，可以看到，AI预测知识的可靠性并非源自结果符合对象，而是根植于AI预测模型及其预测结果分别参与到与其它科学产品和科学共同体的复杂缠结之中。此外，AI预测及其“端到端”的生成模式正重塑着科学知识的形态。

关键词：人工智能　AlphaFold　科学知识　科学实践哲学　可靠性

Abstract: AI-driven scientific research has become a central topic in contemporary science, accompanied by growing concerns about the predictive capabilities of AI and the reliability of its predictions. Existing studies often examine AI reliability from a normative perspective; however, such an approach fails to uncover the sources of reliability in AI predictive knowledge. By contrast, philosophy of science in practice situates reliability within the dynamic processes in which numerous scientific products entangled in concrete scientific practices, thereby offering an illuminating theoretical framework for further exploration of the issue. Using AlphaFold as an example, this paper demonstrates that the reliability of AI predictions does not stem from mere correspondence to external objects, but rather from how both AI models and their predictions participate in the complex entanglement of other scientific products and scientific communities. Moreover, AI predictions and their “end-to-end” mode of generation are reshaping the structure of scientific knowledge.

Key Words: Artificial intelligence; AlphaFold; Scientific knowledge; Philosophy of science in practice; Reliability

中图分类号：TP18; N031　DOI: 10.15994/j.1000-0763.2026.08.002　CSTR: 32281.14.jdn.2026.08.002

近年来，人工智能深度参与到科学研究的各个环节，成为新一轮科技革命和产业变革的重要驱动力量。人工智能驱动科学研究（AI for Science）的兴起为科学知识的生成打开了新的视域。在诸多科研领域中，AI首先被广泛用于预测未知结果、性质或未来发展趋势。在现阶段的科学实践中，AI驱动的科学预测模型（AI-driven scientific prediction model，后文简

收稿日期：2025年12月24日
作者简介：宋奇峰（1998–）男，山西吕梁人，中国社会科学院大学哲学院博士研究生，研究方向为科学实践哲学。Email: 545303764@qq.com

称为AI预测模型）已在生物与医学、天文学、地球与环境科学、物理与化学等研究领域中起到了加速作用。这些AI预测模型以既有数据库为训练资源，在科学研究和实际运用当中预测准确，事实上已成为科学研究不可或缺的一部分。

然而，AI预测模型只是众多AI模型中的一部分。并非所有AI 模型的预测结果都能够像AlphaFold 对蛋白质结构的预测那样被视为科学知识，由此，一个核心问题便浮现出来：那些被视为能够生成科学知识的AI预测模型与一般AI模型相比，其可靠性源自何处？换言之，我们应该如何理解AI预测知识的可靠性？

现有关于AI预测可靠性的讨论通常从两类角度展开：一是聚焦于AI自身机制的分析，二是强调外部验证的方法。前者揭示了AI模型在预测过程中可能出现的各种的不足；后者则主张通过实证手段对预测结果进行验证，为结果的可靠性提供坚实支撑。然而，无论是前者还是后者，本质上都属于规范性的方法：前者体现为理论上的规范性，通过揭示理论层面的困境指出AI预测可能不可靠；后者体现为实践上的规范性，认为预测结果必须经过实验验证方可获得认可。更为关键的是，后者存在潜在矛盾：如果所有预测结果都需依赖实验验证，其高效性将被削弱，违背了发展AI以参与科学实践、提升科研效率的初衷。[1]因此，可以认为，单纯依靠规范性的方法并不能充分回应AI预测可靠性来源的问题。

科学实践哲学主张从实践的视角阐释科学合理性，强调描述性方法而非规范性方法。换言之，科学实践哲学认为，科学知识的可靠性源自具体的科学实践活动。该视角为我们绕开传统的实证检验框架，重新思考AI预测知识的可靠性提供了新的理论路径。本文以AlphaFold为例，从实践的角度考察AI预测知识的可靠性，并指出，其可靠性并非仅依赖模型本身，而是源于AI在参与科学实践时与其他科学产品之间的复杂缠结（Tangle）①。此外，具备新运作模式的AI预测正在深刻地丰富和重塑科学知识的生成与形态。

一、对AI预测知识可靠性的质疑

关于AI预测知识的可靠性的问题已有大量讨论，但现有讨论主要采取规范性视角，聚焦于AI的不可靠性。总体而言，对AI预测知识可靠性的质疑可以归纳为两方面：其一，AI的预测结果在未经实证检验的情况下可能出现错误；其二，AI的预测机制缺乏透明性。下面将分别对这两点进行阐述。

认为AI预测结果可能存在错误的论述可以总结为四种观点。首先，其中最受关注的现象是所谓的“AI幻觉”（AI Hallucination）。AI幻觉指人工智能系统在生成内容时产生的看似合理、但实际上错误、虚构或不存在的信息。对于幻觉产生的必然性，学界主要提出两类论证。[2]第一类基于广义信息论框架，其核心观点是：生成式人工智能模型所包含的信息量不足以准确刻画复杂的经验分布，因此不可避免地会产生错误输出。[3]第二类为几何性（geometric）论证。有学者指出，生成模型通过高维空间中的潜在表示生成输出，但高维空间的几何特性可能导致模型偏离真实数据分布。由于模型在不同数据流形之间插值时可能进入未被训练数据支持的区域，因此会产生“幻觉”等错误输出。

其次，模型高度依赖训练数据，如果数据存在噪声、缺失、偏差或测量误差，预测结果可能不可靠。[4]例如，检测蛋白质三维结构的实验方法通常有三种，分别是冷冻电镜、核磁共振与X射线晶体学。三种不同的方法得到了不同的数据。有研究显示，训练数据来自单独的冷冻电镜或核磁共振的AI模型表现始终低于训练数据单独来自X射线晶体学的AI模型的表现，而当训练数据同时包含冷冻电镜和核磁共振时，这一情况得到改善。这表明不同实验方法用于蛋白质结构测定可能会引入训练数据偏差。[5]

①Tangle本意为纠结、缠绕或混乱，为避免与其他“纠缠”概念混淆，本文将其意译为“缠结”。

第三，AI模型擅长预测与训练数据相似的情况，但对超出训练范围的极端事件或全新体系可能失效。例如，AlphaFold模型的训练需要通过多序列对比（Multiple Sequence Alignment, MSA），即通过对齐大量同源序列，捕捉保守位点与共进化信号，为后续深度学习或物理建模提供关键特征。而现有的数据库中，罕见蛋白或异常结构的同源序列较少，因此AlphaFold对罕见蛋白或异常结构的预测准确率低于常见蛋白。[6]

第四，AI预测是否能够如实反映自然现象与科学对象，也同样值得怀疑。例如，蛋白质是细胞内部动态变化的实体，具有多变的构象及复杂的相互作用关系。AlphaFold 默认预测的是氨基酸序列在热力学上最稳定的构象，并不会自动生成不同功能状态或激活态/非激活态的多种构象。尽管可以通过引入模板或MSA策略等方法诱导生成不同构象，[7]但客观存在的技术局限与模型本身的有限性，使其难以覆盖自然界的无限复杂性。这种功能局限与对自然复杂状况的理性分析共同构成了对AI预测结果可靠性的合理担忧。

另一方面，预测机制的不透明性限制了科学共同体对预测结果的直接理解与评估，从而影响了对AI预测知识可靠性的判断。这一问题在讨论AI是否能够生成科学知识时不可回避。机制的不透明性是人工智能模型长期存在的特性。[8]以功能为导向的AI模型，只要能够执行预定功能，或许可以对机制的不透明性持保留态度。例如，AlphaGo的机制不透明性可能成为研究对象，但并不构成信任其结果的障碍。然而，传统科学研究并不满足于工具性使用。理论解释与实验观察是现代科学研究的重要环节，理论为个别知识提供解释基础。而预测机制的不透明性意味着我们无法理解AI模型在迭代后生成的模拟函数及其参数的理论意义，从而限制了对其科学性的认知。

值得注意的是，对AI预测不可靠性的分析，并不能否认存在可靠的AI预测。原因在于，在讨论不可靠性时，已经预设了一个作为标准的“可靠”，于是讨论自然围绕规范性展开，即形成“三段论”式推理：可靠的科学知识应当如何→AI未满足这些条件→AI不可靠。这种规范性思维在逻辑实证主义之后已被证明难以实现。取而代之的是以科学实践哲学为代表的描述性思维，即关注科学实践本身，而非对科学活动进行预先界定。实践不仅是科学知识得以成立的手段，更是知识及其实在性的界定者。[9]

针对传统规范性论述，科学实践哲学家主张应从具体科学活动中寻找可靠性的依据，而非设立外在的客观标准来判断是否为科学知识。正如拉图尔指出：“我们研究的是行动中的科学，而不是已经形成的科学或技术。为了进行这种研究，一方面我们在事实和机器被变成黑箱之前抵达它们，另一方面我们也紧随把它们重新打开的争论。”[10]换言之，若某一AI预测结果被视为科学知识，并非意味着它符合“科学知识”或“可靠性”的预设标准，而是它在实践活动中被认可为科学知识或具有可靠性时，才被视为科学知识。因此，研究的关键在于回到AI预测所参与的具体科学实践中，探寻其被视为可靠的依据。

二、从科学实践哲学视角看可靠性概念

在讨论AI预测知识可靠性之前，有必要澄清“可靠性”概念的哲学意涵。“可靠性”指在现有证据与方法基础上的可信赖性。该概念并非凭空产生，而是取代或弱化了“绝对真理”“确定性”“绝对客观性”等传统观念，后者将科学视为对自然世界的客观映射。然而，随着相对论与量子力学的发展、历史主义的兴起，以及统计方法与概率论的广泛应用，科学不再被理解为“发现永恒真理”，而是建立在证据与方法基础上的、可重复验证的可靠知识体系。换言之，讨论“可靠性”本身即意味着对科学实践活动的重视，而非单纯关注理论本身。

关于科学知识的可靠性，莱茵伯格（Hans-Jörg Rheinberger）从实验系统兼容的视角提供了启示。他认为，现代经验研究过程中的最小功能单元是实验系统。然而，若实验系统自我

固化，将失去研究功能，仅能进行自我验证，因此必须保持差异性。不同实验系统交汇的地方即为操作（tinkering）的场所，可能产生意想不到的效果，并带来新的知识增量。[11]这表明新的科学知识既诞生于科学实践中，又持续参与其中。一方面，多种技术交融是科学创新的来源；换言之，新的科学成果只有在与已有可靠科学产品成功结合后，才被视为可靠。皮克林（Andrew Pickering）则强调科学知识的稳定实现并非一蹴而就，而是在同一实践个体的后续活动中通过不断的“去稳—再稳”过程获得的。[12]

与这两个观点相似但更整体性的阐释来自卡特赖特（Nancy Cartwright）等人。在理论层面上，卡特赖特主张律则多元论，反对任何种类的基础论。[13]换言之，科学知识的可靠性并不奠基于任何普遍性的基础。他们认为，科学知识作为一种科学产品，其可靠性源于与其他众多科学产品的缠结。“缠结”是理解科学知识可靠性的核心概念，强调知识之间的相互依存和实践网络的重要性。

卡特赖特指出：“任何单个产物的可靠性不仅依赖于其自身的精细细节，还依赖于背景中庞大的其他产物缠结的可靠性。”（[14]，p.11）这一观点涉及两个核心概念：科学“产品”和“缠结”。她以“产品”取代传统意义上的“真理”，强调科学活动的目标在于生产可靠的产品，而非追求绝对真理。这些科学产品在生产时被赋予明确功能标签，可以用于完成特定任务。（[14]，p.12）科学产品涵盖模型、定义、程序、仪器、概念、测量方法等一切在科学实践过程中产生的成果，它们构成科学产品库，为其他研究者提供服务。在此视角下，关注点不再是理论本身，而是科学产品的可靠性及其能够可靠实现的目标。例如，科学家通过实验解析出某蛋白质的三维结构，该结构成为科学共同体的公共产品，其他科学家可以将其用于新的研究，如药物研发。在讨论可靠性时，实际上是讨论科学产品的可靠性。

那么，何种科学产品可被视为可靠？卡特赖特指出：“一个测量方法、模型或概念对于特定目标的可靠性，并不仅意味着它能在其被创造的具体科学活动中与当时特定的一组产物协同完成任务，而是它必须能够在更广泛的活动中被信赖，以实现相同目标、完成相同任务。”（[14]，p.12）这一观点包含两个层面的可靠性界定：其一，科学产品能够在特定语境中重现。例如，在蛋白质三维结构测定中，相关科学产品包括结构生物学理论、测量仪器、实验流程及蛋白质结构本身，这些产品能够重复稳定地协同完成测定工作，即体现实验的可重复性。其二，科学产品能够运用于其他科学活动，实现相同目标。例如该蛋白质三维结构可在药物研发、作用机制研究或蛋白质工程等更广泛的研究中使用。科学产品不仅需在其诞生的语境中重现，更重要的是能够在更广泛的科学实践中保持稳定。

单个科学产品与其他众多科学产品的交互即构成“缠结”，这正是科学产品可靠性的来源。卡特赖特将其描述为：“我们所说的‘缠结’，是指丰富交织的科学创造网络，它们相互制约、相互支持——一个并行的、相互反馈、相互发展的思想、概念、理论、实验、测量、桥接原理（bridge principles）、模型、推理方法（methods of inference）、研究传统、数据与叙事的网络，它构成了科学事业（scientific endeavour），并将触须伸向其他同样丰富的缠结，而它们又反过来依赖于前者的成分。焦点不应在如何为科学论断的真理性提供担保，而应在于如何认证这张工作之网——如何获得信心，确信支撑某一科学产物可靠性的这一系列工作能够胜任。”（[14]，p.5）由此可见，单个科学产品的可靠性并非源于其如实反映现实，而在于其能与其他科学产品互为担保。科学成果并非由某一理论或观点奠基，而是由众多科学产品共同缠结搭建而成。缠结不产生必然性，却提供稳定性与可靠性。卡特赖特的观点事实上具有“物质融贯论”和“整体论”的意味。

有人可能质疑，“缠结”视角虽提供了可靠性的依据，但这种解释是否合理？卡特赖特对此提供辩护。首先，传统意义上的证实理论预设了缠结的存在：实验方法本质上通过使多

种科学产品相互交互来验证被检验对象的稳定性。可以说，实验本身即是一种主动的缠结方法。其次，当某个科学产品被认定为不可靠时，这通常是在分析科学实践失败的过程中发现的，其失败部分源于缠结中应存在的要素缺失或存在缺陷。换言之，对不可靠性的证明依然依赖于缠结的失效作为证据。最后，一个科学产品能够与更多类型的科学产品协同工作，或其自身定义越明确，则其失败可能性越低，在缠结状态中越稳定。（[14]，pp.132–156）大量的缠结为科学产品提供约束机制，这正是缠结支持可靠性的核心。其背后逻辑为：若科学产品确实可靠，即能反复且规律地实现既定目标，则必有其原因，而该原因基于自然界的事实，即便其机制可能超出我们理解。（[14]，pp.165–167）

在科学实践哲学视域下，可靠性概念得到重塑。“可靠性”并非静态规范性属性，而是在具体实践中不断建构和维系的过程性概念。某一科学产品即便被视为可靠的，也并不意味着绝对真理性，其可靠性来源于科学共同体的共享实践领域和科学产品彼此缠结，并依赖参照具体实践活动形成的评价标准。[15]

三、AI在科学实践中实现可靠性——以AlphaFold为例

AlphaFold是一个典型的AI预测系统，其核心任务在于根据蛋白质的氨基酸序列预测其三维结构。我们将以AlphaFold为例，阐释AI预测知识可靠性的内涵。

首先，有必要回顾AlphaFold的发展历程。AlphaFold首次亮相于2018年第13届蛋白质结构预测关键评估（Critical Assessment of Structure Prediction 13, CASP13）。该评估旨在确定当前蛋白质结构预测技术水平。参与者根据已知但未公开三维结构的蛋白质氨基酸序列预测其三维折叠构象。[16]预测结果与“标准答案”对比后，可反映预测水平。CASP13上，AlphaFold的出色表现受到顶级科学期刊和主流媒体广泛关注，被认为是彻底革新结构生物学的突破。然而，在这一阶段，由于自由建模预测难以达到实验结构的精度，且缺乏便捷工具，AlphaFold对结构生物学家日常工作的影响仍有限。[17]

2020年CASP14上，经过算法改进①的AlphaFold2显著领先竞争对手，[18]其后Google DeepMind与欧洲分子生物学实验室–欧洲生物信息研究所（European Molecular Biology Laboratory-European Bioinformatics Institute, EMBL-EBI）联合发布AlphaFold蛋白质结构数据库，目前已收录约2亿条UniProt知识库中编目的蛋白质多肽链结构。[19]如果说AlphaFold在技术上展示了AI的科研潜力，AlphaFold2则在实践中推动了科学研究进展。随后，AlphaFold-Multimer与AlphaFold3的发布将预测范围从单链蛋白扩展至多肽复合物和多分子体系，AlphaFold持续更新迭代中。

从科学实践的视角看，AlphaFold及其预测知识均应被视为科学产品，这种产品只有在与其他科学产品充分缠结后才能被认为可靠。这意味着评判标准不再是AlphaFold预测结果是否符合客观蛋白质三维结构，而应从AlphaFold与预测结果分别与其他科学对象的缠结中寻找可靠性依据。

AlphaFold作为预测工具，其可靠性依赖于与已被认为可靠的科学产品缠结，主要体现在两个方面：算法与数据。

1. 算法保障预测能力的可靠性

算法本身并非AlphaFold，但构成其基础。AlphaFold并非神谕式绝对真理，而是建立在严密算法上的计算模型。AlphaFold2通过深度神经网络对蛋白质序列及其同源序列的MSA数据进行端到端预测，其核心架构由Evoformer模块和结构模块组成。Evoformer模块采用类似Transformer的注意力机制，在MSA轨迹与

①AlphaFold2与AlphaFold都以蛋白质结构的实验解析结构为训练数据，但AlphaFold2不再使用卷积神经网络，而是采用了在自然语言处理领域广泛使用的Transformer深度学习框架。

残基对信息（对偶轨迹）之间反复交互，以捕捉序列共进化特征与空间邻近关系。结构模块引入旋转和平移不变的刚性框架，预测每个残基的三维坐标，利用旋转不变注意力机制和迭代循环优化预测结构。[20]该设计融合了蛋白质几何约束，并通过端到端训练实现高精度预测，同时输出每个残基的置信度评分，为评估预测可靠性提供量化指标。更重要的是，蛋白质结构预测算法始终处于激烈竞争中。DeepMind及其工程师团队以不断提升预测精度和实用性为目标，持续优化技术与算法，将工程调试与科学研究紧密结合，以准确性为标准推动模型进化。算法的可靠性又源于其与其他科学产品的缠结。AlphaFold因其卓越的准确性和适用性脱颖而出，正是因为其算法能够进行良性缠结。

2. 训练数据保障预测结果的可靠性

AlphaFold模型高度依赖公开可获得的蛋白质序列与结构数据，从而与具有可靠性的科学产品发生缠结。正如莱奥内利指出："尽管数字化基础设施在推理与'随机'搜索中发挥重要作用，但对数据生物学意义的解释仍植根于具身的、社会性的知识，这些知识帮助研究者评估并利用生物本体论提供的理论框架。"[21]AlphaFold的训练数据主要来自蛋白质数据库（Protein Data Bank, PDB），这些数据由全球科研共同体在具体科学实践中产生，并经过严格筛选与整理。训练所得模型在进行结构预测时，其结果的生物学意义与丰富的科学实践密不可分。数据的具身性与社会性通过算法被继承并嵌入模型，使AlphaFold与传统科学实践深度融合。算法提供技术保障，而数据的高质量与可靠性进一步奠定预测结果的可靠性。

此外，AlphaFold预测结果的可靠性独立于模型本身。AlphaFold作为可靠的预测技术并不决定其预测结果的可靠性。二者的关联仅是缠结中的一条线，无法起到稳定的支撑可靠性的作用。AlphaFold预测结果的可靠性取决于它与其他的科学产品的相互缠结。进入科学共同体评价体系后，预测结果与科学实践实现深度融合。AlphaFold并非一经诞生即被普遍认可，其地位经历批判性审视与逐步吸收。这一过程对于新想法和新技术的完善至关重要。[22]正如朗基诺（Helen E. Longino）指出："在期刊上发表一个想法或结果，并不能使其成为知识大厦的基石，其被吸收的过程复杂，包括后续引用、他人使用和修改等。实验数据和假设通过观点冲突与整合最终转化为被普遍接受的科学知识。"[22]AlphaFold预测结果本身并非直接等同于科学知识，其可靠性依赖于科学家观察结果的一致性，以及围绕这些观察展开的交流与协商所达成的共识。[23]多项研究表明，AlphaFold预测结构与实验解析结果高度一致，例如CASP14 中主链结构中位均方根偏差为0.96Å，而第二名方法约为2.8Å；[20]与冷冻电镜、[24]核磁共振[25]等实验方法比较，AlphaFold预测误差接近实验测定的不确定度。[24]这些验证使AlphaFold预测逐步获得科学共同体认可，融入科学知识生产与评估实践。

总体而言，从算法优化、数据库选择到实验验证，AlphaFold预测始终与科学家、工程师及其他科学共同体成员保持密切互动，并与既有科学产品深度缠结。AlphaFold并非孤立运作的"外星文物"，AlphaFold已融入科学社会体系，其预测质量通过同行评审验证，与蛋白质折叠及相关生命科学研究紧密相关，并在新药研发等领域获得应用。从科学实践哲学的视角来看，科学知识在实践中通过互动式稳定实现，AlphaFold的预测能力正是在这种"去稳–再稳"的过程中获得科学共同体认可。与其他科学工具类似，AlphaFold并非绝对无误，但通过与科学实践中其他流程和工具的持续整合，能够不断纠正偏差与不可靠数据，提升科学效能。[1]

在对AlphaFold预测可靠性阐述的基础上，我们可以进一步审视更具一般性的AI技术。科学知识的生成本质上是实践性的，因此，AI及其预测结果被视为可靠的科学知识意味着AI必须深度嵌入科学实践之中，并参与与其他科学产品的缠结。具体而言，这包括三个方面：一是工程师需根据科学实践需求对AI构建进行针对性调整与持续优化，不断改进算法；二是AI的训练数据必须严格来源于既有科学实践活

动，确保其与真实科学知识体系衔接；三是AI预测结果需进入科学共同体的评价体系，接受充分的协商、讨论与检验。唯有AI的生成和应用均与科学实践深度耦合，其预测成果才可能被纳入科学知识的范畴。尽管AI预测模型的运作机制对于人类而言可能显得不透明，其预测能力及预测结果的可靠性却通过模型本身与其他科学产品的缠结得以建立，这种相互支撑使得既定科学工作能够稳定顺利地完成，同时也依赖于科学共同体的社会性机制予以保障。[1]

四、AI for Science对当代科学实践的意义

作为一种具有巨大科研潜力的新兴技术，AI及其预测结果在参与科学实践缠结的过程中获得可靠性的同时，也作为新的缠结要素进入科学实践，从认识论层面丰富了科学知识的基础与形态。

首先，AI技术的发展引入了一种依赖概率推理、模式识别与异常检测的新型认知模式。这种新的认知逻辑不再以解释深度为核心，而是以预测性能作为主要评价标准；它以算法一致性补充了传统科学基于反思性论证的认知基础，从而从根本上改变了科学知识的生成方式。[26]其次，预测与模拟作为知识形态的重要性愈发凸显。以AlphaFold为代表的AI系统在科学知识生成中所扮演的角色已超出传统模拟与预测范畴。将AI预测与模拟简单归类为传统意义上的“预测”或“模拟”已显不足，更准确的表述应为“待检验的生成”，既产生新的知识内容，又处于验证机制之中，介于发现与辩护之间。再次，端到端（end-to-end）生成模式为科学知识的形态提供了新视角。AI预测知识的生成过程未明确揭示中间理论或可解释机制，即“从个体到个体”且不依赖中介理论环节。这种由复杂算法结构承载、难以还原为传统理论形式的知识，可视为一种新的知识形态，称之为“整体知识”，既不同于传统可解释理论，也不同于简单的数据堆积。将AI系统视为“掌握理论的黑箱”，并对其机制与产出进行系统研究，有助于探索新型知识生成机制，为人工智能与科学知识关系提供新的思考框架，也为科学方法论的未来演进指明方向。

需要明确的是，尽管AI在科学研究中的应用日益广泛，其预测能力在许多实践场景中表现出极高实用价值，但本质上仍是当前理论体系的延伸与拓展。在以实用性为导向的研究中，结构预测功能强、应用成熟的AI模型确实展现出辅助能力；但在理论创新研究中，AI的作用仍需谨慎评估。此外，由于算法特性，AI更容易采纳主流数据而忽视边缘化数据，而后者往往是新知识的重要来源。[27]AI系统高度依赖既有数据训练，输出虽精确，却可能在理论维度缺乏颠覆性。更重要的是，高度优化的AI输出可能呈现“完整性”假象，使研究者误以为问题已解决，从而削弱对既有理论的质疑与探索。[28]

因此，有必要界定AI在科学研究中的适用范围。正如实验方法需经长期实践与验证才能确立核心地位，AI方法也需经过科学体系检验，逐步嵌入现有研究机制，成为可监督、可验证的可靠工具。[29]AI可极大提升数据处理能力，但数据采集、问题设定与研究方向仍必须由人类科学家主导。换言之，AI可作为科研过程中的“加速器”，但无法取代科学研究中的主体地位。即便预测精度不断提升，AI本质上仍更接近“智能仪器”，可在科研某些环节独立参与，但无法替代理论建构与科学家的创造性思维。最终，AI与科学研究的协同发展将是最可持续的路径，其真正价值在于拓展与增强人类科学探索能力，而非替代。

［参 考 文 献］

[1]Zakharova, D. 'The Epistemology of AI-driven Science: The Case of AlphaFold'[EB/OL]. https://philsci-archive.pitt.edu/id/eprint/26659. 2024–11–09.

[2]Rathkopf, C. 'Hallucination, Reliability, and the Role of Generative AI in Science'[J]. arXiv preprint, arXiv:2504.08526, 2025.

[3]Dryden, D. T. F., Thomson, A. R., White, J. H. 'How Much of Protein Sequence Space has been Explored by Life on Earth?'[J]. *Journal of the Royal Society Interface*, 2008, 5(25): 953–956.

[4]Aldieri, A., Gamage, T. P. B., La Mattina, A. A. 'Consensus Statement on the Credibility Assessment of Machine Learning Predictors'[J]. *Briefings in Bioinformatics*, 2025, 26(2): bbaf100.

[5]Derry, A., Carpenter, K. A., Altman, R. B. 'Training Data Composition Affects Performance of Protein Structure Analysis Algorithms'[J]. *Pacific Symposium on Biocomputing*, 2022, 27: 10–21.

[6]Huang, B., Kong, L., Wang, C. 'Protein Structure Prediction: Challenges, Advances, and the Shift of Research Paradigms'[J]. *Genomics, Proteomics & Bioinformatics*, 2023, 21(5): 913–925.

[7]da Silva, G. M., Cui, J. Y., Dalgarno, D. C., et al. 'Predicting Relative Populations of Protein Conformations without a Physics Engine Using AlphaFold 2'[J]. arXiv preprint, arXiv: 2307.14470. 2023.

[8]Hähnel, M., Hauswald, R. 'Trust and Opacity in Artificial Intelligence: Mapping the Discourse'[J]. *Philosophy & Technology*, 2025, 38(3): 115.

[9]刘鹏. 科学实践哲学: 内涵、根源与意义[J]. 安徽大学学报（哲学社会科学版）, 2019, 43（2）: 36–44.

[10]布鲁诺·拉图尔. 科学在行动——怎样在社会中跟随科学家和工程师[M]. 刘文旋、郑开 译, 北京: 东方出版社, 2005, 418.

[11]Rheinberger, H. J. *Split and Splice: A Phenomenology of Experimentation*[M]. Chicago & London: University of Chicago Press, 2023, 78.

[12]皮克林. 实践的冲撞[M]. 邢冬梅 译, 南京: 南京大学出版社, 2004, 229.

[13]卡特赖特. 斑杂的世界: 科学边界的研究[M]. 王巍、王娜 译, 上海: 上海科技教育出版社, 2006, 36.

[14]Cartwright, N., Hardie, J., Montuschi, E., et al. *The Tangle of Science: Reliability Beyond Method, Rigour, and Objectivity*[M]. Oxford: Oxford University Press, 2022, 11.

[15]劳斯. 知识与权力[M]. 盛晓明、邱慧、孟强 译, 北京: 北京大学出版社, 2004, 128.

[16]Moult, J., Pedersen, J. T., Judson, R., et al. 'A Large-Scale Experiment to Assess Protein Structure Prediction Methods'[J]. *Proteins: Structure, Function, and Bioinformatics*, 1995, 23(3): ii–iv.

[17]Malhotra, Y., John, J., Yadav, D., et al. 'Advancements in Protein Structure Prediction: A Comparative Overview of AlphaFold and Its Derivatives'[J]. *Computers in Biology and Medicine*, 2025, 188: 109842.

[18]Baek, M., DiMaio, F., Anishchenko, I., et al. 'Accurate Prediction of Protein Structures and Interactions Using a Three-track Neural Network'[J]. *Science*, 2021, 373(6557): 871–876.

[19]Varadi, M., Anyango, S., Deshpande, M., et al. 'AlphaFold Protein Structure Database: Massively Expanding the Structural Coverage of Protein-sequence Space with High-Accuracy Models'[J]. *Nucleic Acids Research*, 2022, 50(D1): D439–D444.

[20]Jumper, J., Evans, R., Pritzel, A., et al. 'Highly Accurate Protein Structure Prediction with AlphaFold'[J]. *Nature*, 2021, 596(7873): 583–589.

[21]Leonelli, S. *Data-centric Biology: A Philosophical Study*[M]. Chicago: University of Chicago Press, 2019, 135–138.

[22]Longino, H. E. *Science as Social Knowledge: Values and Objectivity in Scientific Inquiry*[M]. Princeton, NJ: Princeton University Press, 1990, 69.

[23]Ziman, J. *Reliable Knowledge: An Exploration of the Grounds for Belief in Science*[M]. London: Cambridge University Press, 1978, 177.

[24]Terwilliger, T. C., Poon, B. K., Afonine, P. V., et al. 'Improved AlphaFold Modeling with Implicit Experimental Information'[J]. *Nature Methods*, 2022, 19(11): 1376–1382.

[25]Fowler, N. J., Williamson, M. P. 'The Accuracy of Protein Structures in Solution Determined by AlphaFold and NMR'[J]. *Structure*, 2022, 30(7): 925–933. e2.

[26]Shin, D. 'Automating Epistemology: How AI Reconfigures Truth, Authority, and Verification'[J]. *AI & SOCIETY*, 2026, 41(2): 1553–1559.

[27]Cox, A., Thelwall, M. *AI for Knowledge*[M]. CRC Press, 2025, 68.

[28]Cheng, X., Zhang, L. 'AI-generated Literature Reviews Threaten Scientific Progress'[J]. *Nature*, 2025, 641(8064): 852.

[29]Robertson, I. 'AI, Trust and Reliability: I. Robertson'[J]. *Philosophy & Technology*, 2025, 38(3): 94.

［责任编辑　王巍　谭笑］