

AI for Science 的可证伪性挑战:根源与出路

于金龙¹,孙 烨²,叶洛溪¹

(1. 北京航空航天大学人文与社会科学高等研究院,北京 100191;2. 北京服装学院马克思主义学院,北京 100105)

摘要:当前,AI 科学发现因可解释性缺失,导致证伪机制失效;既无清晰理论标靶可供批判,也令科学论辩浅化为对输入输出的表面争论。这揭示出两种不同推理模式,即人类传统科学所依赖的、基于清晰概念与因果的简单性推理与 AI 系统内在执行的复杂性推理之间难以适配。因此,化解这一危机不能仅追求 AI 的透明化,更须拓展科学共同体的认知边界——发展表征复杂性的新概念体系与形式化语言,从而在理解 AI 所揭示的复杂世界的同时,维护科学理论批判与演进的根本机制。

关键词:AI for Science;科学发现;可解释性;证伪主义

中图分类号:G315

文献标识码:A

伴随人工智能的崛起,AI for Science 作为前沿的研究路径,为人类知识生产打开了一扇全新的大门。然而,算力与数据驱动下产出的成果能否直接等同于科学理论却面临深刻拷问^[1]。AI 产出的知识主张,若要被承认为科学,就必须接受科学共同体的严格质询。其中,可证伪性的检验至关重要。因为在人类的科学框架里,只有经得起反驳与检验的论断,才能被接纳为可靠的知识;科学并非对世界的直接复写,而是人类基于特定规则为理解世界如何运作而构建的一套公共模型,这套模型必须能够被公开分享、接受质疑,并被持续修正^[2]。

但是,当前可解释性问题的存在严重威胁了 AI 科学发现的置信度。可解释性问题指的是大多数复杂的人工智能系统(尤其是大语言模型)内部工作原理的不透明性,导致人类难以理解其从输入到输出的推理过程和依据^[3]。AI 能够给出关于“是什么”的精准预测,却无法自动生成“为什么”的逻辑解释^[4]。由于无法理解 AI 得出结论的内在逻辑,科学共同体只能将其发现视为一个有待验证的统计黑箱输出,而非一个可以融入既有知识网络并进行理性辩论的科学主张。因此,AI 科学发现的置信度是脆弱的,它无法唤起科学认知所必需的内在信服,仅仅停留在高性能猜测的层面,难以转化为科学共

同体赖以思考和发展的公共信念。这种脆弱的置信度也使证伪主义所代表的科学核心批判机制遭遇了前所未有的挑战。科学进步依赖于“理论被证伪 – 诊断错误根源 – 提出更优理论”的螺旋上升逻辑,而对于不可解释的 AI,我们只能进行“预测失败 – 调整模型参数”的操作。错误难于被定位和概念化,理论的进化遇到障碍。科学批判依赖于同行对推理过程的审查与辩论,但面对黑箱,审查无从下手,理性对话成为对输入输出有效性的浅层争论,科学强大的自我纠错系统因此受到阻碍。不可解释性导致研究成果置信度低下,低下的置信度连同背后模糊的理论使得有效证伪难以执行,而证伪机制的失效又反过来让该理论无法通过科学的终极测试,从而难以获得坚实的置信度。可解释性、置信度与证伪机制三者环环相扣、形成了一个负向反馈链条。在 AI 驱动科学发现的背景下,如何确保科学理论的可证伪性这一检验功能得以充分发挥,是一个亟待解决的深层次问题。

1 原有证伪路径的失效

证伪作为科学方法论的核心引擎,并非对原有理论的简单否认,而是一个结构严谨的理性批判程

收稿日期:2025-12-31;修回日期:2026-01-23

基金项目:北京航空航天大学研究生教育教学改革研究重点项目(JGXY2025018B)

作者简介:于金龙(1976-),男,副教授,博士。孙烨(1982-),女,副教授,博士,通讯作者,E-mail:13811801311@163.com。叶洛溪(2002-),男,硕士研究生。

序。它由“猜想”和“反驳”两个彼此依存、逻辑连贯的部分构成,共同驱动知识的动态增长。

1.1 解释+找反例:证伪逻辑的双重动力源

波普尔认为科学进步的目标不是追求绝对正确的真理,而是追求可证伪度更高的理论。基于此,波普尔提出以提高可证伪度为核心的“猜想与反驳”四段图式:“问题(Problem, P1) – 试探性理论(Tentative Theory, TT) – 错误消除(Error Elimination, EE) – 新的问题(New Problem P2)”^[5]。这四段图式包含了两层重要逻辑:

第一层,确立可检验的理论对象。在波普尔建构的“猜想与反驳”的科学发现逻辑中,“猜想”绝不仅仅是一个随意的假设,而是一个试图解释世界何以如此的清晰、可理解、可公开辩论的理论体系。该理论由概念与符号构成,它不仅能够描述现象,更要能够阐明现象背后的原因、机制或者统一的原则,也就是说科学理论要具备解释力。波普尔指出:“概率高的陈述在科学上引不起兴趣的原因就在于缺少解释力”^[5]。这一步之所以关键,是因为它为证伪明确了对象与标准,为科学共同体提供了一个公共的、可操作的检验框架,并规定了在何种经验条件下该理论应被否定。没有这样的概念化表达,所谓“科学发现”就只是无法被理性辩驳的预测,其认知价值也无从确立。正是理论的可解释性使系统性的理性批判与经验检验成为可能。

第二层是执行严格的经验检验。只有当理论被明确表述之后,证伪才能进入执行阶段,即主动寻找与理论核心预言相矛盾的经验反例。这不是被动观察,而是针对理论最薄弱环节所设计的严格检验。一个被确证的反例,凭借其与理论预言在逻辑上的直接冲突,构成对该理论的否决性判决,逻辑地宣告该理论在描述世界时的失败。这一过程体现了科学的自我纠错机制,使理论具备了被经验推翻的可能性。

上述两层逻辑共同构成一个整体。概念解释作为证伪的重要前提,使科学主张能够具备在实践中被检验的基本资格和条件;而寻找反例则成为证伪的重要实践手段,它能够借助经验实现对理论的批驳与迭代。

1.2 AI 科学发现的证伪存在解释力不足

AI for Science 与人类传统科学发现都能够完成证伪过程中对反例的寻找,两者都对理论与观察中的矛盾高度关注,并将其作为系统自我修正的重要

动因。证伪主义将这种矛盾视为对理论的否决——一个针对某一理论的确凿反例就能够形成对该理论的决定性证伪^[6]。而贝叶斯推理构成了 AI 驱动科学发现的核心机制。与证伪过程不同的是,贝叶斯推理采用量化的方式来处理反例,每个不利证据都会削弱先验概率的置信度^[7]。尽管对反例的推理逻辑和目标存在分歧,但它们却存在着共性,那就是两种方法都是通过动态迭代过程逼近更优状态:证伪主义通过“猜想 – 反驳 – 新猜想”的循环对理论进行更替,贝叶斯推理通过“先验信念 – 证据更新 – 后验信念”的循环对信念进行渐进式优化。它们的终极目标不是为了证明某个理论的恒久正确性,而是为了能够设计出一套能在经验世界的持续诘问下,始终具备自适应性的算法,从而推动认知体系在猜想和经验检验中不断演化^[8]。

本质上,二者在找反例过程中体现出的一致性源于科学理性内在的防御机制。它们承认认知的主体会犯错,并且主动构建了一整套较为形式化的程序用来确保这种错误不仅能够显现出来,也能够得以系统化地处理,并最终转化为推动认知体系不断成长的养分。在科学发现中,证伪主义承担了仲裁的责任,它要借助批判性检验对那些存在根本性缺陷的理论假设进行识别和淘汰;而贝叶斯推理则体现为一种渐进式的修正过程,在研究中持续动态调整并优化对各类信念的可信度评估。

AI 科学发现虽然在数据层面能够识别出与现有模型预测不符的异常模式,即数据上的反例,却缺失了执行证伪逻辑中所必须的前提——构建可供检验的概念化理论解释。因此, AI 科学发现难以完成从“数据反例”到“理论反例”的跃升。大模型本质是对复杂的数据相关性的高维函数拟合,其表征以缺乏明确语义的高维特征向量为载体,这些载体难以像符号化逻辑体系那样被人类轻易理解。这也使 AI 科学发现难以满足科学理论建构所必需的三类核心认知功能:首先,由于理论命题的缺失, AI 科学发现难以形成可被科学共同体批判并发展的完整主张。作为公共辩论对象的属性是科学理论的核心特征,因此它必须由清晰概念构成且逻辑自洽,并能够接受科学共同体的检视、质询与修正。例如:在“所有物种都从共同祖先逐渐演化而来”这一命题中,不仅包含着“物种”“演化”“祖先”等语义十分丰富的概念,更通过逻辑连接构成了一个可被检验、可被反驳的完整主张。而大模型在学习过程中却依靠统

计模式提取与参数优化,最终产出是输入到输出的映射函数,而非具有命题结构的陈述。模型即使借助自然语言生成技术输出文本描述,这些描述也只是训练数据中语言模式的复现,本身并不承载模型自身的信念或论证。因此,AI 科学发现难以步入到科学辩论的“主张-反驳-辩护”循环中去,其输出也只能是缺乏互动的结论式宣告,而非对话式的理论框架。其次,由于因果解释的缺位,AI 科学发现难以识别出反常与理论内部逻辑冲突,因此难以提供动摇特定理论的逻辑理由。当一个反例被提出时,其说服力不仅来自于经验数据,更来自该数据与核心理论之间的逻辑矛盾。例如,迈克尔逊-莫雷实验的零结果之所以构成对牛顿力学体系的挑战,是因为它在牛顿引力框架内无法被完全解释,从而暴露出牛顿经典力学体系的潜在问题。而 AI 在对异常值的识别过程中,是依据统计偏差发出信号,无法构建异常和理论漏洞之间的因果解释链条。AI 难以将识别到的数据异常关联到特定理论假设的失效或修正上。这也使 AI 发现只能停留在数据与模型预测不符的现象描述,难以进入理论内部对何种预设或推理环节存在问题进行深层次解读,也难以直接驱动理论在此基础上进行批判与更新。第三, AI 科学发现缺乏理论所必需的可证伪性自觉。波普尔指出,科学理论应“自我设限”,也就是要明确承诺在何种经验条件下该科学理论将被视为是错误的^[5]。这种可证伪性成为了科学理论得以进步的风险契约。然而,深度学习模型需要最小化经验风险,其优化过程本质上是在不断消除自身可被证伪的可能性,即模型难以主动声明:若在 X 条件下出现 Y 结果,则证明我的核心机制是错误的,这就与科学理论主动划定边界、承诺在何种条件下自我放弃的风险契约精神不符。相反,它倾向于通过增加参数复杂性或利用数据中的虚假相关对异常值进行覆盖,从而避免被反驳。这种内在逻辑使其产出的成果天然缺乏理论应有的自我颠覆潜能,更像是一个不断自我加固的预言黑箱。

AI for Science 的可解释性问题并非仅仅源自技术本身,而是其方法与人类传统科学理论建构逻辑之间的差异。较之于人类传统科学发现路径, AI 更擅长对异常值的识别,却难以引发理论危机,并完成对固有范式的颠覆。这一局限最终导致了 AI 科学发现的证伪机制失灵。

2 可证伪性挑战的根源:科学发现的复杂性表征突破概念和符号内涵的边界

上述分析表明, AI 在理论构建环节的不足并非偶然的技术瓶颈,而是恰恰指向了可解释性危机的本质:人类固有的简单性认知模式与 AI 的复杂性认知模式之间的错位。可解释性并非单一的、静态的标准,其实现难度与核心矛盾,直接源于人类与 AI 所运用的推理模式之间存在的差异。这种差异可具体归结为两种不同的推理模式的冲突:人类传统科学所依赖的、基于清晰概念与因果的简单性推理与 AI 系统所内在执行的复杂性推理之间的解释鸿沟。

2.1 简单性推理及其可证伪性

本质上,人类围绕因果展开对世界的理解与探索,因此推理过程也本能地倾向于寻找模块化的、可分解的、线性的因果关系^[9],形成以简单性因果链为核心的推理模式,即简单性推理。简单性推理在人类发展的历史上成效显著,曾经帮助人类取得了认知上的重大突破,这成效的背后并非源于简单性推理能够对全部世界的复杂性给以完美地刻画,而在于能够帮助人类在自身所特有的认知架构及外在经验世界之间建立解释的桥梁,从而能够持续的相互印证。

首先,简单性推理的形态受限于人类的生理构造。人类大脑并非一台能够处理无限变量和复杂网络的超级计算机。人脑工作的核心约束来自有限的工作记忆容量,这使得人脑难以并行维持和处理长链的、多要素混杂交织的复杂性交互^[9]。为了能够更加高效地思考,人类在生理的限制下发展出了符号系统和概念压缩两种策略:符号系统能够借助语言、数学符号和图表将思维过程卸载到外部,从而实现认知的分步操作与审查^[10];概念化压缩则能够借助高度抽象的范畴或新的变换方式,将纷繁复杂的现象压缩为少数几个可被人脑认识和理解的基本认知单元。简单性推理正是这种认知偏好的产物。它适用于那些因果结构具有高度可分解性的系统。在这类系统中,整体现象可以被视为局部、单向或弱反馈的因果环节的线性叠加。例如,在经典力学体系,物体的运动被理解为各分力作用的矢量合成;在基础化学中,总反应被视为基元反应的加和。这种模块化、层级化的因果观,使得我们可以将复杂问题分解为一系列独立的、可在工作记忆中顺序处理的

小问题,从而与我们有限的处理能力完美契合。

这种推理模式在操作上体现为一种基于符号与概念的逻辑闭环。我们将辨识出的因果模块,用离散的符号进行定义和表征,概念则概括了跨情境的因果共性,成为我们思维的快捷方式。整个推理过程无论是从理论(因)推演现象(果),还是从现象(果)反推机制(因),都能被展开为一系列符号操作步骤。其最大的优势在于公共可审查性与可逆性:任何推导步骤都可以被冻结、质疑和辩论;任何预测失败都可以被回溯,以定位错误究竟发生在计算过程、边界条件还是核心因果假设本身。这确保了在此范式内的知识增长是严谨、可控且可公共累积的。而这种结构化的简单性推理由于步骤清晰、逻辑可追溯,能够被人类心智把握与理解,进而能够被完整表述并公开检验,从而具有可证伪性。

因此,在波普尔的证伪框架中,“简单性陈述就比不那么简单的陈述得到更高的评价……因为它们更可检验”^[2]。简单性直接关系着可检验性和可证伪度,而可检验性或可证伪度又是好的科学理论所必备的特质。波普尔认为简单理论更易暴露于观察和实验的潜在反驳之下,从而承担了更大的风险,科学的进步并非通过不断累积被证实的事实,而是通过提出更大胆(因而更简单)、更具风险性的猜想,并经受无情的经验检验来驱动。一个试图通过增加复杂条款和特设性调整来规避反驳的理论,是“退化的”;而一个被严格检验后幸存下来的简单理论,则拥有更高的逼真性和认知价值。

2.2 复杂性推理及其证伪难题

与简单性推理以模块化、可分解的认知方式和线性因果逻辑构建对事物的理解不同,复杂性推理则直面具有因果涌现特性的系统本体结构,即整体不等于部分之和,系统的宏观规律也不可还原为诸微观行为的线性加总。这种非线性相互作用导致因果链不可追踪,涌现性使规律难以推导^[11],例如蛋白质折叠过程中从海量结构和作用关系中“选择”其天然态,其内部机制本质上是非线性的,无法通过将组分的行为进行简单叠加的方式来进行预测^[12]。系统的整体行为源于组分间复杂的相互作用网络,而非一条清晰可分步追溯的线性因果链。

面对因果涌现,人类符号系统遭遇失灵。首先,符号的离散性难以回应因果涌现的整体性。因果涌现强调的是系统在宏观尺度呈现出新的特征或者行为,而这些新的特征或者行为是不能通过低层次要

素间的线性叠加关系进行预测的。蛋白质折叠涉及到多尺度下的相互作用,整体功能源于这些部分和要素的非线性作用。因此,人类通过固有的因果逻辑推导难以从中解释和预测蛋白质折叠过程中的涌现。其次,概念的概括性在涌现面前被削弱。在复杂性推理面前,难以存在一个简单的概念能够对所有实例的共性进行抽象概括,因为每个个体都是不同的。例如,每个蛋白质的天然氨基酸序列都是高度特异的^[12]。这些独特性是难以被概念和符号进行简化、变换从而完美复现全部微观现实。复杂性推理在演绎环节面对着呈指数级增长的计算复杂度,这对人类的固有计算能力来说几乎不可能实现^[13]。而复杂性推理在归纳环节也同样遭遇困难,在科学实验中所观察到的折叠现象无法借助概念载体构建形式化假说,因为这类复杂性问题已经超越人类概念化能力,进而也使得基于复杂性推理的证伪举步维艰;科学家在无法理解 AI 模型的“核心假设”的情况下难以设计判决性实验,当理论预测失败时,更无法进行逻辑归因,此时的大模型只能对超参数进行盲调,方法论也就从自循环模式退化为试错、调试模式,科学进步机制失灵。

因果涌现超越了人类固有解释架构,这与人类客观生理基础有着密切的联系:大脑工作时的运行内存容量十分有限,大约能够容纳 4-7 个信息组块,这种生理结构使人脑更加适合处理步骤清晰、可以分解的线性任务^[9]。然而,在面对诸如蛋白质折叠这样涉及数百万原子及其非线性相互作用的问题时,便显得难以为继。复杂系统内部要素间存在海量非线性相互作用,这种非线性相互作用远超人脑串行处理能力的极限。与此同时,人类赖以扩展思维的外部符号工具在面对复杂性推理时也可能失效,因为这些符号工具是线性和离散的,更擅长存储数据与分步推演。但复杂系统内部的非线性交互作用难以分解和简化。人类无法在蛋白质折叠过程的某一个步骤按下暂停键并对其进行孤立检验,因为其中每一步的状态都对全局的宏观状态有所影响,蝴蝶效应正是这种非线性系统中初值敏感性的典型表现。符号系统更擅长对复杂现象进行静态表征,却难以充分呈现其内在的动态演化与复杂交互作用。而以人类目前概念化压缩的能力而言,尚不足对这类复杂性推理给以很好的解决。对于规律清晰、高度可重复的因果关系,人类可以将其抽象为统一概念进行把握。然而,在复杂推理中,却难以强行

将系统多样性都笼统进行概念化压缩,这样做的后果便是系统的异质性难以得到彰显,最终得到的解释也将失去实证内容与预测力,从而背离科学解释的根本目的。

人类认知架构在演化过程中并未发展出直接理解非模块化、强耦合且跨尺度涌现现象的能力。当前人工智能在预测上取得突破,却难以给出符合人类认知的逐步解释,这一现象恰恰暴露了人类认知模式自身的边界,而不仅仅是一项有待攻克的技术缺陷。我们需要从更深层次思考当前挑战:理解复杂性,或许需要超越我们传统基于简单性推理的认知范式与以概念和符号作为载体的解释框架。上述局限进一步影响了实证科学的方法论基础。当人工智能成为知识发现主体时,其产生的“知识”常以无法还原为人类可以独立表述与验证的逻辑命题或公式,这使得依赖可表述假设进行批判检验的证伪主义陷入困境:科学共同体因缺乏可公共审议的理论批判对象,难以开展有效的理论辩驳,科学验证逻辑与知识生成机制之间出现了需要弥合的理论断裂。

3 AI 科学发现证伪的可优化路径

波普尔的证伪主义与贝叶斯方法在科学进步的驱动逻辑上具有内在的一致性,它们都强调反例在深化认识上具有重要作用。无论是证伪主义通过反驳来淘汰理论,还是贝叶斯方法通过概率更新来调整对理论的置信度,它们都将寻找反面证据置于科学检验的核心,并视其为推动理论变革的关键动力。这种一致性表明,通过严格的经验挑战和动态的信念修正来实现知识迭代,在人类科学发现与 AI 科学发现模式中是共通的。正因如此,基于贝叶斯推理的 AI 科学发现,其过程与结论在原则上具备了可检验、可证伪的理论基础。但将这一原则性基础转化为现实的科学理解,仍要应对好可解释性危机的挑战。危机背后是两种认知范式的深刻张力:人类传统科学认知依赖于直觉、因果叙事和追求简约的理论模型,而 AI 则擅长捕捉那些超越人类直觉的理解框架的高维、非线性、复杂的模式关联。因此,这场危机的本质,是人类既有的科学认知框架在理解高度复杂系统时固有局限的显现。AI 并非给出了错误答案,而是给出了一个我们现有概念工具尚无法充分解析的复杂答案。应对这一挑战,不能仅仅单向地要求 AI 变得透明,真正的突破口在于一场双向

的认知重构。

3.1 搭建认知桥梁:在大模型中内嵌知识单元

当 AI 揭示的复杂性超越人类概念体系的直接承载极限时,拓展解释边界的直接途径是主动在技术与方法层面进行架构性创新。其中一个重要方面即利用并整合现有知识工具,结构化一系列中间工具——即搭建知识过渡框架。这些过渡框架的目标,正是将 AI 的复杂性输出,系统地转译为人类传统科学思维能够有效处理与审验的信息,从而在已知与未知之间建立可通行的认知桥梁^[10]。

构建知识增强的 AI 系统能够为该问题提供可行的方案,其实践路径可以借助领域知识图谱与大语言模型的深度耦合。知识图谱作为人类长期以来积淀的结构化、概念化知识体系的载体,能够像一张概念地图为 AI 的复杂模式识别建立一个可以参照和解释的坐标框架。当 AI 从数据中发现一个难解的复杂性关联时,它可以将关联映射在知识图谱的特定节点和路径上,从而生成定位性的解释。大语言模型则可以在此基础上承担转译的功能,将这些结构化的映射转化为连贯的自然语言,并将相关背景进行补充。这一过程的本质,就是使 AI 在数据中挖掘到的统计相关性能够更大限度地与人类固有的概念、符号体系相融通,从而消除 AI 科学知识 with 人类知识之间的隔阂,使人类能够凭借熟悉的认知框架,对 AI 科学发现进行更加深刻的理解、评估,并对其合理质疑。

构建此类知识过渡框架,其本质是在人类既有认知版图与 AI 所揭示的复杂性之间,架设一座能够进行阐释与转化的桥梁。它不执着于一劳永逸地理解并解决复杂性问题,而是采用一种工程化智慧,显著提升 AI 科学发现在当下的可用性、可信用度与可整合性。这是一种至关重要的过渡性范式,它既积极调用原有知识体系的力量来稳住前沿,也为未来更深刻的理论突破与认知融合准备了条件、划定了问题。通过持续迭代与丰富这些过渡框架,我们得以让 AI 强大的发现力,更顺畅、更可靠地注入人类传统科学知识进步的洪流之中,从而实质性地推动人类解释边界向前延展^[14]。

3.2 深度萃取规律:提高对复杂性的压缩效率

面对复杂性的根本挑战,除了上述通过构建外部认知过渡框架来桥接人类与 AI 科学发现之外,还存在另一条并行且更为根本的向内探求的路径,即系统性地提升人类自身的复杂性推理能力,以应对

信息过载与复杂性交互作用所带来的理解屏障。这条路径的核心策略在于寻找到更多变换的形式。其中,“压缩”能够为人类提高 AI 科学发现的可解释性和可证伪性,提供一条更为根本的路径。

压缩绝非简单的简化或选择性忽略。它区别于将复杂现象强行塞进传统解释框架的“削足适履”,而是一种积极的、创造性的认知重构^[15]。其本质在于,通过对复杂系统中深层规律与稳定不变性的主动探索,找到一种兼具信息密度与逻辑一致性的高阶表征,从而实现对系统庞杂性的有效约简。它是主动探索并发现复杂系统中存在的深层规律、稳定不变性以及更高效的表征方式,从而将系统庞杂的信息与高维度的相互作用网络压缩,在信息密度和逻辑一致性上都得以优化的表达形式。这一过程的目标在于对复杂性精华的萃取,最终为了生成一种既保留原系统核心动态特征与预测能力,又在形式上足够凝练、足以被人类心智所把握,并能无缝嵌入既有科学推理与辩论框架的知识成果。这种被压缩后的知识,因其逻辑结构更清晰、变量更少、核心原则更突出,从而天然地具备了更强的可理解性、可传播性、可辩论性,以及至关重要的——可证伪性。

例如:在大模型的压缩过程中,通过 Scaling Laws^[16]捕捉系统内部要素间的相互作用关系,为后续的蒸馏确定目标阈值、算力和性能边界,已经成为大模型提升压缩效率的常见解题思路。Scaling Laws 是指那些看似不相干的对象和各种宏观表征随系统规模变换的规律,这些规律通常可以用一个简单的幂律公式概括。在大模型中,Scaling Laws 能够通过揭示模型规模和模型性能之间的幂律关系,为以蒸馏为代表的压缩技术提供更加高效的思路,使其从无差别的“做减法”转变为一种以复杂性为基础的、目标明确的精炼过程。蒸馏是将大模型(教师模型)的知识传递给小模型(学生模型)的一种压缩技术^[17]。作为一种简化方法,其目标是在减少模型的复杂度与资源消耗的同时,尽量保留其核心性能。而 Scaling Laws 则为以蒸馏为代表的简化的“度”提供量化标准:根据 Scaling Laws,可以在蒸馏之前精准的预期学生模型在缩小规模后的性能下限,从而确定实现最高蒸馏效率的最优参数量;同时,Scaling Laws 也能够揭示计算量与性能关系之间的“饱和点”,从而在蒸馏训练中科学的分配计算预算,避免浪费,提升整个蒸馏流程的效率。

这一混合了整体与局部、数值与符号、可计算性

与可证伪性的表征架构的最终目标在于,确保 AI 驱动的科学发现在理论抽象的每一层级上,都能接受明确的、可操作的证伪性检验。这要求研究者从被动接收模型输出,主动转向对高维关联进行理论建构与精炼^[18]。其核心使命是对 AI 所揭示的高维、弥散的原始关联与模式进行系统性的提炼与升华,将其加工为可供逻辑推演与批判性审视的形式化概念体系,而持续提高压缩效率正是这一建构过程的效率核心。当系统复杂度逼近人类概念化承载的临界阈值,传统证伪的线性反例模式虽显乏力,但其内核恰恰是对抗 AI 意义真空的重要动力。

在 AI 驱动的科学发现时代,证伪主义非但未因智能化时代的来临而褪色,反而在涌现的复杂性面前体现出更根本的价值——它成为抵御“算法幻象”的最后防线。

参考文献:

- [1] 李正风. AI for Science: 强化科研伦理规范防止 AI 技术滥用[J]. 科学新闻, 2024, 26(6): 14 - 16. Li Z F. AI for Science: Strengthening scientific research ethics to prevent misuse of AI technology[J]. Science News, 2024, 26(6): 14 - 16.
- [2] 卡尔·波普尔. 科学发现的逻辑[M]. 北京: 中国美术学院出版社, 2007: 119 - 122. Popper K. The Logic of Scientific Discovery [M]. Beijing: China Academy of Art Press, 2007: 119 - 122.
- [3] 董春雨. 从机器认识的不透明性看人工智能的本质及其限度[J]. 中国社会科学, 2023(5): 148 - 166, 207 - 208. Dong C Y. Exploring the essence and limits of Artificial Intelligence from the perspective of machine cognition opaqueness [J]. Social Sciences in China, 2023(5): 148 - 166, 207 - 208.
- [4] Rudin C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead [J]. Nature Machine Intelligence, 2019, 1(5): 206 - 215.
- [5] 卡尔·波普尔. 猜想与反驳[M]. 北京: 中国美术学院出版社, 2003: 3 - 82. Popper K. Conjectures and Refutations [M]. Beijing: China Academy of Art Press. 2003: 3 - 82.
- [6] 王晶金. 证伪主义与科学史[M]. 西安: 西北工业大学出版社, 2019: 29. Wang J J. Falsificationism and the History of Science [M]. Xi'an: Northwestern Polytechnical University Press, 2019: 29.
- [7] 赵汀阳. 人工智能还给人类的思维难题[J]. 中国社

- 会科学,2024(8):101 – 123,206. Zhao T Y. Artificial Intelligence and the challenges to human thinking [J]. Social Sciences in China, 2024(8): 101 – 123,206.
- [8] 孙昌璞. 贝叶斯哲学视角下实验“证实”理论的科学认知重构[J]. 自然辩证法研究,2025,41(9):17 – 25. Sun C P. Reconstruction of scientific cognition of experimental confirmation theory from the perspective of Bayesian Philosophy [J]. Studies in Dialectics of Nature, 2025,41(9): 17 – 25.
- [9] Cowan N. The magical number 4 in short – term memory: A reconsideration of mental storage capacity [J]. Behavioral and Brain Sciences, 2001, 24 (1): 87 – 114.
- [10] Wilson M. Six views of embodied cognition [J]. Psychonomic Bulletin & Review, 2002, 9(4): 625 – 636.
- [11] 刘劲杨. 当代整体论的形式分析 [M]. 西安:西安交通大学出版社,2018. Liu J Y, Formal Analysis of Contemporary Holism [M]. Xi'an: Xi'an Jiaotong University Press, 2018.
- [12] Jumper J, Evans R, Pritzel A, et al. Highly accurate protein structure prediction with AlphaFold [J]. Nature, 2021, 596(7873): 583 – 589.
- [13] Abramson J, Adler J, Dunger J, et al. Accurate structure prediction of biomolecular interactions with AlphaFold3 [J]. Nature, 2024, 630(8016): 493 – 500.
- [14] Deng X T, Li H Y. How large language models need symbolism [J]. National Science Review, 2025, 12 (10): nwaf339.
- [15] Dele tang G, Ruoss A, Duqenne P, et al. Language Modeling is Compression [C]. International Conference on Learning Representations, 2024.
- [16] Kaplan J, Mccandlish S, Henighan T, et al. Scaling laws for neural language models [EB/OL]. arXiv:2001.08361, 2020 [2025 – 12 – 27]. <https://arxiv.org/abs/2001.08361>.
- [17] Hinton G, Vinyals O, Dean J. Distilling the Knowledge in a Neural Network [EB/OL]. arXiv: 1503. 02531 [stat. ML], 2015 [2026 – 02 – 01]. <https://arxiv.org/abs/1503.02531>.
- [18] Wang H, Fu T, Du Y, et al. Scientific discovery in the age of artificial intelligence [J]. Nature, 2023, 620: 47 – 60.

The challenge of falsifiability in AI for Science: Roots and pathways

YU Jin – long¹, SUN Ye², YE Ming – xi¹

(1. Institute of Advanced Studies in Humanities and Social Sciences, Beihang University, Beijing 100191, China;

2. School of Marxism, Beijing Institute of Fashion Technology, Beijing 100105, China)

Abstract: With the rise of artificial intelligence, AI driven scientific discoveries have become a new paradigm in scientific research. However, whether the research path that relies on data and algorithms can truly constitute “scientific discoveries” still faces fundamental philosophical and methodological questions. Scientific epistemology requires that knowledge claims must undergo falsifiability testing. However, current AI scientific discoveries often appear as “black boxes” due to the lack of interpretability, with outputs mostly statistical predictions rather than logically complete rational claims, resulting in the failure of falsifiability mechanisms; there are no clear theoretical targets to criticize, and scientific debates have become superficial debates on input – output. This reveals a deeper crisis: the traditional scientific cognition based on human intuitive reasoning is difficult to adapt to the complexity patterns revealed by AI. Therefore, resolving this crisis cannot only pursue the transparency of AI, but also requires promoting the expansion of the cognitive boundaries of the scientific community – developing new conceptual systems and formal languages that represent complexity, so as to understand the complex world revealed by AI while maintaining the fundamental mechanisms of scientific theory criticism and evolution.

Key words: AI for Science; scientific discovery; interpretability; falsificationism