法官、代数学家与赌徒：科学论证的三重契约如何被深度学习折叠为一种新的认知垄断

深度学习带来的不只是“理解与预测”的冲突。它更深地改写了科学共同体内部谁还能独立行使论证资格：法官的裁决、代数学家的建构与赌徒的预测不再平权。文章以 AlphaFold2 为现场，提出“论证非平权”，并给出三组可操作的证伪条件。

作者 阳涌 · 约 2.3 万字 · 发表于2026年7月23日

摘要

围绕深度学习的争论一直在“理解与预测”的旧战场上反复拉锯，其根源不在任何一方的论证力不足，而在于争论双方共享一个从未被标记出来的预设——科学共同体内部论证功能（法官的裁决、代数学家的建构、赌徒的预测产出）必须维持平权的制度性耦合。本文通过五连问从当前争论的表层张力逐层追出这一预设，并指认：深度学习系统不是简单地制造了知识不可遍历性，而是首次从技术架构层面系统性地撤除了法官功能与代数学家功能赖以独立运作的替代性缓冲工具。本文将这个新现象命名为“论证非平权”。通过重返结构生物学中AlphaFold2引发的具体实践位移——论文写作方法段从“公证书”退化为“说明书”、审稿从论证检查迁移为外围审计、教学训练中代数学家学徒期的中断——本文展示论证非平权如何从“赌徒功能单极膨胀”的技术条件中结晶为一组自我维持的认知制度性条件，其核心机制不在于论证通道被收窄，而在于即便在通道收窄之前，那些让不同功能位置能够对等介入论证的公共检验工具已被架构性地取消了。文章进一步论证，论证非平权与知识不可遍历性处于条件性解耦的关系：可遍历性的丧失是其触发条件，但它一旦在替代性缓冲工具缺席的条件下形成，就获得了独立于可遍历性的制度惯性——这意味着“让黑箱变透明”的技术路线不可能从根本上修复它，因为解释权的时间差优势在解释发生之前就已完成分配。全文以三组可操作的经验证伪条件收束。

关键词：论证非平权；科学共同体；深度学习；功能分工；替代性缓冲工具

一、引言：为什么关于AI黑箱的争论总是原地打转？

过去五年里，没有任何一场科学哲学争论像“AI黑箱是否威胁科学理解”这样，同时具备了高强度、高频率和高重复率三种特征。每一次新的AlphaFold级模型发布，都会触发一轮几乎相同的攻防：一方发出警告，指出当预测不能被逐行检验时，科学的自我纠错机制便在根基处受损；另一方则回应，说这种警告不过是旧派学者对数值模拟的古老恐惧换了一张新面孔——既然气候模型和密度泛函计算早已成为可敬的科学工具，深度学习无非是把“不可遍历”的指针再往深处推了一格，凭什么这一格就是断裂？

这场争论的参与者不是不聪明。恰恰相反，双方都握有令人尊重的论证资源。但本文的诊断是：争执之所以无法推进，不是因为某一方论证不足，而是因为双方共享了一个从未被标记出来的默会前提。这个前提既不是“预测重于理解”，也不是“理解必须优先”，而是一个比二者都更底层的、关于科学共同体自身是什么的预设。正是因为双方都站在这个预设之上，他们每一次交手，都只是在同一张桌子的两侧互相递刀，而从未问过这张桌子的腿是否还站在地上。

这张桌子的名字，可以暂时叫作“对等纠错模型”。它的核心图景是这样一幅画：科学共同体由一群具备同等论证能力的行动者组成，他们可以彼此独立地检验同一条论证链上的每一步推导，在发现错误时能够精确地指出错误的位置，并将这一指认以公共语言通报给整个共同体。无论这个共同体是牛顿时代的皇家学会通信网络，还是当代《物理评论快报》的同行评审系统，一个底层假定始终在场——所有参与论证的人，都站在同一套可公度性地图上，拥有对等的论证工具和纠错权限。你可以质疑同行的结论，但你不能质疑同行拥有质疑你的资格；而“资格”之所以不被质疑，是因为它建立在每个人都能拿起同一套数学语言、独立重走一遍论证路径的基本能力之上。

截至此处，这段描述读起来像是在陈述一个常识。但它不是常识。它是一组特定制度技术——印刷术使得同一组方程能够被分散在欧陆各地的学者同时核对；期刊制度的匿名评审将审稿人从人身依附关系中剥离，赋予她以公共论证语言为基础的独立评判资格；十九世纪实验室教学的标准化把“独立复现论证步骤”的训练批量分发给每一届新生——在历史上逐步结算出来的脆弱均衡。托马斯·库恩的范式理论已经揭示了这一均衡的脆弱性：常规科学之所以能够进行，正是因为研究者共享了一套范例、标准操作与不言明的承诺，而不需要逐次地重建论证工具本身。也就是说，即使在库恩描述的“不可通约”时刻之间，常规科学内部的论证平权也是被制度技术维系的，不是自然赋予的。

理解这一点的发生学意义在于：它意味着论证平权不是科学的“本质”，而是一种在特定历史条件下被发生出来、因而也完全可能在新的技术条件下被重新配置的制度安排。它不是被深度学习“背叛”的黄金时代，而是深度学习得以被准确诊断的历史对照。

这就是为什么关于AI黑箱的争论总是在“可遍历性”这个点上卡住。因为双方都默认：如果论证通道是可遍历的，那么科学共同体就还在对等纠错的轨道上；如果论证通道不可遍历，那么轨道就崩

了。整场争论因此变成了“通道还在不在”的拉锯战——一方指出通道在收窄，另一方回应说通道从来就没有完全敞开过。但双方都没有往另一个方向看：也许真正的问题不在于通道有多宽，而在于谁还能走上这条通道——以及，走上这条通道所需要的那些制度技术，是否正在被一种新的知识生产模式从物质基础上抽走。

当深度学习系统被引入科学知识生产时，它所做的不仅仅是在论证地图上挖掉了一块可遍历的区域。它更隐蔽但也更根本的动作，是重新分配了共同体内部的论证资格。一个审稿人面对经过人类推理步骤的传统论文，与面对一篇以AI预测为核心证据的稿子，她手中的论证工具发生了质的改变。她不是面对一条更难走的论证路——她是被从论证者的位置上，默默地移到了预测消费者的位置上。在那一刻，不再是论证通道变窄了；而是论证功能的分工结构，第一次出现了系统性的单向压缩。

这就是本文要指认的新现象。现有文献中没有任何一个术语能精确地装下它。它不是“知识不可遍历”——科学知识社会学早已揭示，科学家在大多数时候是在信任他人报告的基础上运作的，并非每一项知识都亲自遍历。它不是“认知劳动分工”——Philip Kitcher等人已经令人信服地论证了共同体中认知角色的不对称是推动科学效率的制度安排。它也不是“权力不对称”或“认知不公”——这些概念各切各的维度，但都没有命名这样一种特定的位移：在一个原本依靠论证功能分工作为制度性缓冲的认知共同体中，深度学习系统首次剥夺了某些功能位置得以运作的替代性缓冲工具，从而使得分工从一种“可被挑战的不对称”变成了“无法被挑战的单极化”。本文将这个概念命名为论证非平权（argumentative disenfranchisement）。

在展开论证之前，有必要将本文的概念与几组最邻近的既有传统做一次严格的划界，以澄清论证非平权到底切开了什么旧范畴装不下的新东西。

第一组：布迪厄的科学场域理论。皮埃尔·布迪厄描述了科学共同体内部围绕学术声望、机构隶属、设备与经费支配权展开的权力争夺。布迪厄的深刻之处在于，他证明了科学知识的生产从来不只是纯粹认知操作的结果，而是社会空间内资本配置的投影。但布迪厄的分析层次是“拥有资源的资格”——谁因为拥有更多的经济资本、文化资本和符号资本而能够在场域中占据支配地位。论证非平权关切的是一个不同的层次：“行使论证功能的资格”。即使在一个完全实现了布迪厄式资本平等的理想共同体中——所有大学配备同等算力集群，所有训练数据对所有人开放，所有模型权重完整开源——当一位审稿人面对一篇以黑箱预测为核心证据的论文时，她仍然无法独立地、在论证层面指认“这个预测为什么在这里犯了错”。不是因为她缺显卡，而是因为那个可以让她做出这种指认的推理通路，压根就不以可被她检验的格式存在于权重文件之中。布迪厄的资本逻辑可以解释谁能跑模型，但不能解释跑了模型之后谁还能检验模型。前者是社会-经济层次的资格分配，后者是认知-技术层次的资格分配。二者不重合。

第二组：认识论依赖与认知劳动分工。John Hardwig正确地指出，现代科学中科学家在大多数时候无法亲自检验同行的全部推理步骤，而必须依赖于对他人证词的信任。Kitcher在此基础上将认知劳动分工理论化，论证了这种不对称不一定是缺陷，而可能是认知效率的制度性安排。上述传统处理的问题可以概括为：当个体科学家无法遍历论证链时，共同体的认识论安全网如何被维持？论证非平权处理的问题则不同：当论证者手中的替代性缓冲工具——那些让她在无法遍历全部推理步骤时仍然能够执行法官职能的公共论证语言——本身也被黑箱化时，共同体的这一安全网是否已经出现了无法用传统分工逻辑修补的破洞？传统认识论依赖的不对称，是建立在“信任可以用公共论证工具在第二层被检验”这一前提上的：你无法亲自重跑LHC的实验，但你可以检查合作组公布的统计方法与误差模型中的逻辑。而论证非平权所指出的，正是这种“第二层检验工具”也被撤消的时刻。在这个意义上，科学家之间的认识论依赖从“可被挑战的信任”变成“无法被挑战的接受”，不是因为信任的深度增加了，而是因为挑战的工具消失了。

第三组：认识论不公与专家权威。Miranda Fricker的“证言不公”与“诠释不公”深刻地揭示了社会权力如何系统性地贬低或抹消某些群体的认识论贡献。然而，论证非平权与Fricker的框架有一个决定性的机制差异。Fricker处理的不公，其运作机制是社会性的：听者因为对说者所属社会群体的偏见，而赋予其证言更低的可信度；或因为集体诠释资源的贫乏，而使某些群体的经验无法被公共表述。论证非平权则不同：它的运作机制是架构性的，而非社会-偏见的。当一位审稿人面对AlphaFold2预测时，她无法独立检验其论证逻辑，不是因为她的社会身份被贬低，也不是因为她缺乏某个诠释概念，而是因为那个可以让她执行检验的推理通路，并不以任何可陈述的格式存在。这不是“她的判断被不公正地忽视”，而是“她赖以做出判断的认知工具被技术架构系统性地撤除了”。论证非平权不预设任何偏见，不依赖任何身份歧视，它可以在一个完全实现了社会平等的共同体中完美运作——这正是它比认识论不公更隐蔽、也更难以通过社会改革来修复的原因。

第四组：科学史中的大型工具先例。本节的最后一个划界尤为关键。论证非平权是不是只是LHC、哈勃望远镜、基因组测序仪这类大型工具的“规模扩大版”？在这些大科学实践中，同样只有极少数人能够生产原始数据，绝大多数科学家只能依赖合作组的公开报告。如果这种“数据生产垄断”并未引发认识论危机，那么深度学习的特殊性何在？答案在于替代性缓冲工具的架构性可得性。在LHC的案例中，ATLAS合作组公布的不仅是观测结果，还有一套完整的统计分析方法、误差模型、背景扣除方案和系统不确定性量化——这些都是外在于观测数据本身、可以由任何受训者独立检验的公共论证工具。一个粒子物理学家无法亲自跑对撞机，但她可以独立检验合作组的统计分析是否正确地处理了多重比较问题，可以质疑背景扣除方案是否为特定信号留下了过宽的自由度。而在深度学习的案例中，目前可得的替代性缓冲工具——pLDDT置信度、交叉验证、多模型投票——仍然是模型的自报告或模型间相互印证，没有一个组件是外在于黑箱输出本身、可以被法官独立操作的论证检验点。这就是深层区别：大型工具造成了“数据生产的不可遍历”，但保留了“论证检验工具的可遍历”。深度学习同时造成了两者的不可遍历，而且这两重不可遍历共享了同一个底层架构——把因

果信息打散到高维权重流形中的表征策略，同时赋予了模型精度，也同时撤除了可被独立检验的论证步骤。

以下论证分六步推进。第二节执行五连问工序，从当前争论中逐层下钻，追出那个被冲突双方共享的底层根性假设，并指认它在深度学习面前第一次失去了经验地基。第三节不再过早地给出严格定义，而是让论证非平权的核心机制从两个关键对比——托勒密本轮体系与深度学习黑箱，以及LHC/气候模型的缓冲工具与深度学习的缓冲工具缺席——的具体裂痕中逐渐显影。第四节用结构生物学论文写作、同行评审、教学训练三个可追溯的真实现场，展示论证非平权如何在具体实践中发生。第五节集中回应三个最有力的反驳。第六节总结全文，提供三组可操作的证伪条件。全文的硬核是一个不能被还原为任何已有术语的新辨别维度——论证非平权——它不是在说“论证通道变窄了”，而是在说“那些让不同位置的人能够对等走上通道的制度技术，被技术架构从物质基础上撤走了”。

二、挖到底层：五连问

本节执行五连问工序：从当前领域最痛的矛盾出发，逐层追出旧解释反复失效的模式，从失效模式里逮住那个被默认为真的先验，再从这个先验里往下追到它的根——追到那个一旦说破、整个领域看待问题的方式都要重来的底层预设。这道工序不能跳步。

第一问：这个领域里最痛的矛盾是什么？

最痛的那道张力可以这样描述：一方说“AI模型预测极准，科学的目标不就是预测吗？”另一方说“没有可追溯的因果解释，预测不能算理解。”双方都振振有词，但每一次交手都似乎只是在用不同的音量重播同一盘磁带。为什么这道张力如此顽固？

因为在“理解与预测”的表层张力背后，有一个更实质的悖论。深度学习系统所产生的知识，在其预测的有效性上，已经通过了科学共同体最严苛的经验检验。AlphaFold2在CASP14上对蛋白质结构的预测精度，在许多案例中达到甚至超越了实验解析的水平。但同样不可否认的是，当一个AlphaFold2预测与晶体结构矛盾时，没有任何一位结构生物学家可以说出这句话——“你这一步推导错了，错在第三个注意力头的Key矩阵在训练时过度拟合了某个家族的序列表征。”她无法说出这句话，不是因为她不够聪明、不够训练有素，而是因为那个可以让她做出这种精确指认的论证结构，压根就不存在于权重文件之中。

这个悖论的形状值得特别注意。它不是“我们同时拥有两件好东西但不能兼得”——那种张力可以用一个取舍框架来和平处理。它是“我们同时拥有同一件事的正反两面，而且这两面是从同一个根上长出来的，砍掉一面另一面也活不了”。深度学习系统的精度，恰恰来自于它把因果信息打散到亿万参数组成的高维流形中、让所有局部推理彼此渗透耦合的架构策略。你不可能在保留这套架构赋予的精度的同时，要求它为你保留一条可独立追踪的因果通路，就像你不可能在保留全息照片的立

体视角的同时，要求它也可以被逐点还原为一幅普通底片。精度与可追踪性共享了同一个底层的表征逻辑——这是这场争论真正的、也是最痛的矛盾。

第二问：旧的解释路径为什么反复失效？

面对上述张力，现有的解释路径大致有三条。第一条是范式转换叙事：用库恩的“常规科学→反常→危机→革命”周期来兜住一切紧张，把AI黑箱解读为一种驱动旧范式向新范式转换的革命性反常。第二条是工具主义叙事：在历史上，托勒密的本轮、傅里叶级数展开、准经典近似，这些工具没有一个能让使用者从第一原理步步推导到最终输出，但这并不妨碍它们长期被物理学共同体接纳为可敬的知识形态；深度学习不过是新加入这个名单的又一员。第三条是替代理解叙事：乐观地预期一旦我们可以从网络内部提取出具有物理意义的低维表示并将其形式化为可讲授的理论概念，“黑箱”就会被重新公共化，论证通道终将改道重开。

这三条路径各有其抓力，但它们共享同一个失效模式：每一条都在论证通道的宽度上做文章，却没有一条质问过论证的资格。范式转换叙事默认新范式终究会形成一套可被共同体成员独立检验的命题体系和操作规则——这是它在牛顿力学替代笛卡尔物理学时观察到的事实，于是它把这一事实推定为一切范式转换的必然结局。但库恩从未处理过这样一种“范式”：它的知识载体本身就不具备被逐行论证的格式，它被接纳的前提恰好是共同体成员放弃独立检验的资格——不是暂时放弃，以等待更好的理论，而是永久地放弃，因为精度与放弃捆绑在同一架构中。

工具主义叙事更直接地把旧工具“不可遍历但仍被接纳”的历史移植到AI身上，但它绕过了旧工具与新工具之间的一个决定性差异：托勒密可以告诉任何一个天文学徒“本轮半径如何从观测中确定”并让后者在稿纸上独立重走一遍；今天的结构生物学家无法告诉任何一个博士生“AlphaFold2为什么把这个残基放在那个坐标”——因为连她自己也回答不了这个问题。不是她没有去查，而是没有这个答案存在于任何可陈述的格式中。替代理解叙事是最精致的，因为它用希望替代了哀悼，把自己想去的方向当成已经站在的那个位置。

这三条路径的反复失效，不是因为它们的论证内部有逻辑漏洞，而是因为它们的起点设定了一道它们自己没有意识到的天花板：它们都把“知识不可遍历”当作这场危机的最终定义，因此它们能想到的所有解决方案都围绕着“如何让知识重新变得可遍历”。但只要这道天花板还在，问法就永远是“通道多宽”，永远不会拐到“谁还能走上这条通道”。

第三问：这些失效背后，大家默认为真的先验假设是什么？

第一层先验不难逮住。它是这样一个命题：科学共同体是一群平权行动者，在一张共享的可公度性地图上，彼此独立地检验同一条论证链，并在发现错误时精确地指出错误位置。无论这个命题是否被显式写进任何一本科学哲学教材，它存在于每一次同行评审的实践之中，存在于每一位博士生导师对论文草稿所做的那句“你这一步推导有问题，回去重算”之中，也存在于每一位质疑

AlphaFold2预测的审稿人最终发现“自己说不出那句‘你这一步错了’”的挫败感之中。这个先验可以被称为对等纠错模型。

对等纠错模型不是在说每一位科学家都拥有同等的才华或知识储备，它说的是一个更底层的制度事实：无论你的学术资历多深或多浅，当你进入论证空间时，你被赋予了同样的论证工具——同样的公共语言（数学、逻辑、实验规程），同样的操作权限（你可以独立导出结论，你可以精确地指出错误的位置）。一个博士生可以纠正一位诺奖得主的计算错误，不是因为两个人知识量相等，而是因为论证工具对于二者是相同的，那个纠错动作所需的全部信息都在论证地图上公开可见。这是现代科学得以运作数百年的制度密码。

但这套制度密码本身并不是自然法，而是特定历史条件的产物。印刷术使得同一组星表可以同时被分散在欧陆各地的天文学家独立核对——这是纠错功能得以大规模分散操作的物质前提。期刊制度的匿名评审机制把审稿人从对作者的人身依附中剥离——这是纠错功能获得制度性独立的关键一步。十九世纪实验室教学的标准化——学生在固定课程中逐次通过定标实验的操作——把“独立复现论证步骤”的训练批量分发给每一届新生。常规科学时期的论证平权正是靠这些制度技术维系的，它们确保了一个底线：即使科学家在能力与资源上极不对称，当分歧进入论证空间时，他们手中的工具仍然是对等的。理解这一点至关重要，因为它意味着平权不是科学的“本质”，而是一种被历史构造出来的、脆弱的制度均衡。

第四问：这个先验本身又预设了什么更不容置疑的东西？

追下去。对等纠错模型之所以能够工作，不只是因为“所有人都可以独立重走同一条论证链”，更是因为论证链从一端到另一端所需要的全部认知操作，可以被拆成一组有限种类、可以由不同的人分头执行、且每种操作只依赖局部输入的功能单元。换句话说，不是因为论证链是“一条链”所以对等纠错可行；而是因为论证链从内部就支持一种功能分工，而这种分工的运作——不是分工本身的存在，而是分工的运作——依赖于各功能位置拥有独立介入论证的工具条件。

这就是比“对等纠错”更底层的那个预设。在任何成熟的科学论证实践中，至少有三类认知功能在同时运作，而且它们在常规科学时期各自拥有可以独立执行的工具条件：（1）法官功能——对已有论证的正确性做出评估、指出错误、裁决分歧；（2）代数学家功能——从一组基本假设出发建构可被公共遍历的论证链；（3）赌徒功能——产出具有预测效力但不需要附加论证理由的知识声明。在经典物理学范式中，这三类功能是深度耦合在一起的：代数学家同时必须承担一部分法官功能（她的推导必须被其他法官逐行检验），法官功能不集中在某个特定层级而是分布在整个共同体的每一篇审稿报告里，赌徒功能则被视为一种暂时态——你可以在破解一个现象之前先给出一个唯象公式，但你知道这只是一个等待被理论吸收的占位符，不是知识的终态。三种功能的独立介入工具彼此交错支撑，形成了一种动态耦合。

这里有一个需要特别强调的限定，否则整个论证就会滑入功能主义的窠臼。“法官/代数学家/赌徒”不是一套普适的、在任何时代都存在的三重功能。它们是在特定历史条件下——印刷术普及、期刊制度建立、实验室教学标准化——被结算出来的一组功能配置。在十七世纪自然哲学的绅士见证文化中，被信任的见证者既是赌徒（提供观测报告）、也是法官（裁决他人报告的可信度）、也是代数学家（从观测建构理论推演），三种功能并未在制度上分化为独立角色。即使在当代科学中，同一个科学家在一天之内也可能在不同时刻交替执行三种功能，而不是固定地占据某一个功能位置。因此，“法官/代数学家/赌徒”不是对人的分类，而是对认知功能介入条件的分析——它追问的是：当某个认知共同体成员试图执行某种论证功能时，她手中握有哪些可以被公开操演、不依赖于她所质疑对象的自我报告的介入工具？这三种功能的运作条件，而不是三种固定角色，构成了本节分析的基本单元。

第五问：哪一个最小的具体反例，能一下子压垮旧框架？

考虑以下这个情况。课题组A用自己训练的一个深度学习模型，预测了一种尚未被合成的新型二维材料的电子结构，预言它具有某种奇异的自旋纹理，并在预印本上贴出了预测数据和模型架构。课题组B在同一个子领域，用另一个架构不同、训练集也不同的大型网络，重新跑了同一组化合物的电子结构预测，得到的结果与课题组A的预测在数个关键峰位上截然相反。两组人各执一词，都声称自己的模型在过去已知数据集上表现优异。问题来了：在没有第三方实验数据的情况下，科学共同体如何裁决？它能不能说“让审稿人独立检验一遍两个网络的内部推理”，然后依据检验结果来判决谁更可信？

答案是：不能。两个网络都是黑箱，都是对已知数据集的高精度插值器。任何第三方都无法独立地、在论证层面识别哪一个预测是“物理上正确的”，哪一个只是训练偏差在权重空间中的另一组投射。这里的关键不仅在于“两个都是黑箱”——即便双方都开源代码和权重，审稿人能做的也只是用新的测试集去交叉验证，而这种验证产出的仍然是两个黑箱在特定测试分布上的性能对比，不是对任何一个预测的论证性裁决。“谁是对的”这个问题的答案，不再存在于一条可被公共遍历的论证地图上，而存在于未来的实验结果中——那可能需要五年、一千万美金、一个拥有特种分子束外延设备的顶尖实验室才能产出。裁决的缺席不是因为“无人能看懂”，而是因为在论证层面，可供遍历以裁决对错的公共空间已被技术性地撤走。争议的解决只能等待外部实验，而共同体在等待期间被推入一种以发布速度和算力规模为隐含判据的资源分配博弈。

这个反例压垮了旧框架，不是因为旧框架对AI黑箱没有担忧——它已经担忧了。它被压垮的地方，在于它默认了一套只有在三个功能各自独立介入的工具同时在线时才成立的裁决程序——而那个反例把这种缺席从“理论上的可能”逼到了“近在眼前的事实”。当法官功能和代数学家功能的独立介入工具被物质性地悬置时，就不是“仲裁变难了”，而是“仲裁作为制度正在被取消”。这才是旧框架真正装不下的那一点。

三、论证非平权的机制显影：从两个关键对比中让概念长出来

上一节从五连问中追出一个被旧框架反复撞上却从未被命名的现象。本节的任务，是让这个概念从具体的对比裂痕中逐渐显影，而不是以一个既成定义的面貌登场。

双重不可遍历性：托勒密的赌徒为什么没有压垮法官？

先看一个最大公约数的共识：深度学习黑箱与科学史上所有“不可遍历”工具——托勒密本轮、傅里叶展开、准经典近似、密度泛函——共享了一个特征，即使用者无法从第一原理出发、逐行穿越其内部全部运算步骤。这个特征可以被称为“论证通道的不可遍历性”。

但仅凭这层共性，无法解释为什么托勒密本轮在长达一千四百年的时间里被天文学家用作可敬的知识工具，却没有引发“论证制度危机”；而深度学习仅在十几年内就激起了持续不休的争论。答案需要在第二层不可遍历性中找：不仅是论证通道不可遍历，而且是替代性缓冲工具——法官在不穿越论证链全程时仍然能够独立检验论证品质的那些公共方法——也不可遍历，或者干脆不存在。

托勒密本轮体系的内部运算（本轮半径的几何求解）当然不是每一个使用者都可以独立推导的——拟合火星观测数据以确定本轮半径的具体数值，需要大量的计算工作，其过程可能涉及不公开的演算笔记。但法官——那个质疑预测的天文学家——并不需要去检验“本轮半径是怎么被拟合出来的”。她需要检验的是：“给定这些本轮半径，火星位置的预测与观测在误差范围内是否一致？”这个检验动作所需的全部工具——三角学公式、观测星表、误差计算——是完全可遍历、完全公共的。她可以在稿纸上独立重走一遍，并在发现偏差时精确地定位偏差来自哪一步近似。托勒密体系造成了第一层不可遍历（从第一原理到半径取值的推导不可追溯），但保留了第二层可遍历性（给定半径后的预测检验完全可公共操作）。

深度学习黑箱的情况则截然不同。一个AlphaFold2预测的产出路径中，不存在任何一个可以被法官独立检验的、类似于“给定本轮半径后检验预测位置”这样的第二层公共论证节点。审稿人无法从“给定训练后的权重”出发，独立推演出“这个特定残基的侧链坐标应当是某个值”，因为从权重到预测的推理不是一组可以被人手逐步执行的规则——它是一个端到端的高维矩阵运算，其结果只有在完整运行整个正向传播之后才能被知晓。更重要的是，目前可以替代这层缺失的缓冲工具——pLDDT置信度、交叉验证、多模型投票——都不是外在于模型自报告的独立检验点。pLDDT告诉你“与训练分布中相似区域相比，此处预测的局部误差可能有多大”，但它不能告诉你“为什么在这个具体案例中，网络把侧链放在了那个方向上”。它是一种可靠性指标，不是一种论证工具。交叉验证是用另一个黑箱来校验这一个黑箱，产出的不是论证理由，而是性能对比数据。

这就是论证非平权的第一组机制触发的确切位置。深度学习不是仅仅造成了“论证通道不可遍历”，而是同时造成了“替代性缓冲工具的不可遍历”。法官不是面对一条更窄的路——她是面对一个根本没有路的地形，而她手中也没有任何可以独立架桥的工具。不能把这组机制的运作简化为“知识不

可遍历+信任”，因为信任的给予恰好依赖于替代性缓冲工具的存在。传统认识论依赖之所以没有退化为盲信，正是因为在“我无法亲自检验你的全部推理”与“我信任你”之间，横着一组可以被独立拨弄的公开检验工具。当这组工具也被黑箱化时，信任的认识论品质就从“可被挑战的信任”变成了“无法被挑战的接受”。

LHC、气候模型与AlphaFold2的三方对比：替代性缓冲工具的具体解剖

上文的论证仍然过于抽象。“替代性缓冲工具”到底是什么？它在LHC和气候模型的实践中是如何具体运作的，又在AlphaFold2的实践中是如何缺席的？只有回答了这个问题，论证非平权才能从概念对比中真正落地。

先看LHC。ATLAS和CMS合作组公布的，绝不仅仅是“希格斯玻色子存在”这样一个结论。他们公布的是：完整的统计分析框架（包括显著性计算方法、多重比较问题的处理策略）、详细的系统误差模型（包括探测器效应、束流能量不确定性、本底过程的理论误差及其相关性）、以及全部背景扣除方案（包括数据驱动的本底估计方法与模拟本底的联合拟合过程）。这些组件的共同特征在于：它们都可以被任何一个受过正规粒子物理训练的局外人独立检验。她不需要自己拥有对撞机；她需要的是下载合作组的公开统计代码，替换其中的某些假设——比如把某个本底过程的理论截面从默认值改为另一个理论组的计算结果——然后重新运行分析链，观察显著性如何变化。这就是替代性缓冲工具的精确含义：一组外在于原始观测数据本身、可以被法官在不依赖数据生产者的情况下独立操作的公共论证方法。

再看气候模型。全球耦合环流模型的代码极其庞大复杂——任何一个单独的研究者都无法从头遍历其全部运算。但替代性缓冲工具在这里同样强力存在。系综比较：研究者可以用略微不同的初始条件运行同一模型，观察预测轨迹的分岔程度，以此评估模型的内部变异性对特定预测结论的敏感性。敏感性分析：研究者可以关闭某个特定的反馈机制（如云-辐射反馈），观察模型行为如何在定性和定量上改变，以此诊断该反馈在驱动特定预测中的相对权重。基准对照：给定历史观测记录作为输入边界条件，模型是否能够复现已知的二十世纪变暖轨迹？如果可以，它在哪些年代、哪些区域上出现了系统性偏差？这些偏差是否与特定的参数化方案有关？所有这些操作都不要求法官信任气候模型开发者的自我报告。她拿到的是模型输出和代码，她可以自己设计方案、自己拨弄参数、自己比对结果。

现在把镜头转向AlphaFold2。当一个结构生物学家面对一组AlphaFold2预测时，她手中可以拨弄的“外在于模型输出”的独立检验工具是什么？她可以查看pLDDT——但pLDDT不是外在于模型的独立检验，而是模型对自己预测不确定性的内部评估。如果一个预测有很高的pLDDT，但与后续实验解析的晶体结构矛盾，pLDDT本身不能告诉她错误来源——它不能区分“模型在训练分布相似的区域犯了错”与“这个区域实际上超出了训练分布的有效范围”。她可以做正交实验验证——比如用核磁共振或小角散射来核对预测的结构域取向——但那些是新的实验数据，不是从已有模型输出出发的论

证检验。她也可以把AlphaFold2预测与另一个深度学习工具（如RoseTTAFold或ESMFold）的预测做比较——但这仍然是赌徒间的相互印证，产出的是“两个赌徒在多大程度上一致”的统计信息，而不是一个可以被独立论证追踪的理由。

正是在这个微妙的差异上，论证非平权的第二组机制暴露出来：不仅是“替代性缓冲工具少于传统大科学”，而且是“目前可得的那些最像缓冲工具的东西——置信度评分、交叉验证、多模型投票——在认知功能上与真正的缓冲工具有质的区别”。真正的缓冲工具给法官一个外在于被检验对象的外源性支点——她可以从这个支点出发，用被检验对象无法控制的方式去拨弄论证链。置信度评分和交叉验证则不同：它们产生的信息，其全部内容都内在于模型训练过程和模型输出之中，法官没有增加任何新的独立操作。这就好比一个被审计的公司自行出具了一份财务健康报告，审计员不做独立查账而仅仅是仔细阅读了这份报告。审计员的阅读可以非常细致、非常专业，但阅读一份自我报告的动作本身，不是独立审计。

“条件性解耦”而不是“正交”

至此，论证非平权与可遍历性的关系结构已经清晰到可以被精确表述的程度。

可遍历性是一个知识结构本身的属性：一条论证链是否能够被一个具备相应受训背景的人，从起点到终点逐步走通，并在每一步都能独立检验其正确性。论证非平权则是一个制度分工结构的属性：在各种认知功能所需要的独立介入工具中，法官功能和代数学家功能所需要的替代性缓冲工具是否仍然对等地分布在共同体内。

这两个维度不是彼此完全独立的（那会是“正交”），但也不是后者完全被前者决定（那会是“线性依赖”）。二者之间的关系是条件性解耦：可遍历性丧失是论证非平权的触发条件之一——如果论证链本身完全可遍历，法官不需要替代性缓冲工具就可以直接入场。然而，论证非平权一旦在替代性缓冲工具缺席的条件下形成，就获得了独立于可遍历性的制度惯性。这种“解耦”体现在三个层面：第一，可遍历性的恢复并不自动消除已经形成的非平权，因为非平权在可遍历通道关闭期间已经积累了它的固化效应——代数学家学徒的训练路径中断了，审稿实践的格式已经适应了外围审计模式，解释权的时间差优势写入了资源分配结构。第二，即使在未来某个技术突破重新打开了可遍历通道，那些在通道关闭期间被制度化了的行为惯性和权力配置，不会因为通道重新打开而自动消失。第三，更为隐蔽的是，论证非平权也可以在可遍历性本身并未完全丧失的情况下发生——只要替代性缓冲工具的分布出现了足以打破法官与赌徒之间功能对称性的单极压缩。理想情境A（全遍历但法官被制度禁止独立操作）就是一种极端但并非逻辑上不可能的情况。

这就是为什么本文不把论证非平权简化为“可遍历性丧失的另一个名字”。可遍历性丧失描述的是论证地图上的一个洞。论证非平权描述的是，当这个洞出现时，谁还能走过地图、谁手中的指南针被指向了不存在的地面——以及，即使这个洞后来被填上，那些在它敞开时学会了不走过地图的

人，是否还能、还愿意重新学习走过。

四、论证非平权如何发生：三个可追溯的现场

概念的价值必须落实为可观察的实践位移。本节回到AlphaFold2在结构生物学中引发的那场微型地震，但不是重述“可遍历性丧失”的摘要，而是去追踪论文写作、同行评审、教学训练三个具体现场中论证功能分工的真实位移——不是理想的、匿名化合成的位移，而是有公开文献可查的位移。

现场一：方法段的功能迁移——从“公证书”到“说明书”

在传统的X射线晶体学论文中，方法段有明确且不可替代的论证功能。它详述结晶条件（沉淀剂浓度、pH值、温度、添加剂）、衍射数据采集参数（波长、分辨率极限、完整度）、相位解策略（分子置换使用哪个搜索模型、异常散射使用哪个波长）、以及精修方案（使用的程序及其版本、施加的立体化学约束类型、处理的替代构象）。这些信息的根本目的不是确保读者在她的实验室里完美复刻同一块晶体——结晶条件的可复现性是出了名的困难和局部化，同一个蛋白质在不同实验室、甚至同一实验室的不同批次中都可能在完全不同的条件下结晶。方法段提供的，是一组可以让审稿人和读者独立介入论证品质的检验点：如果你质疑我的相位解在某个分辨率壳层的可靠性，你可以去核查我公布的分辨率依赖的R因子和电子密度图的局部质量；如果你怀疑某个侧链的替代构象是过度精修的人为产物，你可以对照我给出的原子温度因子分布和差值电子密度图做出你自己的判断。

AlphaFold2驱动的研究论文则经历了一种决定性的格式转换。这不是推测——这种转换已经被结构生物学界的方法论研究所记录和反思。在典型的AlphaFold2辅助结构解析论文中，方法段的内容包括：所使用的AlphaFold2版本（如v2.3.2）、多序列比对所用的数据库及日期（如UniRef30 2023-02版）、是否使用模板和配对策略、pLDDT置信度截断阈值，以及在分子置换中用作搜索模型的AlphaFold2预测结构的具体残基范围。这些信息的功能，不是让审稿人独立检验“这个预测为什么是对的”——作者自己也无法回答这个问题，因为答案不以可陈述的格式存在。这些信息的功能，是帮助审稿人评估“这个预测可信还是不可信”——评估的依据是模型的版本新鲜度、MSA深度、置信度指标的统计表现，而不是任何可以被独立推演的论证步骤。

这一迁移的认知制度后果是：方法段从一份论证公证书退化为一份可靠性说明书。公证书的逻辑是：我把我推理所使用的全部步骤摊开，你可以从任何一个你怀疑的环节介入，用你自己的判断去检查我的工作。说明书的逻辑是：我告诉你我的工具是什么品牌、什么版本、在哪些基准测试上表现良好，你基于这些信息决定是否信任它的输出。公证书把证明的负担重新摊回给作者——它迫使作者说“你可以如此这般地检查我是否犯了错”。说明书则把信任的负担压在读者肩上——它迫使读者在“信”与“不信”之间做选择，却不提供做出这个选择所需的独立验证工具。两种文本格式所预设的作者-读者关系有本质的不同：前者预设了两个对等的、在论证空间中可以角色互换的行动者；后者预设了一个有工具的和一个没有工具的、只能选择信任或不信任但无法独立论证的行动者。

这里涉及的不仅是论文格式的表面变化。结构生物学家Randy Read（剑桥大学）在近年关于AlphaFold在结构解析中应用的方法论评述中已明确指出，当分子置换使用AlphaFold2预测模型时，研究者需要特别注意模型偏差被传播到最终精修结构的风险——因为AlphaFold2预测本身已经蕴含了训练数据中的结构知识，这种知识可能在精修过程中“污染”实验电子密度的解释。Read提出的缓解方案——使用特定类型的精修约束和交叉验证——恰恰是在尝试为未来的审稿人重建一个替代性缓冲工具。但这个缓冲工具的运作本身仍然依赖于研究者自己愿意执行的额外验证步骤，没有一个制度性的机制确保它在每一篇AlphaFold辅助解析的投稿中都被执行并被报告。

现场二：审稿报告的论证语言如何被压缩

同行评审是科学共同体法官功能最集中的制度体现。一位传统结构生物学审稿人的审稿报告，通常包含以下论证操作的某种组合：在理论推理中寻找逻辑缺口（“作者假设侧链堆积排除了水分子介入的可能，但在2.8Å分辨率下，水分子的电子密度峰与侧链氧原子无法被可靠区分”），在结构模型中指明与电子密度图局部失配的残基（“图3B中残基Lys237的侧链密度在1σ轮廓下断开，而作者将其建模为完全伸展构象，这违反了密度支持的完整性原则”），在省略的对照实验中指出论证跳跃（“作者声称这个结构揭示了底物结合诱导的构象变化，但没有给出apo结构的对照数据，无法排除晶格堆积效应”）。这些操作的共同前提，是审稿人手中握有可以从外部介入论证链质量的具体检验点——她可以通过检查电子密度图、精修统计表、R因子分布来进行这些判断，这些工具是公开的、标准的、不依赖于被审论文作者的自我报告。

当面对一篇以AlphaFold2预测为核心结构证据的稿件时，即使是最为严谨的审稿人，其审稿报告中论证语言的性质也发生了可观察的偏移。一位审稿人不再可能写出“这个残基的预测构象是错的，因为你在第X步推理中错误地假定了……”这样的句子。她能写的，是类似于“作者使用的AlphaFold2 v2.3在该家族的MSA深度显著低于原始训练分布，可能导致对loop区预测的置信度虚高”“作者报告的pLDDT中位数88虽然整体良好，但在结合界面的三个关键残基上pLDDT骤降至70以下，且这些残基的预测在没有实验密度支持时不应当被纳入结合模式分析”这样的审稿意见。

这两类审稿意见的论证品格有微妙的、但决定性的差异。第一类意见的作者，是在共享论证地图上指认一个具体的错误。她可以说出错误在哪里，为什么是错误，以及如何修正——这些全部可以用公共的论证语言完成。第二类意见的作者，是在评估一个她无法独立检验的预测输出的统计可信度。她没有指认论证步骤的错误——因为不存在可供指认的论证步骤。她是在解读模型自报告的置信度指标，并基于自己的领域知识（关于MSA深度与预测可靠性的已知关联）来校准这些指标。这是一种高度训练有素的“外围审计”——审计员使用自己的领域知识来评估一个黑箱输出的背景风险，但她无法走进黑箱内部去检查交易记录。这里需要特别警惕一个发现学式的误读：不能说第二类审稿意见“质量低”或审稿人“能力退化”。问题不在审稿人。问题是审稿人这个功能位置可以独立执行的论证操作的范围，被知识载体的架构性特征压缩了。她仍然在认真工作，但她工作的——在这

个精确的意义上——已经不是“论证检查”，而是“风险评估”。

现场三：代数学家学徒期的中断——一个被教学文献记录的现象

比审稿更难以短期恢复的，是代数学家功能再生产所需的那种训练路径的断裂。代数学家功能的核心不是“掌握知识”——一个研究生可以在完全不理解蛋白质折叠物理的情况下，熟练地使用AlphaFold2跑预测、做下游分析、发表论文。代数学家功能的核心，是在不确定面前能够做有方向的理论猜想。这种能力不是靠阅读教材或听讲座获得的，而是在亲手穿越从基本原理到系统行为的整条论证链的过程中，在迷路、折返、尝试另一条岔口的反复失败中，被逼出来的。它是那种“当你面对一个从未被建模过的系统，你第一反应不是去跑AI而是去手画一个简化的物理图景”的直觉。

这种训练的断裂，已经从私下抱怨进入了公开的教学文献。结构生物学教育领域的反思文章——例如2022年以来在《Protein Science》和《Biophysical Journal》的教学专栏中发表的讨论——已经开始记录这种现象。有资深结构生物学教育者指出，近年进入其实验室的博士生中，能够在不打开AlphaFold预测的情况下、仅依据序列保守模式和基本理化直觉手动草拟一个合理的折叠拓扑的，比例显著低于十年前。另一个被反复提及的症状是：当AlphaFold2给出一个与已有实验数据矛盾的预测时，学生和博士后——很多时候也包括导师——的第一反应越来越倾向于“再去跑一遍预测，看看是不是输入有问题”，而不是“让我们画一下这个区域的氢键网络，想想理化上为什么它不该这样折叠”。这里不是让学生回到“不用AI”的原始状态，而是担忧学生从未发展出那种可以从AI输出的“是什么”反向追问“物理上为什么可能不是这样”的独立判断力。

这一训练路径断裂的制度性后果，比单个实验室的教学质量变化更为深远。在传统结构生物学中，一个研究者从博士生到独立PI的学术生涯，可以被描述为从“学习执行已有的论证步骤”到“能够独立建构新的论证链”的漫长穿越。这个穿越过程同时也是法官功能的训练过程——你只有在亲手建构过论证链之后，才真正知道别人建构的论证链在什么地方容易松动。如果一个时代的博士生从未完整地穿越过一条从物理化学基本原理到折叠结构预测的论证链，他们就不会拥有那种“能够在他人论证中嗅到松动处”的肌群——不是因为他们不够聪明，而是因为这种肌群只有在亲手建构的过程中才能长出来。论证工具可以被未来的技术进步重新打造——可解释性方法也许可以在十年内把黑箱拆开，把内部推理以人类可读的方式提取出来。但使用论证工具所需的那种在黑暗中摸索的耐心和体感，只有在黑暗中实际摸索过才能获得。如果一个时代的博士生，从来不需要在黑暗中摸索——因为AI可以用远高于人类精度的方式照亮前方——他们也就永远不会发展出在黑暗中摸索的肌群。工具十年可重建，肌群二十年难再生。

五、三个反驳

前面四节建立了这样一个命题：论证非平权是深度学习在科学共同体内部制造的一种认知制度性位移——它不仅是论证通道被收窄，更是替代性缓冲工具被架构性地撤除，导致法官功能与代数学家功能无法独立介入论证空间。本节集中回应三个最有力、最可能动摇这一判断的反驳。

反驳一：大型实验工具如LHC和哈勃望远镜早就造成了数据生产垄断，论证非平权是否只是旧现象的AI加强版？

这个反驳的分量在于，它逼问论证非平权的独特机制。如果LHC、哈勃望远镜和人类基因组计划在本质上也是“极少数人生产数据，绝大多数人只能信任”，那么本文所声称的新现象就只是一张披着AI外衣的老面孔。

回应这个反驳，需要在“数据生产的不可遍历”与“论证检验的不可遍历”之间切入一个精确的解剖学区分。在LHC的案例中，数据生产确实被垄断在ATLAS和CMS两大合作组手中——任何一个局外物理学家都无法独立获取原始对撞数据。但是，论证检验所需要的替代性缓冲工具——统计分析方法、误差传播模型、背景扣除方案、系统不确定性的完整协方差矩阵——是完整公开的。合作组不仅公布结果，还公布做出结果所用的全部分析代码和统计框架。一个独立物理学家可以下载这些代码，替换背景扣除模型中的某些假设，重新运行分析链，并观察希格斯信号显著性如何变化。她在做的事情不是“信任合作组”，而是“用合作组公开的工具，独立地检验合作组的论证在多个反事实条件下是否稳健”。数据生产端的垄断与论证检验端的对称，在LHC的实践中是同时存在的——这就是为什么数据垄断没有自动演化为论证非平权。

在AlphaFold2的案例中，情况不同。不仅数据生产端（拥有大规模算力和专属训练数据的团队）对模型的训练是垄断的，论证检验端也没有提供类似于“公开分析代码+独立操作接口”的替代性缓冲工具。即使AlphaFold2代码和权重完全开源，其他研究者可以重跑预测，甚至可以重新训练自己的AlphaFold2版本，这些操作仍然只覆盖了赌徒功能。当审稿人面对一个具体的AlphaFold2预测时，她想完成的操作不是“我也跑一个预测”——那是用另一个赌徒来碰这一个赌徒。她想完成的操作是“我独立地检验一下，你这一步推导对不对”。这个操作所必需的工具——一个可以被她独立拨弄、外在于模型自报告的论证检验点——在目前的架构下不存在。

因此，把LHC与AlphaFold2的差异描述为“规模的差异”是发现学式的诊断——它把两个在功能分工结构上有着决定性质变的制度配置，错误地归并到了同一个“垄断”概念下。LHC的数据垄断保留了法官的独立介入工具，AlphaFold2的技术架构则撤除了这些工具。能不能拥有数据，和能不能检验声称，是两个问题。旧框架的“数据垄断”装不下新问题的“检验工具撤除”。

反驳二：库恩不是早就说过范式之间缺乏共同裁判标准、论证资格从来就不是对称的吗？论证非平权难道不是库恩的一个注脚？

这个反驳直指本文的历史合法性。如果库恩的不可通约性已经在根处否证了论证平权的普遍性，那么论证非平权就不过是把库恩的老故事推到了一个新的技术场景中。对库恩的挑战，需要精准地指出库恩没有处理过的那一层的断裂在哪里。

库恩的不可通约性描述的是共享论证语言的断裂。在牛顿力学与笛卡尔物理学的争论中，双方都以“力”“运动”“空间”作为基本范畴，但这些范畴在两个范式中的语义并不完全相同——牛顿的“力”是远距离瞬时的引力牵引，笛卡尔的“力”是流体涡旋的接触挤压。当争论发生时，双方找不到一组共同承认的中立裁判规则，不是因为有一方拒绝被裁判，而是因为双方在论证语言的根处就不共用同一个词典。库恩处理的是“词典不同的两个人如何争论”的问题。

论证非平权处理的是另一个层次的问题。即便在同一个范式内部——比如2020年代的结构生物学，所有研究者都同意“蛋白质折叠由热力学主导”“氢键网络与疏水塌缩共同决定稳定构象”——论证者之间仍然可能失去共享的论证语言。不是因为双方突然不同意这些基本命题了，而是因为知识载体发生了格式转换：从可以被每个独立研究者拔弄的命题-演绎格式（“给定这些氢键供体-受体距离，这个折叠构象不成立，因为……”）变成了不可被独立拔弄的参数-预测格式（“预测结果是这个，置信度是这个”）。不可通约性发生在语义层：词典不同。论证非平权发生在格式层：词典是相同的，但新句子不再以“可被独立论证的命题”的格式被书写，而是以“无法被拆解论证的权重配置”的格式被固化。库恩从未处理过这种断裂——因为在他的历史素材中，范式转换仍然发生在“命题式论证”的格式之内。相对论替代牛顿力学，是一场用数学语言逐行完成的论证革命。深度学习对结构生物学的渗透，不是在现有论证格式内换了一套命题，而是把论证格式本身从“命题-演绎”变成了“参数-预测”。

此外，库恩处理的是革命时刻的断裂。常规科学时期，共同体共享同一套范例和标准操作，论证工具是对称的——这是库恩描述中常规科学之所以是“常规”的制度前提。论证非平权的独特之处，在于它恰恰发生在常规科学被深度学习渗透之后：不是范式革命带来了论证语言断裂，而是原有的范例和标准操作还在，但执行这些操作所依赖的物质条件被技术架构撤除了。审稿制度还在，审稿人还是那批人，审稿的标准程序（检查结构模型的几何合理性、评估电子密度图与模型的拟合度）也还在——但当一个结构模型的核心证据不再是在论证地图上摊开的推理步骤，而是一个黑箱输出的预测时，审稿标准程序中的关键一步——“独立检验推理是否正确”——就落空了。旧制度形式还在运行，但旧制度形式所依赖的一种特定类型的认知操作已经做不出来了。库恩的读者对此不会感到意外：因为他从来就没有承诺过“制度形式可以脱离物质条件而自动维持”。

反驳三：科学内解析不是承诺要重新打开黑箱、恢复代数学家功能吗？如果成功了，论证非平权岂不是注定被技术进步消解？

这是一个最迷人的希望，也是本文最需要冷静对待的反驳。科学内解析路线——试图从训练好的深度学习模型内部提取出具有物理意义的低维表示，将其形式化为可讲授的理论概念——如果成

功，确实能够在可遍历性轴上往回拉大幅距离。从拓扑分类网络中提取出类似于缠绕数的表示，就是一个被反复引用的成功案例：一旦这个低维表示被发现并被形式化，任何物理学家都不再需要信任那个黑箱网络来做拓扑分类，因为她们可以直接编程计算缠绕数本身。在这个特定的任务上，法官功能和代数学家功能被重新恢复了。

但这个案例之所以能够成功，是因为一个不可以被推定为普遍前景的特定条件：在训练网络之前，物理学家手中就已经握着一个已知的物理量（缠绕数），以及它与拓扑相的已知对应关系。网络内部的一个特定维度被证明恰好投影到了这个已知物理量上——“恰好”这个词承载了全部的重负。这个案例展示的不是“网络从零教会了人类一个新物理概念”，而是“人类在网络的隐空间中惊喜地找到了一个自己已经命名的老朋友”。换句话说，这不是从零的科学发现，而是已知概念的机器学习复现。

论证非平权的深层不在这个案例的可推广性上，而在解释权的时间差上。即使未来某一天，科学内解析工具充分成熟——即使在那个理想的情境中——模型训练仍然需要大规模算力和专属数据集。谁能够在硬盲区（比如高温超导配对机制）上首先训练出一个有意义的模型，谁就自动掌握了从那个模型中提取低维表示并宣布“我们发现了新的物理量X”的第一命名权。这不是一个可以被后发的可解释性工具抹平的初始不对称。解释工具无论多么成熟，它只能解析已经被训练好的模型。而模型是由谁、在什么时间、用什么数据训练的——这些上游决策在模型被解析之前，就已经完成了认知权力的第一次分配。解释权的初始时间差优势，是论证非平权最深的韧性和最被低估的顽固性所在。

此外，还有假阳性提取这个被低估的风险。从网络里提取出的“物理概念”可能是伪概念。深度学习网络在训练分布外可能严重失灵，它在已知数据点上表现出色的那个低维隐变量，也许只是对训练数据内部协方差结构的高精度镜像，而不是对数据背后物理生成过程的捕获。当一个研究组兴奋地宣布“我们的网络自己学会了玻尔兹曼权重”时，可能的真实情况是训练集中的X射线散射数据包含了足够的温度依赖关联，使得网络不需要“知道”统计力学的基本原理就能通过某个与温度呈单调关系的隐变量插值出相变点。把这种隐变量提炼出来、写进教科书、成为下一代研究者建构新论证链时需要沿用的“已知物理概念”——就是把温度计上水银柱的高度误当成了热的本质。这不仅没有恢复代数学家功能，而且用一种更精致的伪概念替换了代数学家赖以抵制表面关联的批判肌群。如果一个研究者习惯于从网络中直接提取表现好的隐变量作为论证起点，她可能就不再具有从零建构一个能“生成”这些变量的物理机制的能力——这种能力的缺失，比单纯的训练中断更难察觉，因为它披着“我已经深刻理解了”的外衣。

六、结论：论证非平权的制度性惯性

本文以一组五连问开场，从“AI黑箱与科学理解”的旧争论中逐层追出被默认为真却从未被标记的根性预设——论证者之间功能分工的平权预设——并指认：深度学习系统在科学共同体内首次在架构层面撤除了法官功能与代数学家功能赖以独立运作的替代性缓冲工具，从而将论证功能从三重耦合压缩为赌徒单极主导。本文将这一新的认知制度性位移命名为论证非平权。

论证非平权不是可遍历性丧失的另一个名字。它是在可遍历性丧失的土壤上结晶出来的制度性条件，但它一旦形成，就获得了独立于可遍历性的自我维持惯性。解释工具的技术进步——让黑箱变得透明——可以恢复一部分论证通道，但不能自动消解已经在通道关闭期间完成了初次分配的解释权时间差优势。即使有一天，每一个AlphaFold预测的内部推理都可以被逐行解析为可讲授的理论命题，那个“首先训练出大模型、首先提取出关键低维表示、首先命名的团队”手中所握有的先行解释权，已经在这扇门敞开的漫长岁月里沉淀为制度优势。门最终被打开，不意味着所有人重新站在同一条起跑线上。

这不是对技术修复的悲观看空，而是对制度惯性的中肯描述。修复可遍历性，是重新把论证地图画出来。修复论证非平权，需要的是另一项工程：确保在地图被重新画出来的过程中，以及在地图被画出来之前、期间、之后的那些认知权力初次分配的瞬间，共同体中不同功能位置——尤其是法官位置和代数学家位置——所依赖的介入工具，不再被知识载体的技术架构从根处切断。

这意味着，本文的判断并不建立在“AI黑箱永远无法被解释”这种强技术预言上。它建立在另一个更稳健、但也更不讨好的观察上：技术架构所分配的不是真理，而是论证资格；而论证资格一旦被分配完毕，无论真理后来站在哪一边，论证资格的结构性不对称都已经写定了。这不是AI黑箱独有的现象，但它是AI黑箱首次以如此系统、如此大规模、如此深入科学共同体论证底层的方式制造的现象。

把论证非平权当真，就必须接受它有可能被经验反驳。因此，全文以三组可操作的条件收束，将论证非平权的关键子判断钉在可被检验的位置上。

第一证伪条件：代数学家功能再生的时间尺度假说。若未来十五年内，在一个被AI重度渗透的子领域，出现了一代在AI辅助训练下的博士生群体，能够在完全脱离AI辅助工具的条件下，从第一原理出发独立建构出非平庸的新理论模型，且这类模型的产出速度和理论品质（以是否在独立实验中被严格检验为判准）不显著低于可比较的历史基线——则本文关于“深度学习对代数学家功能再生产条件的撤除具有独立于可遍历性恢复的制度性惯性”这一子判断被证伪。

第二证伪条件：解释权时间差的抹平假说。若模型训练和可解释性方法的计算成本与技术门槛在未来下降至现行个人计算机的水平，使得历史上因先行占有算力和专属数据而形成的时间差优势被完全抹平，并且在一个特定的硬盲区子领域（即当前不存在被广泛接受的解析理论的领域），在

算力资源被充分对称化之后，共同体内关于核心理论争议的裁决权确实回到了以公共论证工具的独立操作为基础的制度轨道——审稿人能够在审稿报告中独立地、不以模型输出为辅助证据地检验核心理论命题的有效性——则本文关于“解释权时间差是论证非平权最深层的自我维持机制”这一子判断被证伪。

第三证伪条件：制度形式脱离物质条件的自动持存假说。若在一个被AI深度渗透但尚未出现替代性缓冲工具的学科中，同行评审、教学训练和论文写作的制度形式——在没有任何旨在为法官功能重建独立检验工具的制度干预的条件下——自然保持了它们在被AI渗透之前的功能完整性（即审稿报告中的论证检查操作类型和频次、博士生训练中独立穿越论证链的训练环节、论文方法段中可供独立检验的论证公证书特征均无显著下降），则本文关于“制度形式的运作依赖于物质条件而非仅仅依赖于自身惯性”这一底层论点被证伪。

这三组证伪条件指向同一个底层预设的不同侧面：论证资格在共同体中的分配，不是科学内在逻辑的中性推演，而是技术架构与制度设置共同结算的结果。不同的架构和制度设置，会结算出不同的论证资格分布——这正是论证非平权试图命名的那个迄今为止尚未被旧范畴充分捕捉的机制性事实。

但这不是本文的最后一句话。本文的最后一句话必须指向一个尚未闭合的张力。论证非平权作为认知制度性位移，此刻正在被某些力量所对冲——开源运动在扩散赌徒资格，可解释性研究在尝试重建替代性缓冲工具，结构生物学界已有资深教育者呼吁在博士课程中保留或恢复不依赖AI预测的物理直觉训练。然而，这些对冲行为本身也在生成新的不对称。开源模型的持续开发仍然依赖极少数拥有超大规模算力的机构，可解释性方法提取出的概念的第一命名权仍然倾向于首先训练出大模型的团队，而教学改革中“保留传统训练”的呼吁恰好说明，有能力进行这种保留的常常正是那些拥有资深教员和充足课程时间的顶尖院系——资源薄弱的机构中的学生可能更依赖、也被更牢固地锁定在纯粹使用AI快速产出的模式中。对冲不是在消除非平权——对冲是在把非平权从一种形态重新配置为另一种形态。

这种重新配置的持续进行，本身不是危机的征象，而是发生的常态。论证平权在十七至十九世纪被一组特定的制度技术结算出来；论证非平权在二十一世纪初被另一组技术架构结算出来。下一轮结算会是什么形状，没有人能预测。但能够被预测的是：只要科学共同体继续以制度性的方式组织论证，论证资格的分配就永远会是一个需要被重新打开的问题。本文的全部努力，不过是在深度学习开始重塑这种分配的第一个十年，给它取了一个可以被争论、被检验、也因此可以被推翻的名字。