界面的退隐：AI时代物理学认知格式的张力与结算

人工智能系统在物理学前沿正大规模地产出一种不携带公共批评格式的认知成果。本文论证：AI对物理学的根本冲击不在于它是否“理解”自然，而在于它首次在集体操作层面抽走了使知识共同体能够在概念层面比较、质疑和裁决彼此判断的认知格式——本文称之为认知界面。论文以量子多体物理中神经网络模型的认知地位争议为核心案例，追溯了界面从牛顿力学公理格式到量子力学希尔伯特空间格式、再到AI分布式表征的三次格式结算，揭示了当前张力的结构性特征：分布式表征的性能优势与其无法生成无损界面的属性，是同一枚硬币的两面。

作者 阳涌 · 约 2.1 万字 · 发表于2026年7月24日

摘要

人工智能系统在物理学前沿正大规模地产出一种不携带公共批评格式的认知成果。本文论证：AI对物理学的根本冲击不在于它是否“理解”自然，而在于它首次在集体操作层面抽走了使知识共同体能够在概念层面比较、质疑和裁决彼此判断的认知格式——本文称之为认知界面。论文以量子多体物理中神经网络模型的认知地位争议为核心案例，追溯了界面从牛顿力学公理格式到量子力学希尔伯特空间格式、再到AI分布式表征的三次格式结算，揭示了当前张力的结构性特征：分布式表征的性能优势与其无法生成无损界面的属性，是同一枚硬币的两面。在与认识论不透明性（Humphreys）、不可通约性（Kuhn）和数据界面（Leonelli）等概念的系统对话中，本文论证“界面”并非既有概念的重新包装，而是从分布式表征与命题格式的具体冲突中逼出的一个新辨别维度——它关注的不再是“个体是否理解”，而是“公共批评的格式通道是否被堵死”。本文正面回应了工具主义、AI可解释性、实用主义和科学社会学建构论的最强反驳，并通过湍流模拟与高能唯象学的真实案例展示界面被绕过的具体机制。结论落在一组尚未完成的张力上：旧界面正在被抽走，新公共格式的可能性条件尚在生成之中，物理学共同体正站在一轮认知格式结算的裂口上。本文给出严格的证伪条件：如果一种建构出分布式表征的系统，同时产出了一套无损的、可被公共批评操作的界面对接方案，本文的核心判断将被宣告失效。

关键词：物理学哲学；人工智能；认知界面；科学批评；分布式表征；认识论不透明性

一、从一次真实的失败出发

本节的任务不是给出“界面”的定义，更不是铺设概念坐标，而是让读者先看到一件事的完整发生：在某一个真实的物理问题面前，两位有能力的研究者确实无法进行那种在传统物理学中再寻常不过的概念批评。这件事首先需要被看见，然后才谈得上被命名。

1.1 一个被堵死的对话

2017年，Carleo和Troyer在《科学》杂志上提出了“神经量子态”方法——将多体量子系统的波函数表示为人工神经网络，通过变分优化逼近基态能量（Carleo & Troyer, 2017）。这一工作引发了量子多体物理领域的一场持续至今的方法论震动。此后数年，生成对抗网络、图神经网络、Transformer架构等深度学习模型相继被引入，在Hubbard模型、t-J模型、阻挫磁体等非平庸系统上取得了传统数值方法难以企及的变分精度。

然而，这场技术繁荣的背面，是一个具体的认知困境正在浮出水面。我们可以从一个真实的方法论争论切入。在二维Hubbard模型中度掺杂区域的基态性质问题上，神经量子态的不同实现给出了相互矛盾的低能激发谱预言：一些模型在热力学极限下倾向于d波超导基态，另一些则在相同的参数区间内预测自旋密度波序占据优势。在传统计算物理中，这种分歧的后续处理是有章可循的：研究者会坐下来，逐项比对各自的近似方案——有限尺寸标度的外推方式、截断误差的控制策略、试探波函数的对称性假设——然后从这些比对中逼出一个新的、能区分两种方案的关键测试。整件事依赖一个再朴素不过的前提：双方的认知成果，被装在一种可追踪推导的公共格式里。

当分歧的双方都是神经网络模型时，这个前提不存在了。研究者当然可以比较泛化误差、交叉验证分数、损失函数收敛曲线，也可以对上百万个权重做降维可视化。但这些操作给不出一个中间地带，让双方能够在概念层面定位分歧的来源。他们不能问“你的模型在哪一步用了什么近似”——因为模型的每一步都是高维空间中的一次非线性变换，它不包含任何可以被单独拎出来检验的“近似”。他们也不能问“你的波函数在长程极限下满足什么对称性”——因为对称性并不是模型显式陈述的条件，而是被隐式地分布在训练后的权重模式中，不同的可解释性方法会从中提取出不同的、甚至相互矛盾的对称性判断。

结果是一个在概念论证史上从未大规模出现过的困境：两位有能力的研究者，面对同一个哈密顿量，使用同一套训练数据，得到了预测精度同样高的模型，却无法在概念层面说清他们何以分歧。他们不是在等待对方给出更好的论证——他们是找不到论证可以抓住的把手。

1.2 这困境到底新在哪儿

一个合理的质疑是：物理学家过去也常常“不懂”自己的工具，这难道不是常态吗？

并非如此。以密度矩阵重整化群（DMRG）为例。大多数使用DMRG的计算物理学家并不亲手推导密度矩阵的谱分解，也不逐项检查截断算符的矩阵元。但他们知道，DMRG的理论基础来自量子信息论中的纠缠熵面积律；他们可以指出矩阵乘积态表象的物理动机是“在实空间中对低纠缠态的紧致表示”；当计算结果出现偏离时，他们可以根据最大截断维度、截断误差和纠缠熵增长速率来诊断问题是出在键维度不足还是数值噪声。这些操作不需要通晓DMRG的全部内部结构——它们只需要一个概念化的、可公共传递的简化模型。

问题出在神经网络模型上。一个深度残差网络对多体波函数的变分表示，它在训练动态中形成的内部表征结构，与DMRG的矩阵乘积态之间，有一道迄今无法桥接的鸿沟。DMRG的简化模型之所以能当简化模型，是因为它的架构设计本身就嵌入了明确的物理约束：矩阵乘积态的矩阵维度直接对应纠缠熵截断；一维性的成功与二维性的困难，可以从纠缠熵面积律在实空间维数下的行为得到解释。而神经量子态的架构设计——无论是受限玻尔兹曼机、卷积网络还是Transformer——它们所施加的约束（如平移不变性、局域连接性）并不天然携带此类直截了当的物理解释。一个训练完成的Transformer变分态，对不同量子相的敏感性可能以一种极不直观的方式分布在自注意力头的权重里——这种现象确实被观察到了，但它目前尚不可被无损地翻译为“这个模型在物理上假设了什么”的陈述。

这就逼出了一个实质性的断裂：神经网络模型的可批评性，与它的预测精度，不是彼此独立的两个属性，而是被同一个深层架构——分布式表征——决定了此消彼长的东西。精度越高，内部表征越复杂，可追踪推导就越困难。这个断裂，是传统科学哲学工具箱没有准备好处理的。它既不是“模型不可理解”（那是个体认知层面的老问题），也不是“不同范式之间缺乏共同语言”（那是范式断裂层面的老话题），而是在一个范式内部——在同一批基础物理定律、同一套训练数据、同一个实验检验标准之下——两种内核一致的认知产物，却无法在概念层面进行批评性对话。这个困境是新的。它需要一个新的辨别工具。

这个工具，本文称之为认知界面。

本节只做了一件事：把一个真实的困境摊开。下一节将追溯这个困境的历史前身——不是为它建立坐标谱系，而是为了看清：那些今天被我们怀念的界面格式，它们自己也是从更早的张力中被挤出来的，它们从来不是“天然就在那里”的东西。

二、界面的三次结算：从公理推导到分布式表征

上一节从Hubbard模型的现场逼出了一个判断：AI的分布式表征，在大规模产出预测精度的同时，抽走了物理学家赖以进行概念批评的公共格式。本节的任务是把这个判断放到一个更深的时间尺度上去校准。那个被抽走的东西——“界面”——是怎样一步一步长成后来那个样子的？

这个追问不是为了怀旧，也不是为了画一张从完整到受损的退化图。它的发生学动因很具体：如果我们看不清界面是如何从历史上两次重大转型中被逼出来的，就会把AI 时代的困境误判为一次偶然的技术故障，从而低估它的结构性。必须看清，界面从来不是自然给予的；每一代物理学家都不得不在新的认知压力下，重新结算出他们自己的界面格式。

2.1 牛顿格式：界面的第一次结算

在十七世纪后半叶以前，物理知识的生产并不存在一个公认的公共批评格式。自然哲学文本的论证方式高度依赖于作者的个人权威、修辞技艺和形而上学承诺。笛卡尔和莱布尼兹可以就“运动的量度”争论数十年不休，不是因为双方缺乏实验数据，而是因为双方争论所依赖的概念（“力”“运动量”“实体”）本身没有一套可被第三方独立操作的公共格式。

牛顿力学的真正革命性，并不优先在于解释了什么新现象，而在于它第一次生产出一种可被匿名化操作的批评格式：通过公理化推理从少数原理出发导出可观测量。《原理》前两卷的结构——定义、公理或运动定律、命题与定理的演绎——不是一种写作风格，而是一种认知公共设施的发明。这个设施的核心要件有三项：概念边界通过定义锁定（质量、动量、力的操作性定义）；逻辑步骤以命题序列展开，每一步都可在纸面上追踪；导出结论以几何形式呈现，与特定实验的量化结果直接对应。

这三项要件合在一起，构成了物理学史上第一个可公共操作的认知界面。一个在巴黎的物理学家，不需要认识在伦敦的同行的任何私人属性，不需要共享他的形而上学倾向，甚至不需要重复他的实验，就可以从纸面上推导一遍他的论证，定位出“第三步的近似是问题所在”，并据此设计一个能区分两种假设的新实验。这种操作在今天的物理学界被认为是理所当然的，但在十七世纪，它是被牛顿力学这套特定的认知格式发生出来的，它依赖于一个在纸面上可传递的推导链。

这个界面的功能边界也需要被诚实划定。它的代价从一开始就内置其中：牛顿力学要求所有需要被批评的概念都必须被翻译成公理化语言，否则它们就不在界面上。那些未曾被这样翻译的东西——比如“力”作为实体因的直觉、绝对时空的形而上学背景——很快就在批评实践中被边缘化了，不是因为它们被驳倒了，而是因为它们无法在界面格式上被操作。界面的第一次结算，同时是某些问题的可见化和另一些问题的不可见化。

2.2 量子力学的格式结算：界面的变形而非退隐

二十世纪初的量子力学面临着一种与AI 时代表面相似、其实质结构完全不同的认知压力：旧有的公理化推导格式无法容纳新的物理内容。波函数既是粒子又是波，测不准原理禁止同时以任意精度标定位置和动量，态矢量在测量行为中获得确定值的那个瞬间似乎跳过了一连串中间推导——每一个新现象都在撕扯牛顿—麦克斯韦格式赖以运作的因果连续性预设。

1925到1927年间海森堡、玻恩和约当发展矩阵力学，紧接着薛定谔给出等价的波动力学；1927年狄拉克完成了变换理论的统一表述，将态矢量置于希尔伯特空间中，把物理可观测量定义为作用在该空间上的厄米算符；1932年冯·诺依曼在《量子力学的数学基础》中将这一结构严格公理化。这个过程持续了大约七年，它的最终产物不是退隐，而是一套新的界面格式：以希尔伯特空间的线性代数结构为公共语言，以算符的对易关系为推论引擎，以本征值问题为从原理到观测量的稳定通道。

这套格式提供了牛顿格式所不能提供的东西，同时也丧失了牛顿格式曾经拥有的某些便利——直观的可视化几何推演被放弃，代之以抽象算符运算。但这并不意味着“界面退隐了”。正相反，一个物理学家可以拿过一篇陌生同行的量子力学论文，从哈密顿量的形式出发，沿着算符对易关系追踪能谱的推导，定位出对方在哪里用了微扰展开、在哪里做了截断，并据此设计一个能区分两种近似方案的新测量。批评的格式通道非但没有被堵死，反而更加精确了：界面的信息带宽提高了。代价是，进入这个界面的门槛也提高了——不是任何受过普通教育的人都能操作的，它需要一段特定长度的训练。

这个历史案例直指一个关键区分：界面的“变形”（格式的更替，同时保持批评的可操作性）与“退隐”（批评操作本身在格式内找不到把手），是两种完全不同的动力学。量子力学的格式结算属于前者——旧格式被瓦解，但新格式迅速从瓦解的张力中长了出来。AI的分布式表征撞上的，是一种更深的麻烦：它瓦解了旧格式，却未必在执行同样的重建。不是时间不够——量子力学从混乱到定型只用了七年——而是分布式表征的结构本身，不承诺将知识组织成可追踪推导的形态。这正是下一节要展开的核心。

现在可以看清“界面的三次结算”全貌：牛顿格式是第一次结算——从零散的形而上学争论中逼出一套可匿名化操作的公理推导界面。量子格式是第二次结算——在因果连续性的旧格式被撕裂时，重新在希尔伯特空间上建立起一套更抽象但同样可操作的批评界面。AI的分布式表征碰上的，是第三次：前两次的格式都是命题式的——无论多抽象、多不直观，知识最终都被装入了可追踪推导的命题容器。而现在，这个容器本身的有效性，正在被一种不依赖命题容器的认知格式大规模地挑战。

三、谁跟界面最像——以及本文要说它们没说的那个东西

到这里为止，我一直没有给“界面”下一个定义。这不是疏忽——在一个概念的发生论证没有走完之前，过早定义会被旧概念的语言惯性强按进旧坐标。前面两节做的事情，是让读者和我一起走到了这一步：有一个东西，在历史上曾经以不同格式存在过，正在当前的特定张力中被抽走；它既不是个体的理解感，也不是理论的内部属性，而是在公共批评中可以操作传递的格式通道。现在这个名字已经被局部的用法锚定得足够具体了，可以开始做区分：把“界面”从那些跟它长得像的既有概念里剥出来。

这不是学院式的概念清理。如果“界面”确实能被某个旧概念1:1替换，那么本文最多只是在为别人的框架提供一个脚注式的案例应用。更关键的是，如果旧概念已经够用，那么AI对科学批评的冲击就可以被既有的科学哲学工具箱稳妥收纳，而本文所要诊断的那种新型张力根本就不存在。本节必须同时做两件事：指出旧概念的边界，并展示这个边界恰好就是当前的真实困境所在。

3.1 不是“不透明性”：批评格式不是理解感的集体版

在现有的科学哲学文献中，讨论AI黑箱最常被调用的资源是Paul Humphreys的“认识论不透明性”概念（Humphreys, 2004）。Humphreys的诊断是：当计算科学依赖大规模的数值模拟时，科学家在个体认知层面“无法理解其工具的内部运作”——这一判断对AI加持的物理学研究似乎完全适用，因为一个深度网络内部的分布式表征确实无法被个体研究者从纸面上追踪推导。

但本文的诊断并不停在这里。一个人的局限是认知问题；两个人无法比较彼此认知成果中分歧的来源，是格式问题。不透明性概念的锚点是个体心智与认知对象之间的关系——一个认知主体是否拥有对该对象的“理解”。而认知界面的锚点是共同体内多个认知产出之间的关系——两个判断是否有某种公共格式，使得它们能被放置在同一个批评空间中进行比较和裁决。

这两者的分界可以用一个具体场景来严格标识。两位物理学家各自使用相同的从头算电子结构软件包进行密度泛函计算，但得到不同的反应势垒。他们并不需要知道交换关联泛函的解析形式，也不追踪Kohn-Sham方程的迭代求解全过程——他们在认知个体层面上对这个计算工具在很大程度上是“不透明”的。但他们仍然可以彼此批评：对比使用的泛函类型、基组截断、布里渊区k点取样——这些操作变量都装在命题格式里，可以被单独拎出来比较。批评发生了，格式通道完好，尽管个体认知并不透明。

现在将工具替换为两个神经网络模型——比如两个不同的神经量子态变分——在同一个哈密顿量上给出了矛盾的相图预言。操作变量“泛函类型”“基组截断”“k点取样”——这些命题粒度的比较项消失了。只剩下权重、激活、损失曲线。批评的格式通道断了，不是因为个体更不理解了，而是因为知识没有被装进那种能被公共操作的命题容器里。

这个对比给出了本文定义界面的第一块基石：界面不是个体理解程度的描述，而是认知产出之间能否发生可公共操作的比较批评的格式条件。它追问的不是“你懂不懂这个模型”，而是“你能不能和另一个模型的使用者坐下来、从概念层面定位你们之间的分歧”。Humphreys的不透明性概念可以很好地描述前电子结构案例中的个体状态，但没有概念资源去区分前后两个案例之间的实质差异。那个差异，正是“界面”所要命名的东西。

这同时明确了一件事：界面不等于元语言。两个黑箱模型可以通过一个共同的基准测试平台进行比较——一组相同的输入，一组约定的绩效指标。在这个意义上，基准测试平台确实提供了一种公共批评平台。但它不提供概念层面的分歧定位：它能告诉你“模型A在X区域预测错了”，但不能告

诉你“模型A为什么在X区域预测错了”。“为什么”的追问需要的是格式通道——一种能让双方从自己的模型内部提取出可传递、可检验的论证步骤的操作格式。元语言（共同的数据集、共同的评估标准）可以启动批评，但只有可追踪推导格式才能让批评穿透模型之间的概念分歧。界面的功能恰恰在于后者。

3.2 不是“不可通约性”：表面近似的两种不同结构

库恩的不可通约性概念（Kuhn, 1962/1970）描述的是科学革命前后不同范式之间缺乏共同语言的状态。表面上看，两个给出相反预测的神经网络模型争不出对错，与爱因斯坦和洛伦兹争论时空本质时的情形确实有某种表面的相似——双方都“无法用对方的概念陈述自己的判断”。

然而这两种“无法对话”的结构截然不同。库恩式不可通约性发生在不同范式之间的语言断裂：同一术语（“质量”“同时性”）在不同范式中的指称不同，以至于双方无法在术语层面直接互译。中断发生在语义层。

AI黑箱之间的“无法批评”不是这种中断。两个神经网络模型完全共享同一套物理学概念语言——哈密顿量、掺杂浓度、反铁磁序、费米面——它们的分歧，恰恰是在这些概念的语言内部被陈述出来的。双方都说“我预测这个掺杂浓度下会出现反铁磁序”，这套物理语言本身没有任何歧义。中断发生在操作层：即使双方使用完全相同的概念语言，甚至共享同一套基础理论（多体薛定谔方程），也无法从各自模型的内部提取出“我之所以这样预测，是因为……”的论证过程。不是因为语词不好使了，而是因为知识的组织格式本身不具有可提取论证步骤的结构。

这是两个层级上的困难。不可通约性问的是：“我们说的是同一件事吗？”界面退隐问的是：“我们能不能在概念上论证出分歧的来源？”不可通约性在语义层制造断裂，界面的退隐在格式层堵塞批评。界面的缺失，不是概念语言被撕裂了，而是批评动作需要抓住的那些把手——假设陈述、近似步骤、推导链条——根本不存在于这种认知格式之中。

3.3 不是“数据界面”：可操作批评格式与可共享数据通道的区分

Sabina Leonelli在《数据为中心的生物学》中提出了“数据界面”概念——在跨实验室、跨共同体的科学数据流通中所依赖的标准化格式、元数据协议与本体系统（Leonelli, 2016）。这一概念极为接近本文的关切，且共享“界面”这一术语，因此区分义务尤为沉重。

Leonelli的关切是：科学数据在什么条件下能够从一个生产语境中释放出来，进入另一个研究语境并被有效地再利用？答案的关键是“数据包装”——通过标准化分类系统与元数据协议，让数据携带着足够的语境信息，使得不同的研究团队能够“读懂”并信任彼此的数据。在这个意义上，数据界面确实促成了跨主体的公共批评：一个团队可以利用另一个团队的数据去证伪一个假说，或去质疑数据采集过程中的分类偏倚。

但本文的界面概念锚定的是一个不同的操作。Leonelli的数据界面所促成的那种批评是数据驱动型的批评：它让被批评者无法以“你的数据不靠谱”为由拒绝对话——因为数据界面已经提供了数据的出处、处理流程与质量判据。而本文的界面促成的是论证驱动型的批评：它让批评者可以从对方的理论推理链条中找出具体的逻辑漏洞或隐式假设，并据此设计一个能将两种方案拉开距离的关键实验。数据界面确保“你用的数据是可查的”，认知界面确保“你的推理链是可追踪的”。

当两个神经网络模型在Hubbard模型上给出相反预言时，双方使用的是同一套训练数据——数据界面的功能在这里是被满足的。批评仍然无法发生，因为堵死的不是数据通道，而是论证通道。这个区分给出了界面的第二块基石：界面是推理格式的可公共操作条件，不是数据格式的可公共流通条件。

3.4 收紧互锁：界面到底是什么，以及不是什么

上述三个区分可以收紧为对“界面”的一组必要与充分条件。

一个认知成果C具有界面，当且仅当：

- （必要）C的内部认知组织方式允许一个非C的建构者从C中提取出独立的、可逐项检验的论证步骤；
- （必要）这些论证步骤以某种公共可传递的符号格式（不限于数学，但必须允许符号操作而不依赖对建构者私人状态的通达）呈现；
- （充分）上述操作可以定位出C与另一个认知成果C′之间分歧在论证链中的概念来源；
- （充分）定位这一来源之后，可以根据该来源设计一个独立的、可区分C与C′的操作性测试。

这个条件集既有包含，也有排除。它包含了牛顿公理化推导、量子力学的希尔伯特空间形式、密度泛函计算中交换关联泛函的类型比对——因为这些格式都允许上述操作的发生。它排除了纯粹的数据共享平台（它们满足第二条的符号呈现，但不满足第一条的提取论证步骤），排除了基准测试的比较（满足第二条但不满足第三条——它能告诉你谁更准，但不能帮你定位为什么不准），也排除了对黑箱模型输出绩效的外部比较（同理）。

由此，“界面退隐”的含义得以精确界定：它不是指“理解变得困难”，也不是指“失去了共同语言”，更不是指“数据无法共享”。它特指：在分布式表征这一特定的认知格式中，知识的组织方式不携带可提取的论证步骤，因此即使两个认知产出共享完全相同的基础理论、概念语言与训练数据，它们之间的分歧也无法在概念层面被定位、无法通过独立的操作测试得到裁决。批评的格式通道，在这种格式内部找不到操作把手。

四、绩效与批评不可兼得的结构性根源

前一节从概念层面把“界面”与几个最接近的既有概念剥离开来。但这只是完成了概念划界，还没有碰触一个更根本的问题：分布式表征为什么就产不出界面？如果界面只是“暂时没被开发出来的附加功能”，那么工具主义者就可以轻松回应：“等可解释性技术成熟了，问题自然解决。”本节的任务是论证这件事比“暂时不成熟”要深——分布式表征能在高维空间中捕捉旧概念尚未命名的模式，这正是它精确预言的优势所在，也恰恰是它无法将这些模式无损压缩进命题式界面的代价。优势和代价来自同一个根。

4.1 两个真实的案例

先让这两个案例把问题逼到可操作的程度。

案例一：湍流边界层的神经网络壁模型

在大涡模拟中，近壁湍流的直接解析成本极高，传统做法是依赖基于物理假设的壁模型来参数化近壁通量。近年来，研究者开始用全连接网络和图神经网络替代传统的代数壁模型，在高雷诺数槽道流和边界层模拟中取得了优于传统模型的预测精度。有研究者训练了若干不同架构的神经网络壁模型，在大致相当的训练数据和预测精度下，在极高雷诺数（试验数据尚未充分覆盖的区域）的预测上给出了分歧：有的预测对数区的外推保持经典的卡门常数，有的则倾向于壁面律的斜率在不同粗糙度下发生漂移。

分歧出现之后，双方的诊断路径跟量子多体案例完全平行。研究者可以比对网络架构（全连接还是图结构）、输入特征集（是否包含压力梯度历史项）、损失函数的各项加权方式——所有这些比对，都是表层操作条件的比对，不是内部知识格式的比对。被问的是模型被设定的操作条件，而不是模型在这些条件下学到了什么。没有一个操作可以沿着论证链条进入到“你的模型到底在近壁区编码了什么样的速度-涡量耦合模式”这一层——因为那个模式被分布在数百万个权重里，没有任何单独一层或单独一组神经元充当它的命题级替身。

案例二：高能喷注的深度学习分类器

在大型强子对撞机的物理分析中，区分由夸克或胶子引发的喷注是一项核心任务。传统方法依赖QCD第一性原理构造的可解释的喷注子结构变量，这些变量的物理动机清晰，但区分能力存在瓶颈。深度学习分类器——从卷积网络到粒子流网络再到基于注意力的点云Transformer——相继被引入，在标准基准上不断刷新区分性能的纪录。

然而，当将这些深度学习分类器应用于蒙特卡洛生成的喷注样本之外的真实碰撞数据时，一个麻烦浮现了出来。不同的生成器给出的夸克/胶子组分比例不同，不同的部分子簇射和强子化模型给出的喷注内部关联也不同——这些差异在训练数据中已经存在，网络将它们一并学到了内部的分布

式表示中。当两个在不同生成器上训练的分类器在同一批真实数据上给出相反的标签判定的比率差异时，问题的性质并不是哪一个分类器“错了”，而是这两个模型在内部以极不相同的结构混合了“来自物理信号的模式”和“来自生成器特定假设的模式”，而这两种模式在分布式表征中是纠缠在一起、不可分离的。

研究者当然可以在各种性能指标上比较两个分类器。但当一个分类器将某个特定喷注标记为夸克起源、另一个标记为胶子起源时，你无法去问“分歧的原因是这个喷注的运动学变量被两个模型以何种方式、依据什么特定的假设映射到了不同的判定边界上”。那个映射存在——它就在权重里——但它不被命题格式封装。批评被堵死的不是数据的可及性，不是对问题定义的分歧，也不是物理概念语言的不一致——而是在所有这些条件都齐备的情况下，论证步骤本身无法被提取出来。

4.2 来自认知格式本身的约束：分布式表征无法自我命题化

这两个案例指向的是同一个结构。在传统命题式认知格式中，物理学家做的是这样的操作：把对自然的假设置入可追踪推理链中，推导出若干可观测后果，然后用实验去检验。检验的对象是一个可被单独拎出来的假设——一个有限粒度的、可被修改而不必然推翻其余推理链环的命题。批评和被批评都围绕着这些独立的、可被检验的环。

分布式表征的认知操作方式完全不同。在深度网络中，输入被逐层非线性变换，最终映射到输出。整个映射是一体化的：从输入到输出的映射函数没有可被单独勾连的中间命题，只有一层层的数值变换。网络学到的东西被分布在这些变换的参数里——同一个“概念”（如果可以这么叫的话）对应的是参数空间中的一支分布，而不是一个句子。当网络学到了一种复杂的相分类模式，这个知识不是作为一个可被陈述的命题被存下来的，而是作为一组使得某些输入产生某些输出的权重被存下来的。

优势就出在这里。命题格式要求知识在进入容器之前，必须先被离散化为有限粒度的概念。在高维连续空间中的复杂模式——比如量子临界区的连续相变形——在命题化时必然丢失它在连续变形过程中的那些精细结构。分布式表征不进行这种离散化，因此它能够完整地保留对高维模式的编码。这是它精确预测能力的关键来源。

代价也出在这里。同样的这种不承诺离散化，堵死了一种基本操作：从认知成果的内部，把知识拆成若干个环，单独拎出一个来说“我们在这个假设上不一致”。这种操作的前提，是知识被封装为一组命题的集合——因为只有命题，才能被单独拎出来而不让其余推理坍塌。分布式表征从一开始就没有把知识切成这种环，它把整个映射函数揉成一个连续体。在这连续体内部，无论用什么后处理技术去投影、提纯，只能得到近似——这些近似在某些区域可能有用，但恰好在物理学家最需要批评的那些极限区域，它们是最不可靠的。

优势与代价是一体两面。这不是说分布式表征存在一种逻辑缺陷，能在未来被某个更好的技术消除；而是说，分布式表征的特定结构，使得它在一类问题上获得精确预言的优势时，同时必然地丧失了另一项功能——让知识能被拆解为可公共追踪的命题链。这两者是同一种组织的两种后果。增强后者的尝试（更精致的可解释性工具）本质上是在让网络的一部分运作被强行翻译成命题格式，而翻译操作在高维连续模式的最复杂边界——正是界面驱动的批评最有价值的地方——必然遗漏信息。这并非说批评被推迟了。它是说：在这种认知格式的给定结构下，将知识格式无损地从分布式转化到命题式，在数学上无法得到保证。

五、反驳与回应：最尖锐的四把手术刀

至此，全文的核心判断已经立出。现在需要请出它可以预见的最强反驳，正面回应。回避最强反驳的论证是自我安慰，不是学术判断。

5.1 工具主义反驳：预测足够，批评可以等数据

最强反驳来自工具主义立场。一大批计算物理学家——尤其在凝聚态物理和高能唯象学的前沿——对黑箱模型持有完全实用的态度：只要模型能给出足够精确的可实验检验的预言，区分两个相竞模型的任务就应当交还给实验本身。等有了新的实验数据，谁对谁错自然分明。在这个过程中，“概念层面的批评”是一种过时的人文执念，而不是科学实践的必要环节。如果实验才是最后判官，为什么还要在乎实验前的批评？

这一反驳在直觉上极强，而且它抓住了本文的一个潜在风险：把“界面”做成了一种对旧有认知习惯的多愁善感，一种“物理学家需要理解感”的心理主义辩护。

我的回应放弃心理主义，转而追问功能性后果：如果批评的界面被大规模废置，知识共同体将丧失哪些它当前高度依赖的运作能力？这些能力能否被实验数据的高频供给所完全替代？

此处需要一组历史比较。托勒密天文学的本轮—均轮体系在预测行星位置方面达到了相当高的精度——可以误差控制在一度以内——而它对天体运动机制的物理描述（地球位于宇宙中心，行星沿本轮运动）被证明是错的。为什么这些高精度预测没能保住地心说？不是因为预测不够准，而是因为在没有假设天体物理本质的前提下，本轮体系无法回答一个操作层面的问题：“当连串相互嵌套的本轮模型在同一数据上给出精度相当而本轮数目与层级设定明显不同的预言时，究竟哪一个设定了合理的本轮配置？”当新的观测数据——如金星相位的变化——与地心说的预言出现系统性偏差时，体系无法定位是哪一个基设级而非几何可调的层级引发了偏差。面对新数据时，体系的推演能力已经耗尽：每一次修正都需要添加一层新的本轮，但没有任何公共流程可以预先决定本轮应该加在哪里、加多少层。批评界面的缺失使精确预测机器无法自我修正。

同样，标准模型的量子场论在希格斯粒子发现前已经对大量观测做出了极高精度的预言——但它之所以不是“粒子物理的本轮体系”，正因为它内部嵌入了可批评的界面：电弱对称性破缺机制是一组独立的假设，微扰幺正性的自洽约束是一组独立的条件，两者的联合可以被用来“在撞上去之前就预言”如果希格斯机制是正确的，什么样的能区应该出现什么样的共振峰。批评界面具备的恰是实验数据尚未触及时的推演与限缩能力——它帮助共同体在面对未知之前，预先锁定了何种新数据能够使理论存活以及何种新数据足以使其致命。

AI黑箱模型目前既没有托勒密式的可紧急堆叠本轮的灵活性，也没有标准模型那种可超前预判“应当去寻找什么”的推演性。当两个模型在尚未实验探索的区域给出相反预言，工具主义者说“等实验数据”。但实验数据在若干物理学前沿的更新周期是十年级别的——高温超导的配对机制、中微子质量排序、暗物质的非引力检测——在这些领域，如果知识共同体在长达十几年的数据空白中失去了在概念层面争论和限缩可能空间的能力，认知生态会发生什么变化？

它会从“共同推理—共同探索”的组织模式，滑向“各自猜测—等待判决”的组织模式。在这种模式下，研究者的工作不再是彼此批评以逼出现有知识的裂缝，而是各自提交一个预测，等待未来的实验来宣布胜出者。理论评估，也就从共同体内持续的论证性批评，坍缩成了单次事后打分的竞赛。概念创新——新物理观念被逼出的那个过程——赖以生存的张力消失了。那些最重大的概念创新——狄拉克的反粒子、南部—戈德斯通定理、威尔逊的重整化群——没有一个是纯粹从数据里冒出来的。它们是在批评界面的运转中，从旧概念的裂缝里被挤出来的新形式。如果批评的界面在实验中场休息期间退场，这种裂缝将不再被看见。

5.2 AI可解释性反驳：界面正在重建的路上

第二个强反驳来自可解释性AI的前沿。当前已经有概念激活向量方法可以从神经网络内部提取出对反铁磁序敏感的激活子空间；有基于梯度的显著性映射可以标示出输入中的哪些特征驱动了分类决策；有符号回归算法可以从训练后的网络中提取近似的解析模型（如从分子动力学轨迹中逼出简化的哈密顿量）。这些可解释性工具构成了一条重建界面的技术路径：不是直接从权重推导，而是从权重中提取出命题级的、可被追踪的近似概念。

如果这条路径走得通——如果未来有一种可解释性方法能从任何一个训练良好的深度网络中提取出一套无损的、与网络原始输出在所有输入区间上都一致的命题化模型——那么本文的核心判断就被推翻了。这正是本文在结论部分将给出的证伪条件之一：一旦可解释性技术达成了从分布式表征到命题格式的无损对接，界面退隐就被逆转了，本文直接失效。

但这个可能性有多大？

从深度网络内部提取出来的“概念”（如激活子空间对某一物理序参量的选择性响应），其定义高度依赖于提取方法的选择——概念激活向量方法中“概念示例”的选取、显著性映射的平滑半径、

符号回归时的目标函数形式，每一个方法论的选择都相当于在翻译之前先规定了“你要翻译出什么类型的物理概念”。对于量子反铁磁序这种人类已经命名且范例清晰的概念，提取是相对有效的。但对于量子临界区中那种处于两种序的连续变形过程、人类尚未命名、没有清晰范例的模式，翻译工具从一开始就缺少命名目标。旧概念尚未融化时，AI可解释性是有效的。正是在旧概念正在融化的前沿——这是批评界面最有价值的地带——可解释性方法丧失了锚定翻译目标的能力。

这又与分布式-命题化翻译的结构性流失这一论点形成了互锁。可解释性方法不是在原表征的连续体中任意漫步，它必须有所选择：哪些模式被翻译、哪些被忽略。而选择标准本身，依赖人类已有的概念词典。在人类概念词典的半径之外，网络上可能已经编码了对某些奇妙物理模式的敏感性，但没有任何可解释性方法能发现它——因为能发现的只能是方法预设了命名目标的那类东西。那些真正新鲜的知识，将永远被封存在分布式参数的沉默里，只有等到实验数据已经摆到桌面上，它们才通过“预测得对”这一事实间接地发声。但那时，已经不是批评在引导新知识的发现，而是实验在事后宣告某种知识已经被发现。

批评的导引功能——康德到爱因斯坦到南部的那条通过论证从旧概念中提前逼出新概念的路线——正是这个功能，在可解释性方法的结构性局限下，无法被回收。

5.3 实用主义反驳：把批评分散到各种外围比对即可

第三个反驳比前两个更细腻。它并不声称黑箱模型本身是可追踪的，而是声称：即使黑箱内部不可追踪，物理学家可以——而且事实上正在——通过一系列对黑箱的外围操作来维持有效的公共批评。这些操作包括但不限于：比较不同架构在相同数据上的归纳偏置与泛化行为、设计能暴露模型脆弱性的对抗样本、在干预式输入下测试模型的行为一致性、比对模型在不同子群体上的失效模式等。

这一反驳的核心逻辑是：批评界面的功能可以由一系列“外围开放性测试”的集合来替代，即使内部推导链不可追踪，批评的手感仍然可以从外部找到。这相当于一个论证：当无法看到黑箱内部的齿轮时，一群人可以通过从不同角度摇晃箱子、敲击外壳、投喂特殊输入，听声音辨质量，在大量实验的实证基础上形成一套“外部批评”的实践。

我的回应认可这些外围操作的部分价值，但指出它们与界面之间存在着不可桥接的功能落差：这些操作不能定位概念分歧的来源。

比较归纳偏置可以告诉你“这批神经网络架构在这类数据上系统性地优于那批”，但无法告诉你“它们在哪个物理假设上与真实物理一致/不一致”。测试对抗样本可以揭示“在这个特定输入方向上模型不稳定”，但无法帮你判断这种不稳定性对应于模型内部学到了什么以及未学到什么。外围操作能揭示行为差异——模型A和模型B在某个区域输出不同；但不能揭示概念差异——A和B对这同一个区域的物理到底被推定为不同在什么假设上。

界面驱动的批评，与外围绩效比对的批评，两者的功能差异在最关键的那个节点上被拉开：当你需要从模型的分歧中逼出一个新的可检验预测时，你需要知道分歧的原因在哪个概念节点上。外围操作可以提供“这里有一个行为差异”的位置标记——在设计关键测试的第一步上，它是有效的。但在从“这里有一个行为差异”到“所以我们应该去测以下这个物理量”的那一步上，它需要有关于模型内部表征的概念假说才能推进。而这个假说，恰好是被黑箱格式封存的。

外围批评把批评分散成了一连串行为层阶的检查，却未能在概念层阶把分属不同模型的两套知识放在同一个论证空间中连接起来。它更像是各模型之间的“黑箱互操作性测试”，而不是“概念平面上的推理论辩”。这种测试当然有用——但在实验数据稀缺的前沿地带，它不能替代界面曾经在执行的那个推演性功能。

5.4 建构论反驳：界面只是共同体的社会协定

一个更根本的质疑来自科学社会学方向。根据这一立场，科学批评所依赖的“公共格式”从来不是认知格式的自然属性，而是科学共同体在特定历史条件下创建并强制执行的一套社会协定。一个东西“算不算合理的批评”以及“怎样算独立的检验”，本身就是共同体内部磋商的结果，而非被认知格式决定的。如果分布式表征导致了批评困难，那么随着AI深度卷入物理学研究，共同体自然会重新协商出适应新认知生态的批评实践。所谓“界面的退隐”在实质上是对一段转型期不适应的社会学误诊。

这一反驳的力度在于，它的确准确命中了“界面”在与历史缠结的维度上的边界。界面确实是历史性的：牛顿的公理化格式在17世纪末刚出现时曾遭到笛卡尔派的猛烈抵制——不只是认识论的抵制，而是对“什么才算有效的批评”的规范性抵制。量子力学的希尔伯特空间格式也曾引发“直觉失落”的哀叹——但它最终被接受为新的公共批评平台，并且运作良好。建构论正确的一点是：界面的规范地位可以被共同体重新协商。

但它错误的一点是：共同的协商本身也需要一个能够承载协商过程的界面。当爱因斯坦在索尔维会议上批评玻尔的量子力学时，他依靠的正是量子力学格式所提供的可追踪推导结构：他能指出某个具体实验安排的推导在何处与测不准原理发生具体冲突。这个操作之所以可能，是因为量子力学的格式允许他将推理分解为可独立批评的步骤——这是一个格式条件，而不是一个社会学协定。“共同重新协商一个新界面”这句话在暗中预设了协商能够发生的条件——协商者各自都能从自己的认知产出中提取出可被对方独立操作的论证。而分布式表征在当前以权重空间的形式承载知识时恰好抽走了这个前提。建构论的批评若在本文的逻辑框架中站住脚，它首先需要说明“重新协商”如何在没有格式界面的条件下完成——它不能让协商本身依赖于它旨在论证会自然产生的那种格式。

简言之：不是共同体“有没有打算协商”的问题，而是“有没有可以用来协商的公共格式”的问题。前者是社会学的，后者是认知格式的。界面的退隐，是指后者开始消失，而前者的意愿无论多么充

沛，也不能在真空中维持批评的动作。

六、界面在过去并非从未遭遇震荡——以及这一次有什么不同

前三节的核心论证已经做完，但一个合理的历史性质疑尚未被认真面对：物理学在历史上经历过多次“旧批评格式被震碎”的时刻，而每一次最终都结算出了新的格式。凭什么说这一次就特别麻烦？

这个问题问得对。本文不声称“AI黑箱是物理学史上唯一的不透明工具”——正相反，前文已经承认，密度泛函、蒙特卡罗模拟、甚至重整化群本身在历史的某些阶段都曾被指责为“无法被直观理解的黑箱”。但在那些案例中，工具的运作逻辑即使对个别的使用者而言在认知上是遮蔽的，对共同体而言仍然保有可提取论证步骤的界面。是分布式表征，把“提取论证步骤”本身的通道堵死了。这一节展示这并非理论上的杞人忧天，而是两个真实记载中都能够清楚看见、正在发生的操作级断裂。

6.1 湍流边界层壁模型：被架空的物理解释

前面已经用湍流壁模型展示过两套黑箱在操作表层的比对。这里再往下推一层：当界面的退隐从外部绩效分歧卷入概念裂缝，它最先是透过什么被察觉到的？正是在“物理解释被架空”这个具体切面上。

传统的代数壁模型——比如基于对数律的壁面函数——有一个可被批评的物理内核：在充分发展的湍流边界层中，惯性子区与黏性子区的通量匹配假设是作为可被独立质疑的命题被陈述的。如果一个模型在某个流向压力梯度的区域内崩溃，诊断路径是清晰的：检查压力梯度项在壁面律推导中的量阶假设是否在那个区域内失效。批评能够沿着这个命题定位出故障的位置，并根据这个物理机制设计纠正方案——比如换用非平衡壁模型。

神经网络壁模型没有这种可被隔离开来独立批评的物理内核。训练完成之后，被网络学到的近壁湍流表征不是一个或几个命题的集合，而是一种高度纠缠的输入—输出映射。当网络壁模型在某个雷诺数下漂移了壁面律的斜率，你无法去问：“这是因为它学到的近壁条带结构错误，还是因为它对逆压梯度的响应函数在训练数据的分布尾部没有被充分约束？”

物理“解释”在这里被架空——不是没有解释，而是解释失去了它传统上承担的那个批评功能。解释本是批评的工具：一个好的物理解释能告诉你，如果某个机制是错的，你应该在什么实验条件下观察到什么样的系统性偏离。神经网络壁模型可以被后验地“解释”——可以把输入—输出关系做一个敏感性分析，说“这个输入特征贡献了最大梯度”。但这种后验解释不能承担批评的推演功能——它不能告诉你如果模型错了，错误的概念来源在哪儿，以及为了暴露错误，你应该去设计一个什么样的新测试。因为解释此时所解释的对象不是推理链中环环相扣的命题，而是一个连续的、非命题的映射。

6.2 高能喷注：被纠缠的物理与人工

QCD喷注分类器将这个问题从“单个黑箱”推进到“多个黑箱之间的批评”，从而把界面退隐的操作性后果展示得更清楚。

不同的蒙特卡洛生成器——比如Pythia和Herwig——对部分子簇射的角序处理和强子化模型的参数化方式有着不同的物理假设。当一个深度学习喷注分类器在Pythia样本上训练后，它的内部表征混入了Pythia特有的、不一定在所有生成器中都成立的喷注关联模式。当这个分类器与另一个在Herwig上训练的分类器在同一个真实数据的喷注上给出相反标签时，发生器特定假设与真实物理信号之间的边界，在分布式表征内部是缠在一起的。它们不是作为可分离的先验被存下来的，而是作为联合决定最优决策边界的因素，被网络混在一起优化的。

这种纠缠的操作性后果是：无法把分歧的源头转化为一个可独立检验的新测量。在命题格式下，我可以说：“我们的分歧来自于你是否在推导的第三步假定了角序。让我们去设计一个对角序敏感的喷注形状变量，来独立检定它。”界面允许你根据分歧的概念来源创造出这个新变量。在黑箱格式下，双方都会同意“分歧来自于Pythia和Herwig在某些关联上的处理不同”——但没有任何一方能说清，“在我的网络里，这种不同是通过什么具体机制转换成输出反差的”。你的网络内部将Pythia偏差与真实信号的差异变成了一个高维非线性映照——你无法把自己模型的决策从其训练数据的特定假设中剥离出来推给另一方看。

七、界面在什么条件下恢复——与在什么条件下本文失效

经过前面几节的分析，我们站在了一个清晰却不安的位置。本节直面这个位置，不回避任何张力，也不粉饰未完成之处。

AI黑箱的高精度预测正在物理学前沿迅速扩散。这一扩散不是偶然的——它在那些数据丰富但理论滞后的领域（如湍流壁面律、高能喷注分类、量子多体相分类）展现出了传统方法难以匹敌的工程优势。物理学家正集体性地走入一个认知生态：知识的大规模生产，正在从它所依赖的那套公共批评格式中脱嵌出来。

这并非一篇哀悼文。界面从来不是被物理学一次性享用的永恒礼物。本节的任务是指明当前正在堆积的那些张力，以及在这种张力中可能使新的公共批评格式得以出现的条件。

7.1 界面结算是怎么发生的

回看量子力学的格式转型。从海森堡1925年发表第一篇矩阵力学论文，到冯·诺依曼1932年完成希尔伯特空间的严格公理化，只过了七年。在这七年里，发生了两件事的同步推进。第一，一种新的实践——用矩阵和算符而不是轨迹和速度来导出可观测量的数值——在大量具体问题上被证明极度有效，迫使所有严肃的物理学家承认必须学会这种新语言，无论它多么不直观。第二，这种新实践本身的运作规则被迅速提炼为一套可公共传递的格式：态空间、算符、对易关系、概率解释、测量公设。这七年间，批评非但没有暂停，反而在每一次会议和每一封信中猛烈进行——玻尔和海森堡、薛定谔和玻恩、爱因斯坦和玻尔之间的每一次交锋，都依赖着正在成形中的这套格式。他们在争论的同时，也在建造争论所需的舞台。旧界面的瓦解和新界面的建立，不是先后发生的两件事，而是同一个过程的两面。

现在回到AI物理学。分布式表征已取得了足以让严肃的物理学家学习它的工程有效性，但它尚未从自身的实践中长出一套可公共操作的概念格式。没有这种格式，批评的剧场就没有舞台。

7.2 新界面结出的可能条件

从量子力学的先例中可以提取出若干新界面结出的条件，这是一组被历史检验过的动力学变量，不是规范性指令。

第一，分歧的压迫必须足够痛苦。在量子力学早期，旧量子论（玻尔—索末菲模型）对光谱线的预测已经达到了很好的精度，但它对氢原子谱线精细结构、反常塞曼效应和氦原子基态等问题的系统性失败制造了一种不可忍受的认知痛苦——不是“模型不够准”，而是“现有框架内部怎么改良都碰不到那些问题的结构”。波动力学和矩阵力学之所以能迅速被接受为新界面，正是因为它们在旧界面碰到的那些问题上打开了旧工具打不开的新操作空间。

当前AI黑箱模型的痛苦可能还没有积累到触发下一轮格式结算的阈值。在绝大多数当前的应用中，AI模型的预测误差仍然在可接受的工程范围内；真正尖锐的、让共同体感到“旧批评格式完全失灵”的冲突，目前集中在少数前沿地带——如量子临界区的基态竞争、喷注分类的生成器依赖性——尚未以足够广泛的规模扩散。这并不意味着冲突不真实，只意味着它尚未变得不可忽视。

第二，新表征必须从实践中自己长出元语言。矩阵力学能够成为界面，不只是因为物理学家“决定使用它”，而是因为它内部的运作规则——矩阵乘法、对易关系、本征值问题——本身就构成了一套可被独立操作的符号体系。物理学家可以通过学会这套规则来批评彼此的推导，而这套规则是从矩阵力学的具体实践中被抽象出来的，不是从外部被贴上去的。

分布式表征目前缺少这种自发的元语言。当前使用的各种可解释性工具，本质上是外部贴上去的翻译器——它们用旧概念语言（“这个神经元群编码了反铁磁序”）去描述网络内部的某些局部模式，而不能从网络自身的运作规则内部生成一套批评语言。新界面的出现需要一个在分布式表征系

统的内部运作与可公共操作的论证语言之间建立严格映射的翻译机制——这种映射尚未产生。

第三，共同体需要被逼到必须协作建设的境地。量子力学的格式结算不是某一个天才单独完成的——它是玻尔、海森堡、薛定谔、狄拉克、冯·诺依曼和其他数十位物理学家在激烈的批评性对话中共同建造的。他们来自不同的背景、使用不同的数学工具、对旧物理学的可靠性持有不同的态度——这一多样性恰恰是格式结算的资源，因为每一方的视角都能暴露其他方的盲区。

AI物理学共同体现在还在使用AI工具的不同子群之间处于相对隔离的状态。凝聚态物理学家使用神经量子态，高能物理学家使用喷注分类器，等离子体物理学家使用强化学习控制器——他们各自撞上了各自的界面退隐困境，但尚未形成一个集体的压力，促使不同的子群坐在一起建造一套共享的批评格式。量子力学的先例表明：界面的结算需要不同工具箱的使用者在同一张台上激烈碰撞，工具的分歧才能逼出格式的创新。

7.3 本文的证伪条件

一篇诚实的论文，必须给出自己的死亡方式。这不是修辞姿态，而是把判断从不可证伪的哲学断言拉到可操作的科学讨论层面的唯一步骤。本文的判断是：分布式表征的结构决定性地使得从这种表征内部提取可公共批评的论证通道无法做到完全无损。这条判断在遇到以下任何一种情况时将被宣告失效：

1. 技术证伪条件。在任何一个已经确立的物理问题上（量子多体、湍流、高能唯象或其他），有团队展示了一种可解释性工具，从训练好的神经网络模型中提取出了一套完整的、可被非网络建构者的独立研究者操作的命题化物理推导链，并且该推导链在所有相关参数区间上的预测都与原网络一致、且不依赖该可解释性工具的参数选择（如示例选取、降维方法、目标函数权重）。一旦此条件被满足，界面退隐即被逆转，本文失效。

2. 实践证伪条件。如果出现了一个有可靠文献记载的真实案例，其中两个竞争性的AI模型在概念层面成功进行了批评——即研究者们在不依赖新数据的情况下，从两个模型的内部表征中找出了分歧的概念来源，并据此设计了一个能将两个模型拉开距离的独立操作测试，且该测试的语言和步骤是任何受过相应学科训练的第三方都能独立重复操作的。此案例必须发生在本文论证所指的核心领域——即模型的认知产出以纯分布式表征的方式存在，而非以混合式（分布式加符号约束）的方式存在。此条件的满足，意味着界面可以在分布式表征内部被重建，本文的核心判断失效。

3. 历史证伪条件。如果在本文发表后的相当长一段时间内（比如一代研究者的职业周期），物理学界在持续大量使用AI黑箱模型的同时，并未观察到批评实践的系统性衰退——例如，概念创新并未减缓、共同体在实验数据稀缺区间仍能通过论证性批评限缩理论空间、研究实践并未从“共同推理”滑向“各自预测、等待裁决”——那么本文对影响的诊断虽然未必被逻辑推翻，但确实被经验削弱到不再具有现实重要性。这个条件不是让论文在逻辑上失效，而是让它变得无关紧要——对于一篇

试图诊断真实张力的论文来说，这两种失效在认知价值上是等价的。

这三组证伪条件覆盖从技术层到实践层再到历史层的递进层级。它们可以被独立检验，互不依赖，任何一个的满足都构成对本文的不同层级但同等严肃的挑战。本文邀请读者用这些条件去检验它的核心判断——这不是在说客气话，这是对论证的真正尊重。

八、结论：一道裂缝的诚实素描

让我们回到一开始的那个逼问：两个黑箱模型给出相反预言，而实验数据要等上好几年——在这段沉默的间隙里，知识共同体能做什么？

本文的回答是：在分布式表征的结构性地抽走了批评的格式通道之后，共同体能做的是意识到这种抽走的发生——然后直面由此堆积的张力。这不是一个悲观主义的结论。悲观主义者说“再也回不去了”，而本文说的是：旧界面不是被破坏的，而是被特定认知格式的特定优势所自然扬弃的；新界面如果要出现，只能从这种扬弃所制造的张力之中被挤出来，而不是靠怀旧被唤回。真正紧迫的问题不是“界面退隐是不是灾难”，而是“如果新公共格式要在当前的结构条件下被发生出来，它必须包含哪些旧格式所不具备的组件”。

这些组件中，至少有两样是可以从历史类比中猜到其必要性的。第一，一种能够从分布式表征内部被自动提取——而不是从外部被贴上——的可公共操作的概念粒度的论证语言。量子力学的希尔伯特空间格式之所以能成为界面，正是因为它无需外部翻译：态矢量、算符、对易关系，这些就是理论自己的运作概念本身。第二，一种能将多个异构分布式模型的内部表征，映射到同一个公共批评空间中的元结构。这比前一个更难，因为它要求在模型之间建立严格的可比较性，而不只是从单个模型内部提取近似的命题式替身。

目前，这两样组件都还不存在。承认这一点，是诚实的。指出它们的必要性，是对未来可能性的一个发生学追问——不是预言，而是标定它们如果出现，必须满足的功能条件。

在现有的认知生态中，物理学正以共同体尚未充分承认的速度，把知识生产的基础设施从命题式格式迁移到分布式格式。这不是说命题式思维会被取代——在可预见的未来，任何物理学家仍然会用纸面上的推导来设计实验和检验论点。而是说，在数据最密集、预测压力最大的那些前沿地带，知识的大规模近产已经开始运行在一种与公共批评格式脱嵌的认知底层之上。这种脱嵌不是故意的，也不是恶意的——它是工程选择自然而然导向的后果。但它一旦发生，就会制造一种表面上难以察觉的裂缝：物理学家仍然在发表论文、引用彼此、在会议上争论，但争论依托的不再是可被公共操作的概念论证，而是对彼此模型的外部绩效的比对。这不是批评消失了——消失的是批评的概念把手。

这道裂缝是否会成为新一轮格式结算的推力，取决于它在未来能否变得足够痛苦——足够让共同体意识到，旧批评方式已经运作在一种被抽空了格式支持的虚空之中。本文的全部论证，不过是为这个意识提供了一组可操作的辨别工具。至于裂缝会往哪个方向裂开，诚实地说，我不知道。但我知道的是：一旦你能看见裂缝，你就不再能假装它不在那里。而这，是任何结算得以发生的前提。

证伪条件陈述（作为结论的最后一段）——本文的核心判断会被以下任何一种情况的出现所推翻：（1）有技术手段从训练后的深度网络中提取出完全无损的、可被独立操作的命题式推导链；（2）有真实文献记载的案例展示了两个纯分布式表征模型在概念层面成功进行了批评，且批评的语言是公共的、可被第三方独立操作的；（3）在长期实践中，物理学共同体的批评能力并未随黑箱模型的扩散而出现可观测的衰退。三者满足其一，本文失效。

———

参考文献

Carleo, G., & Troyer, M. (2017). Solving the quantum many-body problem with artificial neural networks. Science, 355(6325), 602–606.

Humphreys, P. (2004). Extending Ourselves: Computational Science, Empiricism, and Scientific Method. Oxford University Press.

Kuhn, T. S. (1970). The Structure of Scientific Revolutions (2nd ed.). University of Chicago Press. (Original work published 1962)

Leonelli, S. (2016). Data-Centric Biology: A Philosophical Study. University of Chicago Press.