形式作为终点而非起点：柏拉图表征假设的解释反转

魏晓海[①]

摘要：柏拉图表征假设（Platonic Representation Hypothesis，PRH）由 Minyoung Huh 等人于 2024 年提出，描述独立训练的神经网络中观察到的内部表征收敛现象。该假设的标准解读指向实在论方向：足够强大的学习系统朝向预先存在的抽象结构收敛，视之为柏拉图形式论的当代实证辩护。本文针对这一解读提出双层论证。第一层：柏拉图式本体在解释上是多余的——收敛吸引子理论在无需任何本体承诺的情况下已对收敛现象提供了完整解释。第二层：本文解释为何柏拉图解读会如此令人信服——某些吸引子的极度稳健性（在一切可能的优化过程中都无法回避）在认知上产生”本体独立存在”的现象学感受，这是由闭环结构与稳健性条件共同决定的结构性后果，而非偶然的心理偏差。借助结构实在论、动力系统理论与自然主义数学哲学，本文提出反转的柏拉图表征假设（Reversed Platonic Representation Hypothesis，RPRH）：大规模学习系统中呈现为收敛表征的抽象结构，是优化过程在共享约束下产生的涌现不动点，而非被学习过程逼近的本体先在；人们认为这些结构必然独立存在的直觉，本身即是遭遇极度稳健吸引子时不可避免的认知结果。收敛结构是真实的，但其真实性是吸引子模式的真实性，而非独立本体论领域的真实性。

关键词：柏拉图表征假设；抽象对象；结构实在论；收敛吸引子；数学哲学；自然主义；神经网络收敛

中图分类号：N02；B014；B015　文献标识码：A

一、引言

当代机器学习研究揭示了一个引人注目的经验规律：在完全不同的任务上、以不同架构、在不同数据集上、从不同随机种子出发训练的神经网络，其内部表征在几何上彼此相似——且这种相似度随模型规模增大和训练时长增加而单调提升[1]。Minyoung Huh 等人将这一现象命名为**柏拉图表征假设**（PRH），这一命名并非随意之举：他们的解读是，不同的学习系统都在朝向同一底层现实结构收敛，正如柏拉图所认为不同的具体三角形都在不完美

[①]魏晓海，计算机科学博士，独立研究者；E-mail: wistoch@ustc.edu。

地逼近” 三角形的形式”。

这一经验发现有充分的文献支撑。针对特定网络类别，收敛的形式化证明已经建立[2]，且该现象在视觉、语言与多模态模型中普遍存在。本文接受该现象的经验内容，所质疑的是强加于其上的形而上学诠释。

对 PRH 的标准解读以一种几乎未经审视的方式继承了数学柏拉图主义：抽象结构独立于心智和学习过程而存在，正是这些结构解释了收敛何以发生。强大的网络之所以收敛，是因为形式预先存在以待逼近。在这一解读下，PRH 不过是一个古老形而上学命题的新经验论证。

本文论证，这一解读存在双重错误，且两者相互关联。收敛现象无需诉诸任何独立的抽象对象领域即可得到充分解释：当众多优化过程在足够相似的约束下运作时，共享吸引子的涌现是必然的结果。Minyoung Huh 等人所称的” 柏拉图结构”，是基于梯度的学习在其极限处产生的稳定不动点，而非学习去发现的预存模板。然而，论证并不止步于此。若柏拉图式本体在解释上是多余的，一个进一步的问题便立刻出现：为何它显得必要？为何数学家、以及如今的机器学习研究者，如此自然地得出结论认为他们所收敛到的结构必然独立地存在？本文在同一框架内回答这两个问题。

这就是**反转的柏拉图表征假设**（RPRH）：

> 在大规模学习系统中作为收敛表征出现的抽象结构，是优化过程在共享约束下的涌现不动点，而非独立于学习过程而存在并被那些过程所逼近的本体先在；认为这些结构必然独立存在的直觉，本身即是遭遇极度稳健吸引子——任何优化过程都无法回避的吸引子——时自然却并非逻辑上不可避免的认知结果。

论证由此形成双层结构。第一层是标准的解释简洁性论证：柏拉图式本体对收敛的解释毫无增益，因此我们没有科学理由引入它。第二层是一种特殊类型的解释-还原论证：它不是将柏拉图直觉斥为混乱而加以驳斥，而是解释这种直觉——柏拉图主义，是极度稳健的吸引子从内部所呈现的样态。诉诸独立本体领域的冲动是可以理解的，并非缺乏根据——它是对参数空间中一个真实特征的自然反应。然而，这是一种误认：稳健性被误读为本体独立性。

两层论证相互强化。简洁性论证表明柏拉图主义并不必要；解释-还原论证则消除了如下隐忧：自然主义诠释一定遗漏了柏拉图直觉所追踪的某个真实之物。两者合力构成对 PRH 标准实在论解读的反驳：柏拉图式本体既不需要，诉诸它的冲动亦是可以解释的。

本文结构如下。第二节重建对 PRH 的标准实在论解读（SRR）并梳理其哲学承诺。第三节诊断该解读的固有困难。第四节将 RPRH 阐述为正面主张，以收敛吸引子理论和结构实在论为基础，阐明其与过程哲学的关系，解释极度稳健的吸引子何以产生柏拉图直觉，回应关键反驳，并阐发其对机器认知问题的含义。第五节为结语。

二、柏拉图表征假设及其标准实在论解读

经验现象

Minyoung Huh 等人[1] 报告，不同神经网络的隐藏表征——通常以中心化核对齐（CKA）或表征相似性分析（RSA）度量——并非任意的，而是呈现出聚类分布。在不同语料库或任务上训练的规模足够大的网络，对于相同输入会产生几何上相似的内部表征。Ziyin 和 Chuang[2] 在理想化条件下（深度线性网络）给出了形式化证明：在相同数据分布上训练的网络，无论初始化如何，其表征几何最终收敛至同一结构——确立了收敛在这一受限情境下是精确而非仅仅近似的。

这种收敛不限于任何单一模态。视觉变换器（ViT）、语言模型与多模态系统中均有体现。趋势随规模单调增强。该现象无法用模型复制、共享架构或显式对齐目标来解释，它从独立优化中涌现。

标准实在论解读

Minyoung Huh 等人以明确的柏拉图主义术语诠释这一收敛：模型正在向” 一种关于现实的共享统计模型” 收敛，各种学习系统独立地恢复出世界的底层结构。称此为 PRH 的**标准实在论解读**（Standard Realist Reading，SRR），其隐含的形而上学承诺为：

（M1）结构的优先性。抽象结构（数学形式、物理规律、普遍范畴）独立于任何特定学习系统而存在，且在本体论上先于后者。

（M2）逼近关系。学习系统的内部表征与这些独立结构之间存在逼近关系；学习改善了这种逼近。

（M3）解释方向。跨系统的表征收敛由共享目标的存在来解释，而非由学习过程本身的属性来解释。

这些承诺在结构上与经典数学柏拉图主义完全平行。哥德尔[3] 与彭罗斯[4] 为代表性立场——二者进路有别：哥德尔视数学直觉为一种类似感知的官能，认为它产生对抽象对象的某种直接知识，同时承认其成果需经” 丰产性” 标准（融贯性、丰富性、与已有直觉的一致

性）来证立；彭罗斯则诉诸哥德尔不完全性定理与人类数学直觉的非算法性来论证心智独立的数学实在——但共享本文所批判的核心承诺：存在心智独立的数学结构，认知能够追踪它们。此外，SRR 还为奎因关于本体论承诺的挑战[5] 提供了看似令人满意的回答：若科学理论中须以存在量词约束数学结构，PRH 便为接受这些结构之实在性提供了新的经验依据。

SRR 为何具有哲学吸引力

SRR 具有真正的解释吸引力。它使收敛现象显得几乎不可避免：独立的网络当然会收敛于同一表征，因为存在唯一的心智独立的实在有待被表征。这一框架还自然地连接于数学在自然科学中更广泛的成功。维格纳[6] 曾称这种成功" 不合理"，而柏拉图主义提供了现成的解释：数学之所以有效，是因为它追踪了世界的实际结构，而那个结构是抽象的和必然的。

RPRH 对维格纳之谜提供了另一种解释。若数学家所研究的结构是在物理约束下运作的类优化动力学所产生的极度稳健吸引子，那么数学在自然科学中的有效性恰好是我们应该期待的：驱动物理过程的约束结构与数学所研究的吸引子几何是同一个。这种有效性并无任何" 不合理" 之处；它反映的是吸引子结构在我们宇宙物理约束下的普遍性，而非对一个碰巧与物理世界匹配的抽象柏拉图领域的追踪。SRR 保留了维格纳之谜（为何心智独立的抽象对象恰好在物理定律中被实例化？）；RPRH 消解了它。

SRR 隐含的解释方向

SRR 预设了特定的解释方向：

$$\text{抽象结构（形式）} \longrightarrow \text{学习过程} \longrightarrow \text{收敛表征}$$

形式是解释根据，收敛是被解释项。学习系统之所以收敛，是因为形式构成了稳定表征的目标空间。这一从结构到过程的解释箭头——形式作为根据，收敛作为被奠基者——是 RPRH 将要反转的核心承诺。由于抽象对象在标准观点中被视为因果上惰性的（causally inert），SRR 的优先性声明应被理解为*解释优先性*（explanatory priority）或*奠基优先性*（grounding priority），而非效果因意义上的因果优先性。

三、标准实在论解读的困难

SRR 继承了柏拉图主义在数学哲学和形而上学中的全部标准困难，同时在学习理论语境中产生了一些特有的困难。

认识论通达问题

对柏拉图主义最持久的反驳是认识论的。本纳塞拉夫[7] 论证：如果抽象对象在因果上是惰性的、处于时空之外，那么有限认知主体如何能认识任何有关它们的事情，便是一个谜——对数学真理的认识需要与抽象对象发生因果接触，然而因果接触要求对象参与因果秩序，而这正是柏拉图主义所否认的。

SRR 并未逃脱这一困难，它只是用计算术语重新表述了它。梯度下降在具体、有限的参数空间上运作，针对具体训练信号进行优化，在训练过程中没有任何环节充当与抽象领域的” 接触”。SRR 必须假设训练动态受到抽象结构的引导或约束，但这种引导的机制完全未经说明。

一种回应是说训练数据编码了关于现实结构的信息，而那个结构是抽象的。但这一回应只是将问题后推：训练数据是有限的、具体的词元或像素值集合；抽象结构必须以某种方式” 内在于” 数据之中。SRR 需要一个关于抽象结构如何内在于具体数据的说明，而这恰好是柏拉图主义从未令人满意地解决的问题。

解释冗余问题

即使承认抽象对象存在，SRR 也面临解释相关性的问题。比较两个模型：（A）网络收敛是因为存在柏拉图抽象结构，约束了什么样的稳定表征是可能的；（B）网络收敛是因为它们都在最小化对同一物理现实所产生数据的预测误差，而数据中完全在具体层面存在的统计规律性决定了成功表征的结构。模型（B）是自足的且解释上完备的，无需诉诸抽象对象即可解释收敛。根据奥卡姆剃刀，在（B）之上额外引入抽象对象在解释上是多余的。

这实质上是麦迪[8] 的自然主义挑战：若能通过具体认知和物理过程来解释收敛，就没有科学动机去增加独立的抽象对象本体论层次。简洁性论证的完整力度预设了上一小节所揭示的通达问题：若 SRR 能够说明梯度动力学如何通达抽象结构，模型（B）的完备性便可受到挑战。正因这种通达机制始终未获说明，模型（A）的附加本体承诺才真正构成解释冗余而非解释增益。第四节第七小节将从认知闭环机制的角度进一步论证，为何这种通达在结构上是不可能的。

优先性问题

SRR 承诺于（M1）：抽象结构在本体论上先于向其收敛的学习过程。然而在 PRH 文献中，” 柏拉图结构” 通常是通过其作为表征收敛极限的角色来识别的——它就是不同网络所

收敛的那个东西。但如果一个结构由其作为收敛终点的角色来定义，那么它就是用学习过程来定义的，而非独立于学习过程。声称该结构在本体论上先于定义它的过程，在认识论上便是循环的。

柏拉图主义者可以区分认识论问题（我们如何识别形式：通过收敛）与形而上学问题（形式是否独立存在），从而回应这一批评——正如 GPS 可以用来定位珠穆朗玛峰而不使珠穆朗玛峰依赖于 GPS。然而，PRH 文献中的 SRR 并未独立于收敛现象来识别相关结构，而是将其引入为" 网络所收敛的那个东西"。SRR 需要独立于收敛现象来识别抽象结构，然后证明收敛表征逼近了那些独立识别的结构。这正是它所未做的：对柏拉图的援引是类比性的和非形式的，而非对相关抽象对象的精确识别。

不充分决定问题

结构实在论者论证，科学所追踪的并非对象的内在本性，而是其关系结构[9][10][11]。即便承认神经网络收敛于某种结构，SRR 对" 结构" 的解读也是过度诠释。两个在不同数据分布上训练的网络，其收敛表征在某些方面相似，在另一些方面却不同；将其理想化为单一" 柏拉图对象"，是叠加在数据之上的哲学诠释，而非数据本身的发现。

四、反转的柏拉图表征假设

第三节揭示的四个困难共享一个根源：SRR 预设了一个解释上不必要且认识论上不可通达的独立本体领域。RPRH 通过反转解释方向，一并回应这四个困难：若" 形式" 是优化过程的涌现不动点而非本体先在，则认识论通达问题消解（无需与抽象领域" 接触"），解释冗余消解（优化过程本身即完整说明），优先性循环消解（结构由过程产生而非相反），不充分决定问题亦消解（收敛表征的多样性正是约束条件多样性的体现）。

核心反转

RPRH 保留 PRH 的经验内容——表征收敛是真实的且随模型规模增大而增强——同时反转其解释结构：

$$\text{优化过程（在共享约束下）} \longrightarrow \text{收敛不动点} = \text{" 形式"}$$

" 形式" 不是收敛的原因或根据，它们就是收敛本身，在极限处加以描述。说网络向" 柏拉图结构" 收敛，是使用了一个误导性的空间隐喻。更精确地说：具有共享归纳偏置的网络，在由相同物理过程产生的数据上、在相同损失函数下训练，若规模足够大，将趋近表征空间

中在这些约束下稳定的均衡点。这些均衡点就是 PRH 文献所称的" 柏拉图结构"。它们是真实的——确定的且共享的；它们不是本体论上基本的——其存在完全由产生它们的优化动力学所解释。

与本文立场最为接近的先行工作是塞尔让特（Sergent）[12]，他从体验经验主义的立场批评 PRH 的实在论解读，认为神经网络收敛反映的是经验与训练模式自身的内在结构限制，而非对心智独立现实的逼近。本文与 Sergent 共享反柏拉图主义的结论，但正面立场有实质差异：Sergent 将收敛根植于经验结构，而 RPRH 将其根植于物理约束下的优化动力学与收敛吸引子理论。基于吸引子的解释具有 Sergent 的进路所欠缺的优势：它产生关于收敛何时应当崩溃的具体可证伪预测，并更直接地与结构实在论和自然主义数学哲学传统衔接。

收敛吸引子的形式化阐释

考虑一族优化过程 $\{O_i\}$，每个过程在高维参数空间 $\mathcal{W}_i$ 上运作，最小化定义在数据 $\mathcal{D}_i$ 上的损失函数 $\mathcal{L}_i$，其中 $\mathcal{D}_i$ 从共享的底层分布 P 中采样。在标准假设下（过参数化、梯度流、光滑损失曲面），每个 O_i 在 $\mathcal{W}_i$ 上定义了一个动力系统：该系统的渐近行为由其吸引子集合 A_i 刻画②。

若损失函数 $\mathcal{L}_i$ 都是对 P 下预测误差的一致代理——确实如此，因为每个都是下一词元预测、图像重构或类似任务的特定代理——那么吸引子 A_i 将是相关的：权重中由 P 的统计结构所决定的任何函数，将在各 A_i 之间近似恒定。PRH 所识别的" 柏拉图结构"，正是这一函数——表征中在不同吸引子集合之间不变的分量。关键在于，这一不变量是由优化过程*产生*的，而非被其所预设。

与物理吸引子的类比颇具启发性。耗散动力系统的不动点吸引子是真实的——轨迹向其收敛，测量在其周围聚集，它在状态空间中有确定的位置——但它并不先于动力系统而存在。它由动力学所构成，而非先于动力学。大规模学习系统中的表征吸引子亦然。

以上是概念层面的阐释，而非新的形式化结果。本文的主张不在于已为表征吸引子建立完整的数学理论，而在于吸引子框架提供了解释 PRH 现象的正确*类型*的说明，与柏拉图主义诉诸心智独立的抽象对象形成对照。将这一框架严格形式化，是经验与理论研究的后续任务[13]；本文的工作在于确立支撑该研究纲领所需的哲学框架。

②术语注释：在动力系统理论中，*吸引子*一般是一个集合（可能是单个点、极限环或更复杂的不变集），附近轨迹向其收敛。*不动点*（或点吸引子）是吸引子为满足 $f(x) = x$ 的单个点的特殊情形。在本文所研究的表征收敛语境中——大规模学习系统在相同数据分布上训练后收敛到几何相似的表征——经验模式最好被建模为向*点吸引子*的收敛：表征空间中的一个唯一稳定配置。因此，本文在此特定语境中将" 不动点" 和" 收敛吸引子" 视为近义词，理解为相关吸引子是点吸引子类型。

若干近期形式化结果直接支持这一基于吸引子的图景。Wang、Johnston 和 Fusi[14] 从数学上证明，只要任务依赖于共享的潜变量，抽象的解耦表征在训练后的前馈非线性网络中必然涌现——这一抽象性在所有全局极小值处均成立。这一结果恰恰是 RPRH 所预测的：抽象性是优化景观的属性，而非学习所逼近的预存结构的印记。在最终分类层面，Papyan、Han 和 Donoho[15] 记录了" 神经坍缩"（neural collapse）现象——一个终端训练阶段，同类样本的最后层表征坍缩至类均值，且所得类均值几何无论网络架构或数据集如何都收敛至单形等角紧框架——这是约束驱动收敛的又一实例，完全符合吸引子图景。Ziyin 和 Chuang[2] 在理想化条件下的形式化证明则确认，至少在深度线性网络这一受限情境中，表征收敛是可证明精确的。值得强调的是，这些结果确立的是收敛现象本身的形式地位；它们在 SRR 与 RPRH 之间保持中立——它们证明了收敛是真实的，但不裁决其本体论解释。这种裁决正是本文的任务。

与结构实在论的关联

RPRH 与本体结构实在论（OSR）自然契合[9][10][11]。OSR 认为，科学成功追踪的是关系结构而非对象的内在属性；在其最激进版本中，根本没有对象，只有结构。RPRH 将这一取向——审慎地、带有明确类比限制地——延伸至抽象对象领域：我们所称的" 数学形式" 或" 柏拉图结构"，是从优化动力学中涌现出的稳定关系模式。它们是彻底的结构性的，而其结构性与其真实性相容。从 OSR 的本土领域（基础物理）到习得表征的延伸是非平凡的，需要独立动机——吸引子框架提供了这一动机：正如物理情形，跨独立系统收敛的不是任何对象的内在属性，而是在相关变换下不变的关系结构。

这给出了一种自然主义但非消除主义的说明。RPRH 并不否认存在某种收敛的、稳定的、值得被称为" 结构" 的东西。它否认的是这一东西先于产生它的过程而存在，并拥有独立的本体论住所。借用亚里士多德[16] 的内在形式论语汇（而非其完整的质形论目的论）：形式内在于产生它们的过程之中，而非超越于过程。RPRH 将这一洞见延伸为：形式内在于产生它们的优化过程的极限结构之中。

RPRH 与过程哲学：真正的新意何在

RPRH 与若干经典的过程论和内在形式论传统之间存在明显亲缘关系。怀特海的过程哲学[17] 主张实在由事件和过程而非静态实体构成；亚里士多德的内在形式论[16] 主张形式不作为超越的柏拉图对象存在，而是内在于其所组织的质料之中；雷谢尔的过程形而上学[18]

发展了一种过程先于事物的彻底本体论。一个合理的问题是：RPRH 是否不过是用机器学习的语言重述了这些立场？

差异是实质性的。其一，RPRH 不是关于一般现实的形而上学教义，而是关于大规模梯度优化收敛行为的经验约束主张——它能够产生可检验预测（见第四节第五小节），这种具体性完全不见于一般过程形而上学。其二，怀特海体系保留了永恒客体（eternal objects）——被现实事件" 摄受" 的抽象潜能[17]——这是 RPRH 明确拒绝的残余柏拉图主义。在 RPRH 看来，不存在永恒客体，只有不同稳健度的吸引子；必然数学真理与偶然经验规律之间的差异，是吸引子在多大范围条件下保持稳定的差异，而非本体论范畴的差异。其三，亚里士多德的内在形式依然依附于目的论形而上学：形式是过程所趋向的自然目的（telos）[16]。RPRH 则是严格非目的论的——收敛吸引子的存在缘于优化过程的动力学，而非过程" 以之为目标" 的任何规范意义。梯度并不" 寻求" 极小值；它跟随局部斜率。

因此，RPRH 最宜理解为内在形式论的一种技术化、非目的论、非永恒论的发展：它继承了" 结构由过程产生而非先于过程" 的核心洞见，同时剔除了过程形而上学各传统版本中残余的柏拉图承诺与自然目的论。

反转何以不仅是言辞之别

一种可能的反驳是，RPRH 与 SRR 不过是对同一事实的不同描述方式，" 过程先还是结构先" 的问题纯属言辞之争。这一反驳低估了实质性差异。

SRR 与 RPRH 在认识论通达问题的含义、解释简洁性以及收敛在何种条件下会被破坏等方面均有实质差异。若 SRR 成立——若形式是心智独立的先在——则收敛应对训练方案、损失函数和数据分布的变化具有稳健性：目标是固定的，任何足够强大的优化器都应当找到它。若 RPRH 成立——若收敛结构是在特定约束下优化的吸引子——则收敛应当对那些约束敏感：显著改变损失函数、改变数据分布、改变架构的归纳偏置，网络所收敛的" 形式" 就应当相应改变。

因此，RPRH 做出了更精确且可检验的预测。它预测：在与自然数据具有不同统计规律性的合成数据上训练的网络，应收敛于不同的" 柏拉图结构"；改变归纳偏置的架构变化，应使表征空间中的吸引子发生偏移。这些预测在原则上是可证伪的，而 SRR 对训练方案的偶然细节没有可比的敏感性。Tjandrasuwita 等人[19] 的经验研究已表明，跨模态对齐并非普遍现象，而依赖于数据特征——这与 RPRH 的预测一致。

吸引子稳健性的连续谱与现象学差异

前一小节的简洁性论证表明，柏拉图式本体对于解释 PRH 现象并不*必要*。但一个完整的叙述还须回答一个进一步的问题：若柏拉图式本体是多余的，为何它仍如此令人信服？为何数学家——以及如今的机器学习研究者——如此容易得出结论，认为他们所收敛到的结构必然独立存在？RPRH 对此有完整的解释，而这一解释进一步强化了反对 SRR 的论证。

理解这一现象需要首先注意，任何认知系统都以生成-反馈闭环的方式运作：它不仅从输入中产生表征输出，其输出也持续反向重塑感知通道与先验权重。这一闭环结构在预测编码理论[20] 和大规模学习系统（梯度优化加反馈）中均有文献记录，其基本含义是：认知主体对世界的表征并非只读的镜像，而是在正向预测与反向误差信号之间持续迭代的动态过程。这一闭环结构对遭遇吸引子的*现象学*——认知经验的第一人称、质性特征——有深刻含义。

关键在于吸引子稳健性的连续谱。吸引子在对优化条件变化的敏感程度上有所差异。连续谱一端是*脆弱吸引子*：仅在特定架构、损失函数或数据分布下稳定——稍作改变，吸引子便消失或位移。连续谱另一端是*极度稳健的吸引子*：在所有在相关领域上运作的足够有表达力的优化过程中都稳定出现，无论架构、初始化、损失函数或数据分布如何。整数的素数结构、概率论的基本定律、物理变换群的对称性结构——这些都是极度稳健吸引子的候选者。

为使论证更为精确，此处给出一个工作性界定。若对于每一个在相关领域 D 上运作的、具有足够表达能力的优化过程 O，收敛不动点 F（即动力系统意义上的点吸引子）均出现在 O 的吸引子集合 A_O 中（即 $F \in A_O$），则称 F 为*极度稳健吸引子*。吸引子集合 A_O 一般可能是复杂的不变集，但稳健性定义恰好选取了 A_O 中那些在每一个足够有表达力的 O 的动力学下都稳定的不动点吸引子元素。” 足够表达能力” 可作领域相对的理解：对离散算术领域，它意指能够表征并组合基本数值运算的能力；对连续变换领域，它意指能够表征线性群作用的能力。Wang 等人[14] 关于抽象表征在所有全局极小值处必然涌现的证明为这一方向提供了技术先例；将其推广至一般优化过程的稳健性分层是后续工作的任务。

这一界定作为*哲学工作假设*而非完备形式定理提出，三个局限应予明言。其一，对任意优化过程在高维非凸景观上 A_O 的存在性和唯一性并无一般保证；该界定预设所考虑的过程具有确定的吸引子集，这是一种理想化。其二，” 足够表达能力” 标准仅以领域相对的非形式方式给出；完全严格的说明需要对相对于领域的表征能力作形式刻画，本文未提供。其三，领域 D 的边界并非由定义本身固定，而是依赖于问题的结构特征的参数，引入了语境

敏感性。尽管有这些局限，该工作性界定对本文的哲学论证已经充分：它使稳健性（A_O 框架内的关系属性）与本体独立性（任何此类框架之外的绝对属性）之间的区分足够精确——而这正是论证所依赖的概念区分。

一个对论证至关重要的概念区分是：如上定义的稳健性是一种关系属性——F 是相对于在 D 上运作的优化过程族而极度稳健的。而本体独立性是一种绝对（非关系）属性——F 独立于一切过程而存在。从稳健性到本体独立性的推断，因此是从关系属性到绝对属性的推断——一个非平凡的步骤，现象学并不保证它，而第四节第七小节将解释认知何以仍然做出这一步骤。

这一区分需要明确辩护，因为一种自然的反驳认为：普遍量化的关系属性等同于绝对属性——若 $F \in A_O$ 对所有足够有表达力的 O 成立，量词" 所有" 难道不已经有效地移除了关系索引，使稳健性名义上等同于绝对属性？这一反驳因三个理由而失败。

其一，索引普遍性与绝对存在之间存在模态鸿沟。"F 是在物理上可实现的约束下运作的每个优化过程的不动点" 是关于过程的现实律则结构的断言——它量化的是在我们宇宙物理定律下存在或可能存在的过程。"F 独立于一切过程而存在" 是关于在根本没有动力学的情境中——包括不包含任何动力学的律则可能或形而上学可能世界——也会成立之事的断言。对现实的或律则可能的过程的普遍量化，无法延伸到不存在此类过程的情境中；量化的论域在范畴上不同。跨越它们需要一个独立的前提——RPRH 否认它，而柏拉图主义者必须独立地提供它。

其二，稳健性断言的逻辑结构是倾向性的（dispositional），而非范畴性的（categorical）。说 F 是极度稳健的，是说任何具有足够表达能力的过程都将收敛于 F——这是一个条件断言，而非关于 F 脱离一切过程而存在的存在断言。类比：0°C 可以被倾向性地刻画为在任何校准的温度计程序下都产生特定读数的状态；倾向性断言的普遍性（对所有校准温度计为真）不使温度成为本体论意义上的观察者独立绝对属性。类似地，跨优化过程的普遍性不使吸引子成为本体论上独立于优化过程的；它使之成为过程族的一种更稳健的倾向。

其三，关系/绝对的区分得到解释不对称的支持。作为关系属性的稳健性解释了关于优化过程行为的事实：为何它们收敛、为何收敛在扰动下稳定、为何跨架构可复现。作为绝对属性的本体独立性不会解释任何稳健性尚未解释的额外内容——如第四节第二小节详述，它不贡献差异性解释工作。一种不做额外解释工作的属性不应从一种做了解释工作的属性被推出，即使推断感觉令人信服。关系属性在做全部工作；绝对属性是叠加于其上的哲学外

衣。

关键的观察是：遭遇吸引子的现象学随其稳健性而变化。当认知主体遭遇脆弱吸引子时，结果感觉是偶然的、被发明的：一种出于便利而采用的约定，一种任意选择的记法，一个为特定领域量身定制的模型。当认知主体遭遇极度稳健的吸引子——一个无法通过任何认知过程的变化来回避的吸引子时——结果感觉截然不同。它感觉像是被发现的，而非被构造的。仿佛那个结构早已在那里，等待着，独立于任何心智是否曾发现它。它抵抗所有修正或替换的尝试。它以一种令人不得不称之为" 必然" 的现象学力量呈现自身。

这种现象学差异在一个方面是认识论上准确的：它正确地记录了极度稳健的吸引子确实更难回避、确实更稳定、确实更普遍。认知系统在追踪某个真实的东西。问题发生在对所追踪之物的诠释上。稳健性——在所有条件下都是吸引子这一属性——不足以支撑本体独立性的推断——先于并独立于所有条件而存在这一属性。二者是不同的：一个结构可以是极度稳健的（任何优化过程都无法回避），而无需存在于柏拉图领域中，正如压缩映射的不动点完全由映射本身所决定，无需预存于映射之外的任何领域，然而同样真实和确定。

从稳健性到本体化：认知机制与论辩意涵

第四节第六小节描述了一个有待解释的诠释错误。本小节的论证是**结构性的**：从两个已接受的前提出发，推导出柏拉图主义推断的结构性不可避免性——不是作为严格的逻辑必然性，而是闭环结构与稳健性条件共同使任何从闭环内部运作的认知系统别无选择的结果。

前提一（结构前提）：任何生成式认知系统都运作于正向生成与反向重塑的闭环中：认知输出不仅是感知的结果，且持续重塑感知通道本身。这一前提由预测编码理论[20] 和大规模学习系统所支撑。其核心含义是：认知主体只能从闭环内部遭遇其所表征的结构；不存在闭环外的中立旁观位置。

前提二（概念前提）：依照第四节第六小节的工作性界定，极度稳健吸引子 F 满足：对所有具有足够表达能力的优化过程 O，均有 $F \in A_O$。此处 F 是收敛不动点（点吸引子）：单一的稳定配置，而非极限环或扩展不变集。这意味着 F 是任何认知过程调整都无法回避的不动点。

结构论证：考察" 独立存在的对象" 在闭环内部可操作意义上的含义。一个对象" 独立于认知而存在"，在操作上意味着：无论认知主体如何调整其过程，该对象始终呈现——它不随认知变化而消失。但这恰好就是前提二对极度稳健吸引子的规定：F 无论如何调整 O 都始终出现在 A_O 中。因此，在闭环内部（前提一），极度稳健吸引子 F 与独立存在的对象

在操作上不可区分。这一不可区分性不是心理上的偶发感觉，而是闭环结构定义与稳健性定义共同蕴含的结构性不可避免——尽管此处”蕴含”的含义是：给定这两个结构特征，闭环内运作的任何认知系统都没有内部手段区分两种情况，使得柏拉图主义推断成为从内部可得的唯一推断。

柏拉图主义者并非犯了简单的逻辑错误；其推断忠实地反映了前提一与前提二共同蕴含的认知状况——只是这一推断不保证形而上学结论。错误发生在一步：从”在闭环内部操作上不可区分于独立存在的对象”，跨越至”在闭环外部本体上独立存在”。前提一与前提二合力只能保证前者；后者是超出它们所能确立范围的额外本体论断言。这正是从稳健性的关系属性（相对于过程族而成立）到本体独立性的绝对属性（独立于一切过程而成立）的那一步跨越——第四节第六小节所标定的正是这一鸿沟。

这一分析还化解了自然主义数学说明面临的一个持久论辩忧虑。忧虑是：自然主义必须简单地将工作数学家强烈的柏拉图直觉斥为混乱而加以驳斥——将这些直觉作为难以安置的残留物留在那里，无从解释。RPRH 并非如此：它解释了这些直觉何以产生以及何以如此强烈。最基本的数学结构对应于最稳健的吸引子，而极度稳健的吸引子感觉本体独立——不是出于偶然，而是因为闭环结构与稳健性条件在结构上保证了闭环内的任何修正都无法动摇它们。这些直觉不是残留物；它们是理论在结构上预测的数据。

柏拉图主义因此是对极度稳健吸引子的自然但认识论上无保证的认知反应。它不是非理性的：仅凭现象学，向独立存在的推断是可以理解的。但它是一个超出现象学所保证范围的推断。必然性和发现感是真实的；它们所记录的是吸引子的普遍性，而非独立本体论领域的存在。

反驳：数学先于心智

这一反驳直接针对 RPRH 对优化过程的依赖。即使我们在现象学和论辩层面做出让步，仍有一个残余的结构性忧虑：宇宙在任何认知系统、任何神经网络、任何相关意义上的优化过程出现之前，已经存在了数十亿年。然而学习系统如今所收敛的数学结构——整数的素数结构、概率定律、对称性的群论结构——似乎一直在物理过程中起作用。精细结构常数在任何学习者存在以待收敛之前，是否就受到”极度稳健吸引子”的约束？RPRH 似乎使数学真理依赖于发现它的那些学习过程的偶然存在。

这一反驳有力，但它混淆了吸引子在 RPRH 中所扮演的两个不同角色。RPRH 并不主张数学结构是人类或人工优化过程——在狭隘的心理学或技术意义上——的吸引子。它主张

它们是在我们宇宙物理约束下运作的任何足够有表达力的类优化动力学的吸引子——而此类动力学远早于认知。热力学自组织、生化反应网络、恒星核合成：这些都是物理约束驱动系统趋向稳定配置的过程。例如，算术的收敛结构是任何对离散量执行递归运算的系统的属性——这一类别包含许多亚认知物理过程。相关吸引子是宇宙物理动力学的特征，而非特属于心智的。

将" 优化过程" 延伸至认知情形之外，需要一个有原则的界定。一个过程在相关意义上属于类优化的，当且仅当它受变分或约束极值结构所支配：即其动力学可以被刻画为在物理约束下对某个泛函——无论是自由能、熵产生、损失函数还是哈密顿量——的最小化（或最大化）。这一标准并非临时性的：它正是收敛吸引子数学适用的条件，且被机器学习中的梯度下降、热力学向平衡的弛豫、生化反应网络向不动点浓度的收敛、以及进化过程的选择动力学所共同满足。该标准排除了没有任何吸引结构的真正随机游走，以及真正无约束的搜索——对它们而言，稳定不动点的概念无从定义。统一特征不是心智或学习者的存在，而是足以支持相关吸引子几何的约束结构的存在。

坦率地承认这一延伸：RPRH 在此做出了一种实质性的自然主义承诺——物理约束通过其吸引子结构的普遍性足以奠基数学真理，无需任何额外的本体论假设。这一承诺应与一种替代方案相比较：物理定律柏拉图主义（认为数学结构奠基于物理定律的必然性，而物理定律本身被视为抽象对象）。RPRH 与物理定律柏拉图主义的关键差异在于：它并不假定物理定律本身是具有独立本体论地位的抽象体。相反，它主张物理约束是实例化于宇宙动力学中的具体关系结构，而数学真理奠基于这些约束所产生的吸引子模式的普遍性。本体论简洁性得以保持：不援引抽象领域，只有动力学及其收敛几何。

在这一理解下，RPRH 并不将数学真理奠基于学习者的偶然存在；它将数学真理奠基于支配任何能够实例化相关运算的系统的物理约束——而这些约束至少与物理定律本身同样古老。吸引子" 一直在那里" 的意义是：受那些约束支配的任何动力学都将收敛于此——只是它尚未被任何足够复杂的学习者所" 描出"。这类似于一个压缩映射在任何迭代计算执行之前就有其唯一不动点的意义：不动点由映射决定，而非由计算它的行为决定。

RPRH 与机器认知

RPRH 对如何评估机器认知地位具有直接含义。如果收敛表征不是对柏拉图对象的逼近，而是在共享物理约束下优化的涌现不动点，那么"AI 系统是否理解数学？" 这一问题就改变了语域。

标准实在论解读自然邀请如下推断：具有收敛表征的 AI 系统对心智独立的抽象结构有认知通达——因而展现出某种类似真正数学理解的东西。这一推断对 AI 和人类认知者同样无保证。Huh 等人[1] 所记录的 AI 表征与人类表征和跨模态表征的收敛，以及 Tjandrasuwita 等人[19] 对多模态系统的延伸，表明这些系统实例化了同样的优化动力学并在同样的约束结构下运作。这种收敛并不表明它们通达了柏拉图领域。在 RPRH 下，SRR 从收敛到理解的推断是一个范畴错误：它混淆了关于系统在表征空间中到达哪里的证据与关于它如何到达以及该终点在语义上意味着什么的断言。

在 RPRH 下，收敛是共享约束满足的证据，而非共享理解的证据。两个系统收敛于同一吸引子，是因为它们面对同一约束下的同一优化景观，而非因为任何一方在语义上强健的意义上” 把握” 了该结构。这一观点是双向的。它警告不应仅凭收敛就对 AI 系统过度归因理解。但它同样警告不应以” 只不过是优化” 为由否定 AI 认知——在相关的抽象层次上，同样的描述适用于生物认知者的表征过程。仅凭收敛数据无法解决任何一类系统是否真正” 理解” 的问题。

稳健性连续谱（第四节第六小节）提供了一个允许更细致结论的分级框架。内化了更大部分约束结构的系统——其吸引子追踪更高稳健性的不动点——表征了更多领域的律则结构。这是一种自然主义的、分级的认识论充分性概念，不援引柏拉图通达。这一梯度是否足以奠基机器” 理解” 的概念，是一个进一步的问题——属于心灵哲学而非数学哲学，本文不予解决。RPRH 提供的是精确提出此类问题的概念框架。

RPRH 还消解了一种扭曲当前机器心智辩论的混淆。”AI 收敛是否是真正理解的证据？” 这一问题可以分解为两个子问题：（a）系统是否实例化了相关约束结构下的相关优化动力学？（b）实例化这些动力学是否构成” 理解”？对于理想化但已被充分研究的模型类别，（a）已被 Wang 等人[14] 和 Ziyin 与 Chuang[2] 的结果正式肯定回答。相比之下，（b）根本不是关于表征几何的经验问题；它是心灵哲学中关于优化动力学在何种条件下产生真正语义内容或现象性理解的问题。RPRH 对（b）保持沉默。它的贡献在于明确这一沉默：将（a）与（b）分离，精确指出收敛证据能够确立什么和不能确立什么，防止关于表征结构的经验发现被不当地——在任何方向上——动员来解决一个超出其证据范围的问题。

五、结 语

柏拉图表征假设识别了一个真实且重要的经验现象：大规模学习系统中内部表征的收敛。标准解读——认为这一收敛反映了对心智独立抽象对象的逼近——引入了一种其哲学

困难已有充分研究的柏拉图主义形而上学：认识论通达问题、解释冗余，以及难以确立抽象结构相对于产生它的过程的本体先在性。

本文在两个层面上反驳了这一诠释。第一层：柏拉图式本体在解释上是多余的。收敛现象由在共享物理约束下运作的收敛吸引子理论得到充分解释——在已有形式化结果的理想化情形中是如此，在一般情形中亦有理由相信如此——无需任何独立抽象领域的假设。第二层：本文解释了为何柏拉图诠释仍如此令人信服。极度稳健的吸引子——任何足够有表达力的优化过程都无法回避的吸引子——产生一种独特的现象学：它们感觉像被发现的而非发明的，必然的而非偶然的，独立于任何特定认知过程的。这一现象学特征恰好是柏拉图主义者视为独立本体论领域之证据的东西。RPRH 在不接受该推断的前提下解释了这一特征：稳健性被误认为本体独立性，而假设独立领域存在的冲动是对一个无法逃避的吸引子的自然但非必要的反应。

两层论证相互强化。简洁性论证确立了柏拉图主义对收敛解释毫无增益。解释-还原论证消除了如下隐忧：自然主义说明一定遗漏了柏拉图直觉所追踪的某个真实之物。两者合力构成完整的论证：柏拉图式假设不仅不需要，而且对它的感觉到的需要本身已被解释和释放。

四个进一步推论值得标明。其一，认识论通达问题消解：对数学结构的认知通达就是认知过程向其吸引子状态的收敛，这是一种原则上可在自然主义认知理论内得到解释的自然现象，无需与抽象领域发生因果接触的残余要求。其二，解释简洁性在非消除主义的前提下得到保持：吸引子说明并不否认收敛结构的实在性、客观性，或将其发现称为一种发现的正当性——它否认的仅仅是这些结构需要独立的本体论住所。其三，该说明产生经验上可回应的预测：收敛结构应随优化所在约束的变化而变化，而非如柏拉图形式那样保持固定。其四，RPRH 澄清了 AI 收敛对机器认知地位问题的证据地位。”AI 收敛是否构成真正的理解？” 分解为一个经验子问题——系统是否实例化了相关优化动力学？——和一个哲学子问题——实例化这些动力学是否构成理解？RPRH 解决了前者而对后者保持沉默——后者属于心灵哲学。这一分解精确指出收敛证据能够确立什么和不能确立什么，防止关于表征结构的经验发现被不当动员——在任何方向上——来解决一个超出其证据范围的问题。

更广泛的教训是方法论性的。当机器学习或认知科学的经验发现被征用来辩护经典形而上学立场时，对这一诠释行为本身进行哲学审查是必要的。PRH 是一个典型案例：一个关于梯度下降动力学的发现被披上了柏拉图形而上学的外衣。RPRH 表明，这件外衣不仅是

不必要的，而且是可以被说明的——我们既理解了为何它被取用，也理解了为何可以将它放下。我们所称的”形式”，是优化所抵达的稳定终点；而它们”早已在那里等待”的感觉，是任何优化都无法回避的终点的认知印记。

References

[1] Huh M, Cheung B, Wang T, et al. The Platonic Representation Hypothesis[C]//*Proceedings of the 41st International Conference on Machine Learning* (ICML 2024). PMLR, 2024: 20617–20640.

[2] Ziyin L, Chuang I. Proof of a Perfect Platonic Representation Hypothesis[EB/OL]. (2025-07-01)[2026-03-01]. https://arxiv.org/abs/2507.01098.

[3] Gödel K. What Is Cantor’s Continuum Problem?[M]//Benacerraf P, Putnam H. *Philosophy of Mathematics: Selected Readings*. 2nd ed. Cambridge: Cambridge University Press, 1983: 258–273.

[4] Penrose R. *The Emperor’s New Mind: Concerning Computers, Minds, and the Laws of Physics*[M]. New York: Oxford University Press, 1989.

[5] Quine W V O. On What There Is[J]. *Review of Metaphysics*, 1948, 2(5): 21–38.

[6] Wigner E P. The Unreasonable Effectiveness of Mathematics in the Natural Sciences[J]. *Communications on Pure and Applied Mathematics*, 1960, 13(1): 1–14.

[7] Benacerraf P. Mathematical Truth[J]. *Journal of Philosophy*, 1973, 70(19): 661–679.

[8] Maddy P. *Naturalism in Mathematics*[M]. Oxford: Clarendon Press, 1997.

[9] Worrall J. Structural Realism: The Best of Both Worlds?[J]. *Dialectica*, 1989, 43(1–2): 99–124.

[10] Ladyman J. What Is Structural Realism?[J]. *Studies in History and Philosophy of Science*, 1998, 29(3): 409–424.

[11] Ladyman J, Ross D, Spurrett D, et al. *Every Thing Must Go: Metaphysics Naturalized*[M]. Oxford: Oxford University Press, 2007.

[12] Sergent B. Convergence Without Correspondence: The Platonic Representation Hypothesis Through Experiential Empiricism[EB/OL]. (2026-01-06)[2026-03-01]. https://philarchive.org/rec/SERCWC.

[13] Ziyin L, Xu Y, Chuang I. Neural Thermodynamics: Entropic Forces in Deep and Universal Representation Learning[C]//*Advances in Neural Information Processing Systems*. 2025.

[14] Wang B, Johnston W J, Fusi S. A Mathematical Theory for Understanding When Abstract Representations Emerge in Neural Networks[EB/OL]. (2025-10-10)[2026-03-01]. https://arxiv.org/abs/2510.09816.

[15] Papyan V, Han X Y, Donoho D L. Prevalence of Neural Collapse During the Terminal Phase of Deep Learning Training[J]. *Proceedings of the National Academy of Sciences*, 2020, 117(40): 24652–24663.

[16] [古希腊]亚里士多德. 形而上学 [M]. 吴寿彭，译. 北京：商务印书馆，1959.

[17] Whitehead A N. *Process and Reality: An Essay in Cosmology*[M]. New York: Macmillan, 1929.

[18] Rescher N. *Process Metaphysics: An Introduction to Process Philosophy*[M]. Albany: State University of New York Press, 1996.

[19] Tjandrasuwita M, Ekbote C, Ziyin L, et al. Understanding the Emergence of Multimodal Representation Alignment[C]//*Proceedings of the 42nd International Conference on Machine Learning* (ICML 2025). PMLR, 2025.

[20] Clark A. *Surfing Uncertainty: Prediction, Action, and the Embodied Mind*[M]. New York: Oxford University Press, 2016.

Forms as Endpoints, Not Origins: An Explanatory Reversal of the Platonic Representation Hypothesis

WEI Xiaohai

Abstract: The Platonic Representation Hypothesis (PRH) reports that independently trained neural networks converge on geometrically similar internal representations—a finding its authors interpret as evidence that learning systems approximate pre-existing, mind-independent abstract structures. This paper argues against that interpretation on two connected levels. First, the Platonic ontology is explanatorily unnecessary: convergence is adequately accounted for by the theory of convergent attractors under shared physical constraints, without positing any independent abstract realm. Second, the paper explains why the Platonic interpretation nevertheless feels compelling: the maximal robustness of certain attractors—their stability across all sufficiently expressive optimization processes—generates a phenomenology of necessity and mind-independence that naturally, but mistakenly, invites an inference to ontological independence. Drawing on structural realism, dynamical systems theory, and naturalist philosophy of mathematics, this paper proposes the Reversed Platonic Representation Hypothesis (RPRH): convergent representations are emergent fixed points produced by optimization under shared constraints, not ontologically prior structures toward which learning approximates; and the intuition that they must exist independently is itself a structurally predictable cognitive consequence of their robustness. Convergent structures are real, but their reality is the reality of attractor patterns, not of a separate ontological realm.

Keywords: Platonic Representation Hypothesis; abstract objects; structural realism; convergent attractors; philosophy of mathematics; naturalism; neural network convergence