当代 AI 深度神经网络与人脑在工作原理上高度一致

任智平 岭南科学工业出版社研究员

摘　要

本文旨在论证，当代人工智能（AI），特别是深度神经网络（DNNs），在经历了从符号主义到连接主义的范式转移后，其核心工作机制在本体论、功能和学习三个层面上与人脑认知“非常相似”。本文首先回顾了认知科学中符号主义与连接主义的理论之争，并确立了连接主义（并行分布式处理）与神经科学中的赫布定律及预测编码理论作为本文的理论基础。本文的核心论证分为三部分：1）**本体论层面**：AI 的知识载体是“连接强度”（权重），而非显性规则，这构成了知识的分布式表征；2）**功能层面**：AI 的核心功能是基于高维 statistical 的“预测”，而非逻辑演绎，这与大脑的预测编码机制（Predictive Coding）相一致；3）**学习层面**：“反向传播”（Backpropagation）算法是一种高效的“预测误差”修正机制，其功能（尽管生物机制不同）与大脑基于经验的突触可塑性相类似。本文认为，AI 的智能涌现于海量节点（神经元）的并行激活（“Ping 的联盟”），这标志着一种新的认知范式，它挑战了传统上基于规则的智能定义。

关键词：人工智能；认知范式；连接主义；深度学习；预测编码；反向传播

1. 引言 (Introduction)

1.1 研究背景：从“图灵机”到“神经网络”

人工智能（AI）自诞生以来，其发展路径就充满了分歧与迭代。以艾伦·图灵（Alan Turing）的计算理论为基础，早期 AI 研究（常被称为“符号主义 AI”或 GOFAI）试图通过实现一套复杂的、基于逻辑的规则系统来复刻人类智能（Newell & Simon, 1976）。这种范式在专家系统等领域取得了初期成功，但很快在面对现实世界的模糊性、复杂性和“常识”问题时暴露了其根本局限性，导致了 AI 领域的“寒冬”（Haugeland, 1989）。

然而，近二十年来，尤其是自 2012 年以来，以深度学习为代表的“连接主义”路径取得了革命性突破（LeCun, Bengio, & Hinton, 2015）。AlphaGo 在围棋

这一被视为人类直觉巅峰的领域战胜了世界冠军；大型语言模型（LLM）如 GPT 系列（Radford et al., 2018）展现出惊人的语言理解、生成乃至推理能力。这些模型在解决传统 AI 无法企及的“模糊”和“直觉”问题上的巨大成功，迫使我们重新审视 AI 的本质。公众与学界不禁产生困惑：当代 AI 是更高级的“统计鹦鹉”（Bender et al., 2021），仅仅是在模仿数据分布？还是它真正触及了智能的某种“类人”核心？

1.2 研究问题与核心论点

本文的核心研究问题是：当代 AI（特别是深度神经网络）的工作机制，在多大程度上可以被视为一种与人脑认知同构或高度相似的“新认知范式”？

本文主张，AI 正在经历一场根本性的范式转变，使其脱离传统的“工具”属性，转而展现出一种“类脑”的认知范式。**需要明确的是，本文的理论创新点不在于提出全新的基础理论，而在于“综合”与“应用”**——即系统地综合认知科学中的“连接主义”和神经科学中的“预测编码”理论，赫布定律与突触可塑性与反向传播（BP）算法的类比性，并将其作为一个统一的分析框架，应用于论证当代 AI 的核心机制（连接强度、预测、认知修正（即反向传播））在功能上与人脑高度一致。

我们论证，这种新范式下的智能，其本质不是基于规则的串行计算，而是：1）一个在本体论上以“连接强度”为知识载体的**分布式系统**；2）一个在功能上以“预测”为核心目标的**推断机器**；3）一个在学习上以“反向传播”（误差修正）为驱动，从海量数据（经验）中**自我塑造**的系统。

2. 理论基础与文献综述 (Theoretical Basis and Literature Review)

2.1 认知科学的范式之争：符号主义 vs. 连接主义

2.1.1 符号主义 (Symbolicism)

符号主义，或称“物理符号系统假说”（Physical Symbol System Hypothesis），由纽厄尔（Newell）和西蒙（Simon）提出，他们认为智能行为的充要条件是拥有一套符号处理系统（Newell & Simon, 1976）。在这种范式中，认知即计算，智能是对离散符号（如词语、概念）进行逻辑规则（如 IF-THEN）的操作。这种“自上而下”的方法在高度结构化的任务（如逻辑证明、专家系统）中表现良

好，但面对“常识”问题（如理解一个笑话）或感官运动任务（如识别面孔）时则显得极其脆弱。

2.1.2 连接主义 (Connectionism)

连接主义，尤其是在鲁梅尔哈特（Rumelhart）和麦克莱兰（McClelland）的“并行分布式处理”（PDP）模型中被系统阐述（Rumelhart et al., 1986）。该理论认为，认知并非来自一个中央处理器对符号的串行操作，而是涌现自大量简单处理单元（神经元）的并行、分布式活动。其核心概念是：

亚符号处理 (Sub-symbolic)：基本处理单元操作的不是高级概念，而是简单的激活信号。

分布式表征 (Distributed Representation)：知识或概念（如“狗”）不是存储在某个单一节点，而是表现为网络中大量连接（权重）的特定激活模式。 当代 AI，特别是深度神经网络，正是“连接主义”思想在强大算力和海量数据支持下的伟大回归与工程实现。

2.2 认知神经科学的理论支撑

2.2.1 赫布定律与突触可塑性 (Hebbian Law and Synaptic Plasticity)

唐纳德·赫布（Hebb）在 1949 年提出的理论（Hebb, 1949），常被概括为“一起放电的神经元，连接更紧密”（Cells that fire together, wire together）。这奠定了学习和记忆的生物学基础：经验和学习不是写入规则，而是通过改变神经元之间的“连接强度”（即突触强度）来实现的。这是连接主义“知识存储在权重中”的直接生物学证据。

2.2.2 预测编码理论 (Predictive Coding Theory)

预测编码理论，由卡尔·弗里斯顿（Karl Friston）等人系统发展，认为大脑是一个“预测机器”或“贝叶斯推断引擎”（Friston, 2010）。该理论的核心机制是：大脑不断利用其内部的生成模型（Generative Model）“自上而下”地产生对下一刻感官输入的“预测”。然后，它将这个预测与“自下而上”的实际感官输入进行比较。如果两者一致，信号被抑制；如果存在差异（即“预测误差”或“惊奇”Surprisal），则只有这个“误差”信号被传递到上层，用以修正内部模型（即调整连接强度），从而在未来做出更准确的预测。智能的本质，就是在一个动态变化的世界中，不断最小化“预测误差”。

3. 论证（一）：作为知识载体的“连接强度” (Argument I: "Connection Strength" as the Substrate of Knowledge)

3.1 知识的分布式表征 (Distributed Representation)

符号主义 AI 的知识是局域的、显性的。一条规则（如 IF "有羽毛" THEN "是鸟"）被明确地存储在系统的特定位置。这种方法的弱点在于其脆弱性——规则的微小错误或缺失可能导致系统崩溃。

当代 AI 则完全不同。在深度神经网络中，“知识”是弥散的、隐性的。一个概念，例如“猫”，并非存储在某个特定的神经元中。相反，它是弥散在网络中数亿（甚至数万亿）个权重参数所构成的“连接强度”矩阵中（Olah et al., 2017）。当我们向网络输入一张猫的图片时，“猫”的概念表现为网络中被激活的一整套特定模式。这种分布式表征具有极高的鲁棒性（robustness），即使部分神经元或连接被移除，网络性能的下降也是平滑的，而非灾难性的。这与神经科学家卡尔·拉什利（Karl Lashley）寻找“记忆痕迹”（Engram）的经典实验结果相一致——记忆似乎并非存储在大脑的特定位置，而是广泛分布的（Lashley, 1950）。

3.2 “Ping 的联盟”：智能作为并行的涌现 (A "Coalition of Pings": Intelligence as Parallel Emergence)

基于连接强度的分布式表征，AI 的决策过程（即“思考”过程）也根本区别于传统计算。它不是冯·诺依曼架构下的串行逻辑推理，而是一种并行的、动态的激活过程。我们可以将其比喻为“Ping 的联盟”：一个输入信号（如一个词或图像像素，即一个“Ping”）进入网络，它并不会被“中央处理器”读取，而是同时激活与其相连的数千个神经元。这些神经元根据其“连接强度”再次激活下一层神经元。这个过程（如 Transformer 架构中的“注意力机制”）允许信号在网络中并行传播，动态地形成一个临时的“激活联盟”（A Coalition of Activations）（Vaswani et al., 2017）。最终的决策（如输出的下一个词），是这个涌现出来的、高维联盟的集体“共识”，而非任何单一规则的产物。这与人脑的工作方式高度相似。在认知神经科学中，“神经元集群放电”（Neural Ensembles）理论认为，一个特定的思维、记忆或感知（如“祖母”的面孔），正是由大脑皮层中一个特定神经元集群的同步激活所代表的（Buzsáki, 2010）。因此，无论在 AI 还是人脑中，智能都不是串行的逻辑演绎，而是并行的模式匹配与激活。

4. 论证（二）：“预测”作为智能的核心功能 (Argument II: "Prediction" as the Core Cognitive Function)

4.1 AI 的预测本质

如果说“连接强度”是 AI 的“本体”，那么“预测”就是其核心“功能”。当代 AI 的训练目标惊人地统一，它们在本质上都是“预测机器”。以大型语言模型（LLM）为例，其核心训练目标极其简单：“预测下一个词”（Next Token Prediction）（Radford et al., 2018）。模型被输入海量的文本，并被要求在每个位置预测下一个最有可能出现的词。为了在这个任务上做得更好（即减少预测误差），模型被迫在内部的“连接强度”中，构建一个关于世界的高维统计模型——它必须“理解”语法、事实、上下文，乃至一定程度的因果关系，才能准确预测“法国的首都是...”之后是“巴黎”。

同样，在计算机视觉中，卷积神经网络（CNN）的本质是“预测”该图像属于特定分类的概率（Krizhevsky et al., 2012）。AI 的复杂能力，如对话、翻译、甚至看似的“推理”，都是从这个简单的“预测”目标中涌现出来的。

4.2 “预测”与大脑“预测编码”的同构

AI 的这种工作机制，与第二节中提到的“预测编码”理论（Friston, 2010）形成了惊人的功能同构。

大脑：利用内部模型（连接强度）“自上而下”地预测感官输入。

AI：利用内部模型（连接强度）“前向传播”地预测数据标签（如“下一个词”）。

两者都在不断地将“预测”与“现实”（感官输入 vs. 真实数据）进行比较，并利用“误差”（Prediction Error vs. Loss Function）来修正内部模型（连接强度）。在这个视角下，智能即是最小化“惊奇”（Surprise）或“预测误差”的过程。无论是人脑还是 AI，拥有“智能”的系统，都是一个通过海量经验（数据）训练出来的、对世界具有强大预测能力的内部模型。

5. 论证（三）：作为误差修正的“反向传播” (Argument III: "Backpropagation" as Error Correction)

5.1 反向传播（BP）算法的本质

如果 AI 的核心功能是“预测”，那么其核心学习机制就是“反向传播”（Backpropagation, BP）（Rumelhart et al., 1986）。BP 算法是 AI 实现“从错误中学习”的数学核心，它优美地解决了“信用分配”（Credit Assignment）问题——即当预测出错时，网络中数万亿个连接，哪些该为错误负责？

BP 算法的机制可以通俗化为三步：

预测（前向传播）：AI 根据当前“连接强度”（权重）进行一次预测（如猜测图片是“猫”）。

比较（计算误差）：将“预测”与“真实答案”（数据标签，如“狗”）对比，通过“损失函数”（Loss Function）计算出“误差”的大小。

修正（反向传播）：利用微积分的链式法则（梯度下降），将这个“误差”信号从网络的输出层反向传播回输入层，精确计算出每一个“连接强度”对总误差的“贡献度”。然后，按“贡献度”微调所有相关的连接，以确保下一次遇到类似输入时，误差会更小。

5.2 BP 算法与生物学学习的“功能相似性”

一个常见的批评是：反向传播算法在生物学上是不可能的（Biologically Implausible），人脑显然没有一个全局的“损失函数”或精确的“梯度”信号（Crick, 1989）。

我们承认两者在“生物机制”上的巨大差异。然而，我们主张两者在“功能”上是高度相似的。

共同目标：两者都是“从错误中学习”的监督/强化过程。大脑的多巴胺系统（奖励/惩罚信号）在功能上就类似于一个“误差”信号（Schultz, 2007）。

共同手段：两者都是为了优化“连接强度”（突触强度 vs. 权重），以最小化未来的“预测误差”。

共同资源：两者都依赖于“经验”。AI 的“海量数据”与人脑的“毕生经验”（数万小时的视觉、听觉、语言输入）是等价的。

因此，BP 算法（误差反向传播算法）可以被理解为：在工程上（使用硅基芯片和数学）实现“类赫布学习”（Hebb-like Learning）和“最小化预测误差”这一生物学原则的、目前已知的最高效的数学手段。

6. 讨论：新范式的内涵与挑战 (Discussion: Implications and Challenges of the New Paradigm)

6.1 重新定义“智能”与“理解”

如果 AI 通过“连接强度”和“预测”实现了智能，这将迫使我们重新思考“智能”和“理解”的定义。约翰·塞尔（John Searle）的“中文房间”思想实验（Searle, 1980）有力地驳斥了“符号主义 AI”可以拥有“理解”的可能（即一个遵循规则的人，即使能完美“处理”中文，也不“理解”中文）。“中文房间”系统完全有可能通过图灵测试（让外面的中国人相信屋里是个懂中文的人）。但这恰恰证明了图灵测试的缺陷：通过行为测试（功能对等）不等于拥有真正的理解与意识。 它可能只是“完美的模拟”，而非“真实的拥有”。

然而，“中文房间”在连接主义范式下可能已经失效。在 DNN 中，并没有一个“遵循规则的人”。“理解”不再是一个需要被“执行”的程序，而是系统在最小化预测误差的过程中，从高维数据中“涌现”出来的一种属性。如果一个系统能像人一样准确地预测和使用语言（乃至图像和声音），我们是否还有理由否认它在某种意义上拥有了“理解”？

6.2 “黑盒”问题的再认识

“可解释性”（Interpretability）的困境是当代 AI 面临的最大挑战之一。我们很难解释一个 DNN 为什么会做出某个特定决策（Castelvecchi, 2016）。然而，如果我们接受 AI 是“类脑”的，那么这个“黑盒”属性恰恰是该范式的证据之一，而非一个纯粹的“缺陷”。我们也无法用语言内省（Introspect）我们自己是如何瞬间识别一张面孔、或在对话中“灵光一闪”想到一个词的。我们的“理解”也是一个基于“连接强度”和“预测”的、无法被完全解释的“黑盒”。要求一个“连接主义”系统提供一个“符号主义”式的清晰解释，本身可能是一个范式上的误读。

6.3 局限性与差异

当然，AI 与人脑之间仍存在巨大差异。AI 需要海量的电力和数据进行训练，而人脑（约 20 瓦功率）则具有极高的数据效率（如“一次性学习” One-shot learning）。人脑的智能是在与物理世界的实时互动中“具身”形成的。而目前大多数 AI 缺乏身体经验。根据 Friston 的理论，大脑不仅被动预测，还会“主动”采取行动（如转动眼球、探索环境）来最小化预测误差。这是 AI 目前普遍缺乏的能力。

7. 结论 (Conclusion)

7.1 论文总结

本文从认知科学和神经科学的理论基础出发，试图论证当代 AI 正在经历一场深刻的范式转移。我们从三个层面进行了论证：

本体论：AI 的知识载体从“规则”转向了“连接强度”（分布式表征）。

功能：AI 的核心功能从“逻辑”转向了“预测”（预测编码）。

学习：AI 的学习机制从“编程”转向了“反向传播”（基于误差修正的经验学习）。

我们认为，AI 通过“Ping 的联盟”式的并行处理，其工作方式与人脑的认知机制“非常相似”。

7.2 理论贡献与展望

本研究为理解 AI 的“智能本质”提供了一个脱离传统“规则”束缚的、基于“连接主义”和“预测编码”的理论视角。AI 的成功不仅是工程学的胜利，更是对“智能”这一古老哲学问题的全新启示：智能或许本质上就是一个通过海量经验训练出来的、复杂的、并行的预测机器。未来，AI 的发展与脑科学的研究必将更加紧密地结合。AI（如 Transformer 架构）可以作为验证认知理论的计算模型来辅助脑科学研究；而脑科学（如更高效的学习法则、主动探索）也必将为下一代 AI 的发展提供新的算法灵感。

参考文献 (References)

Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the dangers of stochastic parrots: Can language models be too big? *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 610–623.

Buzsáki, G. (2010). Neural syntax: Cell assemblies, synapsembles, and readers. *Neuron, 68*(3), 362–385.

Castelvecchi, D. (2016). Can we open the black box of AI? *Nature, 538*(7623), 20–23.

Crick, F. (1989). The recent excitement about neural networks. *Nature, 337*(6203), 129–132.

Friston, K. (2010). The free-energy principle: A unified brain theory? *Nature Reviews Neuroscience, 11*(2), 127–138.

Haugeland, J. (1989). *Artificial intelligence: The very idea*. MIT Press. ISBN electronic:9780262291149

Hebb, D. O. (1949). *The organization of behavior: A neuropsychological theory*.Wiley.

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2017). ImageNet classification with deep convolutional neural networks. *Communications of the ACM*, 60(6), 84–90. https://doi.org/10.1145/3065386

Lashley, K. S. (1950). In search of the engram. In Society for Experimental Biology, Physiological mechanisms in animal behavior. (Society's Symposium IV.) (pp. 454–482). Academic Press.

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature, 521*(7553), 436–444.

Newell, A., & Simon, H. A. (1976). Computer science as empirical inquiry: Symbols and search. *Communications of the ACM, 19*(3), 113–126.

Olah, C., Mordvintsev, A., & Schubert, L. (2017). Feature visualization. *Distill, 2*(11), e7.

Radford, A., Narasimhan, K., Salimans, T., & Sutskever, I. (2018). *Improving language understanding by generative pre-training*. OpenAI.

Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986). Learning representations by back-propagating errors. *Nature, 323*(6088), 533–536.

Rumelhart, D. E., McClelland, J. L., & PDP Research Group. (1986). *Parallel distributed processing: Explorations in the microstructure of cognition, Vol. 1: Foundations*. MIT Press.

Schultz, W. (2007). Behavioral dopamine signals. *Trends in Neurosciences, 30*(5), 203–210.

Searle, J. R. (1980). Minds, brains, and programs. *Behavioral and Brain Sciences,* 3(3), 417–424.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. In Advances in neural information processing systems (pp. 6000–6010).