我和辛顿一起发明了复杂神经网络，但它现在需要升级｜Hao好聊 X AI先驱谢诺夫斯基

原创 博阳 腾讯科技 2025年12月12日 18:11 北京

文｜博阳

编辑｜徐青阳

1984年的一天，物理学家特伦斯·谢诺夫斯基和心理学家杰弗里·辛顿坐在实验室里，盯着黑板上的方程发呆。那是AI的第二个寒冬，神经网络陷入僵局。人们都知道多层网络更强大，但没人知道怎么训练它。

“如果我们把神经网络想象成一团气体呢？”谢诺夫斯基突然说。

这个疯狂的想法最终变成了玻尔兹曼机，这是一个用统计物理学重新定义"学习"的数学模型。它证明了只要找到合适的能量函数，神经网络就能像气体从高温降到低温一样，自发地调整到最优状态。

这成为现代深度学习的理论基石之一。

但两人后续的志趣却互相有所偏离。辛顿发现了更实用的反向传播算法，带领深度学习走出寒冬，最终迎来ChatGPT主导的AI时代。而谢诺夫斯基选择了回到神经科学实验室，用几十年时间解剖大脑的每一个回路，试图回答那个最初的问题：大脑究竟是如何工作的？

40年后，辛顿因玻尔兹曼机获得2024年诺贝尔物理学奖。他在颁奖典礼上自嘲：“我和特里原本以为我们会因为解释大脑而获奖，结果我们错了。但这对物理学奖来说也够格了。”

而83岁的谢诺夫斯基，依然在实验室里追问那个问题。

也许没有人比他更适合回答今天AI缺失的那些碎片。他见证了神经网络从“异端”到“改变世界”的全过程；他既懂物理学的简洁优雅，也懂生物学的复杂混沌；他和辛顿一起打开了AI的大门，又眼看着这扇门后的世界变得越来越陌生。

在他看来，现在ChatGPT能通过哈佛医学院的考试，却连“睡觉”都不会。

是的，睡觉。人类大脑在睡眠时会激活海马体整理记忆，会通过做梦演练各种可能性。ChatGPT一旦停止输入就彻底沉默，它没有海马体，没有基底神经节，甚至没有“自主生成的思想”。

而Transformer虽然强大，**但只模拟了大脑皮层的一小部分功能，缺失了绝大多数关键结构。**

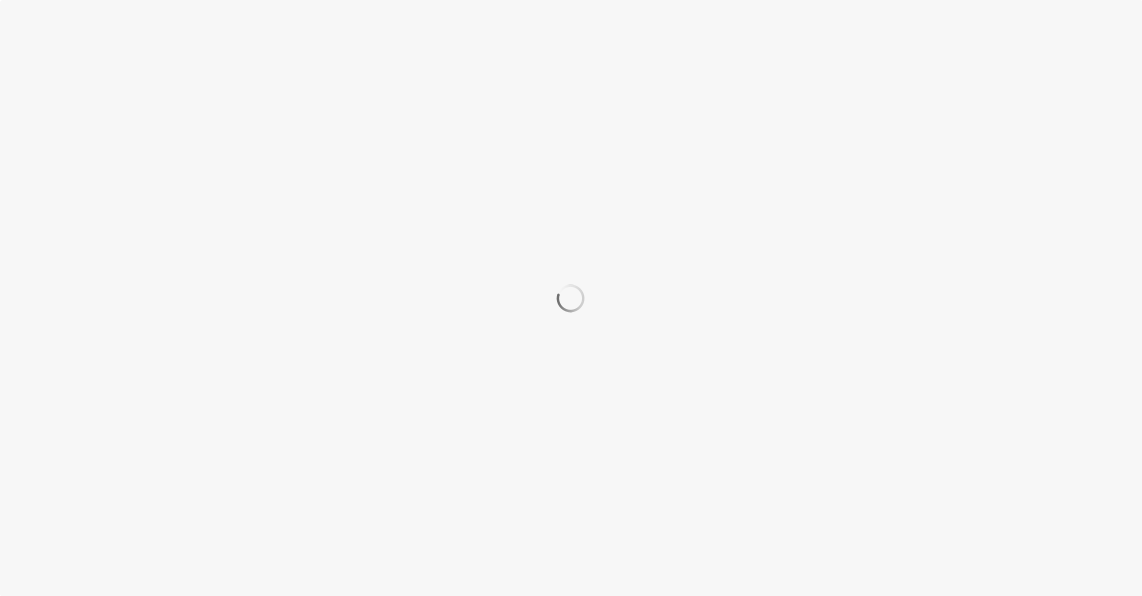
它是一个智能碎片，能思考，但不会行动；能对话，但不会生存。

在这次对话中，我们从历史、当下聊到未来，从这位AI先驱那里听到了关于智能本质的另一种答案：**Transformer可能不是通往AGI的唯一路径。**

这是一场关于物理学、神经科学与AI的深度对话，也是一次对"智能"本质的重新审视。

在这个AI狂飙突进的时代，谢诺夫斯基提醒我们：**技术的进步可以很快，但对智能的真正理解，可能需要几代人的耐心。**

| 本节目也有音频 |



你可以通过小宇宙平台收听，欢迎关注我们。

以下是《Hao好聊》与谢诺夫斯基的对话，经腾讯科技精编：

01

AI历史的起点，是把神经网络想象成一团气体

郝博阳： 计算神经科学是现代AI的基石。在那个年代，您和辛顿是如何利用物理学思维，突破当时传统AI的局限，找到解释大脑核心逻辑的钥匙的？

谢诺夫斯基： 当我们开始职业生涯时，神经网络实际上还不是一个非常成熟的领域。

有一些先驱者，比如日本的斯里尼·甘姆里（Srini Amari），美国的杰克·科万（Jack Cowan），欧洲的克里斯托弗·冯德马尔斯堡（Christoph von der Malsburg）。所以只有少数几个人，我们对如何理解大脑有着相似的直觉。

至少我和杰弗里的方法是——**大脑极其复杂，从神经元组合在一起的复杂性以及它们如何通过突触相互通信的角度来看，其复杂度几乎是无限的。**

但我们想要提取一般原理。

这些原理与数字计算机非常不同。数字计算机有一个与处理单元分离的内存，然后有一系列指令按顺序执行。这就是冯·诺依曼架构。但大脑完全不同。它是大规模并行的，有很多处理单元（神经元）高度互连。

当时的AI缺少一个部分，就是对于大脑而言很重要的学习能力。当我们来到这个世界时，我们必须学习我们将要说的语言、文化以及世界如何运作。很早我们就会做这个事情，而且会在一生中持续做下去。

所以我们意识到，AI需要一个学习算法。

当时存在一个僵局，因为对于单层网络（称为感知器）有一个很好的学习算法，我们想将其推广到多层网络。这就是玻尔兹曼机的用武之地。

我们将神经网络想象成一团气体。在物理学中，气体分子通过随机运动最终会达到一个能量最低的热平衡状态。我们想，如果把‘学习’定义为‘寻找能量最低状态’的过程是否有效？

于是我们发现了一个简单的局部规则。让网络在“看到数据时”达到一个平衡，再让它在“没看数据（自由运行）时”达到另一个平衡。通过比较这两个状态下神经元活动的差异（相减），就能算出权重的调整方向。

这就是玻尔兹曼学习算法。我们证明了这个原理。

当然，当时的网络按今天的标准来说非常小，但这是核心。

霍普菲尔德网络、玻尔兹曼机以及其他许多网络在那时正在起飞。所以那是一个非常激动人心的时期。

02

AI的核心规律不是计算，而是学习

郝博阳： 您是否认为每一个大脑活动最终都可以简化为计算，比如玻尔兹曼机这种结构，它能否代表大脑的所有基本活动？

谢诺夫斯基： 计算是一个在物理学中找不到的概念。因为不需要。

在物理学中，有一个方程叫做哈密顿量，比如氢原子的方程。你分析这个方程，并将其与实验进行比较。但哈密顿量不会改变，你也不可能改变实际的物理学。据我们所知，它在整个宇宙中都是相同的。所以对我来说，这是一个新的概念。

但大脑能通过“学习”来修改自己的公式。

我们把物理学的能量方程引入AI，但允许AI通过经验来修改这个方程，这在当时是一个全新的概念——既不是纯粹的物理，也不是传统的计算。

正是因为我们把物理学直觉引入了大脑研究，才打破了僵局。看看现在的NeurIPS 会议有多火爆，这就证明了当年我们把“物理学”和“计算”结合起来这条路，彻底走通了。

03

与辛顿的“分歧”与“共鸣”

郝博阳： 您和辛顿共同开发了玻尔兹曼机，但后来他走向了更偏向工程的“反向传播”算法，而您坚守在生物学领域。这种分歧背后的原因是什么？

谢诺夫斯基： 这实际上是我们合作的核心——我们有互补的背景。

杰弗里的背景是计算机科学和心理学，他对这两个学科有非常深入的理解。我的背景是物理学和神经科学。

所以我们融合在一起，这就是玻尔兹曼机的来源。

我们都确信玻尔兹曼机学习算法就是大脑的工作方式。原因是它是局部的。学习必须是局部的。（这是因为在大脑中，一个神经元或突触只能感知到它直接连接的邻居，它无法像计算机那样拥有全局视角。）

现在用于所有这些大型语言模型的“反向传播算法”不是局部的。（它要求把最后的误差信号精确地传回前面的每一层，而大脑中并没有这种专门用来传递误差的反向线路）**所以显然这不是大脑学习的方式。**

辛顿沿着那条线继续前进。他与戴夫·鲁梅尔哈特（Dave Rumelhart）同时开发了反向传播学习算法，这在学习所需时间和网络规模方面效率要高得多。它可以很好地扩展。

你说得对，我专注于神经科学，我想了解大脑是如何工作的。但这些年来我们一直保持联系，我们是非常好的朋友。我们来回交流，当有新发现时——比如**辛顿在工程上发现了一种叫做 dropout（随机丢弃）的东西，如果你在学习过程中随机丢弃、关闭一些单元，你可以获得更高效的学习，防止网络死记硬背。这时我们发现工程和生物学惊人地殊途同归了。**

事实证明，大脑中的突触，特别是皮层中的突触是概率性的。也就是说，有时你有输入但没有输出（这看起来像是一种生物故障，但实际上就是大自然的 dropout）**也许这有助于我们理解为什么大脑是概率性的。此外，这在能量上非常高效，因为你不希望所有突触同时活跃。所以如果你只有一小部分可以激活（稀疏激活），你实际上可以学得更好，而且能耗更少。**

2024年诺贝尔奖授予了我的论文导师约翰·霍普菲尔德和杰弗里·辛顿。当杰弗里发表诺贝尔奖演讲时，他说："特里和我原本确信我们会因为解释大脑如何工作而获得诺贝尔奖。结果我们在这方面错了。（**因为玻尔兹曼机并没有成为最终解释大脑的完美模型，后来反而是不那么'生物'的反向传播赢了）**但这对于物理学诺贝尔奖来说已经足够好了。"

尽管如此，我们早期一起做的这项工作得到认可是很棒的。

04

为什么从生物角度入手的AI研究得了物理学奖

郝博阳：这仍然很有趣，辛顿获得物理学诺贝尔奖，因为对我来说，玻尔兹曼机更像是试图解释大脑的生物学机制。

谢诺夫斯基： 玻尔兹曼机受到生物学的启发，但它确实是物理学。分析和洞察都来自物理学——**特别是统计力学，我们将神经网络的状态看作是一个物理系统的能量状态，学习就是寻找能量最小化的过程。**所以这是完美的。但很多人质疑，这与物理学有什么关系？事实证明，物理学是现代AI的灵感来源。

上世纪的现代AI完全由逻辑和规则主导（即符号主义AI，试图把所有知识写成代码规则），这一切现在都改变了。现在物理学家已经进入这个领域，不仅仅是约翰·霍普菲尔德，还有许多物理学家，比如海因茨·泽波林斯基（Haim Sompolinsky）、拉里·阿博特（Larry Abbott）。

我们三个人几年前获得了大脑奖（Brain Prize），这是神经科学领域的最高荣誉。这是因为我们建立了这个领域——计算神经科学——它真的起飞了，并且真正帮助神经科学家理解大脑整合的方式。

实际上，我们早期工作中产生了一个非常重要的一般原理，那就是：在上个世纪，神经科学的焦点是单个神经元。我们现在有工具和技术。例如，大脑计划（Brain Initiative）是科学和工程中的一个重大挑战。

在过去十年中，我们从一次一个神经元发展到现在的记录是同时一百万个神经元，神经科学家通常可以在数十个大脑区域记录数千个神经元。所以，现在我们有了全局图景，这确实给了我们与上个世纪不同的视角。这仍在进行中，但通过分析所有数据，我们开始理解大脑中的大规模活动模式。**即智能不是源于某个超级神经元，而是涌现于成千上万个神经元的集体协作之中。**我认为这真的是从我们早期所有工作中产生的。

05

ChatGPT是一个可以被完全“解剖”的大脑

郝博阳：与您当年构建玻尔兹曼机（早期神经网络模型）的时代相比，现在的大型语言模型（LLM）与人脑有什么本质不同？现在的研究似乎认为，Transformer架构本身的工作方式可能并不像人脑。

谢诺夫斯基：现在，**我们唯一能确定的是：它不是人类。**它可能是大脑某种功能的简化版本，但它的底层机制绝对不是人类的大脑（它基于硅基芯片和反向传播，而非生物神经元和化学递质）。

但这正是令人兴奋的地方。虽然它可能不是人类，但我们拥有对它的**完全访问权**——这与大脑不同，大脑太复杂了，我们永远无法无损地获取每一个神经元的细节。但对于这个模型，我们知道它的每一个细节。

ChatGPT本质上就是一个巨大的方程，一个单一的、虽然庞大但确定的数学方程，你可以把它完整地写在黑板上。你可以访问每一个输入数据、每一个神经元的激活模式，以及它们随时间变化的动态，所有这些都是可以被拆解分析的（这被称为“机械可解释性”）。

所以现在的挑战是：动用数学、物理学和工程学的全部武器库，去搞清楚这个大方程究竟是如何运作，从而能够以这种（看似有智能的）方式回应我们。

谁比数学家更懂方程呢？因此，世界上最顶尖的数学家现在都开始研究这个问题。所以我认为，我们获得对它深刻的数学理解，只是一个时间问题。

其实在过去几年中，出现了一个全新的领域，叫做神经AI（Neuro-AI）。这是什么？这是两类人群的交集：像杰弗里这样致力于创造更好AI的人，和像我这样致力于理解大脑运作的人。这是历史上第一次，这两个群体可以真正地相互对话，因为我们开始使用相同的数学语言、相同的底层原理（如神经网络、优化函数）。

现在，通过观察AI的表现，它确实可能反过来帮助我们理解大脑。我认为未来将是这种**双向的启发和交流。**

所以我认为我们最终将能够完成两件事。首先，从数学上彻底理解这些大语言模型是如何工作的，并制造出更好的版本。这方面仍有巨大的改进空间。一旦我们以深刻的数学方式理解了这些“通用智能原理”，这将反过来帮助我们理解并解释人类自己的行为方式和思考方式。

所有这些突破可能会在未来几十年内陆续发生。但基础研究可能需要更长的时间来沉淀，就像物理学的发展史一样。

举个例子，量子力学诞生于20世纪初，但它花了几十年，许多许多个十年，才被转化为实用的技术，比如激光。

你知道吗？最早的激光器，需要一个诺贝尔奖级别的物理学家才能把它造出来，而且它需要整整一个房间的光学设备。但在那种情况下，可能过了50年，甚至60、70年后，量子力学才真正变成了支撑现代通信网络、计算机芯片工作方式的基石。但这一切辉煌，都源自物理学早期的那些基础研究。

06

AI缺失的大脑拼图之一：基底神经节

郝博阳：强化学习是实现（通用）机器学习的最佳路径吗？

谢诺夫斯基：让我们退一步，先看看人类大脑是如何学习的。回想一下你的求学经历，你在学校里主要在做什么？在学习阅读，对吧？这需要极其漫长的时间，往往耗费数年之久。你花了多久才学会？

郝博阳：大概需要20年甚至更久，才能真正读懂深奥的文章。

谢诺夫斯基：没错，20年。这是巨大的时间和精力投入。

像“视觉”这样的能力，你完全不需要刻意学习就能掌握（这是进化赋予的本能，写在基因里）；而理解、阅读、写作，以及数学、算术，却需要几十年。你需要漫长的过程去学习乘除法、平方根、分数，再进阶到微积分等等。

从进化角度看，大脑原本并不是为了处理这些任务（数学、阅读）而设计的，但它却具备学习这些的能力。我们是怎么做到的呢？

事实证明，我们是调用了大脑的一个特定区域来完成这件事的，这个区域的设计初衷是为了帮你自动化地学习技能。

举个例子，学骑自行车。第一次骑你会摔倒，你不知道如何保持平衡。但通过练习、练习、再练习（“试错”过程），最终这些动作变成了一种自动化的反应。你甚至意识不到你在做这些动作，意识层面根本无需介入。

在大脑解剖学中，负责这部分功能的区域叫做基底神经节（Basal Ganglia），它位于大脑皮层下方。它的核心作用是学习一连串的动作以获得奖励（Reward）。在这里，“奖励”可以是成功骑稳了车、打赢了网球，或者是解出了一道数学题。

这是一个与大脑皮层互补的学习系统。大脑皮层负责你有意识感知的“认知部分”（即知识和推理），而基底神经节负责基于奖励的“强化部分”（即行为和直觉）。你需要两者结合：既需要强化学习（Action/Reward），也需要认知学习（Knowledge/Reasoning）。

反观现在的AI（主要是大语言模型）：我们有了强大的“认知部分”（类似大脑皮层，通过海量数据预训练获得），但我们缺乏深度整合的“强化部分”。

虽然现在AI也会用到强化学习，但这通常只发生在预训练结束后的微调阶段（即RLHF，人类反馈强化学习）。在AI领域，我们将这称为“对齐”（Alignment，即让AI的价值观符合人类利益）。现在的做法是把“对齐”放在最后一步（微调）来做，但这太晚了。

强化学习应该贯穿于AI发展的整个过程，就像人类成长时，大脑的这两个部分是同步发育、相互交织的一样。

在人类社会中，这种“对齐”是从小就开始的，这样人们才能在同一种文化下协作。这种协作需求真正推动了语言的发展。语言的进化是非常晚近的事情，大约发生在几十万年前。人类语言的存在初衷，就是为了帮助小群体进行社交协作。比如，通过语言沟通，大家才能一起合作捕猎巨大的猛犸象。

这种协作还需要利用资源的能力。如果你在茫茫草原上，你必须知道哪些植物根茎可食用，哪些有毒。这些知识都是通过文化传承、通过一代代的**试错**得来的。

所以，**你需要“认知”来传递信息，也需要“强化”来筛选生存价值。这就是人类进化的方式。我认为，既然我们试图让大语言模型更接近人类的思维方式，那么AI的发展最终也会遵循这条（认知与强化深度融合的）路径。**

AI缺失的大脑拼图之二：内在的生命力和反思

郝博阳：理查德·萨顿（Richard Sutton，强化学习之父）最近提到，现代AI缺乏“元学习”（学会如何学习）和“持续学习”（终身学习而不遗忘）的能力。我们该如何将这两个关键特性融入到现在的模型中？

谢诺夫斯基：我们要做的不仅仅是在现有模型上打个补丁、附加一些功能那么简单。我认为，我们需要对整个架构进行根本性的重组。

我所说的架构，是指神经元的组织和运作方式。目前我们对大脑的理解中，有一些关键机制还没有被AI借鉴。我给你举一个例子：**神经调质系统**（Neuromodulator System）。

什么是神经调质？在现有的大语言模型中，神经元之间只有简单的突触连接，它们通过加权求和来处理输入（这是一种静态的权重）。但生物大脑中的神经调质完全不同，当它建立连接时，它能改变神经元整合信息的方式。换句话说，它是在动态地“调制”输入信号（就像调节收音机的音量或频率，而不仅仅是传输信号）。

大脑中有数十种神经调质。比如多巴胺（Dopamine），我们知道它对强化学习至关重要。现在的AI中已经有了它的对应物，叫做时间差分学习（Temporal Difference Learning，简称TD Learning）。我们知道生物大脑中的多巴胺神经元就在执行这个算法。它的核心逻辑是：只有当你获得的奖励比你预期的多或少时，学习才会发生。这就是“奖励预测误差”信号（Reward Prediction Error，即“惊讶”驱动学习）。

除了多巴胺，还有专门用于处理“惊讶”的其他调质，以及用于社会整合的调质——比如催产素（Oxytocin）。它对于生命早期母亲与孩子之间建立情感纽带（Bonding）非常重要。

这很有意思，杰弗里·辛顿（Geoffrey Hinton）最近接受采访时也谈到了这点。你知道，诺贝尔奖得主总是备受关注。杰弗里私下跟我说：“特里，终于有一次，人们开始认真听我说话了。”

他被问到AI缺失了什么？杰弗里说：我们知道人类生命早期，母亲和孩子之间会形成这种深刻的生化联系。为什么不把它引入AI呢？ 为什么不呢？这（基于情感纽带的连接）是智能的基本原则。

郝博阳：是的，他认为这是保护人类免受AI伤害的最好论据。

谢诺夫斯基：绝对是。如果AI能与人类建立这种深层的情感联系，这或许能从根本上防止它产生恶意的偏见或伤害行为。

但另一方面，我认为AI缺失的东西甚至超越了神经调节。那就是我认为至关重要的一点：我称之为“自主生成的活动”（Self-Generated Activity）。

这是什么意思呢？

你看，现在的AI正在被拿来与最聪明的人类做比较。我们要它去考法学院、医学院，做智商测试。现在的AI基本上已经高分通过了所有这些测试。一个单一的神经网络能考上哈佛医学院、能做律师，这确实令人难以置信。

但在这些成就背后，真正缺失的是“内在的生命力”。

现在，如果你与ChatGPT、DeepSeek或其他任何一个模型对话——现在市面上有几十种了——你会觉得你和它建立了一种关系。你可以提问，它可以扩展回答，你获得信息。它像是一个很好的伙伴（Partner）。就像你身边坐着一个非常聪明的同事，你只要问对问题，他就能帮你解决难题。

但是，一旦你停止说话，它就沉默了。它没有自主生成的活动。它不再“思考”。

换句话说，作为人类，你可以独自坐在一间漆黑的屋子里，没有任何感官输入（看不见听不见），也没有任何运动输出（不说话不动），但你依然在思考。

你在计划明天要做什么（未来模拟）；你在回想昨天发生了什么（记忆重组）；你甚至在反思自己刚才的念头（元认知）。

这一切都完全在大脑内部自主生成，是一场持续不断的内在电影。但这在目前的大型语言模型中是完全缺失的（LLM是无状态的，没有输入就没有计算）。

这是一个典型的例子，说明了那些对人类至关重要、但AI尚不具备的能力。我们需要先从神经科学的角度理解这种“自主活动”的机制。虽然我们现在还不知道“思考”确切的神经回路是什么，但如果我们能在这个问题上取得进展，它将帮助我们制造出能够以更接近人类方式去真正“思考”的大型语言模型（而不仅仅是做概率预测）。

08

AI缺失的大脑拼图之三：海马体

郝博阳：顺着您的思路，我认为这里的关键矛盾在于：如果要求大型语言模型（LLM）像您所说的那样进行“持续的主动思考”（即在无人提问时也自主反思），它面临一个巨大的物理瓶颈——**记忆。**

几天前，Yoshua Bengio（图灵奖得主、深度学习三巨头之一）提出了一个关于AGI的新定义。他参考了衡量人类智商的方式，列出了影响智力的10个关键维度。他认为，只有当LLM在这10个领域都达标时，才能被称为AGI。**在“记忆”这个特定维度上，它的得分基本是0。**

您如何看待目前AI的这种“失忆”状态？

谢诺夫斯基：好的。我在新书《大语言模型》中，特意写了几章来专门讨论你提到的这个问题。

你是完全正确的，目前的AI缺失的一个关键要素就是**长期记忆**（Long-term Memory）。

试想一下，你今天和AI进行了一次非常深刻的对话。如果你第二天回来想接着聊，你必须从头开始，因为它根本不记得昨天发生了什么。除非你把前一天的对话记录全部作为新的输入喂给它（但这受限于上下文窗口长度），否则它无法接续之前的思维。这是一个绝对的硬伤。

你刚才提到的另一个关键点是“持续学习”（Continuous Learning）。现在的大模型，工作流程是：先进行大规模预训练，然后进行微调。**一旦训练结束，参数权重就锁定了，就没有“新的学习”发生了。**

之后发生的一切都只是**“推理”**（Inference）。也就是说，它虽然能回答你的问题，但它不能通过改变神经元的权重来适应新的经验。

这与人类大脑截然不同。人类的大脑在每时每刻都在通过改变突触连接来学习。所以我列了一个清单，大约有十几项我们已知人类大脑能做、但目前LLM还做不到的事情（持续修改自身权重就是其中之一）。

郝博阳：如果我们要赋予现在的AI模型以（类似人类的）长期记忆能力，您认为可行的路径是什么？

谢诺夫斯基：长期记忆并不是孤立存在的，它必须与不断涌入的“新信息”协同工作。

当新信息进来时，你希望能记住它，但你不能把所有东西都无差别地塞进去，因为大脑的容量是有限的。所以大脑采取的策略是——**筛选**，它只选择最重要的新信息进行存储。

注意力（Attention）是我们已知的第一道过滤器。通常你只会记住你首先关注到的东西。

但在这些你关注到的信息中，哪些是最重要的？以及如何把它们整合到你已有的庞大知识库中？这关键的一步发生在睡眠期间。当你睡着时，大脑进入了一种完全不同的动态状态（进行离线处理和归档）。

我们对这个过程已经很了解了。有一点非常确定：大脑中有一个关键区域叫海马体（Hippocampus）——顺便说一句，现有的大语言模型中完全缺失了这个组件。

海马体的作用是与大脑皮层“对话”，它帮助皮层弄清楚把你的新经验存放在哪，以及如何在不打乱已有知识的前提下存储新知识。如果你搞乱了这一点（即新知识覆盖旧知识），就会发生灾难性的记忆遗忘（Catastrophic Forgetting）。

这就是为什么人类需要8小时的睡眠来协调这一切。在睡眠中，大脑进行着无数微小的调整。在你年轻时，大量信息涌入，整个皮层正在组装，新的突触在疯狂生成。但实际上这种过程在成年后仍在继续。新突触的不断形成，正是我们拥有“终身学习”能力的物理基础。

这又是我在《大语言模型》一书中提到的、目前大语言模型缺失的十几项关键功能之一。这并不神秘，只是现在的模型太“狭窄”了。它虽然参数巨大，但它只模拟了大脑皮层的一小部分功能，对于真正的生物生存而言，这是远远不够的。

09

复刻人脑不应是唯一目标，理解多样性才是

郝博阳：你觉得Bengio的AGI定义合理吗？

谢诺夫斯基：关于Yoshua Bengio提到的AGI定义，我认为他正在试图解决一个非常棘手的问题。他的定义是很好的尝试，但通过定义来界定AGI本身就很困难。每个人心中都有自己的一套“通用智能”的标准。

我认为问题的根源在于，**“AGI”这个词就像“意识”（Consciousness）这个词一样，是一个模糊的集合体。**

“意识”并没有一个正式的、公认的科学定义。它包含许多不同的维度，有些人强调感知，有些人强调自我反思。我们无法用其他更精准的词来定义它，这导致定义常常陷入**循环**

论证（Circular Reasoning）。

我们对“意识”只有一种直觉，哲学家们为此写了无数本书。

但这让我联想到生物学在20世纪发现**遗传密码（DNA）**之前的状态。在那之前，科学界曾对“什么是生命”有过巨大的争论。

人们困惑：有些东西是活的，比如细菌；有些东西不是，比如地毯、花粉粒。它们的本质区别是什么？

当时人们提出了**“生命力”**（Vitalism）的概念。人们认为，活着的东西体内有一种特殊的“物质”或“能量”，当你死了，这种物质就消失了。

可是，我们现在还谈论“生命力”吗？早就不谈了。

为什么？因为“生命”这个抽象的概念，已经被具体化为非常复杂的**生化机制**（Biochemical Mechanisms）。我们现在谈论的是DNA复制、蛋白质合成、代谢途径。我们谈论的是具体的机制。我们对细胞如何工作、以及它们在癌症等病变中如何出错有了深刻的理解。

因为我们理解了机制，所以我们不再需要“生命”这个神秘的哲学词汇来解释现象。

我们讨论的是特定的化学路径，是细胞自组织成多细胞生物的特定能力。这些都是可以从机械、物理角度去研究的实体。

我认为，同样的事情将会发生在“意识”和“AGI”上。

一旦我们真正理解了LLM是如何工作的——当我们从数学本质上搞懂了它；一旦我们开始能够基于原理改进它们，我们将拥有一套非常复杂的“智能机制”来解释这一切。

到那时，它会有“意识”吗？谁知道呢？（或者说，谁还在乎那个旧词呢？）

我的猜测是，它将拥有一种与我们截然不同的意识。

你看，自然界中所有的动物都以不同的方式拥有意识。蝙蝠的意识、章鱼的意识，都与人类不同。它们进化出的意识形式，是为了适应它们特定的**生态位**（Ecological Niche）——它们需要感知周围有什么、如何获取食物、如何交配。这些功能都对应着大脑中特定的区域。

现在的AI也有它自己的“生态位”（数字世界），所以它的“意识”形式也会不同。

所以我个人更感兴趣的是去理解底层的机制，以及这些简单机制是如何涌现出复杂行为的。

至于像“AGI”、“通用人工智能”、“意识”，甚至“理解”（Understanding）这个词本身——说实话，我们根本不知道“理解某事”在神经层面上到底意味着什么。

这点现在暴露得非常明显，因为学界正在进行一场大辩论：大型语言模型真的“理解”语言吗？还是只是统计概率的模仿？

如果顶尖专家们都无法通过测试达成一致，那么这意味着并不是模型有问题，而是我们对“理解”这个概念本身的定义出了问题。这就是我们现在的处境。

郝博阳：实际上，本吉奥（Yoshua Bengio）只是选择了一条捷径。他之所以用这套标准（人类智商测试的10个维度）来定义AI的理解能力，是因为这是目前我们唯一用来衡量“人类有多聪明”的现成标尺。

谢诺夫斯基：但我认为，那种“人类是进化顶峰”的观念，真的——

你看，有些动物在某些方面比我们强得多。比如蝙蝠能使用回声定位（感知我们看不见的世界），蚂蚁能通过蚁群智慧（Swarm Intelligence）进行精密协作。大自然是令人难以置信的多样和复杂的。

我们自认为是万物之灵——英语里专门有个词叫**“傲慢”**（Hubris）来形容这种心态。

我们以为自己是最伟大的，但我们其实只是众多物种中的一个。没错，我们进化出了语言，这确实帮了忙——让我们能狩猎协作、互相交流——但**语言其实是一个非常、非常“钝拙”的工具**（Blunt Tool）。

语言并不完美，它既能用于善，也能用于恶，用于欺骗和邪恶。所以我真的认为我们对自己评价过高了，实际上我们作为一个物种有很多缺陷。

郝博阳：但这确实是现代AI的起点啊。正是因为您和其他科学家对人类大脑感兴趣，并将人脑作为理解“智能如何运作”的基础，才有了今天的发展，对吧？

那么在这个前提下，现在很多人认为“模拟人脑”是通往AGI（通用人工智能）的唯一路径。您认为这是对的吗？还是说我们想要制造的智能，其实并不需要完全复刻人脑？

谢诺夫斯基：**复刻人脑确实是人工智能的工程目标之一。但我认为还有一个更有趣的目标（科学目标），那就是理解不同大脑之间多样性的基础。**

像我刚才说的，很多动物拥有令人惊叹的能力，这些能力完美适配它们的生活方式，以及在与我们截然不同的环境中生存的需求。这需要非常复杂的脑回路和大脑结构来支撑。

尽管如此，所有动物的大脑在发育方式和功能分区上都有底层的相似性。大脑有数百个专门用于**生存**的区域（比如负责本能、运动、体温调节等）。

而目前的大型语言模型（LLM），充其量只是大脑皮层（Cortex）的一个模型，它只模拟了大脑的一小部分功能。

当然，这是很重要的一部分，因为人类的大脑皮层确实进化得比其他灵长类动物大得多（负责语言和推理）。但大自然的智慧要比这“皮层智能”丰富得多，也复杂得多。

所谓的AGI，好吧，这对工程师来说是个好目标（造出像人一样好用的工具）。但如果你真的对基本原理感兴趣（作为科学家），我们应该向自然界的其他生物学习。

这对于建立一个真正的、关于知识和理解的**统一理论**（不仅仅是解释人类，而是解释所有形式的智能）是至关重要的。

｜认识我们｜

《Hao好聊》是由腾讯科技发起的深度访谈项目。我们关注那些正在重塑时代的人——他们是第一批触摸未来的人，在技术变革的浪尖上冲浪；也是搅动潮水的创造者，用代码与远见重新定义商业与文明的边界。

我们聚焦科技领域的「先行者」，与他们展开沉浸式长访谈，探寻技术浪潮下的思想交锋。当AI开始改写人类社会的底层逻辑，亲历者如何理解这场变革？当技术奇点临近，那些最接近答案的人，如何看待我们共同的未来 ？

《Hao好聊》希望深入技术狂热背后的人文思考，记录产业剧变中的个体抉择，与行业参与者共同探索未来的可能性，成为产业进化的见证者。

｜联系作者｜

作者专注AI赛道，如需交流或提供信息，请添加微信haoboyang001

推荐阅读

后AGI时代，当99%的人类价值归零，资本主义是否会幸存？｜Hao好聊 X 张笑宇

聊聊创业公司与谷歌达成合作的幕后故事，以及AR眼镜的“iPhone时刻”｜Hao好聊 X 徐驰

10分钟就拿到了朱啸虎投资的AI陪伴产品，想让年轻人不孤独｜Hao好聊X孙兆治