**李秀红**：无问芯穹副总裁，致力于研发高性能的大模型推理基础设施，提供多种主流模型和多种国产芯片之间的 M*N 中间层支持。加入无问芯穹之前，在北京大学担任助理研究员，研究领域为计算机体系结构、异构计算和深度学习系统，在 ISCA、MICRO、HPCA、TC、PPoPP 等相关领域国际顶级期刊会议发表论文 20 余篇，相关研究成果以第一作者或通信作者获得 CCF A 类会议 ASPLOS 2024 最佳论文、CCF A 类会议 PPoPP 2019 最佳论文提名。

专家名片

# DeepSeek 大模型：技术突破与产业变革的核心驱动力

■ 李秀红

2025 年春节期间，中国人工智能领域迎来标志性突破—DeepSee 公司推出的 DeepSeek-R1 大模型，凭借“更高智能、更低成本、更开放生态”三大核心优势，迅速成为全球 AI 领域焦点。

## DeepSee 火出圈的核心优势

人类智能包括两大系统，第一类系统是“大脑快速、自动、直观的方法”，第二类系统是“思维的慢速，理性占据主导地位的分析模式”。

更高智能，从直觉到推理的跨越。DeepSeek 在多个基准测试中表现优异，展现出强大的数学推理和代码生成能力。其关键在于从“快系统”直觉感知向“慢系统”逻辑推理的升级，通过思维链（Chain of Thought）技术，将复杂问题拆解为多个子步骤，依赖规则逐步生成 Action，显著提升复杂任务的解决能力。

更低成本，性价比的革命性突破。据相关数据显示，在训练成本上，DeepSeek-V3 仅为 557.6 万美元，远低于 GPT-4 的 6300 万美元和 Llama 3.1 405B 的 5800 万美元。推理成本方面，其每百万 Token 的 API 定价仅为 0.27 美元（输入）和 1.10 美元（输出），较 GPT-4o 的 2.5 美元（输入）和 10 美元（输出）降低约 90%。这种成本优势源于软硬件协同优化。

更开放生态，技术平权的推动者。DeepSeek 通过开源技术报告和模型，降低行业准入门槛，吸引全球开发者参与生态建设。其技术文档详细披露了 Multi-Head Latent Attention（MLA）、DeepSeekMoE 架构等核心技术，以及 FP8 训练、多 Token 预测等优化策略。开源生态的构建不仅加速了技术迭代，还推动了人工智能在中小企业和垂直领域的普及，成为行业发展的重要基础设施。

## 突破算力瓶颈的关键路径

从稠密到稀疏的范式转变。面对稠密模型在 72B 参数规模后性能饱和的问题，DeepSeek 采用混合专家模型（MoE），通过动态选择部分专家参数参与计算，实现 2~3 倍算力撬动 10 倍模型规模的效果。

稀疏注意力技术进一步优化计算效率，如 Native Sparse Attention（NSA）通过层次化 Token 压缩和块状 Token 选择，在长文本处理中实现 10 倍加速，推理速度从原始注意力的近千秒缩短至一百秒左右，显著提升长上下文场景的处理能力。

从训练到推理的全流程优化。训练框架方面，DeepSeek 优化了分布式训练中的通信和计算重叠，如通过 DualPipe 技术实现数据并行、模型并行、流水线并行的混合策略，512 卡扩展效率达到 76%。推理框架则采用分页式内存管理（如 vLLM 的 PagedAttention），减少显存碎片，提升服务吞吐量，Llama3-8B 模型请求服务率达 35req/s。

应对后摩尔时代的挑战。AI 芯片从指令驱动的 CPU/GPU，发展到数据流驱动的存算一体芯片（如 Cerebras WSE-2）和神经形态芯片（如 Intel Loihi），能效比提升 5 个数量级。然而，摩尔定律放缓和美国禁令导致先进制程（14nm 以下）受限，芯片制造公司推动晶圆级芯片和先进封装技术（如 3D 堆叠、芯粒互连），突破单芯片面积和良率瓶颈，实现 P 级算力集成。

应对工艺墙的系统方案。针对制程工艺瓶颈，芯片制造公司联合国内产业链探索 12/7nm 节点的优化方案，通过设计创新弥补制程差距。在封装层面，采用芯粒（Chiplet）技术实现异构集成，将计算芯粒与存储芯粒高速互连，提升带宽和能效，为国产芯片突破封锁提供新路径。

## 从消费级到战略行业的全面渗透

2C 场景，重塑生活与生产力工具。在文化领域，中电信文宣科技接入 DeepSeek 后，游客复购率提升 27%，文化体验满意度提高 35%；教育领域，某智慧校园学情诊断系统使教师备课效率提升 40%，高风险学生干预成功率提升 65%；娱乐领域，短视频平台日均产出创意内容超 10 万条，互动率提升 22%；效率工具方面，代码生成速度比 GPT-4 快 3.7 倍，会议纪要生成准确率达 98%。

2B 场景，驱动行业智能化转型。能源行业，DeepSeek 融合气象、地理数据构建动态安全域模型，优化分布式能源管理；制造业，通过工业知识图谱和多模态处理提升故障诊断准确率 30%，良品率提升 10%~20%；金融业，某银行信贷审核误判率降低 58%，基金公司策略收益提升 23%；医疗行业，加速药物研发周期 70%，提升临床决策效率。

AI Agent，打通垂直场景的智能桥梁。以 Manus 为例，其在 GAIA 基准测试中工具调用成功率达 94.7%，显著高于 OpenAI Agent 的 72.3%。在生活场景中，5 步以上复杂任务成功率比 OpenAI 方案高 23%；金融分析中，用户干预后任务成功率提升至 83%；政务场景中，深圳“AI 公务员”日均处理 1.2 万个咨询，座席减少 60%。

## 未来趋势：算力竞争与国产化闭环构建

开源生态，AI 领域的“Android 时刻”。DeepSeek 的开源模式打破技术垄断，推动行业从闭源走向开放，类似 Android 对移动应用的赋能，其基准测试表现比肩闭源模型，GitHub 星标数超越 OpenAI，日均 API 调用量突破 2000 万次，成为全球开发者的重要选择。

算力需求激增，端云协同的新基建。云侧推理需求爆发，短期全国活跃用户数预计达 1.5 亿，日均 Token 用量达 11.25 万亿，推动新一代推理集群向资源池化、动态调度发展；端侧通过定制芯片实现高能效推理，7B 模型推理性能>150tokens/s，能效>20tokens/J，助力智能终端普及。

技术挑战，效率与成本的持续优化。尽管 DeepSeek 在成本和效率上取得突破，仍需应对算力异构、数据出域、电价差异等问题。未来需进一步优化稀疏化、低比特量化技术，提升端云协同效率，降低推理成本，推动人工智能从“奢侈品”变为“必需品”。

DeepSeek 大模型的崛起，标志着人工智能从理论探索走向产业落地的关键阶段。其技术创新不仅突破了算力和成本瓶颈，更通过开放生态和端云协同，推动 AI 与各行业深度融合。面对中美技术竞争和全球产业变革，DeepSeek 的实践为国产化闭环构建提供了路径参考，预示着人工智能将进入效率提升、成本下降、应用爆发的黄金时代。未来，随着软硬件协同的持续深化，AI 有望成为驱动社会进步的核心基础设施，开启智能时代的新篇章。

**科学导报记者马骏根据录音整理**

# 机器智能的另类性：从科技到社会

■ 陈小平

回顾人工智能的发展，特别是考虑大型语言模型的出现，尽管人工智能在过去 70 多年中取得了巨大的进步，但它始终遵循着奠基者图灵和其他先驱者所提出的基本原理，这就是机器智能的另类性。

到底什么是人工智能？怎么理解人工智能？人们通常将“人工”和“智能”这两个概念的常识性理解结合起来，形成对人工智能的常识性理解。然而，图灵在 1950 年发表的著名论文中第一个自然段，就否定了这种字面意义上的理解。可以认为图灵的一个重要理由是——任何科学技术不应局限于常识概念。以牛顿力学为例，它并非基于物理常识，而是基于四条基本原理，即牛顿运动三定律和万有引力定律。通过数学和逻辑推理，从四条基本原理可以推导出牛顿力学的所有定理。

在图灵的机器智能观中，首先有一个潜在的假说，可以通过计算机即图灵机来模仿人的智能行为，如推理、决策、学习等功能的组合。但是，他并没有直接使用假说这个词，为什么呢？因为，如果把这句话当作一个假说，就犯了一个错误，也就是他一开始就否定的那个错误，即用常识性概念来定义机器智能。因此，他通过图灵测试来替代图灵假说。图灵测试本身包含许多内容，但并没有完全表达出图灵的真实想法。图灵的真实想法是通过图灵测试来证明他的假说是否成立。在 1948 年的报告中，图灵认为，机器智能的工作原理可以与人类智能的工作原理相同，也可以不同。我将这种相同称为“原理模拟”，而将不同称为“功能模仿”。这是图灵关于机器智能的一个基本观点，正是这个观点决定了机器智能的另类性。

根据图灵机器智能观进一步分析后，我认为每一种智能原理都有其先天的局限性，因此是不同的智能，不同的智能不会完全等同。所以，不必焦虑机器智能完全取代人类智能。但是，机器智能能够取代人类的大量工作，这才是真正值得重视的问题。

既然人工智能的原理与人类智能的原理不同，那么它的原理究竟是什么？每当开发出一个人工智能系统比如大模型，它就有一套相应的原理。这些原理是什么？是否突破了智能的传统边界？人们该如何把握它们？这些问题非常重要。

根据对大模型底层原理的分析，我提出了一个科学假说，称之为 LC 假说。当时支持这一假说的证据较少，然而自从在学术期刊上发表后，深度测试越来越多，而且大部分深度测试结果都支持这一假说。大量深度测试结果表明，传统的科学理论、数学、计算机科学以及传统人工智能中的强力法，都是基于概括性和强共识性的。这表明，人工智能研究已经超越了科学技术的传统边界，已经步入了“无人区”。

大模型技术体系主要包括三大块：一是预训练；二是后训练；三是激发（通常称为提示）。通过分析，我发现这三个部分中有两个部分共享一个共同的基础设施或基本机制，即关联度预测。

从类 LC 的三条公理出发，经过逻辑和数学推导，可以得出关联度预测的一些基本性质。其中一个性质被称为实例性或语境纠缠性。这个性质意味着，大模型的底层机制具有语境纠缠性。然而抽象运算如计数和逻辑否定不是语境纠缠的，比如一个文本里有 27 个“~”，这个数目不受文本里其他符号的影响，不会由于文本里其他符号的改变而变多或变少。这是所有抽象运算的一个基本特性。然而大模型却无法确保正确地实现抽象运算，因为它的数学原理决定了它必然是语境纠缠的。此外，大模型在奇偶性、因果关系和反思能力等方面也存在同样的问题。

事实上，图灵在 1948 年的报告里就指出，机器的行为和这些行为给人的感觉都是“智能”的组成部分，这恰恰构成了智能的复杂性。分析表明，大模型的这一表现证明了图灵在 1948 年的判断是正确的。

**陈小平**：中国科学技术大学教授，中国科学技术大学机器人实验室主任、广东省科学院人工智能首席科学家、中国人工智能学会人工智能伦理与治理工委会主任、CAAI Fellow。获中科大“杰出研究校长奖”、机器人世界杯冠军、最佳论文奖、行业年度十大科技进展及其他国内外学术荣誉 20 余项。

专家名片

工业革命的核心是机器在许多特定领域中替代了人类的体力劳动。这些机器的体力功能并不是通用的，而是专用的，但它们取代了人类的大量体力工作，引发了工业革命。我在《企业改革与发展》2024 年 11 期的文章中提出了一个观察：工业革命之所以成功，依赖于熊彼特经济闭环。这个闭环的运作机制是：随着生产力的提升，机器数量增多，而机器的增多又需要更多的人来操作，这导致就业增加；就业增加后，工资随之增长；工资增长又带动消费增长；消费增长创造了更大的市场；而更大的市场需求又刺激生产力进一步提升。这样就形成了熊彼特经济闭环。由此可见，工业革命的逻辑简单明了。

那么，智能革命的特点又是什么呢？预期在许多特定的领域中，机器不仅在体力功能上超越人类，而且在智力功能上也超过人类，于是大多数机器不再需要人的操纵，这是根本性的不同。人们常常担心通用人工智能带来的问题，其实不必担心，因为专用的人工智能强大到一定程度之后，问题就已经出现了。一旦大量专用智能机器超过多数人的能力，熊彼特经济闭环就不再成立，即生产力的提升不再带来更多的就业机会。如果真是这样，那么未来世界就会发生根本性变化。

人工智能将如何重塑人类的文明？我将这个想法表达在《企业改革与发展》文章的最后一段话中：“工业革命向智能革命的过渡将导致‘科技进步’向‘科技重塑’的变迁。智能革命不是用更强大的技术手段，在旧世界里展开更残酷的内卷，而是颠覆旧世界、再造新世界。人类面临的最大课题是：通过智能革命，重新塑造一个什么样的新世界？”

近年来，中国在人工智能和大模型研究中进步很快，取得了一大批成果，但主要集中在技术创新方面。对于大模型这样的大型 AI 系统，工程创新比技术创新的作用更大。工程创新需要以应用为导向，组织实施大量技术创新并组合技术创新成果，以实现大模型性能的较大提升。在大模型的工程创新方面，国内一些企业进行了积极探索，取得了很好的进展。

**科学导报记者杨洋根据录音整理**

# 脑机接口技术革命：人工智能时代的下一场科技浪潮

■ 周 程

**周程**：北京大学哲学系二级教授、博士生导师、科学技术哲学教研室主任、中国科协—北京大学科学文化研究院副院长。兼任国务院学位委员会第八届学科评议组成员（科学技术史）、全国应用伦理专业学位研究生教育指导委员会委员、中国发展战略学研究会副理事长。

专家名片

从让瘫痪患者重获行动能力，到实现人类与机器的“意念交互”，这项曾被视为科幻小说的技术正在实验室和商业市场中快速落地。

脑机接口技术（BCI）的本质，是通过直接捕捉、解码和转化大脑神经活动信号，建立人脑与外部设备的实时交互通道。其核心目标在于“绕过人体外周神经与肌肉系统的限制”，常用于替代、修复、补充、增强人体的感觉、运动、语言功能或提升人机交互能力，主要由神经信号采集器、解码器与效应器构成。

根据技术路径的差异，脑机接口被分为三类：

非侵入式：通过头皮贴附电极（如 EEG 设备）采集脑电信号，安全性高但信号分辨率低，多用于消费级产品（如专注力监测头环）；

半侵入式：将电极植入颅腔但置于大脑皮层外，平衡了安全性与信号质量，适用于医疗场景；

侵入式：直接将微电极阵列植入大脑皮层，可获取高精度神经信号，但存在手术风险与排异反应，代表案例为马斯克旗下 Neuralink 的“脑芯片”。

近年来，第四类技术——“介入式”脑机接口崭露头角。通过血管介入手术将柔性电极输送至大脑特定区域，既降低了开颅风险，又显著提升了信号采集能力。

据评估机构测算，2023 年，美国脑机接口市场规模为 4.9 亿美元，预计到 2033 年将达到约 23.7 亿美元，2024~2033 年的复合年增长率为 17.68%。据该机构评估，2023 年全球脑计算机接口市场规模为 23.5 亿美元，预计到 2033 年将超过约 108.9 亿美元，2024~2033 年的复合年增长率为 16.55%。其中，美国作为技术策源地，市场规模从 2023 年的 4.9 亿美元增至 2033 年的 23.7 亿美元，医疗康复、军事、游戏娱乐成为三大核心应用领域。

资本热度印证技术潜力。自 2016 年马斯克创立 Neuralink 以来，美国脑机接口领域融资额逐年攀升。2021 年，该赛道单年融资总额突破 7.5 亿美元，创历史峰值。另一家明星公司 Synchron 则通过血管介入技术，率先获得 FDA 批准开展临床试验。

2021 年闭环神经调控治疗重度抑郁，2023 年 BrainGate2 恢复了渐冻症患者的沟通能力，2024 年 Neuralink 完成无线柔性脑机接口的植入。2024 年 Precision 的植入电极数达 4096 个。全球竞争格局已从技术研发竞赛，转向临床落地与商业模式的比拼。

中国脑机接口产业虽起步较晚，但凭借政策红利与市场潜力，正加速跻身全球第一梯队。国家“十四五”规划将“脑科学与类脑研究”列为重大科技项目，北京、上海、杭州等地相继设立脑机接口创新中心。

中国科研团队屡创里程碑。上海交通大学研发的“脑控外骨骼康复系统”，主要用于辅助中风患者进行手部的康复训练；清华大学团队开发全球首套“无线脑控自主喝水系统”，通过侵入式电极让渐冻症患者恢复基础生活能力；浙江大学完成国内第一例侵入式脑机接口临床研究，一位高位截瘫的高龄受试者通过侵入式脑机接口控制机械臂完成了喝水、进食等动作，2024 年，该受试者又以相同的方式在白板上成功地完成了多个汉字的书写。

资本市场上，脑机接口企业备受青睐。华为、科大讯飞等科技巨头已成立专项实验室，其数据算法优势或将重塑行业生态。

在 2014 年的巴西世界杯上，一位瘫痪青年经过 6 个月的训练，头戴电极帽，身披美国杜克大学医学院研发的“机械战甲”，用意念开出第一球。2016 年 10 月，志愿者通过脑机接口成功地利用意念控制了一台机械手臂，并与时任美国总统奥巴马进行了“握手”。2018 年 9 月，美国 DARPA 生物技术办公室主任透露，DARPA 于 2015 年启动了人脑控多架小型无人机系统研制项目。经过多年研究，受试者已可以借助意念同时操控 3 架不同类型的模拟飞行器，而且这些飞行器发出的信号可以直接回传至受试者的大脑，使受试者能够感知飞机器周围环境，并据此调整飞行控制策略。2019 年在亚洲消费电子展上，日产汽车展示了能解读大脑信号的脑控车系统，通过对驾驶员车辆操控意图（如左转、右转、刹车等）的预判，提升车辆的操控性能……

面向大众的非侵入式设备快速普及：美国初创公司 OpenBCI 推出的 Galea 头戴设备，可实时监测注意力水平，用于教育场景优化学习效率；中国强脑科技（BrainCo）开发的 Focus 头环已进入数千所学校，帮助学生提升课堂专注度；索尼 PlayStation 7 将搭载脑电感应手柄，玩家可通过“意念”控制游戏角色动作。

尽管前景广阔，脑机接口的大规模应用仍面临严峻挑战。

人脑每秒产生约 1TB 数据，但现有技术仅能捕捉百万分之一。信号采集精度、设备长期稳定性、算法泛化能力是三大技术壁垒。例如，侵入式电极可能因免疫反应导致信号衰减；非侵入式设备易受环境噪声干扰；跨个体、跨场景的脑电解码模型仍不成熟。

当技术触及“思维读取”与“认知增强”，伦理争议接踵而至：例如，隐私泄露风险：脑电波可能暴露个人情绪、意图甚至潜意识；技术鸿沟加剧：富人通过 BCI 提升智力或体能，或引发社会公平危机身份认知冲击：当机器深度介入神经活动，“人机边界”或将彻底模糊。

目前，仅美国 FDA、欧盟 CE 等机构制定了脑机接口医疗器械审批框架，但对消费级增强设备仍缺乏监管。中国虽在 2024 年出台《脑机接口研究伦理指引》，但技术分类、应用红线、责任认定等细则尚未明确。

面对挑战，我们呼吁建立“技术—伦理—政策”协同创新体系。科研层面：推进神经科学、材料学、AI 算法的跨学科攻关，重点突破介入式技术、柔性电极、自适应解码算法；产业层面：医疗机构、科技公司与监管部门共建临床试验平台，加速医疗场景落地；社会层面：开展公众科普与伦理讨论，避免技术滥用与社会割裂。

脑机接口不是一场“替代人类”的颠覆，而应成为“拓展人类可能性”的工具。唯有坚持技术向善、包容发展，才能让这场革命真正服务于人类福祉。

**科学导报记者王小静根据录音整理**