# **人人都在谈论AI Agent时，在谈论什么**

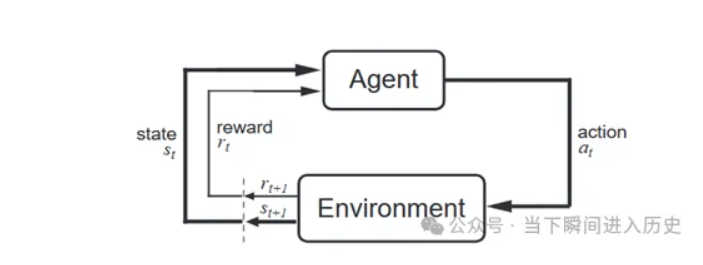
原创 Eagle 当下瞬间进入历史 *2025年01月16日 11:33* *北京*

过去一年里，AI Agent的概念越来越火，是2024年AI领域最重要的概念应该争议不大，而最近老美两位重要商业领袖的观点更进一步推高了这个热度。

微软CEO提出Agent可能将重写所有的SaaS应用，考虑到SaaS应用在整个行业的重要性，这个论断是惊人的。Nvidia CEO老黄说Agent是个几万亿美元的机会，说的更加简单直白。

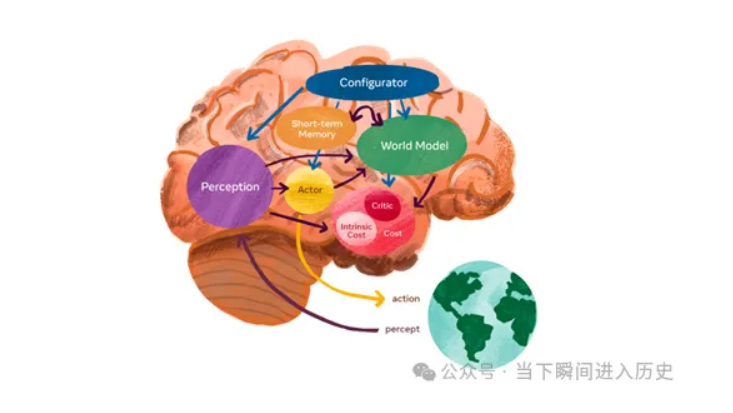
*Agent****是行业最初的理想***

几乎任何一本AI相关的教科书都会介绍一个Agent和外部环境互动构成的经典模型。这个模型简单而强大，在今天最复杂的AI系统中也会看到它的影子。



但像所有的经典模型一样，这个模型似乎过于简单了，简单到像个玩具，任何一个现实问题面临的复杂性，都很难用它简单直接去解决；同时这个模型又是如此强大，蕴含着简洁深刻的思想和丰富的可能性。

图灵奖得主Yann LeCun建议了一个稍微复杂一些的模型，也引入了更多的假设，而这些假设更像是符号主义的观点，是一个过于结构化的模型，隐含比较多的先验偏置。



Hinton老爷子最近再次强调，智能的本质来源于学习能力，而不是推理能力，推理能力是符号主义的核心观点，历史证明走不通，似乎是对弟子LeCun观点的批评。当然历史上没走通，不代表未来一定不可以。

近期有趣的发现是人工神经网络训练以后展现了和生物神经网络相似的一些结构，也许神经结构是习得的而不是先验的，如果真的这样，联结主义和符号主义也许以一种意外的方式实现了共存。

*AI Agent****火热兴起背后的技术进步***

经典的Agent模型作为一个理想参考模型存在了很多年，除了极少数简单的系统取得了有限的成功，这个模型更多用于大家沟通对智能体系统的理解。

但近期以大模型技术为代表的AI进步重新刺激了人们对于Agent的想象，大模型涌现出来的多个能力似乎都可以帮助解决原先无从着手的Agent构建难题。

借用Ilya的总结，这一轮AI技术进步的核心是利用自回归预测的方法，将互联网数据压缩到一个Transformer超级模型中，这个超级模型几乎可以作为世界模型（world model）的代表。

这个超级模型除了拥有远比普通人渊博的知识以外，研究人员还发现了两个有趣的现象：即模型拥有很强的上下文学习能力（In context learning）和思维链推理（Chain of thought）能力。围绕着这两个核心能力，发展出提示工程（Prompt engineering）和智能体（Agent）系统，最终引发了利用这些能力来构建AI原生应用的爆发。

和所有科技史的发展一样，我们目前对于超级模型本质的理解是粗浅的。比如，所谓上下文学习能力，就是一个比拟的不够准确的说法，因为这个过程中模型没有进行任何学习，模型的参数没有任何变化，模型的能力是固化的，学习在这里显然被滥用了。

当然大师Rich Sutton已经在呼吁构建持续在线学习的AI模型。我们人类目前也许唯一还可以自信的是，我们一直在终身持续学习。

如果从数学角度诠释，可以认为模型学习到数据分布，上下文学习（In context learning）的实质是通过控制输入条件来改变模型的生成结果（conditional sampling）。如果从信息论的角度看就更直接，通过上下文输入了更多的信息。

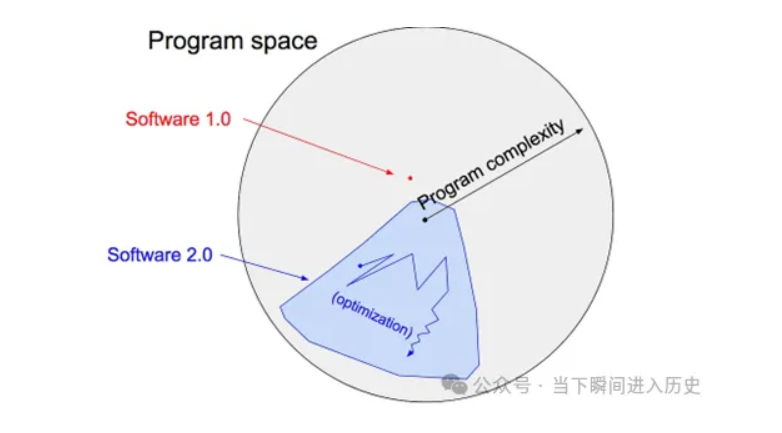
和AI Agent变热最直接关联的是模型表现出来的的思维链推理（Chain of thought）能力。对这个现象本质的理解更加表面，可以和上下文学习现象一样用控制模型条件生成来解释。

稍微深入一点的解释也许可以从计算视角看，上下文学习和思维链都是在增加模型的计算（prefill和decoding）。Stephen Wolfram也曾经用计算视角讨论过，他认为神经网络的层级决定了推理的深度，现有神经网络深度在二百层左右，限制了推理的能力。这个看法很有意思，但他忽略了解码部分的循环计算，如果考虑到解码的循环，可以认为网络计算层级要更深。今年以O1代表的推理模型增加了推理时计算也是这个思路的升维性升级。利用思维链，采样生成和强化学习结合，进一步增加了计算的循环层级，从而提高推理的深度。在这个方向上，强化学习究竟能发挥多少作用，关键在于奖励模型（Reward model）的构建，代码和数学等一些容易形式化验证的领域可能很快会被强化学习突破，但更为开放的领域依然是一个开放问题，Andrej Karpathy就认为RLHF不是真正的RL。

*AI Agent****软件模式***

一年的时间AI智能体就从一个较新的概念变成人人谈论的热点，背后有多重原因。大模型核心能力尤其是表现出来的推理能力进步是核心，利用模型的推理能力可以构建更加自主性的软件系统，和以往完全依靠程序员基于逻辑规则构建的软件系统是完全不同的范式。

从八年前Andrej Karpathy提出当时惊人的Software 2.0概念，到今天大量的AI Agent实践和总结，未来软件架构的演变肯定已经开始。

2024年初Andrew Ng 提议了四种AI Agent模式：Reflection/Tool use/Planning/Multi-agent collaboration，年底Anthropic也有一篇更详细的讨论Agent模式的博客，各种Agent编程框架更在激烈竞争程序员的关注，让人自然想起三十年前“Gang of Four”软件设计模式的总结对软件系统设计的影响。

***未来的可能***

Ilya在今年NeurIPS时间检验奖获奖演讲时讨论了他追求的超级智能（Superintelligence）。他认为现在的模型还远远谈不上是真正的智能体，推理能力也很有限，虽然在测试集上表现良好，未来的超级智能还需要真正的理解和自我意识。当然什么是真正的理解和自我意识，都更像是一个哲学问题。

Ilya提醒大家，虽然不知道超级智能什么时候到来，那个未来什么样是难以预测的（Unpredictable），但肯定会和今天完全不一样。



​