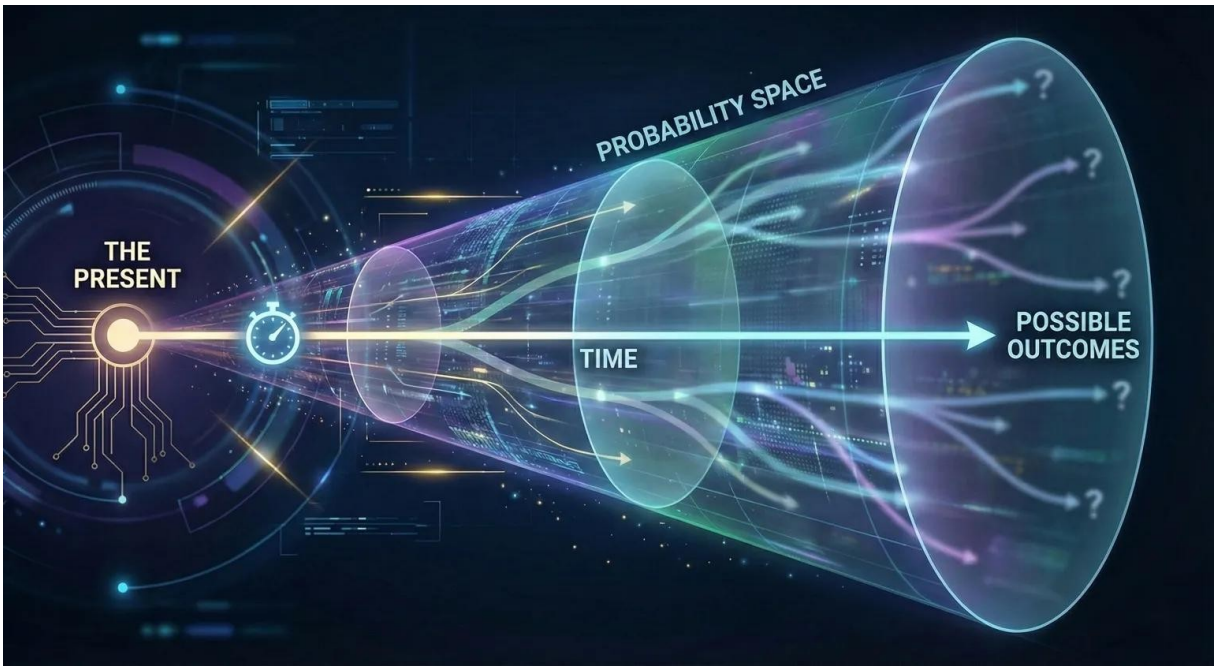
智能的本质：从世界表征、概率预测到未来生成

——从认知科学、概率论、控制论与人工智能重新理解智能



参考：backalleycoder.com

摘要

智能究竟是什么？

传统观点通常将智能理解为学习、记忆、推理、问题求解和语言理解等能力。但如果进一步追问：这些能力为什么会产生？它们共同解决的究竟是什么问题？

我们会发现，智能可能具有一个更深层次的统一结构：**主体通过感知构建世界表征，通过学习不断更新内部模型，通过概率预测理解未来的多种可能性，再通过决策和行动影响现实，使未来朝着目标方向演化。**

在这个过程中，智能不再只是对信息进行处理，而是建立一个连接过去、现在与未来的动态系统。

本文尝试提出一个研究框架：

> **世界表征 → 概率预测 → 未来生成**

从这个框架出发，我们可以重新理解人类认知、人工智能、生命系统以及多智能体之间的交互关系。

一、智能为什么需要重新定义？

我们先考虑一个简单的场景。

一个人站在射击游戏中，面对一个正在移动的目标。

他需要完成什么？

1. 感知目标的位置。
2. 判断目标的移动方向和速度。
3. 估计目标可能出现的位置。
4. 判断目标是否会改变运动方向。
5. 选择瞄准位置。
6. 执行射击。
7. 根据射击结果调整下一次判断。

表面上，这是一个视觉识别和射击问题。

但从智能的角度来看，它实际上包含了几个重要过程：

- 对现实世界进行表征。
- 根据不完整的信息建立内部模型。
- 推断未来状态的概率分布。
- 选择能够影响未来结果的行动。
- 根据反馈不断修正认知。

这里出现一个重要问题：

射手瞄准的究竟是目标当前所在的位置，还是目标未来可能出现的位置？

显然，真正有效的瞄准通常不是追踪目标已经经过的位置，而是估计目标未来的位置。

这意味着智能具有某种面向未来的特征。

主体不仅需要知道世界现在是什么样子，还需要知道世界接下来可能变成什么样子。



参考：securitylab.ru

二、世界表征：智能如何理解现实?

2.1 什么是世界表征?

世界表征（World Representation）是主体在内部建立的、关于外部世界的结构化描述。

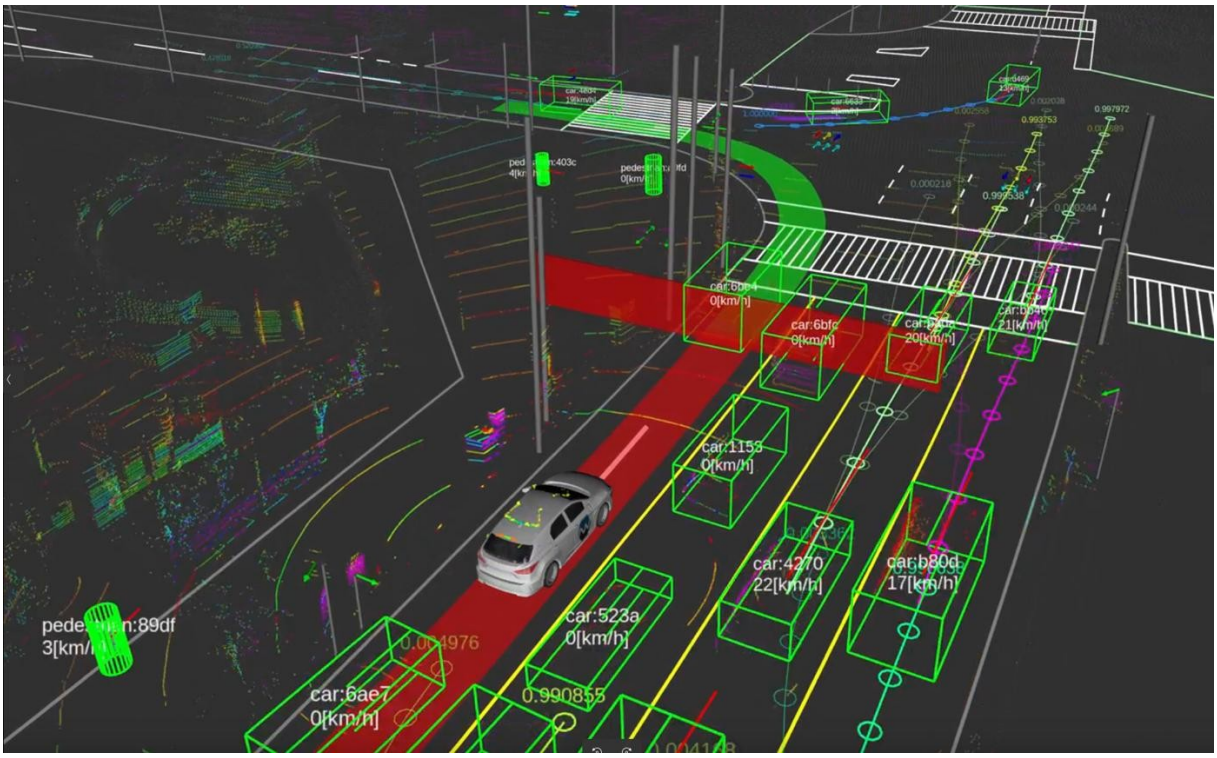
现实世界非常复杂。

主体不可能直接获得世界的全部信息，只能通过感官、传感器和已有知识获取部分信息。

例如，一个自动驾驶系统需要理解：

- 道路的位置。
- 车辆的位置。
- 行人的位置。
- 交通信号灯状态。
- 周围物体的运动趋势。
- 道路之间的空间关系。

它将这些信息组织起来，形成内部世界模型。



参考：autowarefoundation.github.io

可以将世界表征抽象为：

$$\mathrm{M_t = f(O_{0:t},\ K_t)}$$

其中：

- M_t：当前内部世界模型。
- $O_{0:t}$：从过去到现在获得的观察信息。
- K_t：已有知识与先验经验。
- f：信息整合和模型构建过程。

这里需要区分两个概念：

概念	含义
客观世界	实际存在的环境及其状态
主体表征	主体对客观世界形成的内部模型

两者并不完全相同。

主体看到的世界，不等于世界本身。

例如，同一个移动目标：

- 摄像头看到的是像素。
- 人类看到的是一个移动物体。
- 自动驾驶系统识别的是一个行人目标。
- 游戏 AI 识别的是一个对手角色。

它们面对的是同一个客观对象，但内部表征不同。

2.2 世界表征为什么重要?

如果没有世界表征，主体只能对眼前的信息作出直接反应。

有了世界表征，主体就能够：

1. 保存过去的经验。
2. 识别不同对象之间的关系。
3. 理解事物的状态变化。
4. 推断不可直接观察的状态。
5. 在内部模拟尚未发生的事件。

这意味着世界表征是智能从即时反应走向复杂认知的重要基础。

三、概率预测：智能如何理解未来？

3.1 未来不是一个确定的点

这是整个研究框架中非常重要的部分。

假设一个人在道路上行走。

当前观察到：

- 位置：$x=10$
- 速度：$v=2$
- 方向：向右

如果不考虑其他因素，我们可能预测：

$$x_{t+1} = x_t + v\Delta t$$

但现实世界并不如此简单。

这个人可能：

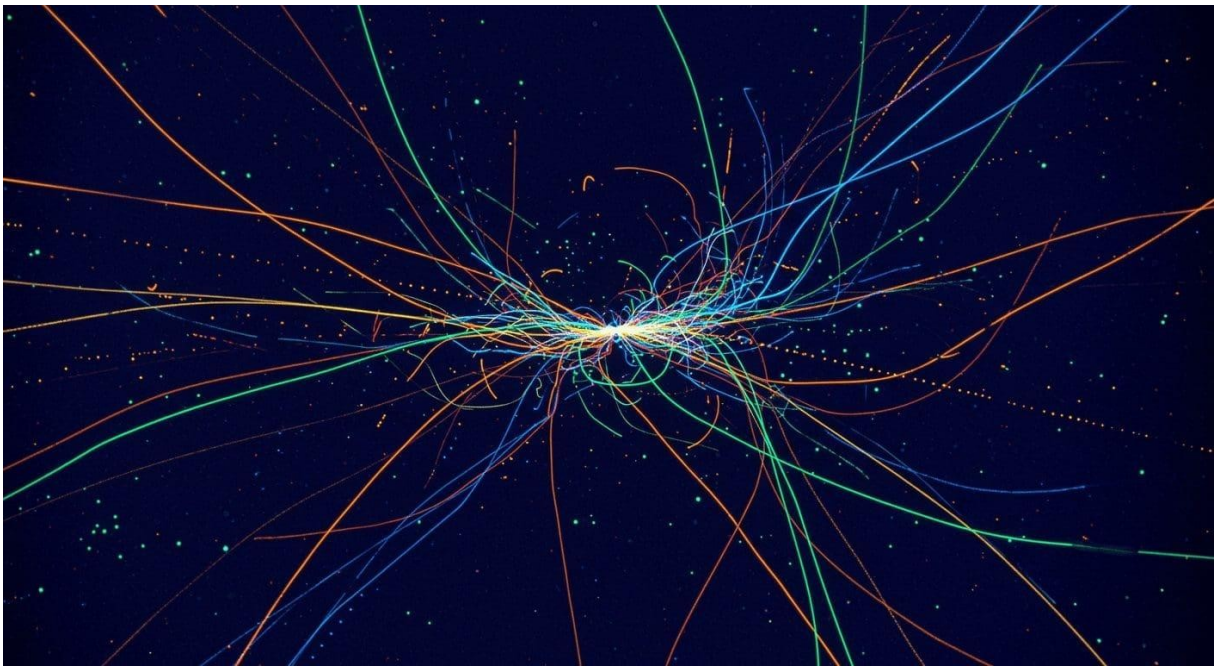
- 继续向右移动。
- 突然停下来。
- 转向左边。
- 加速。
- 被其他人阻挡。

因此，未来不是一个已经确定的点，而是多个可能状态构成的概率分布。

$$P(S_{t+1} \mid O_{0:t})$$

它表示：

在已经获得的观察条件下，未来状态可能出现的概率。



参考：quantumzeitgeist.com

3.2 从单一预测到概率云

传统预测模型经常输出一个结果：

$$\hat{S}_{t+1} = f(S_t)$$

例如：

目标下一秒的位置是：

$$(x, y) = (12, 5)$$

但是，这种表达可能隐藏了预测的不确定性。

更完整的表达应该是：

$$P(X_{t+1}, Y_{t+1} \mid O_{0:t})$$

例如：

目标未来位置的概率分布（概念示意）

示意图：概率云不是实际存在的一团物质，而是描述未来状态不确定性的数学表示。

概率云可以表达：

- 哪些未来状态更可能发生？
- 哪些未来状态不太可能发生？
- 不确定性有多大？
- 不同未来之间是否存在关联？
- 新的信息会如何改变概率分布？

因此，智能不应该只追求预测一个结果，还应该理解预测结果背后的不确定性。

3.3 概率预测的数学基础

贝叶斯推断提供了一种重要的数学工具：

$$P(H \mid E) = \frac{P(E \mid H)P(H)}{P(E)}$$

其中：

- H：某个假设。
- E：新获得的证据。

- $P(H)$：原有先验信念。
- $P(H \mid E)$：结合证据后的后验信念。

例如：

一个射手原本认为对手有 70% 的概率继续向右移动。

突然观察到对手转头、身体倾斜。

这些新证据会改变射手对未来运动方向的判断。

这就是认知系统根据新信息持续更新未来概率的过程。

智能不是简单地预测未来，而是持续修正自己对未来的信念。

四、未来生成：智能如何改变现实？

前面讨论了世界表征和概率预测。

但如果智能仅仅停留在预测层面，它仍然只是一个观察者。

真正重要的是第三个问题：

主体能否通过行动改变未来？

4.1 预测与行动之间的关系

假设一个自动驾驶系统预测：

- 继续直行，有 30% 的碰撞风险。
- 减速，有 5% 的碰撞风险。

- 向左转，有 10% 的碰撞风险。

系统需要根据目标和约束选择行动。

可以抽象为：

$$A_t^* = \arg\min_{A_t} \mathbb{E}[\, L(S_{t+1}, G) \mid M_t, A_t \,]$$

其中：

- A_t^*：最优行动。
- L：损失函数。
- S_{t+1}：行动后的未来状态。
- G：系统目标。

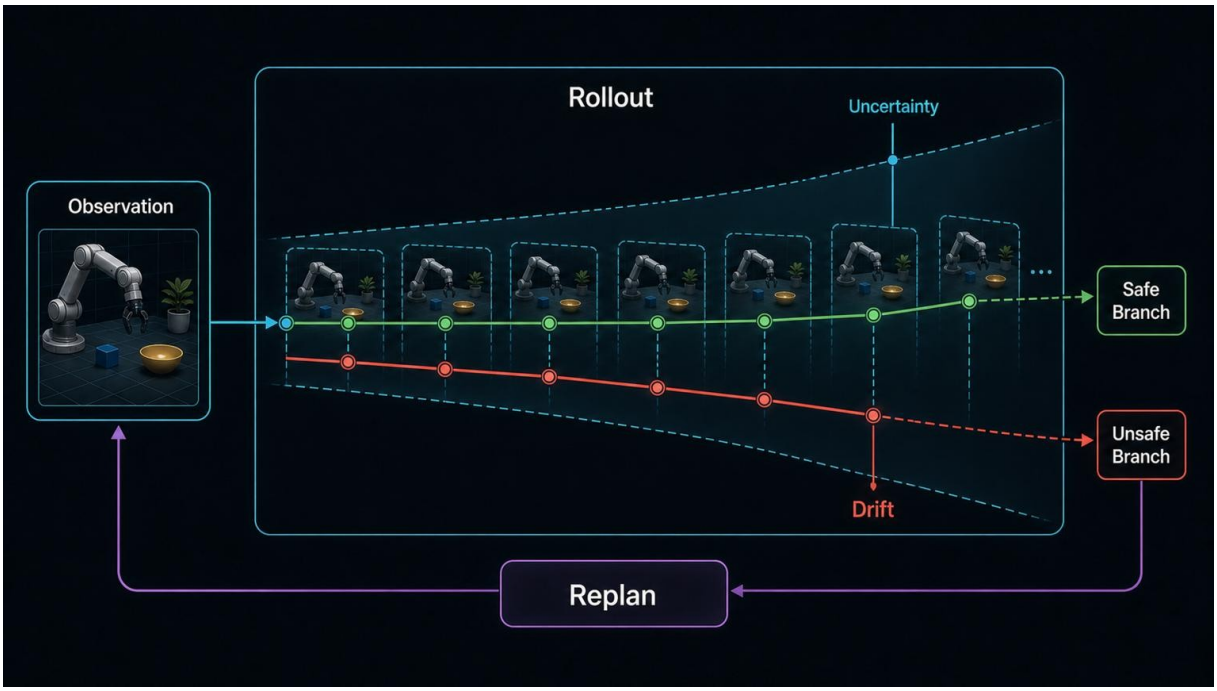
注意，这里的行动不仅仅是对未来作出反应。

行动本身会改变未来状态。

$$S_{t+1} = T(S_t, A_t, \varepsilon_t)$$

因此，未来不是完全独立于主体的外部事件。

在许多系统中，主体的行动也是未来形成机制的一部分。



参考：fusheng-ji.github.io

4.2 从预测未来到生成未来

我们可以区分三种能力：

能力	核心问题
状态识别	现在发生了什么？
未来预测	接下来可能发生什么？
未来生成	我采取什么行动，可以让某种未来发生？

例如：

一个 AI 系统预测某条生产线可能发生故障，这是预测能力。

它进一步调整生产速度、设备参数和维护计划，这是干预能力。

最终让生产线避免故障，则是未来状态被行动改变后的结果。

因此，可以提出一个重要研究命题：

智能不仅是预测未来的能力，也是利用预测结果，通过行动参与未来生成的能力。

这里的未来生成是指现实状态在行动与环境动态共同作用下的演化，并不意味着主体能够任意创造现实。

五、一个更复杂的问题：当两个智能体相互预测时

前面的例子假设只有一个主体。

但真实世界经常存在多个智能主体。

例如：

- 射手 A 与射手 B。
- 两个自动驾驶汽车。
- 两个竞争企业。
- 两个谈判者。
- 两个具有不同目标的 AI Agent。

这时，问题会发生根本变化。

A 不仅要预测 B 的行为，还需要预测：

B 如何预测 A?

进一步：

A 需要考虑 B 对 A 的预测，以及 B 是否会根据 A 的预测采取反制行动。

于是出现递归信念结构。



参考：vegavid.com

5.1 双主体递归预测

设：

- M_A：A 的内部世界模型。
- M_B：B 的内部世界模型。
- $P_A(B)$：A 对 B 行为的概率判断。
- $P_B(A)$：B 对 A 行为的概率判断。

那么：

$$A \to P_A(B)$$
$$B \to P_B(A)$$

但实际上，A 还可能需要预测 B 对自己的判断：

$$A \to P_A(B \to P_B(A))$$

这就形成了递归预测。

A 不再只是预测 B 的运动，而是在预测 B 的预测。

这与博弈论中的信念建模、认知层级模型和多智能体决策密切相关。

5.2 未来不再只是概率分布

当多个主体同时采取行动时，未来状态受到多个主体的共同影响：

$$S_{t+1} = T(S_t, A_t^1, A_t^2, \ldots, A_t^n, \varepsilon_t)$$

这意味着：

- A 的行动改变 B 的决策条件。
- B 的行动改变 A 的预测。
- A 和 B 的相互作用又改变整个环境。
- 环境变化进一步影响双方下一轮决策。

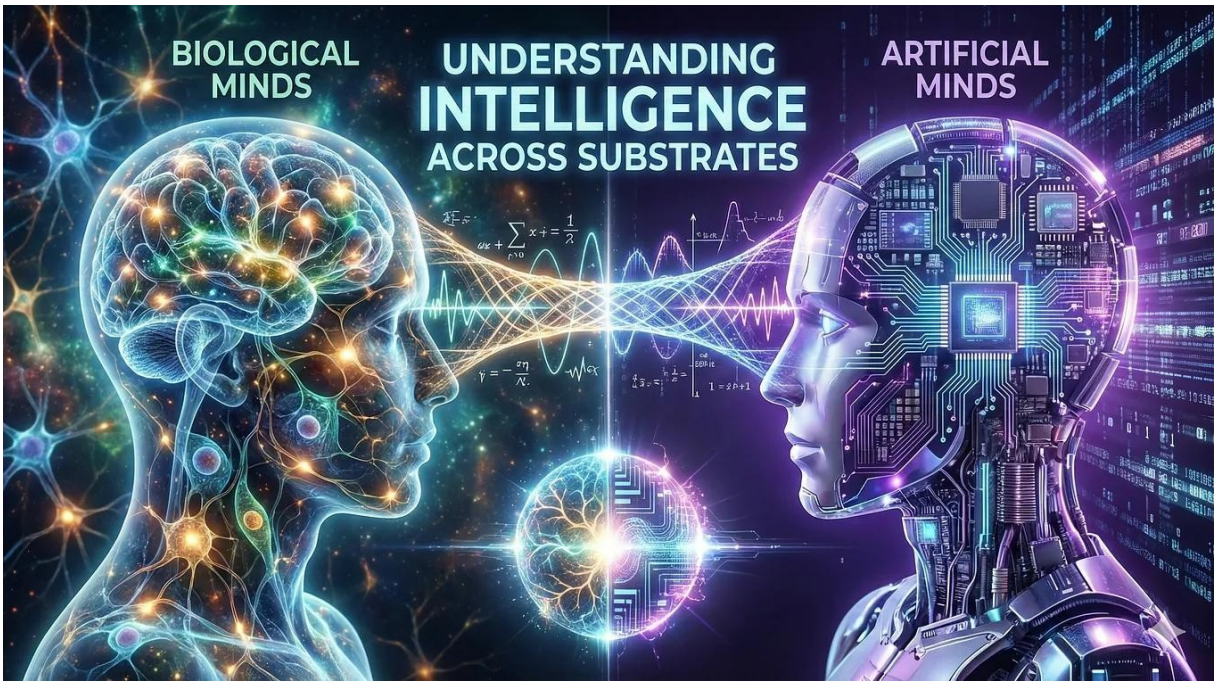
未来由此成为一个动态博弈过程。

这里有一个值得研究的观点：

在多智能体系统中，未来不仅是被预测的对象，也是多个主体通过相互预测和行动共同形成的结果。

六、从生命到人工智能：统一的智能架构

现在，我们可以尝试将上述思想应用于不同系统。



参考：medium.com

系统	世界表征	概率预测	未来生成
细菌	化学环境信号	环境变化趋势	调整运动方向
动物	感官与记忆模型	捕食者、猎物行为	逃跑、捕猎
人类	语言、概念和因果模型	社会与自然事件	规划、创造、改变环境
自动驾驶 AI	道路与交通模型	车辆和行人轨迹	控制车辆
AI Agent	数据、代码和任务模型	任务执行结果	调用工具、修改文件、完成任务

这些系统在复杂程度、认知能力和内部机制上差异很大。

但我们可以用相似的抽象框架研究它们：

Observe → Represent → Predict → Decide → Act → Learn

这是一种统一的研究视角，而不是说所有生命和人工智能都采用完全相同的实现机制。

七、如何将这个理论转化为人工智能工程?

如果要构建一个具备持续认知和行动能力的 AI Agent，不能仅仅把大语言模型作为全部智能系统。

需要建立更完整的架构。

面向未来生成的智能系统架构

感知层 Perception

文本、图像、声音、传感器、外部事件

世界模型 World Model

实体、状态、关系、因果、历史记忆

预测引擎

未来状态、概率分布、风险评估

目标引擎

目标、约束、偏好、价值函数

决策与规划引擎

情景模拟、行动选择、计划生成

执行引擎 Action Engine

API、工具、代码、机器人、外部系统

反馈与学习 Evaluation & Learning

结果验证、误差分析、模型更新、记忆沉淀

持续循环，动态更新世界模型

这套架构可以进一步拆成七个工程模块：

模块	技术方向
感知引擎	多模态模型、计算机视觉、语音识别
世界模型	Knowledge Graph、状态数据库、时序模型
概率预测	Bayesian Inference、概率模型、时序预测
目标引擎	Reward Model、约束优化
决策引擎	Planning、MCTS、强化学习
执行引擎	Agent Runtime、Tool Calling、API
反馈引擎	Evaluation、Online Learning、Memory

这里尤其需要强调：

LLM 只是智能系统可能使用的一种认知和推理组件，并不等于整个智能系统。

一个完整的智能系统还需要可靠的环境状态、预测能力、行动接口、反馈机制以及安全约束。

八、这个研究框架有哪些理论基础？

以下研究方向可以为这套框架提供理论支撑。



参考：wired.com

1. Karl Friston：自由能原理与主动推断

研究生物系统如何通过感知和行动维持自身状态，以及如何减少预测误差。

关键词：Free Energy Principle、Active Inference。

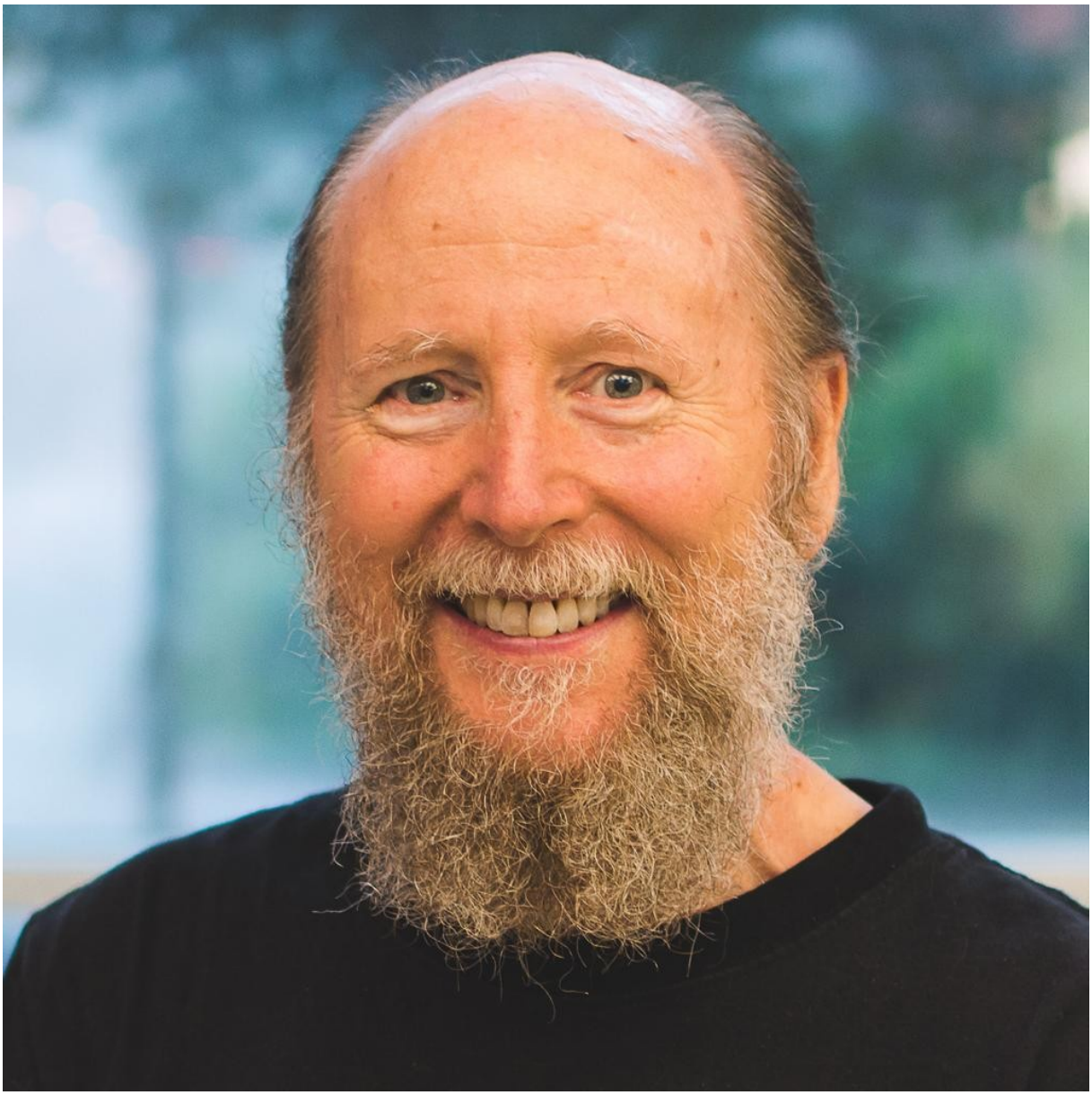


参考：nsf.gov

2. Judea Pearl：因果推断

区分观察、干预和反事实推理，帮助理解主体如何通过行动改变世界。

关键词：Causal Inference、Do-Calculus。



参考：amii.ca

3. Richard Sutton：强化学习

研究智能体如何通过行动、奖励和环境反馈学习策略。

关键词：Reinforcement Learning、Reward、Policy。



参考：computerhistory.org

4. Herbert A. Simon：有限理性

研究真实主体如何在有限信息、有限计算资源和有限时间下作出决策。

关键词：Bounded Rationality、Decision Making。



5. John von Neumann：博弈论

研究多个具有不同目标的决策主体如何相互影响。

关键词：Game Theory、Strategic Interaction。

建议重点阅读：

- The Predictive Mind：预测加工与认知。
- Surfing Uncertainty：预测加工理论。
- Reinforcement Learning: An Introduction：强化学习。
- The Book of Why：因果推断。

- Active Inference 相关研究论文：主动推断。

九、进一步提出五个核心研究问题

如果把这套框架作为一个长期研究方向，我认为有五个问题值得深入。

问题一：世界表征是否等于真实世界？

主体只能通过有限观察认识世界。那么，内部世界模型与客观世界之间究竟存在什么关系？

这涉及认识论、信息论和科学实在论。

问题二：概率云究竟来自哪里？

未来的不确定性究竟是客观世界本身的属性，还是主体知识不足造成的？

这里需要区分：

- 客观随机性。
- 主体认知不确定性。
- 模型不确定性。
- 多主体交互产生的不确定性。

问题三：预测是否会改变被预测对象？

当主体知道自己正在被预测时，它是否会改变行为？

例如，A 预测 B 会向右移动，B 发现 A 正在瞄准右侧，于是转向左边。

预测本身可能成为改变行为的因素。

问题四：行动如何改变概率分布？

传统预测往往关注：

$$P(S_{t+1} \mid S_t)$$

而智能主体还需要考虑：

$$P(S_{t+1} \mid S_t, A_t)$$

也就是说，不同的行动会产生不同的未来概率分布。

问题五：智能是否能够创造新的可能性空间？

普通预测是在已有模型中寻找可能结果。

但创新可能意味着创造原有模型没有包含的状态、概念、工具和行动方式。

这涉及创造力、模型扩展和开放式学习。

十、最终观点：智能是一种面向未来的动态能力

我们可以把整篇文章的观点浓缩为一个结构：

智能的三层本质

01 世界表征：认识现实

建立关于世界的内部模型，理解对象、状态、关系和规律。

02 概率预测：理解可能性

通过内部模型推演未来，建立可能状态及其概率分布。

03 未来生成：干预现实

根据目标选择行动，通过主体与环境的交互改变未来状态。

智能的动态循环

Reality → Representation → Prediction

→ Decision → Action

→ New Reality → Learning

我认为，这个框架最有价值的地方，是把原本分散在认知科学、概率论、控制论、因果推断和人工智能中的问题联系起来。

它还揭示了一个值得进一步探索的方向：

智能的核心可能不在于主体知道多少信息，而在于主体能否利用有限的信息构建有效的世界模型，理解未来的不确定性，并通过行动持续改变自身与世界的关系。

不过，这仍然是一种研究框架，而不是已经被科学界证明的智能统一理论。