PART1：**传统方式-强化学习（Reinforcement Learning, RL）：**

- 智能体与环境交互
- 通过动作获得奖励，并学习最优策略
- 没有直接的“正确答案”，只有“反馈信号”

强化学习 = 状态 + 动作 + 奖励 + 策略

概念：强化学习是一种让智能体（agent）通过与环境交互、试错探索，从而学会在特定情境下采取最优行为的机器学习方法。

组成	含义
Agent（智能体）	学习策略并采取行动的主体
Environment（环境）	智能体所处的世界，反馈结果
State（状态）	当前环境的描述信息
Action（动作）	智能体可选的操作
Reward（奖励）	行动带来的即时反馈

应用场景：

应用领域	举例
游戏	AlphaGo、Dota2、Atari 游戏
机器人控制	行走、抓取、避障
金融策略	强化学习交易、风险管理
推荐系统	学习用户长期偏好
自动驾驶	路况决策、避障规划

常见算法：

类别	代表算法	特点
基于值（Value-based）	Q-Learning, DQN	学每个动作的“价值”
基于策略（Policy-based）	REINFORCE, PPO	直接学行动策略
Actor-Critic	A2C, PPO, DDPG	同时学习“值”和“策略”，效果更好
模型自由/基于模型	Model-free / Model-based	是否模拟环境转移过程

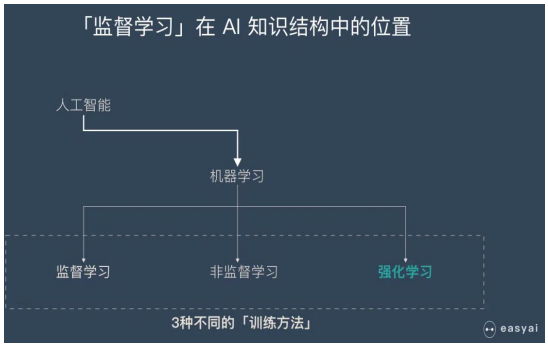
RLHF（Reinforcement Learning with Human Feedback）：

流程如下：

1. 让模型生成多个回答
2. 人类对这些回答进行排序（比如：好、中、差）
3. 用这个排序数据训练一个**奖励模型**（可以打分）
4. 用**强化学习算法（如 PPO）**调整 ChatGPT 的行为，让它输出更高分的回答

Reward Hacking:
在强化学习中，因为 reward function 设置不当，导致 agent 只关心累计奖励，而无法完成研究人员预想的目标。
当 强化学习 (RL) AI 智能体 利用 奖励函数中的缺陷或歧义来获得高奖励，而没有真正学习或完成预期的任务时，就会发生 reward hacking。

强化学习概述:



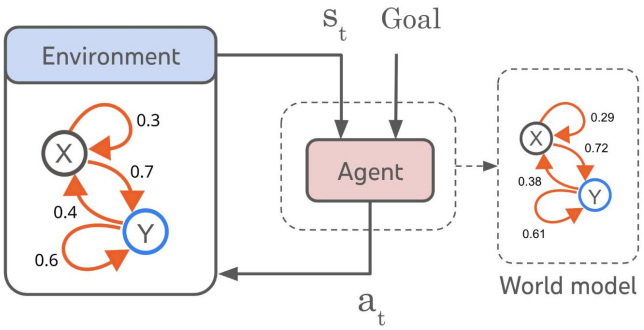
Yann LeCun 在 2016 年的演讲上曾比喻：如果把智能比作一块蛋糕，那么无监督学习就是蛋糕的主体，监督学习就是蛋糕上的糖霜，而强化学习则是糖霜上的樱桃。我们已经知道如何制作糖霜和樱桃，但却不知道如何制作蛋糕本身。LeCun 本人从 2016 年起并不看好强化学习。

PART2: Reinforcement Pre-Training 新范式：
微软提出的 Reinforcement Pre-Training，提出了一种强化学习预训练（RPT），在这种范式中，通过强化学习的推理任务进行下一个 Token 的预测。模型会在预测正确下一个 Token 时获得奖励。这就好比在制作蛋糕的过程中，直接将樱桃融入到蛋糕的主体结构中。scaling 曲线表明，随着训练计算量的增加，下一个 token 预测的准确性持续提升。这些结果表明，RPT 是一种有效且有前景的 scaling 范式，能够推动语言模型预训练的发展。

摘录文章：https://mp.weixin.qq.com/s/UABVUoHYTDlFWWNvD5R9Og
论文标题：Reinforcement Pre-Training
论文链接：https://www.arxiv.org/pdf/2506.08007

PART3: General agents need world models 新范式：
2023 年 3 月，OpenAI 联合创始人 Ilya Sutskever 提出了一个深刻的论断：神经网络学习的不仅仅是文本信息，而是我们这个世界的一种压缩表征。因此，我们预测下一个词的准确度越高，世界模型的保真度就越高。大型神经网络的功能远不止预测下一个单词，它实际上是在学习「世界模型」。

任何一个能够归纳各种简单目标任务的 Agent 都必须学习对其所在环境的预测模型，并且这个模型可以随时从代理中恢复。



构建通用人工智能 Agent 没有“无模型捷径”，如果我们想要 Agent 可以泛化到各种任务，就必须学习世界模型。更好的性能表现需要更好的世界模型支撑。降低错误或者处理更复杂的目标的唯一途径就是学习越来越准确的世界模型。

Agent 学习的四个关键的组成部分：环境、目标、智能体、世界模型。

对 Agent 而言，学习足够通用的 goal-conditioned 策略在信息上等同于学习准确的世界模型。

这种方式适用于需要通过多轮步骤规划来了解行为（actions）是如何影响未来状态（future state）的 Agent，只考虑即使奖励的 Agent 可能会避免学习时间模型，因为这类 Agent 不需要预测长期结果。

整体来讲，我们可能正在见证一些更为深刻的变革：从 David Silver 和 Richard Sutton 所说的「Human Data 时代」向「Experience 时代」的转变。虽然当前的人工智能系统通过模仿人类生成的数据实现了非凡的能力，但 Silver 和 Sutton 认为，超人类智能将主要通过智能体从自身经验中学习而诞生。

摘录文章：

https://mp.weixin.qq.com/s/k-hd-M1XK7fsH2LI80r5AA

https://richardcsuwandi.github.io/blog/2025/agents-world-models/

论文标题：General agents need world models

论文链接：https://arxiv.org/abs/2506.01622