《人工智能通识》（科技素养）

# 第5讲 强化学习

主讲：丛润民

# 章节知识点概览

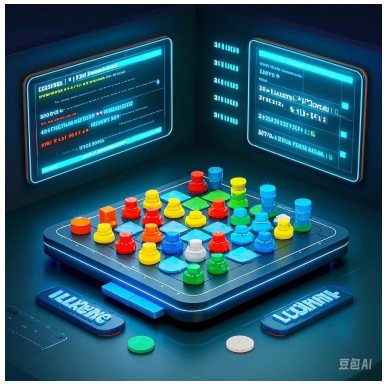
知识点1：马尔可夫决策过程——以井字棋游戏为例

知识点2：解密贝尔曼方程

知识点3：当深度学习遇到强化学习

知识点4：智能机器人与强化学习

# 知识点1：
# 马尔可夫决策过程 — 以井字棋为例

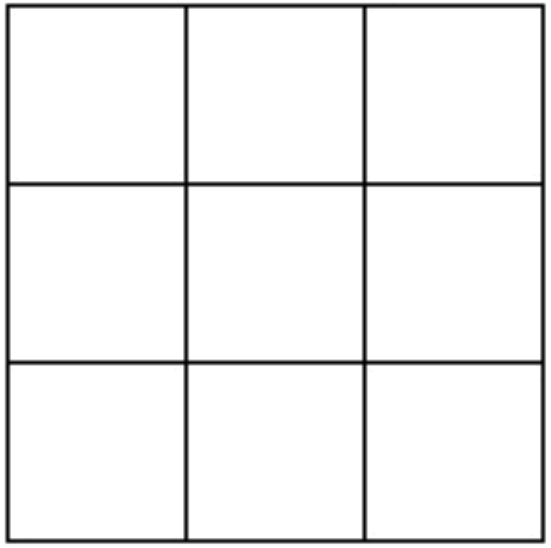


01 马尔可夫性质

02 马尔可夫决策过程

# 一起来玩井字棋！

井字棋，英文名为 Tic-Tac-Toe，是一种在 3×3 格子上进行的两人对弈游戏。游戏双方轮流在空格中放置自己的棋子。先手玩家先放置“X”，后手玩家放置“O”。率先将自己的三个棋子连成一条直线的玩家获胜。如果棋盘被填满后，双方都没有达成三子连线，则游戏为平局。



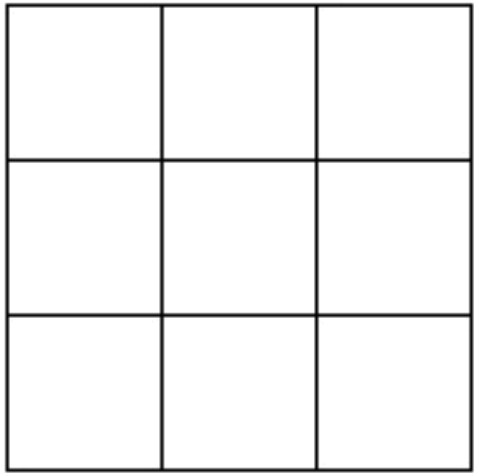
- 每个格子有三种选择，分别为：X、O和空，每个选择都会对结果产生影响
- **只需要关注当前棋盘上棋子的分布状态**，而不需要考虑之前是如何走到这个状态的

# 马尔可夫性质

- 给定当前状态，一个系统的**未来状态**只由**当前状态**和**当前行动**决定，与过去的**历史状态**无关 ⇒ **马尔可夫性质**
- 马尔可夫链：一个满足马尔可夫性质的随机过程，即给定当前状态的条件下，未来状态的概率分布仅依赖于当前状态，而与过去状态无关。

名人堂

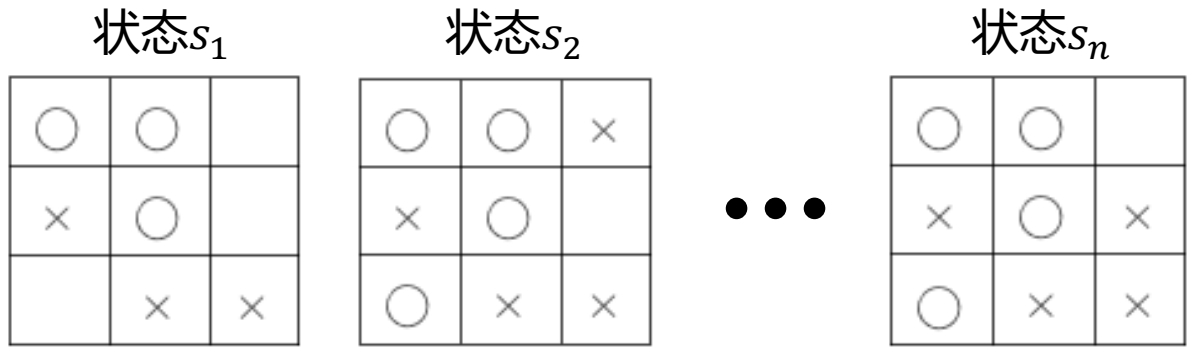
安德雷·安德耶维齐·马尔可夫（1856年6月14日—1922年7月20日），俄国数学家，师从切比雪夫，1886年当选为圣彼得堡科学院院士。马尔可夫1922年逝世于圣彼得堡。著名的马尔可夫决策过程的得名正是为纪念他为马尔可夫链所做的研究。



- ✓ **无记忆性**
- ✓ **状态转移过程只与当前状态和行动有关**

# 马尔可夫决策过程

在井字棋游戏中，我们可以作如下定义：

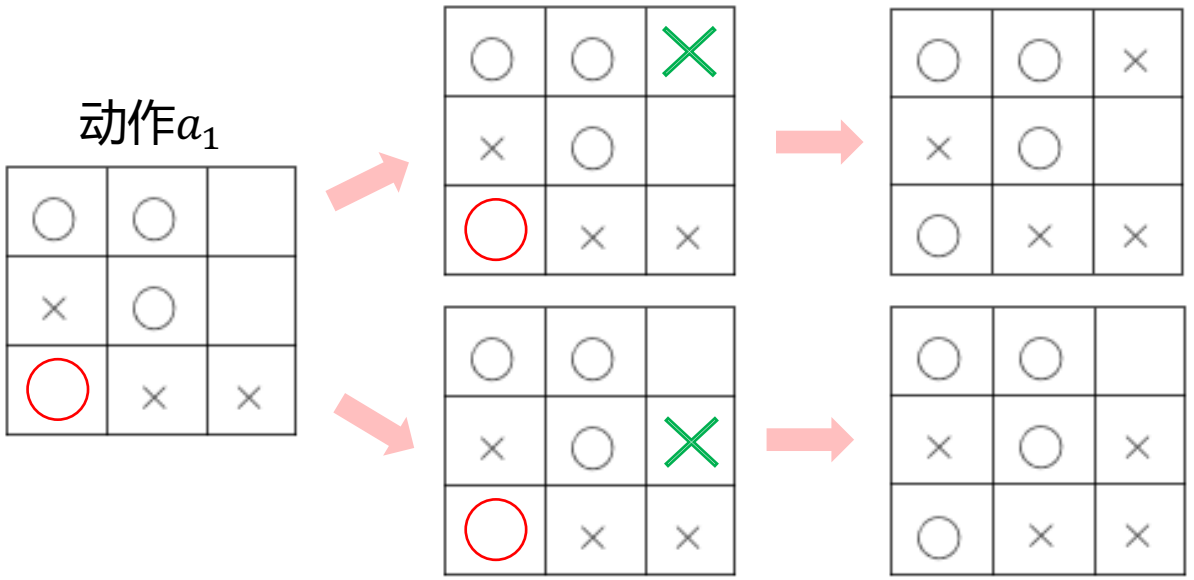



状态空间（State Space，$S$）：

所有落子之前的棋盘分布

动作空间（Action Space，$A$）：

棋手落子的9个位置



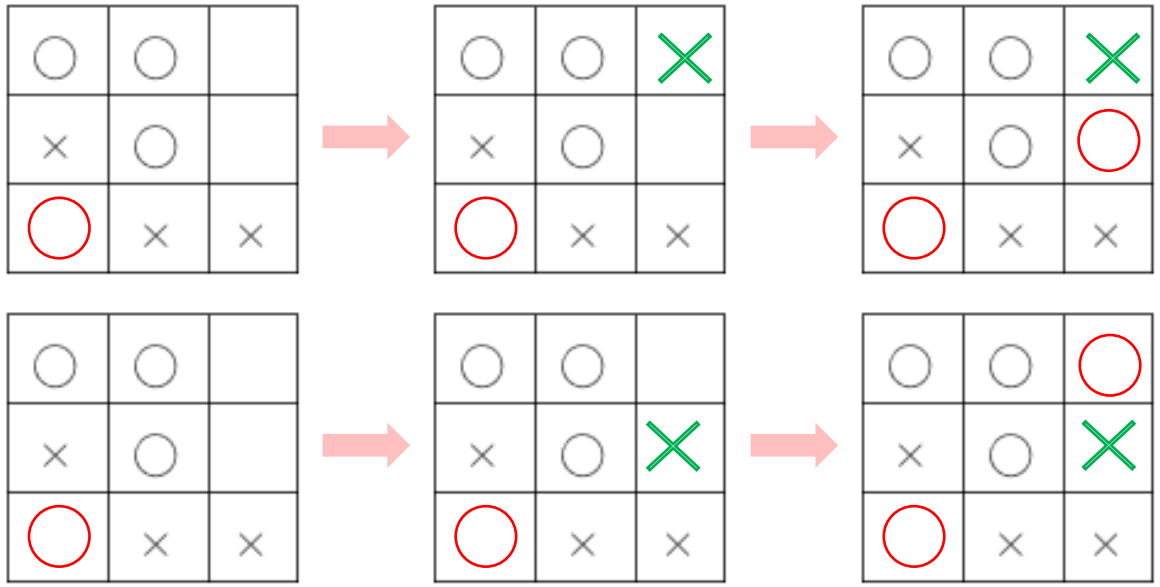

状态转移概率

（Transition Probability，$P$）：

棋手○落子后，棋手×落子位置使棋盘变为新的状态的可能性

# 马尔可夫决策过程

为了增加游戏的趣味性，我们还可以设置奖励和惩罚机制：

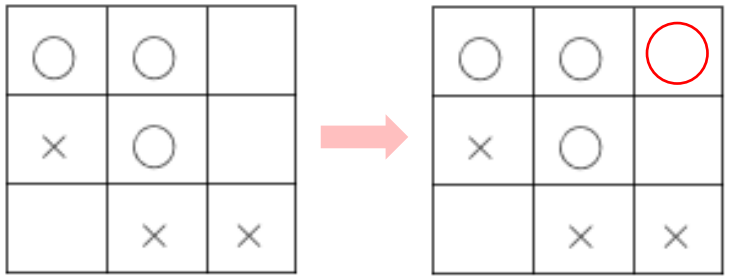


平局：奖励（Reward）$R=0$

获胜：奖励$R=+100$ $R=0.9^4\times100$

失败：奖励$R=-100$

事实上，我们还可以有更快的获胜方式：

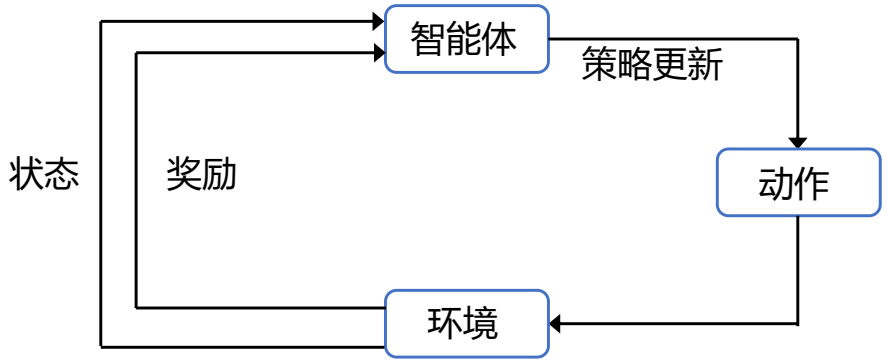


衰减因子$\gamma$：步数越少，奖励越高 $R=0.9^3\times100$

**以最大化游戏奖励为目标，以最小回合数获胜的方式，被称为策略（Policy）**

# 马尔可夫决策过程

通过井字棋，我们可以定义一个马尔可夫决策过程的5个关键组成部分：




1) **状态空间$S$**：所有环境状态的集合，$s_t \in S$表示$t$时刻智能体所处的状态。

2) **动作空间$A$**：智能体可执行动作的集合，$a_t \in A$表示$t$时刻智能体所执行的动作。

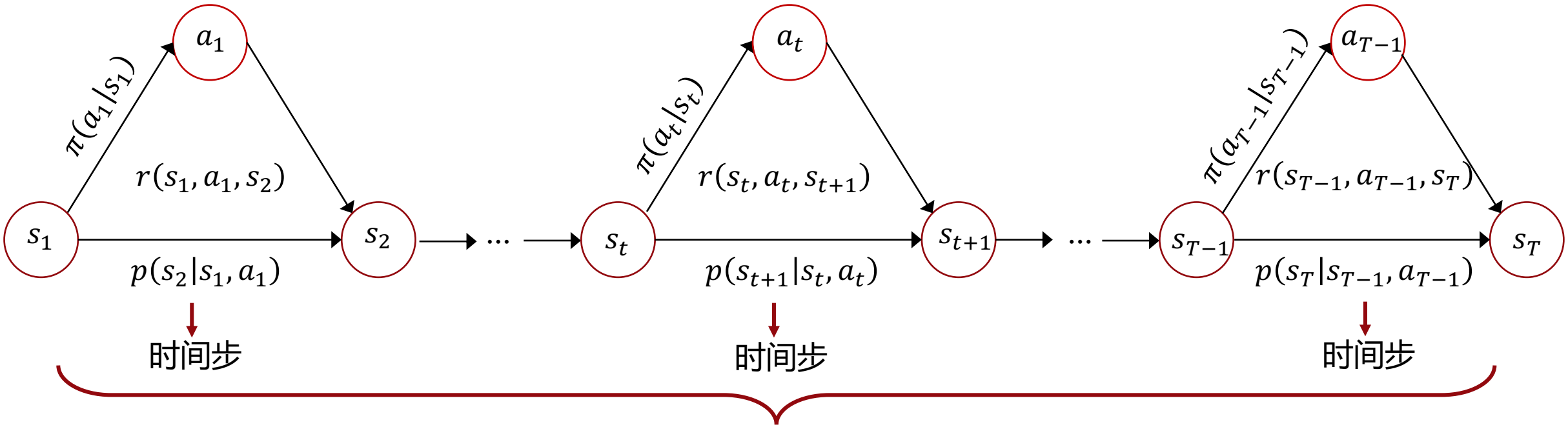
3) **转移概率$P$**：表示在执行某个动作后，从当前状态转移到另一个状态的概率。$p(s_{t+1}|s_t, a_t)$为状态转移概率，表示智能体在状态$s_t$执行动作$a_t$后转移到下一个状态$s_{t+1}$的概率。

4) **奖励函数$R$**：智能体在某个状态下执行某个动作后获得的即时奖励，用$r(s_t, a_t, s_{t+1})$表示。

5) **策略$\pi$**：它是一个从状态空间到行动空间的映射，定义了在每个状态$s$下应该采取哪个行动$a$。

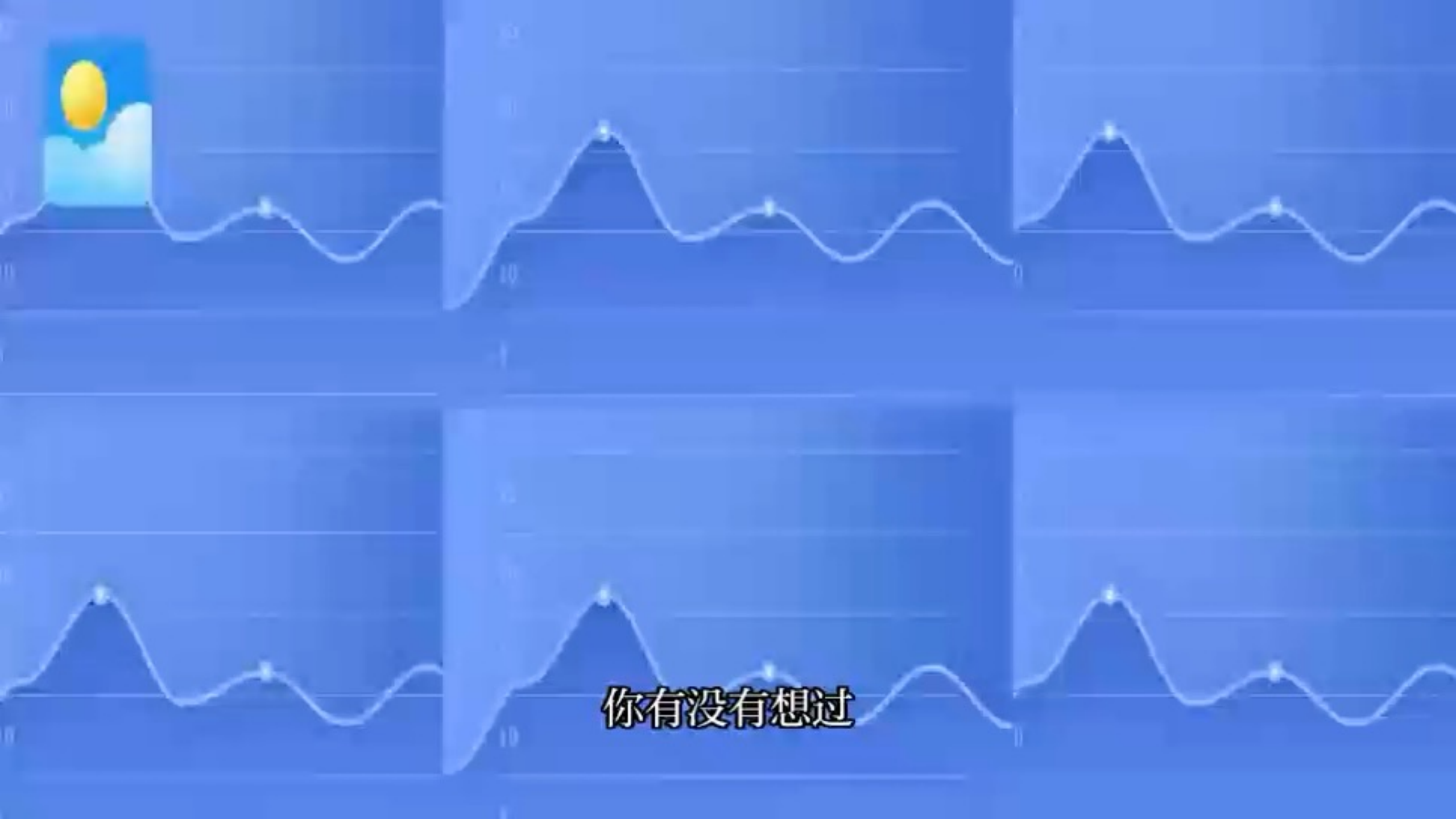
# 马尔可夫决策过程

智能体从起始状态$s_1$开始，根据策略$\pi$执行动作$a_1$，获得奖励$r_1$，并将状态转移到$s_2$，循环往复，直至到达终止状态（或满足某种终止条件）。

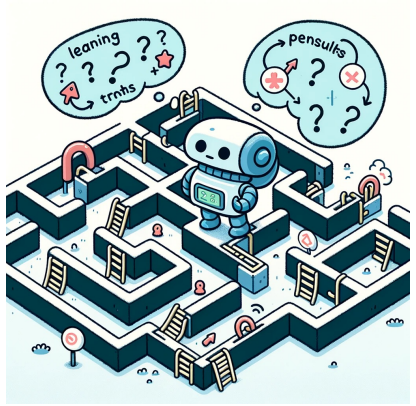



进行一次完整的交互过程称之为回合。

**在这些时间步中，智能体会根据当前状态选择动作，执行动作后收到环境的反馈（即时奖励和下一个状态），直到回合结束。**


你有没有想过

# 知识点2：解密贝尔曼方程




01 状态价值与动作价值

02 贝尔曼方程

# 从迷宫游戏看状态价值与动作价值

我们定义一个 3×3 迷宫游戏，包括起点（Start）、终点（End）、障碍物（#）与可通行区域（1）。游戏的任务是通过上下左右移动，从当前的某一位置最快到达终点。

| | | |
|---|---|---|
| Start | | |
| # | | |
| | End | |

**状态空间$S$**：所有可能的位置。我们用坐标表示状态：起点$(1,1)$、终点$(3,2)$

**动作空间$A$**：智能体在当前状态可能执行的所有动作（如上、下、左、右）

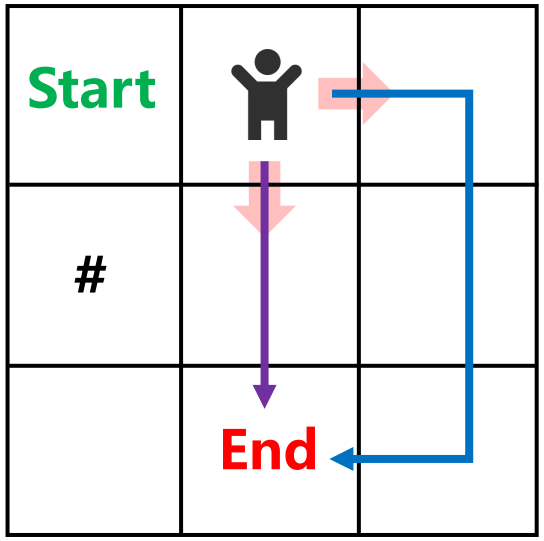
**转移概率$P$**：智能体从当前状态，通过动作$a$（如前进、后退、左转、右转）转移到下一状态的概率

**奖励函数$R$**：遇到障碍物奖励为-100，每移动到相邻格子奖励为-1，到达出口奖励为10，完成任务

**策略$\pi$**：以最大化奖励的方式，最快到达出口路径

**核心目标是找到一个策略 $\pi(a|s)$，使得该策略能够最大化奖励，完成任务。**

# 从迷宫游戏看状态价值与动作价值




以状态 (1,2) 为例，它可以向右 (1,3) 和向下 (2,2) 移动。

若向右移动，至少需要4步才能到达终点。则在这种情况下的最大奖励为：

$$10 + (-1)\times 4 = 6$$

若向下移动，需要至少2步才能到达终点。则在这种情况下的最大奖励为：

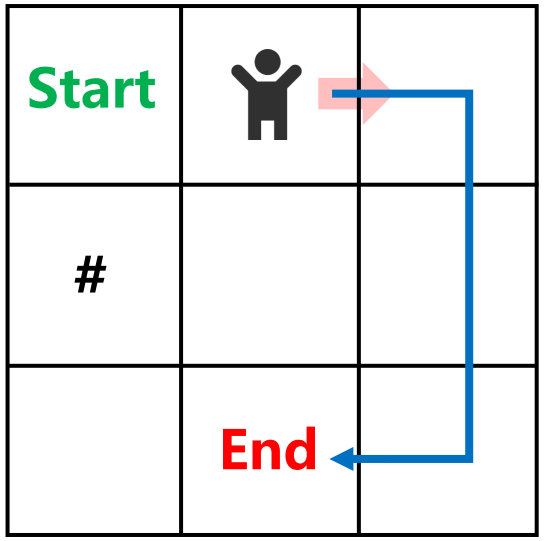
$$10 + (-1)\times 2 = 8$$

选择较大值 8 作为从状态 (1,2) 出发按照最优策略行动的期望奖励：

$$V(1,2) = 8$$


**状态价值函数**（State Value Function）它表示在给定一个状态下，按照某个策略行动所能获得的预期奖励。换句话说，状态价值函数衡量了在某个状态下，遵循某个策略时，能够从该状态开始所获得的期望累计奖励。

# 从迷宫游戏看状态价值与动作价值




以状态$(1,2)$为例，如果采取“向右”的动作，我们还可以计算：

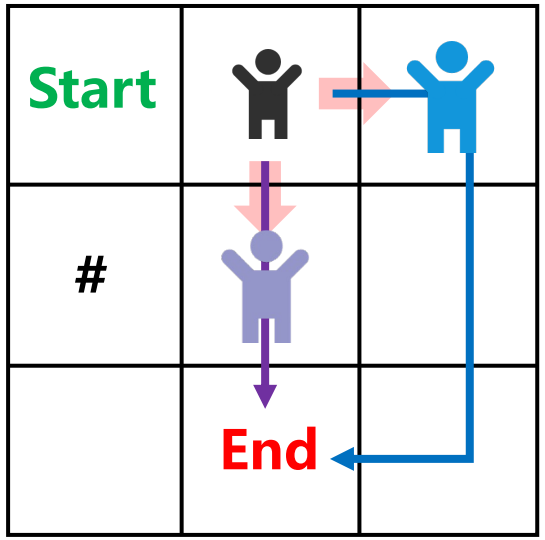
- 即时奖励$R(1,2)$：右移一步$-1$，到达$V(1,3)$
- $V(1,3)$状态价值为$10+(-1)\times 3 = 7$

于是，我们采取“向右”动作的价值为：

$$Q(1,2) = R(1,2) + \gamma V(1,3) = -1 + 0.9\times 7 = 5.3$$


**动作价值函数**（Action Value Function）表示在当前状态下，执行某个动作后，按照最优策略继续行动所获得的期望累积奖励。

# 贝尔曼状态方程




以状态$(1,2)$为例，有“向右”和“向下”两个可行动作

- 采取“向右”动作，到达状态$(1,3)$，转移概率$P((1,3)|(1,2),右)=1$，
- $V(1,3)$状态价值为$10+(-1)\times 3=7$

**动作价值：** $R(1,2)+\gamma P((1,3)|(1,2),右)\ V(1,3)=-1+0.9\times 1\times 7=5.3$

- 采取“向下”动作，到达状态$(2,2)$，转移概率$P((2,2)|(1,2),下)=1$，
- $V(2,2)$ 状态价值为$10+(-1)\times 1=9$

**动作价值：** $R(1,2)+\gamma P((2,2)\,|\,(1,2),下)\ V(2,2)=-1+0.9\times 1\times 9=7.1$

⋮

# 贝尔曼状态方程


**贝尔曼状态期望方程**的核心思想是将当前状态的价值表示为即时奖励与下一状态价值的期望之和。 $V^{\pi}(s)=\sum_{a\in A}\pi(a|s)[R(s,a)+\gamma\sum_{s'\in S}P(s'|s,a)\,V^{\pi}(s')]$

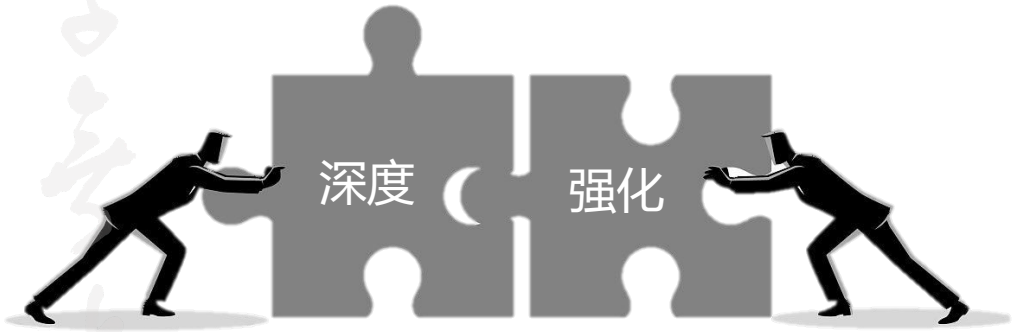
**贝尔曼状态最优方程**的核心思想是找到每个状态的最优价值，从而确定从任意状态出发的最优策略，即如何行动才能获得最大的累积奖励。

$$V^{*}(s)=max_{a}\sum_{s'\in S}P(s'|s,a)\,[R(s,a)+\gamma V^{*}(s')]$$

名人堂

理查德·贝尔曼（英文：Richard Bellman，1920年8月26日——1984年3月19日），美国数学家，动态规划的创始人。贝尔曼先后在布鲁克林学院和威斯康星大学学习数学。随后他在洛斯·阿拉莫斯为一个理论物理部门的团体工作。与1946年获得普林斯顿大学博士学位。贝尔曼曾是南加州大学教授，美国艺术与科学研究院研究员（1975年），美国国家工程院院士（1977年），美国国家科学院院士（1983年）。他在1979年被授予电气电子工程师协会奖，由于其在“决策过程和控制系统理论方面的贡献，特别是动态规划的发明和应用。”

# 知识点3：当深度学习遇到强化学习

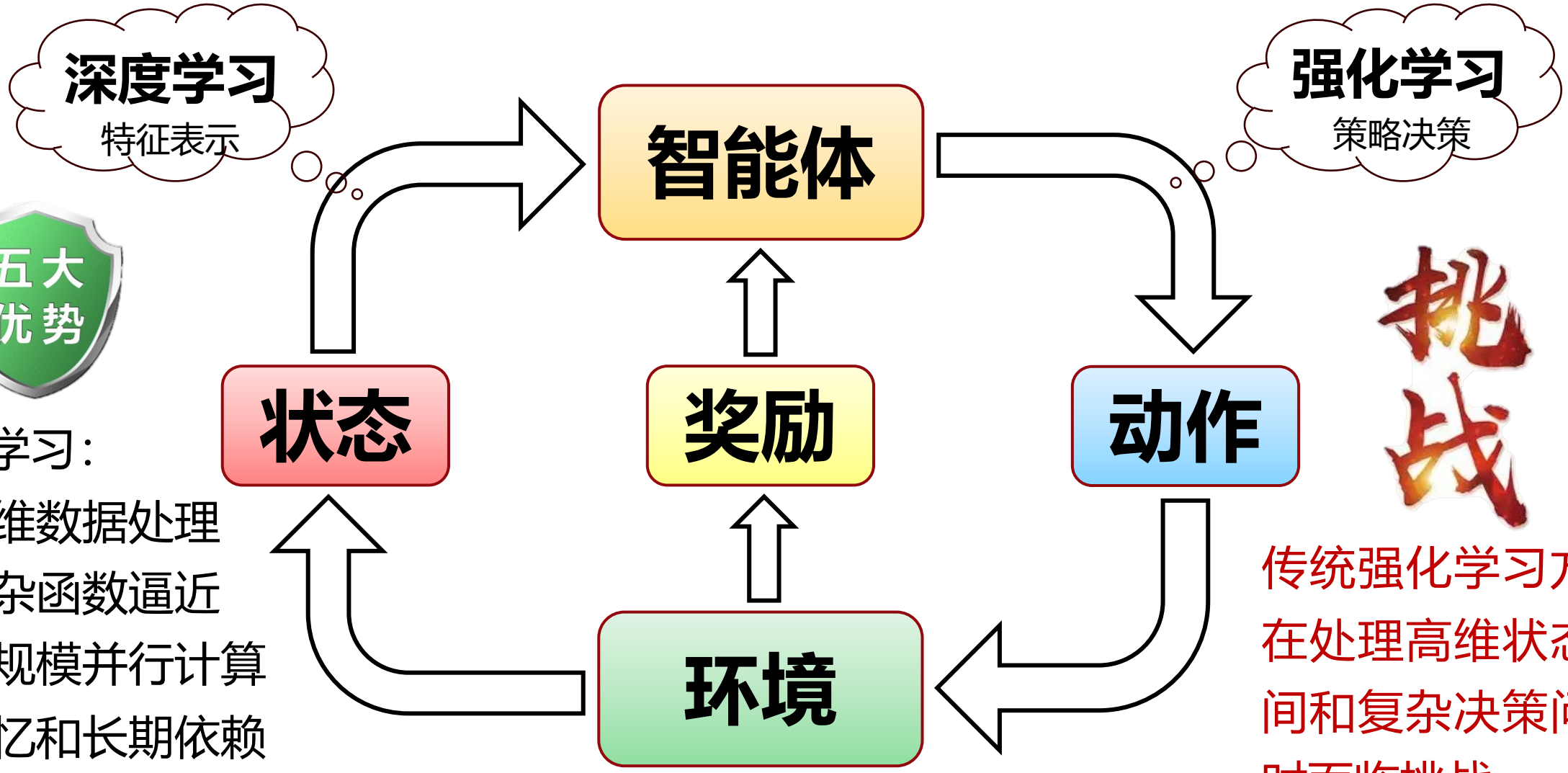



01 深度强化学习概念

02 深度强化学习方法


C:/FILE 26.06.1998
MY VLOGGING LIFE

# 什么是深度强化学习




五大优势

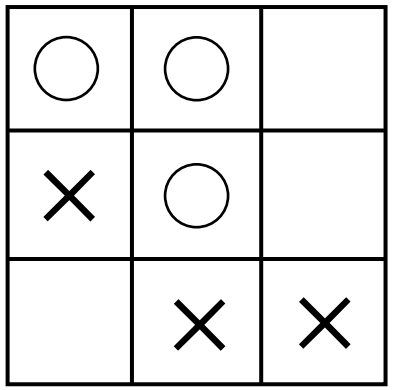
深度学习：

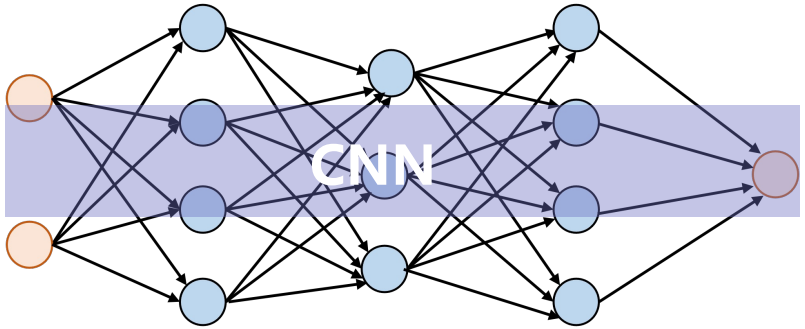
- ✓ 高维数据处理
- ✓ 复杂函数逼近
- ✓ 大规模并行计算
- ✓ 记忆和长期依赖
- ✓ 泛化能力

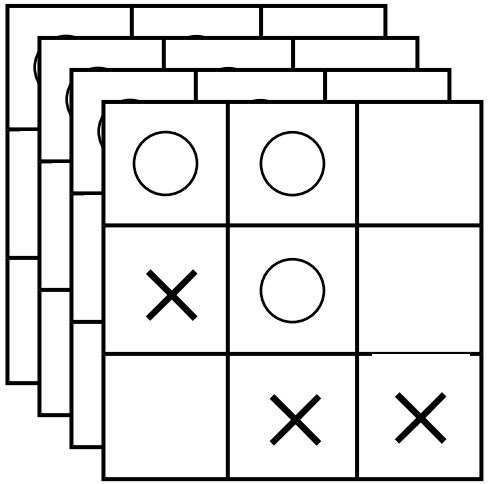
挑战

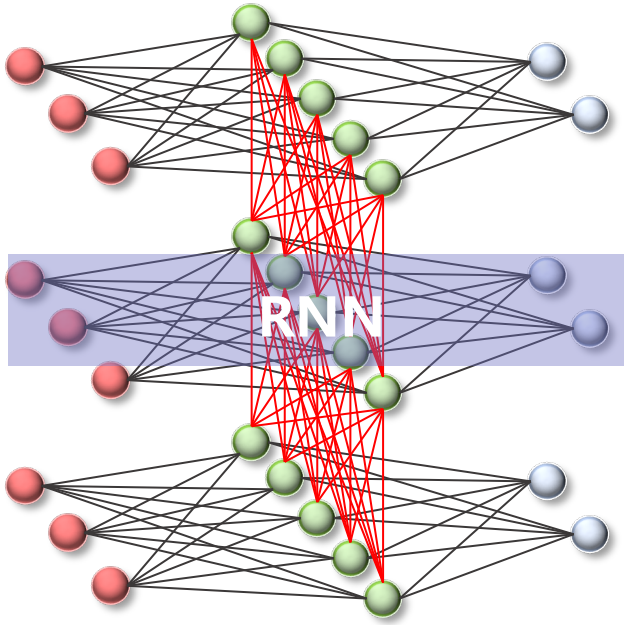
传统强化学习方法在处理高维状态空间和复杂决策问题时面临挑战。

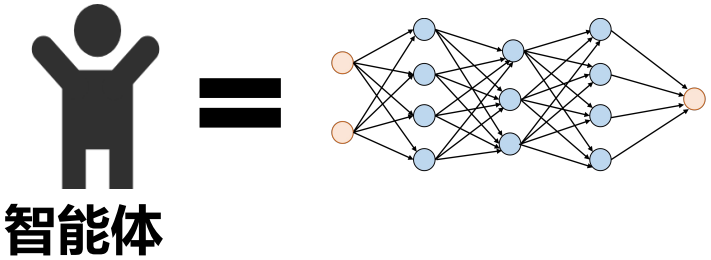
# 使用神经网络构建强化学习主体的方法被称为深度强化学习

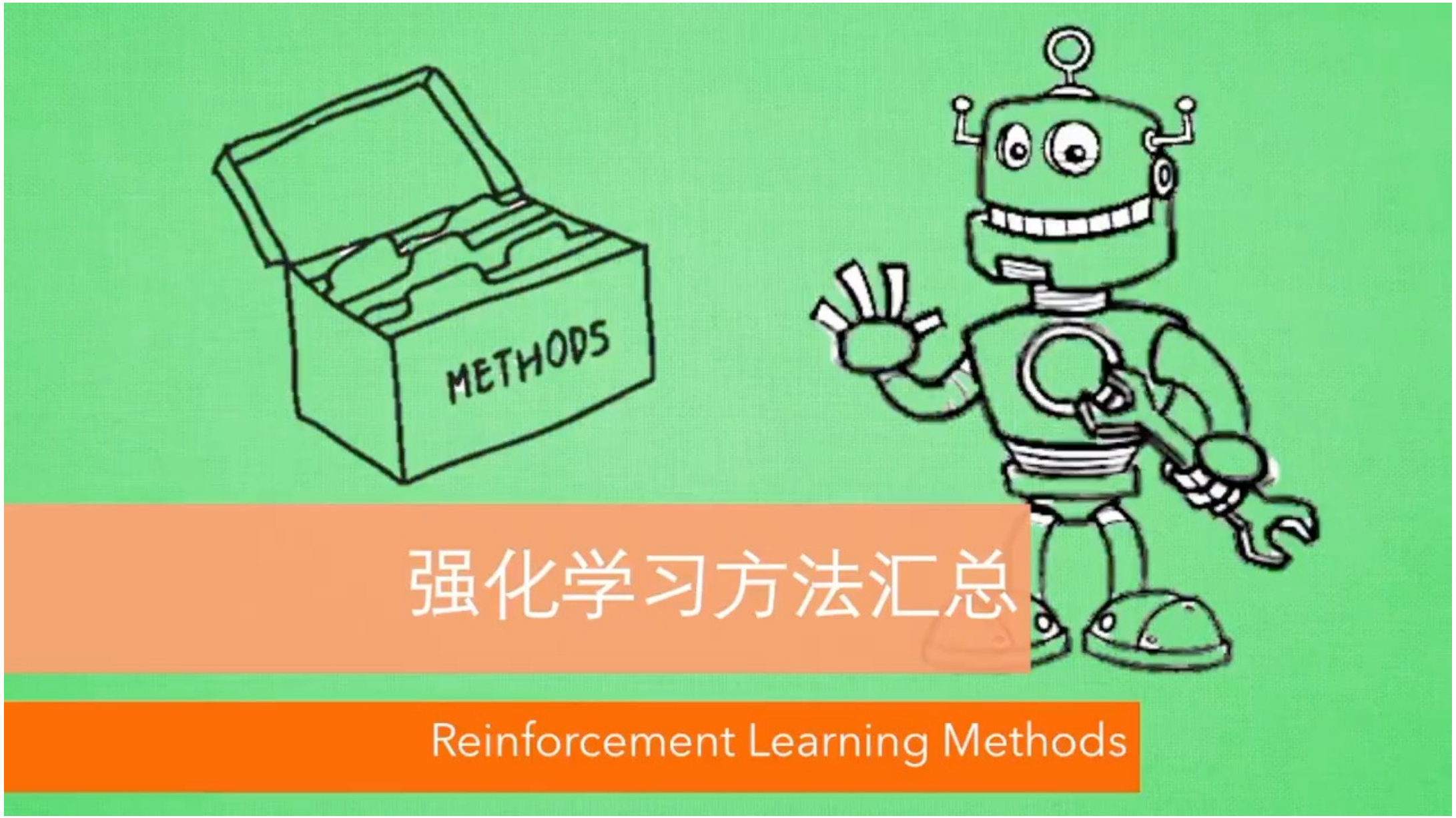






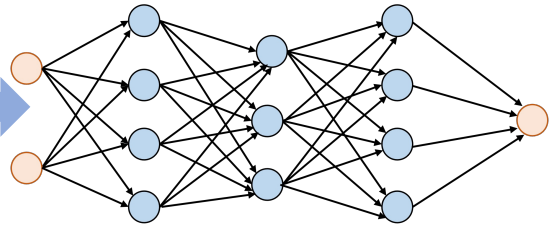




METHODS
强化学习方法汇总
Reinforcement Learning Methods

## ■ 基于价值的方法

| Q-Table | | |
|---|---|---|
| | $a_1$ | $a_2$ |
| $s_1$ | $Q(s_1,a_1)$ | $Q(s_1,a_2)$ |
| $s_2$ | $Q(s_2,a_1)$ | $Q(s_2,a_2)$ |
| $s_3$ | $Q(s_3,a_1)$ | $Q(s_3,a_2)$ |
| $s_4$ | $Q(s_4,a_1)$ | $Q(s_4,a_2)$ |
| $s_5$ | $Q(s_5,a_1)$ | $Q(s_5,a_2)$ |

**DQN方法**



| Q-Table | | |
|---|---|---|
| | $a_1$ | $a_2$ |
| $s_1$ | $Q(s_1,a_1)$ | $Q(s_1,a_2)$ |
| $s_2$ | $Q(s_2,a_1)$ | $Q(s_2,a_2)$ |
| $s_3$ | $Q(s_3,a_1)$ | $Q(s_3,a_2)$ |
| $s_4$ | $Q(s_4,a_1)$ | $Q(s_4,a_2)$ |
| $s_5$ | $Q(s_5,a_1)$ | $Q(s_5,a_2)$ |

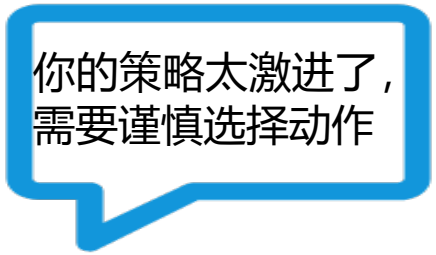
Atari 2600游戏

**Double-DQN**

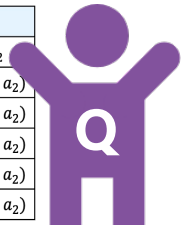
| Q-Table | | |
|---|---|---|
| | $a_1$ | $a_2$ |
| $s_1$ | $Q(s_1,a_1)$ | $Q(s_1,a_2)$ |
| $s_2$ | $Q(s_2,a_1)$ | $Q(s_2,a_2)$ |
| $s_3$ | $Q(s_3,a_1)$ | $Q(s_3,a_2)$ |
| $s_4$ | $Q(s_4,a_1)$ | $Q(s_4,a_2)$ |
| $s_5$ | $Q(s_5,a_1)$ | $Q(s_5,a_2)$ |

价值


动作




| Q-Table | | |
|---|---|---|
| | $a_1$ | $a_2$ |
| $s_1$ | $Q(s_1,a_1)$ | $Q(s_1,a_2)$ |
| $s_2$ | $Q(s_2,a_1)$ | $Q(s_2,a_2)$ |
| $s_3$ | $Q(s_3,a_1)$ | $Q(s_3,a_2)$ |
| $s_4$ | $Q(s_4,a_1)$ | $Q(s_4,a_2)$ |
| $s_5$ | $Q(s_5,a_1)$ | $Q(s_5,a_2)$ |




星际争霸II

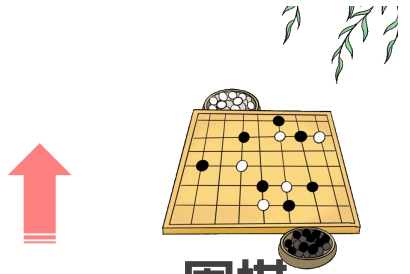
## ■ 基于策略的方法

### 策略梯度方法


策略选择动作

梯度上升

P（动作）



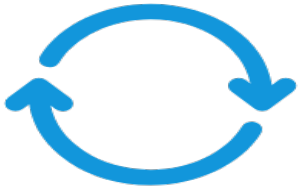
围棋

### 演员-评论家算法



Actor

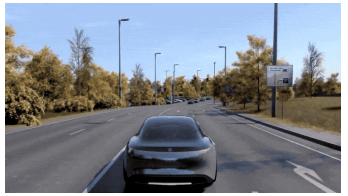
选择动作



价值评估

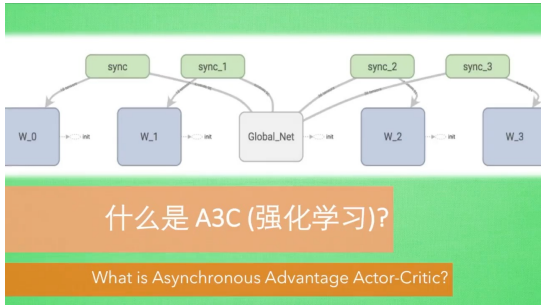

Critic

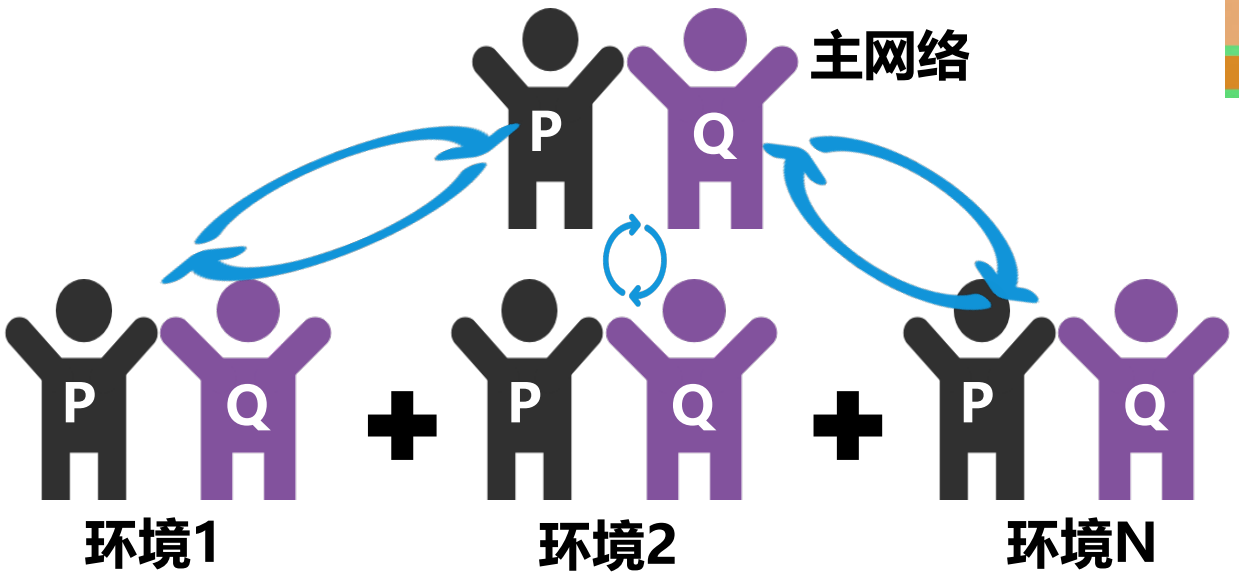
用 Critic 的反馈来优化 Actor 的策略



自动驾驶

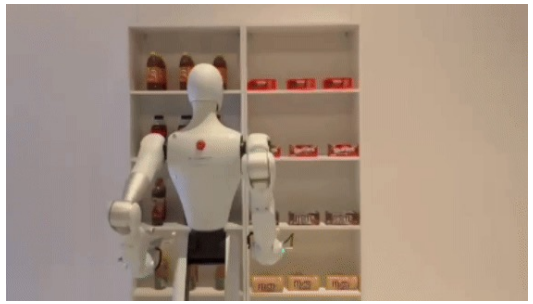
## ■ 基于策略的方法　A3C方法

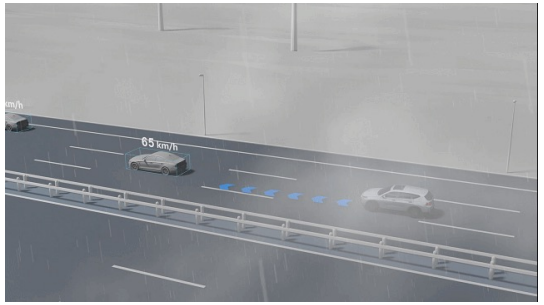


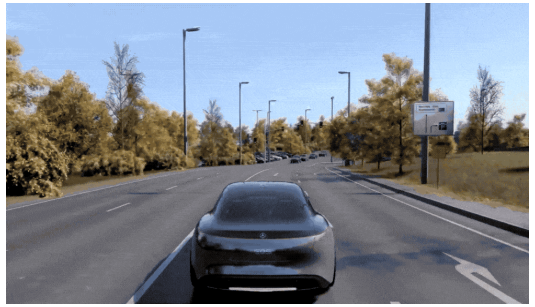



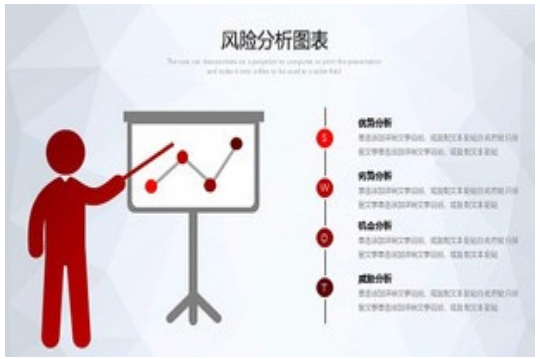

多个 Actor 独立与环境交互并异步更新全局网络

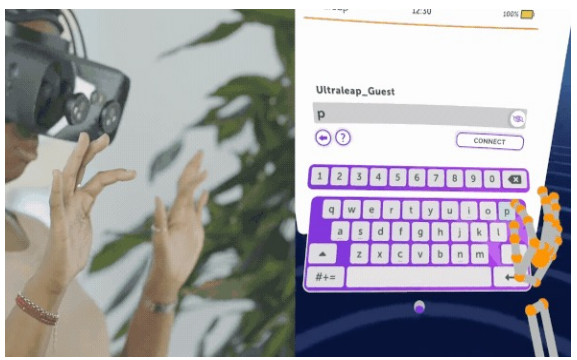
# 深度强化学习应用

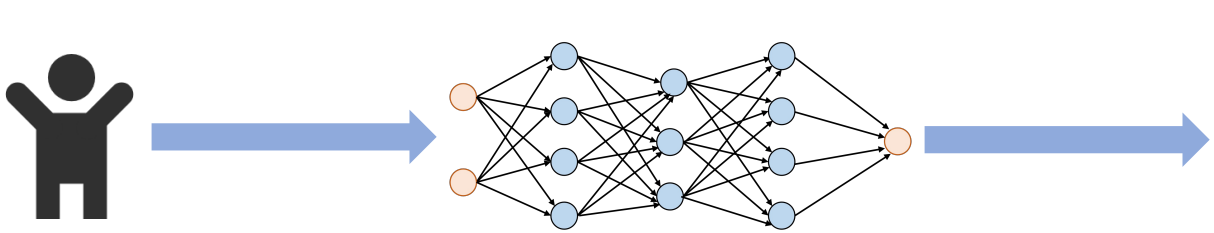




智能体　　神经网络　　决策输出

# 知识点4：智能机器人与强化学习

01 AlphaGo的启示

02 机器人的“学习革命”

# AlphaGo的启示

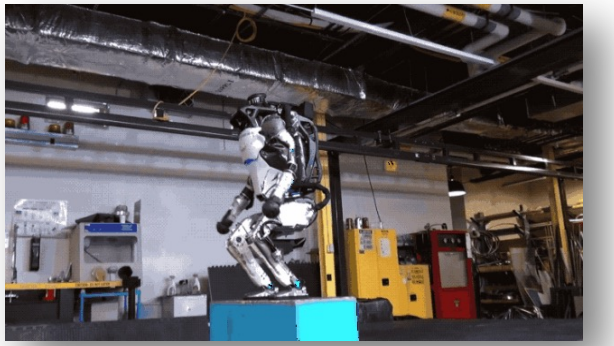
《机器人总动员》

《超能查派》

这些机器人是如何学会行动的呢？

虚拟 VS 现实

强化学习！



波士顿动力机器人Atlas

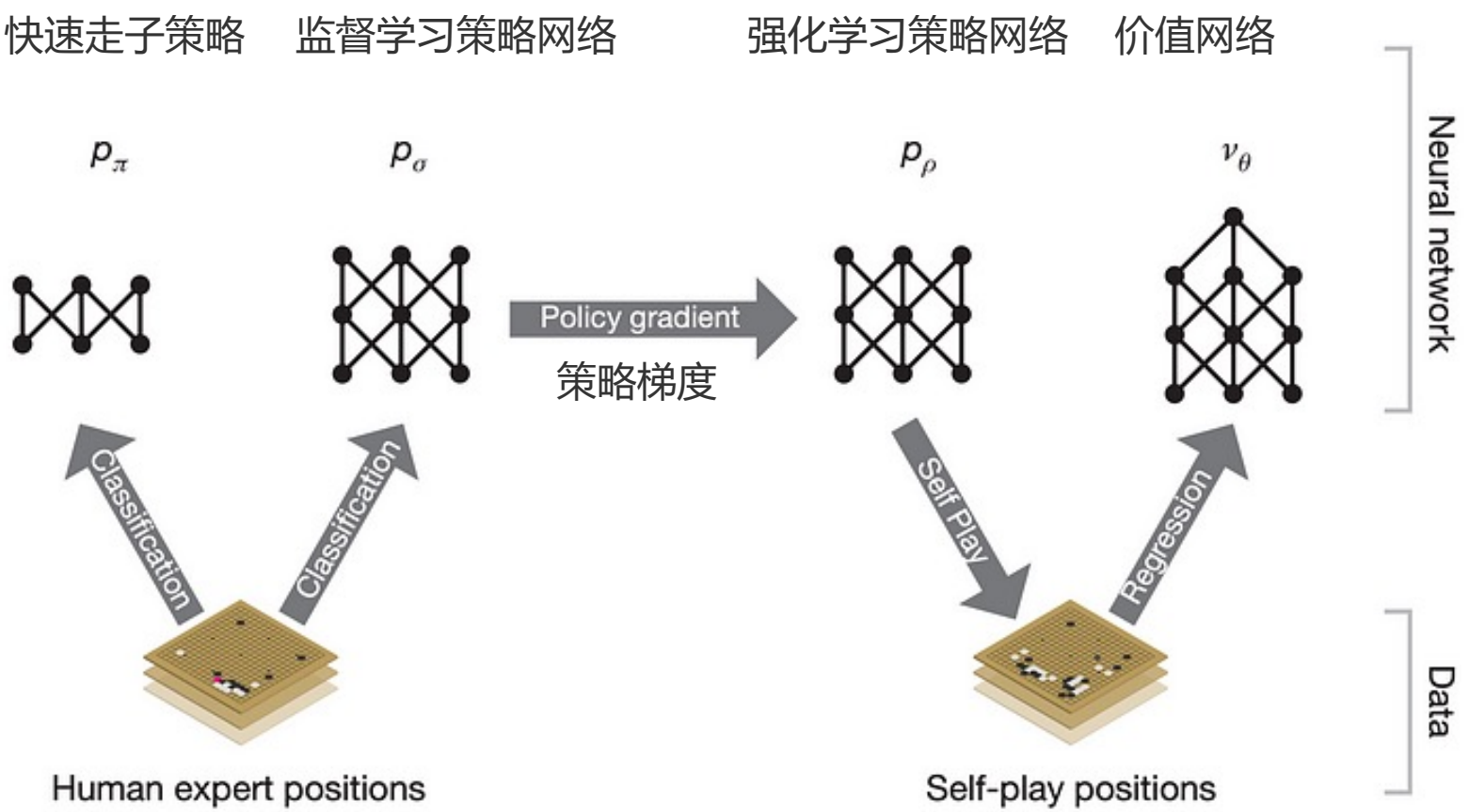

AlphaGo大战李世石

# 围棋有$10^{170}$种可能性，AlphaGo如何“算”出最优解?

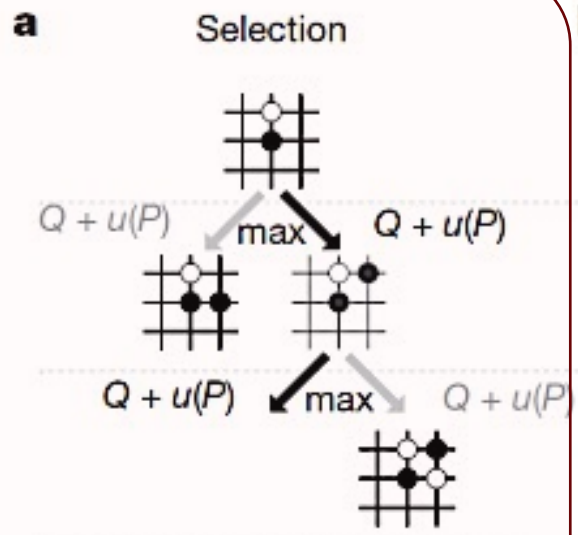
**强化学习**

**蒙特卡洛树搜索**

**制胜法宝**

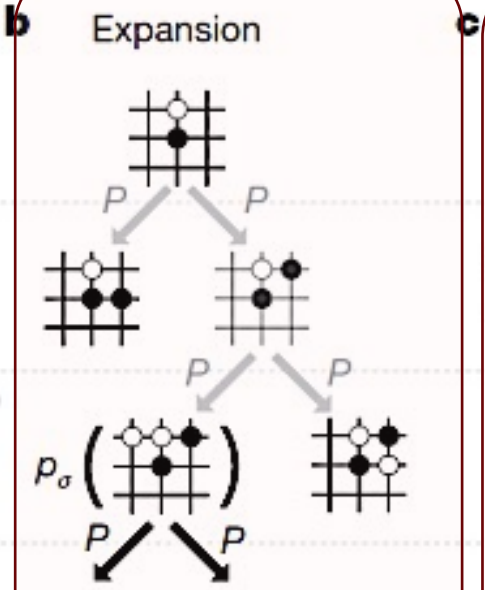


# AlphaGo的启示



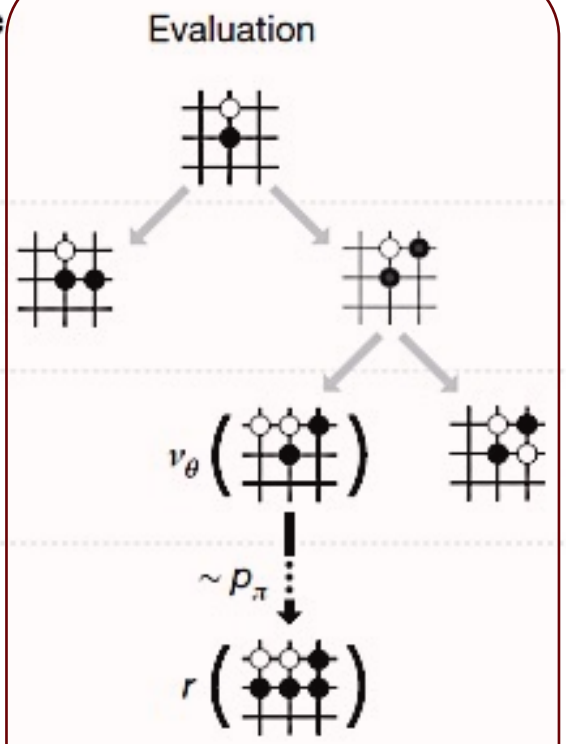

## 选　择

从根节点开始，选择最有潜力的子节点。根节点为当前棋盘状态。



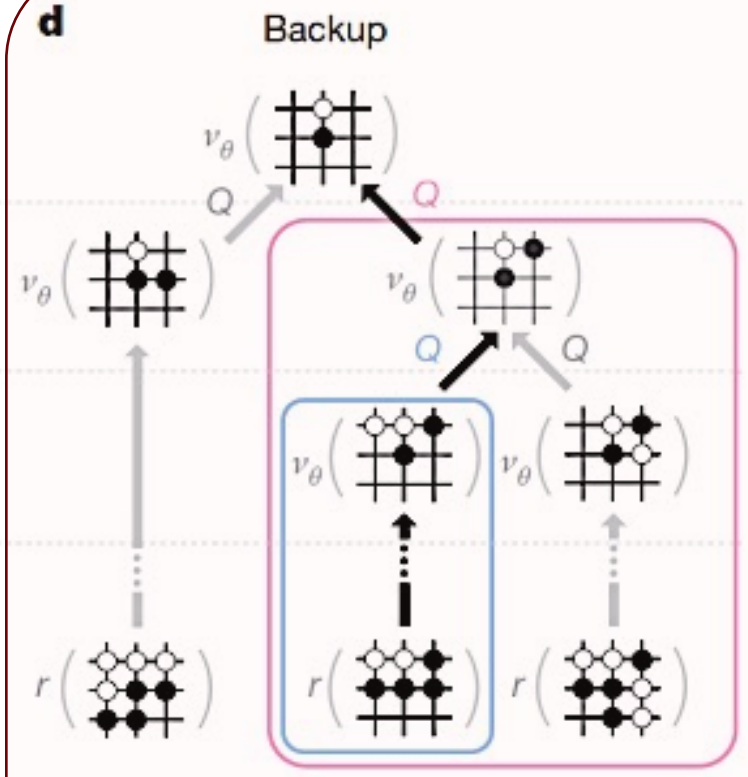

## 扩　展

如果叶节点不是终局状态，使用策略网络生成候选动作，并扩展搜索树。




## 评　估

使用价值网络评估叶节点的胜率。

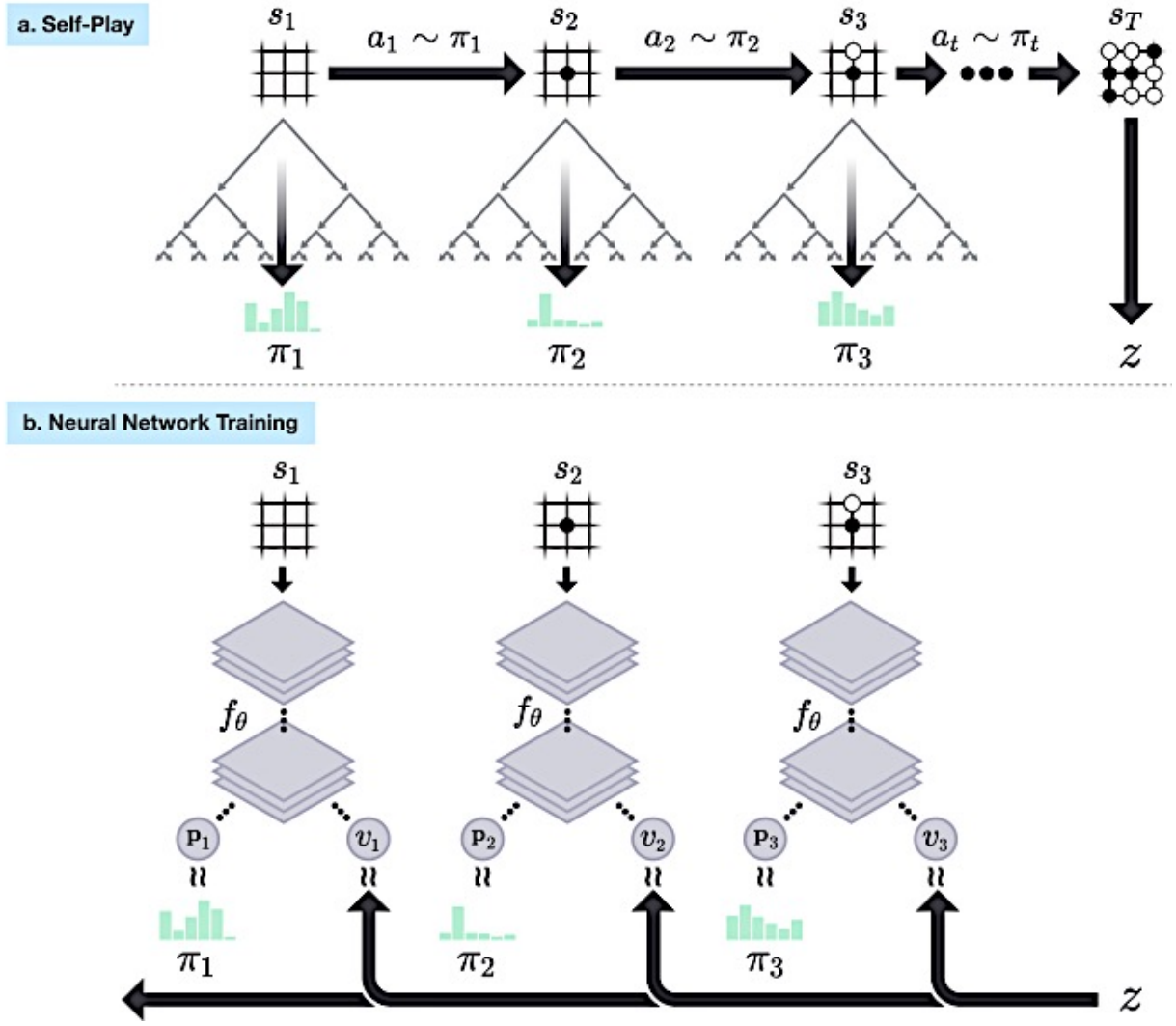



## 回　溯

将评估结果反向传播，更新搜索树中所有节点的统计信息（如访问次数、胜率）。

# AlphaGo的启示

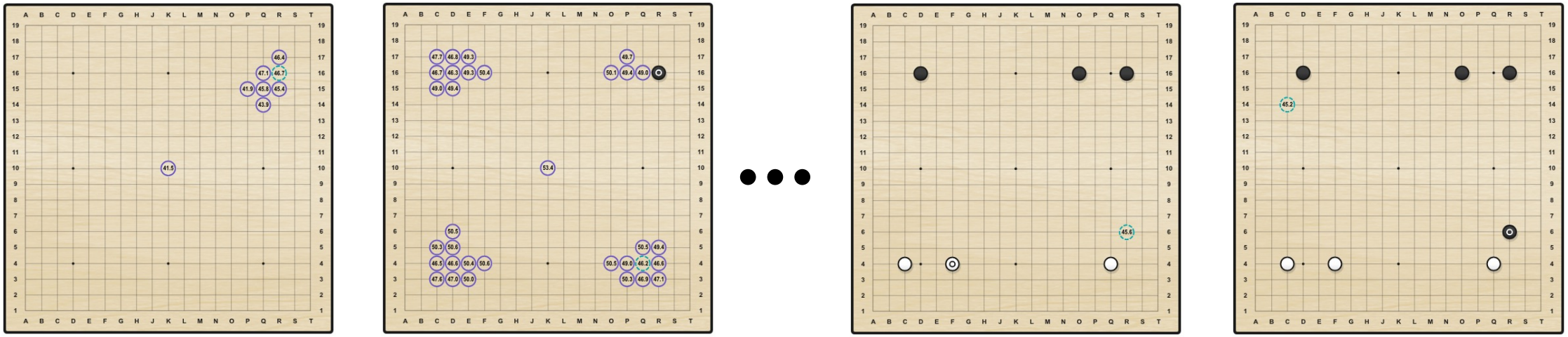
## AlphaGo Zero完全抛弃人类棋谱，100:0 完胜AlphaGo




- **无人类数据**：完全从零开始学习，不依赖任何人类棋谱。通过自我对弈和蒙特卡洛树搜索生成高质量的训练数据，提升学习效率。
- **单一网络**：使用一个统一的神经网络同时预测策略和值函数，简化了模型结构。
- **残差网络**：深度残差网络提升模型表达能力。

# AlphaGo的启示

## AlphaGo围棋教学上线，人类围棋进入AI时代

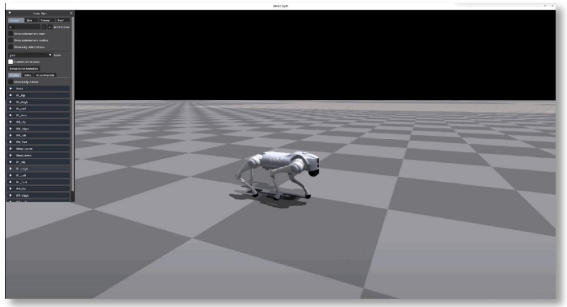


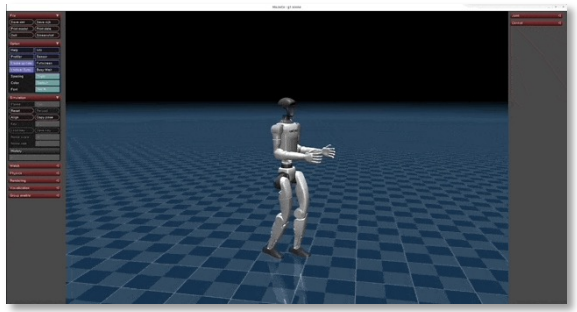
### AlphaGo的成功带来哪些启示呢?

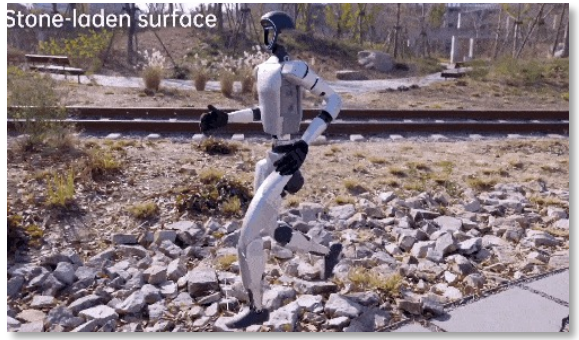
- ✓ 机器可以**从零开始**超越人类千年积累的经验。
- ✓ 解决了传统AI依赖标注数据的瓶颈，开创了**无监督强化学习**的新范式。
- ✓ 实现了**虚拟与现实的动态交互**，通过无监督学习与环境交互实现通用智能。

# 机器人的“学习革命”

## ■ 机器人训练场（行走篇）：如何让机器人适应未知地形?








**在虚拟环境中模拟机器人行走**

**仿真训练**：数千次虚拟跌倒 → 学习平衡策略

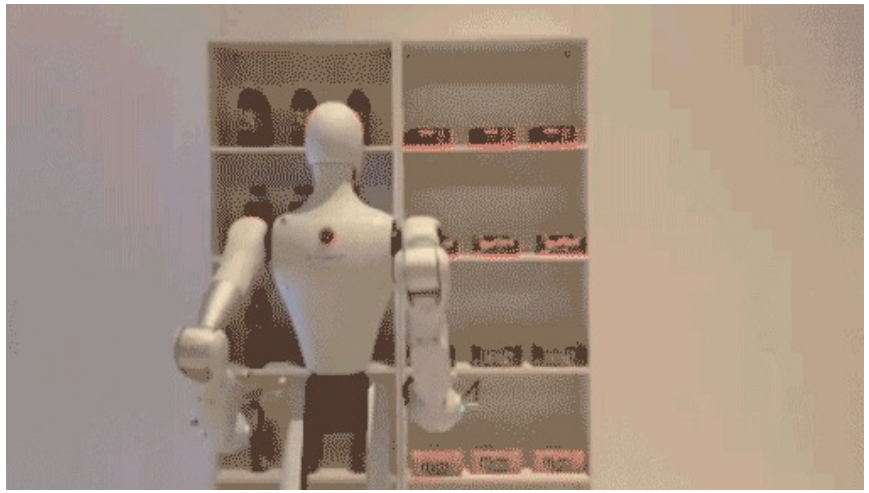
**奖励函数：**

- **基础奖励**：保持平衡 + 前进速度
- **“好奇心”奖励**：鼓励探索新地形

**Sim2Real（仿真到真实的迁移）**

域自适应、域随机化，随机化虚拟环境参数（摩擦系数、地形高度）提升泛化性

## ■ 机器人训练场（抓取篇）：抓取物体



**稀疏奖励**：仅当成功抓取时给予奖励

**自监督预训练**：从随机抓取中学习物理特性

**离散动作**：抓取/松开、移动方向

**连续动作**：夹持器开合力度、6自由度位姿调整

增加内在奖励，鼓励探索新状态（如触碰未知物体表面）

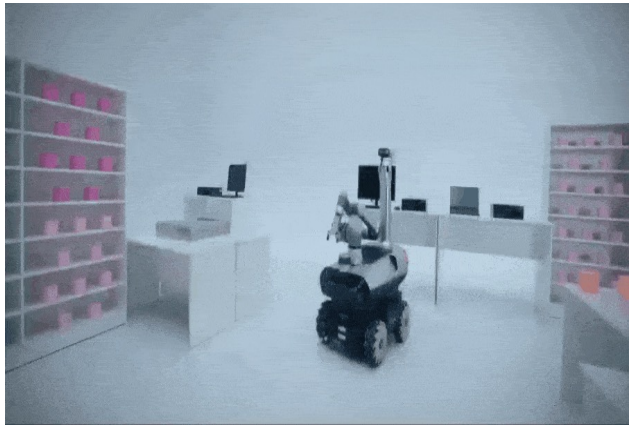
仿真中随机化物体质量、摩擦系数、光照条件，提升泛化性

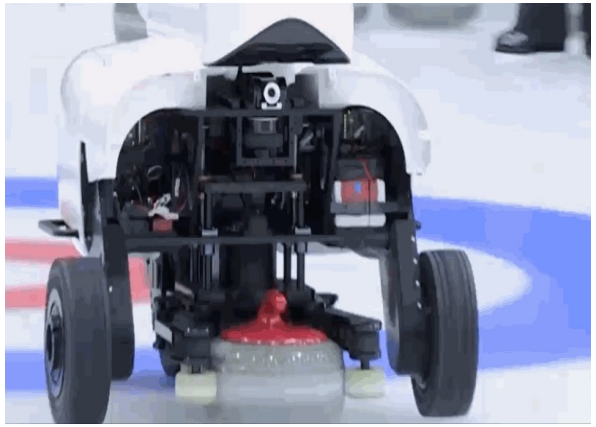
持续生成新物理参数，防止过拟合

**层级强化学习**：高层规划旋转步骤，底层控制手指微调

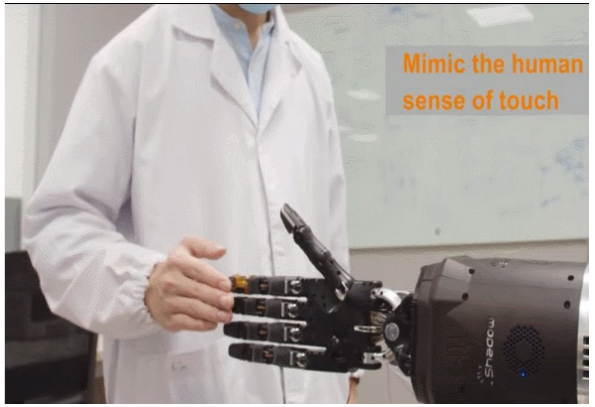
# 机器人的“学习革命”

## ■ 未来趋势






✓ 精细操作　✓ 群体智能　✓ 可变结构　✓ 人机交互　✓ 脑机接口


山东大学
SHANDONG UNIVERSITY

《人工智能通识》AI For Everyone

# 强化学习

学无止境 气有浩然

教育部-华为“智能基座”课程