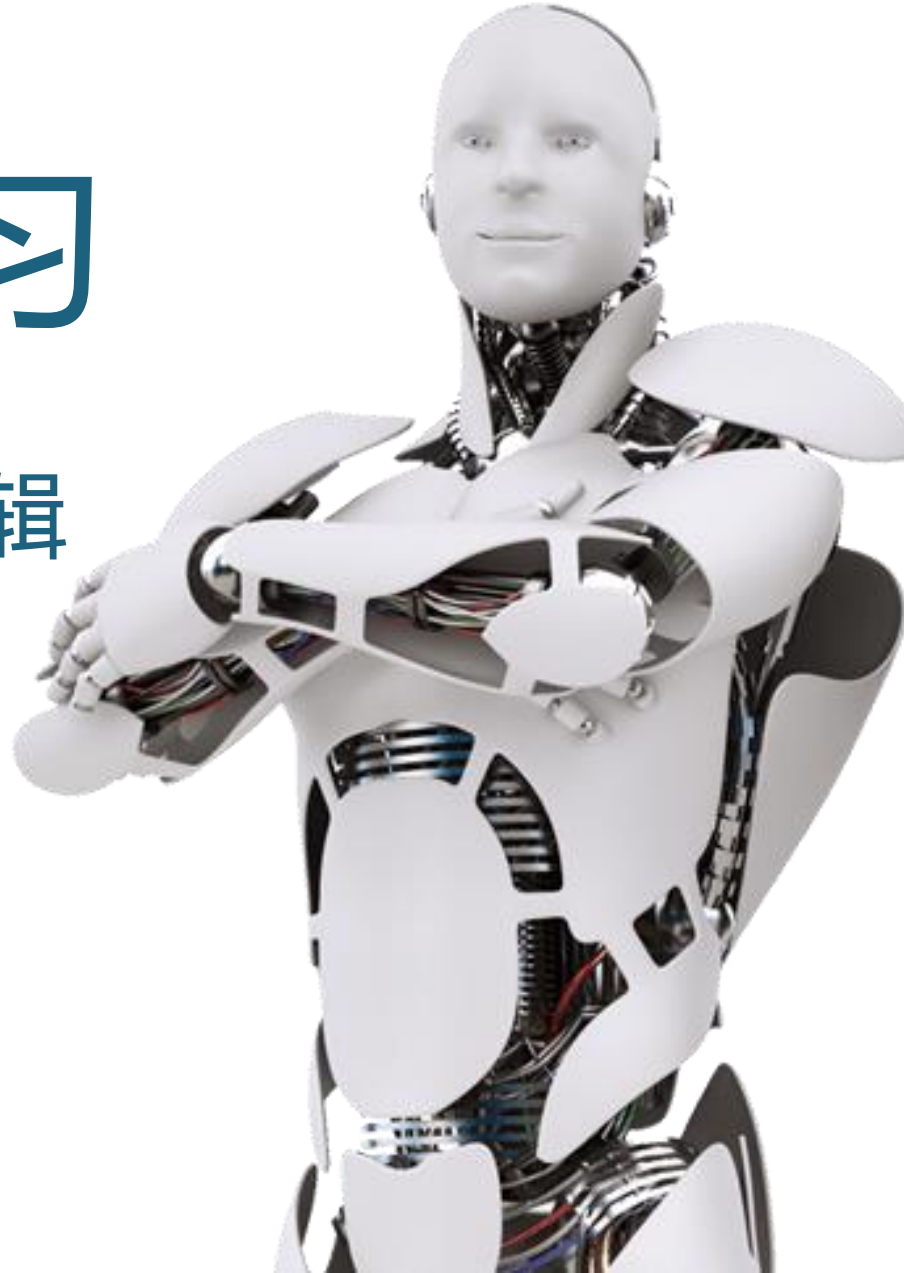
彪哥带你学强化学习

9.策略学习的底层逻辑

DEEPLY UNDERSTAND REINFORCEMENT LEARNING

讲师：韩路彪



策略学习是什么

学习目标：直接学习策略函数π，给定S可以直接输出不同动作的概率

s_1

a_1	0.2
a_2	0.5
a_3	0.3

	s_1	s_2	s_3	s_4	。 。 。
a_1					
a_2					
a_3					

策略学习怎么学

人在学新东西的时候是怎么学的？

1. 跟老师学
2. 借鉴以往类似经验、知识尝试
3. 探索、总结

监督学习

监督学习与强化学习之间

强化学习

爱因斯坦的三个小板凳

新搬家后去哪买水果？　　附近有三家商店：A、B、C

随机选一家，A	买到了	买水果可以去商家A	A:0.5 B:0.25 C:0.25
换一家，C	没买到	买水果不可以去商家C	A:0.7 B:0.28 C:0.02
换一家，B	买到了 比A更新鲜更便宜	买水果尽量去商家B	A:0.39 B:0.6 C:0.01

探索

回报

策略

策略学习思路：根据探索得到的回报不断优化策略

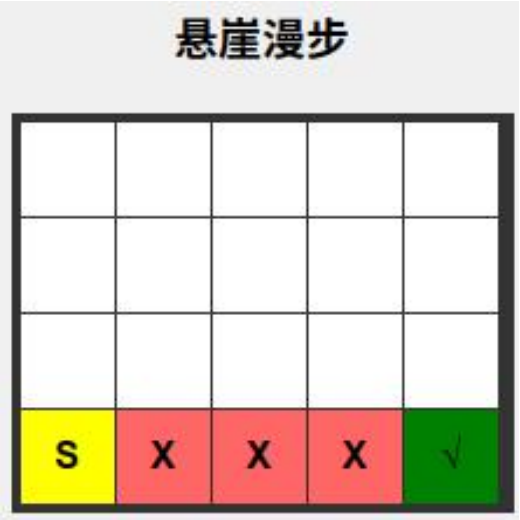
策略学习怎么学

策略学习的底层逻辑：根据回报更新策略，**复盘式学习**，训练过程依赖回报

$$J(\theta) = \mathbb{E}_{\tau \sim \pi_\theta}[G_\tau]$$

如果知道了每个状态的期望（或最大）回报，就变成了监督学习

怎么得到回报？
- 蒙特卡洛采样 — REINFORCE
- 另外单独训练一个价值网络 — Actor-Critic
- 多次采样求均值 — GRPO



策略学习与价值学习对比

价值学习更适合离散动作环境，策略学习可用于连续动作环境

策略学习可按概率采样，比如大语言模型里边的TopP

概率类模型使用策略学习更方便，比如大语言模型

策略学习没有价值学习容易收敛