机器学习

深度强化学习

马尔可夫决策过程与强化学习基础/OpenAI Gym

徐哲　9月29日

滴滴内部
学习资料
请勿外传

扫钉钉群，加入我们

讲师介绍

徐哲

大数据技术部&匹配组

负责交易引擎模型化项目

熟悉计算机视觉、深度学习、强化学习算法

学习受益

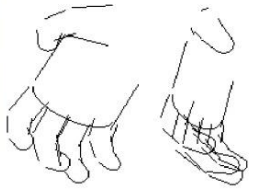
- 了解强化学习算法对实际问题的建模抽象方式
- 理解马尔可夫决策过程(MDP)的定义和主要构成元素

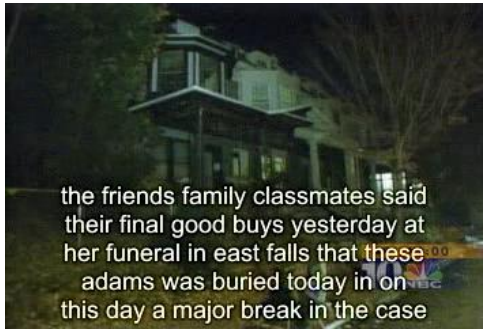
目录
Contents

01 第一章

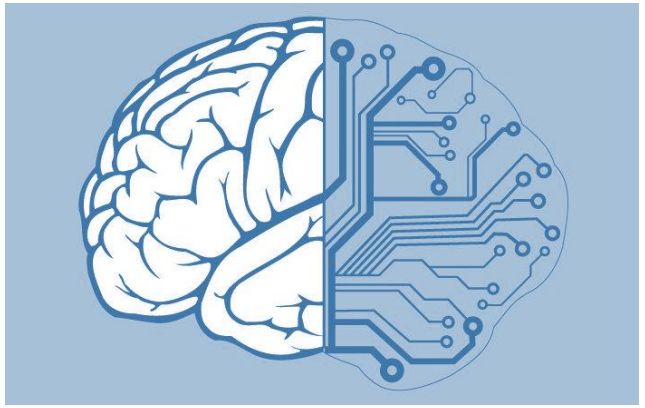
主题：人工智能

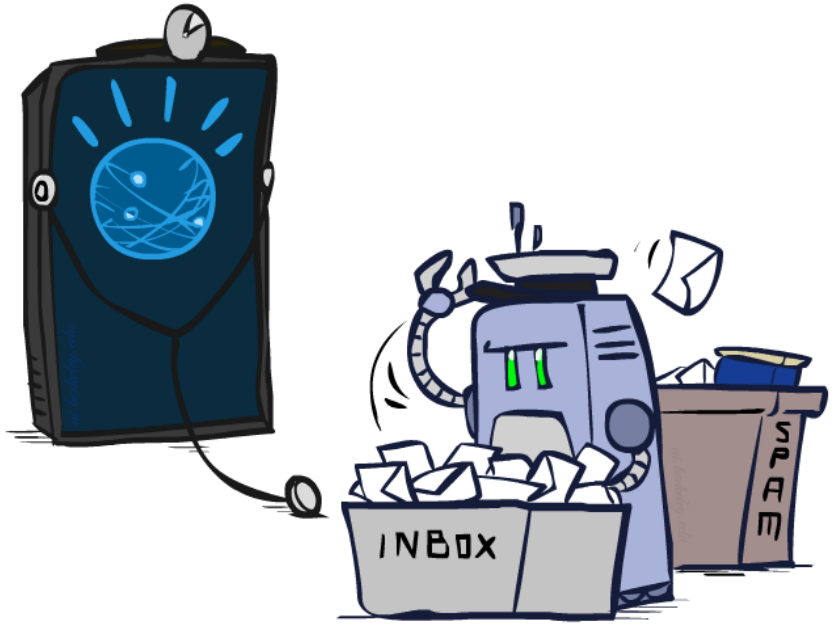
人工智能





人工智能

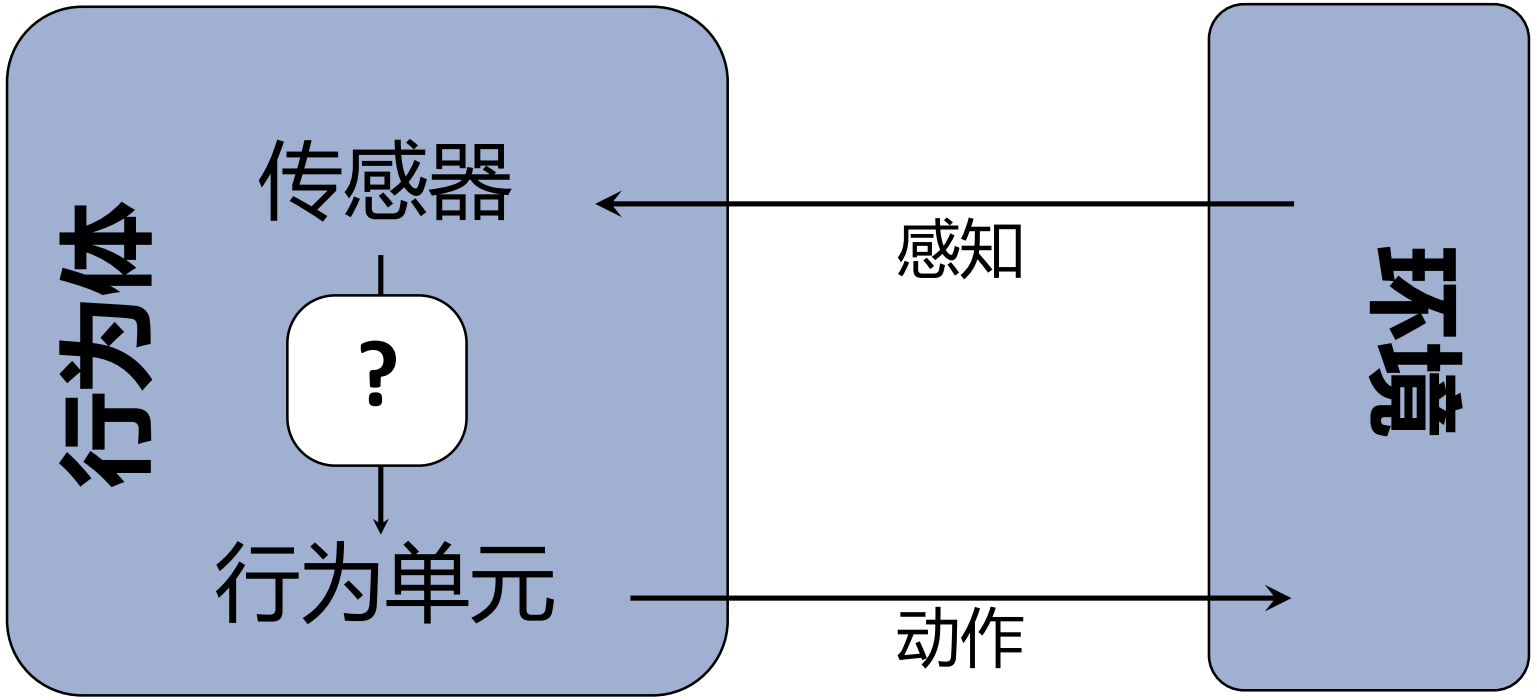




感知

决策

感知与决策

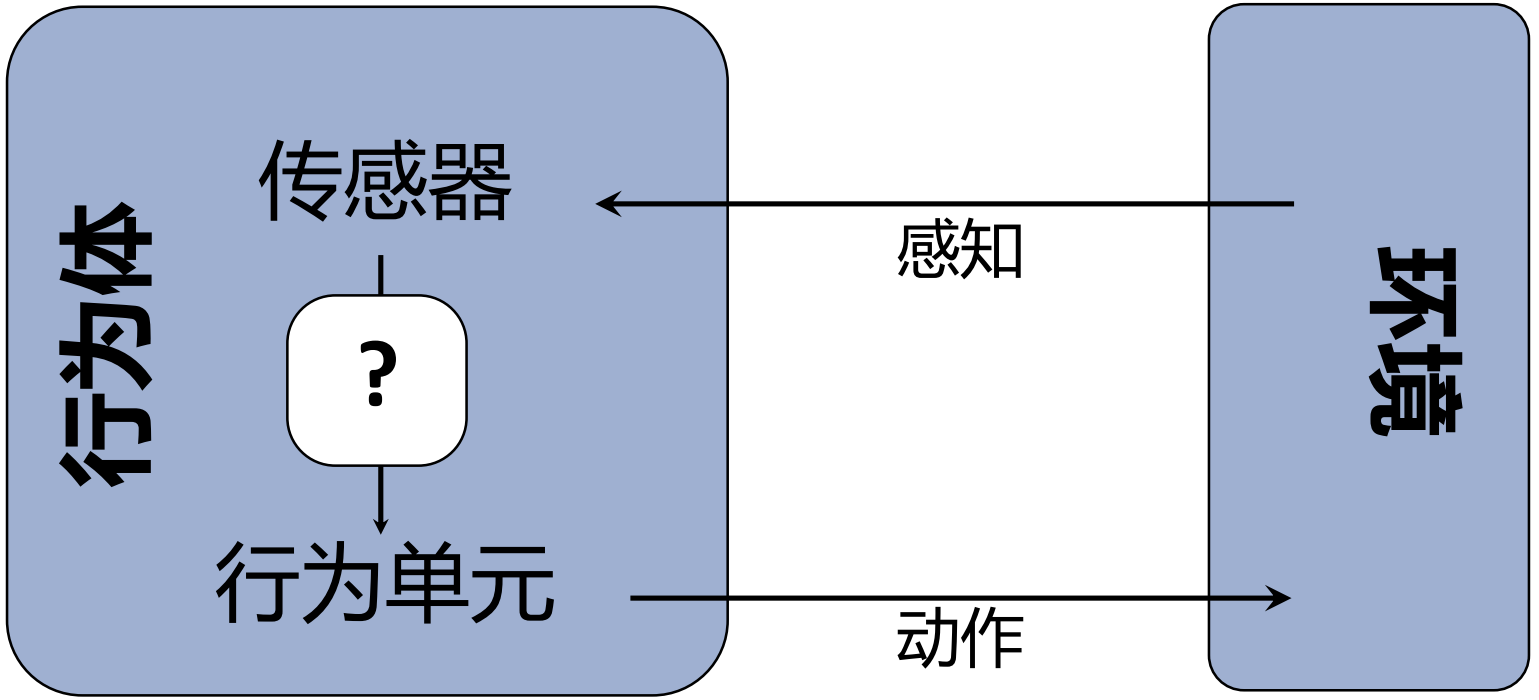


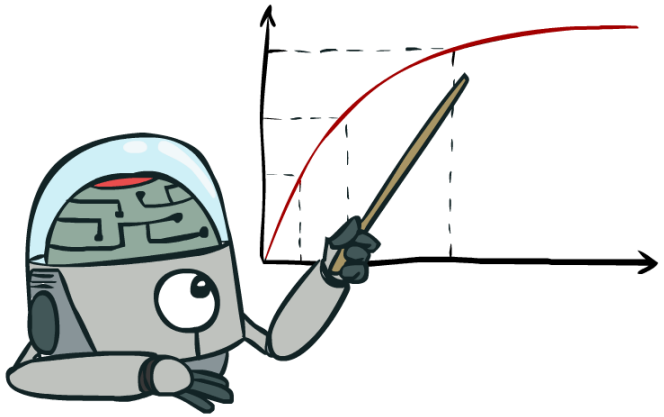
- 行为体(agent)从环境(environment)处通过感知获取信息和激励
- 行为体的动作(action)回过来影响其在环境中的状态(status)

人工智能的目的？

使得计算机

最大化期望效用



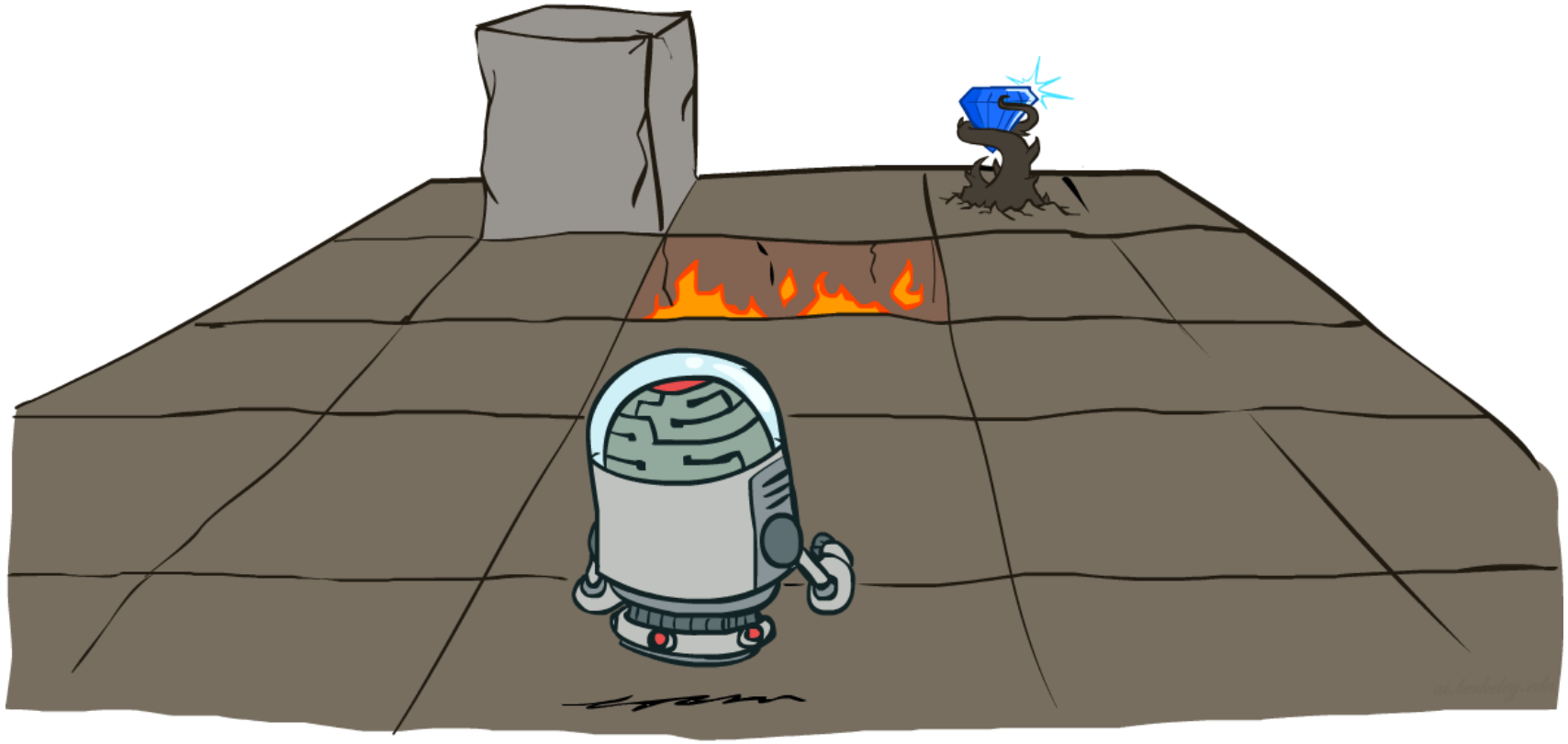


- 人工智能的目标：使行为体能理性地行动，表现为最大化其收益，即期望效用(expected utility)

02 第二章

主题：马尔可夫决策过程-定义

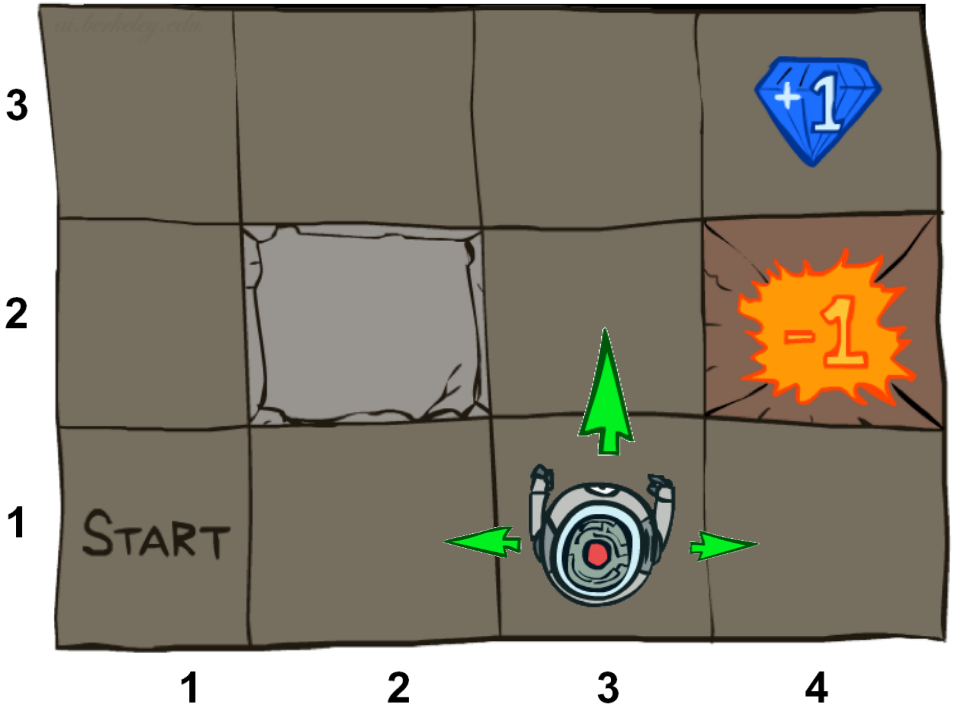
马尔可夫决策过程



PPT插图来源于Berkeley CS188课件 http://ai.berkeley.edu/home.html

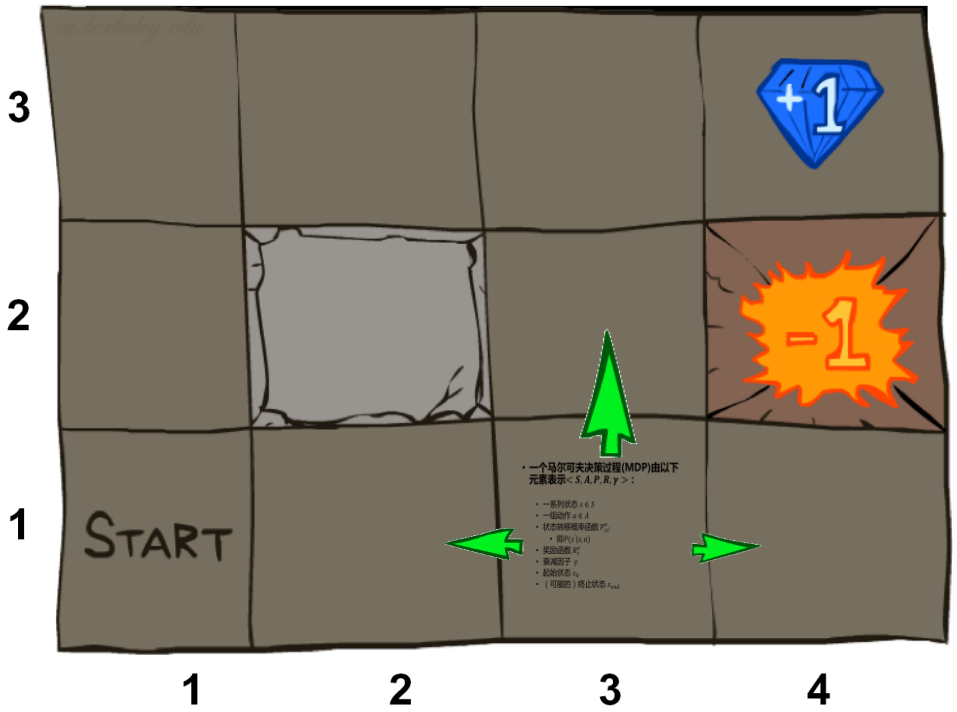
格子世界 (Grid World)

- **类迷宫问题**
  - Agent在以格子为单位存在和移动
  - 不能移动到迷宫范围之外，墙会阻挡agent移动
- **奖励**
  - 部分格子会有特殊的奖励，出现在这些格子会结束一轮游戏
  - 钻石 +1 表示正激励，火堆 -1 表示负激励
  - 每一步可能会有一个小小的生存激励（可能是负的）
- **目标：最大化期望累积收益**



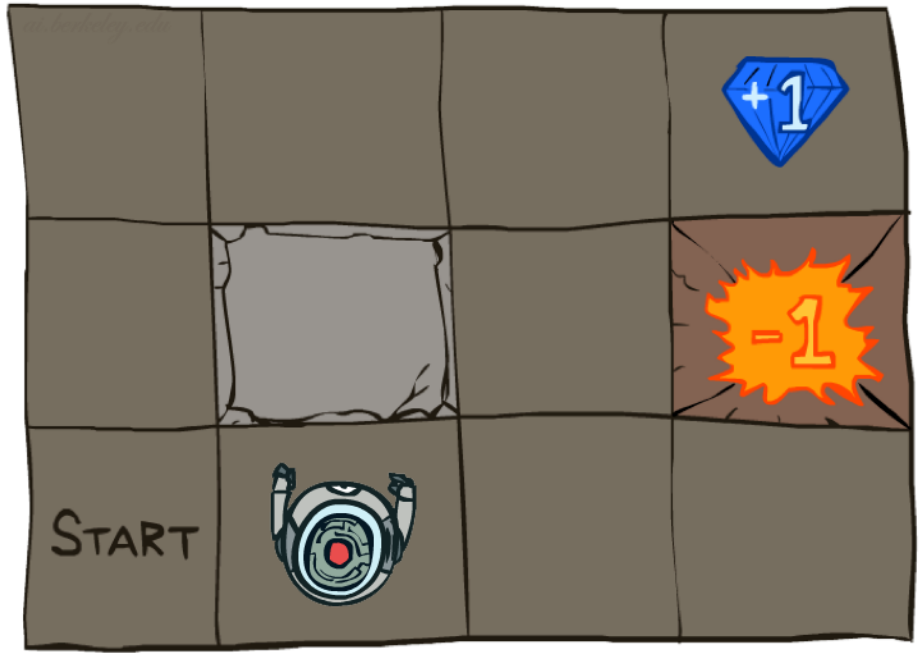
MDP

- **一个马尔可夫决策过程(MDP)由以下元素表示$< S, A, P, R, \gamma >$：**
  - 一系列状态 $s \in S$
  - 一组动作 $a \in A$
  - 状态转移概率函数 $\mathrm{P}_{ss'}^{a}$,
    - 即$P(s'|s,a)$
  - 奖励函数 R_s^a
  - 衰减因子 γ
  - 起始状态 s_0
  - （可能的）终止状态 s_{end}

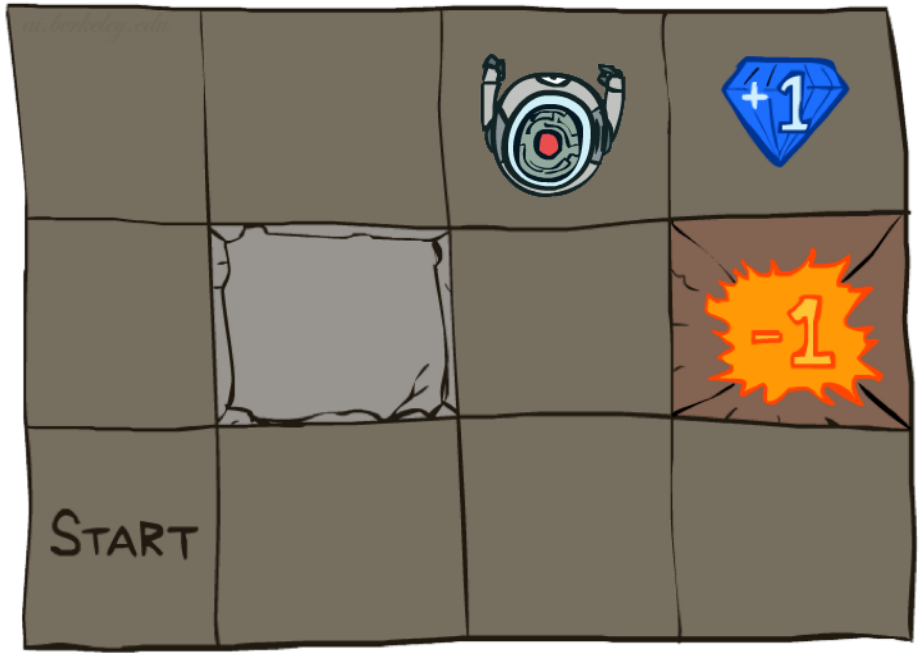


状态 (State, s)

状态（state，这里特指agent state），指agent观测到的状态



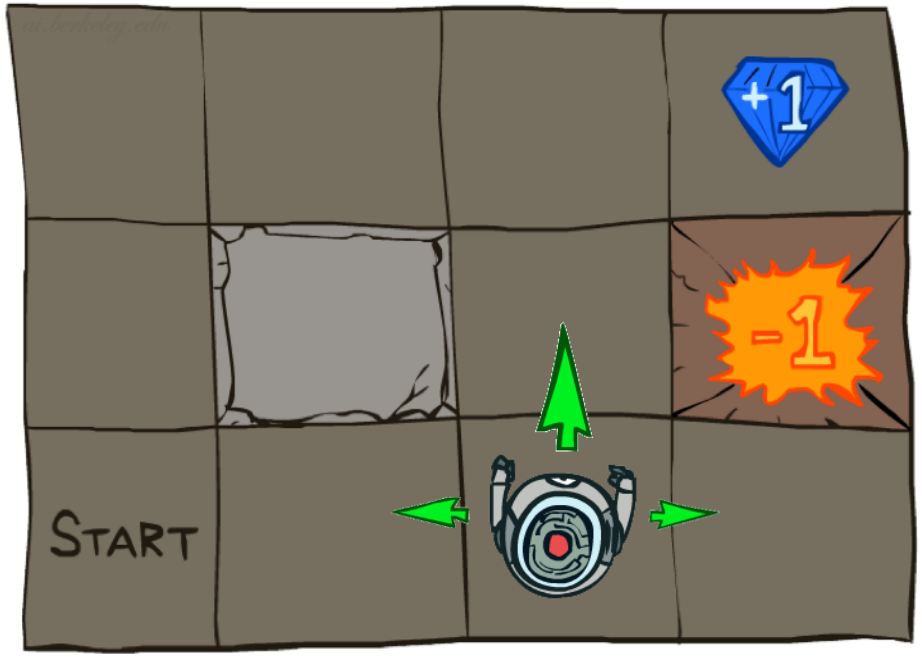
S_1



S_2

动作 (Action, a)

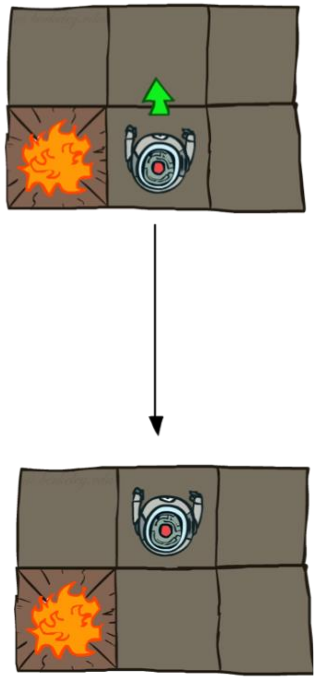
动作(action)，指agent在环境中可以进行的行为



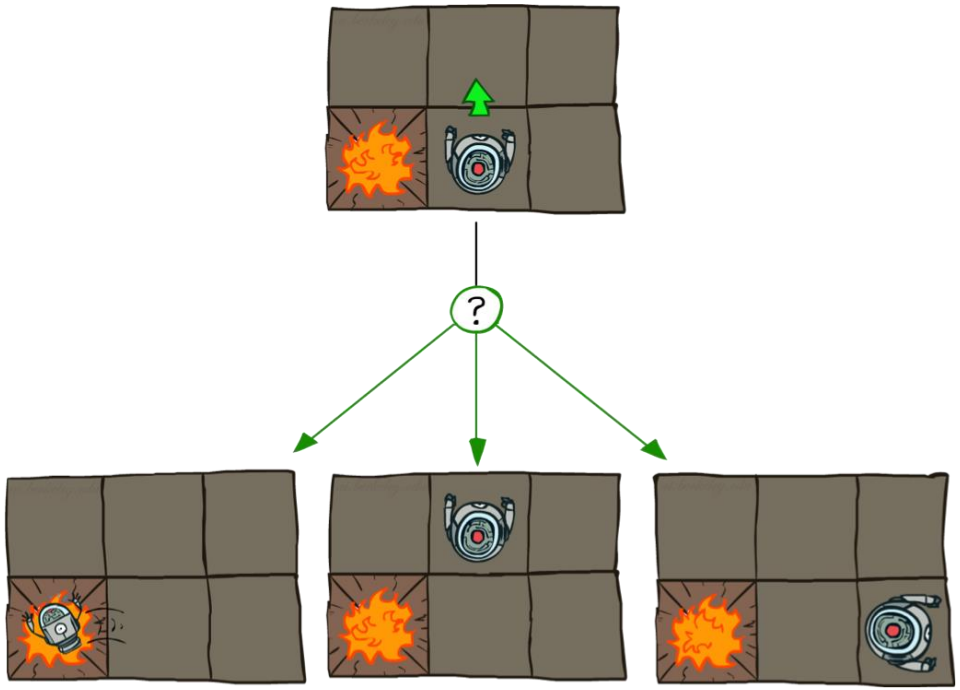
状态转移 ($P_{ss'}^{a}$)

状态转移，即agent进行某行为后，对其状态会发生怎样变化的建模

确定性格子世界

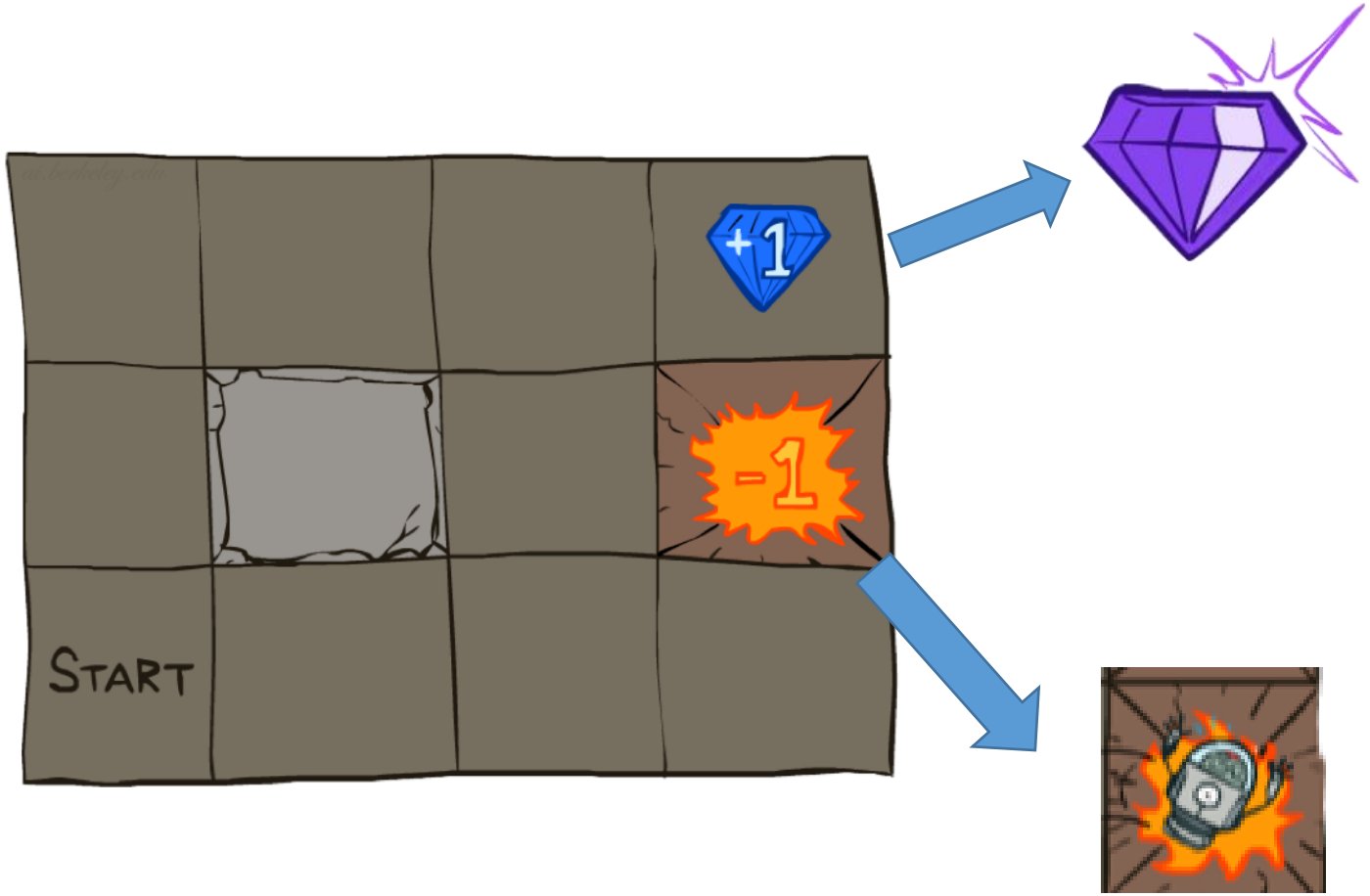


随机性格子世界

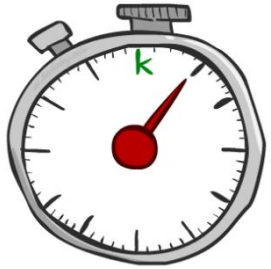


奖励 (R_s^a)

奖励，指环境对agent特定状态/动作的反馈奖赏，agent的目标即获得更多的奖励



奖励可能是稀疏的！



-0.01?

-0.1?

-2.0?

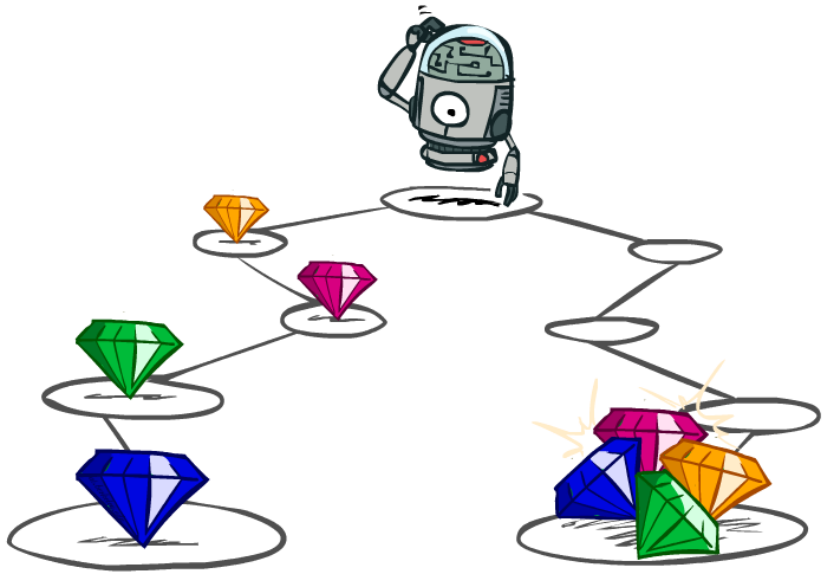
奖励的时效性

- 下面哪个奖励序列是agent更希望拿到的？
- 多还是少?

[3, 2, 2]　　or　　[2, 3, 4]

- 现在还是以后?

[0, 0, 1]　　or　　[1, 0, 0]



衰减因子

- 我们当然可以选择拿到更多奖励
- 我们当然也可以选择尽可能的在当前多拿到奖励
- 解决方案：引入衰减因子，将奖励随时间以指数级衰减

1

现在

γ

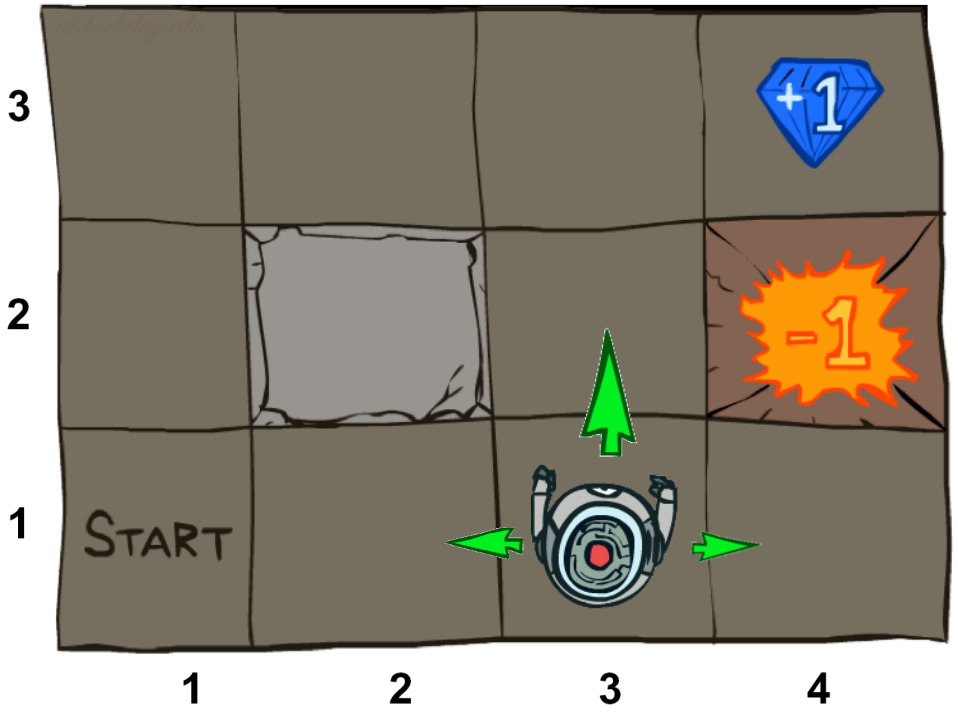
一个时间点后

γ^2

两个时间点后

（回顾）MDP

- **一个马尔可夫决策过程(MDP)由以下元素表示 $< S, A, P, R, \gamma >$：**
  - 一系列状态 $s \in S$
  - 一组动作 $a \in A$
  - 状态转移函数 $\mathrm{P}_{ss'}^{a}$
    - 即$P(s'|s,a)$
  - 奖励函数 R_s^a
  - 衰减因子 γ
  - 一个完整的从起始状态到终止状态的过程，称为一个回合(episode)
    - 起始状态 s_0
    - （可能的）终止状态 s_{end}



目标为最大化累积收益(Return)

$$G_t = R_{t+1} + \gamma R_{t+2} + \cdots$$

为什么叫Markov?

- 马尔可夫性指随机过程中某事件的发生只取决于其上一事件，即“无记忆”的过程

- MDP中，马尔可夫性表现为动作的影响只与当前状态相关，而与历史行为无关

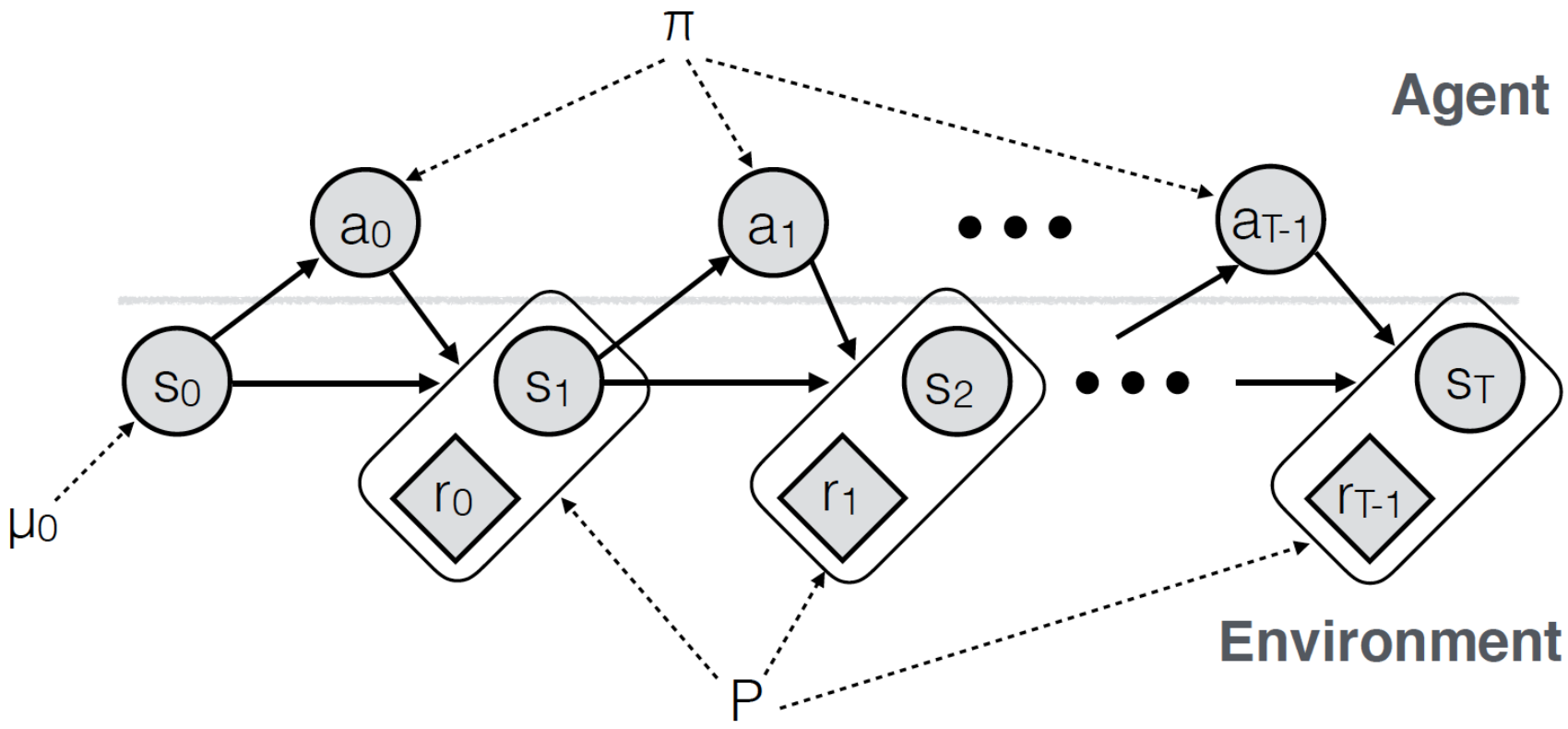
$$P(S_{t+1} = s' | S_t = s_t, A_t = a_t, S_{t-1} = s_{t-1}, A_{t-1}, \dots S_0 = s_0)$$

$$=$$

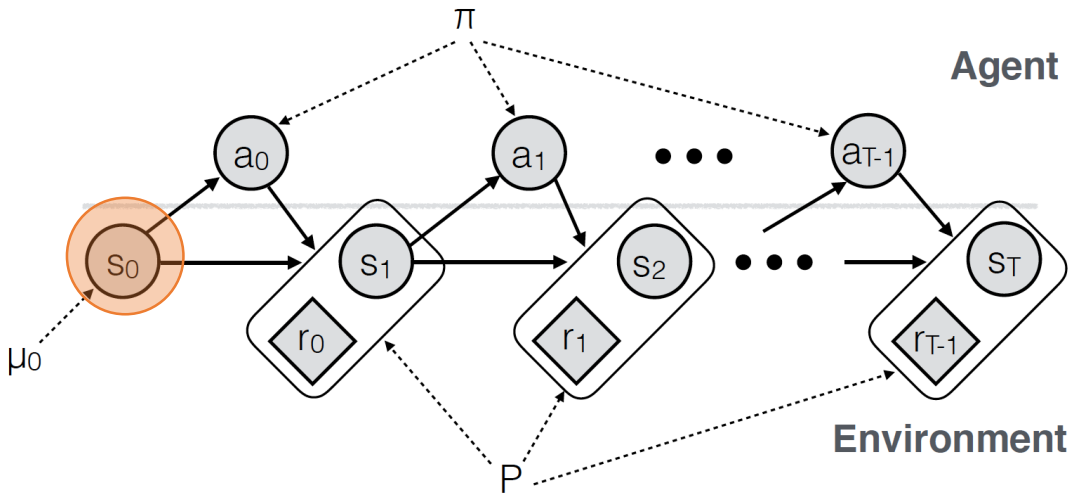
$$P(S_{t+1} = s' | S_t = s_t, A_t = a_t)$$

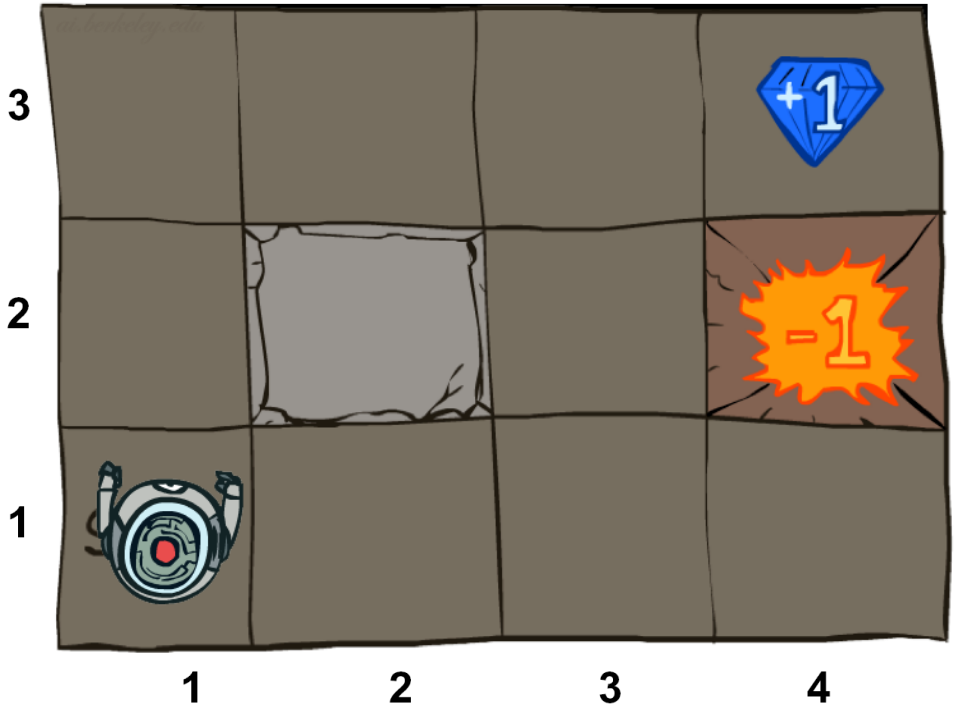
Andrey Markov
(1856-1922)

决策过程 – 序列决策

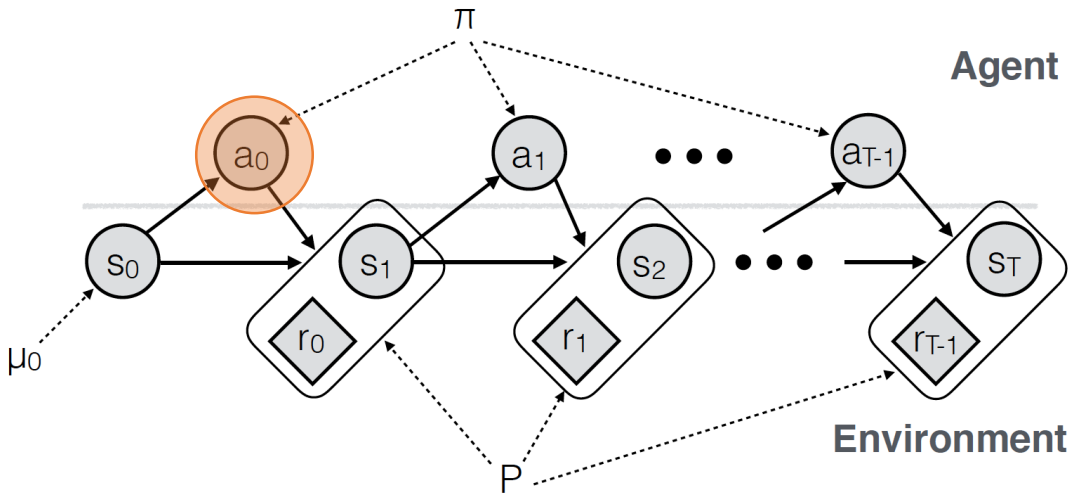


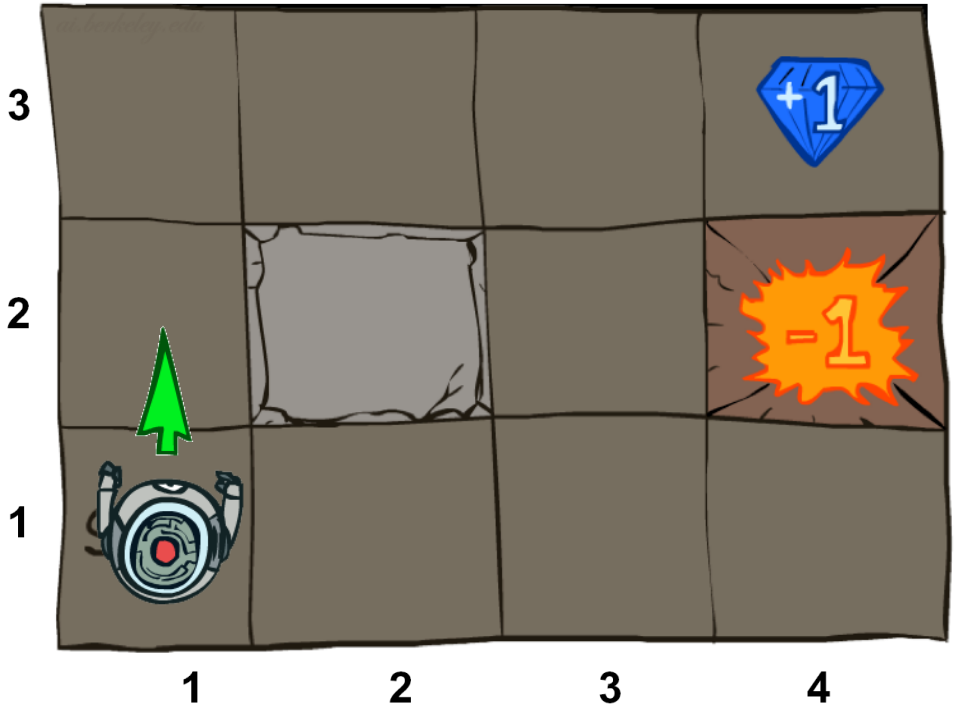
序列决策


π
Agent
a0
a1
aT-1
s0
s1
s2
sT
r0
r1
rT-1
μ0
P
Environment

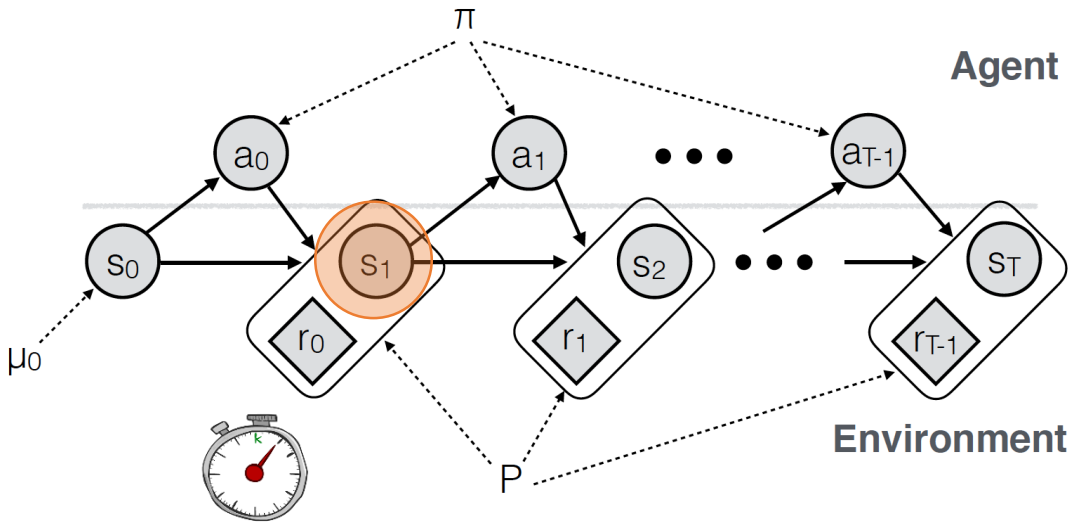

3
2
1
+1
-1
1
2
3
4

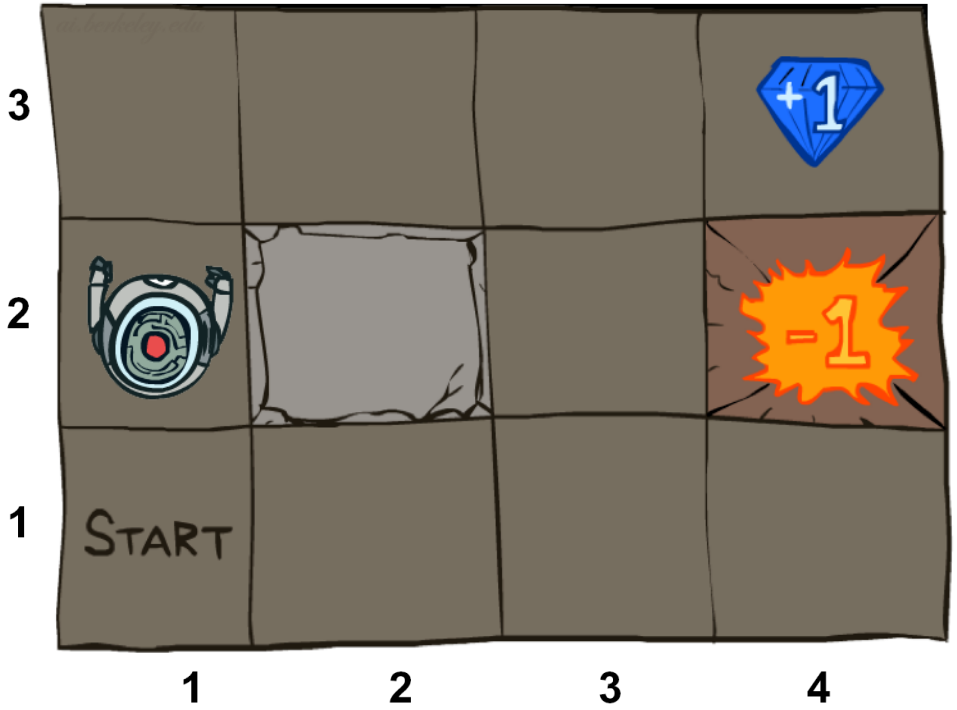
序列决策


π
Agent
a0
a1
aT-1
s0
s1
s2
sT
r0
r1
rT-1
μ0
P
Environment

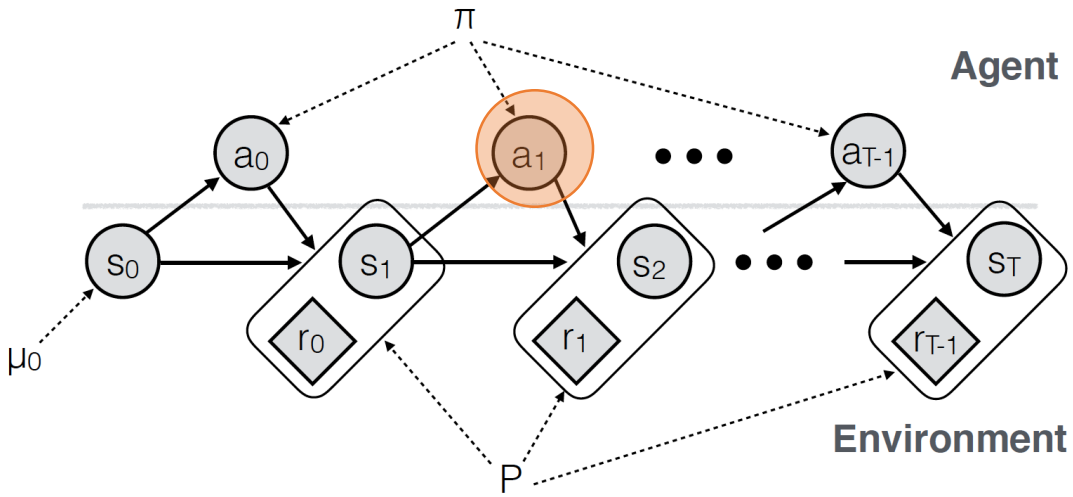

3
2
1
+1
-1
1
2
3
4

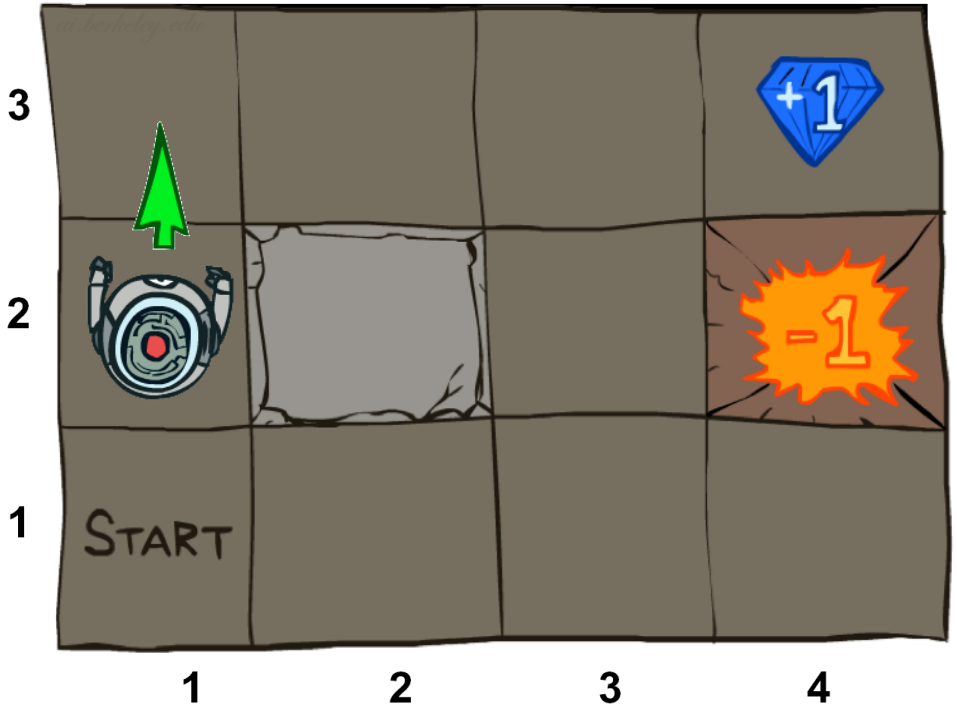
序列决策


π
Agent
a0
a1
aT-1
s0
s1
s2
sT
r0
r1
rT-1
μ0
P
Environment

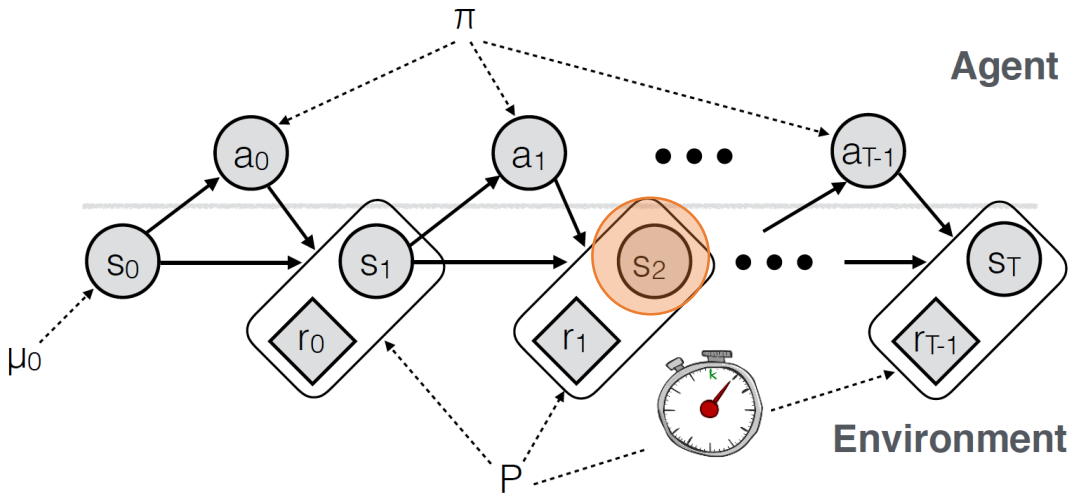

3
2
1
+1
-1
START
1
2
3
4

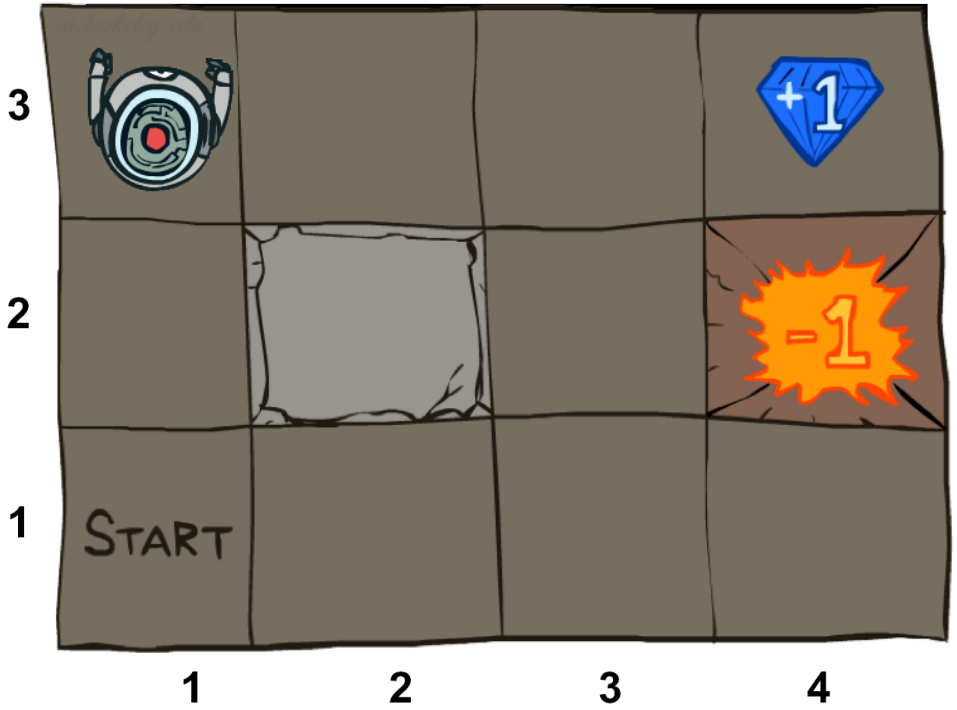
序列决策


π
Agent
a0
a1
aT-1
s0
s1
s2
sT
r0
r1
rT-1
μ0
P
Environment

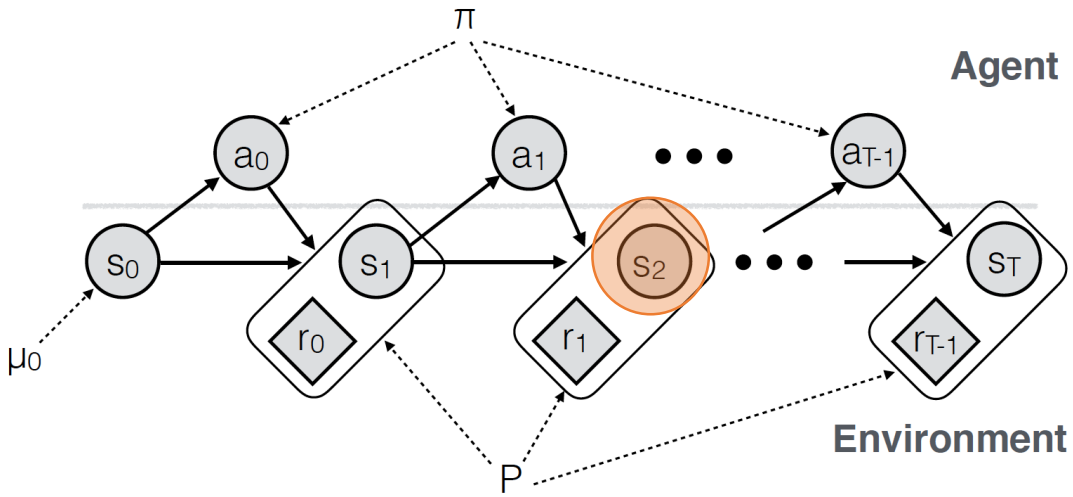

3
2
1
+1
-1
START
1
2
3
4

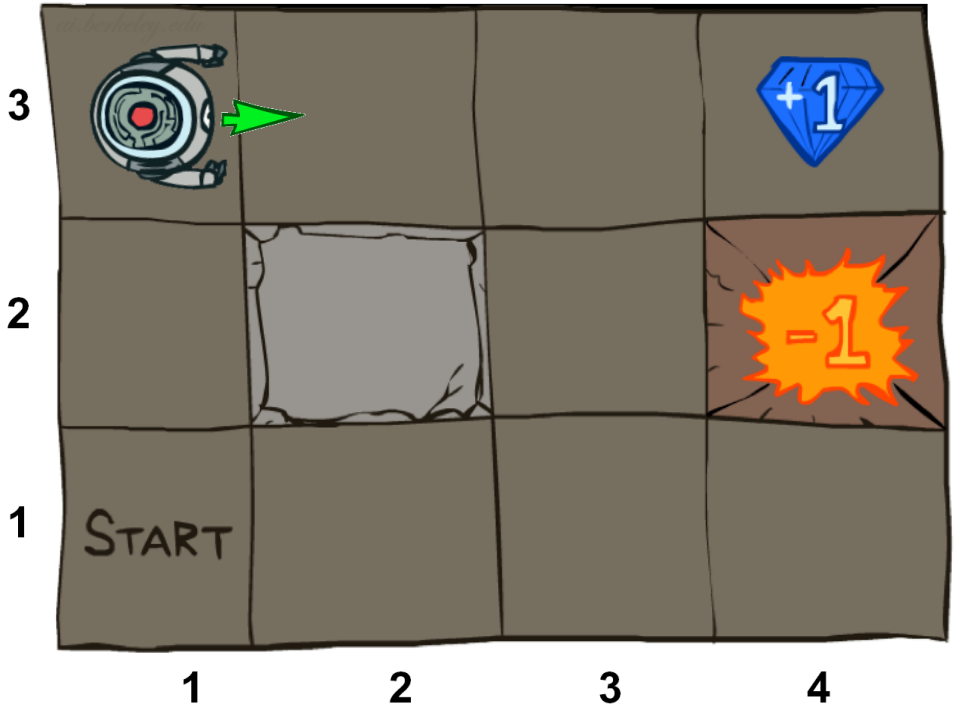
序列决策


π
Agent
a0
a1
aT-1
s0
s1
s2
sT
r0
r1
rT-1
μ0
P
Environment

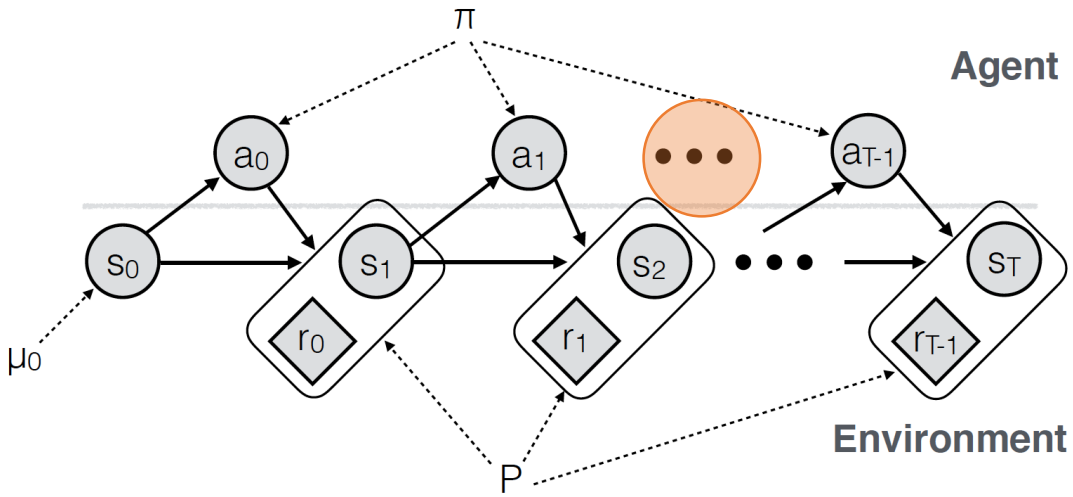

3
2
1
START
+1
-1
1
2
3
4

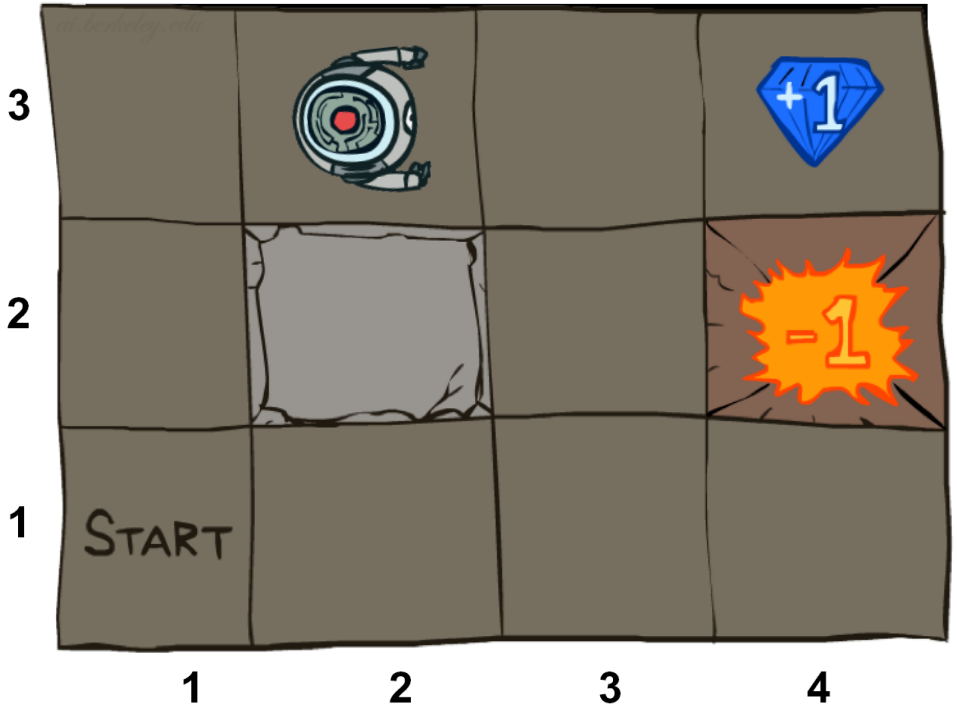
序列决策


π
Agent
a0
a1
aT-1
s0
s1
s2
sT
r0
r1
rT-1
μ0
P
Environment

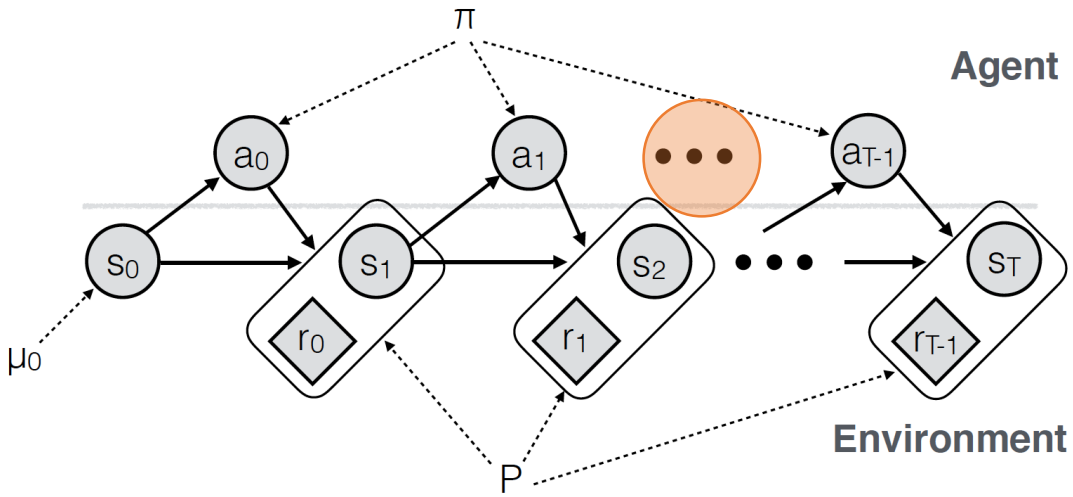

3
2
1
START
+1
-1
1
2
3
4

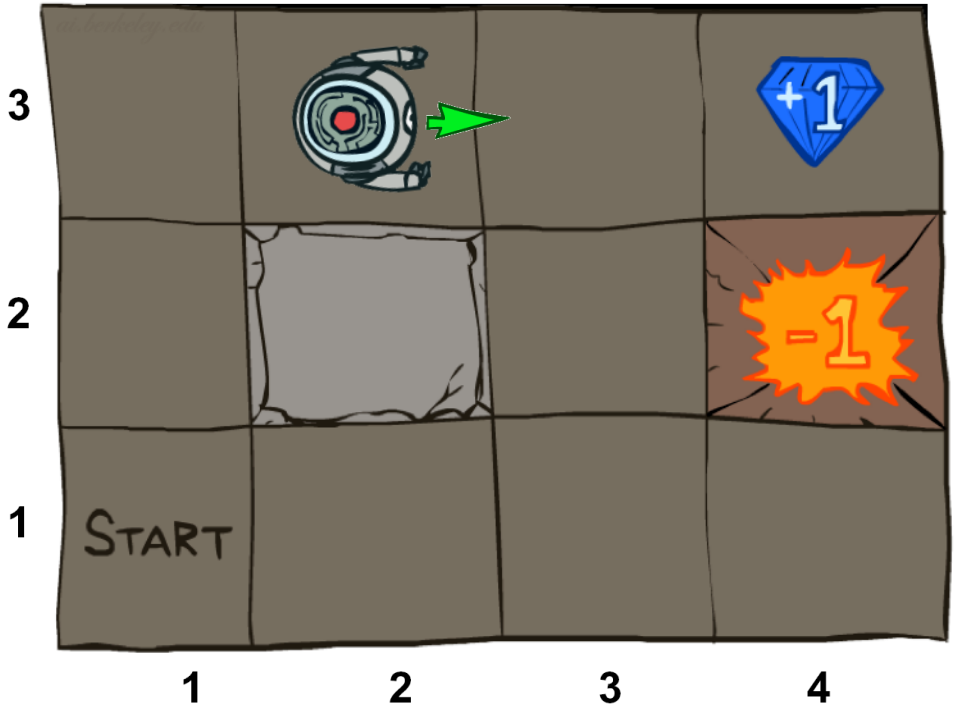
序列决策


π
Agent
a0
a1
aT-1
s0
s1
s2
sT
r0
r1
rT-1
μ0
P
Environment

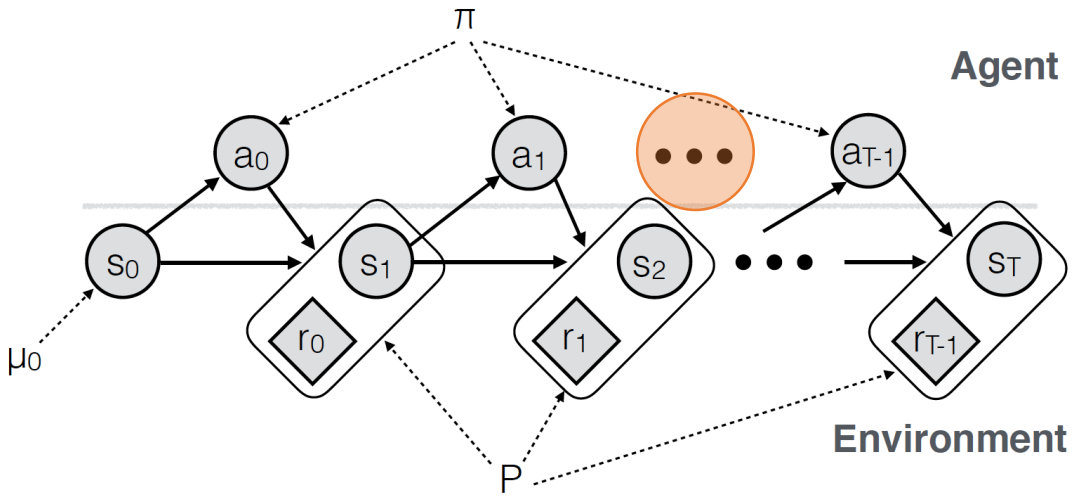

3
2
1
START
+1
-1
1
2
3
4

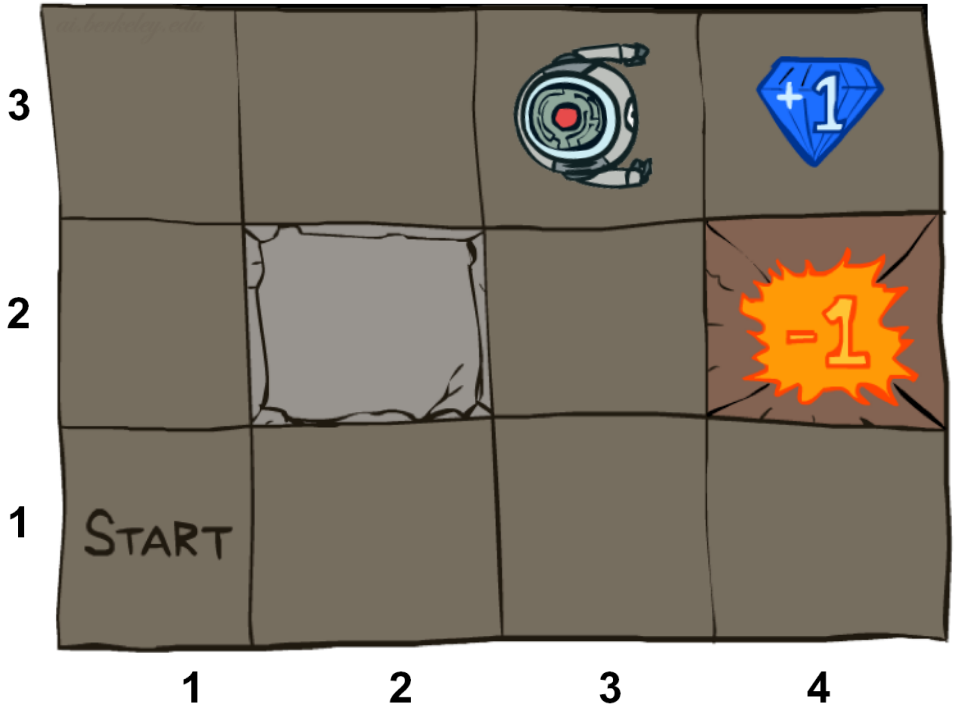
序列决策


π
Agent
a0
a1
aT-1
s0
s1
s2
sT
r0
r1
rT-1
μ0
P
Environment

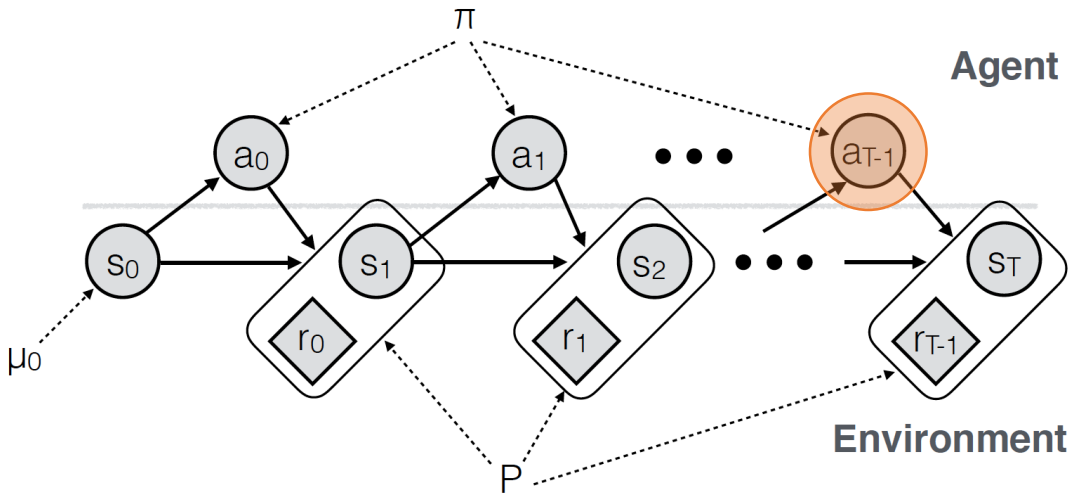

3
2
1
START
+1
-1
1
2
3
4

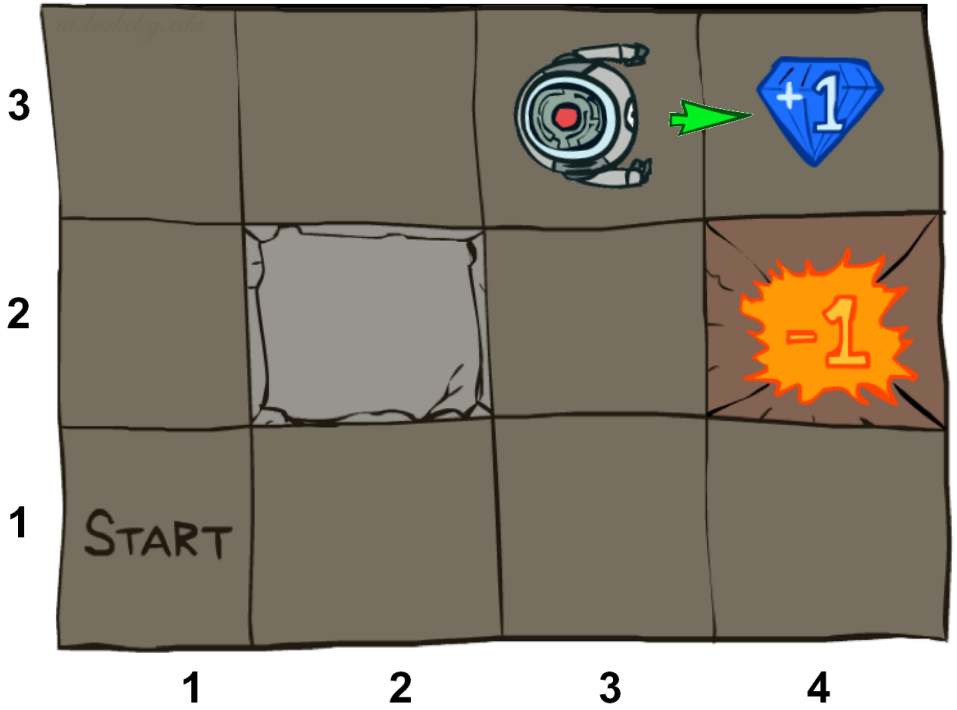
序列决策


π
Agent
a0
a1
aT-1
s0
s1
s2
sT
r0
r1
rT-1
μ0
P
Environment

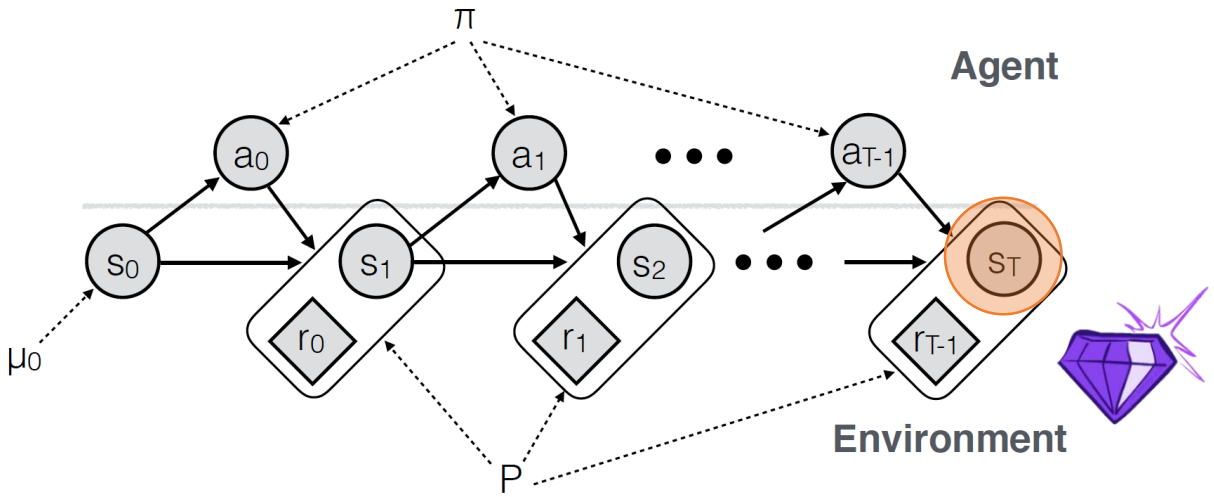

3
2
1
START
+1
-1
1
2
3
4

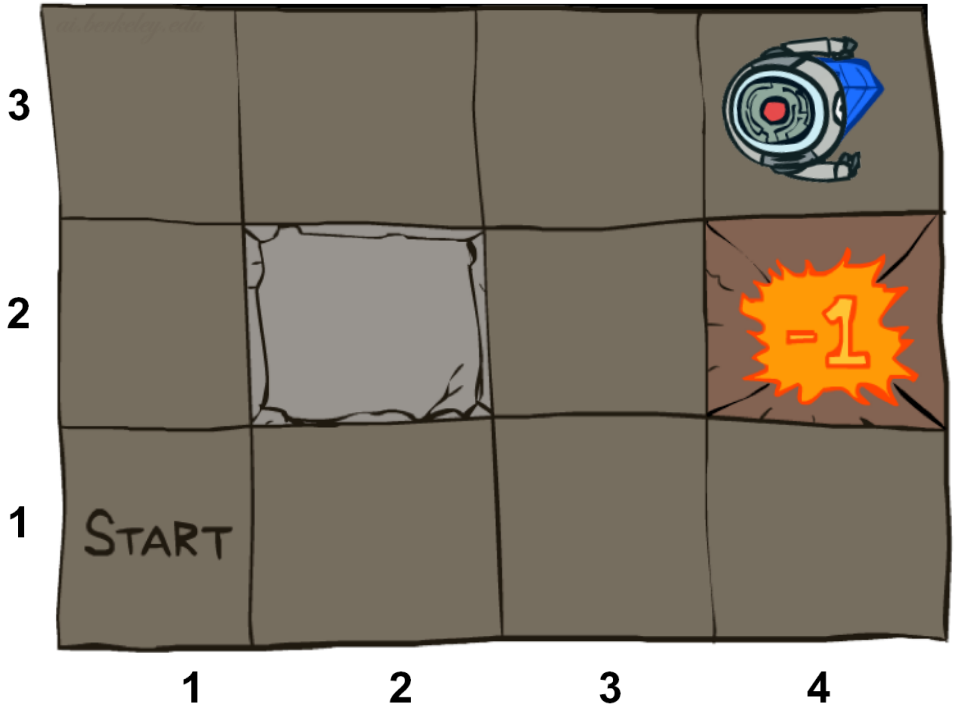
序列决策


π
Agent
a0
a1
aT-1
s0
s1
s2
sT
r0
r1
rT-1
μ0
P
Environment

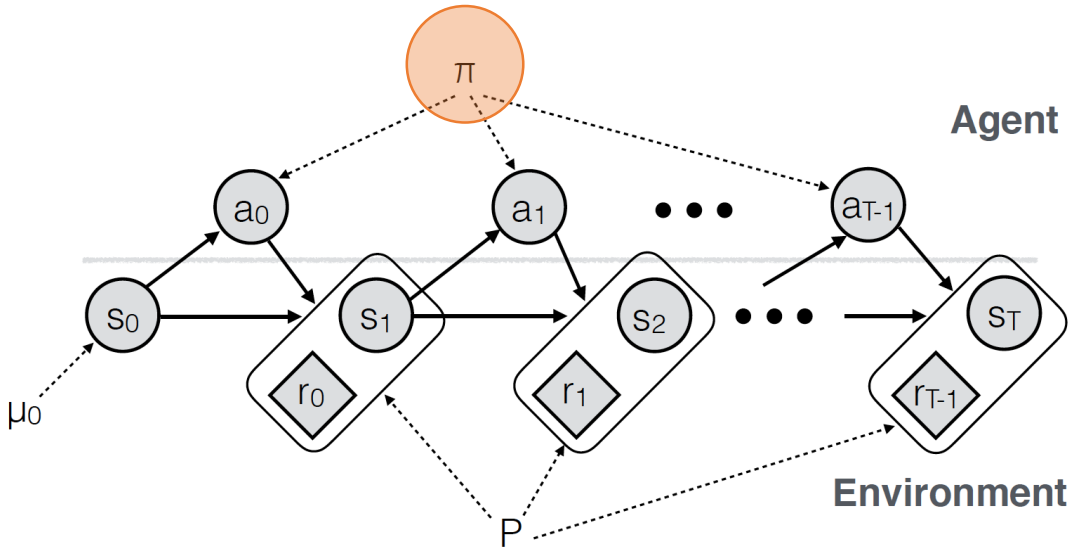

3
2
1
START
+1
-1
1
2
3
4

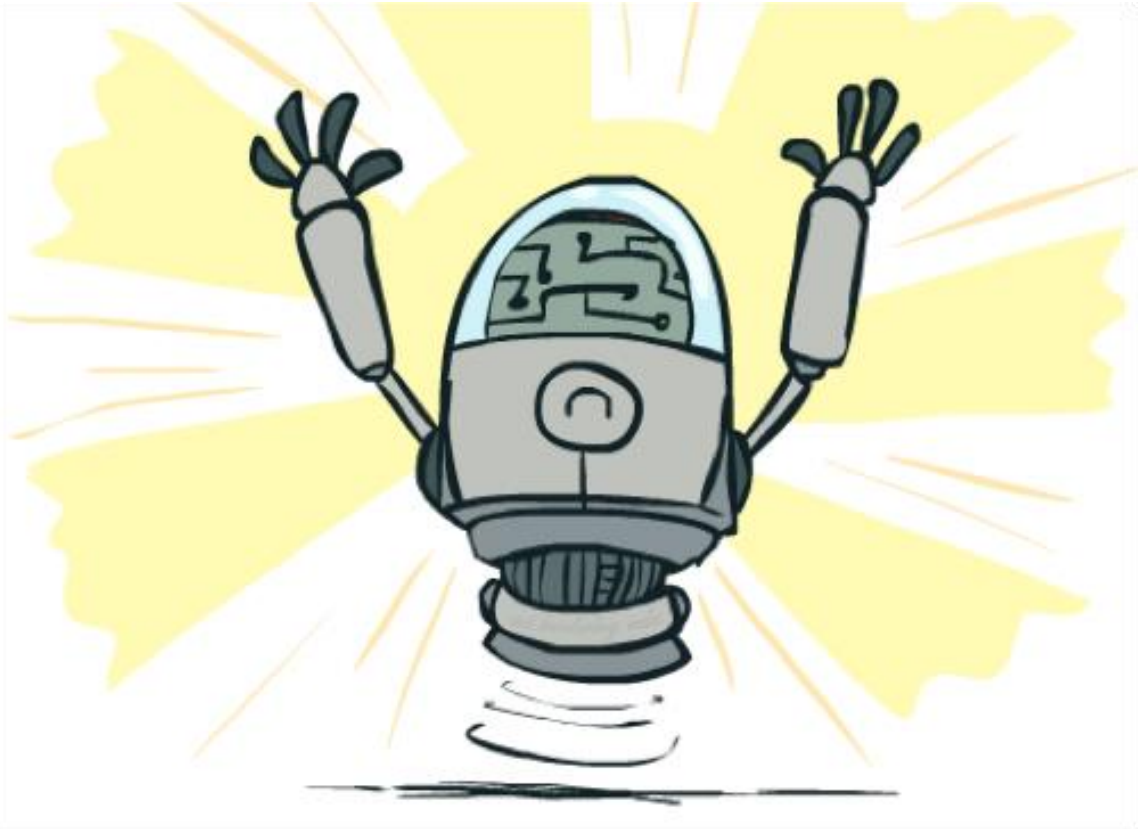
序列决策


π
Agent
a0
a1
aT-1
s0
s1
s2
sT
r0
r1
rT-1
μ0
P
Environment


3
2
1
START
-1
1
2
3
4

序列决策



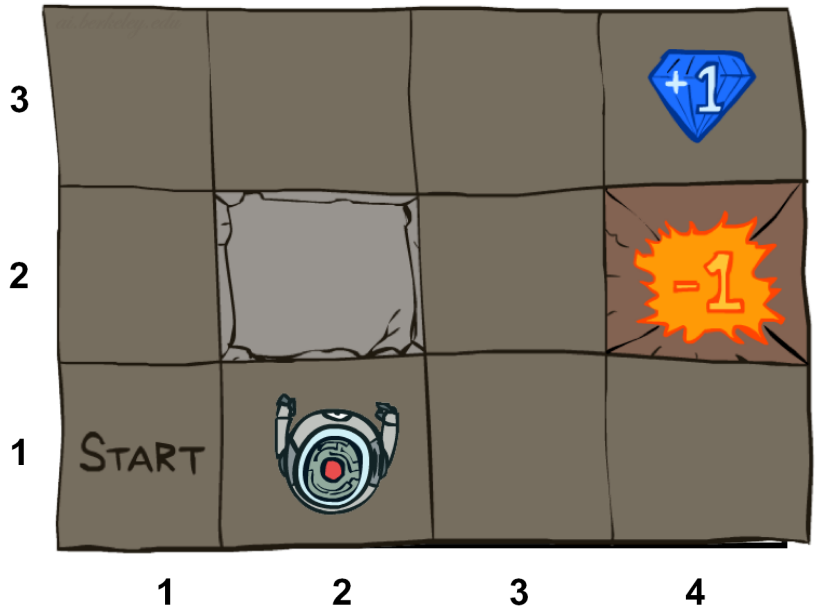


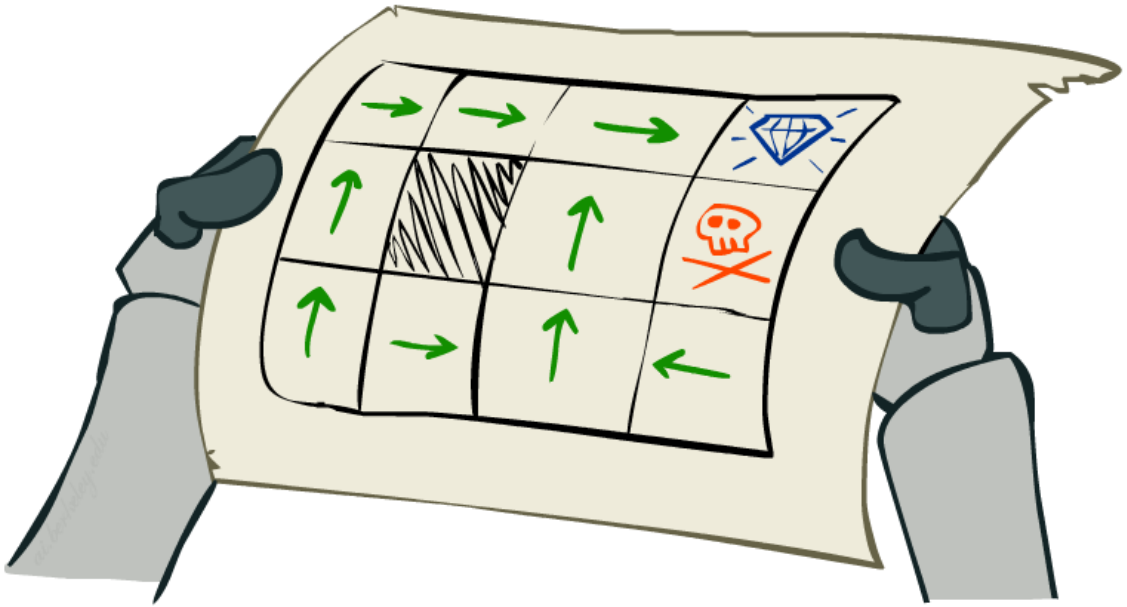
03 第三章

主题：马尔可夫决策过程-策略

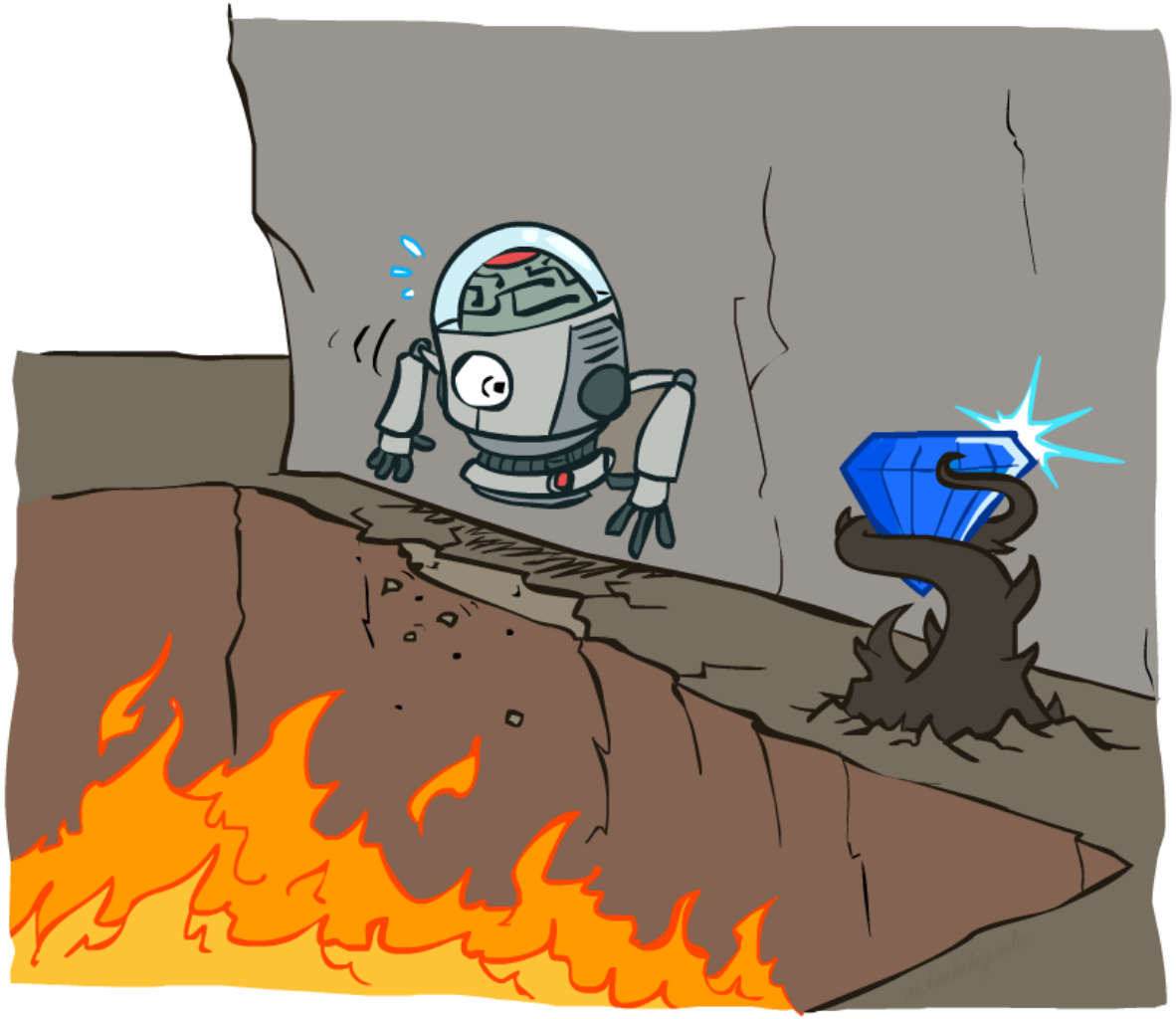
策略

- **策略是agent在环境中行动的“说明书”**
  - 策略 π对每个状态S，决定其应该选择的动作a，即$\pi = P(a|S)$
  - 最优策略 π^* 在任何条件下，永远选择可达到最大效用的动作





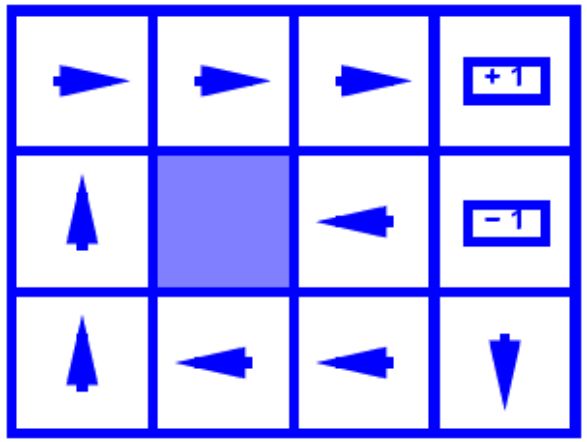
随机性环境下的最优策略



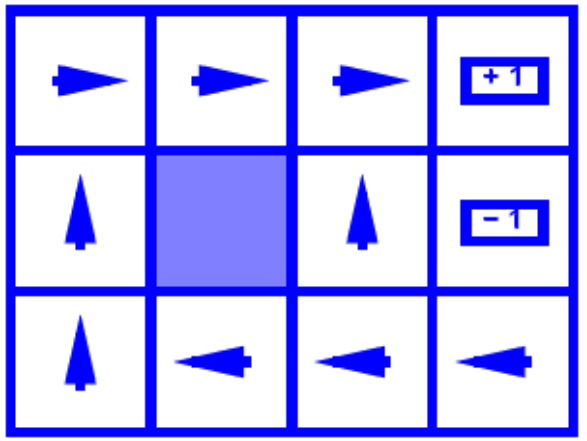
不同设定下的最优策略

- **假设环境中的状态转移具有随机性**

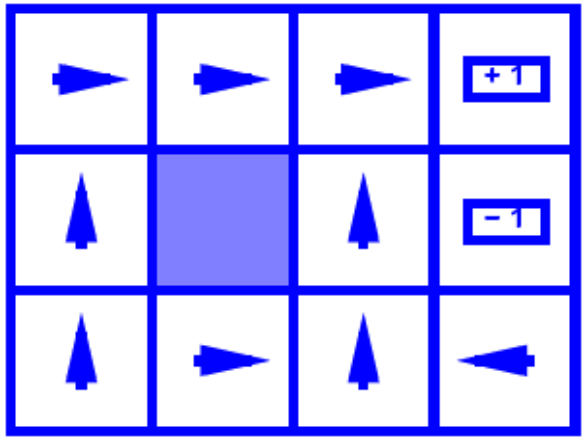
  - 80%概率完成动作，发生预想方向的状态转移
  - 10%概率随机向相邻方向运动



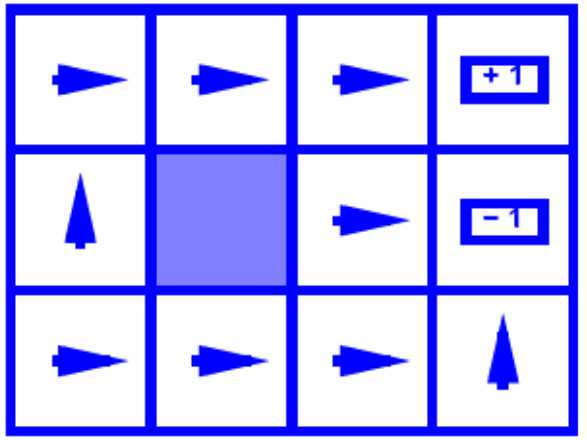
R(s) = -0.01



R(s) = -0.03



R(s) = -0.4



R(s) = -2.0

举例 Atari Games

❑状态

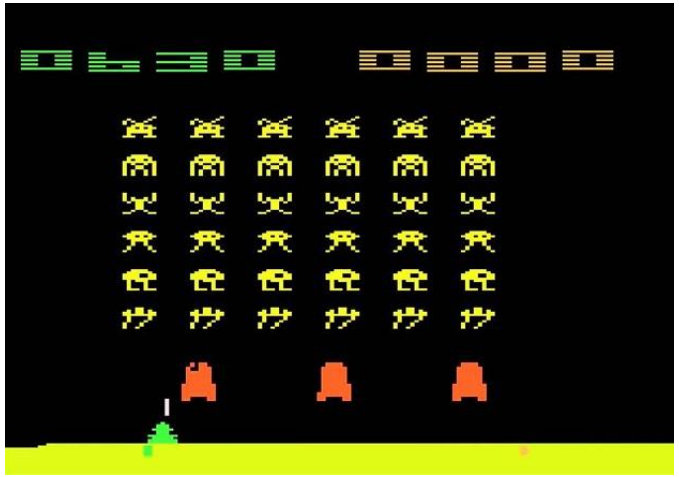
- 原始的游戏视频画帧，或视频中提取的特征

❑动作

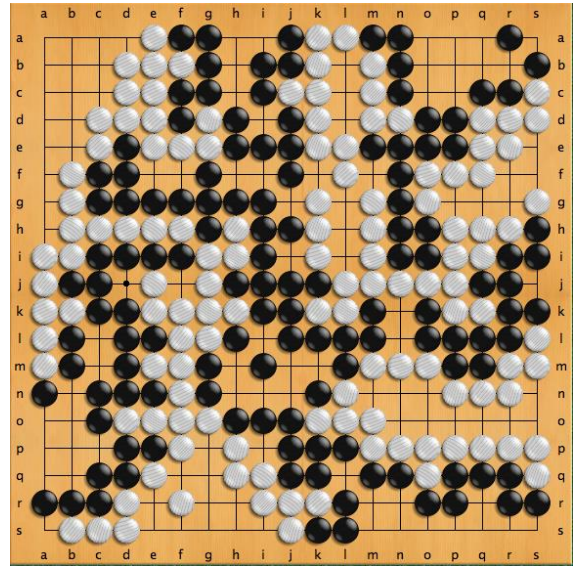
- 手杆方向和开火键

❑奖励

- 屏幕上的得分



举例 AlphaGo



❑状态

- 棋盘上子的位置/状态

❑动作

- 下一子放置的位置

❑奖励

- 最后一步：赢棋(+1)或输棋(-1)
- 其他：0

举例 赛车

❑状态

赛车的速度、位置、方向、加速度等

❑动作

加速、减速、转弯等

❑奖励

和赛道边沿的夹角、时间代价等

04 第四章

主题：OpenAI Gym

OpenAI Gym

https://gym.openai.com/docs

OpenAI gym简介

- **OpenAI gym是一个开源的 开发和评估强化学习(RL)算法的工具平台**
  - Gym environments：轻量级的RL实验环境

Classic control
Classic control problems from the RL literature.

CartPole-v0
Balance a pole on a cart (for a short time).

CartPole-v1
Balance a pole on a cart.

Acrobot-v1
Swing up a two-link robot.

MountainCar-v0
Drive up a big hill.

MountainCarContinuous-v0
Drive up a big hill with continuous control.

Pendulum-v0
Swing up a pendulum.

Doom
Doom environments based on VizDoom.

meta-Doom-v0
(experimental) (by @ppaquette)
Mission #1 to #9 - Beat all 9 Doom missions.

DoomBasic-v0 (experimental) (by @ppaquette)
Mission #1 - Kill a single monster using your pistol.

DoomCorridor-v0
(experimental) (by @ppaquette)
Mission #2 - Run as fast as possible to grab a vest.

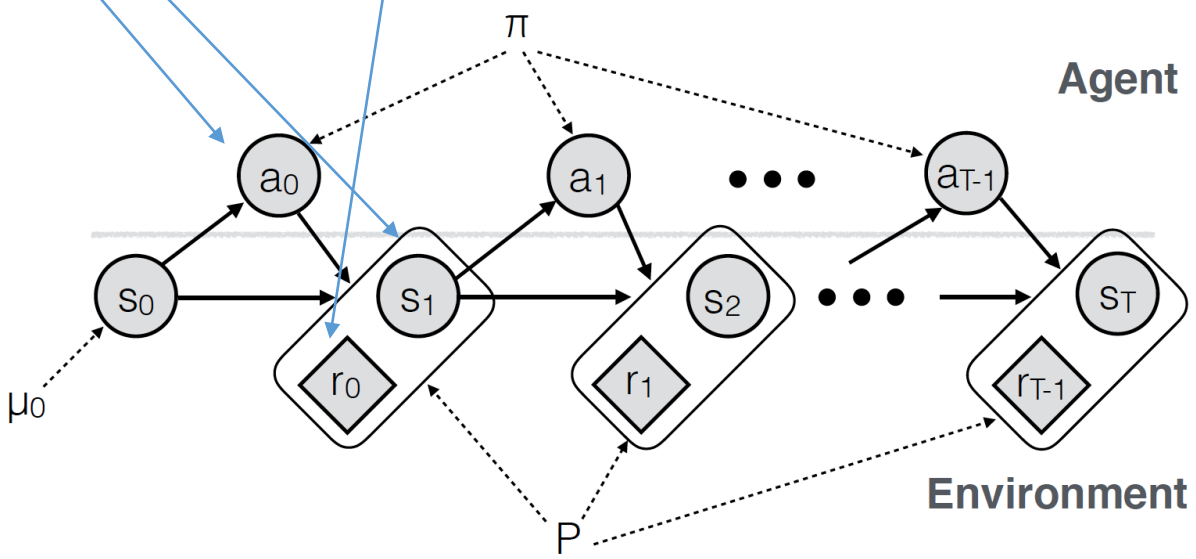
DoomDefendCenter-v0
(experimental) (by @ppaquette)
Mission #3 - Kill enemies

DoomDefendLine-v0
(experimental) (by @ppaquette)
Mission #4 - Kill enemies

DoomHealthGathering-v0
(experimental) (by @ppaquette)
Mission #5 - Learn to

Gym API

```
action = env.action_space.sample()
observation, reward, done, info = env.step(action)
```



小结

1、人工智能与深度学习、强化学习——感知与决策

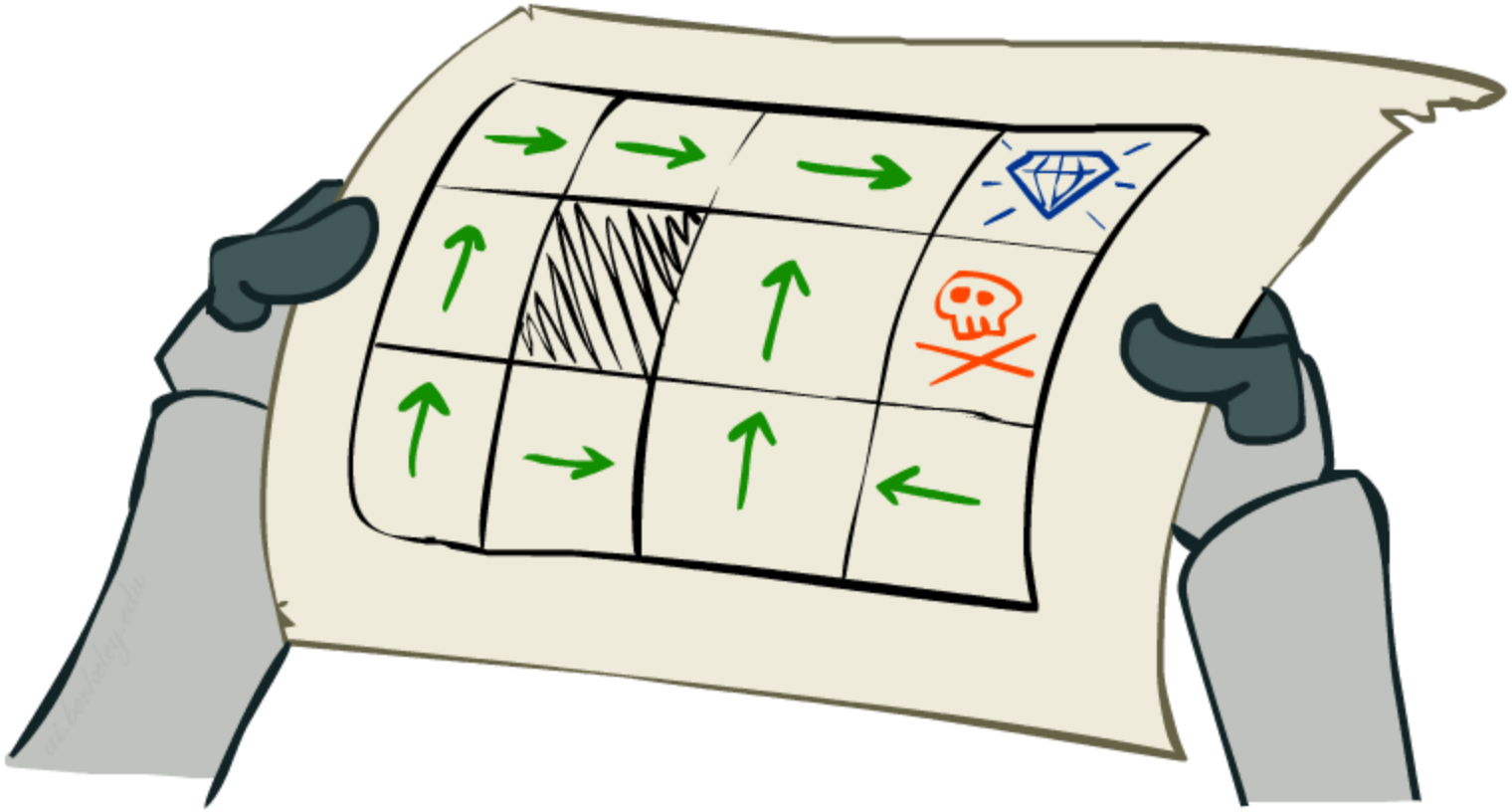
2、马尔可夫决策过程-定义 —— S, a, r, P, γ

3、马尔可夫决策过程-策略 —— π, π^*

4、OpenAI Gym

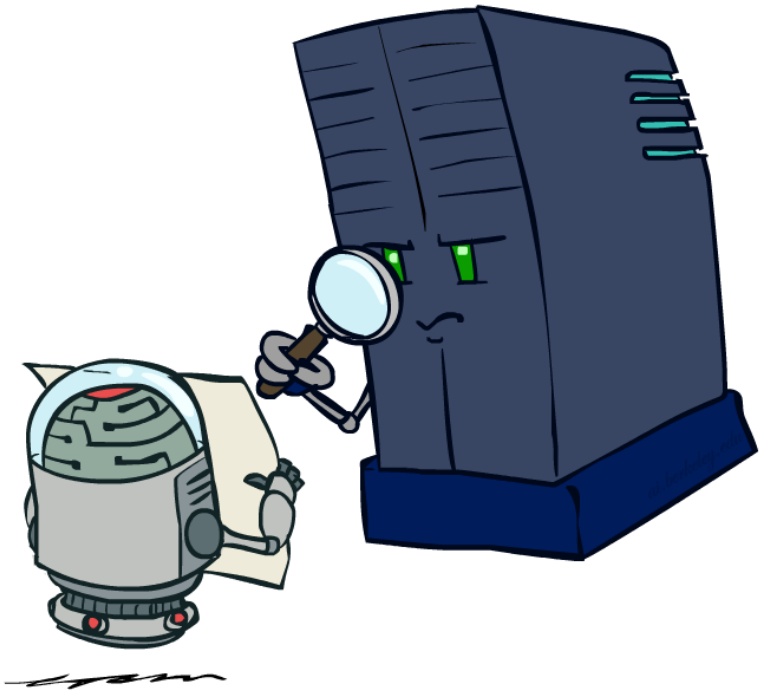
Q&A

下节课内容预告

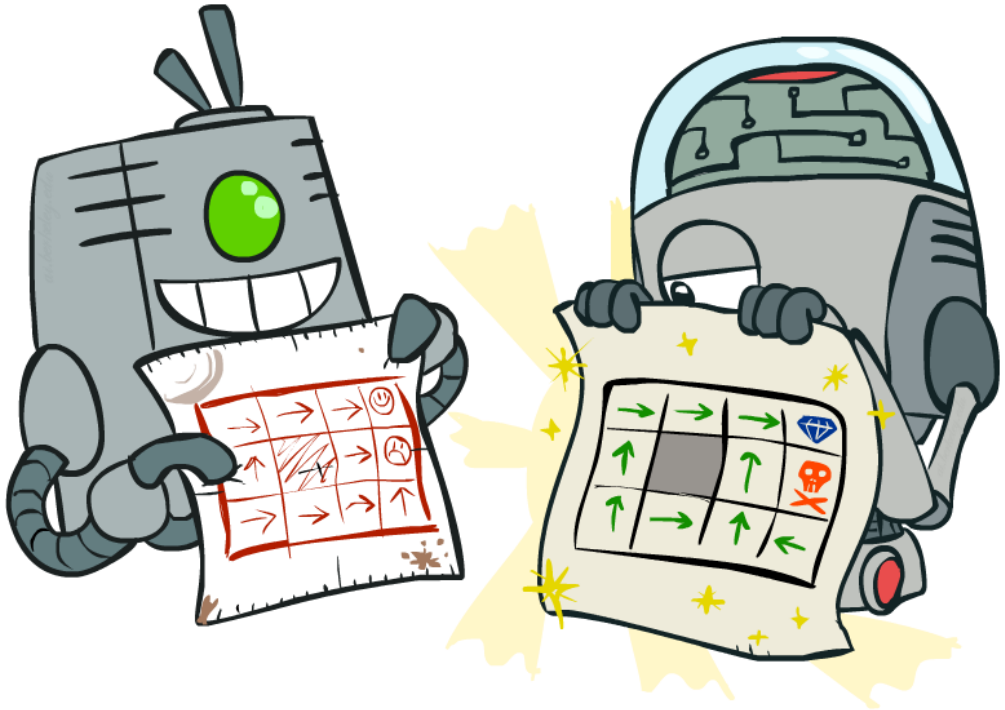


下节课内容预告——MDP求解

评估 (Evaluation)



优化 (control)



Thanks

扫描二维码做课程评估啦~ ~