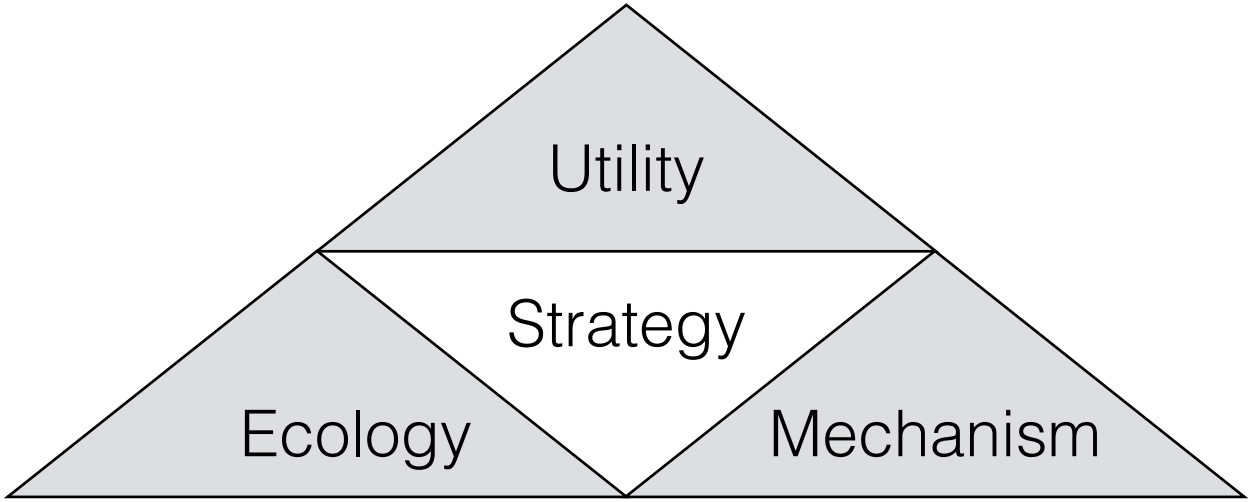
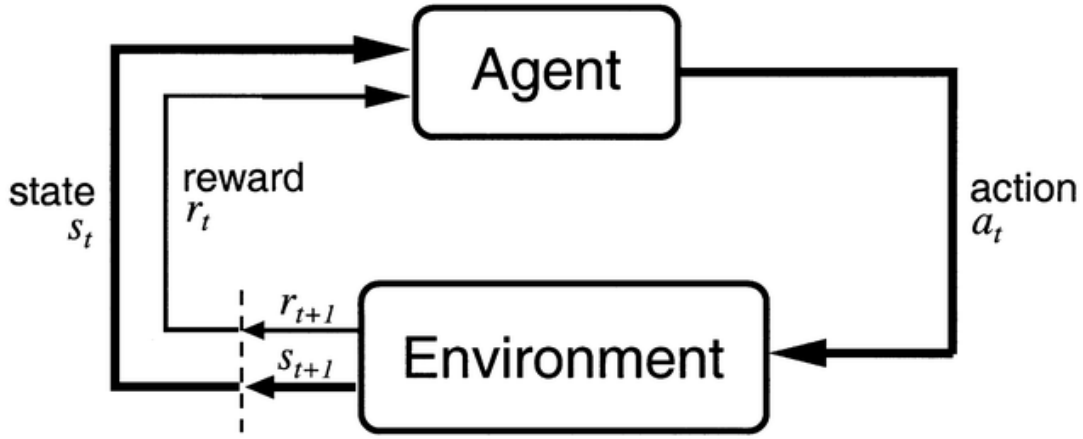
Reinforcement Learning

Andrew Howes
Summer School 2017

Computational rationality



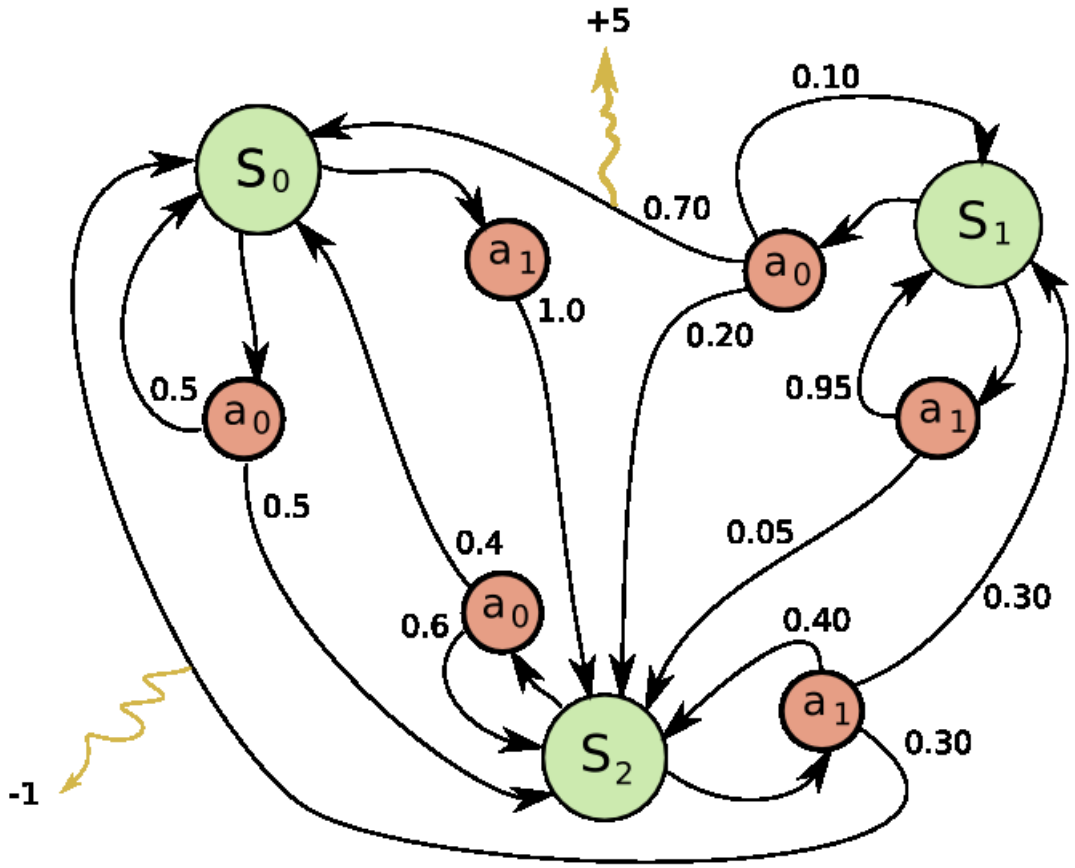
Markov Decision Process (MDP)



a quintuple

- *S* is the set of states
- *A* is the set of actions
- $P_a(s,s')$ the probability that action *a* in state *s* at time *t* will lead to state *s'* at time *t*+1,
- $R_a(s,s')$ is the immediate reward (or expected immediate reward) received after transitioning from state s to state s', due to action a.
- gamma is the discount factor, which represents the difference in importance between future rewards and present rewards.

Markov Decision Process

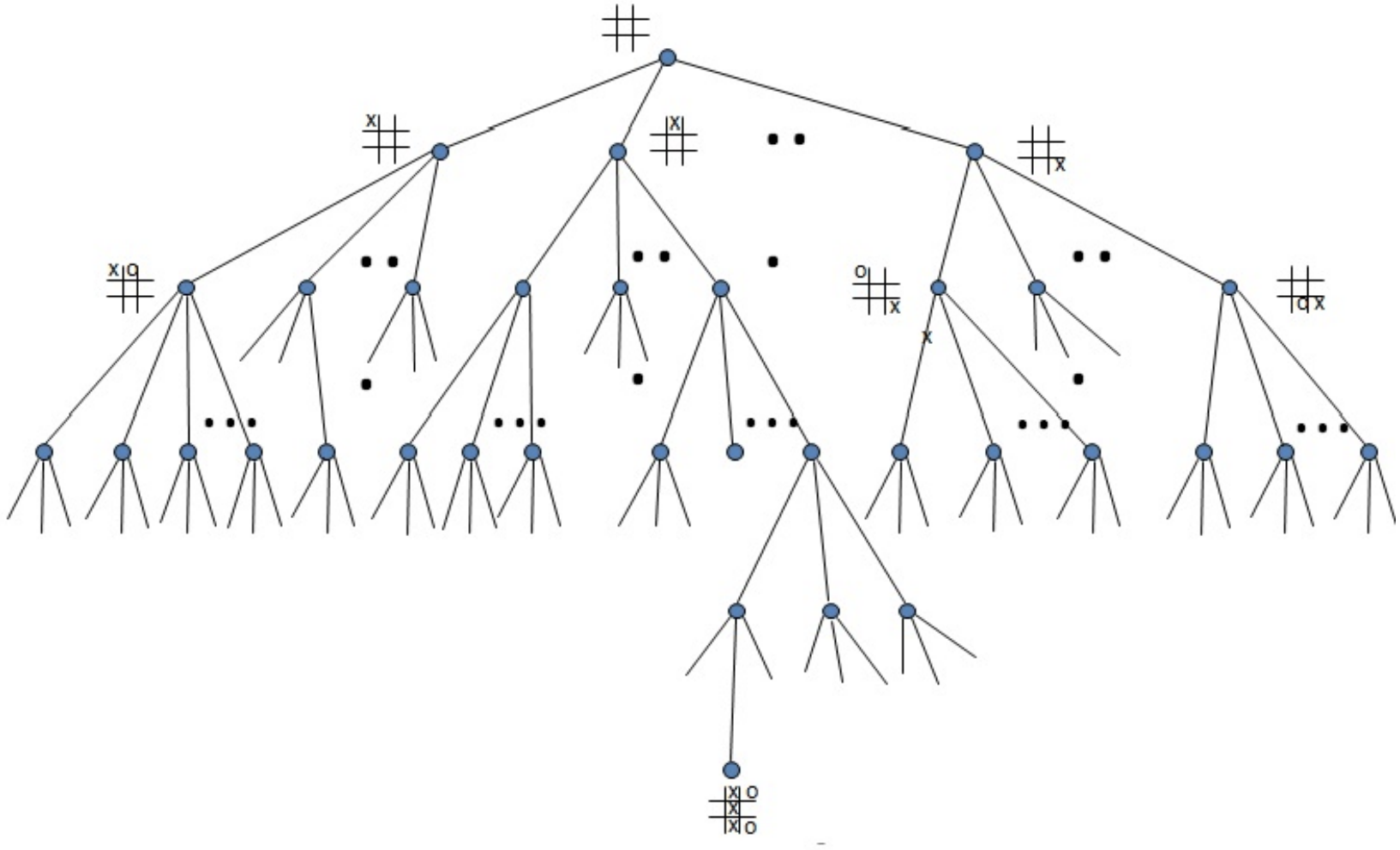


- The problem is to find a **policy** for the decision maker: a function that specifies the action that the decision maker will choose when in state *s*.

- Because of the Markov property the **policy** for can be written as a function of the current state.

Q-learning

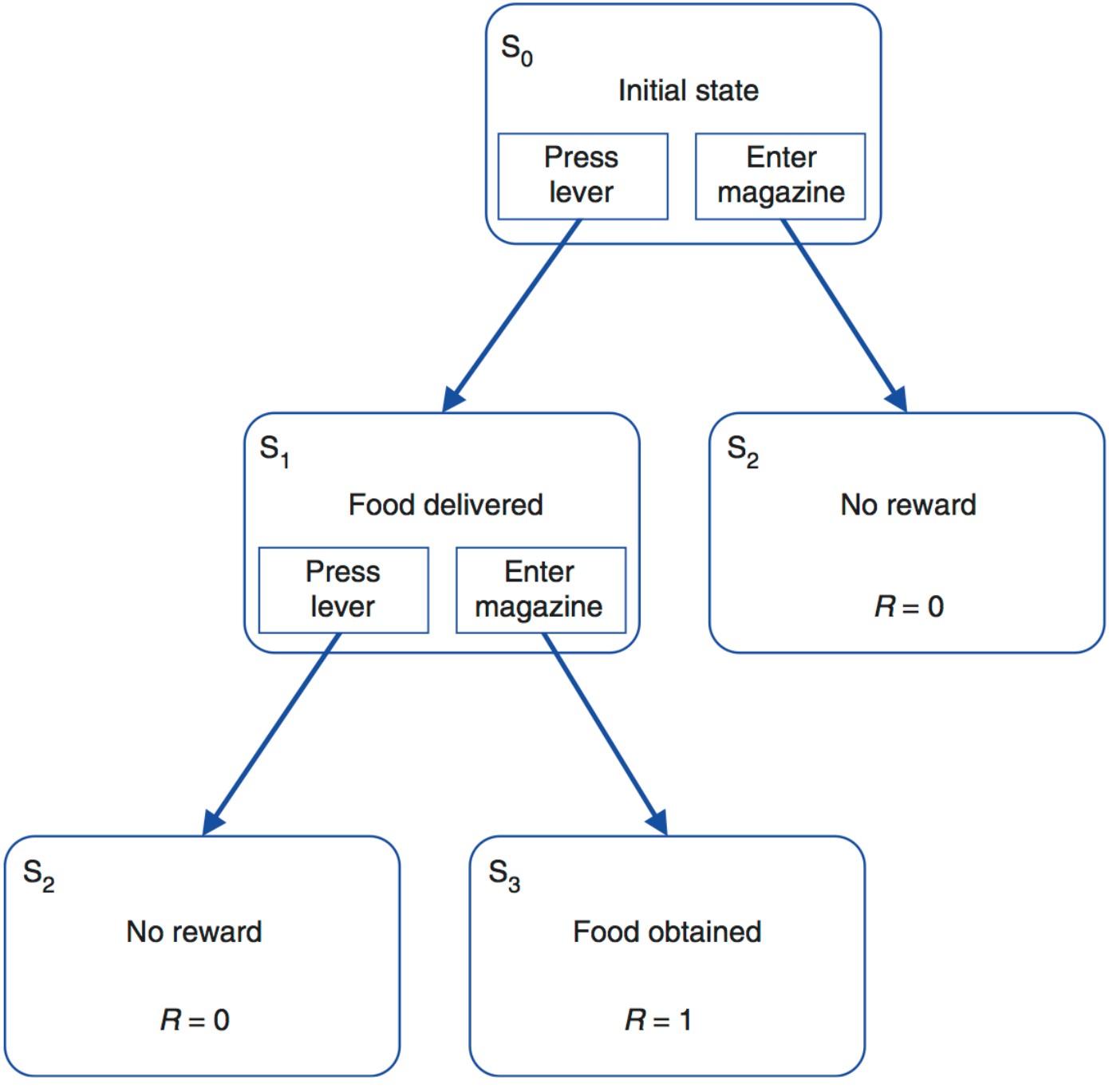
$$Q(s_t, a_t) \leftarrow \underbrace{Q(s_t, a_t)}_{\text{old value}} + \underbrace{\alpha}_{\text{learning rate}} \cdot \left(\overbrace{\underbrace{r_{t+1}}_{\text{reward}} + \underbrace{\gamma}_{\text{discount factor}} \cdot \underbrace{\max_a Q(s_{t+1}, a)}_{\text{estimate of optimal future value}}}^{\text{learned value}} - \underbrace{Q(s_t, a_t)}_{\text{old value}} \right)$$



http://cse-wiki.unl.edu/wiki/index.php/Q_Learning

http://time.com/4720854/vending-machines-25-second-delay-study/

Japanese lettuce vending machine


S_0
Initial state
Press lever
Enter magazine
S_1
Food delivered
Press lever
Enter magazine
S_2
No reward
$R = 0$
S_2
No reward
$R = 0$
S_3
Food obtained
$R = 1$

breakout

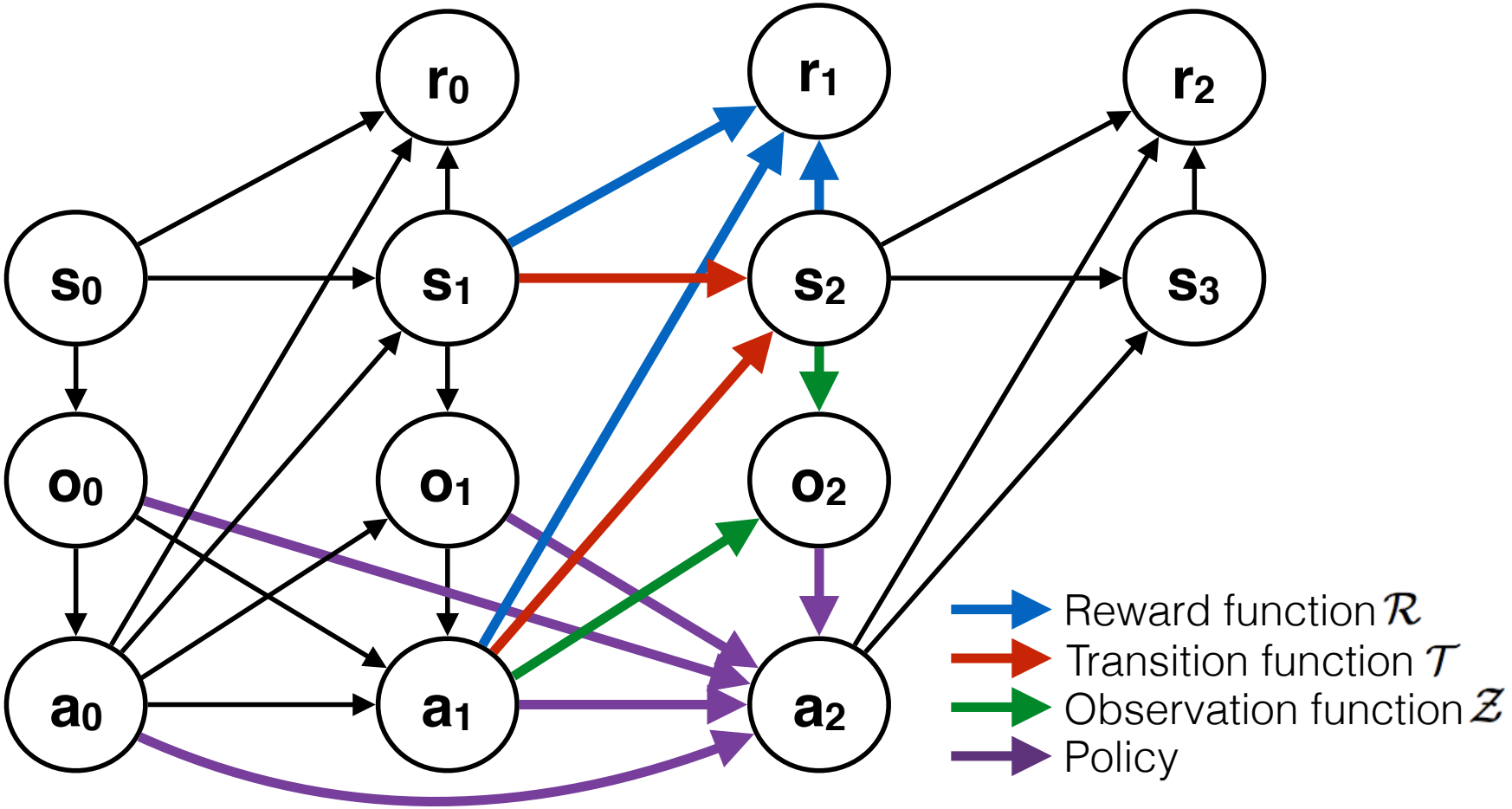
Questions

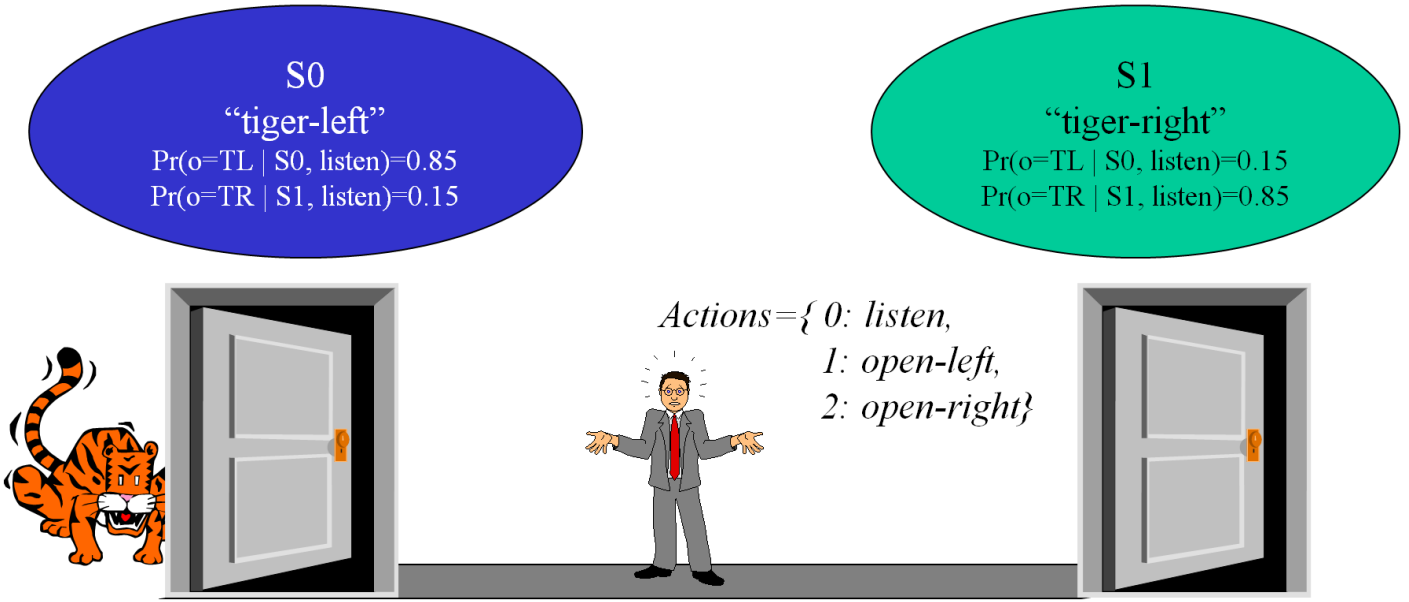
- What is the main difference between the vending machine task and the Iowa Gambling task? Why do we need Q-learning for one and not the other?

- Do you believe that you discount future rewards? Give an example?

- What other tasks might be modelled with an MDP?

- What else might bound human performance on this task?


r0
r1
r2
s0
s1
s2
s3
o0
o1
o2
a0
a1
a2
Reward function $\mathcal{R}$
Transition function $\mathcal{T}$
Observation function $\mathcal{Z}$
Policy



Reward Function

- *Penalty for wrong opening: -100*
- *Reward for correct opening: +10*
- *Cost for listening action: -1*

Observations

- *to hear the tiger on the left (TL)*
- *to hear the tiger on the right(TR)*

solving POMDPs

- State estimation to track the probability distribution as new observations are made.
- learn a policy $\boldsymbol{\pi}$ conditioned on history… or on a probability distribution over possible states.

Reading

- Sutton, R. S., & Barto, A. G. (1998). Reinforcement learning: An introduction. MIT press.