APPLIED MECHANISM DESIGN FOR SOCIAL GOOD

JOHN P DICKERSON

Lecture #10 – 02/27/2020

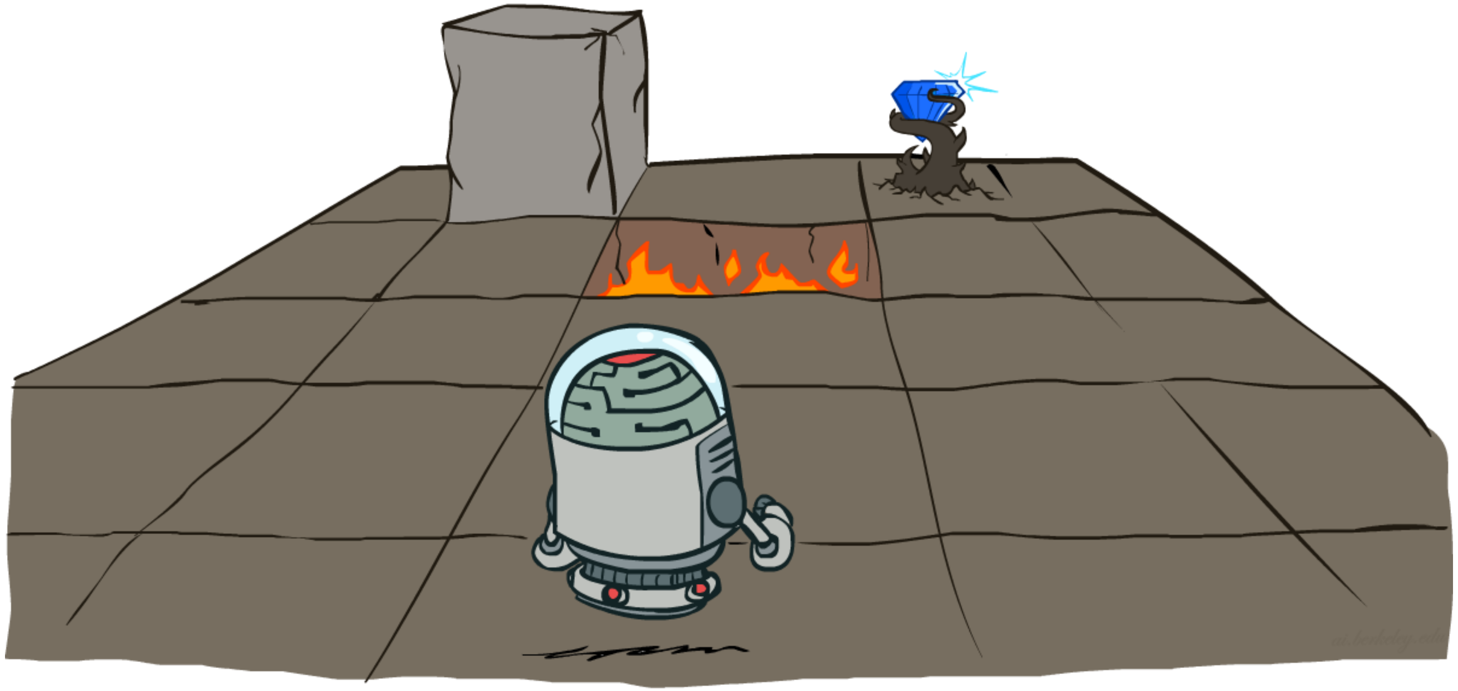
CMSC828M
Tuesdays & Thursdays
2:00pm – 3:15pm

COMPUTER SCIENCE
UNIVERSITY OF MARYLAND

ANNOUNCEMENTS

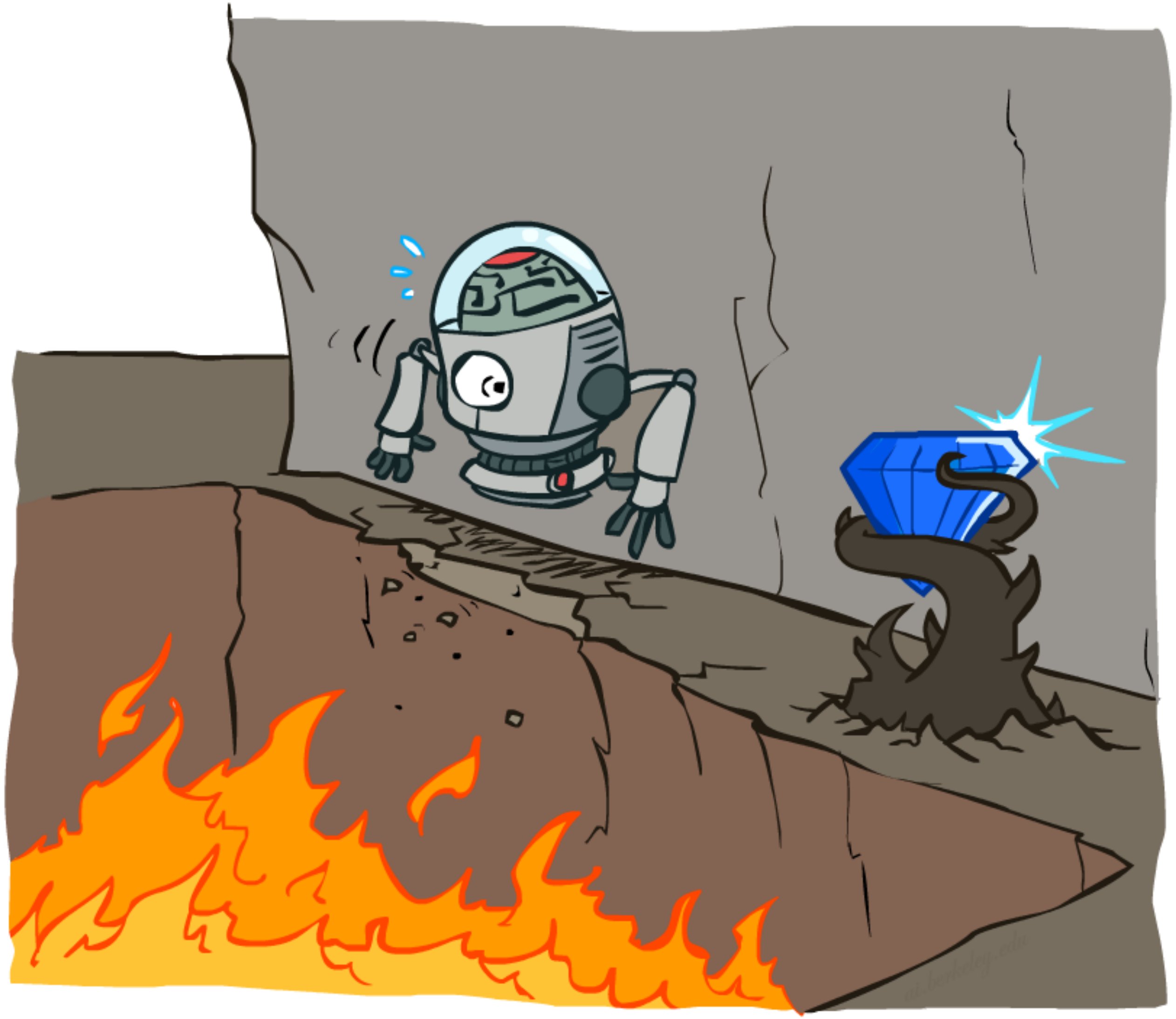
Projects?

MARKOV DECISION PROCESSES

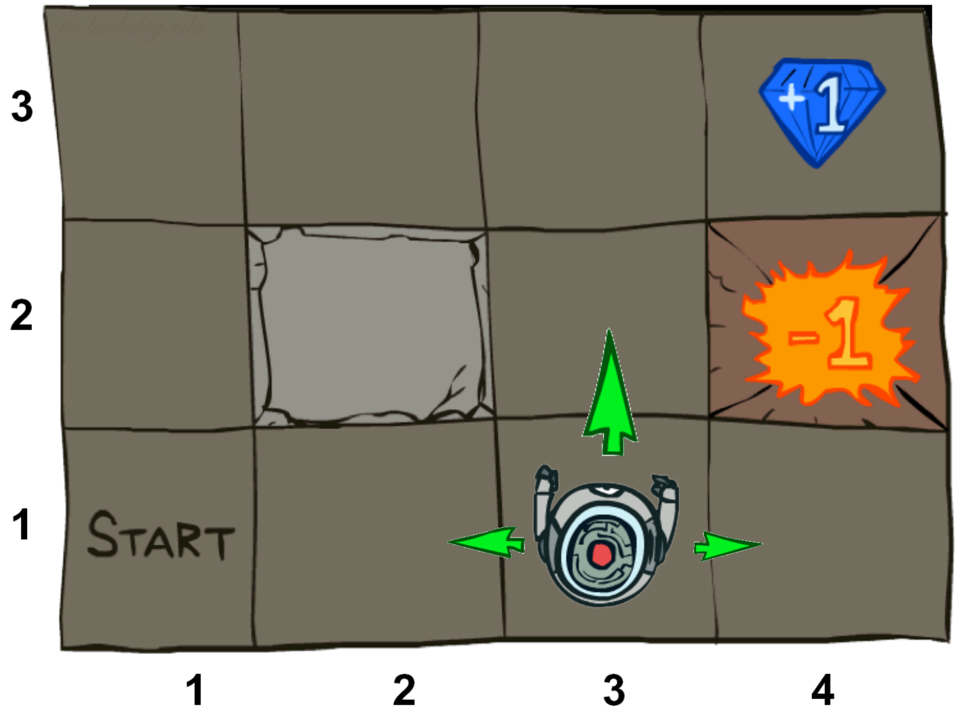


Slide credits: CMU AI and http://ai.berkeley.edu

NON-DETERMINISTIC SEARCH



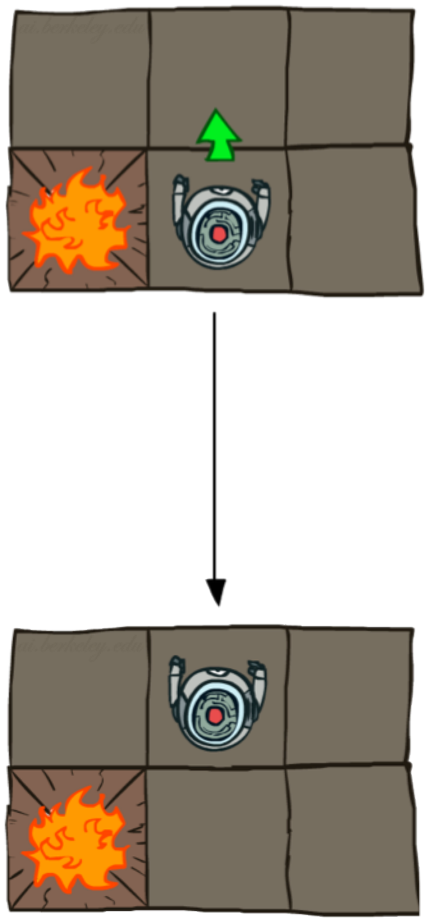
EXAMPLE: GRID WORLD



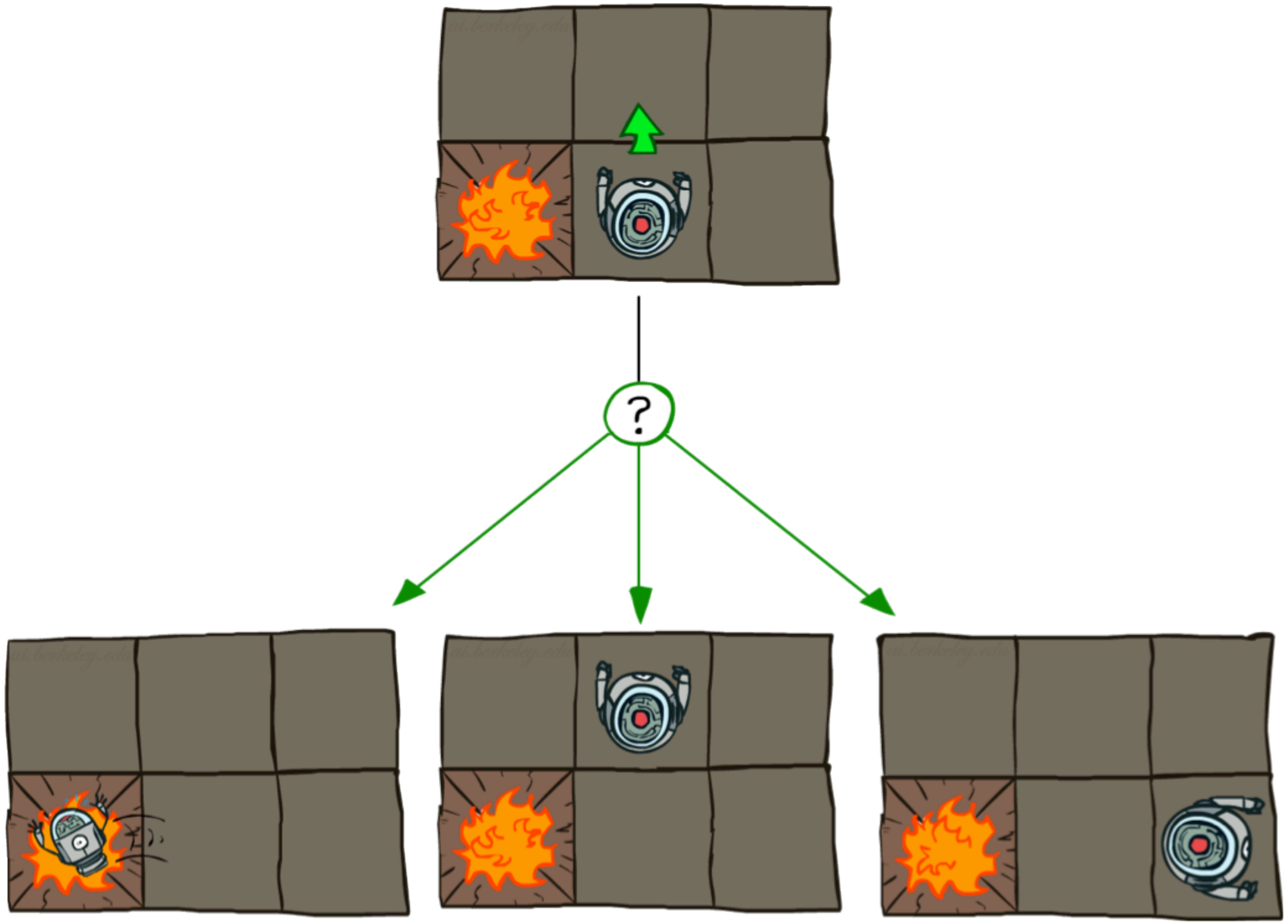
- A maze-like problem
  - The agent lives in a grid
  - Walls block the agent's path
- Noisy movement: actions do not always go as planned
  - If agent takes action North
    - 80% of the time: Get to the cell on the North (if there is no wall there)
    - 10%: West; 10%: East
  - If path after roll dice blocked by wall, stays put
- The agent receives rewards each time step
  - "Living" reward (can be negative)
  - Additional reward at pit or target (good or bad) and will exit the grid world afterward
- Goal: maximize sum of rewards

GRID WORLD ACTIONS

Deterministic Grid World



Stochastic Grid World

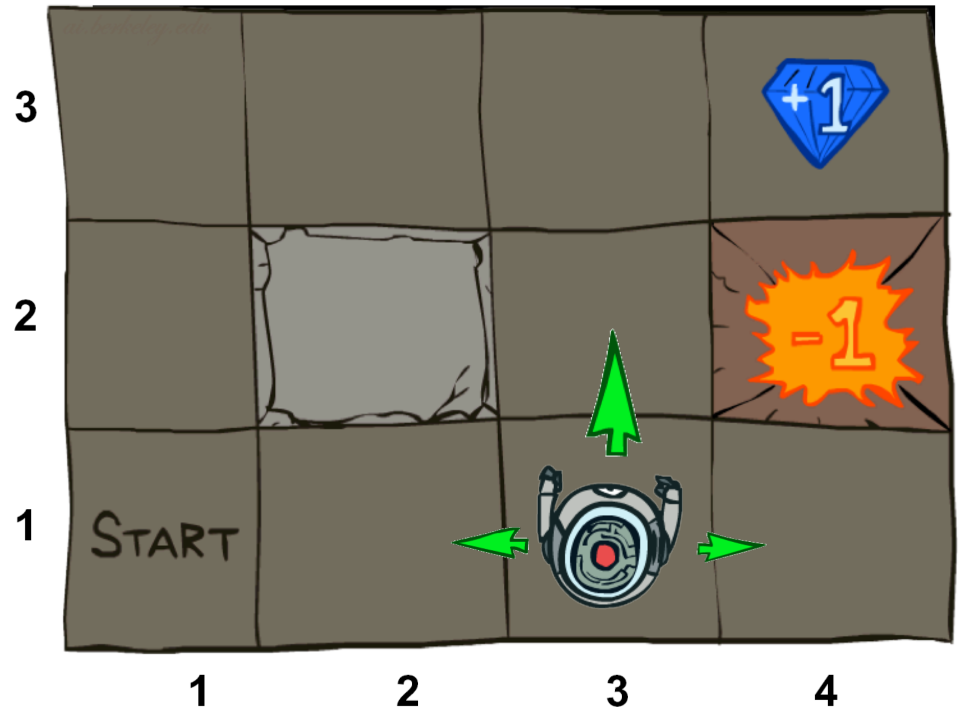


MARKOV DECISION PROCESS (MDP)

An MDP is defined by a tuple (S,A,T,R):

- S: a set of states
- A: a set of actions
- T: a transition function
  - T(s, a, s') where s ∈ S, a ∈ A, s' ∈ S is P(s'| s, a)
- R: a reward function
  - R(s, a, s') is reward at this time step
  - Sometimes just R(s) or R(s')

- Sometimes also have
  - γ: discount factor (introduced later)
  - μ: distribution of initial state (or just start state s_0)
  - Terminal states: processes end after reaching these states



The Grid World problem as an MDP

$$\boxed{R(s_{4,2}, exit, s_{virtual\_terminal})=-1}$$

R($s_{4,2}$)=-1, no virtual terminal state

How to define the terminal states & reward function for the Grid World problem?

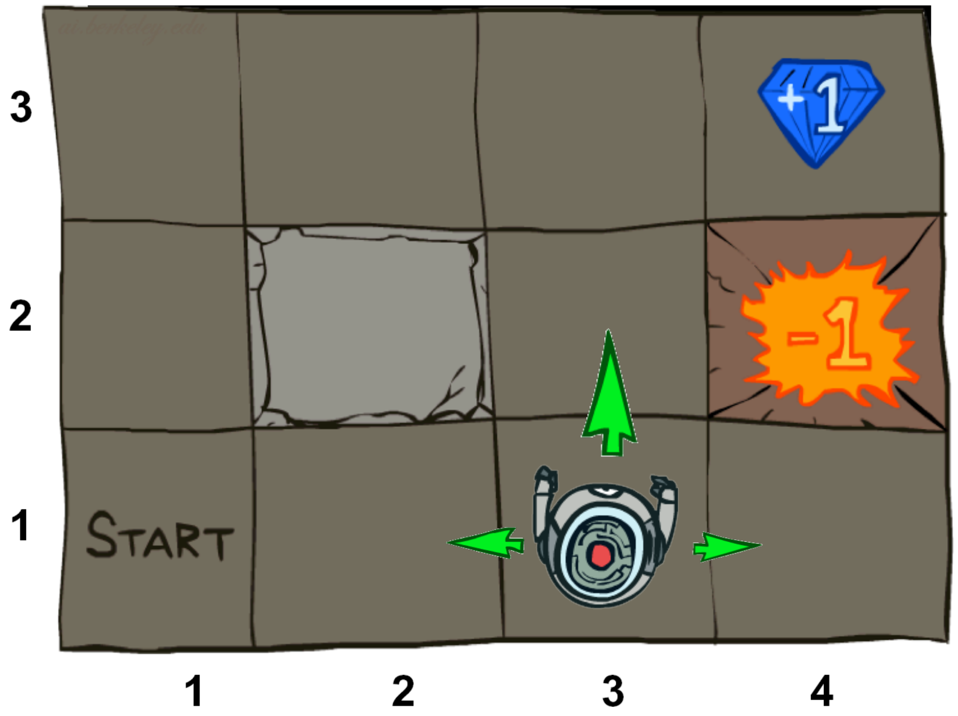
MARKOV DECISION PROCESS (MDP)

An MDP is defined by a tuple **(S,A,T,R)**

Why is it called Markov Decision Process?

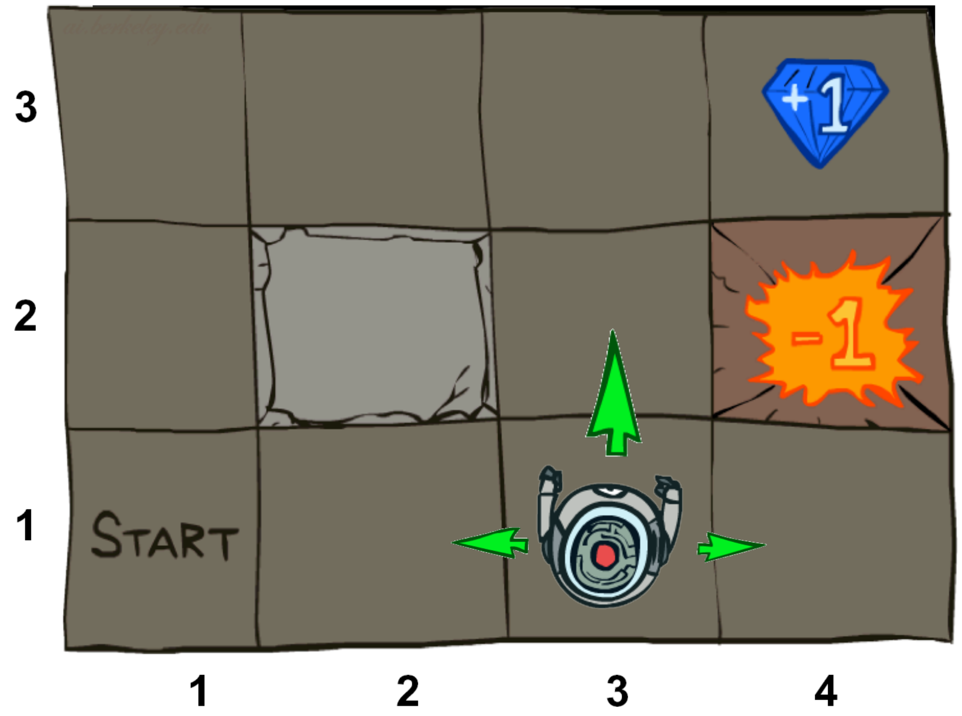
Decision:

Process:



MARKOV DECISION PROCESS (MDP)

An MDP is defined by a tuple (S,A,T,R)



Why is it called Markov Decision Process?

Decision:

Agent decides what action to take at each time step

Process:

The system (environment + agent) is changing over time

WHAT IS "MARKOVIAN" ABOUT MDPS?

Markov property: Conditional on the present state, the future and the past are independent

With respect to MDPs, it means outcome of an action depend only on current state

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

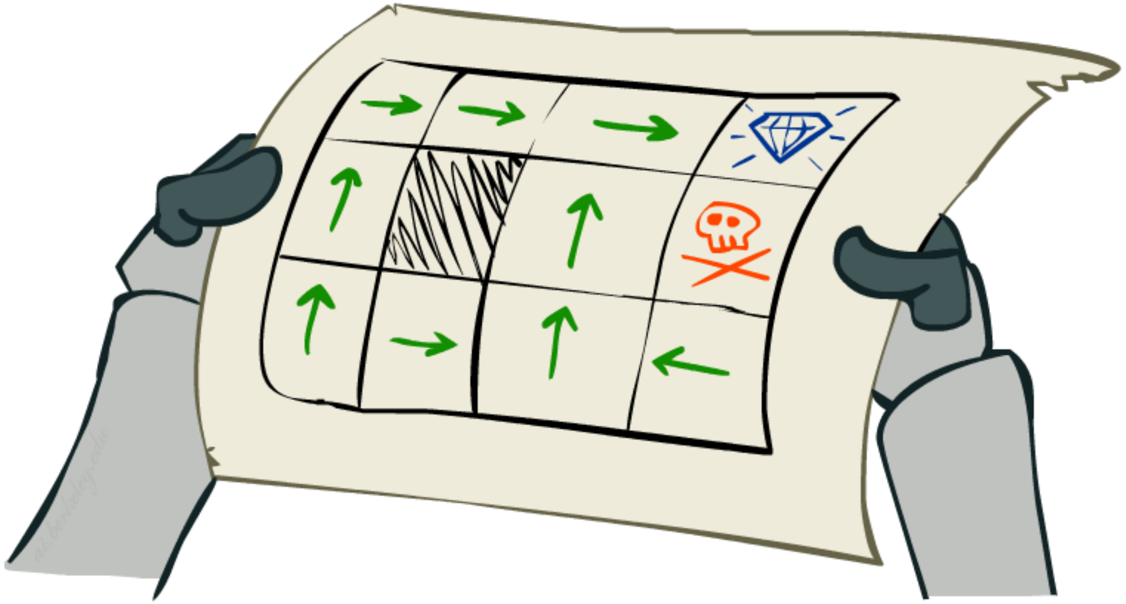
$$=$$

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

Andrey Markov
(1856-1922)
Russian
mathematician

POLICIES

In deterministic single-agent search problems, w sequence of actions, from start to a goal



For MDPs, we focus on policies

- Policy = map of states to actions
- $\pi(s)$ gives an action for state s

We want an optimal policy $\pi^*: S \rightarrow A$

- An optimal policy is one that maximizes expected utility if followed

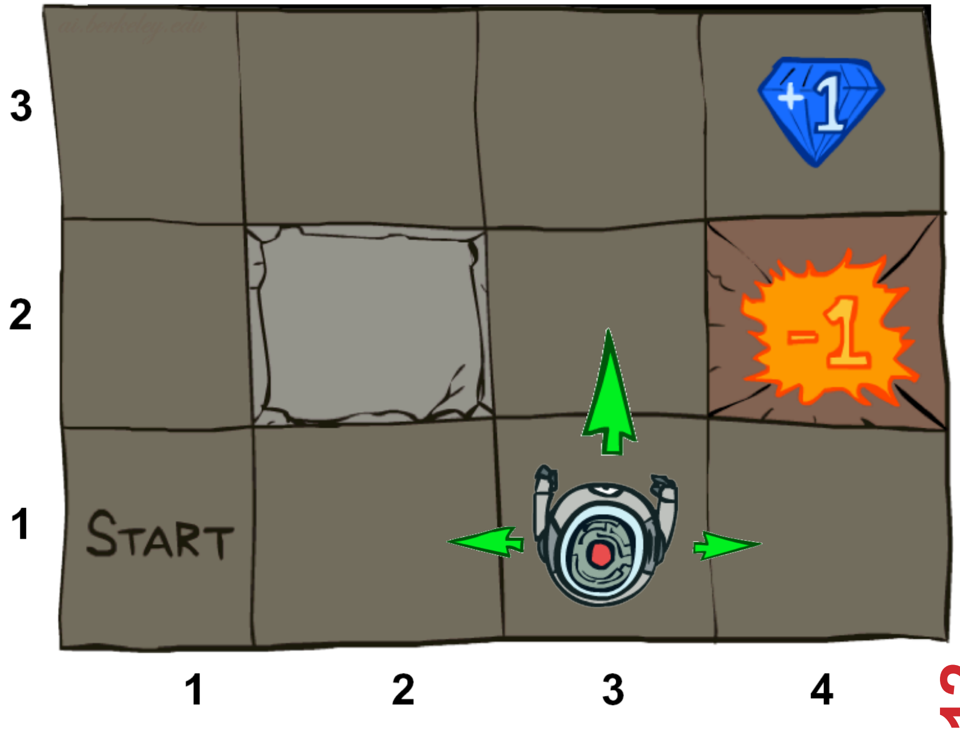
POLICIES

Recall: An MDP is defined S,A,T,R

Keep S,A,T fixed, optimal policy may vary given different R

What is the optimal policy if R(s,a,s')=-1000 for all states other than pit and target?

What is the optimal policy if R(s,a,s')=0 for all states other than pit and target, and reward=1000 and -1000 at pit and target respectively?

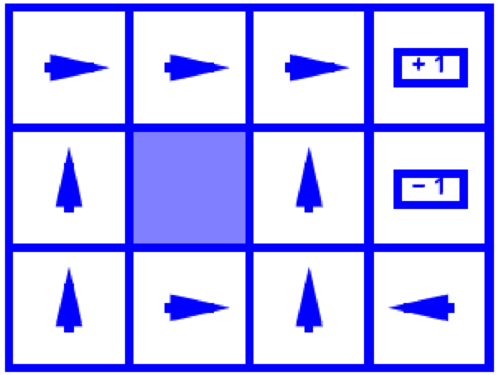


DISCUSSION POINT!

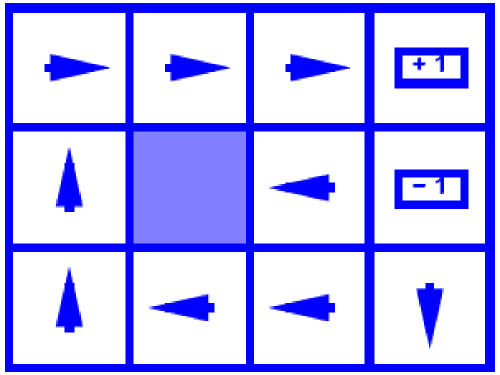
{A, B, C, D} are optimal policies for one of each of the following "reward for living" scenarios: {-0.01, -0.03, -0.04, -2.0}. Which policy maps to which reward setting?

I. {B, A, C, D}
II. {B, C, A, D}
III. {C, B, A, D}
IV. {D, A, C, B}

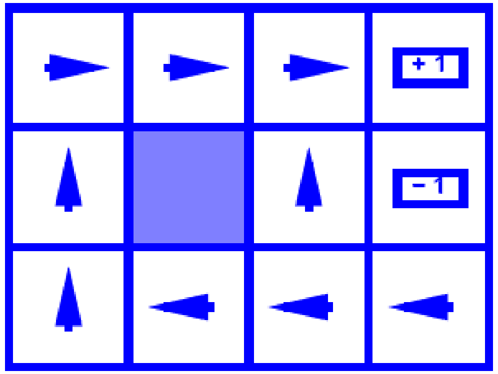
A)



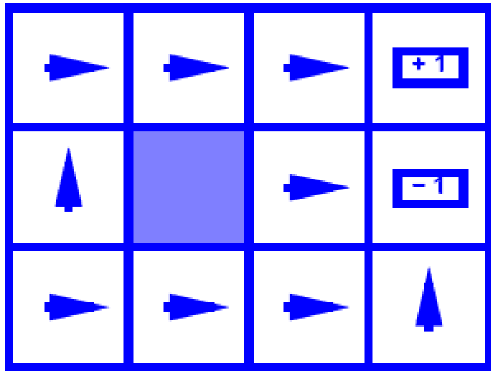
B)



C)



D)

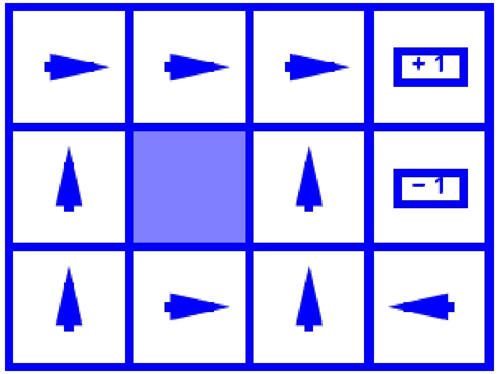


DISCUSSION POINT!

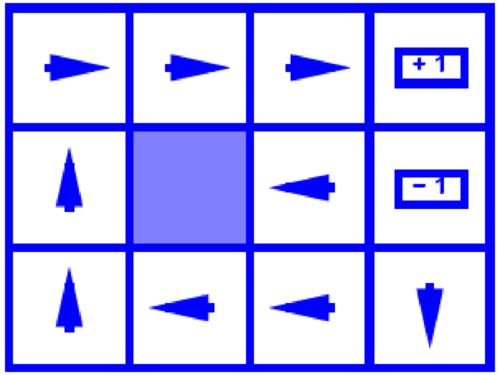
{A, B, C, D} are optimal policies for one of each of the following "reward for living" scenarios: {-0.01, -0.03, -0.04, -2.0}. Which policy maps to which reward setting?

I. {B, A, C, D}
II. {B, C, A, D}
III. {C, B, A, D}
IV. {D, A, C, B}

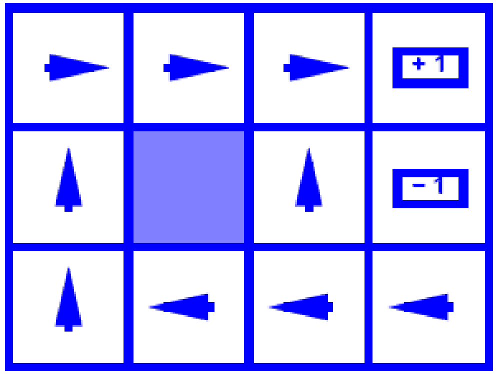
A)



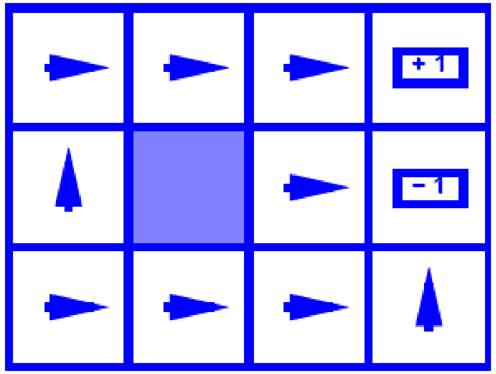
B)



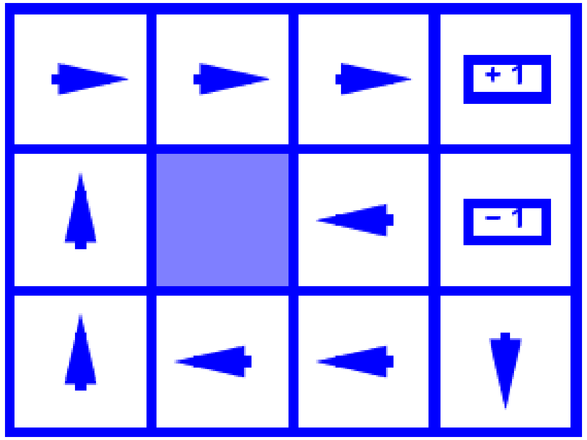
C)



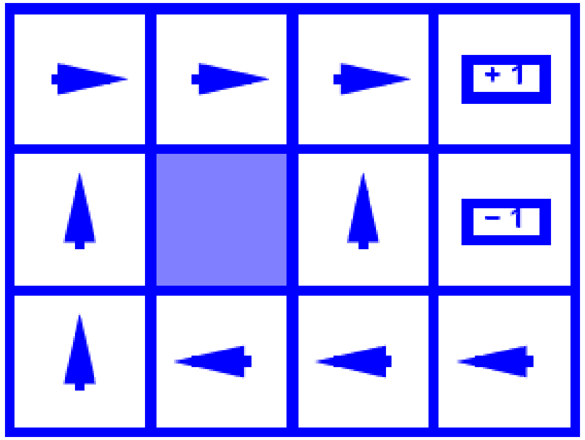
D)



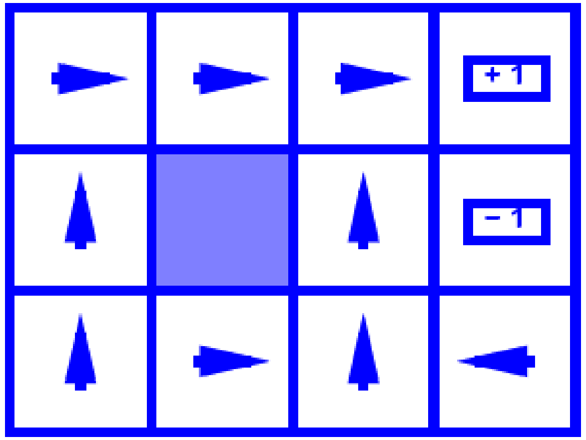
DISCUSSION POINT! POLICIES



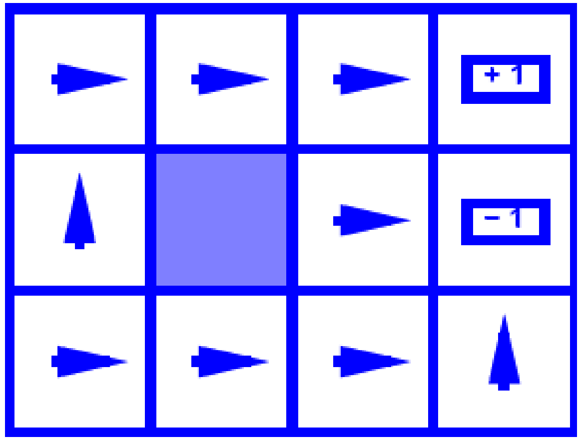
R(s) = -0.01



R(s) = -0.03

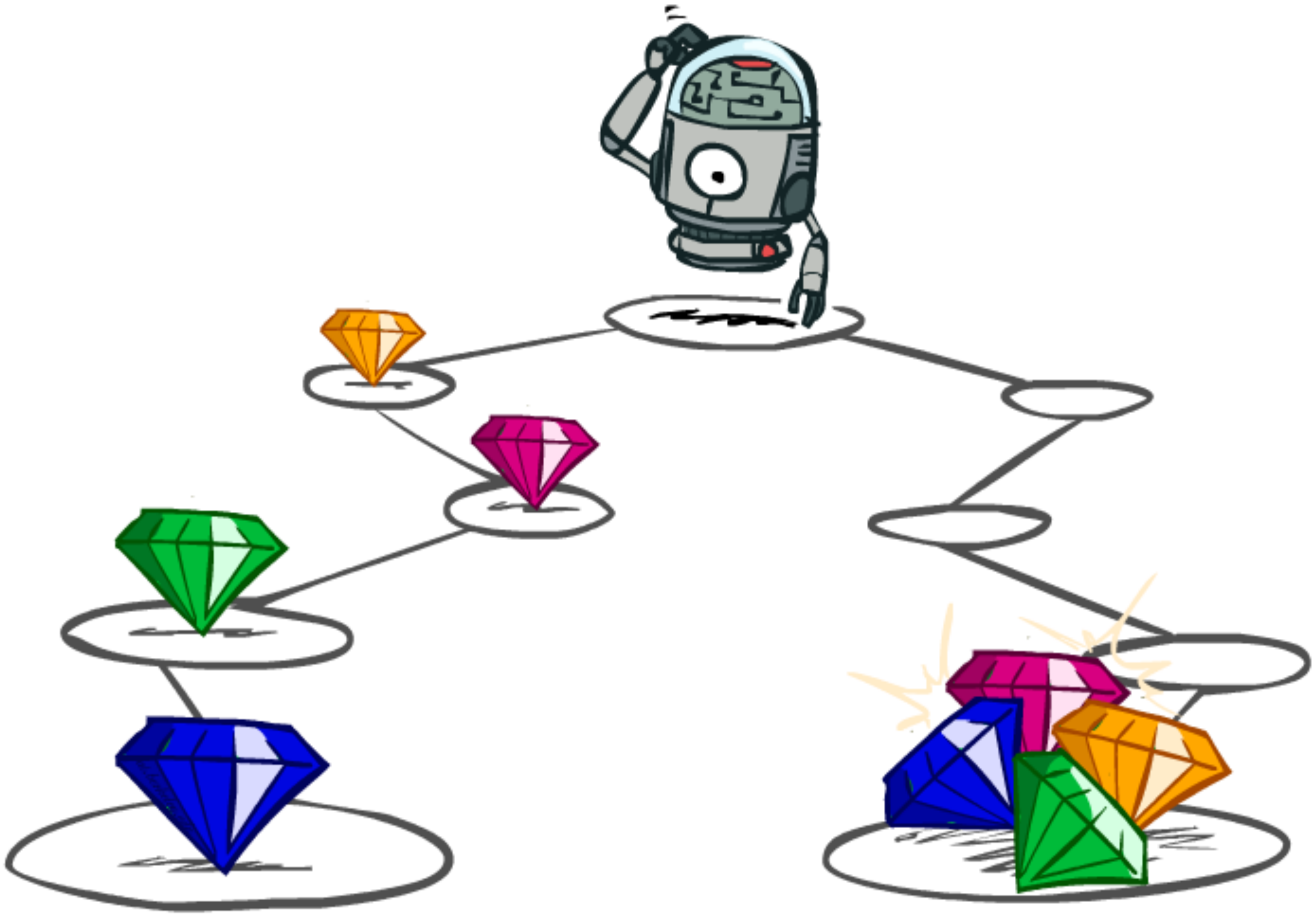


R(s) = -0.4



R(s) = -2.0

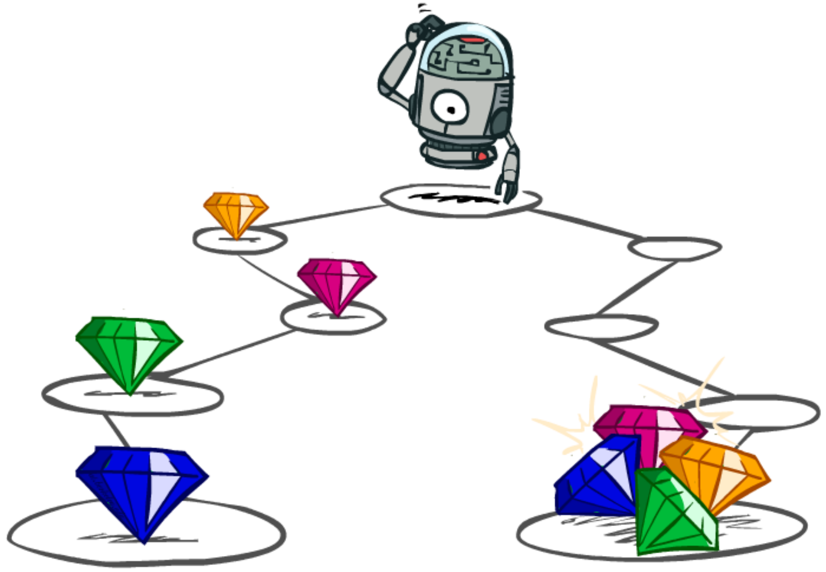
UTILITIES OF SEQUENCES



UTILITIES OF SEQUENCES

What preferences should an agent have over reward sequences?

More or less?	[1, 2, 2]	or	[2, 3, 4]
Now or later?	[0, 0, 1]	or	[1, 0, 0]



DISCOUNTING

It’s reasonable to maximize the sum of rewards

It’s also reasonable to prefer rewards now to rewards later

One solution: utility of rewards decay exponentially

1

Worth Now

γ

Worth Next Step

γ^2

Worth In Two Steps

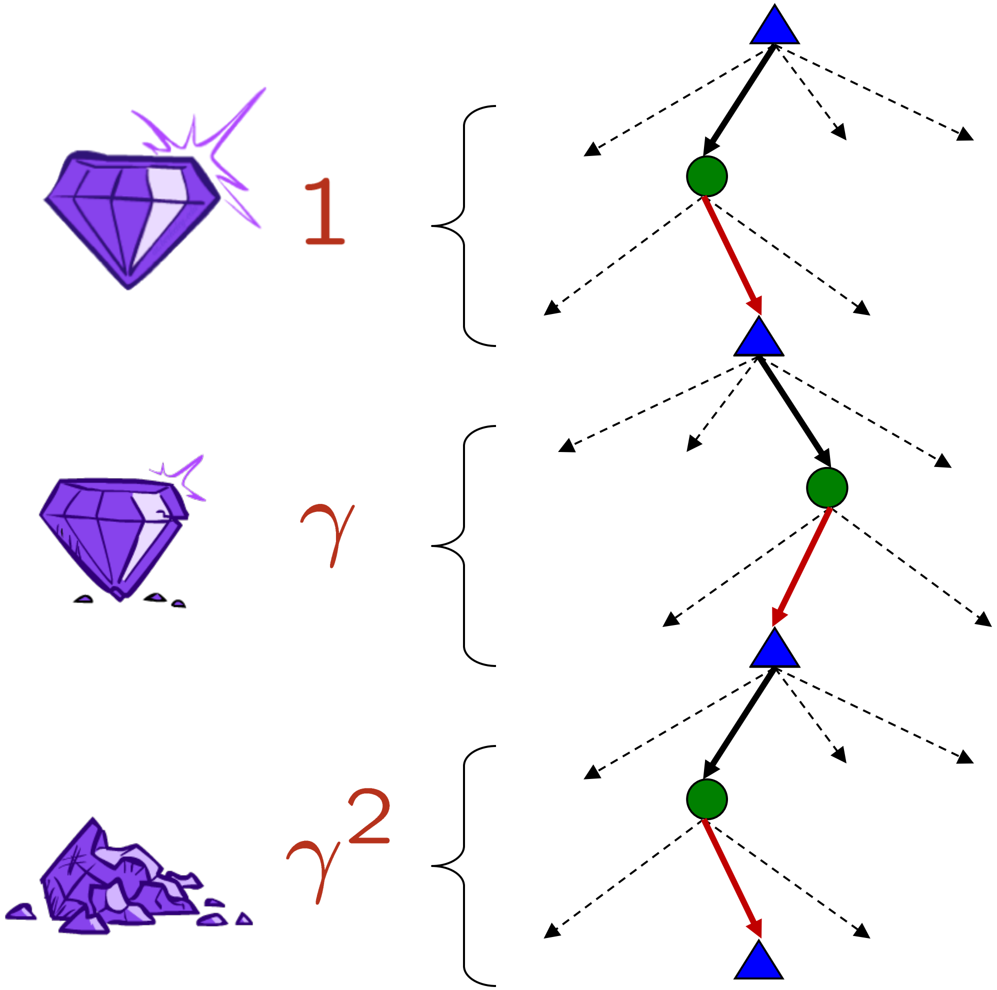
DISCOUNTING

How to discount?

- Each time we descend a level, we multiply in the discount once

Why discount?

- Sooner rewards probably do have higher utility than later rewards
- Also helps our algorithms converge



DISCUSSION POINT!

What is the value of U([2,4,8]) with $\gamma = 0.5$???????????????

(U($\cdot$) is the total utility of a reward sequence.)

- 3
- 6
- 7
- 14

Bonus: What is the value of U([8,4,2]) with $\boldsymbol{\gamma = 0.5}$?????????????

DISCUSSION POINT!

What is the value of U([2,4,8]) with $\gamma = 0.5$???????????????

(U(·) is the total utility of a reward sequence.)

- 3
- 6
- 7
- 14

$$\gamma^0 \times 2 + \gamma^1 \times 4 + \gamma^2 \times 8 = 2 + 0.5 \times 4 + 0.5 \times 0.5 \times 8 = 2 + 2 + 2 = 6$$

Bonus: What is the value of U([8,4,2]) with $\boldsymbol{\gamma = 0.5}$?????????????

$$\gamma^0 \times 8 + \gamma^1 \times 4 + \gamma^2 \times 2 = 8 + 0.5 \times 4 + 0.5 \times 0.5 \times 2 = 8 + 2 + 0.5 = 10.5$$

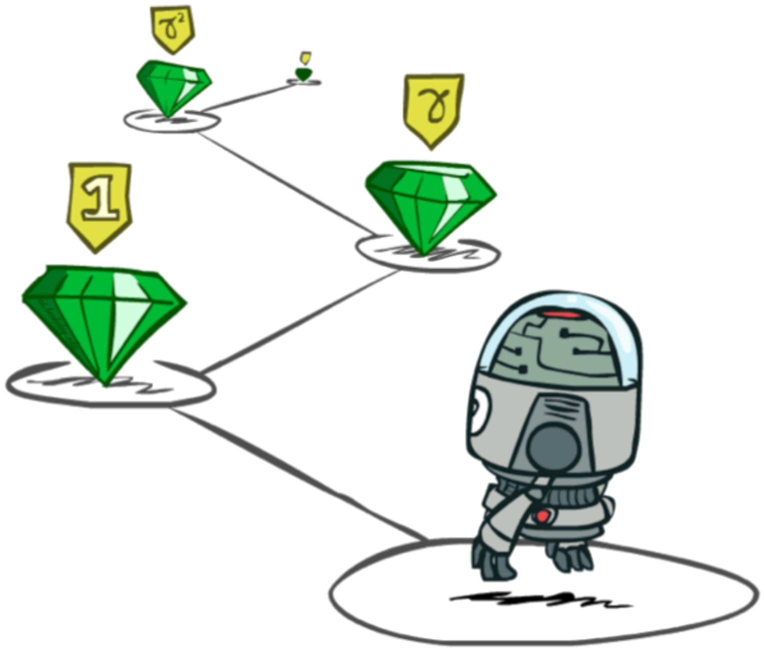
STATIONARY PREFERENCES

Theorem: if we assume stationary preferences:

$$[a_1, a_2, \ldots] \succ [b_1, b_2, \ldots]$$

$$\Updownarrow$$

$$[r, a_1, a_2, \ldots] \succ [r, b_1, b_2, \ldots]$$



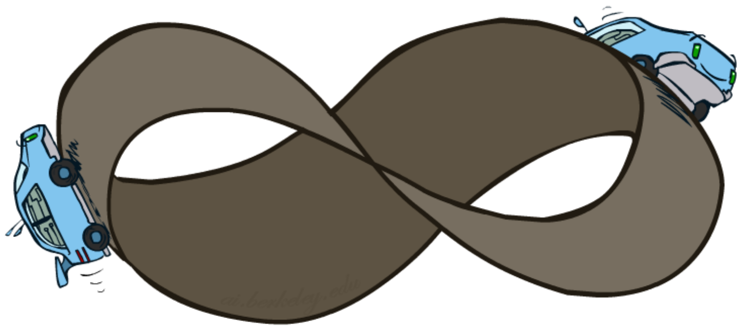
Then: there are only two ways to define utilities

- Additive utility: $U([r_0, r_1, r_2, \ldots]) = r_0 + r_1 + r_2 + \cdots$
- Discounted utility: $U([r_0, r_1, r_2, \ldots]) = r_0 + \gamma r_1 + \gamma^2 r_2 \cdots$

Assume $|r_t| \leq R_{max}$

INFINITE UTILITIES?!

What if the sequence is infinite? Do we get infinite utility?



- With discounting γ where $0 < \gamma < 1$

$$U([r_0, \ldots r_\infty]) = \sum_{t=0}^{\infty} \gamma^t r_t \leq R_{\max}/(1-\gamma)$$

- Smaller γ means smaller "horizon" – shorter term focus

OPTIMAL POLICY WITH DISCOUNTING

Given

10				1
a	b	c	d	e

- Actions: East, West, and Exit (only available in exit states a, e)
- Transitions: deterministic

For $\gamma = 1$, what is the optimal policy?

10				1

OPTIMAL POLICY WITH DISCOUNTING

Given

10				1
a	b	c	d	e

- Actions: East, West, and Exit (only available in exit states a, e)
- Transitions: deterministic

For $\gamma = 1$, what is the optimal policy?

10	←	←	←	1

OPTIMAL POLICY WITH DISCOUNTING

Given

10				1
a	b	c	d	e

- Actions: East, West, and Exit (only available in exit states a, e)
- Transitions: deterministic

For $\gamma = 1$, what is the optimal policy?

10	←	←	←	1

For $\gamma = 0.1$, what is the optimal policy?

10				1

OPTIMAL POLICY WITH DISCOUNTING

Given

10				1
a	b	c	d	e

- Actions: East, West, and Exit (only available in exit states a, e)
- Transitions: deterministic

For $\gamma = 1$, what is the optimal policy?

10	←	←	←	1

For $\gamma = 0.1$, what is the optimal policy?

10	←	←	→	1

OPTIMAL POLICY WITH DISCOUNTING

Given

a	b	c	d	e
10				1

- Actions: East, West, and Exit (only available in exit states a, e)
- Transitions: deterministic

For $\gamma = 1$, what is the optimal policy?

10	←	←	←	1

For $\gamma = 0.1$, what is the optimal policy?

10	←	←	→	1

For which γ are West and East equally good when in state d?

OPTIMAL POLICY WITH DISCOUNTING

Given

a	b	c	d	e
10				1

- Actions: East, West, and Exit (only available in exit states a, e)
- Transitions: deterministic

For $\gamma = 1$, what is the optimal policy?

10	←	←	←	1

For $\gamma = 0.1$, what is the optimal policy?

10	←	←	→	1

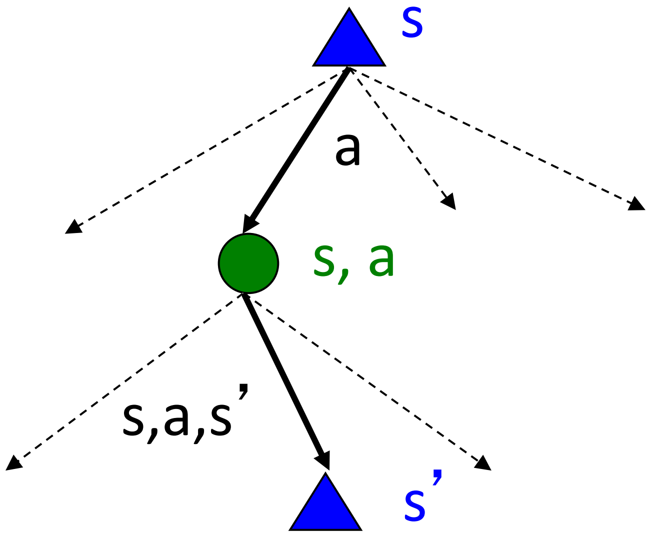
For which γ are West and East equally good when in state d?

$$\gamma^3 \times 10 = \gamma^1 \times 1$$

MDP QUANTITIES (SO FAR!)

Markov decision processes:

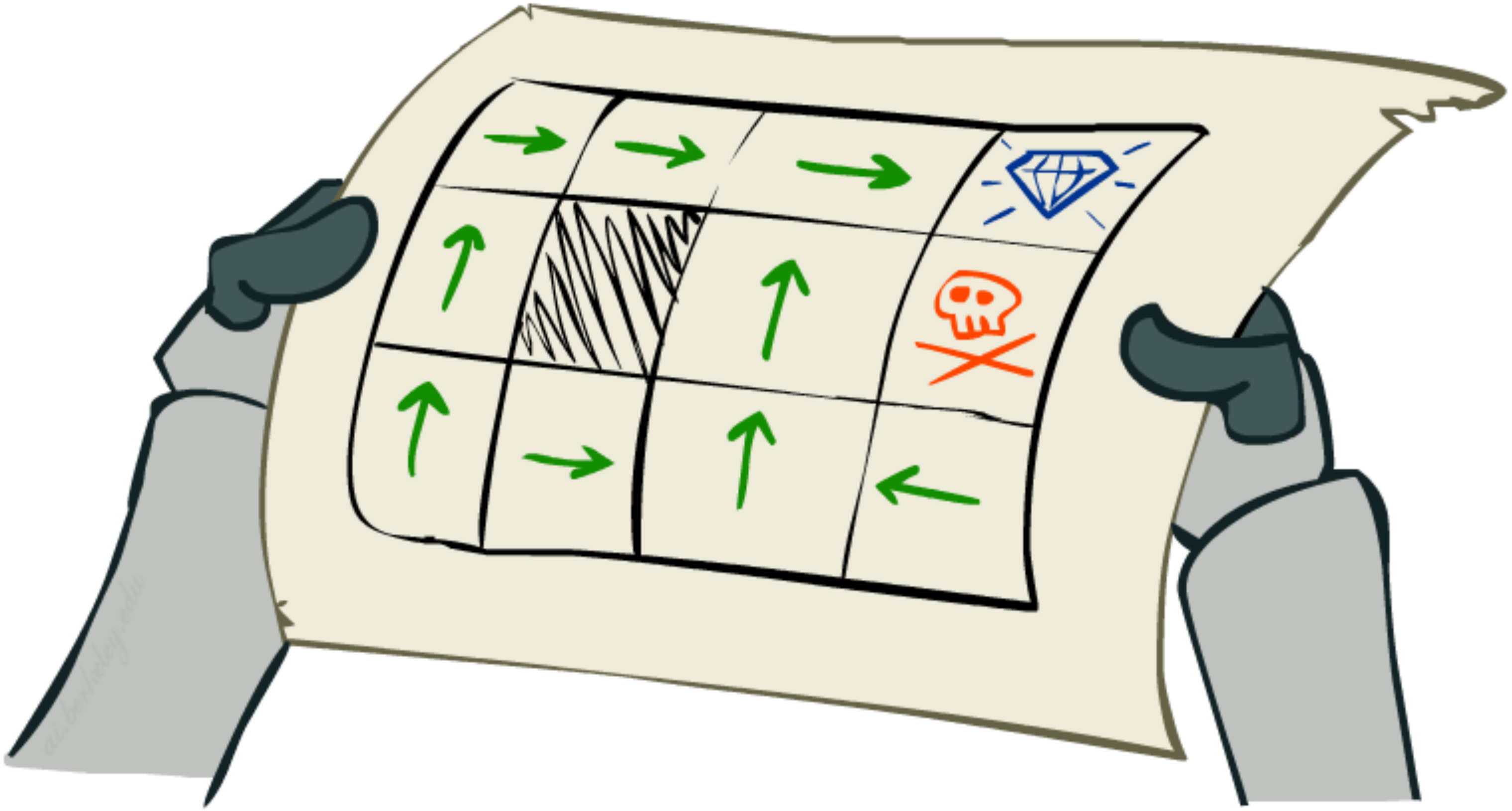
- States S
- Actions A
- Transitions P(s'|s,a) (or T(s,a,s'))
- Rewards R(s,a,s') (and discount γ)
- Start state s0



MDP quantities so far:

- Policy = map of states to actions
- Utility = sum of (discounted) rewards

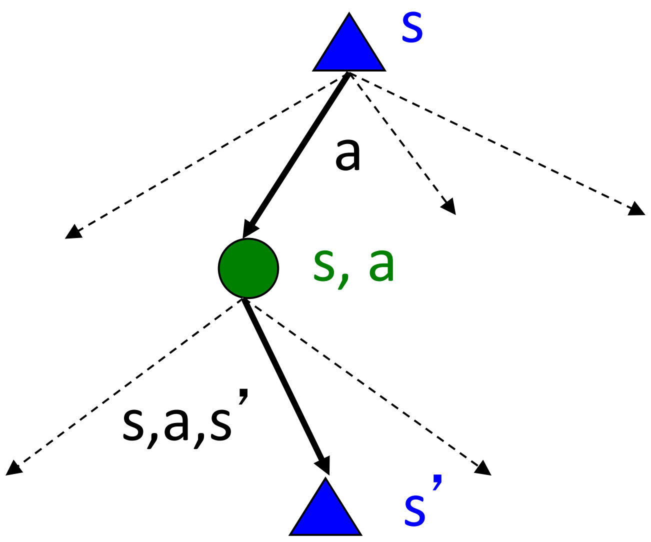
SOLVING MDPS



MDP QUANTITIES

Markov decision processes:

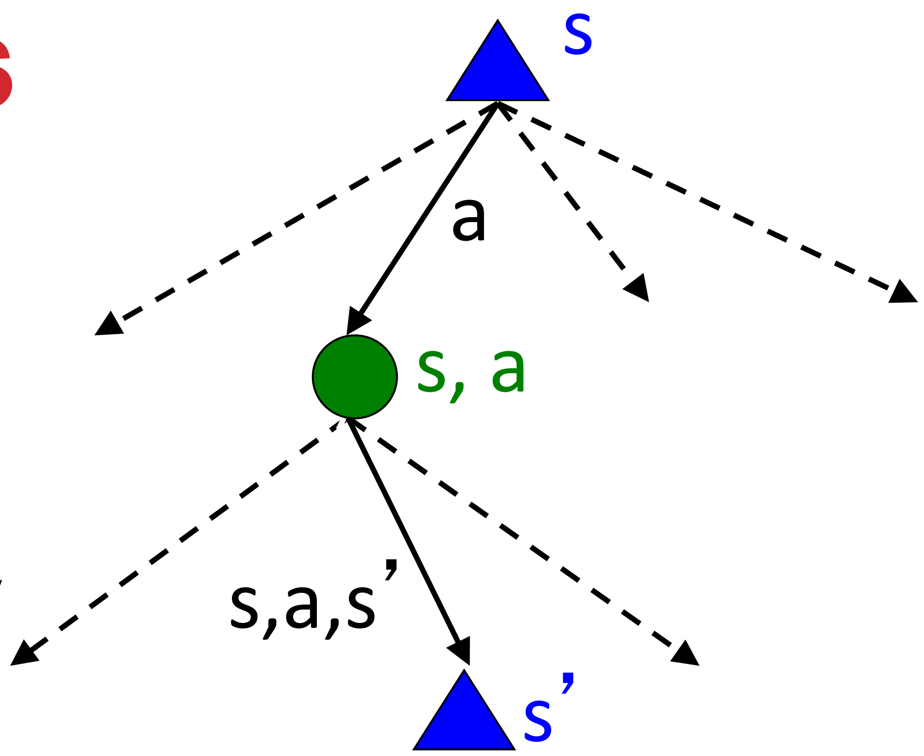
- States S
- Actions A
- Transitions P(s'|s,a) (or T(s,a,s'))
- Rewards R(s,a,s') (and discount γ)
- Start state s_0



MDP quantities:

- Policy = map of states to actions
- Utility = sum of (discounted) rewards
- (State) Value = expected utility starting from a state (max node)
- Q-Value = expected utility starting from a state-action pair, i.e., q-state (chance node)

MDP OPTIMAL QUANTITIES



The optimal policy:

- π*(s) = optimal action from state s

The (true) value (or utility) of a state s:

- V*(s) = expected utility starting in s and acting optimally

The (true) value (or utility) of a q-state (s,a):

- Q*(s,a) = expected utility starting out having taken action a from state s and (thereafter) acting optimally

Solve MDP: Find π^*, V^* and/or Q^*