15-281 AI: Representation and Problem Solving

Markov Decision Processes

Instructors: **Aditi Raghunathan** and Vince Conitzer

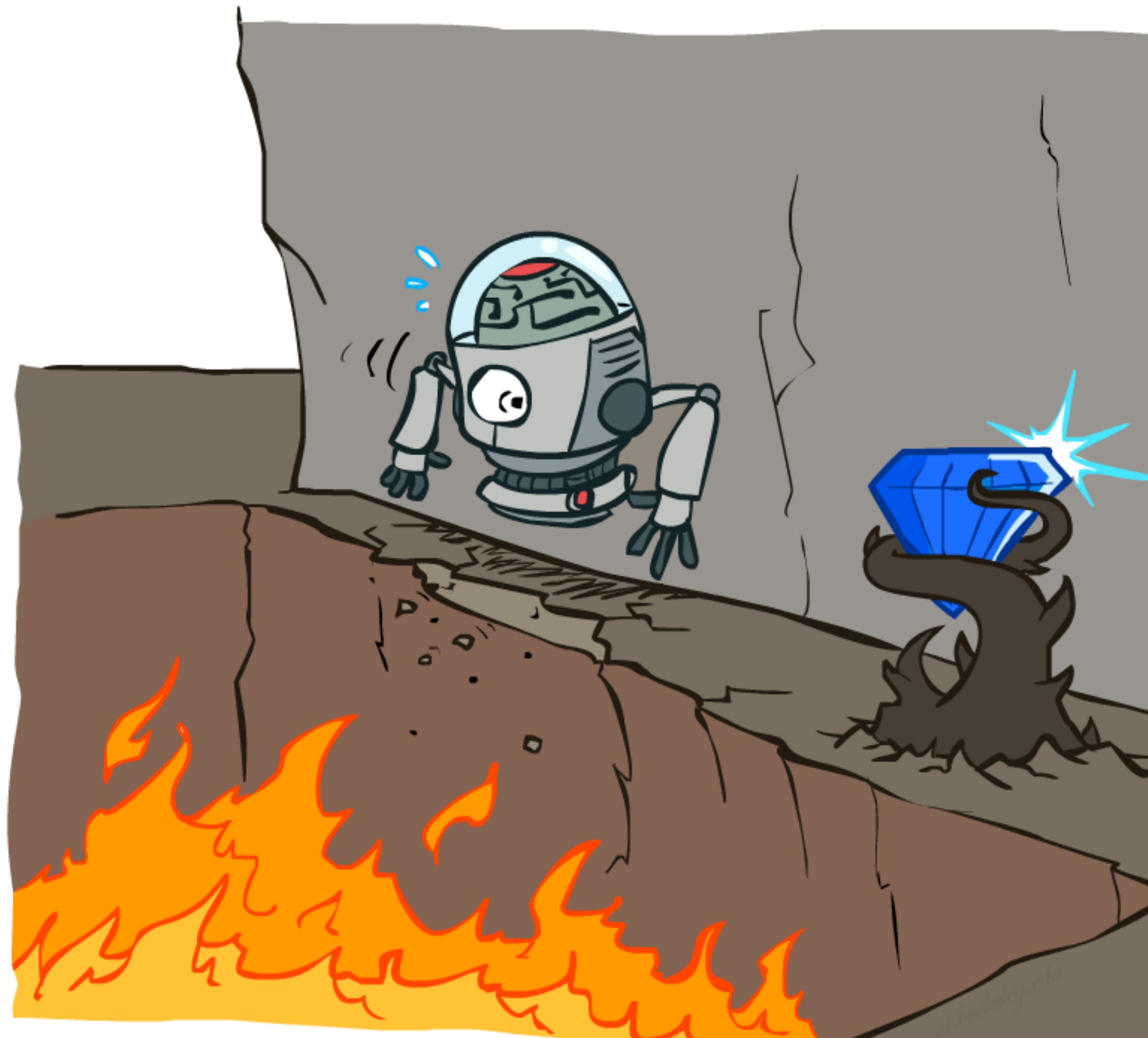
Carnegie Mellon University

[These slides adapted from CMU AI and http://ai.berkeley.edu]

Logistics

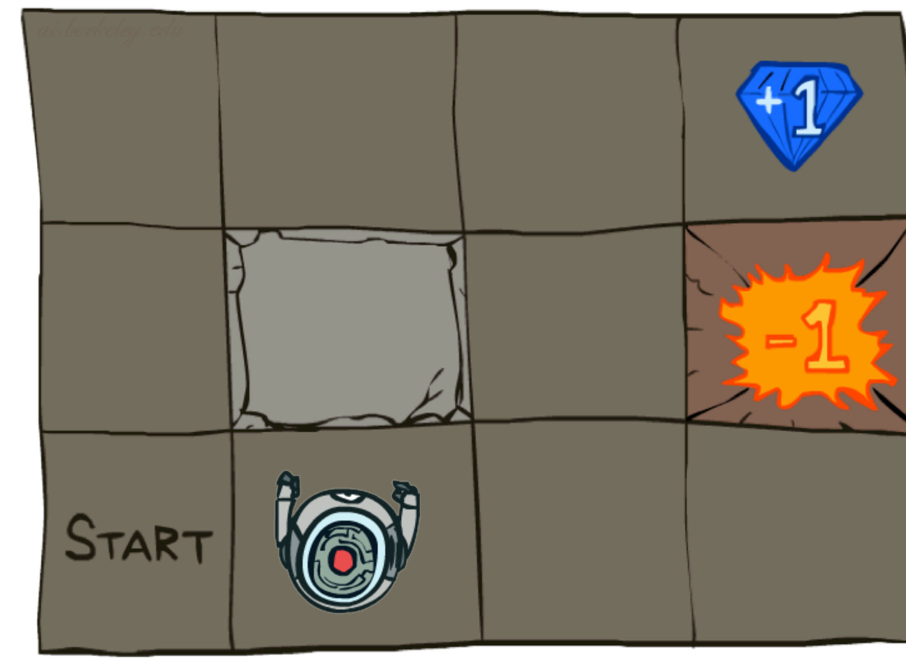
- HW 5 due today (Oct 10th)
- P3 checkpoint due Oct 13th
- Mid-semester feedback - please fill out!
- Fall break next week!

Non-Deterministic Search



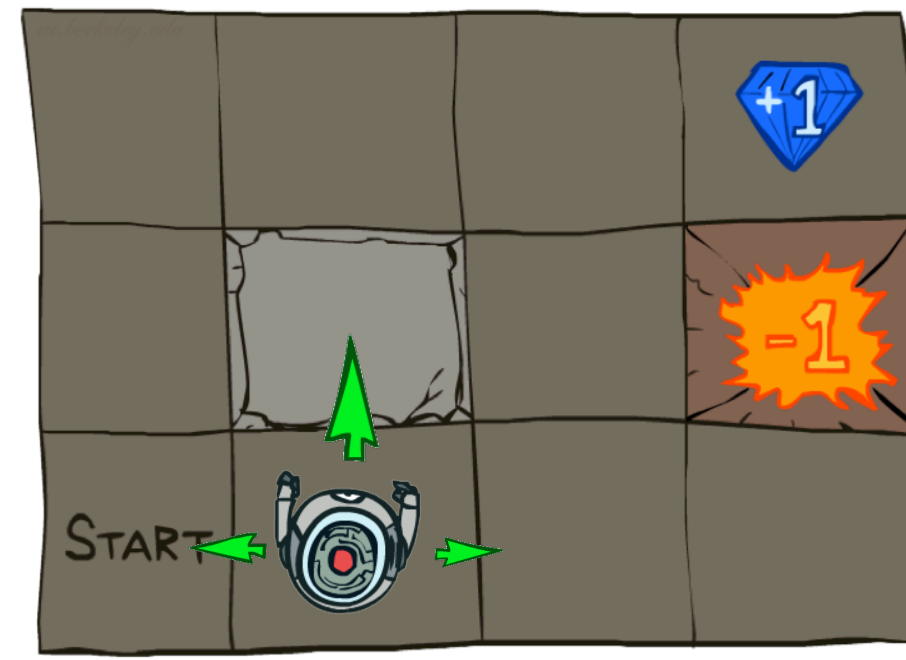
Example: Grid World

- A maze-like problem
  - The agent lives in a grid
  - Walls block the agent's path



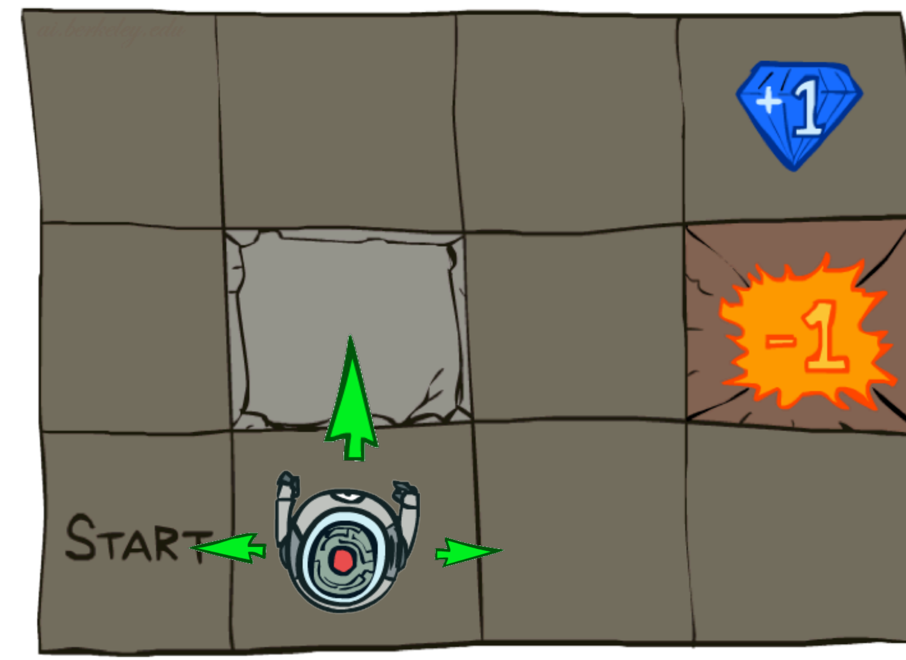
Example: Grid World

- 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
  - 80% of the time, the action North takes the agent North (if there is no wall there)
  - 10% of the time, North takes the agent West; 10% East
  - If there is a wall in the direction the agent would have been taken, the agent stays put



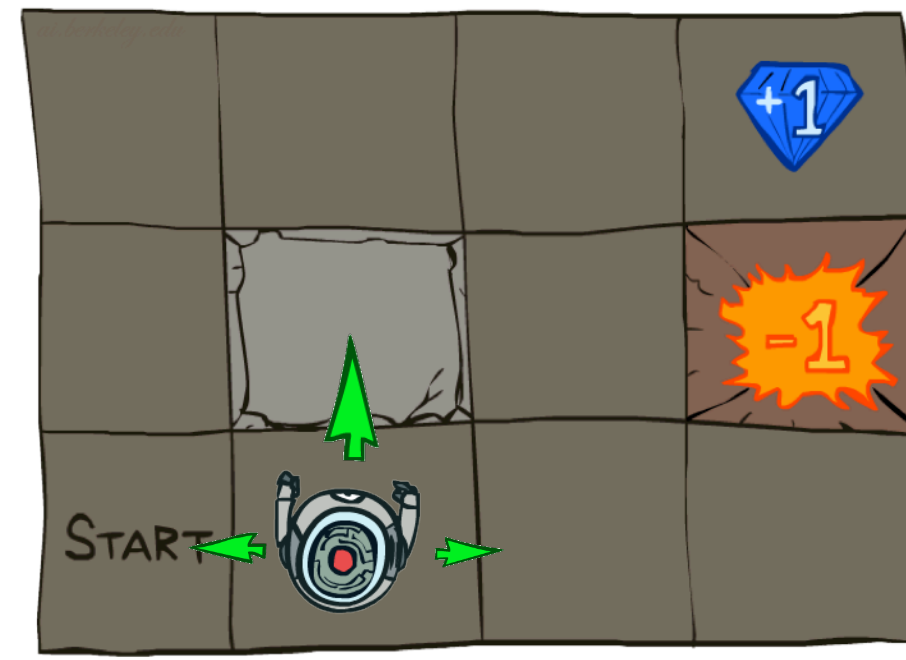
Example: Grid World

- 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
  - 80% of the time, the action North takes the agent North (if there is no wall there)
  - 10% of the time, North takes the agent West; 10% East
  - If there is a wall in the direction the agent would have been taken, the agent stays put
- The agent receives rewards
  - Small "living" reward each step (can be negative)
  - Big rewards come at the **end** (good or bad)



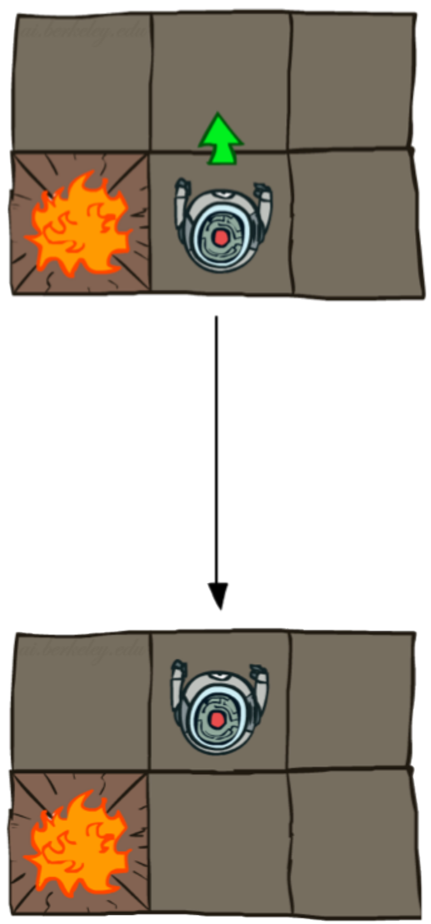
Example: Grid World

- 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
  - 80% of the time, the action North takes the agent North (if there is no wall there)
  - 10% of the time, North takes the agent West; 10% East
  - If there is a wall in the direction the agent would have been taken, the agent stays put
- The agent receives rewards
  - Small "living" reward each step (can be negative)
  - Big rewards come at the **end** (good or bad)
- <u>Goal</u>: maximize sum of rewards



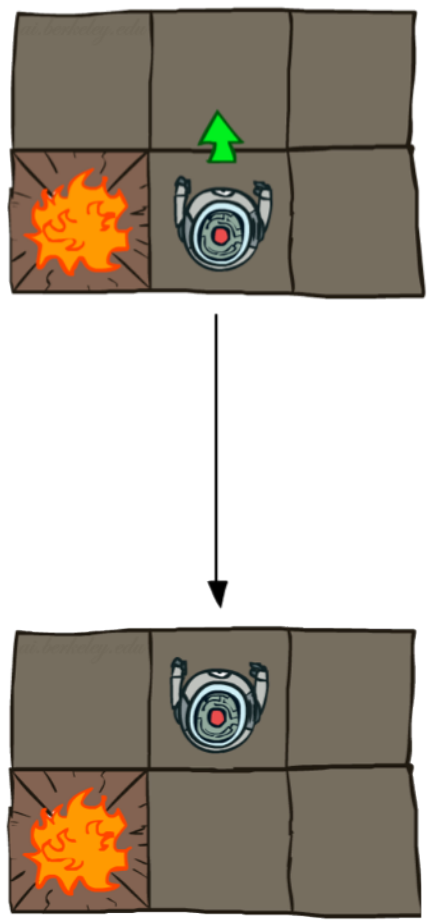
Grid World Actions

Deterministic Grid World

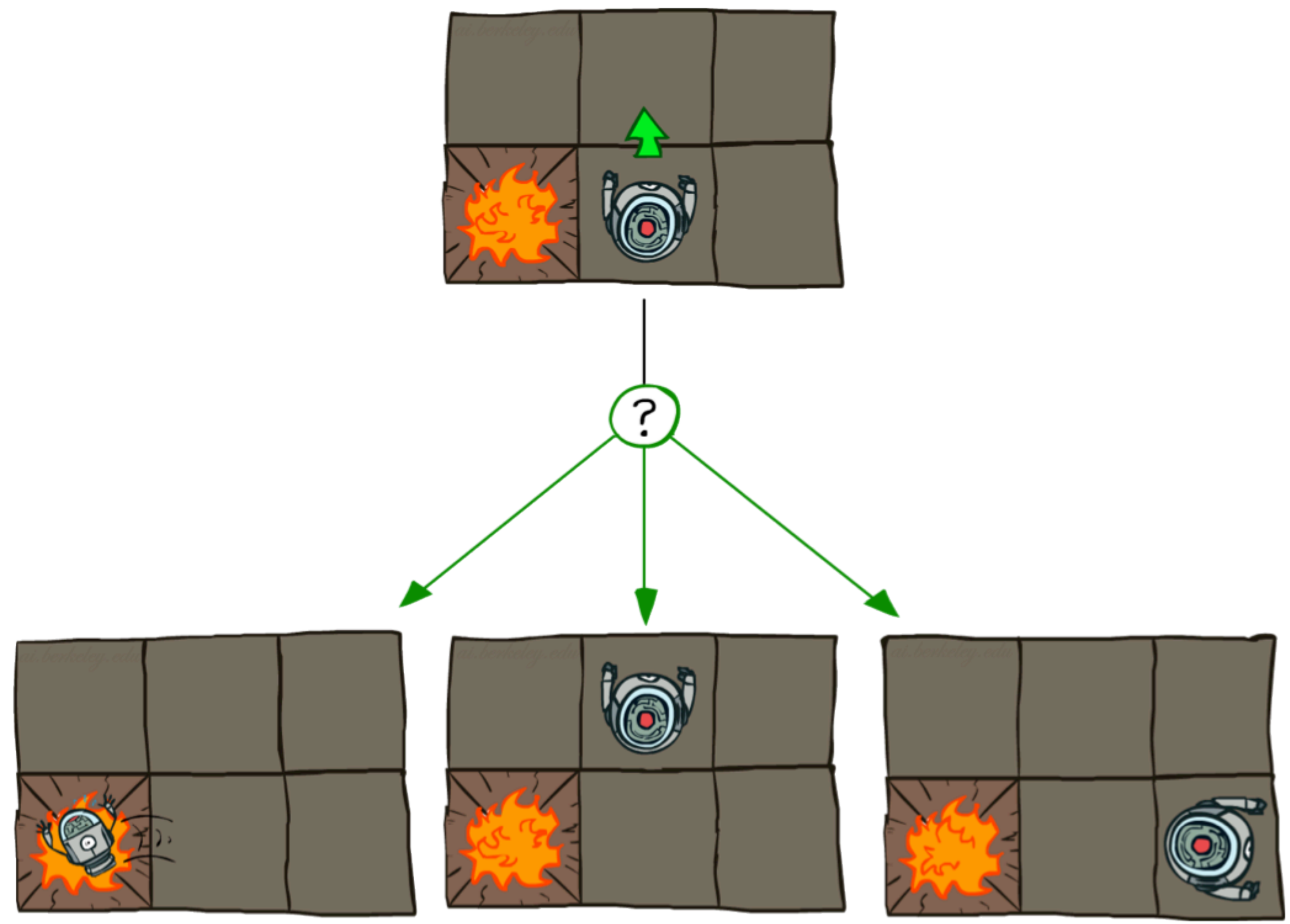


Grid World Actions

Deterministic Grid World

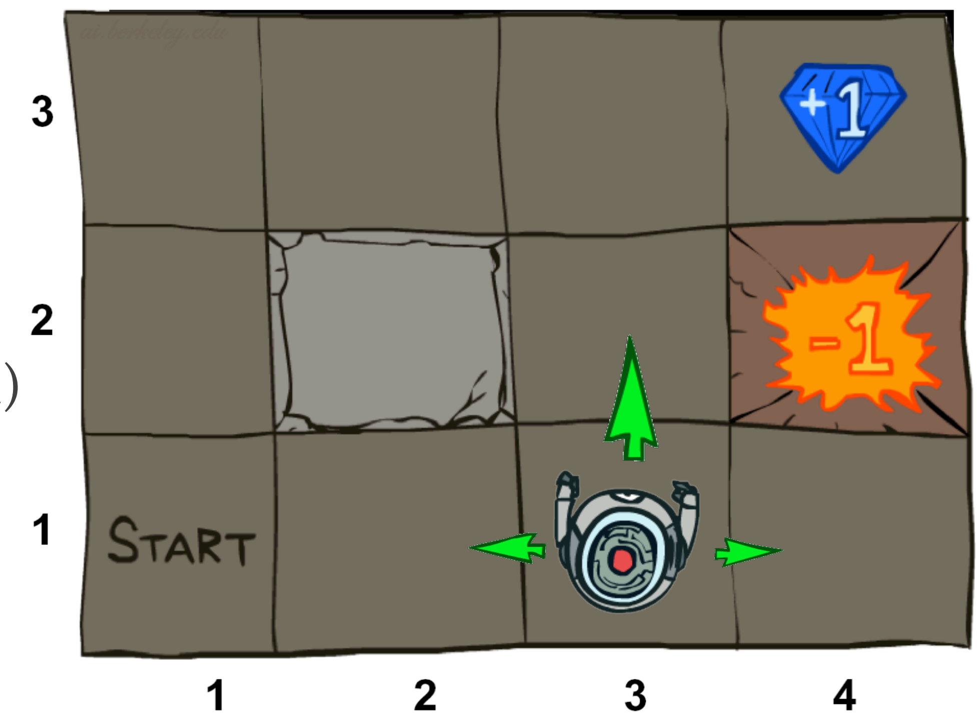


Stochastic Grid World



Markov Decision Processes

- An MDP is defined by:
  - A set of states $s \in S$
  - A set of actions $a \in A$
  - A transition function $T(s, a, s')$
    - Probability that a from s leads to s', i.e., $P(s' \mid s, a)$
    - Also called the model or the dynamics
  - A reward function $R(s, a, s')$
    - Sometimes just $R(s)$ or $R(s')$
  - A start state
  - Maybe a terminal state



What is Markov about MDPs?

- "Markov" generally means that given the present state, the future and the past are independent

- For Markov decision processes, "Markov" means action outcomes depend only on the current state

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

$$=$$

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

- This is just like search, where the successor function could only depend on the current state (not the history)

Andrey Markov
(1856-1922)

Policies

- In deterministic single-agent search problems, we wanted an optimal plan, or sequence of actions, from start to a goal

Policies

- In deterministic single-agent search problems, we wanted an optimal plan, or sequence of actions, from start to a goal

- For MDPs, we want an optimal

Policies

- In deterministic single-agent search problems, we wanted an optimal plan, or sequence of actions, from start to a goal

- For MDPs, we want an optimal

  policy $\pi^*: S \rightarrow A$

Policies

- In deterministic single-agent search problems, we wanted an optimal plan, or sequence of actions, from start to a goal

- For MDPs, we want an optimal

  policy $\pi^*: S \rightarrow A$

  - A policy π gives an action for each state

Policies

- In deterministic single-agent search problems, we wanted an optimal plan, or sequence of actions, from start to a goal

- For MDPs, we want an optimal

  policy $\pi^*: S \rightarrow A$

  - A policy π gives an action for each state
  - An optimal policy is one that maximizes expected utility if followed

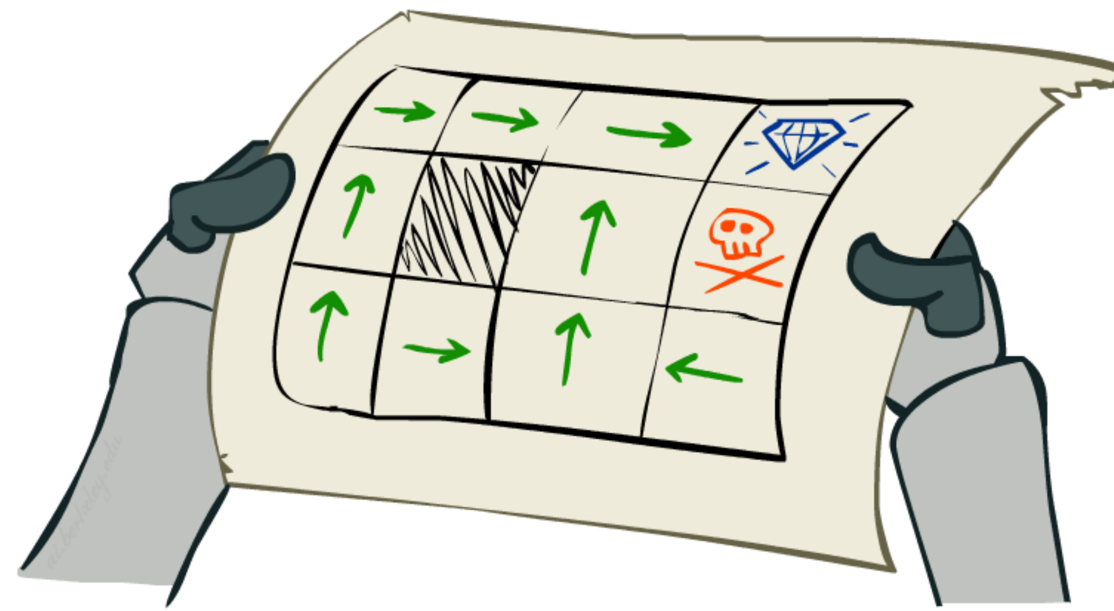
Policies

- In deterministic single-agent search problems, we wanted an optimal plan, or sequence of actions, from start to a goal

- For MDPs, we want an optimal

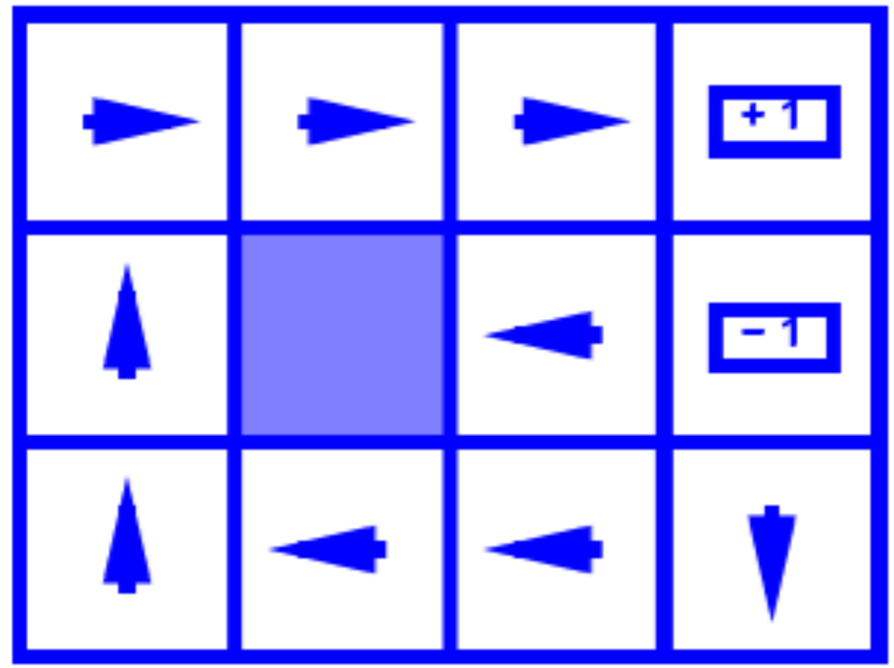
  policy $\pi^*: S \rightarrow A$

  - A policy π gives an action for each state
  - An optimal policy is one that maximizes expected utility if followed



Optimal policy when R(s, a, s') = -0.4 for all non-terminals s

Optimal Policies

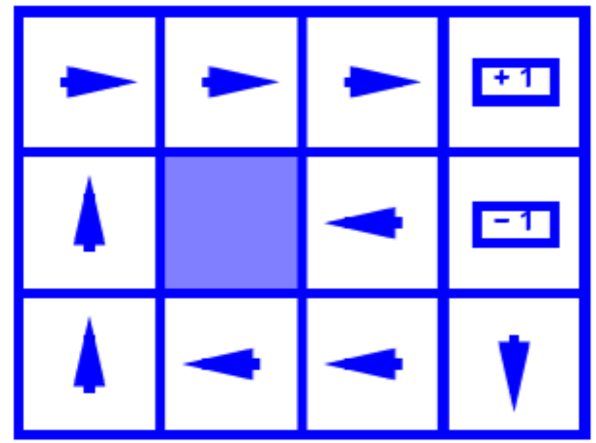


$R(s) = -0.01$

Poll: Optimal Policy for Different Rewards

What is the optimal policy for living reward $R(s) = -2.0$

The others correspond to $R(s) = -0.01$, $R(s) = -0.03$, $R(s) = -0.4$

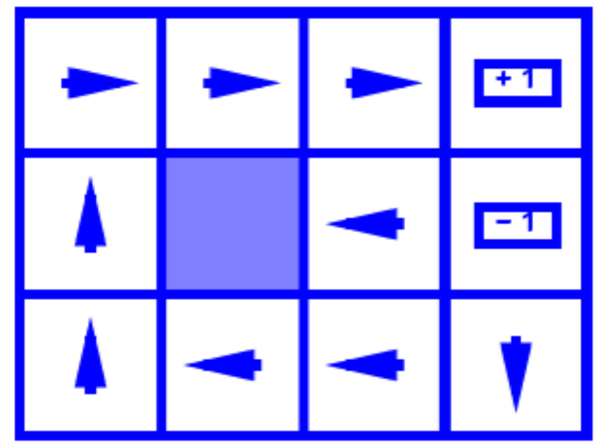


(A)

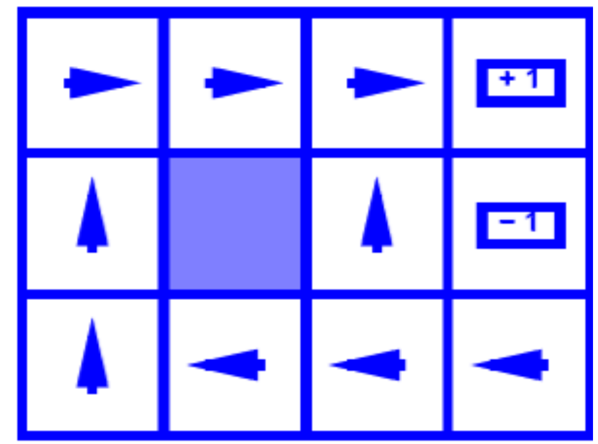
Poll: Optimal Policy for Different Rewards

What is the optimal policy for living reward $R(s) = -2.0$

The others correspond to $R(s) = -0.01$, $R(s) = -0.03$, $R(s) = -0.4$



(A)

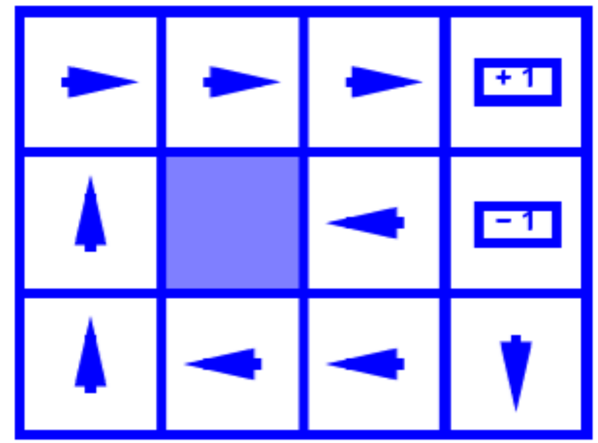


(B)

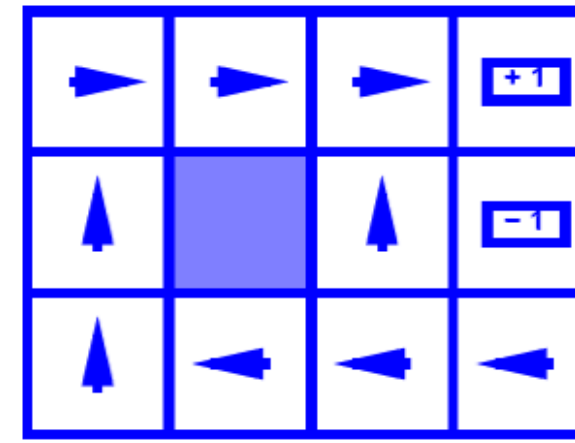
Poll: Optimal Policy for Different Rewards

What is the optimal policy for living reward $R(s) = -2.0$

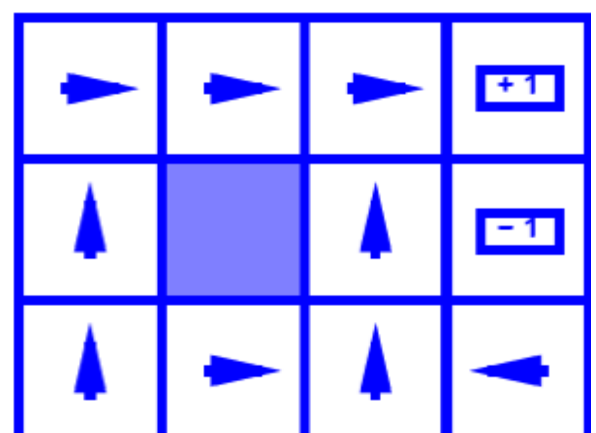
The others correspond to $R(s) = -0.01$, $R(s) = -0.03$, $R(s) = -0.4$



(A)



(B)

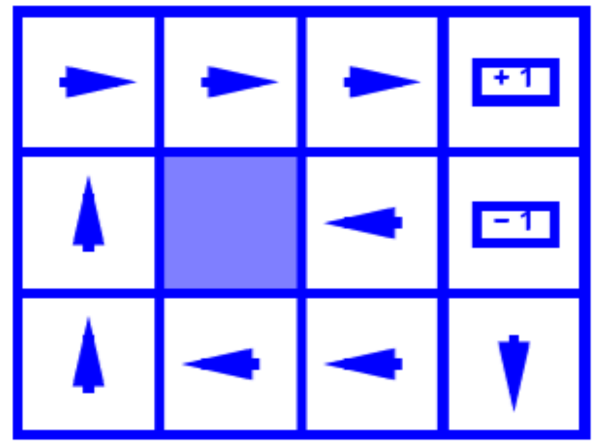


(C)

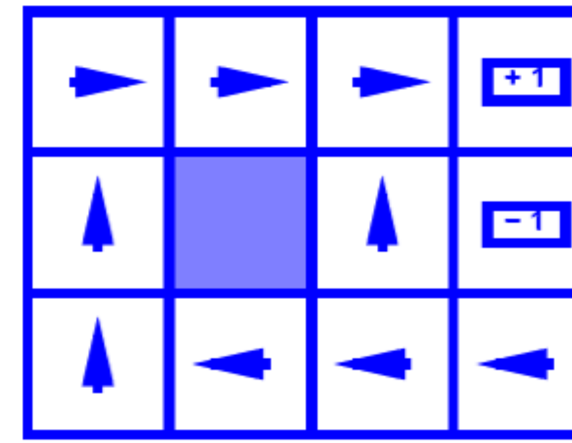
Poll: Optimal Policy for Different Rewards

What is the optimal policy for living reward $R(s) = -2.0$

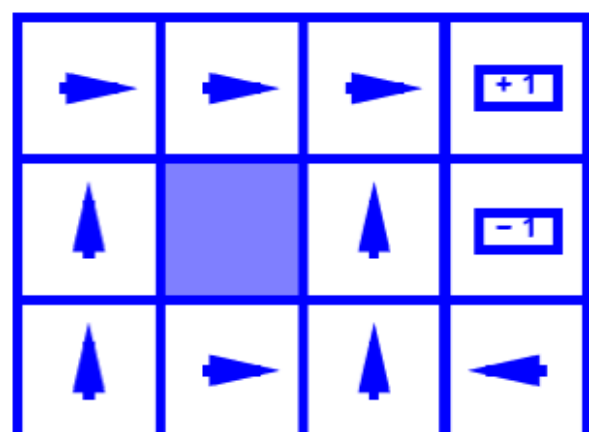
The others correspond to $R(s) = -0.01$, $R(s) = -0.03$, $R(s) = -0.4$



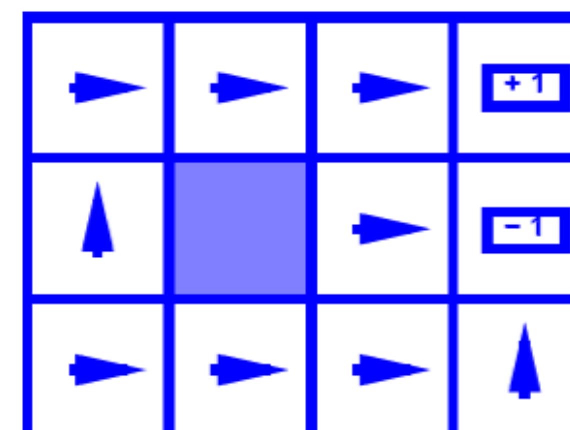
(A)



(B)

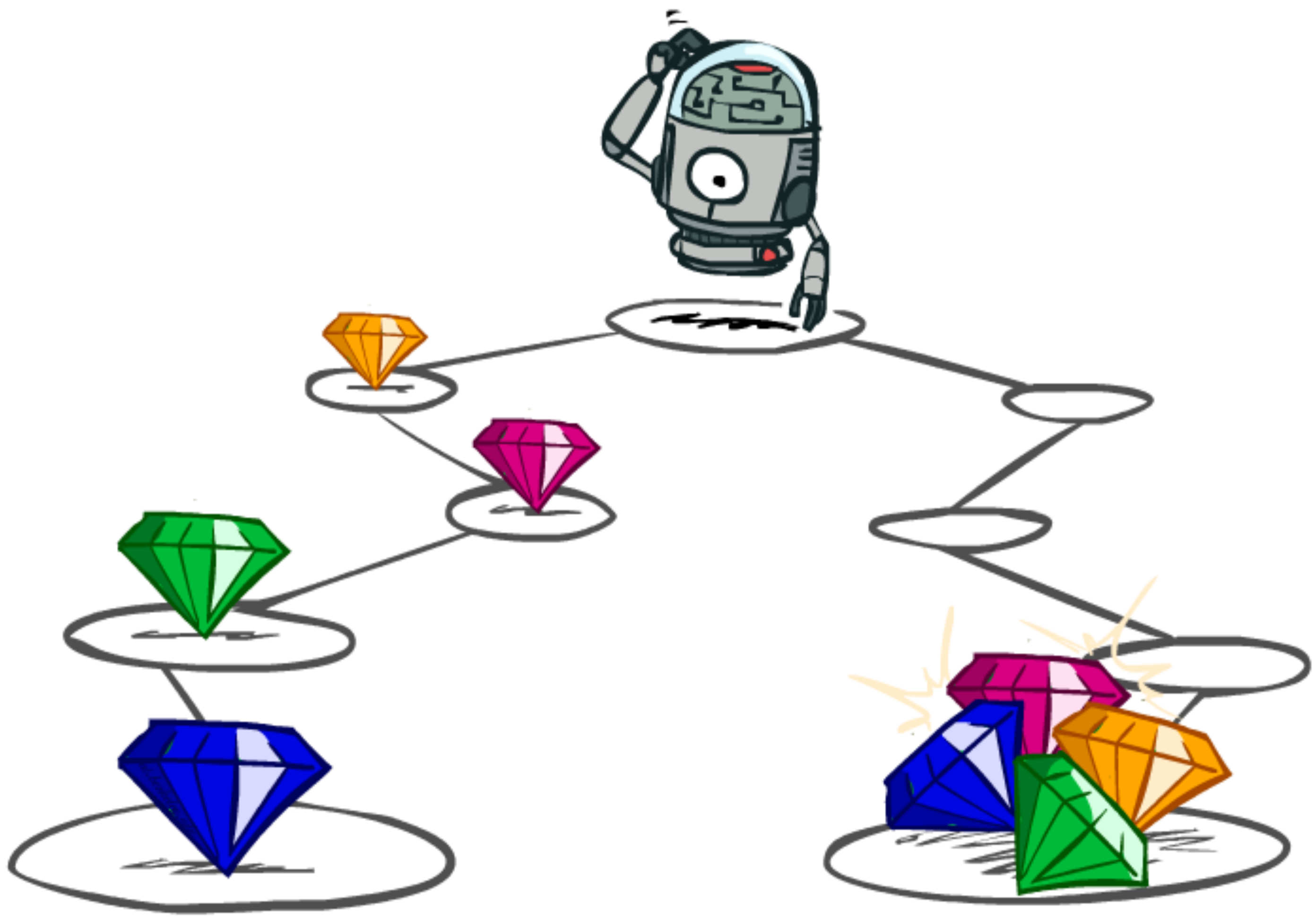


(C)



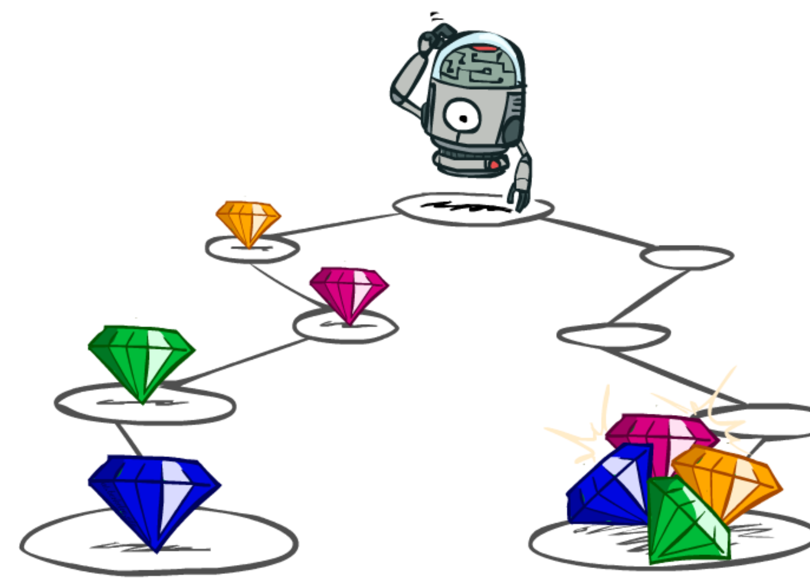
(D)

Utilities of Sequences



Utilities of Sequences

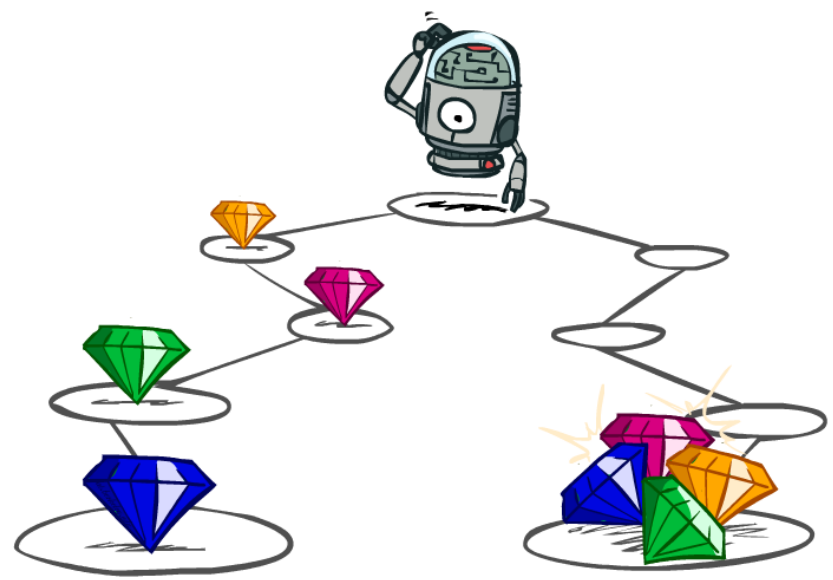
- What preferences should an agent have over reward sequences?



Utilities of Sequences

- What preferences should an agent have over reward sequences?

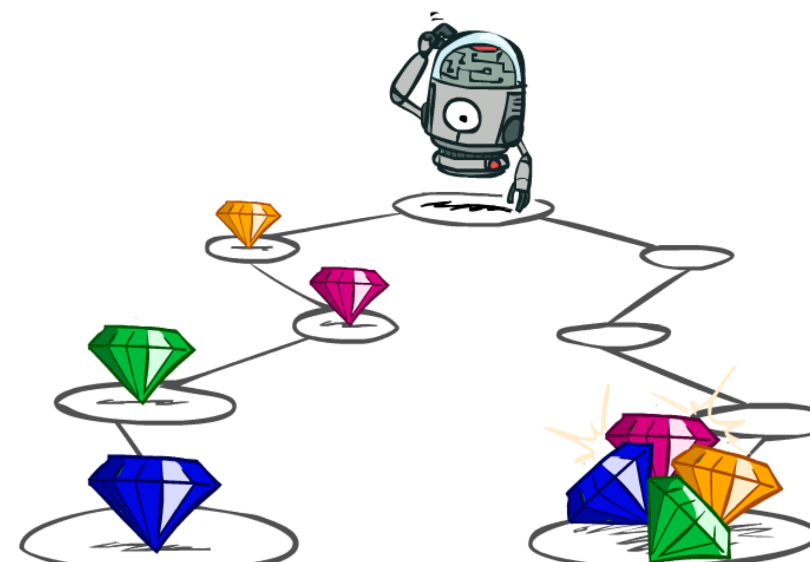
- More or less?



Utilities of Sequences

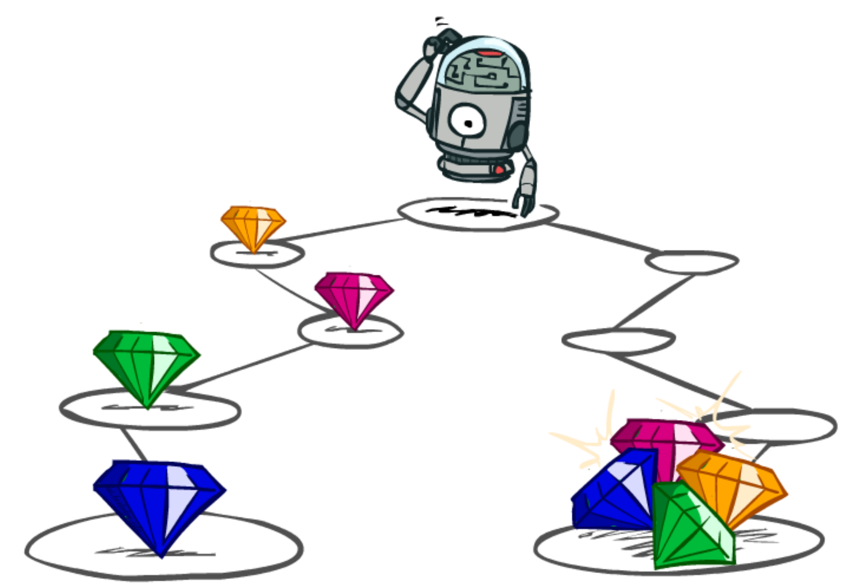
- What preferences should an agent have over reward sequences?

- More or less? [1, 2, 2]   or   [2, 3, 4]



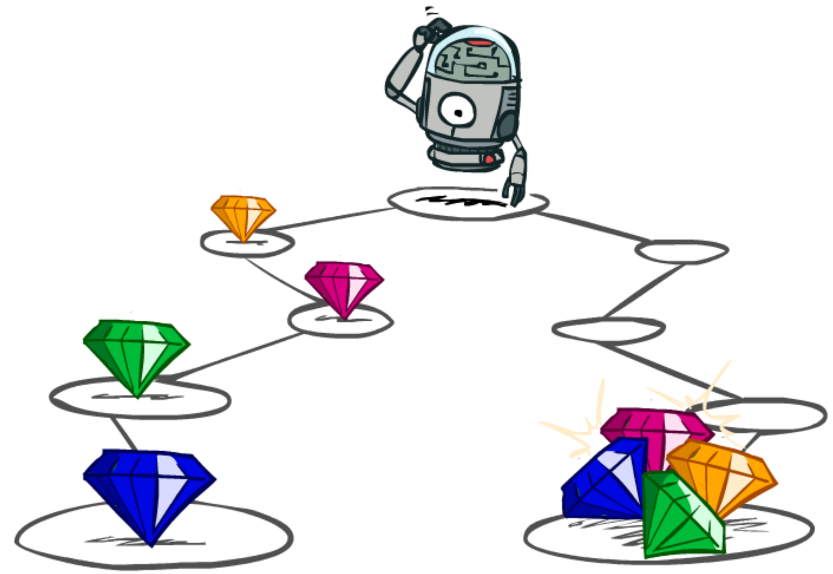
Utilities of Sequences

- What preferences should an agent have over reward sequences?
- More or less? [1, 2, 2] or [2, 3, 4]
- Now or later?



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: values of rewards decay exponentially

Example: discount of 0.5

$U([1,2,3]) = 1*1 + 0.5*2 + 0.25*3$

$U([1,2,3]) < U([3,2,1])$

Discounting

- It's reasonable to maximize the sum of rewards
- It's also reasonable to prefer rewards now to rewards later
- One solution: values of rewards decay exponentially

1

Worth Now

Example: discount of 0.5

U([1,2,3]) = 1*1 + 0.5*2 + 0.25*3

U([1,2,3]) < U([3,2,1])

Discounting

- It's reasonable to maximize the sum of rewards
- It's also reasonable to prefer rewards now to rewards later
- One solution: values of rewards decay exponentially

1

Worth Now

γ

Worth Next Step

Example: discount of 0.5

U([1,2,3]) = 1*1 + 0.5*2 + 0.25*3

U([1,2,3]) < U([3,2,1])

Discounting

- It's reasonable to maximize the sum of rewards
- It's also reasonable to prefer rewards now to rewards later
- One solution: values of rewards decay exponentially

1

Worth Now

γ

Worth Next Step

γ^2

Worth In Two Steps

Example: discount of 0.5

U([1,2,3]) = 1*1 + 0.5*2 + 0.25*3

U([1,2,3]) < U([3,2,1])

Poll: Discounting

- Given:

10				1
a	b	c	d	e

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

(i)

10				1

(ii)

10				1

1. For $\gamma = 1$, optimal policy is (i)
2. For $\gamma = 1$, optimal policy is (ii)
3. For $\gamma = 0.1$, optimal policy is (i)
4. For $\gamma = 0.1$, optimal policy is (ii)

Poll: Discounting

- Given:

10				1
a	b	c	d	e

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

(i)

10	<-	<-	<-	1

(ii)

10				1

1. For $\gamma = 1$, optimal policy is (i)
2. For $\gamma = 1$, optimal policy is (ii)
3. For $\gamma = 0.1$, optimal policy is (i)
4. For $\gamma = 0.1$, optimal policy is (ii)

Poll: Discounting

- Given:

10				1
a	b	c	d	e

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

(i)

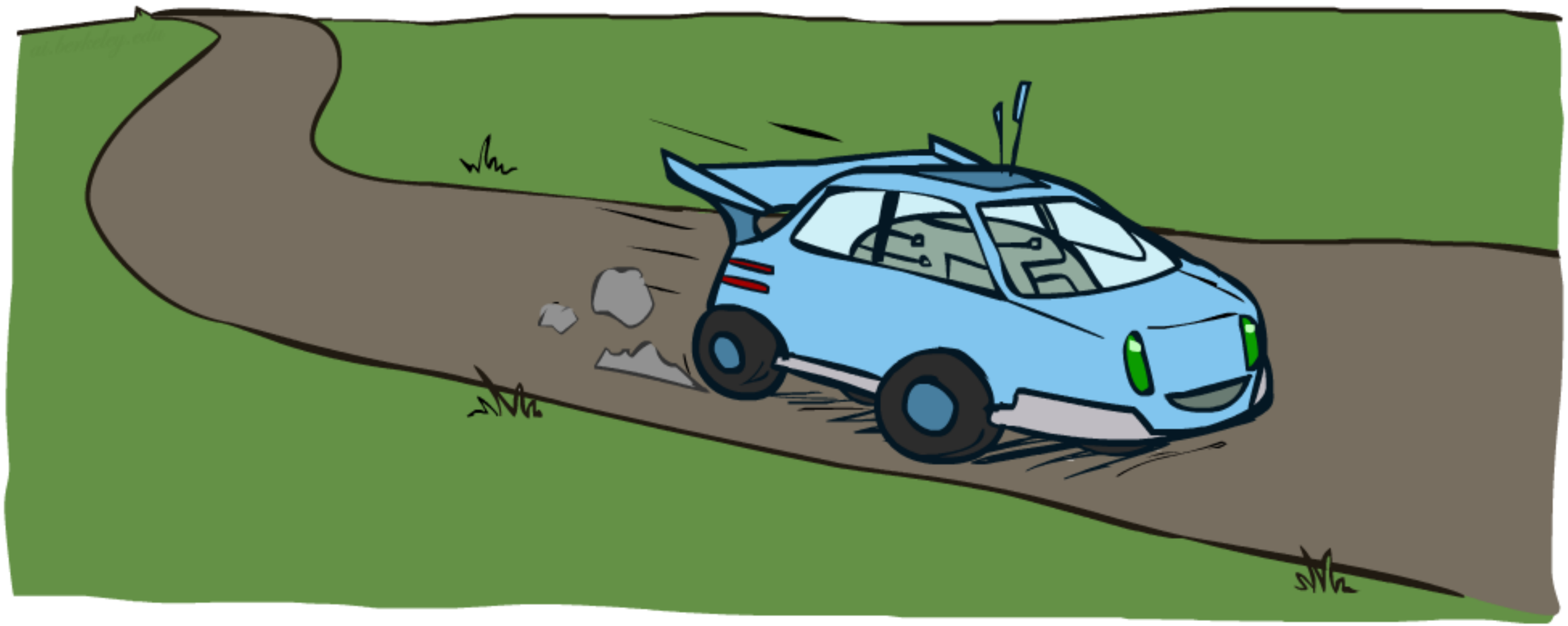
10	<-	<-	<-	1

(ii)

10	<-	<-	->	1

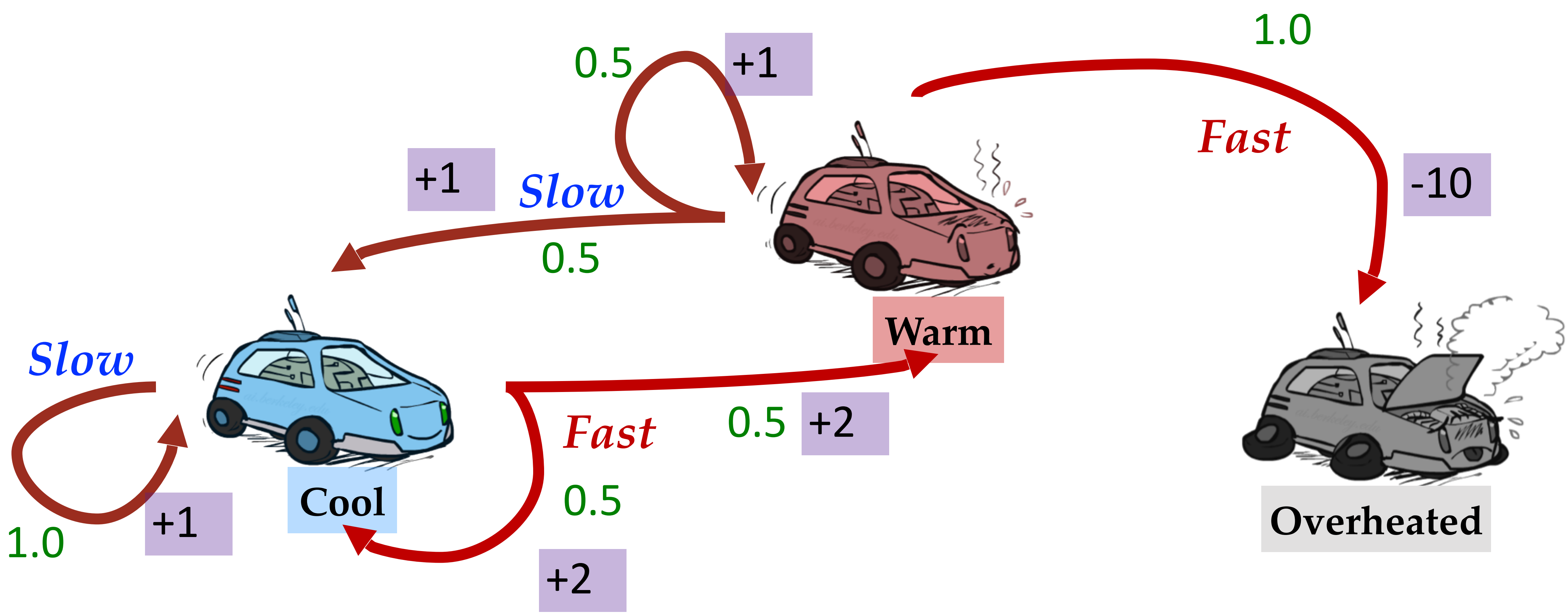
1. For $\gamma = 1$, optimal policy is (i)
2. For $\gamma = 1$, optimal policy is (ii)
3. For $\gamma = 0.1$, optimal policy is (i)
4. For $\gamma = 0.1$, optimal policy is (ii)

Example: Racing



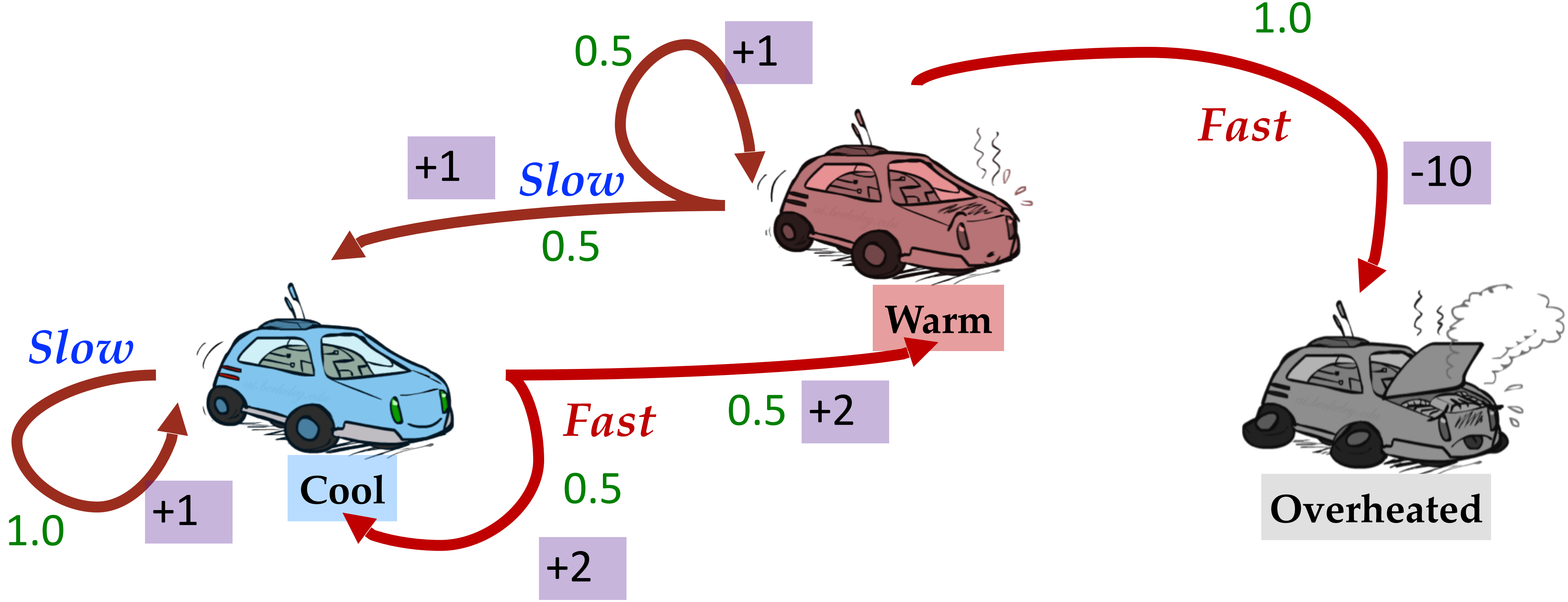
Example: Racing

- A robot car wants to travel far, quickly



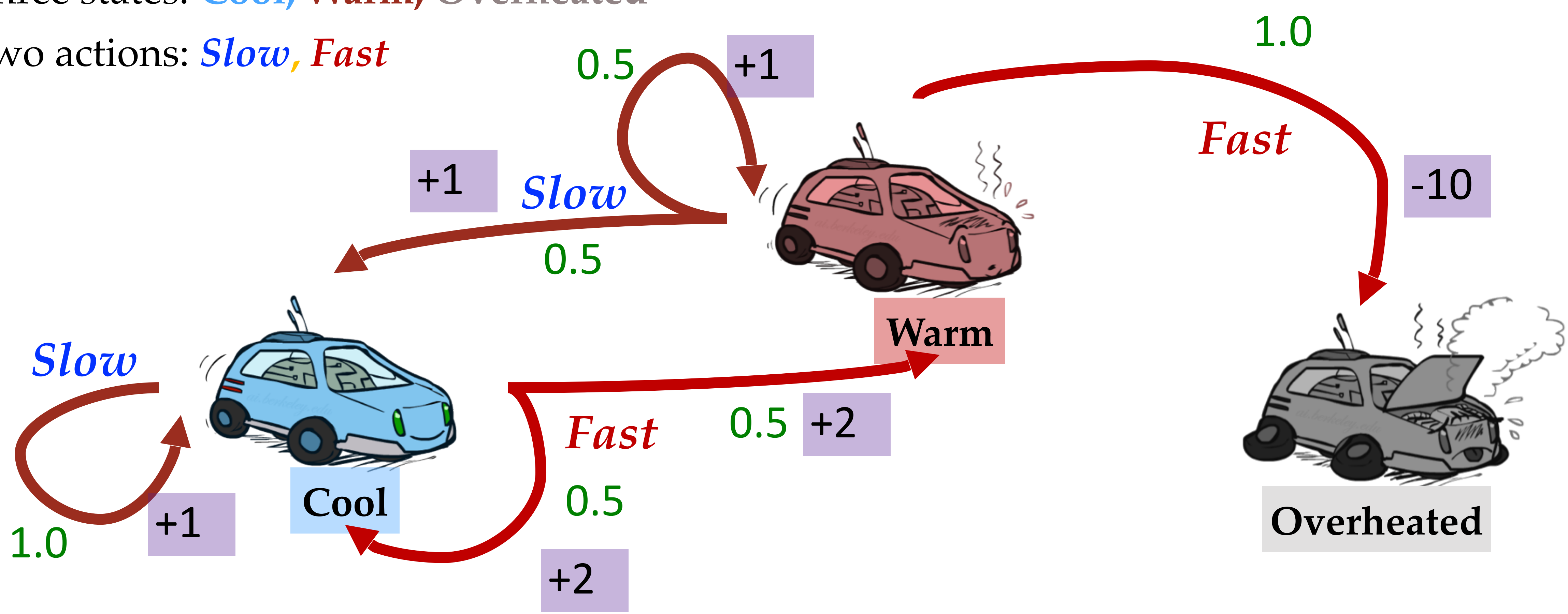
Example: Racing

- A robot car wants to travel far, quickly
- Three states: **Cool**, **Warm**, **Overheated**



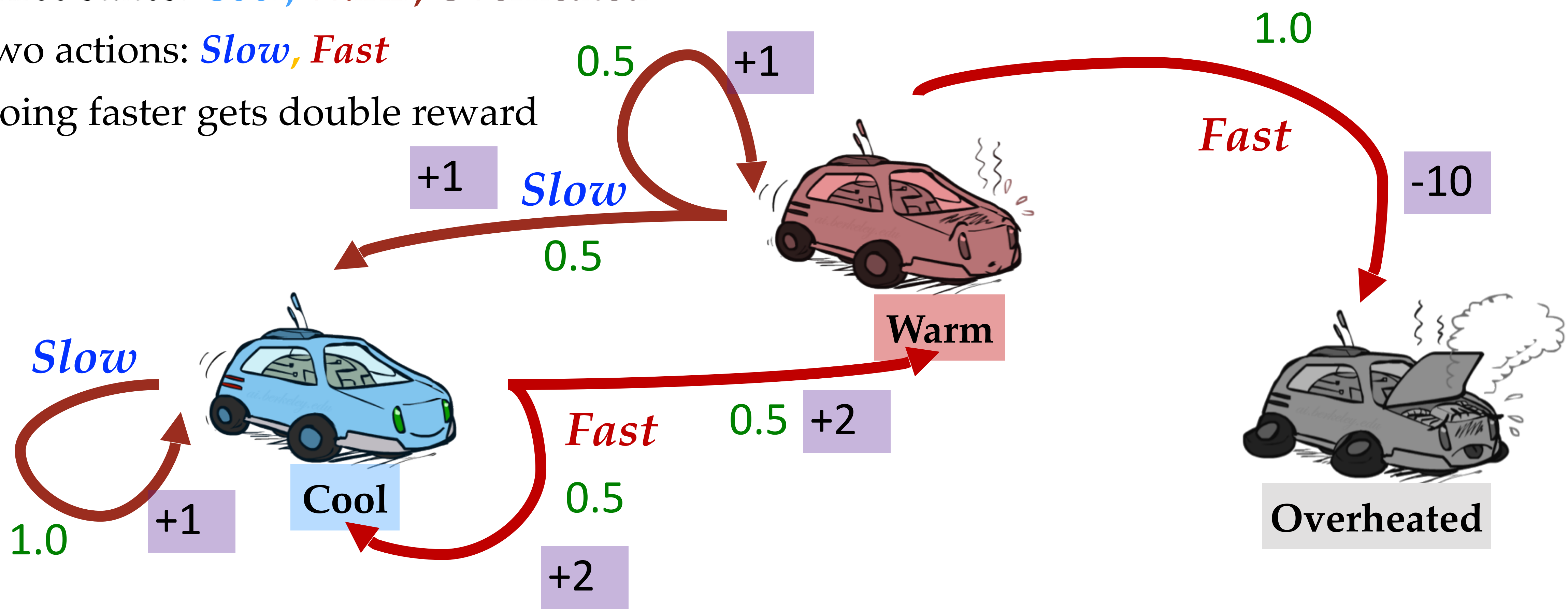
Example: Racing

- A robot car wants to travel far, quickly
- Three states: Cool, Warm, Overheated
- Two actions: *Slow*, *Fast*

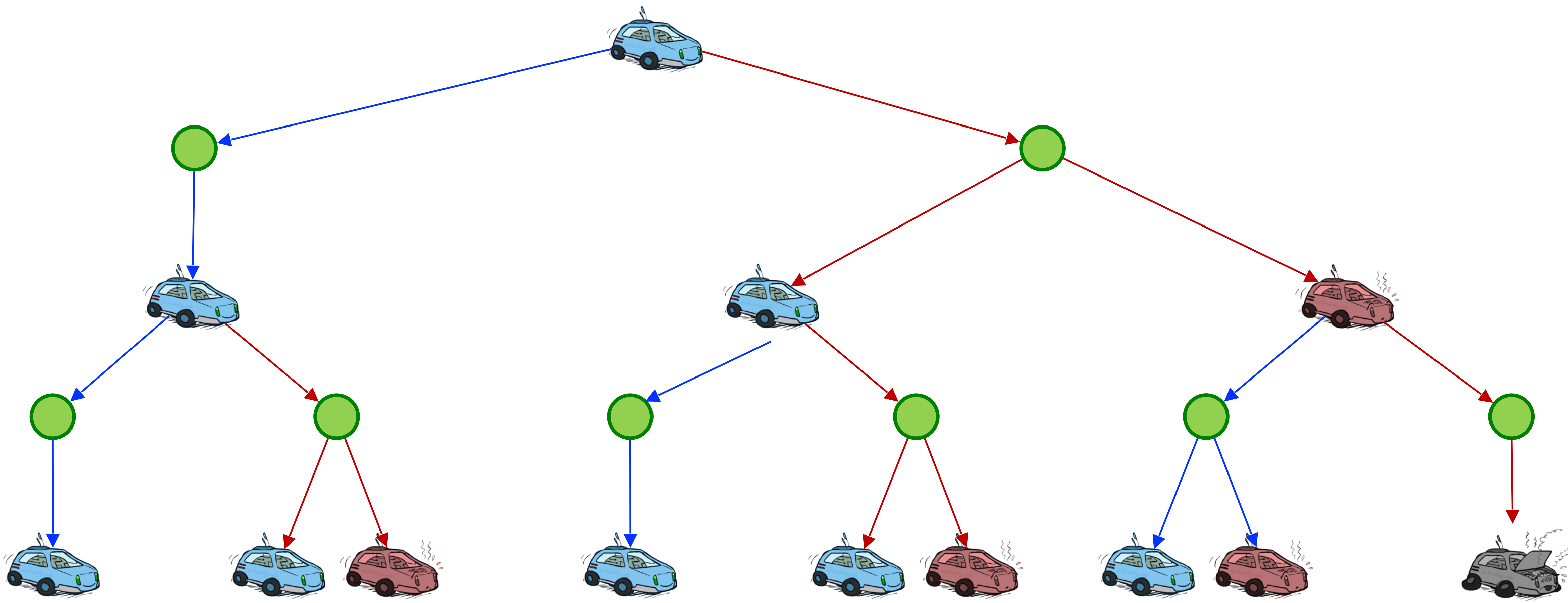


Example: Racing

- A robot car wants to travel far, quickly
- Three states: **Cool**, **Warm**, **Overheated**
- Two actions: ***Slow***, ***Fast***
- Going faster gets double reward

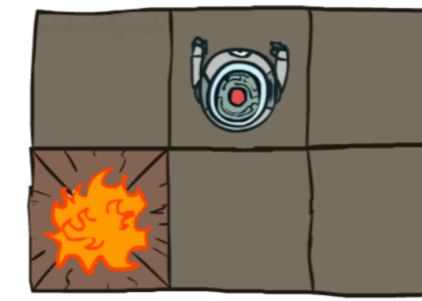


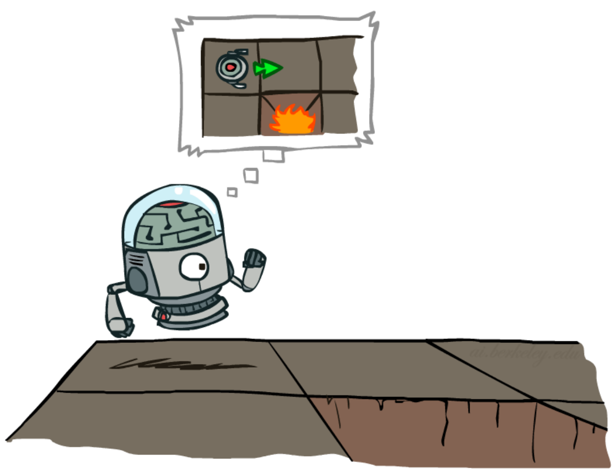
Racing Search Tree

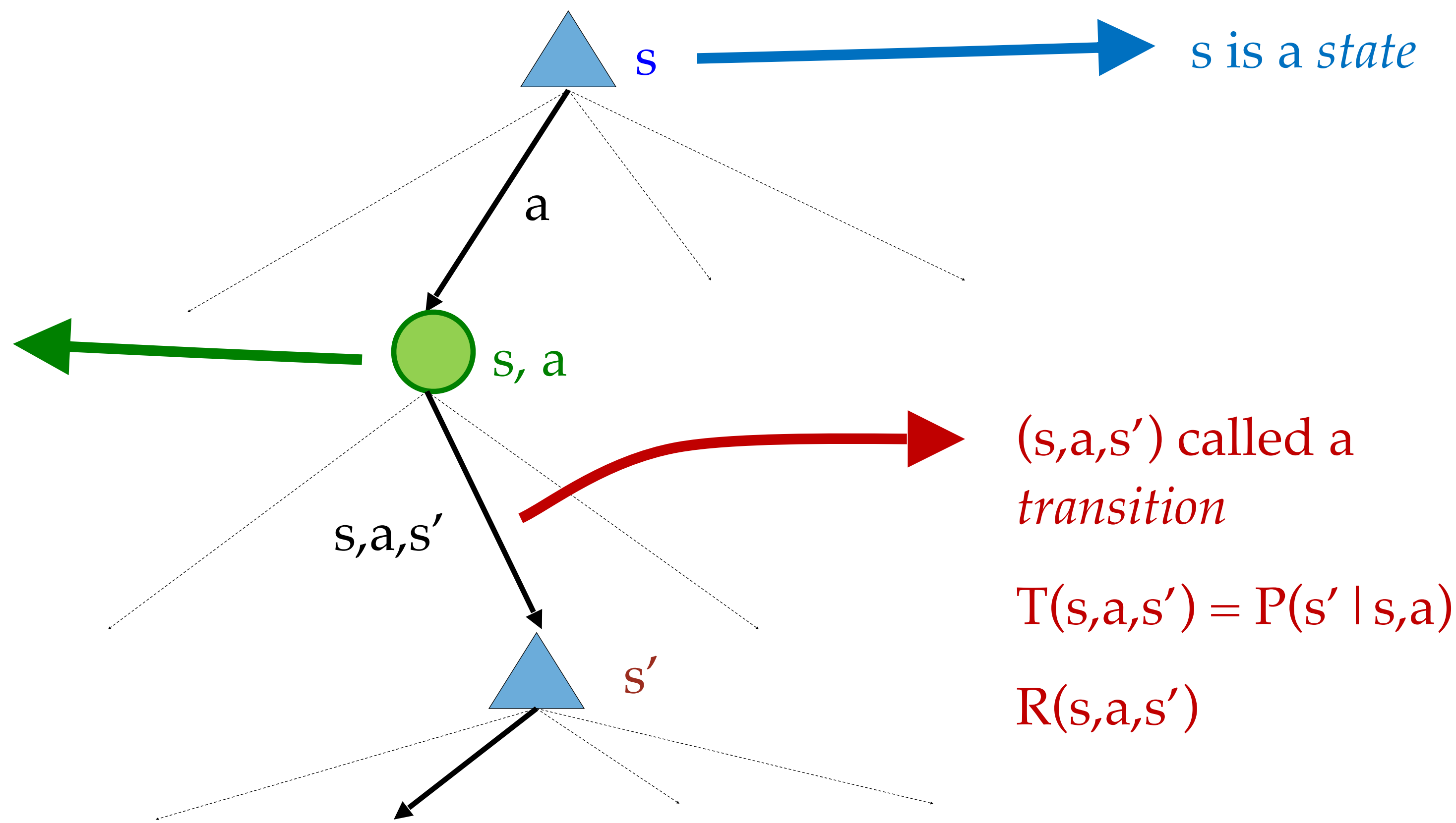


MDP Search Trees

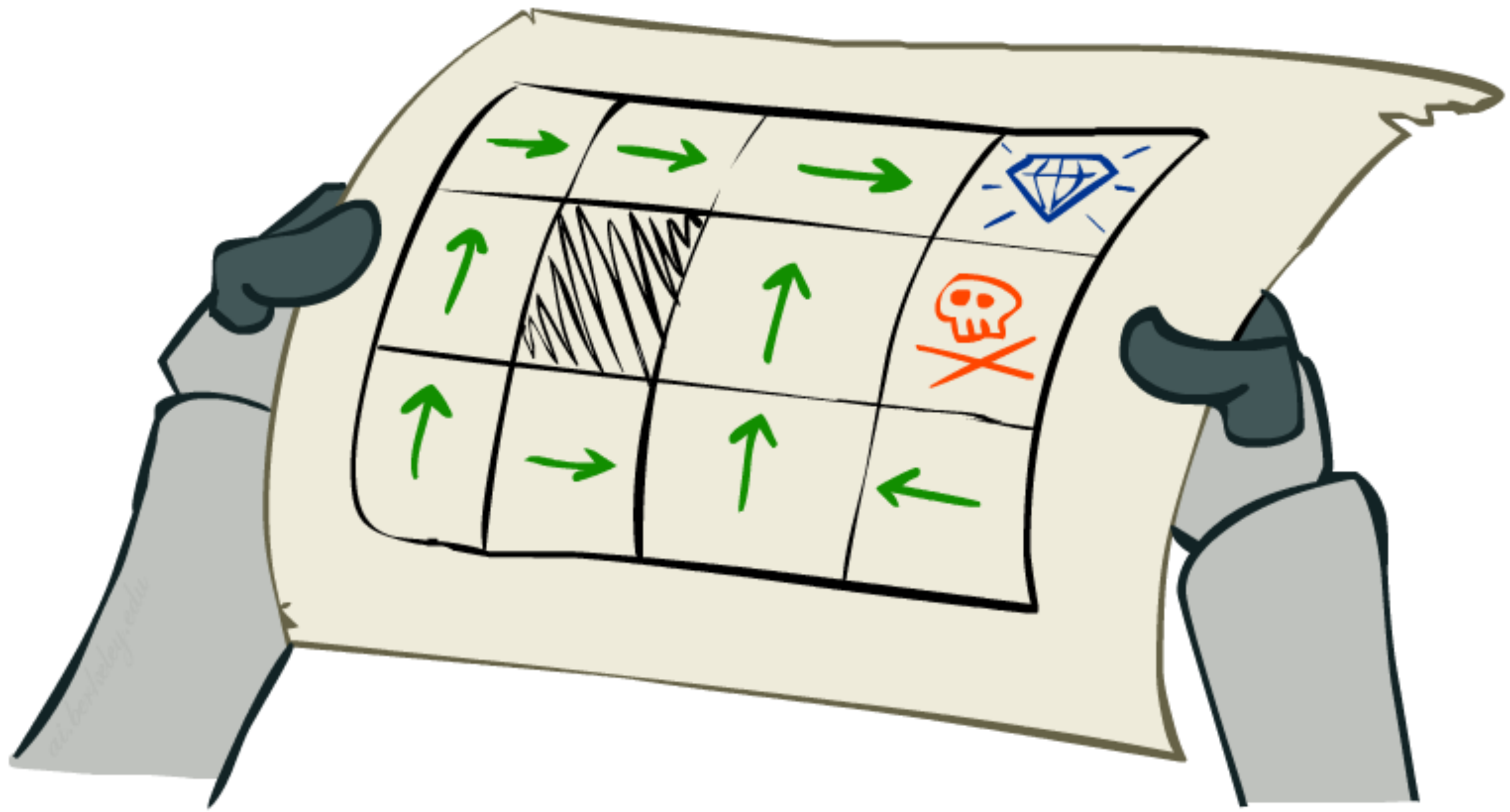
- Each MDP state projects an expectimax-like search tree







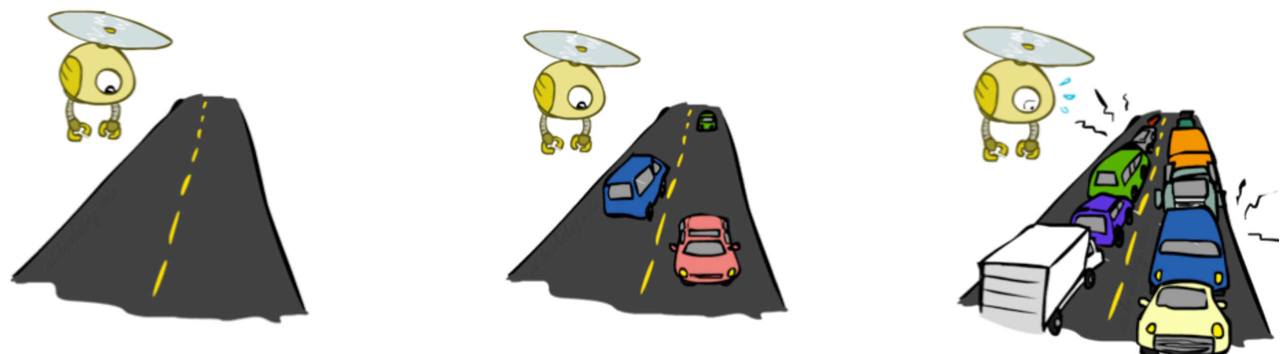
Solving MDPs

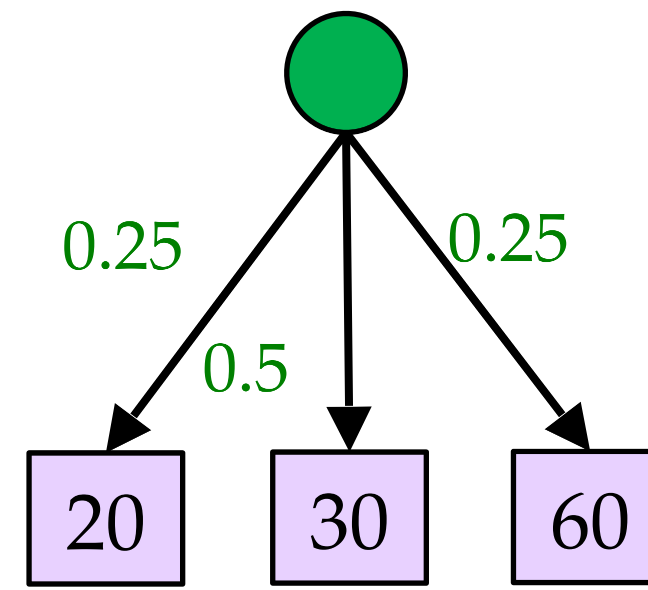


Finding Optimal Policy

- Expectimax algorithm! (studied in search module)

Time:	20 min		30 min		60 min
	x	+	x	+	x
Probability:	0.25		0.50		0.25



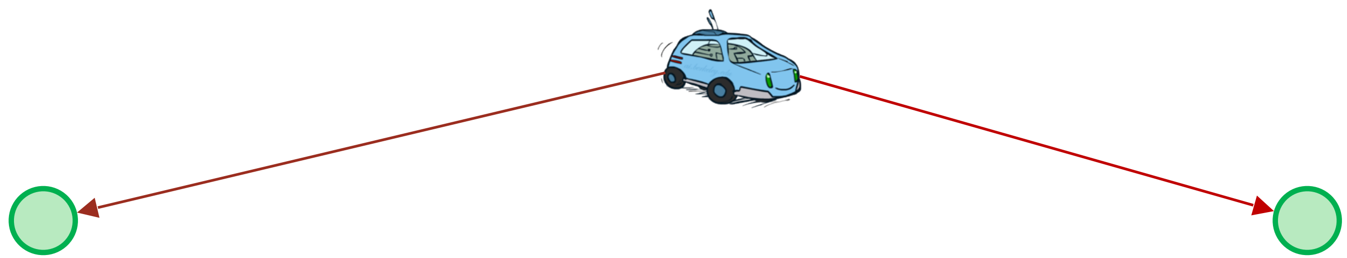


Chance node notation

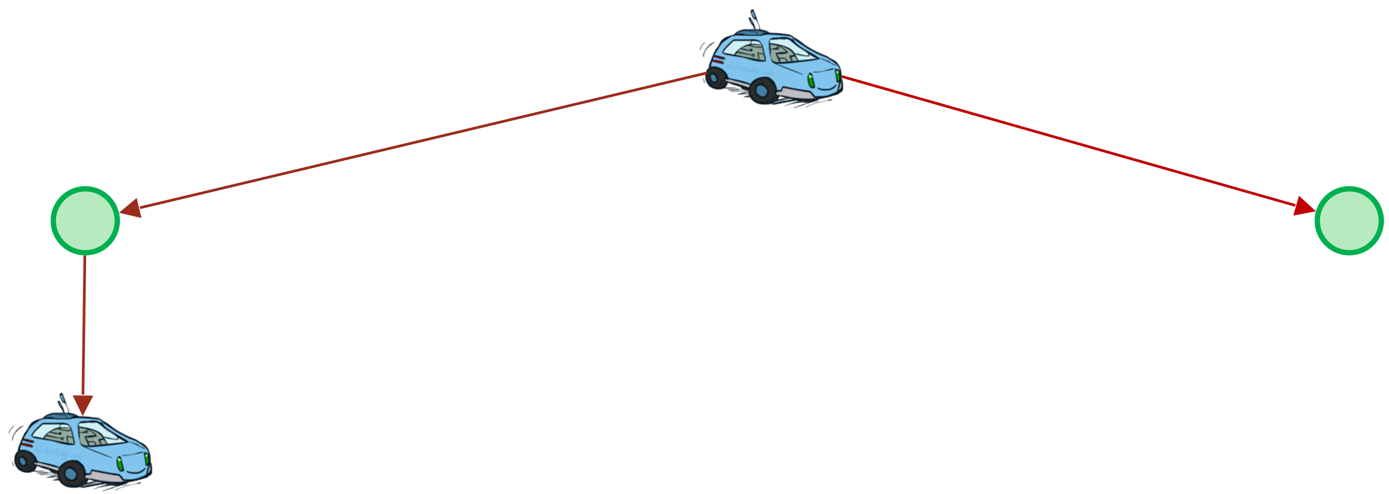
$$V(s) = \sum_{s'} \left[P(s')\, V(s') \right]$$

Racing Search Tree

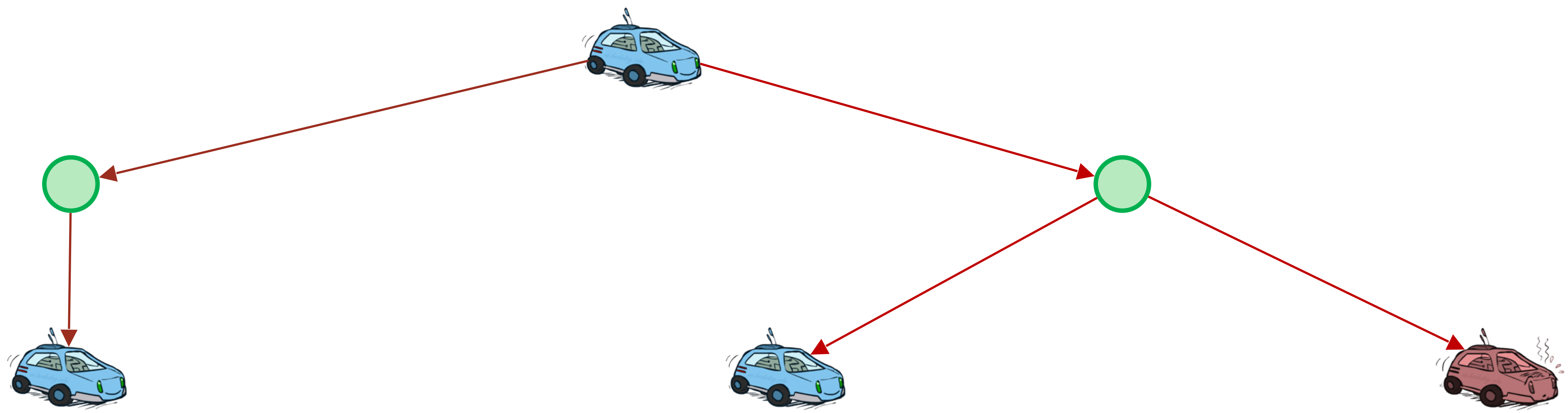
Racing Search Tree



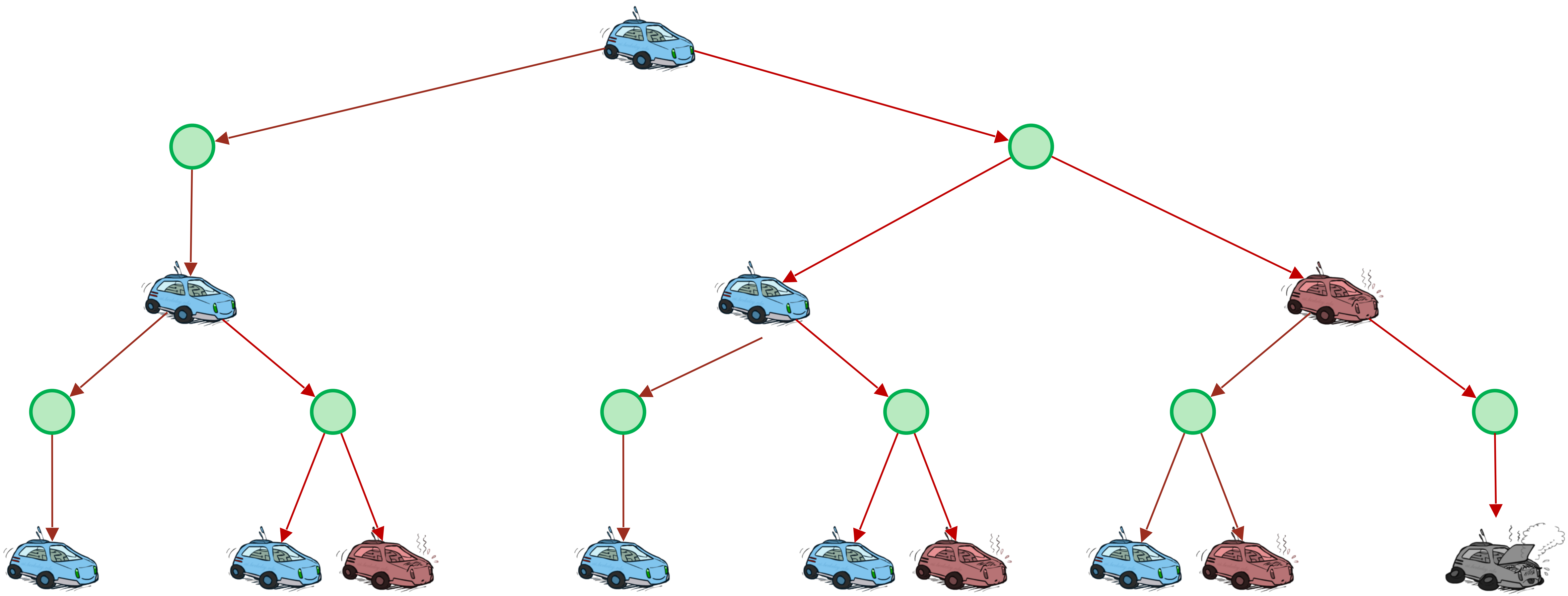
Racing Search Tree



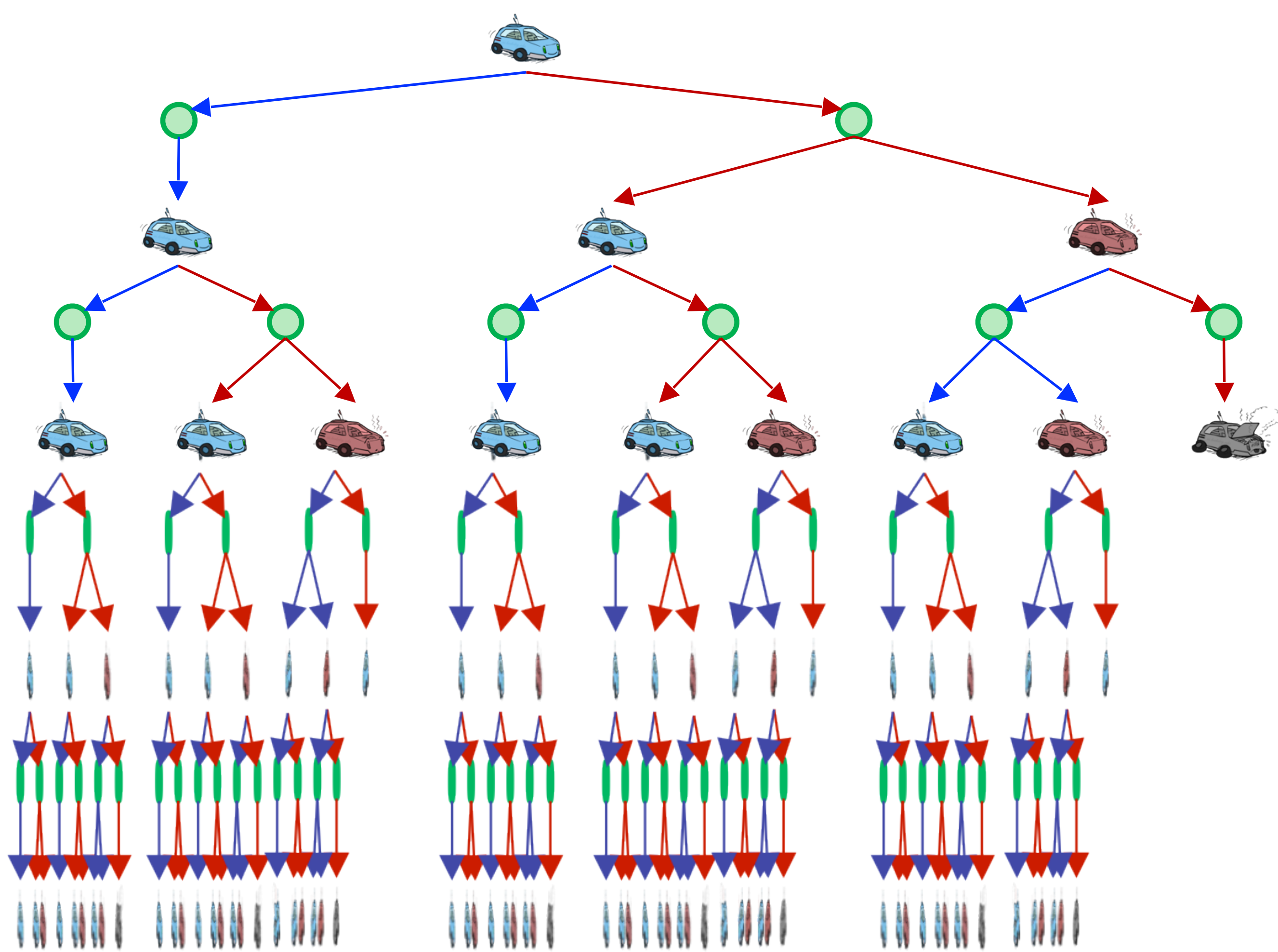
Racing Search Tree



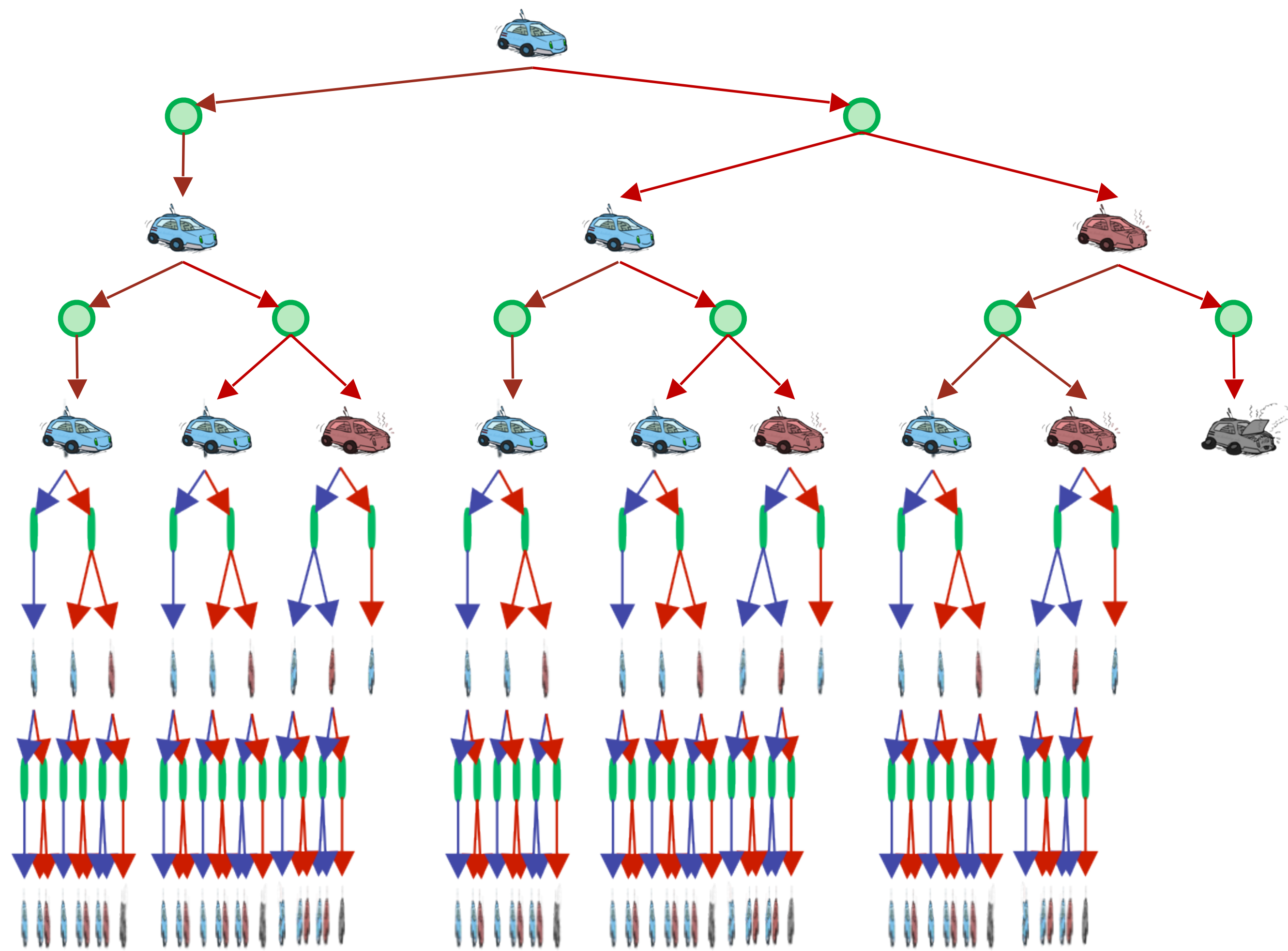
Racing Search Tree



Racing Search Tree

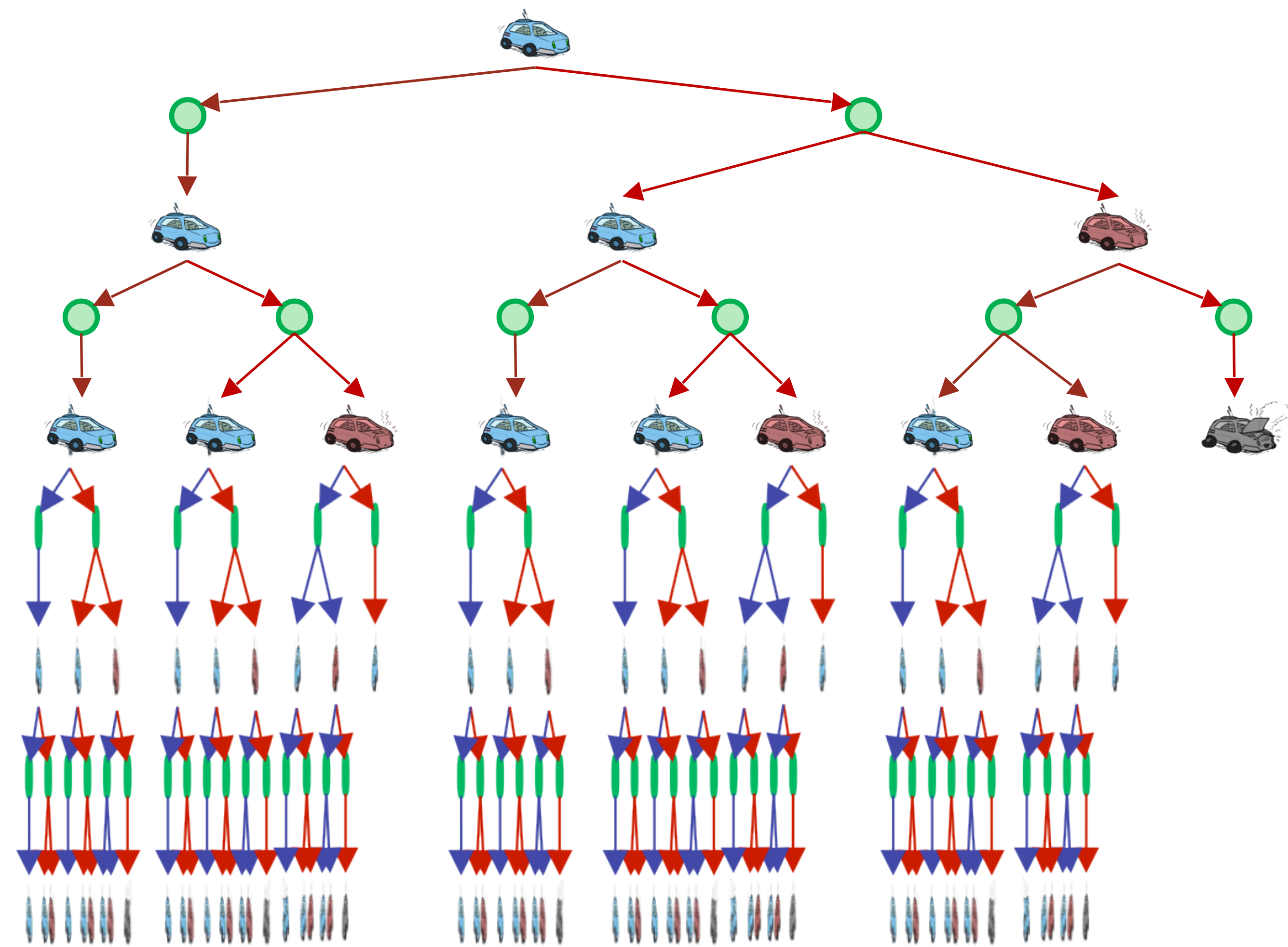


Racing Search Tree



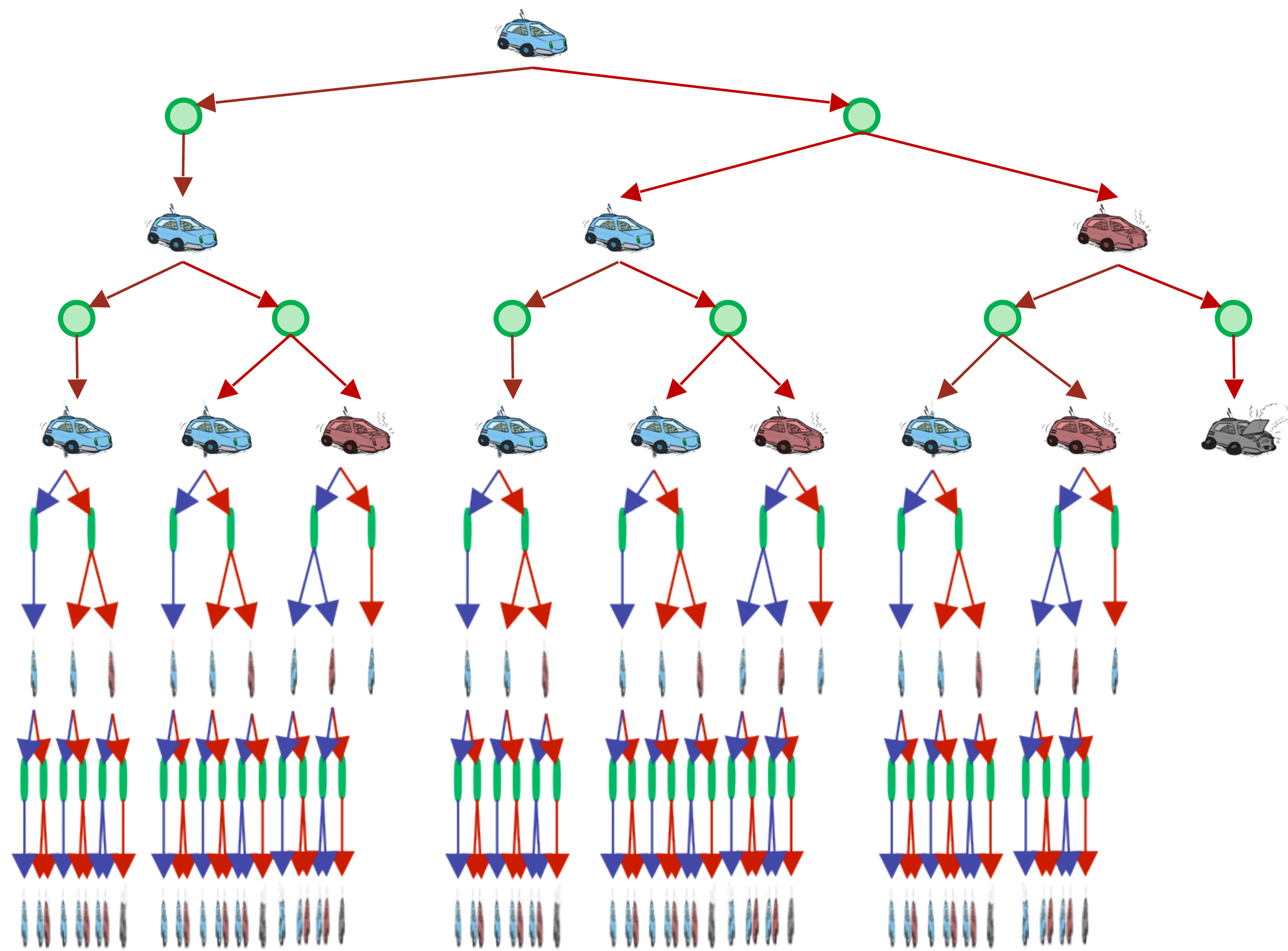
Racing Search Tree

- We're doing way too much work with expectimax!
- Problem: States are repeated
  - Idea: Only compute needed quantities once



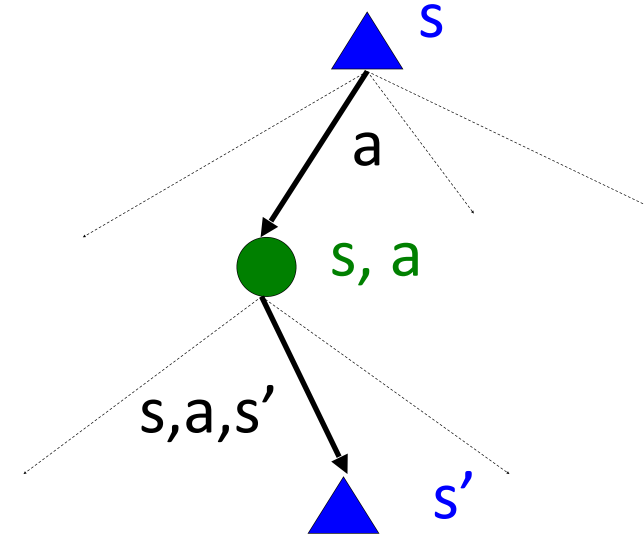
Racing Search Tree

- We're doing way too much work with expectimax!
- Problem: States are repeated
  - Idea: Only compute needed quantities once
- Problem: Tree goes on forever
  - Idea: Do a depth-limited computation, but with increasing depths until change is small
  - Note: deep parts of the tree eventually don't matter if $\gamma < 1$



Recap: Defining MDPs

- Markov decision processes:
  - Set of states S
  - Start state s_0
  - Set of actions A
  - Transitions P(s' | s,a) (or T(s,a,s'))
  - Rewards R(s,a,s') (and discount γ)



- MDP quantities so far:
  - Policy = Choice of action for each state
  - Utility = sum of (discounted) rewards

Optimal Quantities

- The value (utility) of a state s:

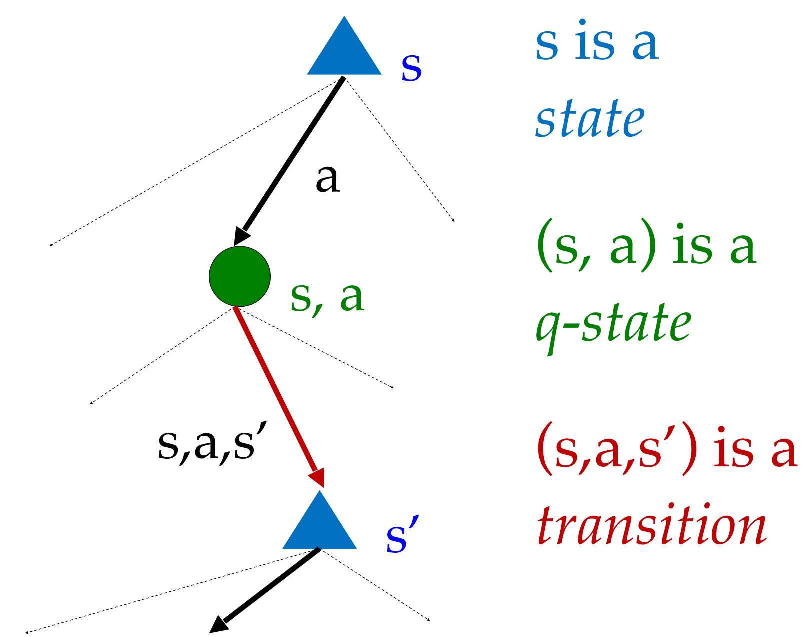
  $V^*(s)$ = expected utility starting in s and acting optimally

- The value (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

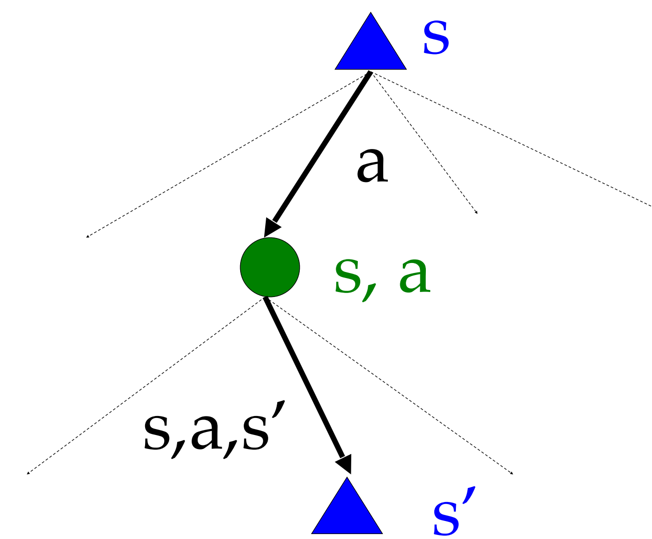
- The optimal policy:

  $\pi^*(s)$ = optimal action from state s

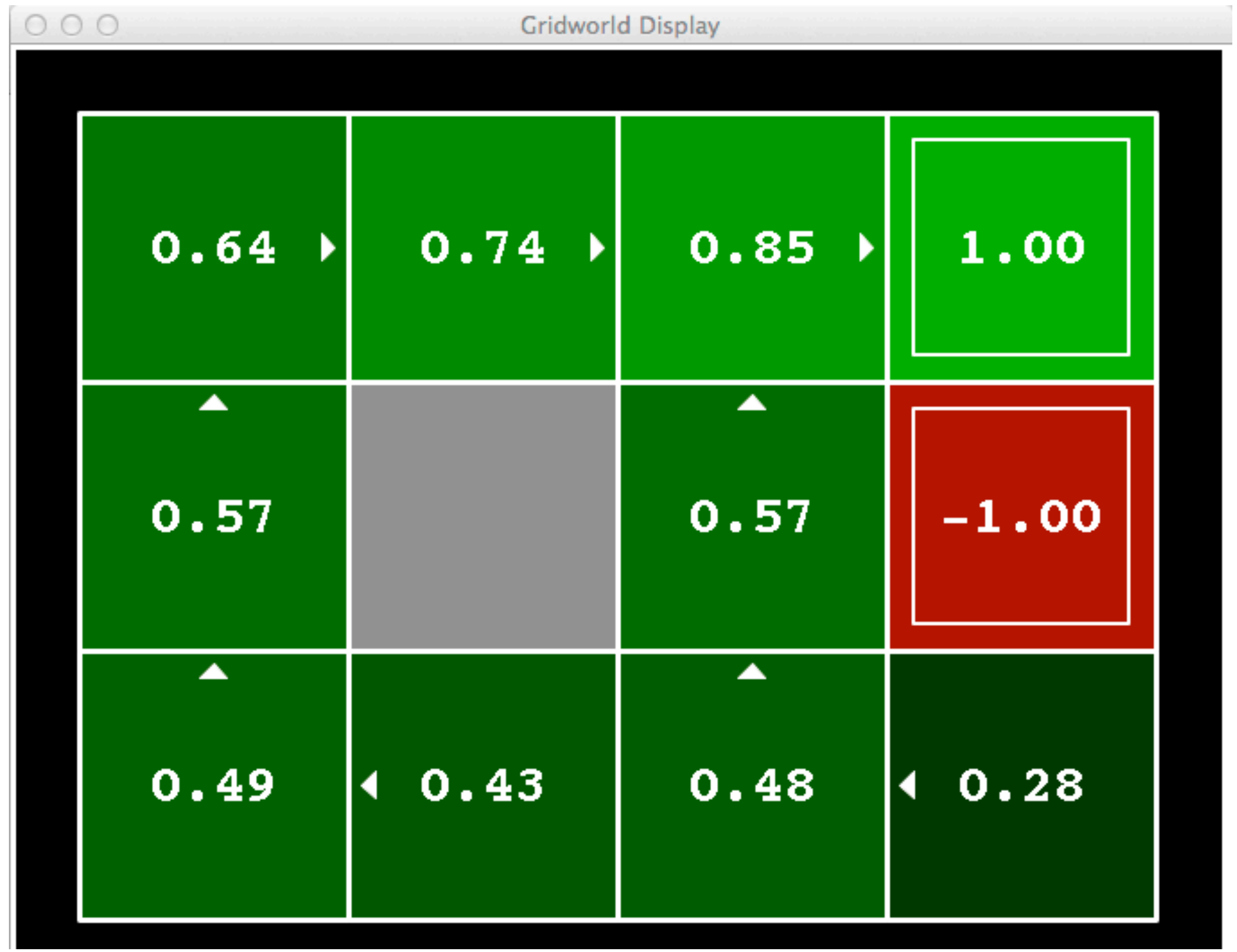


Relationship b/w Optimal Quantities

- $V^*(s)$ in terms of $Q^*(s, a)$
- $Q^*(s, a)$ in terms of $V^*(s)$
- $\pi^*(s)$ in terms of $Q^*(s, a)$

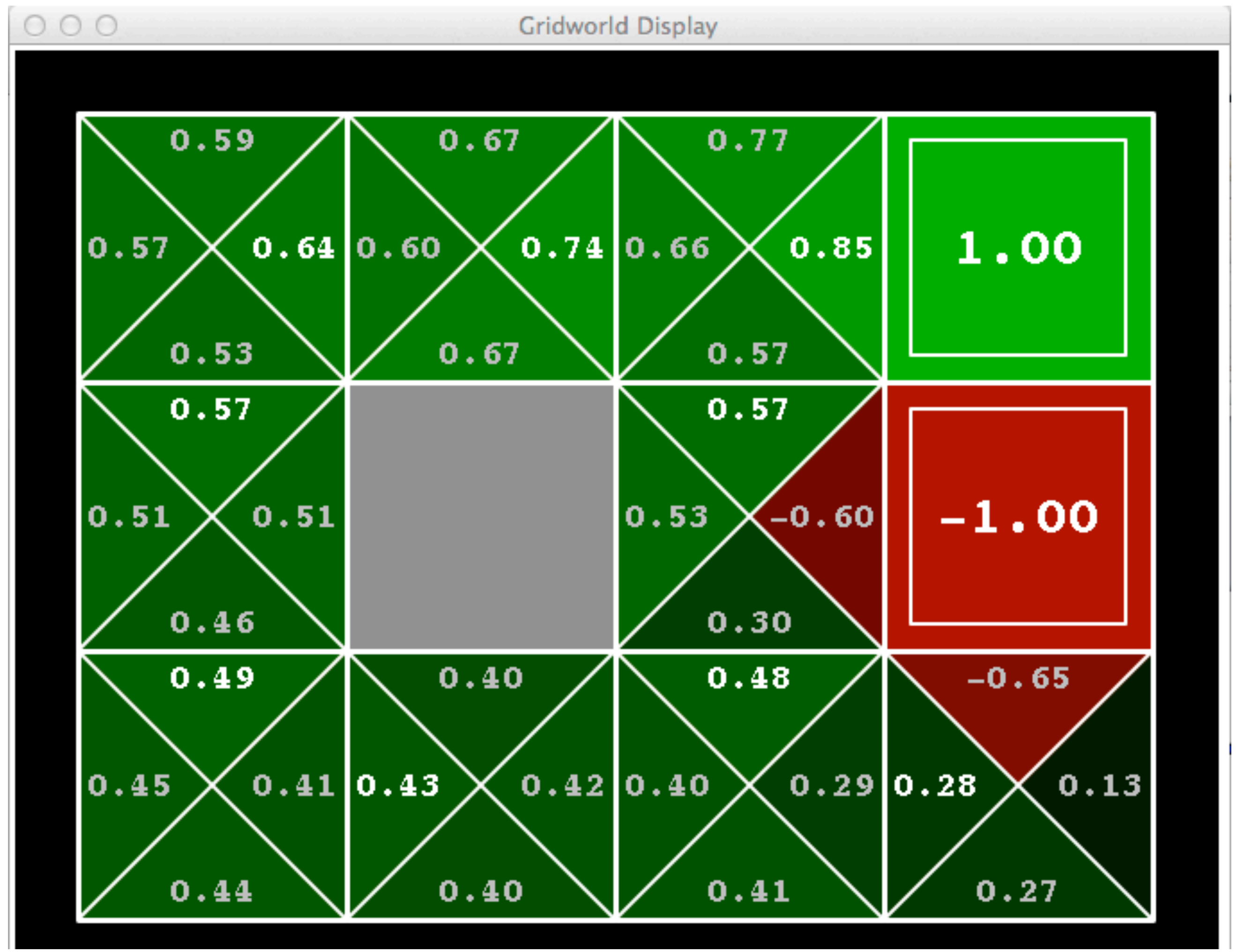


Gridworld V* Values



Noise = 0.2
Discount = 0.9
Living reward = 0

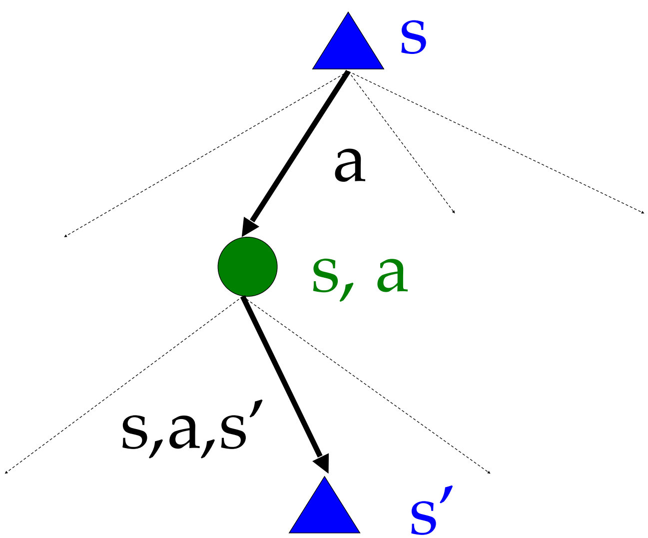
Gridworld Q* Values



Noise = 0.2
Discount = 0.9
Living reward = 0

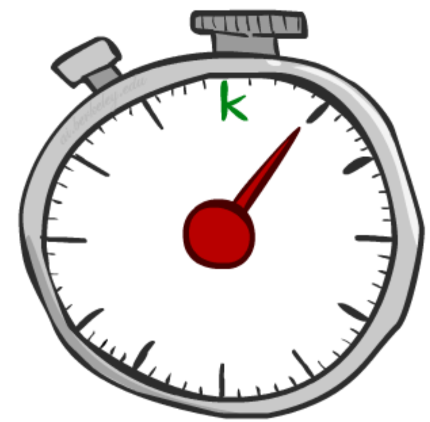
Relationship b/w Optimal Quantities

- $V^*(s)$ in terms of $Q^*(s, a)$
- $Q^*(s, a)$ in terms of $V^*(s)$
- $\pi^*(s)$ in terms of $Q^*(s, a)$
- Recursive definition for V^*



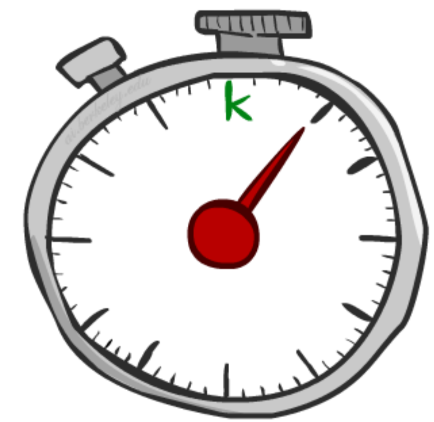
Time-Limited Values

- Key idea: time-limited values



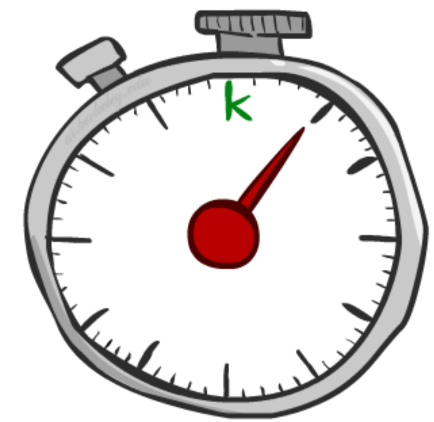
Time-Limited Values

- Key idea: time-limited values
- Define $V_k(s)$ to be the optimal value of s if the game ends in k more time steps



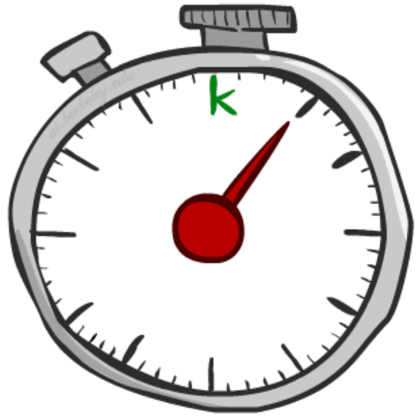
Time-Limited Values

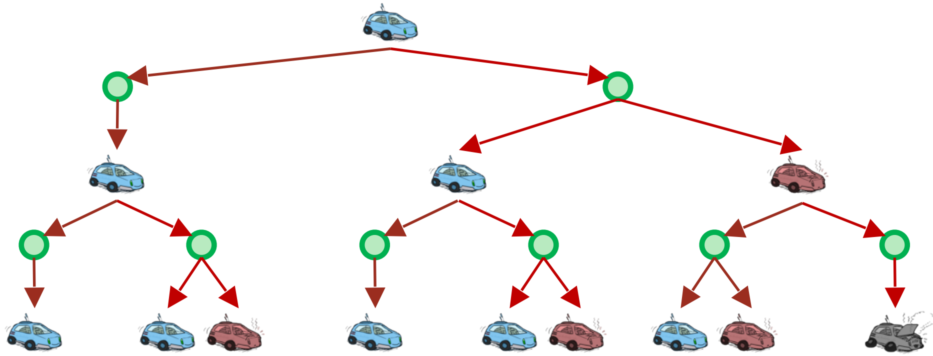
- Key idea: time-limited values
- Define $V_k(s)$ to be the optimal value of s if the game ends in k more time steps
  - Equivalently, it's what a depth-k expectimax would give from s



Time-Limited Values

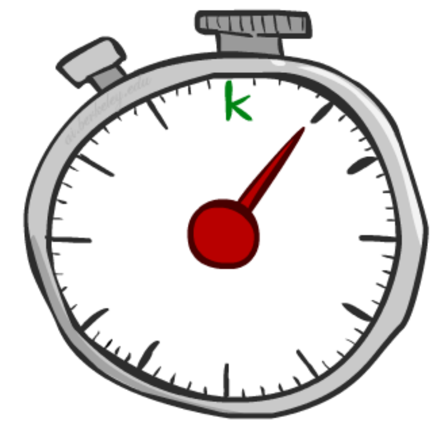
- Key idea: time-limited values
- Define $V_k(s)$ to be the optimal value of s if the game ends in k more time steps
  - Equivalently, it's what a depth-k expectimax would give from s

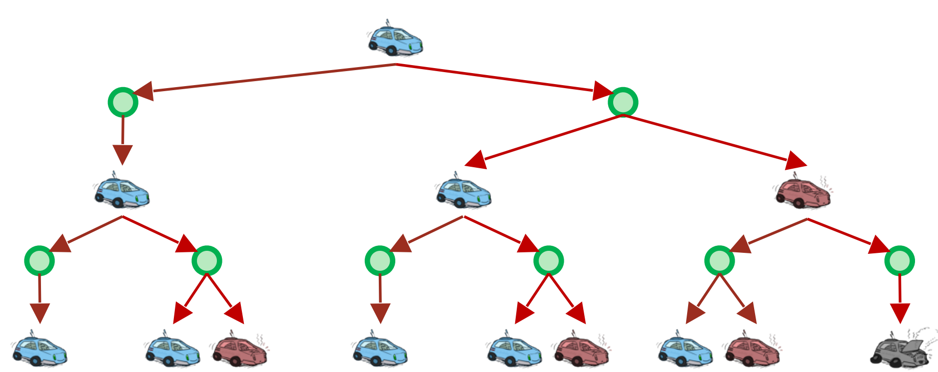


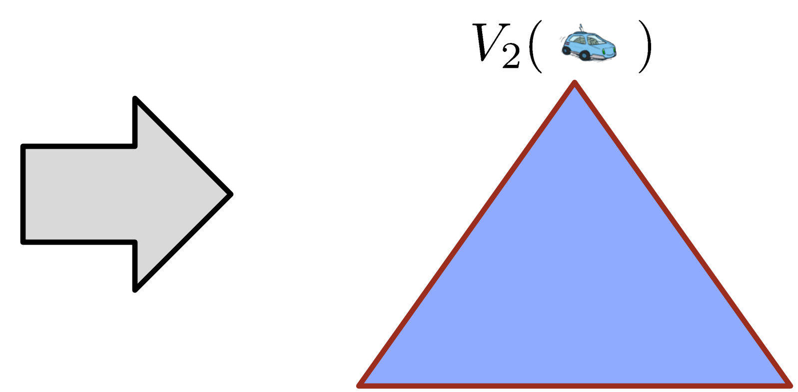


Time-Limited Values

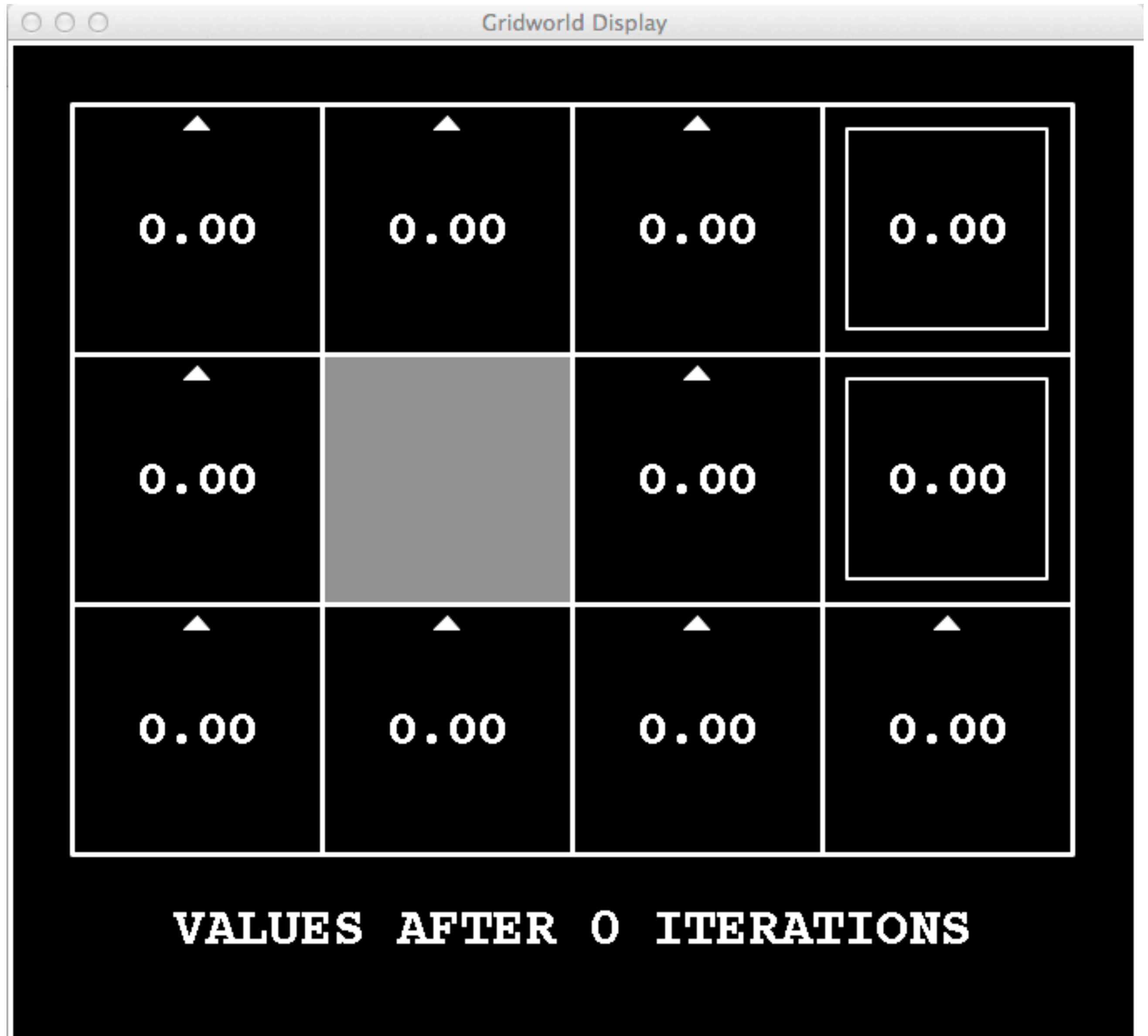
- Key idea: time-limited values
- Define $V_k(s)$ to be the optimal value of s if the game ends in k more time steps
  - Equivalently, it's what a depth-k expectimax would give from s





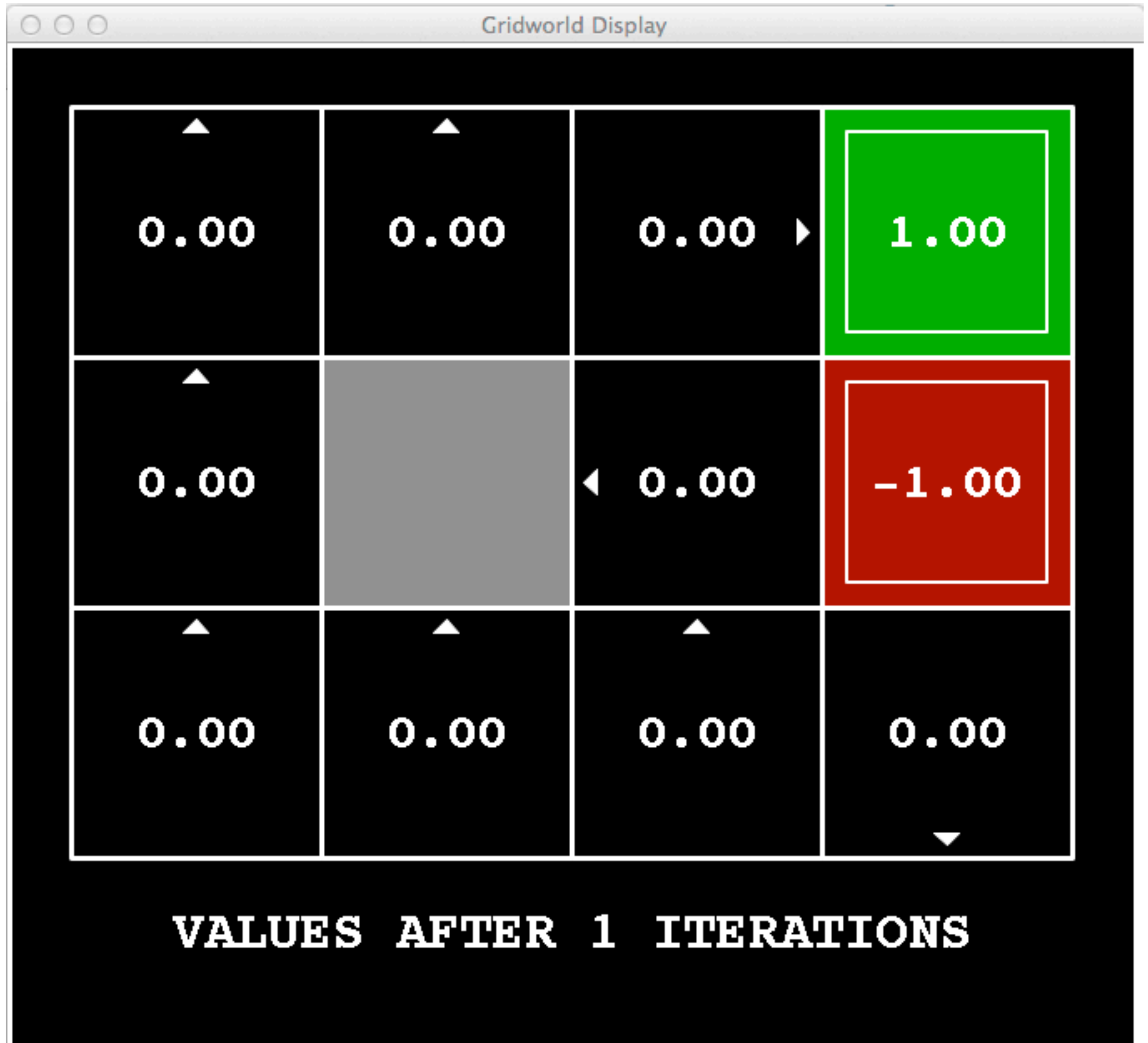


k=0



Noise = 0.2
Discount = 0.9
Living reward = 0

k=1



Noise = 0.2
Discount = 0.9
Living reward = 0

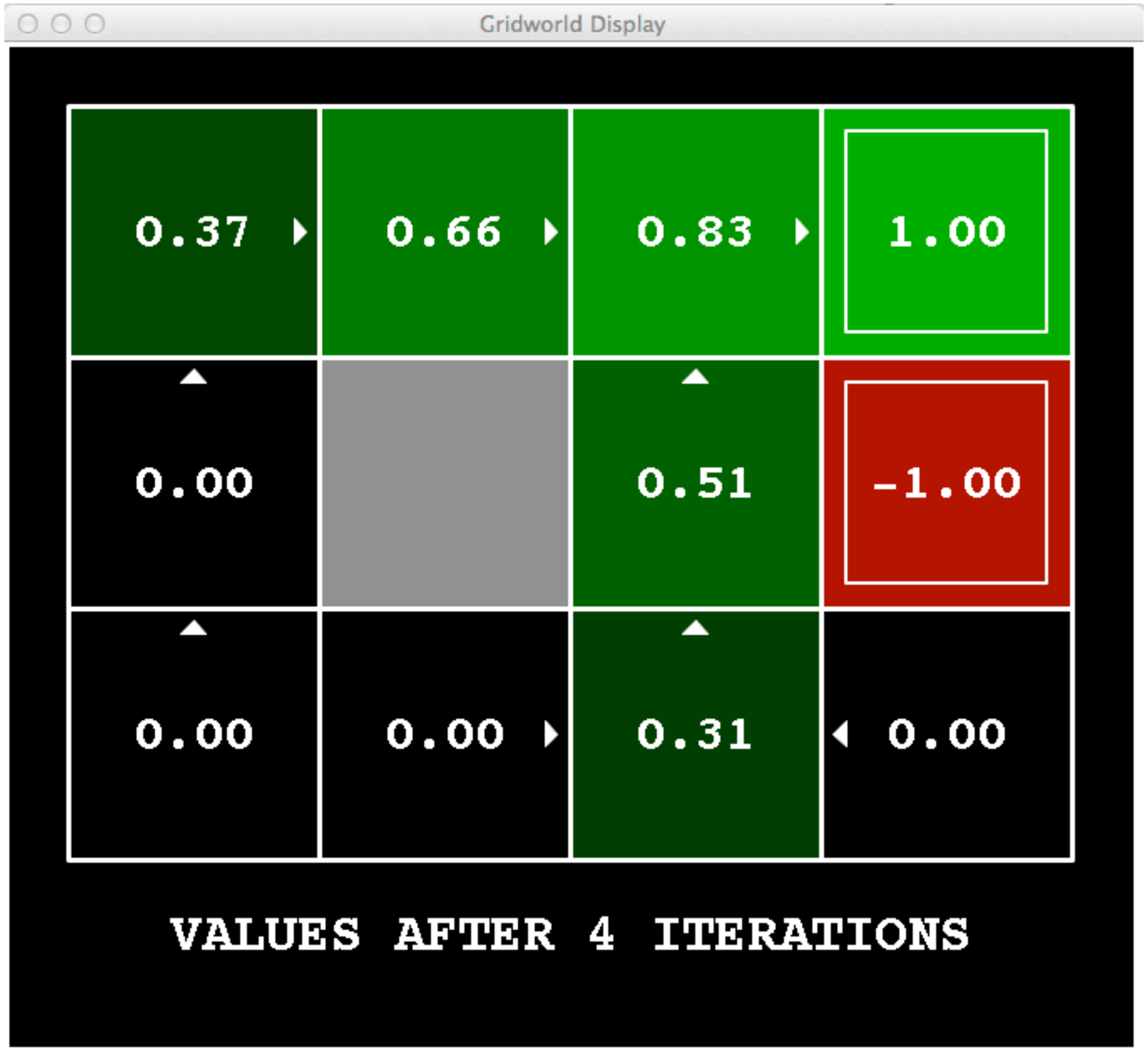
k=2

Noise = 0.2
Discount = 0.9
Living reward = 0

k=3

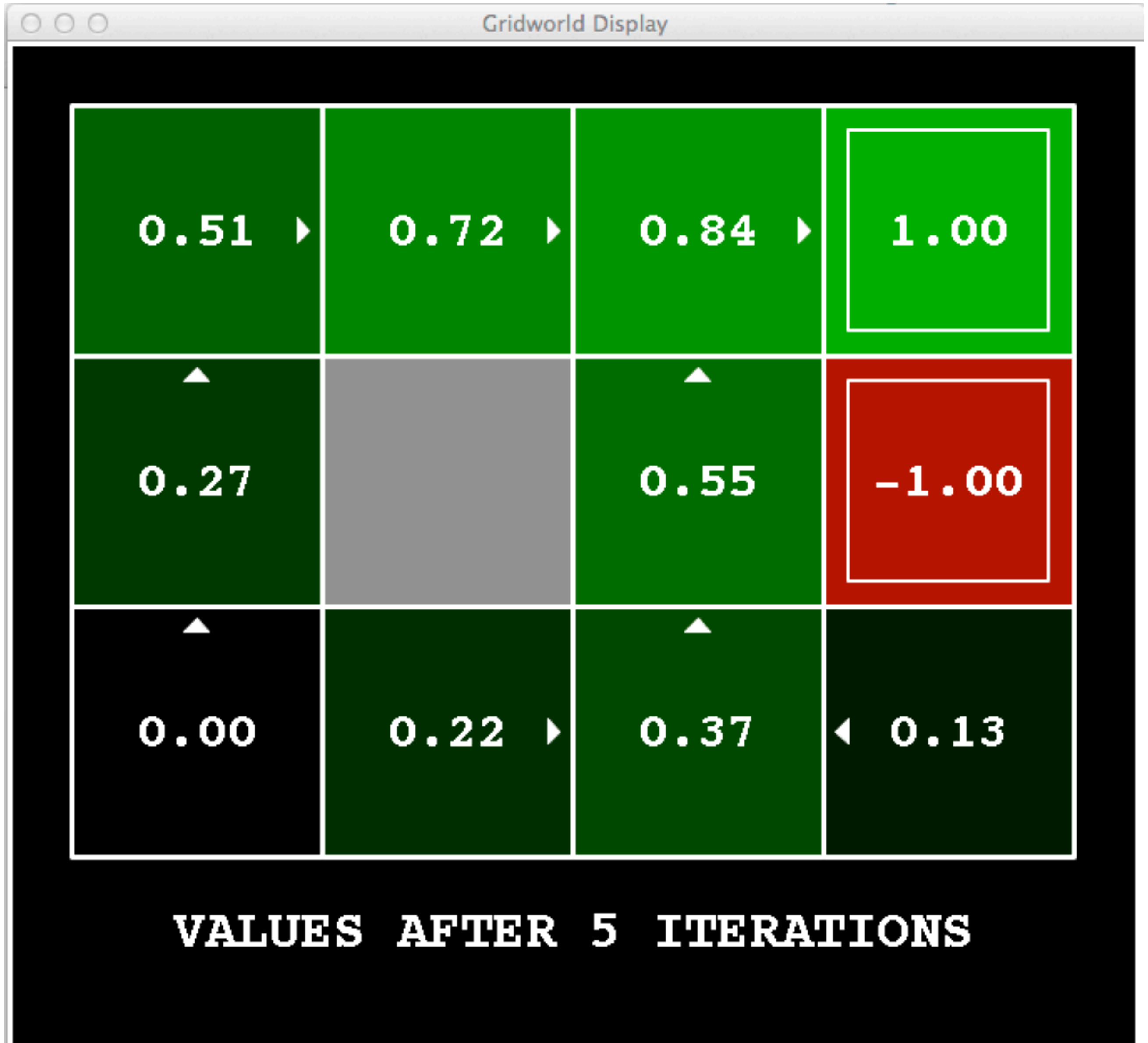
Noise = 0.2
Discount = 0.9
Living reward = 0

k=4



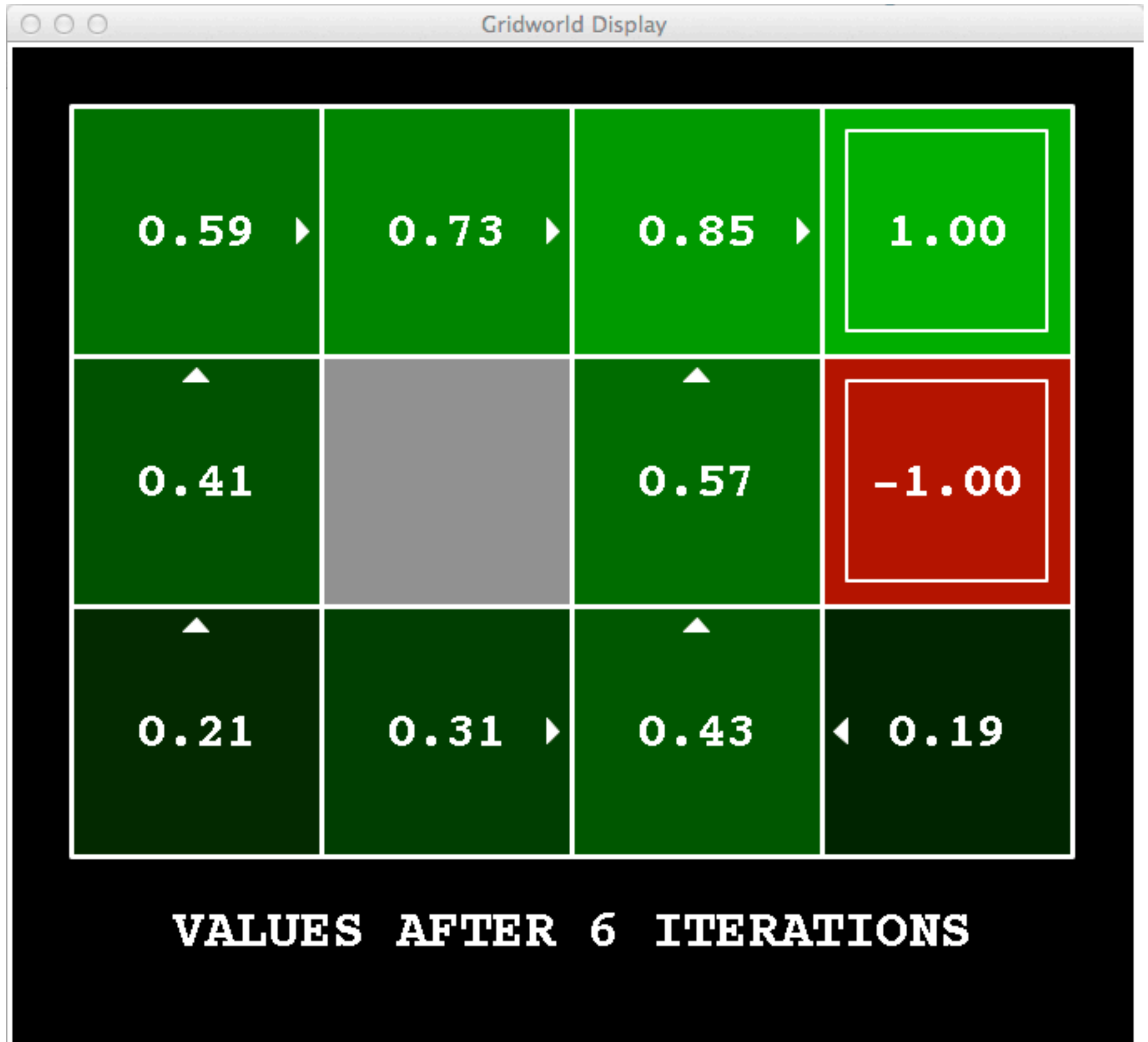
Noise = 0.2
Discount = 0.9
Living reward = 0

k=5



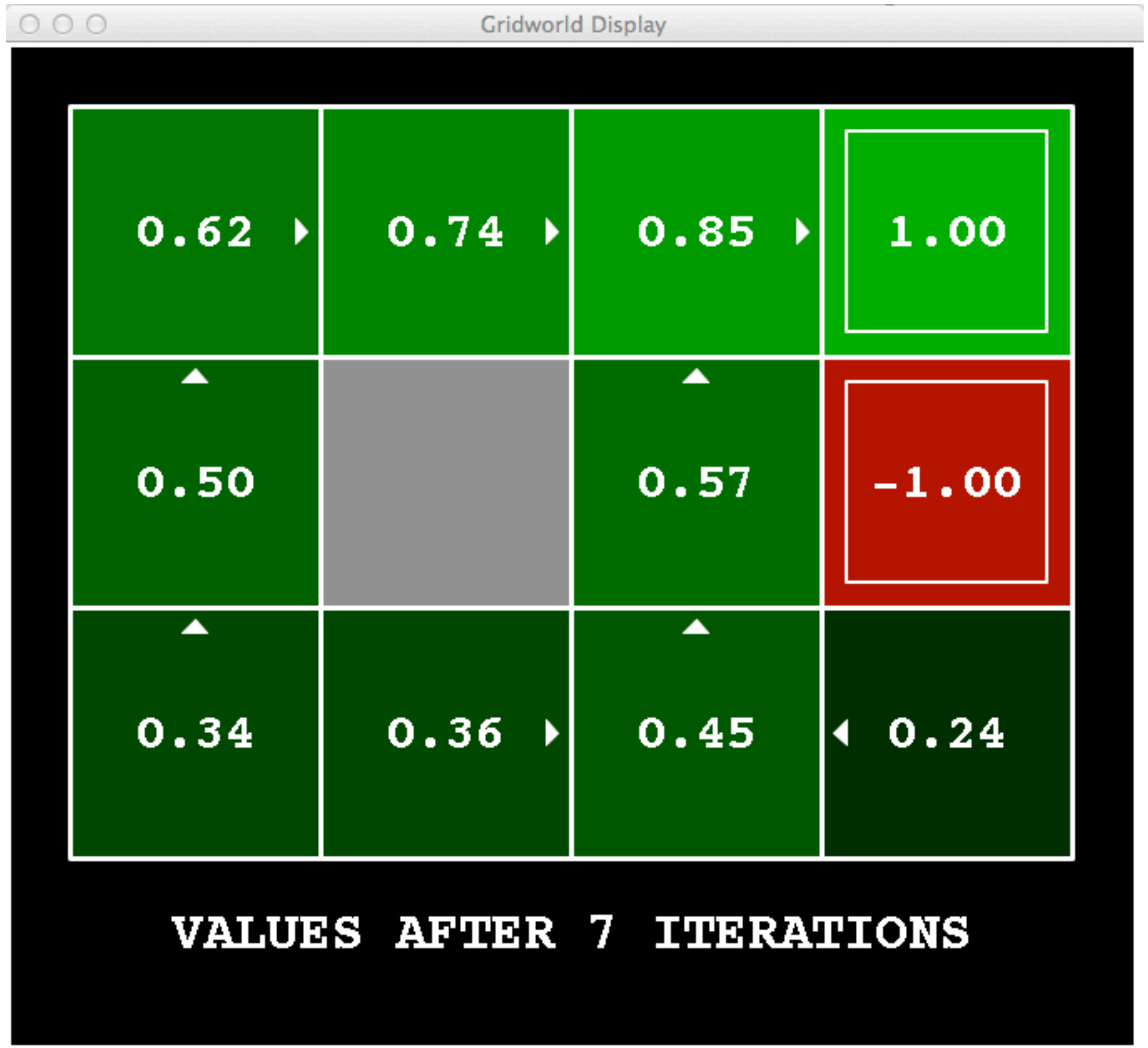
Noise = 0.2
Discount = 0.9
Living reward = 0

k=6



Noise = 0.2
Discount = 0.9
Living reward = 0

k=7

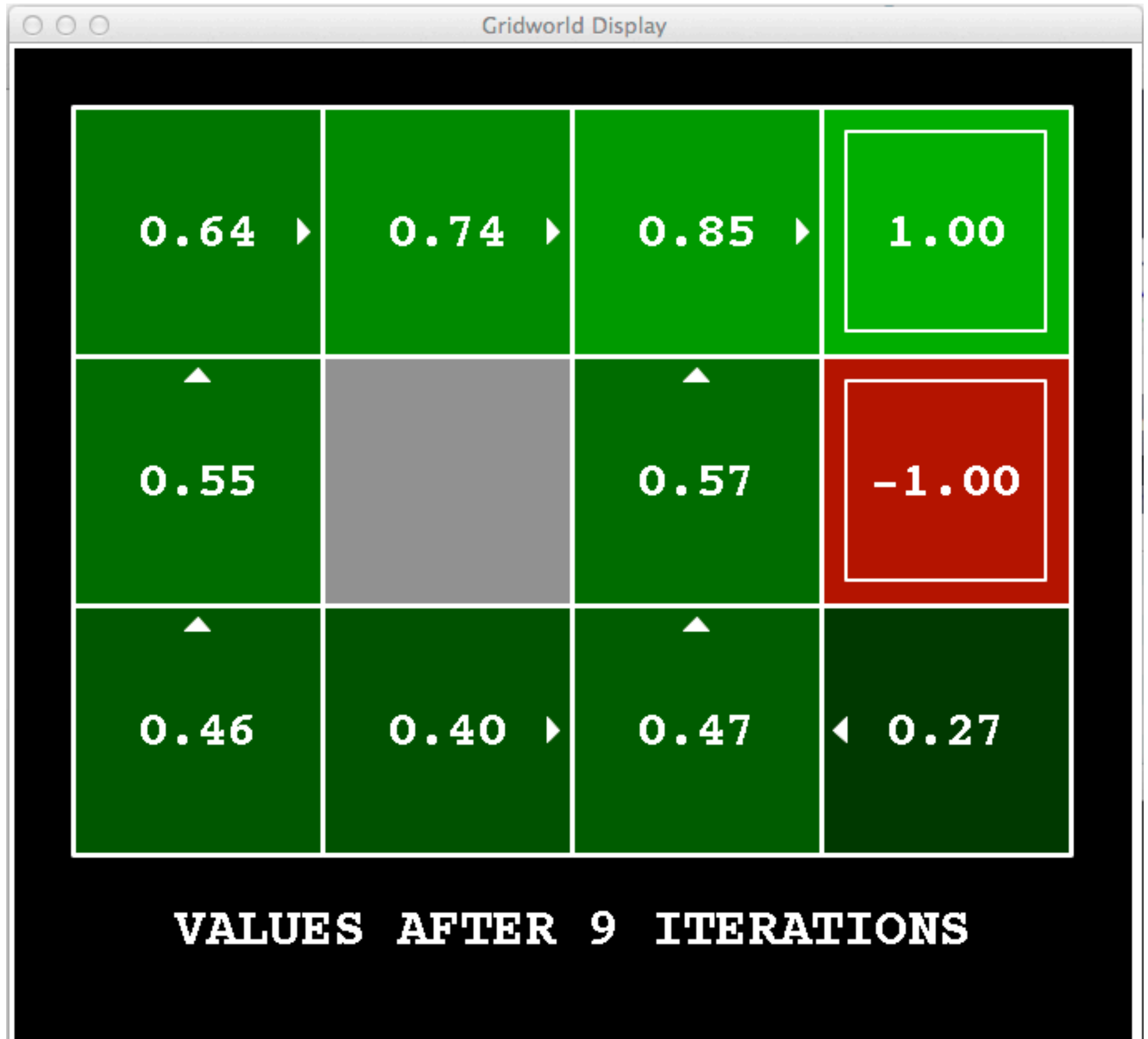


Noise = 0.2
Discount = 0.9
Living reward = 0

k=8

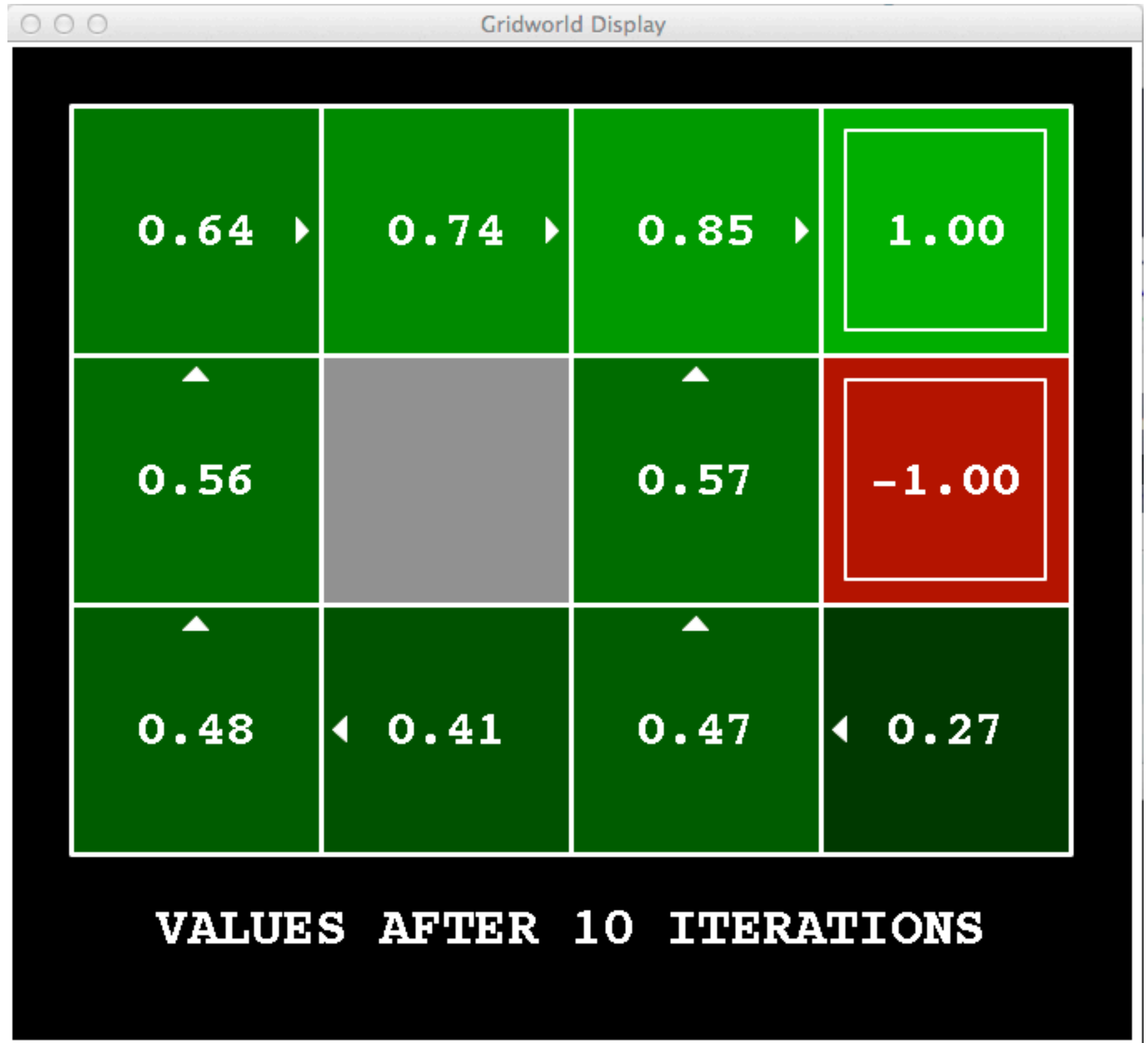
Noise = 0.2
Discount = 0.9
Living reward = 0

k=9



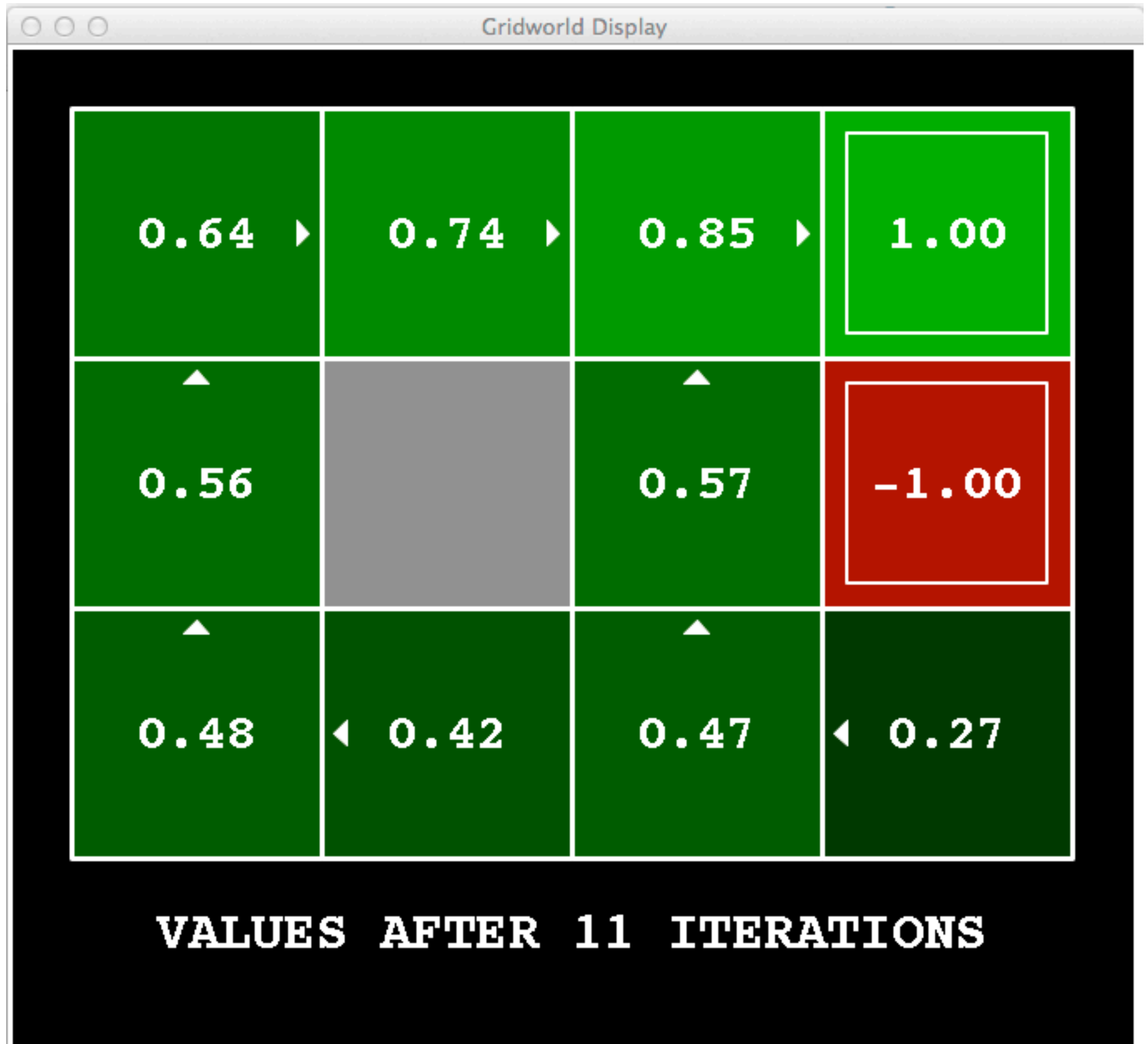
Noise = 0.2
Discount = 0.9
Living reward = 0

k=10



Noise = 0.2
Discount = 0.9
Living reward = 0

k=11

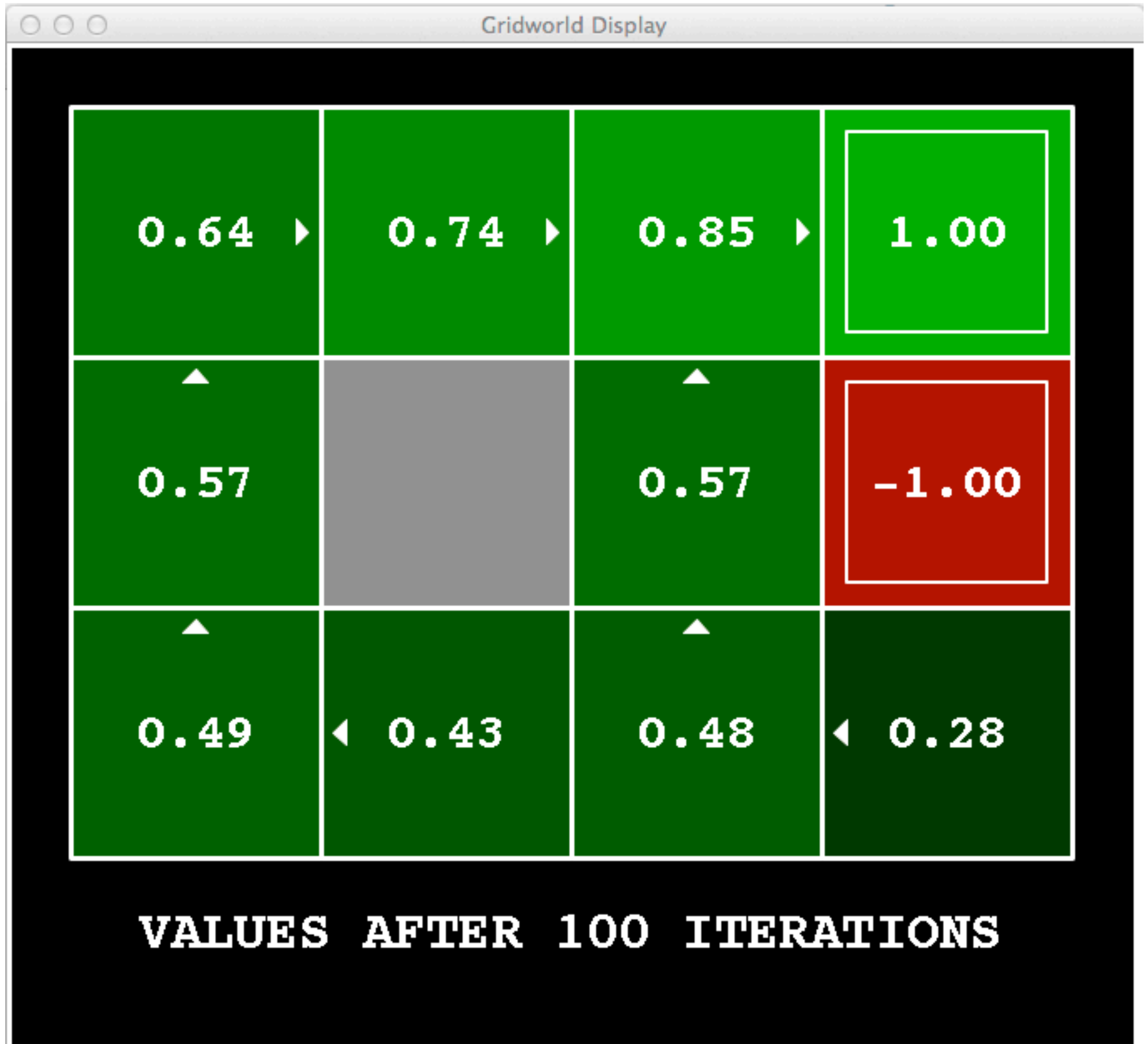


Noise = 0.2
Discount = 0.9
Living reward = 0

k=12

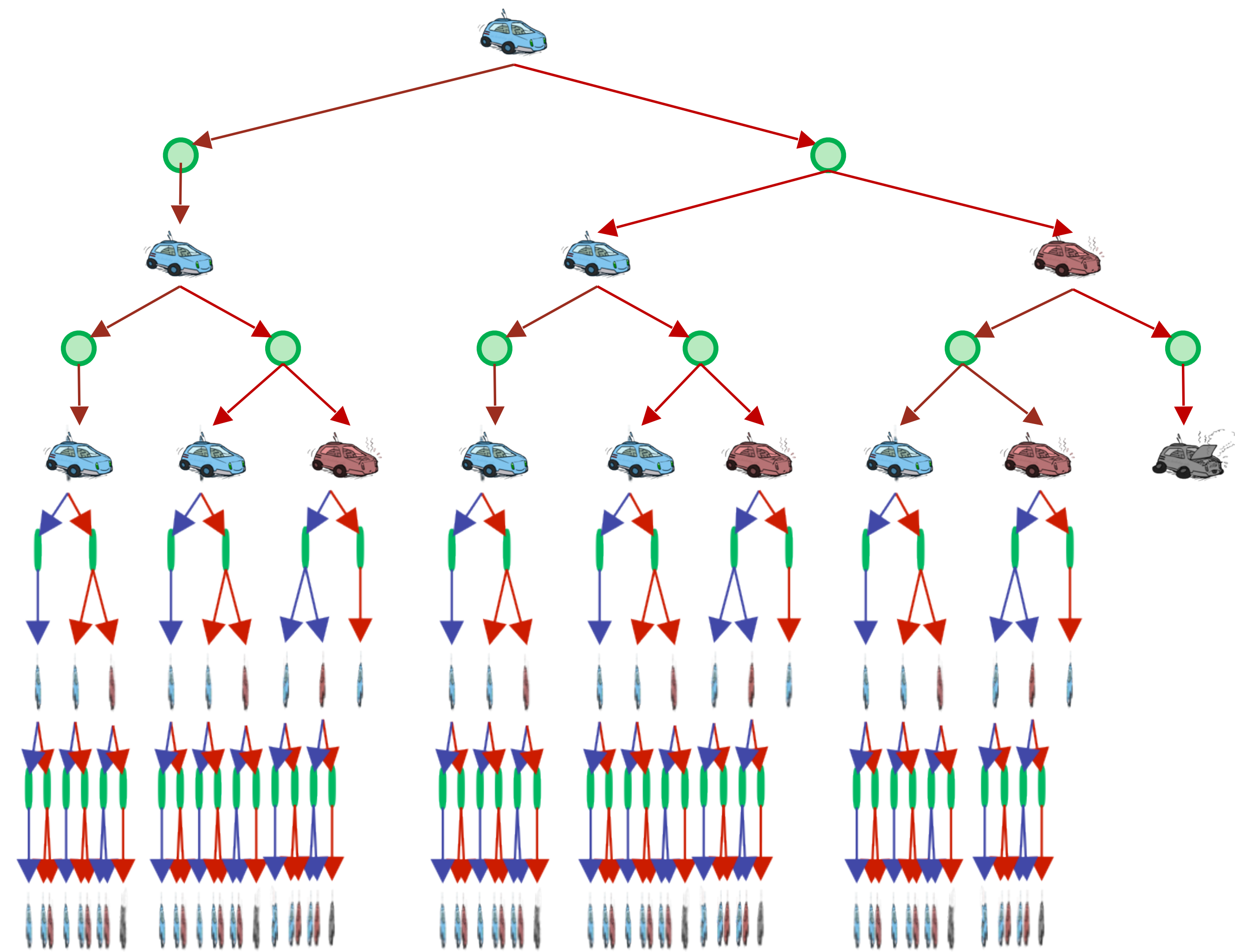
Noise = 0.2
Discount = 0.9
Living reward = 0

k=100

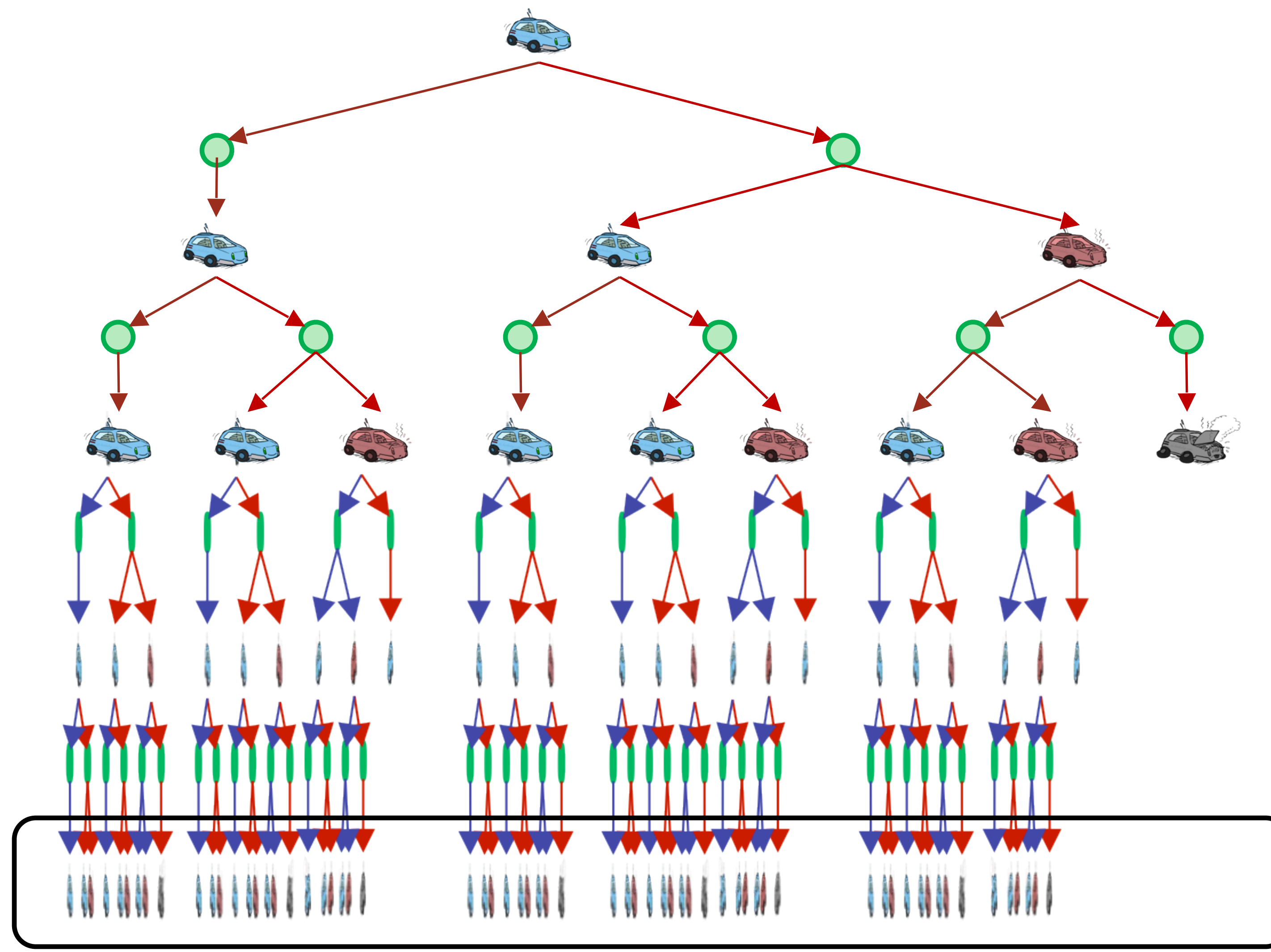


Noise = 0.2
Discount = 0.9
Living reward = 0

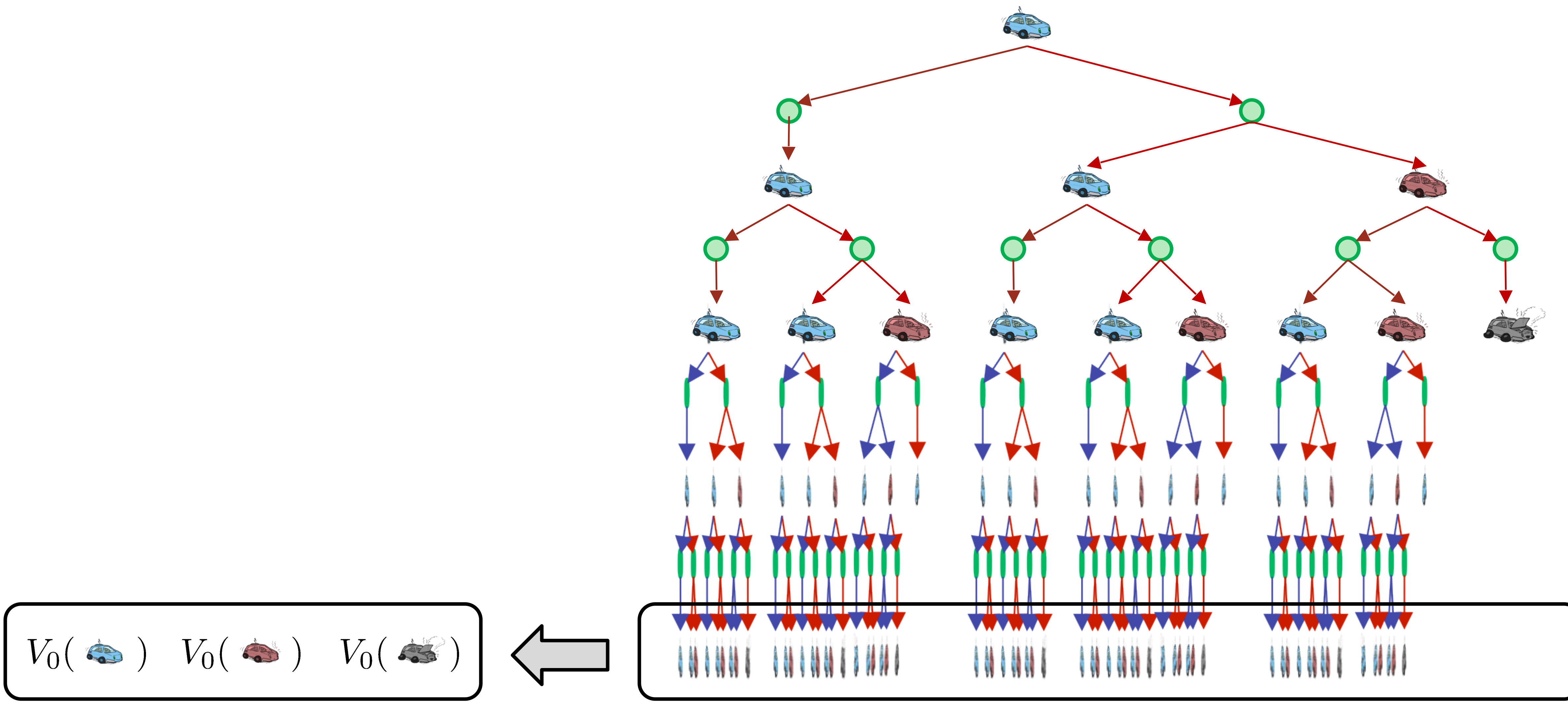
Computing Time-Limited Values



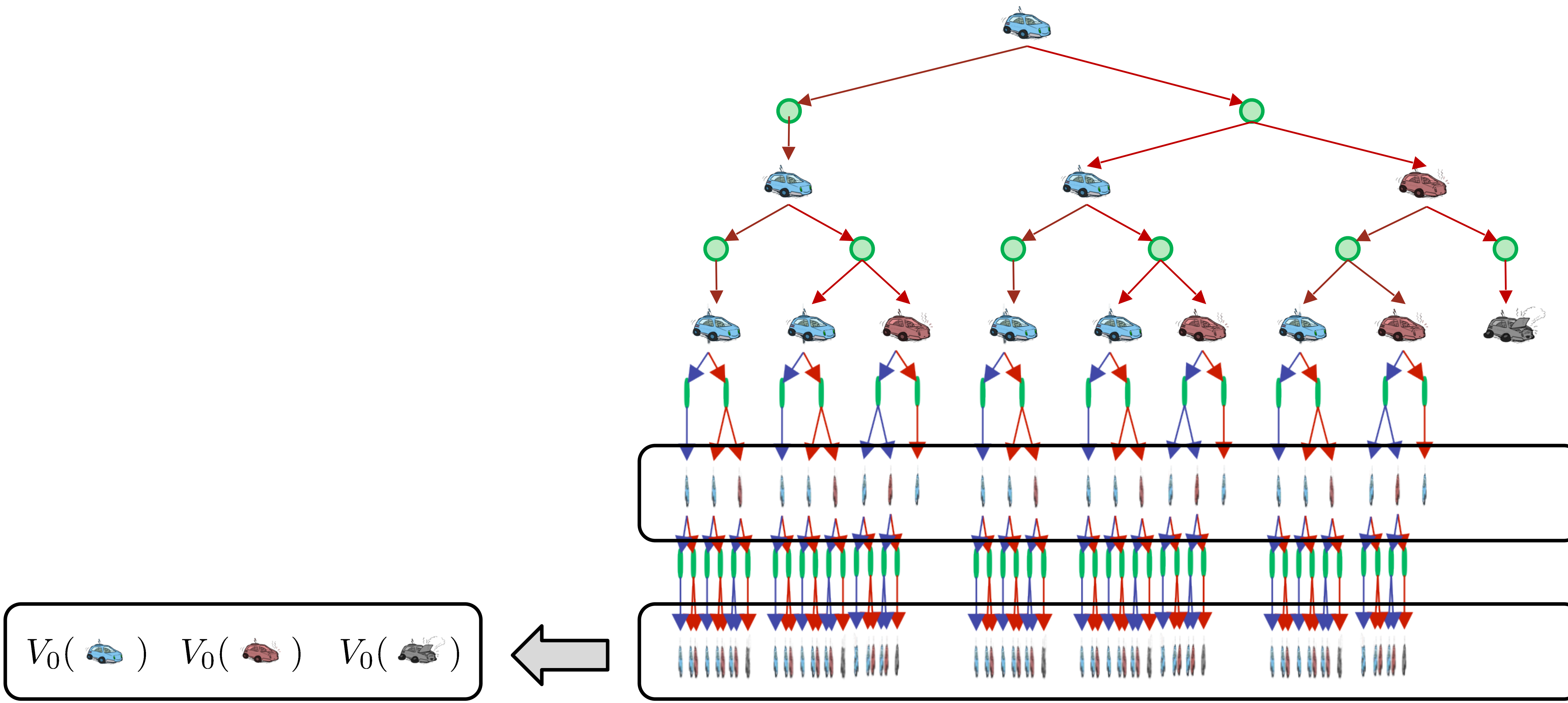
Computing Time-Limited Values



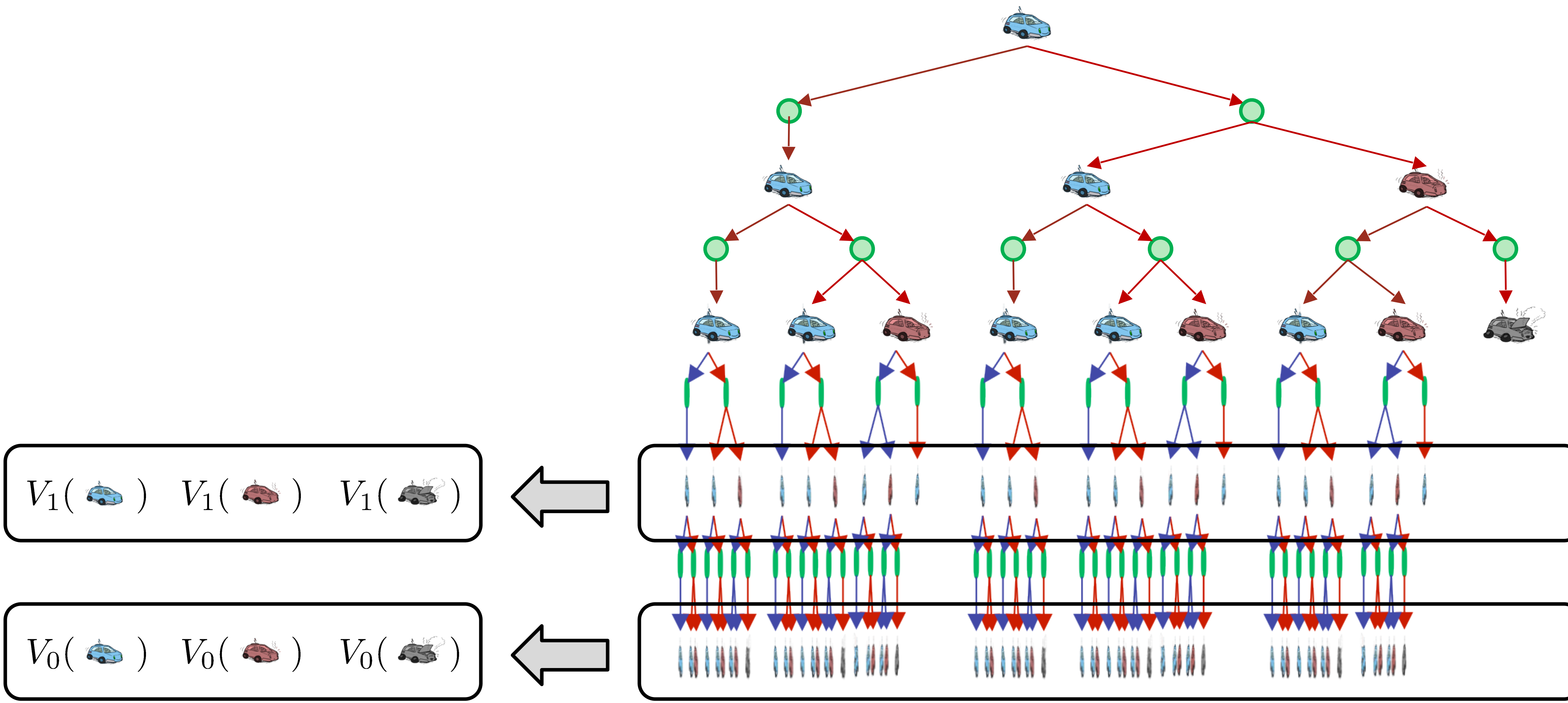
Computing Time-Limited Values



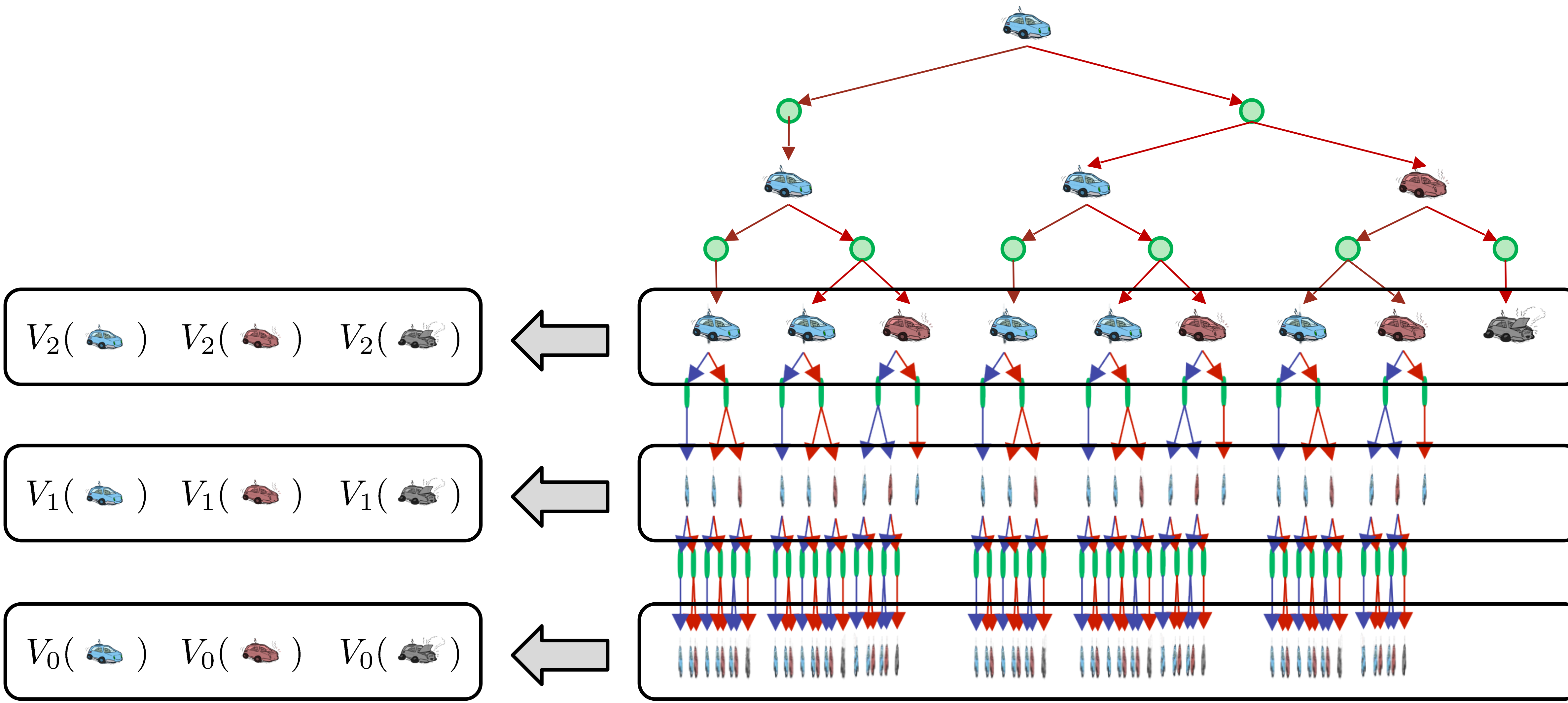
Computing Time-Limited Values



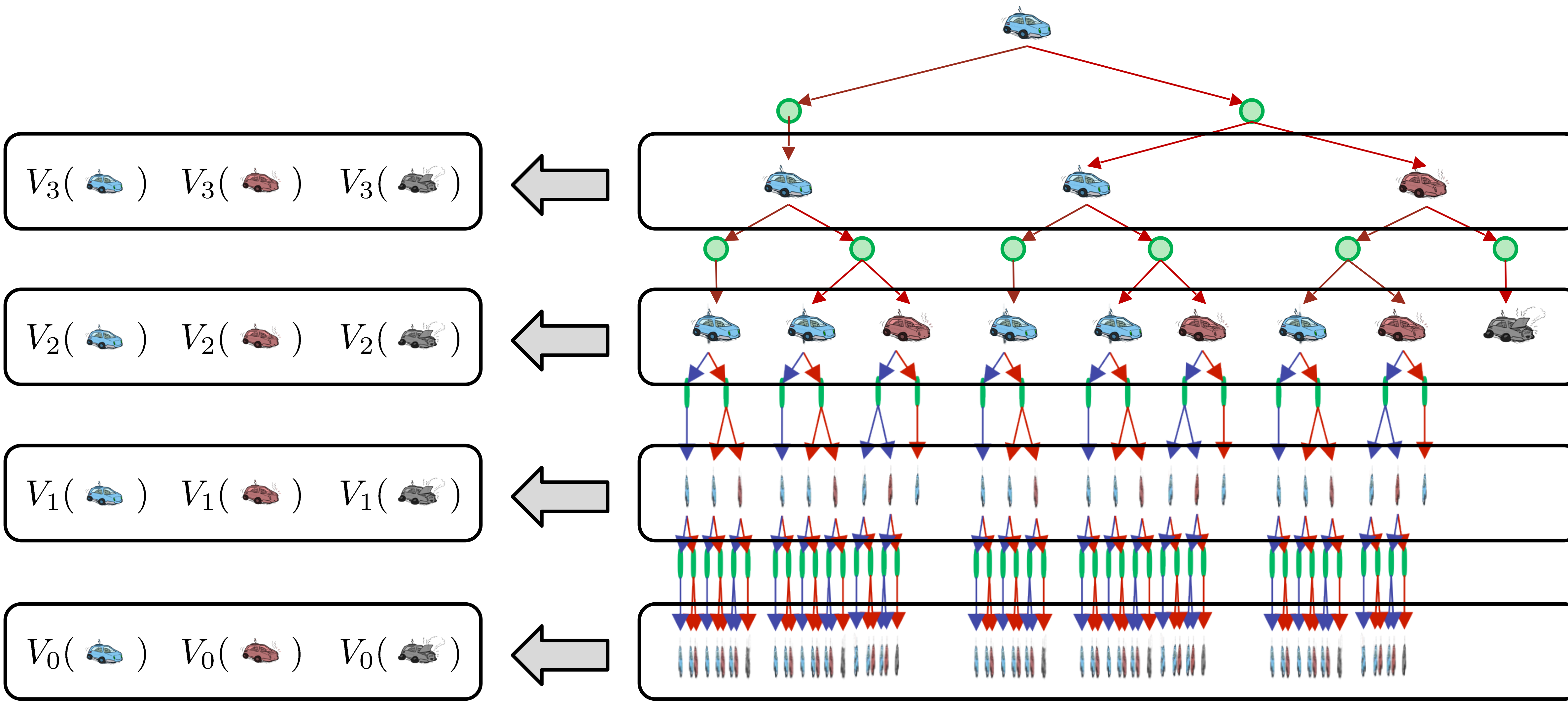
Computing Time-Limited Values



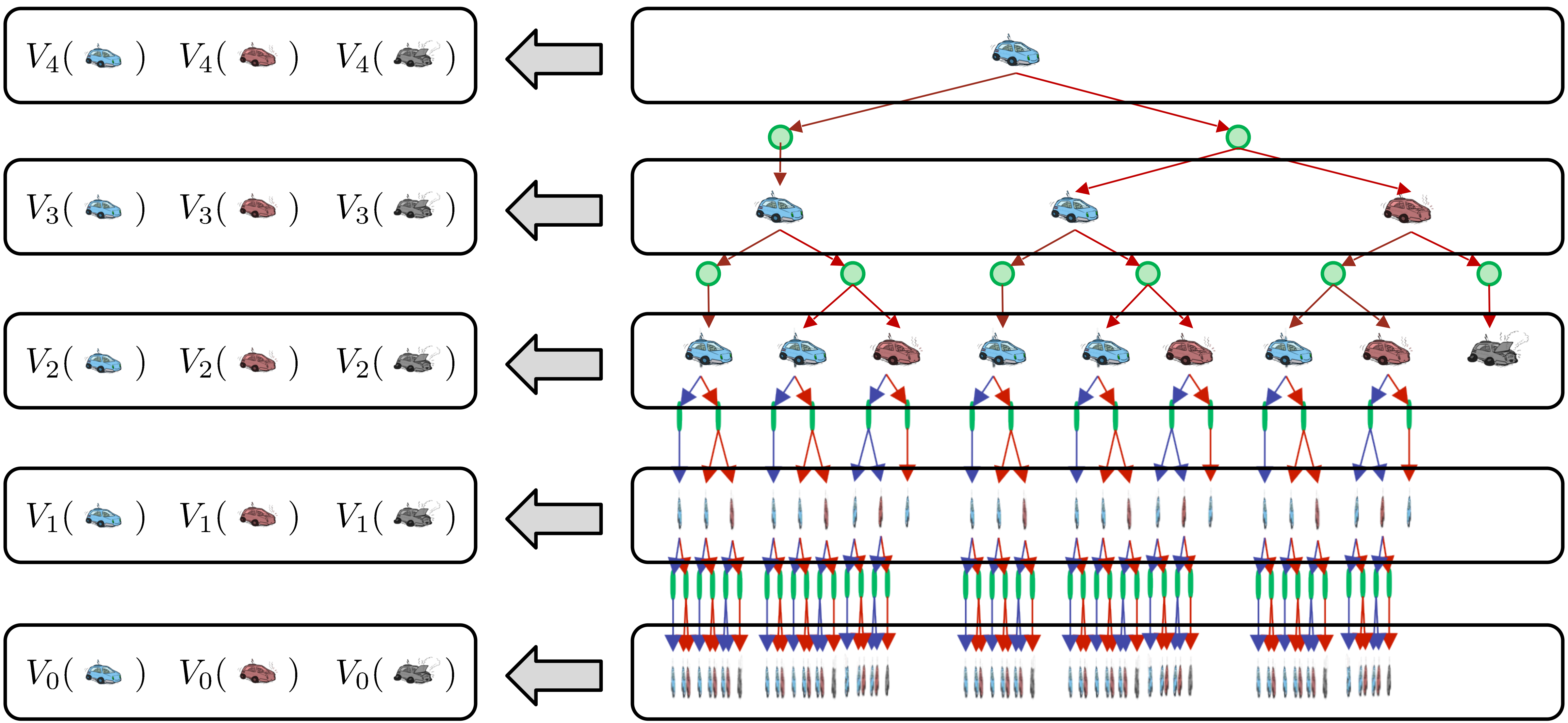
Computing Time-Limited Values



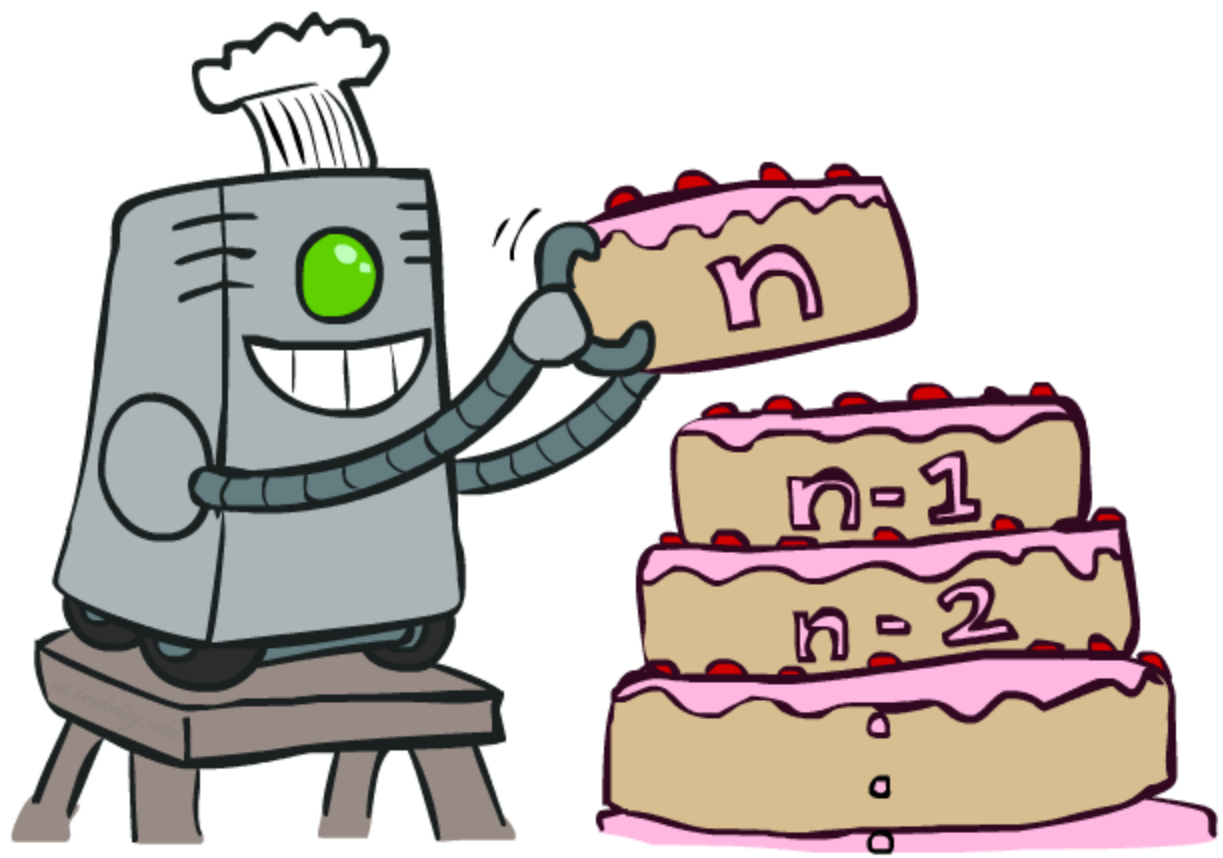
Computing Time-Limited Values



Computing Time-Limited Values



Value Iteration



Value Iteration

Value Iteration

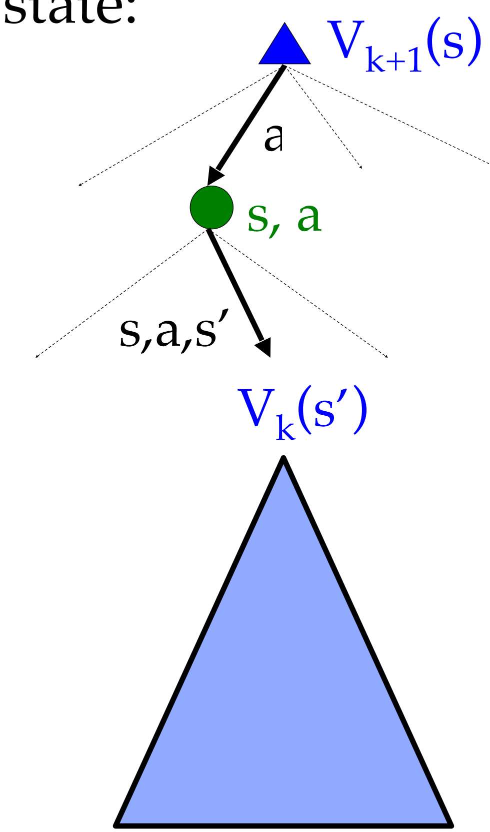
- Start with $V_0(s) = 0$: no time steps left means an expected reward sum of zero
- Given vector of $V_k(s)$ values, do one play of expectimax from each state:

Value Iteration

- Start with $V_0(s) = 0$: no time steps left means an expected reward sum of zero
- Given vector of $V_k(s)$ values, do one play of expectimax from each state:

$$V_{k+1}(s) \leftarrow \max_a \sum_{s'} T(s,a,s') \left[R(s,a,s') + \gamma V_k(s') \right]$$

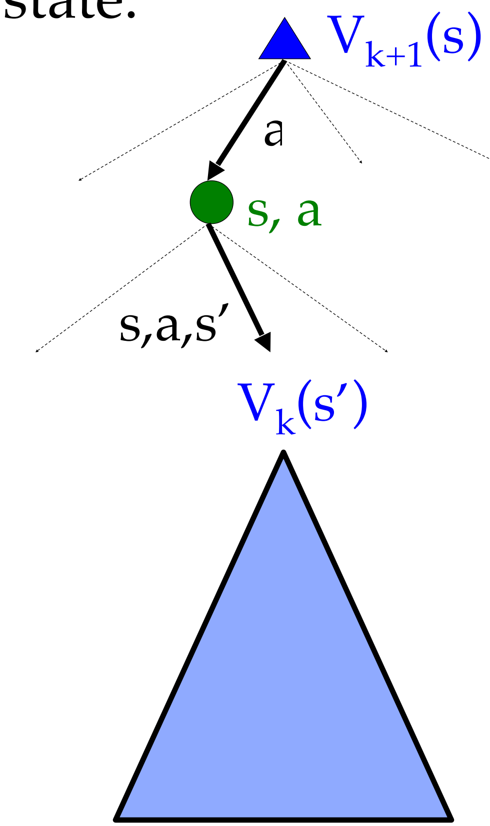
- Repeat until convergence, which yields V*
- Complexity of each iteration: $O(S^2A)$



Value Iteration

- Start with $V_0(s) = 0$: no time steps left means an expected reward sum of zero
- Given vector of $V_k(s)$ values, do one ply of expectimax from each state:

$$V_{k+1}(s) \leftarrow \max_a \sum_{s'} T(s,a,s') \left[R(s,a,s') + \gamma V_k(s') \right]$$



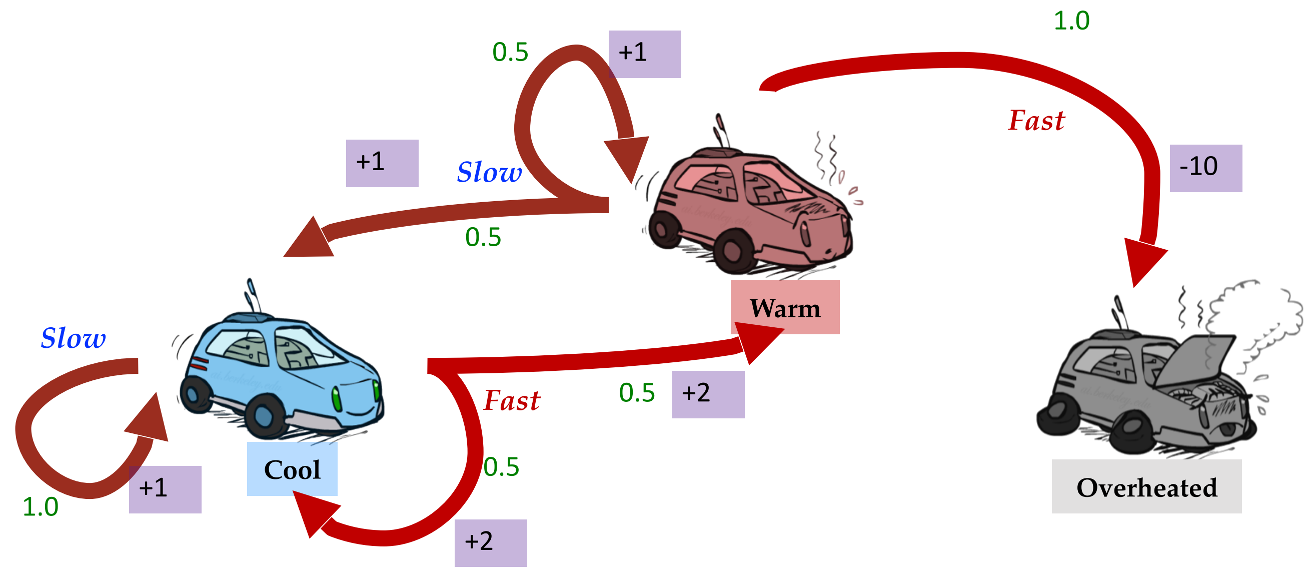
- Repeat until convergence, which yields V*
- Complexity of each iteration: $O(S^2A)$
- Theorem: will converge to unique optimal values
  - Basic idea: approximations get refined towards optimal values

Example: Value Iteration

V_2

V_1

V_0

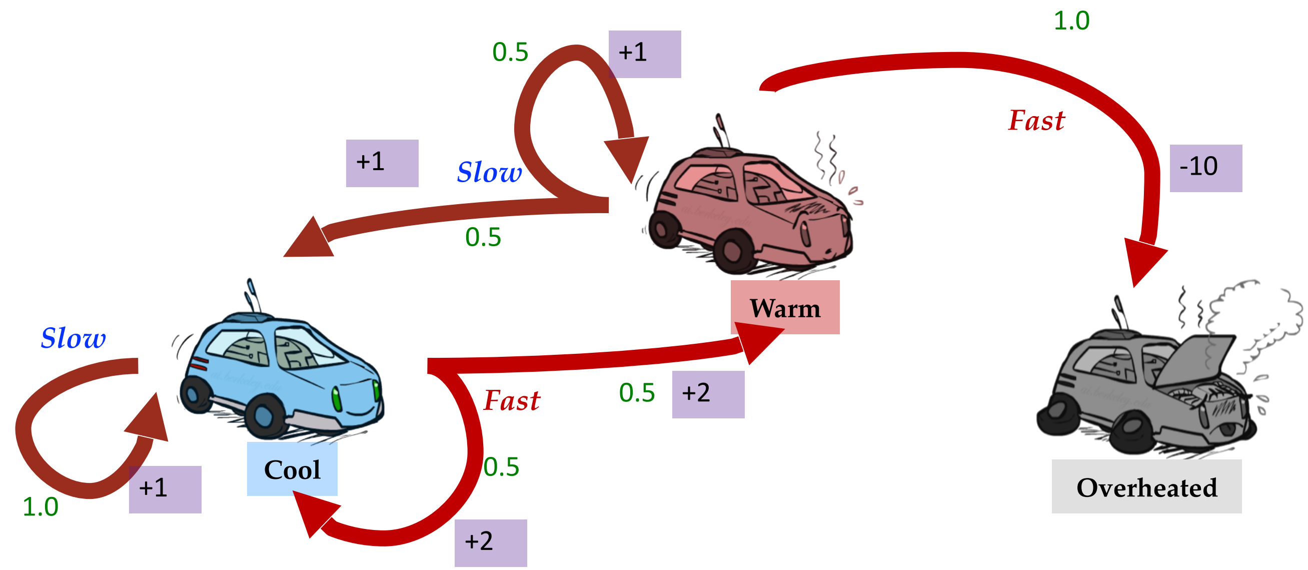


Assume no discount!

$$V_{k+1}(s) \leftarrow \max_a \sum_{s'} T(s,a,s') \left[R(s,a,s') + \gamma V_k(s') \right]$$

Example: Value Iteration

V_2			
V_1			
V_0	0	0	0



Assume no discount!

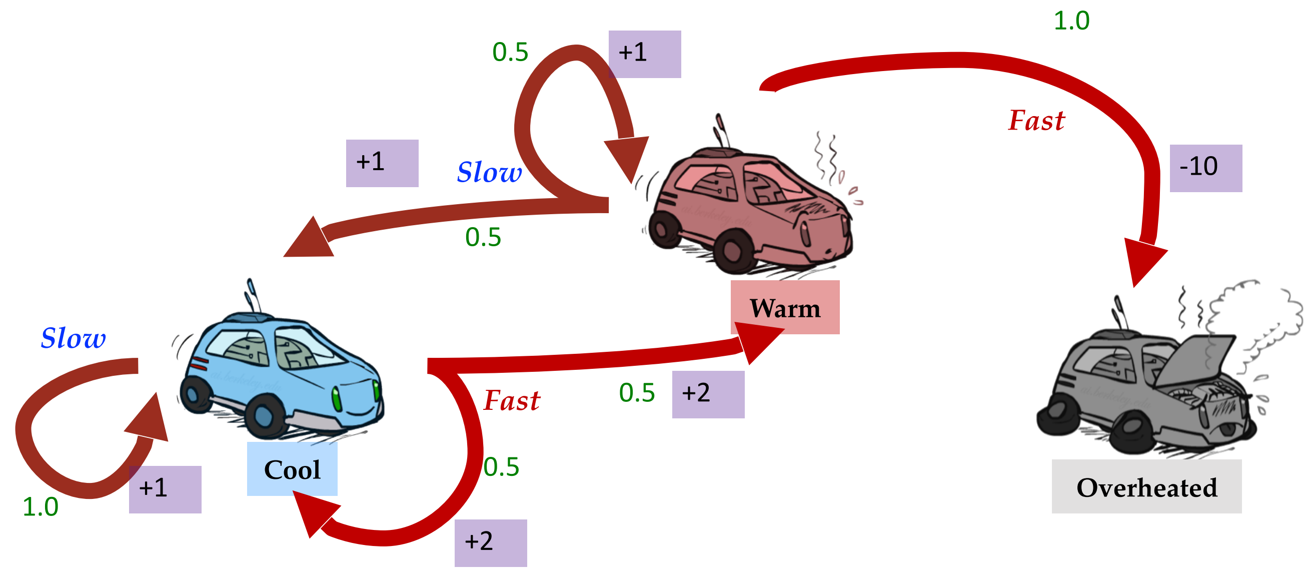
$$V_{k+1}(s) \leftarrow \max_a \sum_{s'} T(s,a,s') \left[R(s,a,s') + \gamma V_k(s')\right]$$

Example: Value Iteration

V_2

V_1 — S: 1

V_0 — 0 0 0



Assume no discount!

$$V_{k+1}(s) \leftarrow \max_a \sum_{s'} T(s,a,s') \left[R(s,a,s') + \gamma V_k(s')\right]$$

Example: Value Iteration

V_2

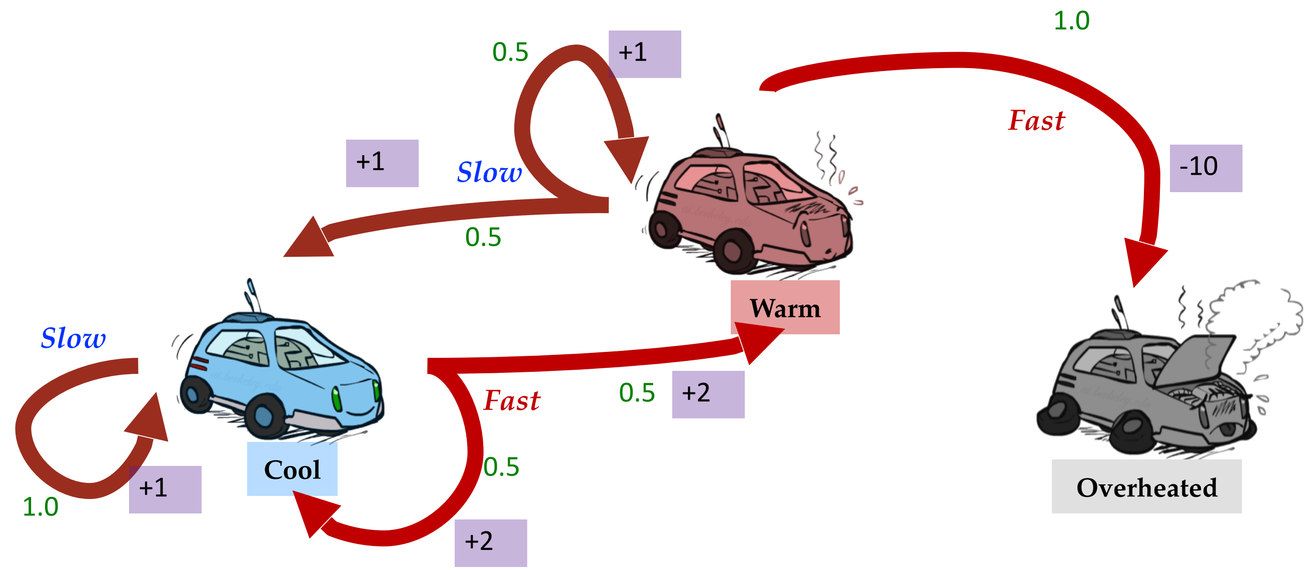
V_1

S: 1

F: .5*2+.5*2=2

V_0

0 0 0

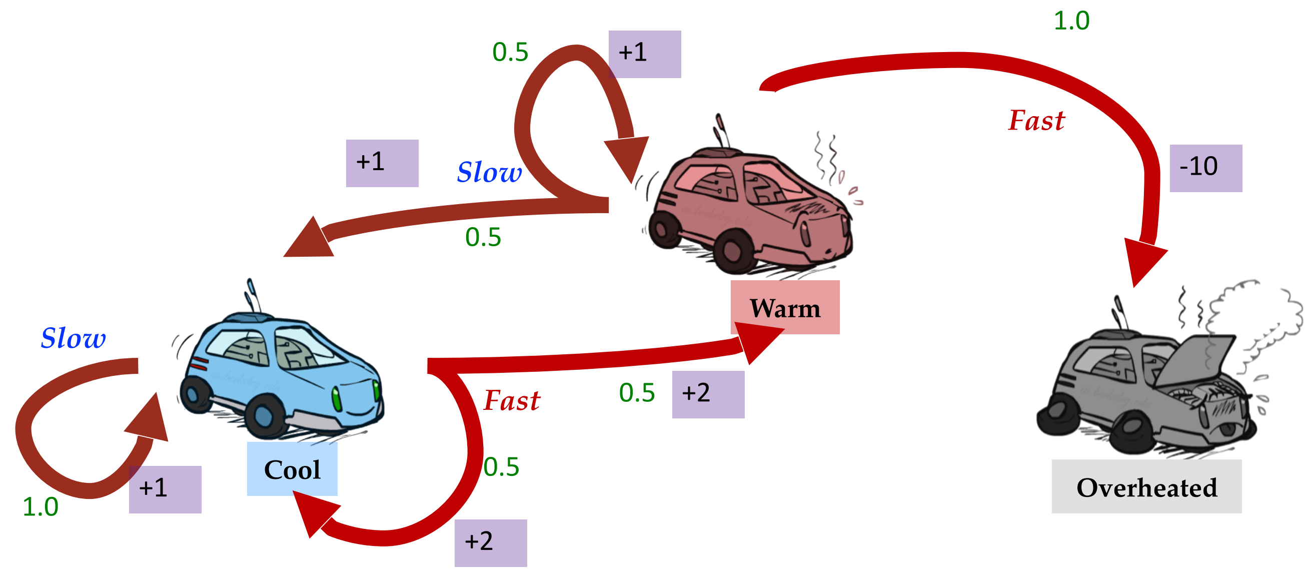


Assume no discount!

$$V_{k+1}(s) \leftarrow \max_a \sum_{s'} T(s,a,s')\left[R(s,a,s') + \gamma V_k(s')\right]$$

Example: Value Iteration

V_2			
V_1	2		
V_0	0	0	0

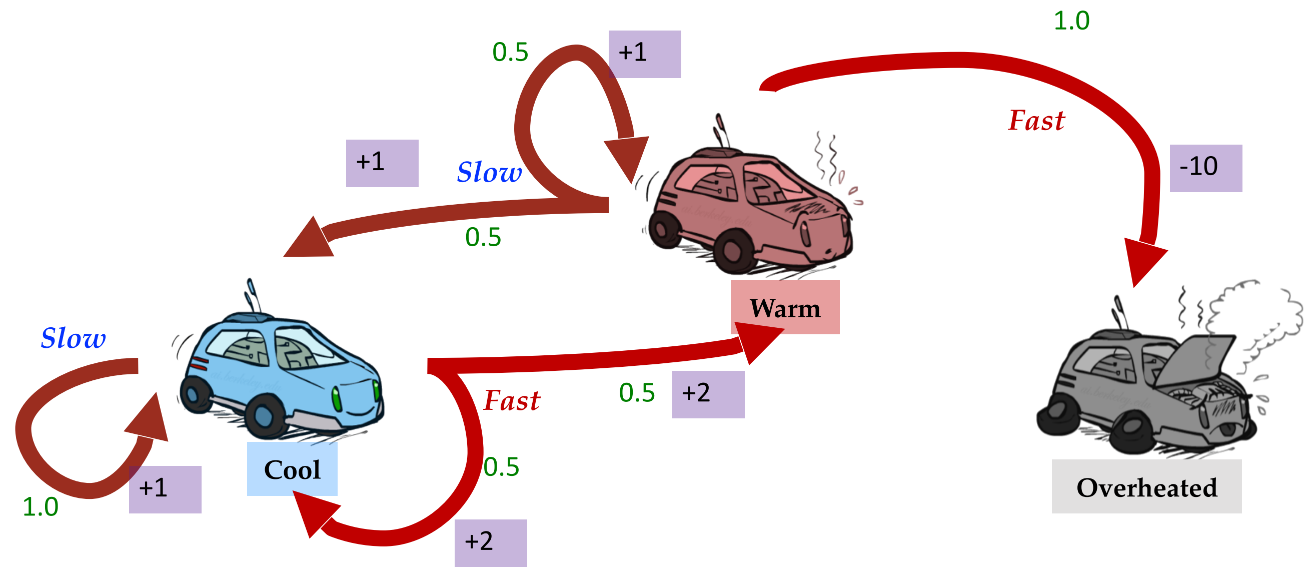


Assume no discount!

$$V_{k+1}(s) \leftarrow \max_a \sum_{s'} T(s,a,s') \left[R(s,a,s') + \gamma V_k(s')\right]$$

Example: Value Iteration

V_2	
V_1	2 S: .5*1+.5*1=1
V_0	0 0 0



Assume no discount!

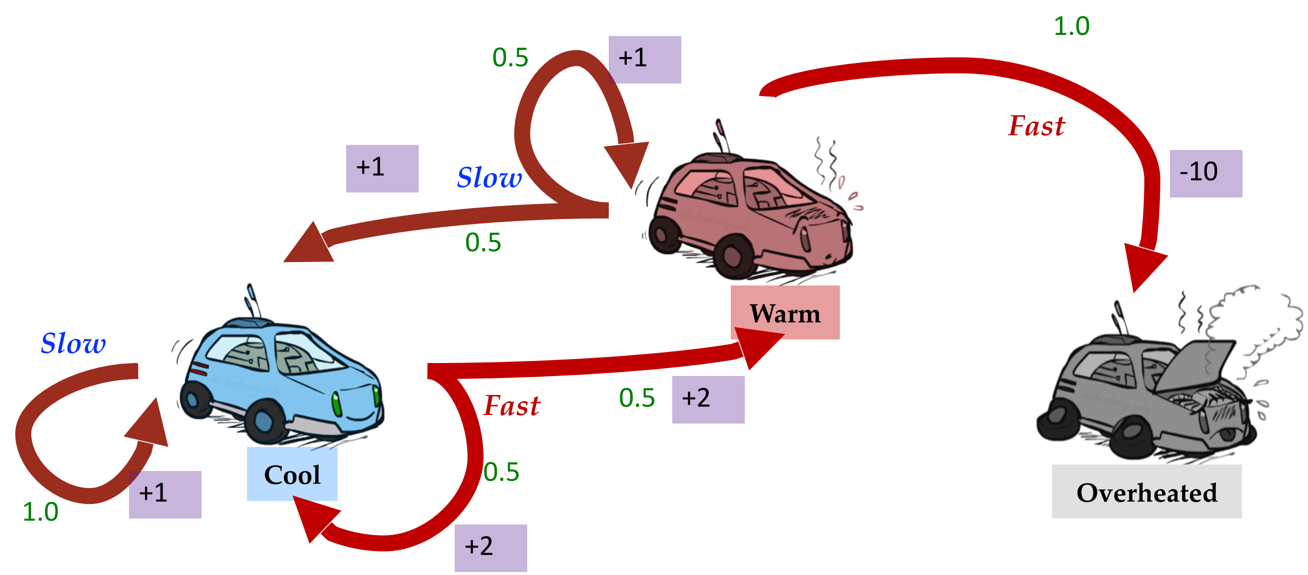
$$V_{k+1}(s) \leftarrow \max_a \sum_{s'} T(s,a,s') \left[R(s,a,s') + \gamma V_k(s') \right]$$

Example: Value Iteration

V_2

V_1 2 — S: .5*1+.5*1=1, F: -10

V_0 0 0 0

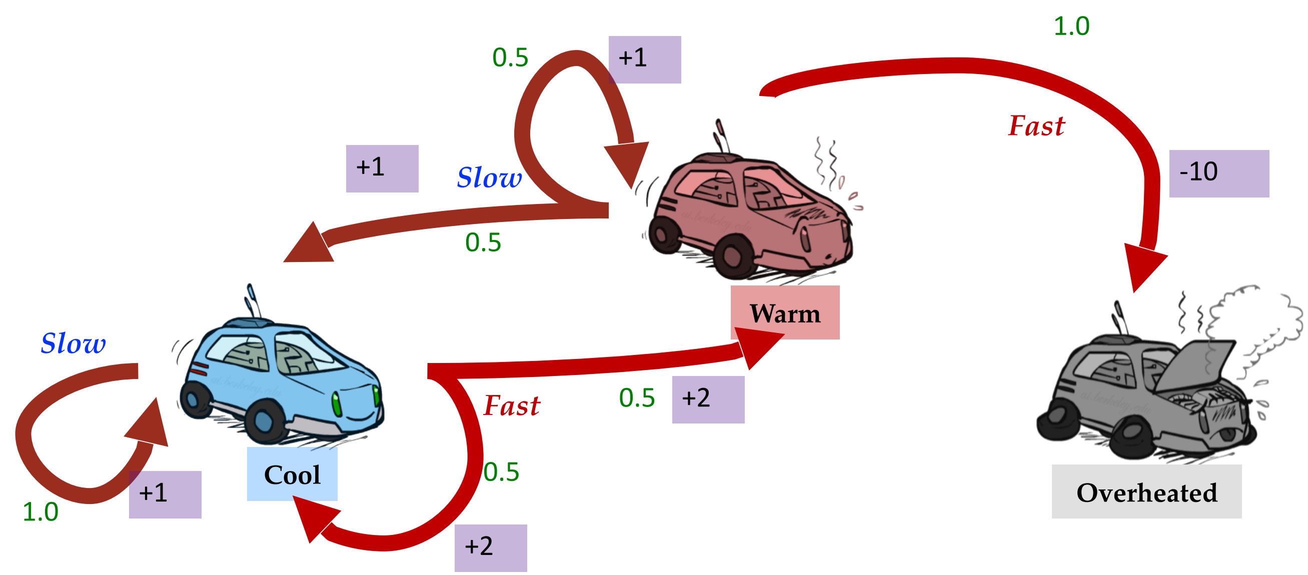


Assume no discount!

$$V_{k+1}(s) \leftarrow \max_a \sum_{s'} T(s,a,s') \left[R(s,a,s') + \gamma V_k(s')\right]$$

Example: Value Iteration

V_2			
V_1	2	1	
V_0	0	0	0

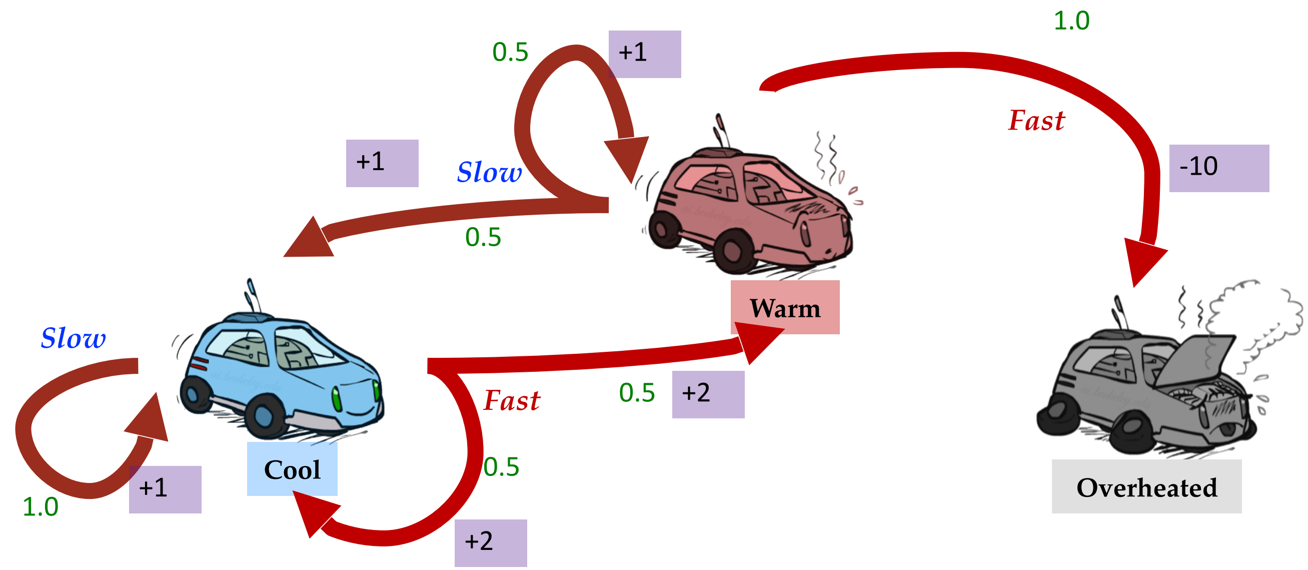


Assume no discount!

$$V_{k+1}(s) \leftarrow \max_a \sum_{s'} T(s,a,s') \left[R(s,a,s') + \gamma V_k(s')\right]$$

Example: Value Iteration

V_2			
V_1	2	1	0
V_0	0	0	0

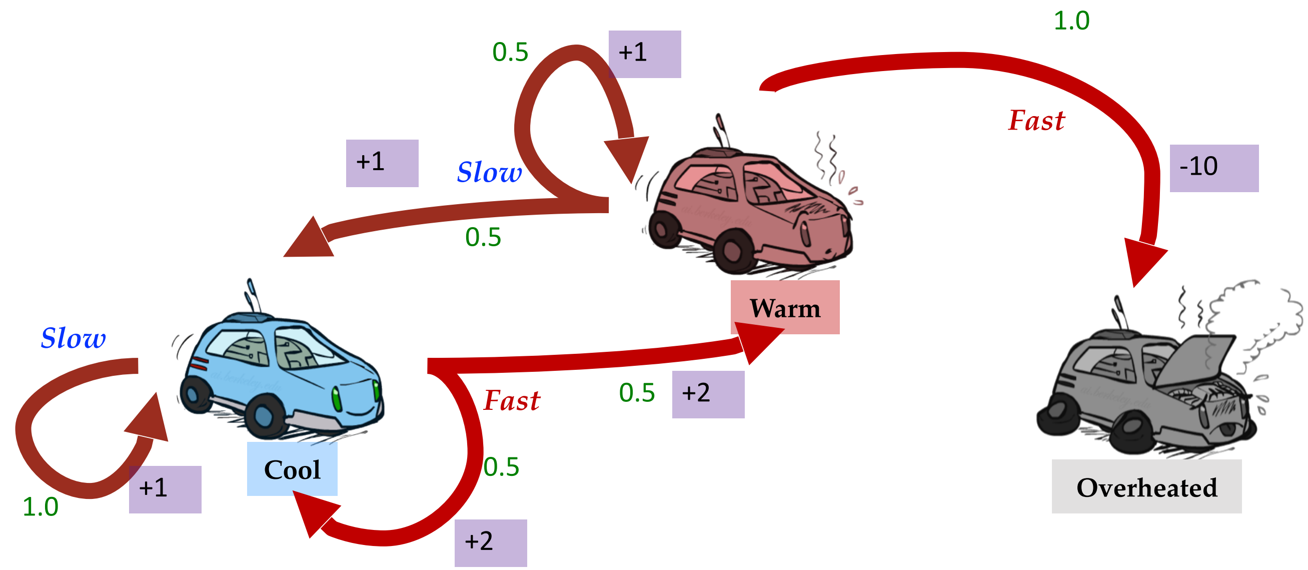


Assume no discount!

$$V_{k+1}(s) \leftarrow \max_a \sum_{s'} T(s,a,s') \left[R(s,a,s') + \gamma V_k(s') \right]$$

Example: Value Iteration

V_2	S: 1+2=3		
V_1	2	1	0
V_0	0	0	0

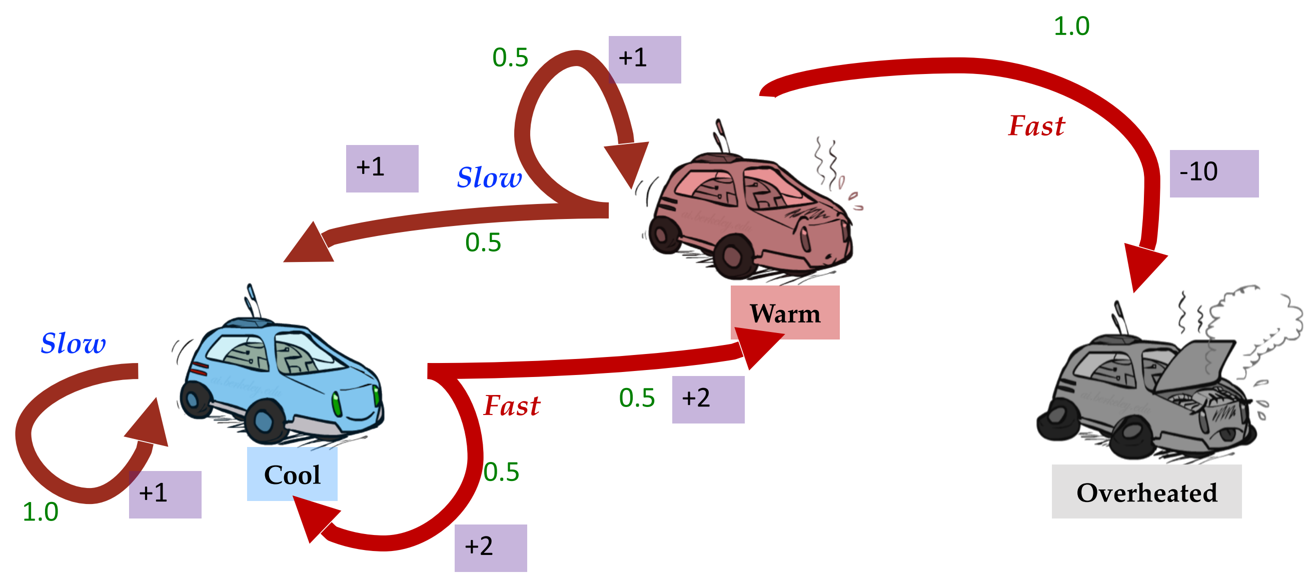


Assume no discount!

$$V_{k+1}(s) \leftarrow \max_a \sum_{s'} T(s,a,s') \left[R(s,a,s') + \gamma V_k(s')\right]$$

Example: Value Iteration

V_2	S: 1+2=3 F:.5*(2+2)+.5*(2+1)=3.5		
V_1	2	1	0
V_0	0	0	0

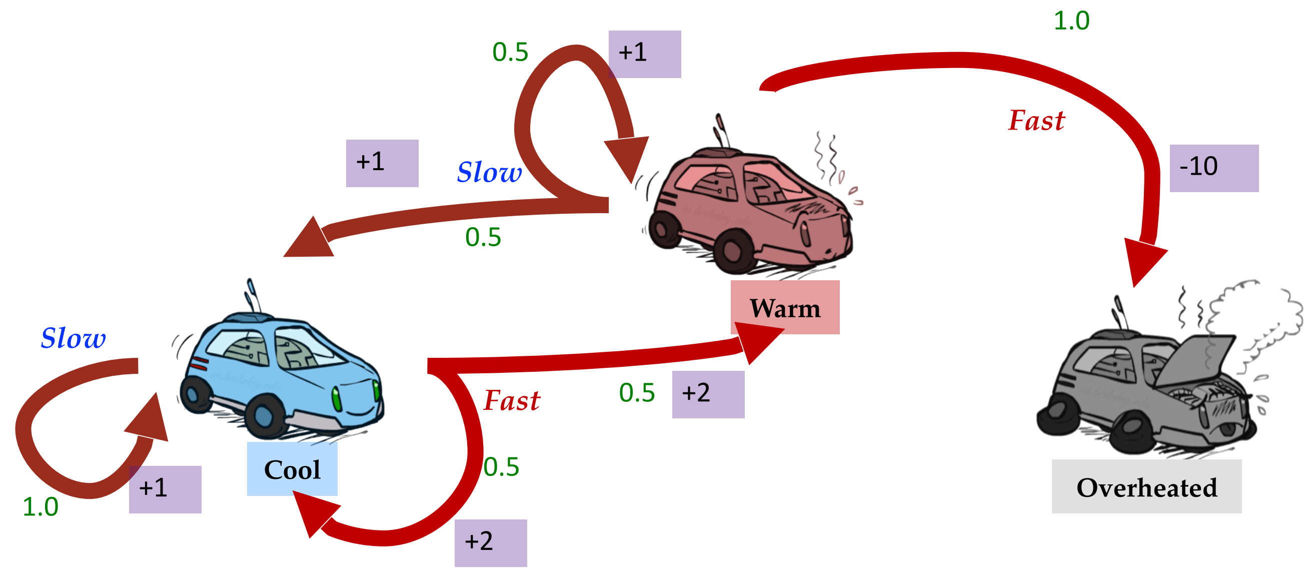


Assume no discount!

$$V_{k+1}(s) \leftarrow \max_a \sum_{s'} T(s,a,s') \left[R(s,a,s') + \gamma V_k(s') \right]$$

Example: Value Iteration

V_2	3.5	2.5	0
V_1	2	1	0
V_0	0	0	0

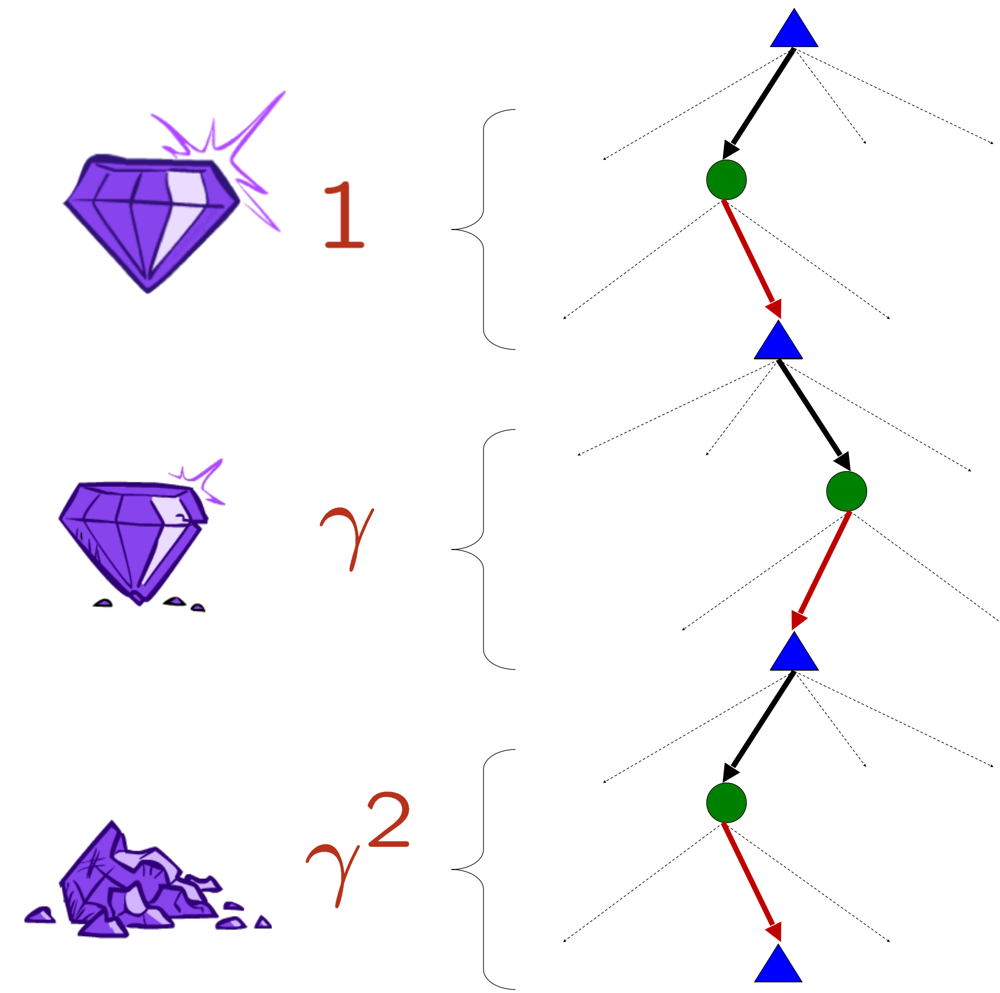


Assume no discount!

$$V_{k+1}(s) \leftarrow \max_a \sum_{s'} T(s,a,s') \left[R(s,a,s') + \gamma V_k(s')\right]$$

Discounting

- How to discount?
  - Each time we descend a level, we multiply in the discount once
- Why discount?
  - Reward now is better than later
  - Can also think of it as a $1 - \gamma$ chance of ending the process at every step
  - Also helps our algorithms converge

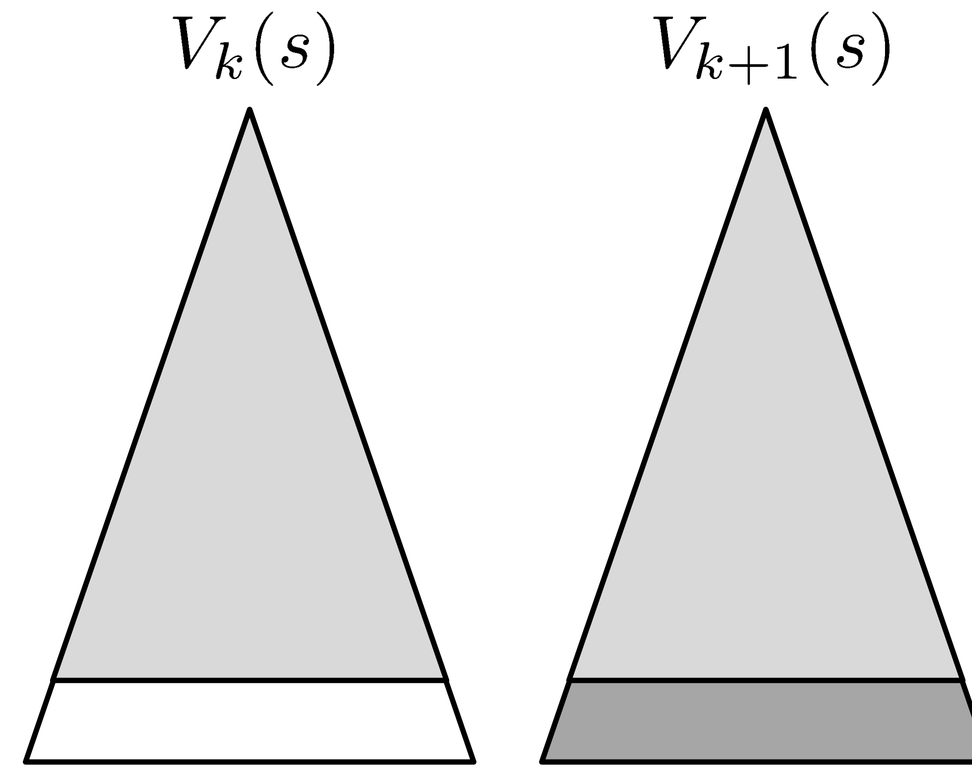


Convergence*

- How do we know the V_k vectors are going to converge? (assuming $0 < \gamma < 1$)

Convergence*

- How do we know the V_k vectors are going to converge? (assuming $0 < \gamma < 1$)
- Proof Sketch:
  - For any state V_k and V_{k+1} can be viewed as depth k+1 expectimax results in nearly identical search trees
  - The difference is that on the bottom layer, V_{k+1} has actual rewards while V_k has zeros
  - That last layer is at best all R_{MAX}
  - It is at worst R_{MIN}
  - But everything is discounted by γ^k that far out
  - So V_k and V_{k+1} are at most $\gamma^k \max|R|$ different
  - So as k increases, the values converge



Next Lecture: Policy-Based Methods

Policies may converge long before values do