R L & A I

Reinforcement Learning and Artificial Intelligence

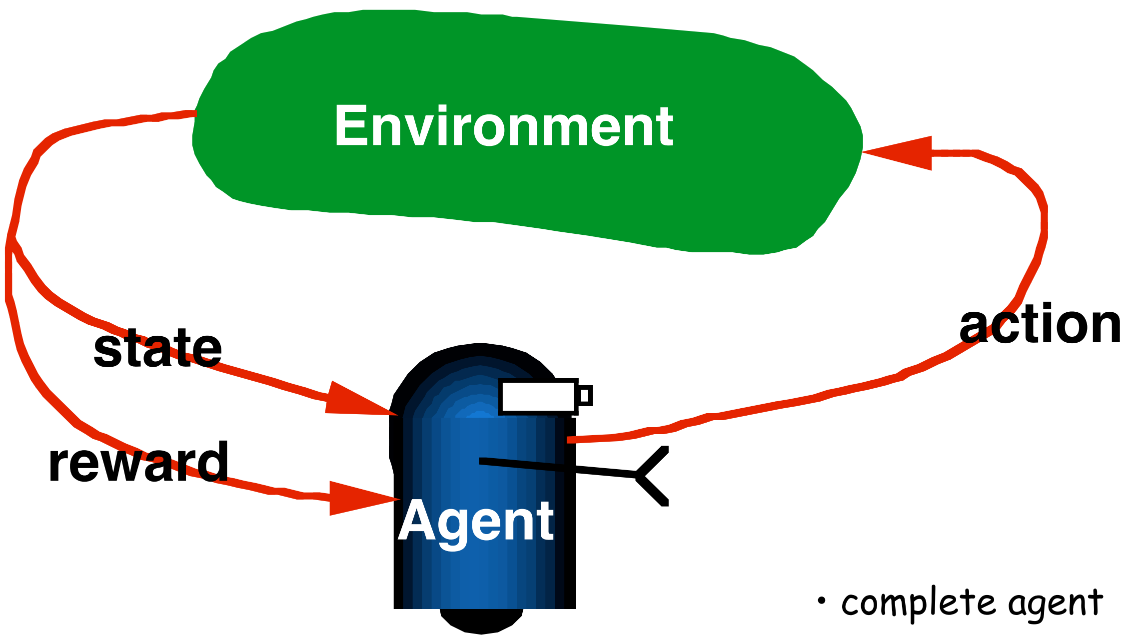
PIs:
Rich Sutton
Michael Bowling
Dale Schuurmans
Vadim Bulitko

University of Alberta
Quaecumque Vera

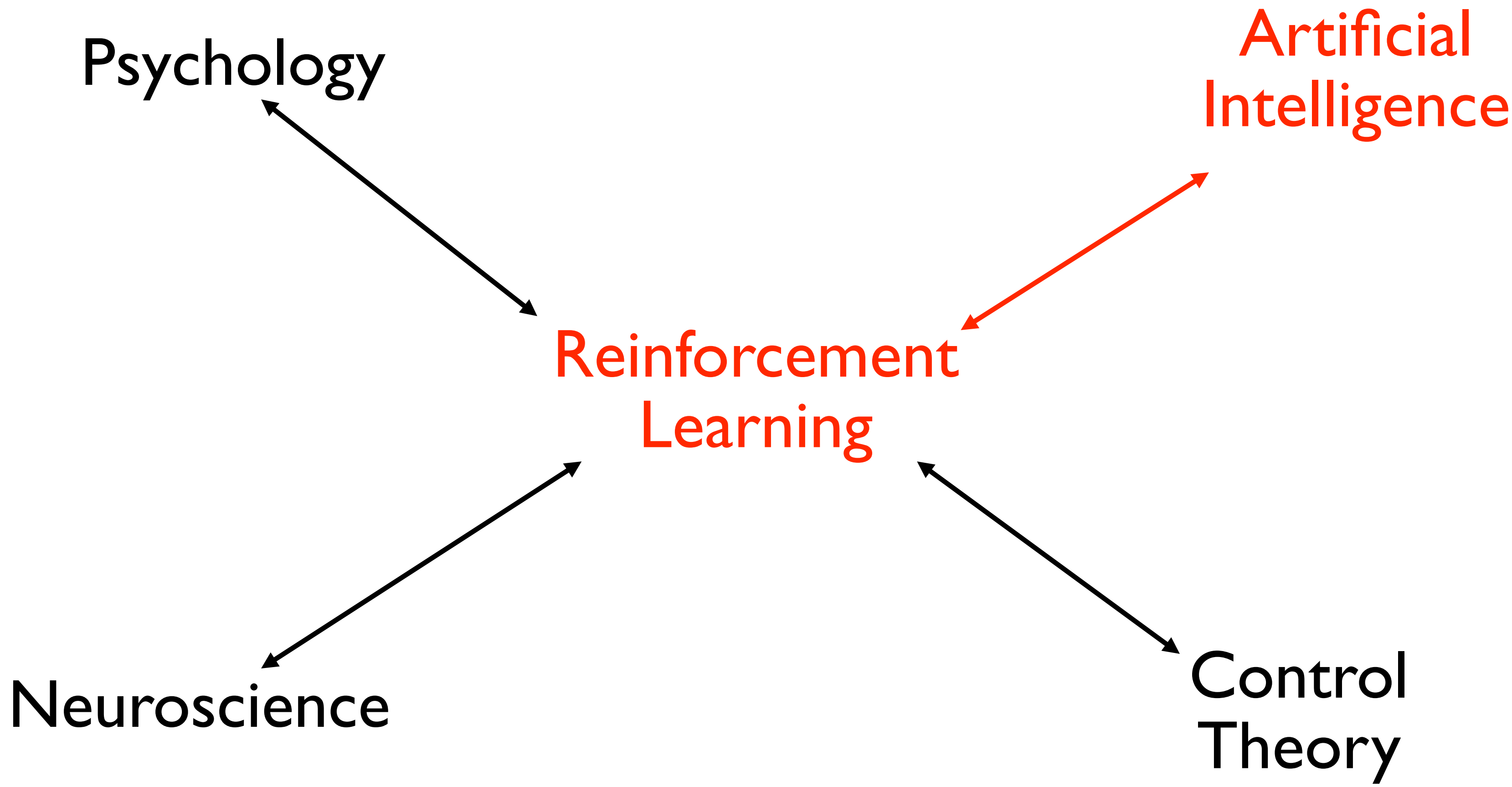
Alberta Ingenuity Centre for Machine Learning

Informatics iCORE Circle of Research Excellence

Reinforcement learning is learning from interaction to achieve a goal



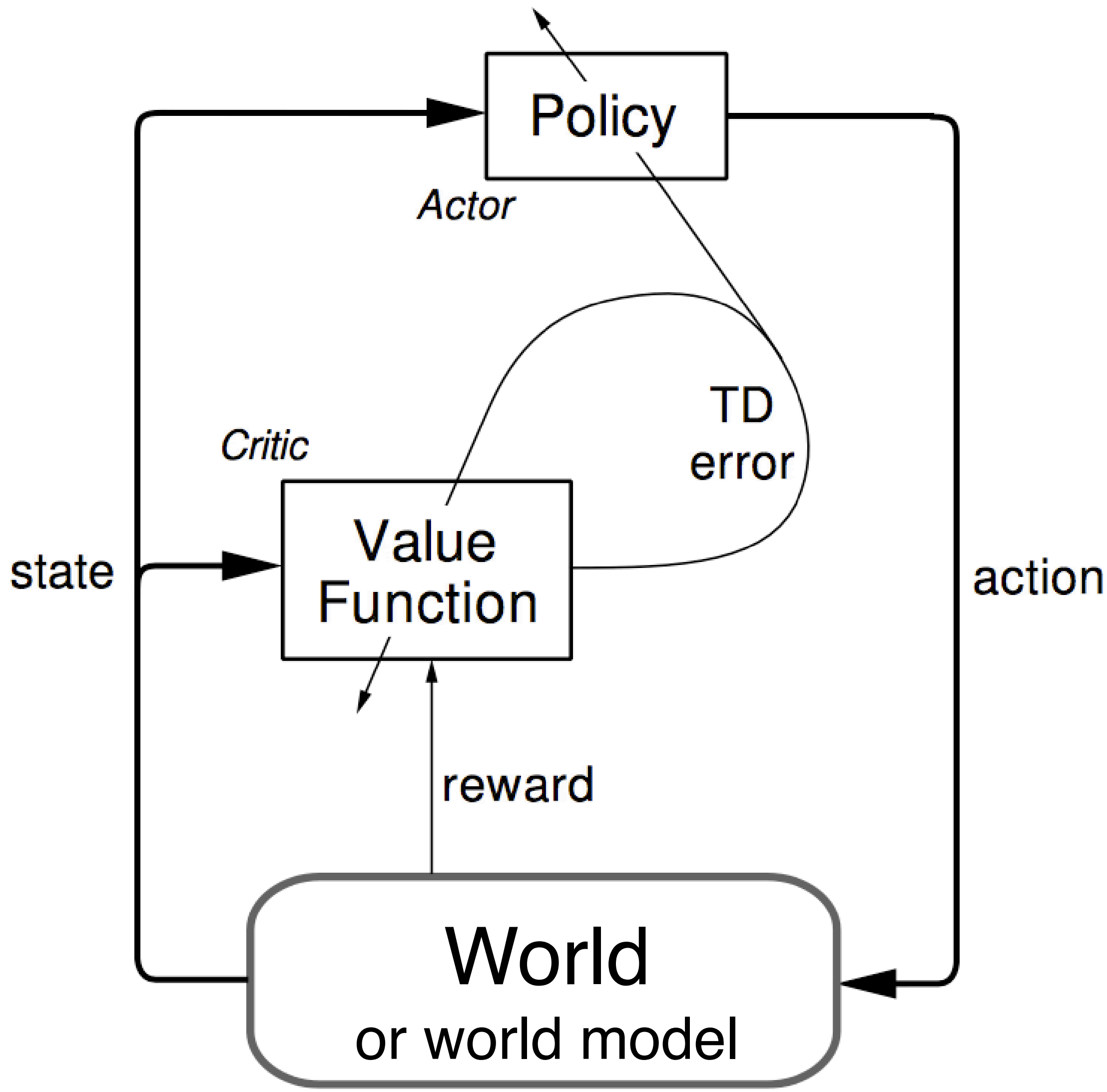
- complete agent
- temporally situated
- continual learning & planning
- object is to affect environment
- environment stochastic & uncertain


Psychology
Artificial Intelligence
Reinforcement Learning
Neuroscience
Control Theory

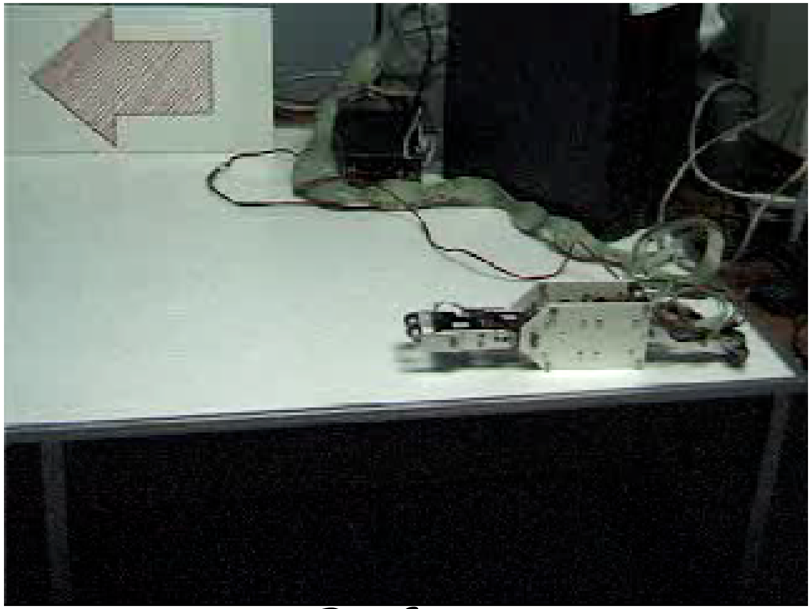
Intro to RL

with a robot bias

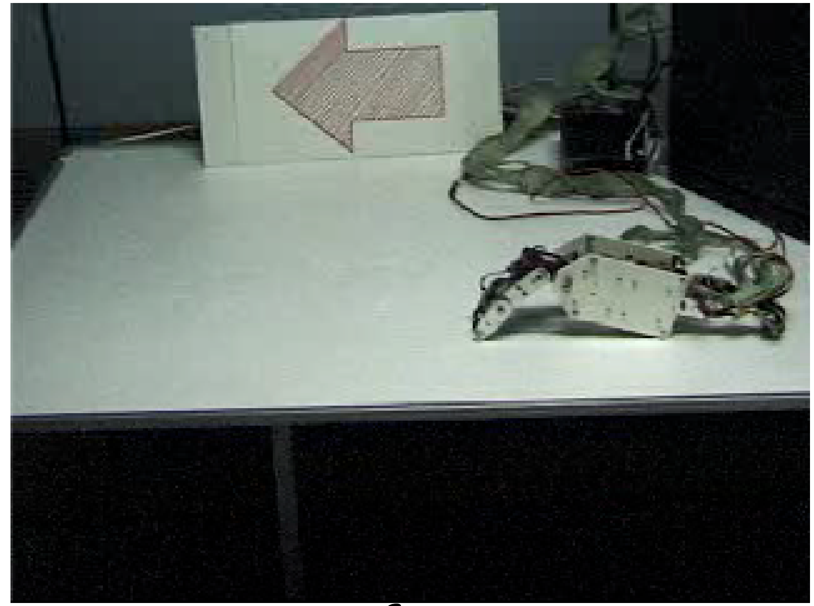
- policy
- reward
- value
- TD error
- world model
- generalized policy iteration


Policy
Actor
Critic
TD
error
Value
Function
state
action
reward
World
or world model

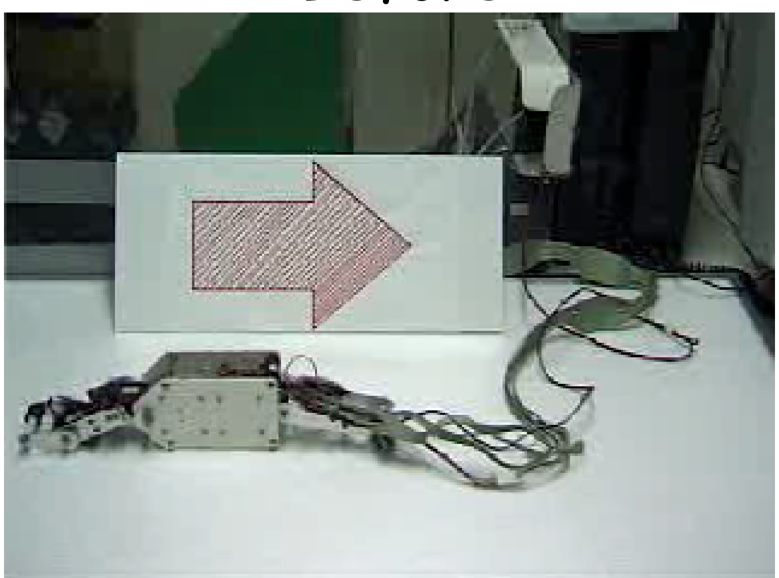
Hajime Kimura's RL Robots



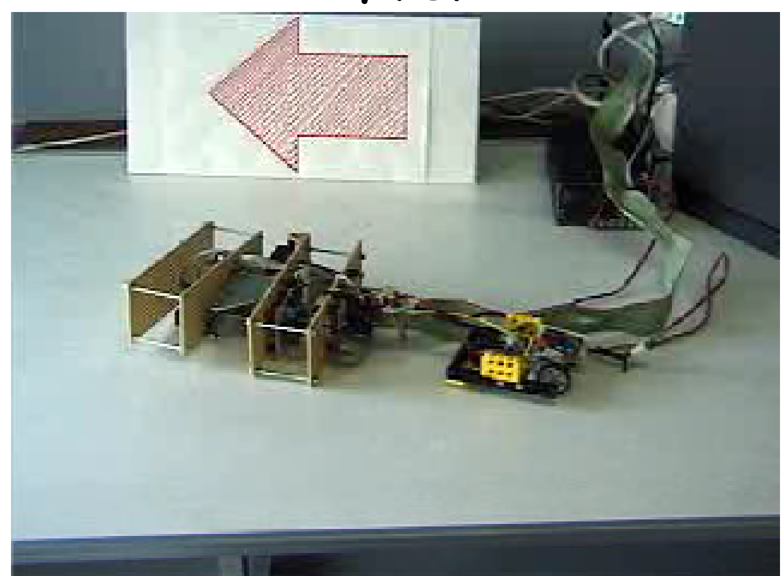
Before



After

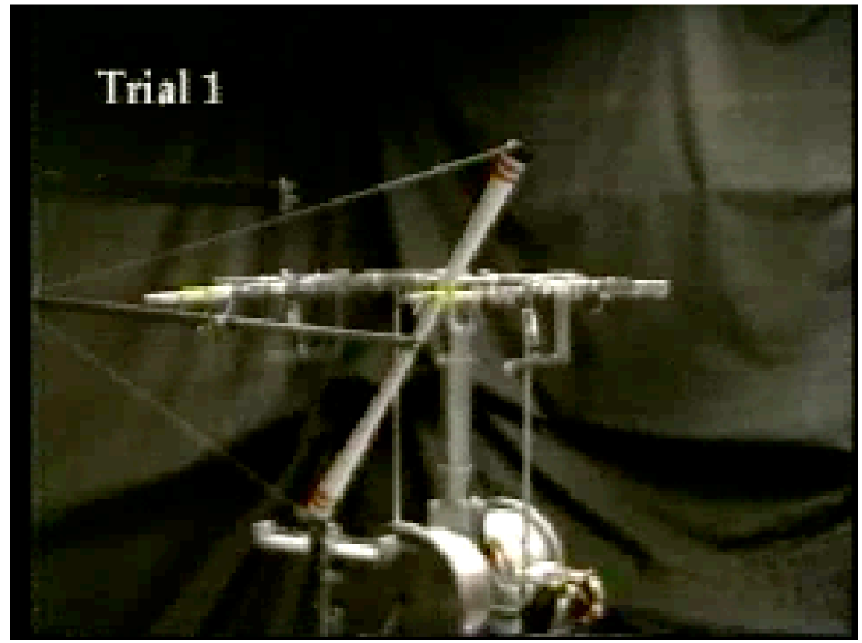


Backward



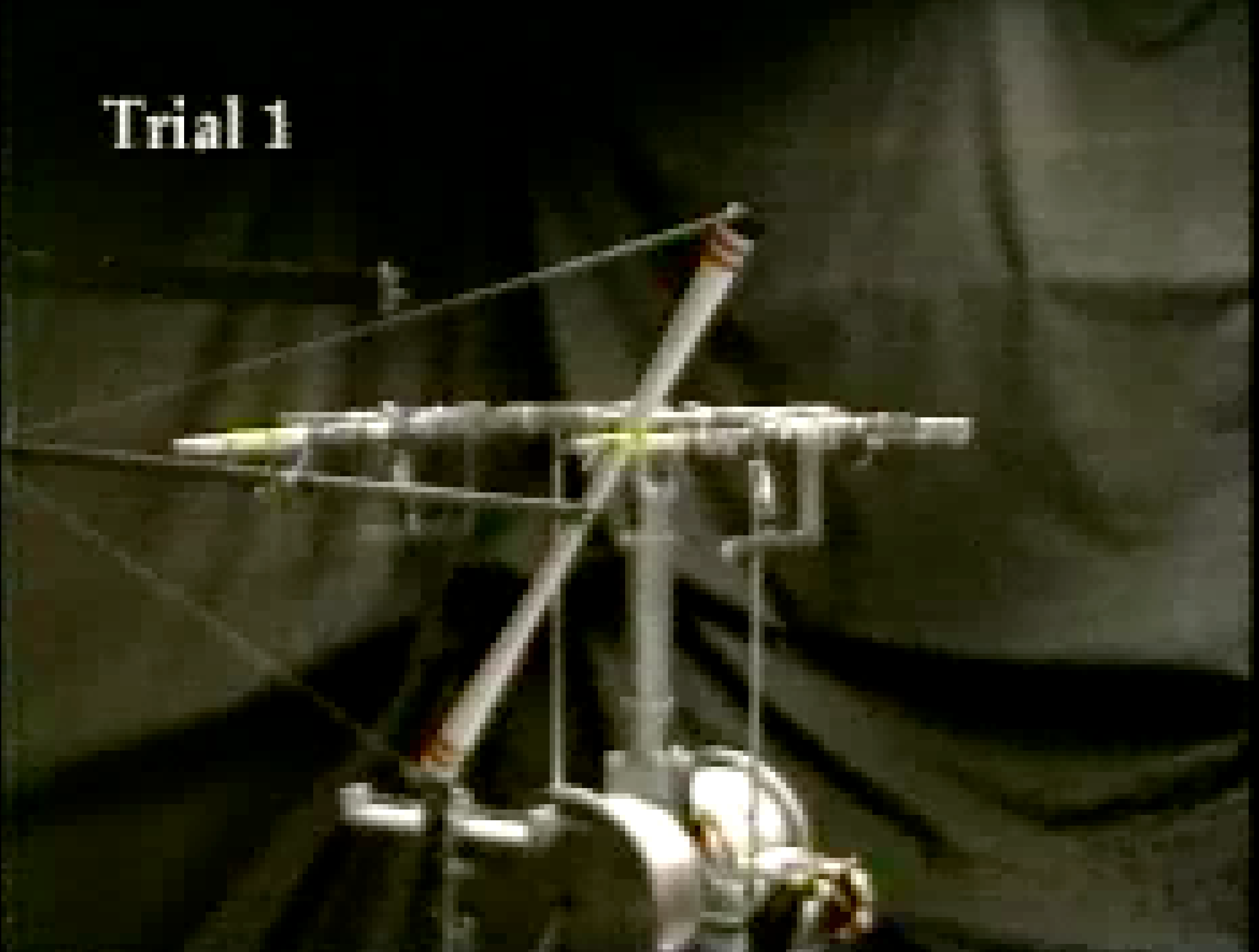
New Robot, Same algorithm

Devilsticking



Finnegan Southey
University of Alberta

Stefan Schaal & Chris Atkeson
Univ. of Southern California
“Model-based Reinforcement Learning of Devilsticking”


Trial 1

The RoboCup Soccer Competition

Autonomous Learning of Efficient Gait

Kohl & Stone (UTexas) 2004

Policies

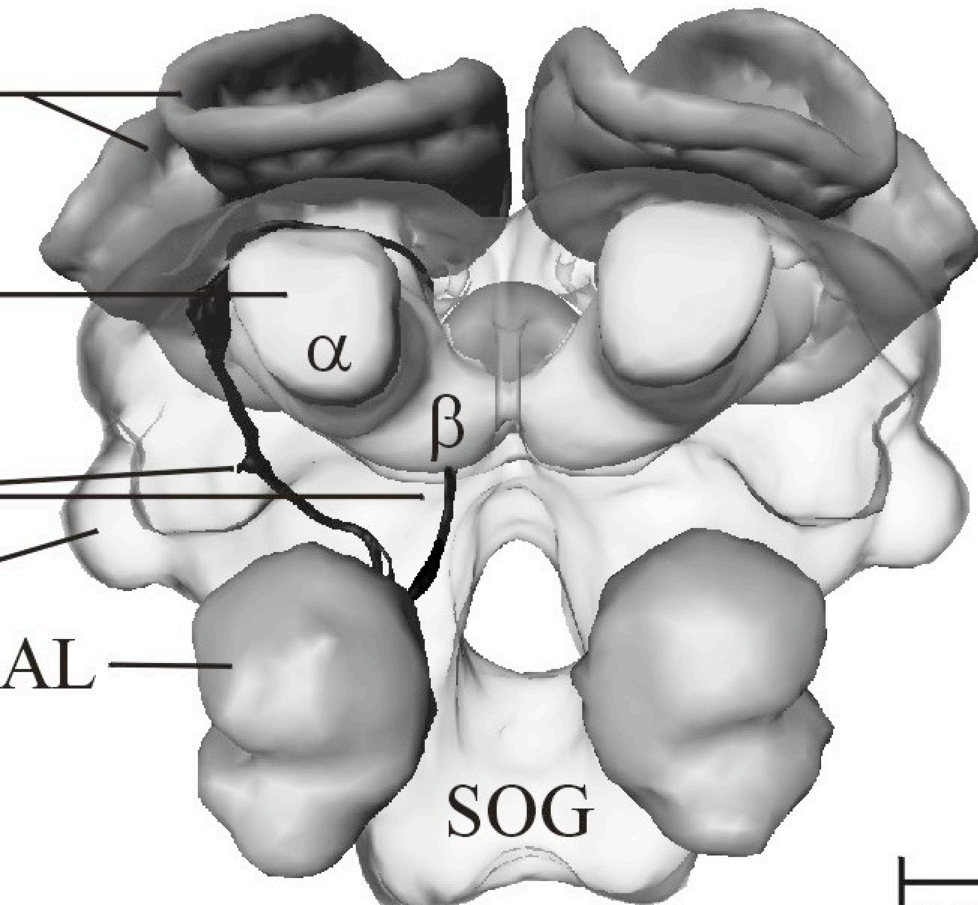
- A policy maps each state to an action to take
  - Like a stimulus–response rule
- We seek a policy that maximizes cumulative reward
- The policy is a subgoal to achieving reward

The Reward Hypothesis

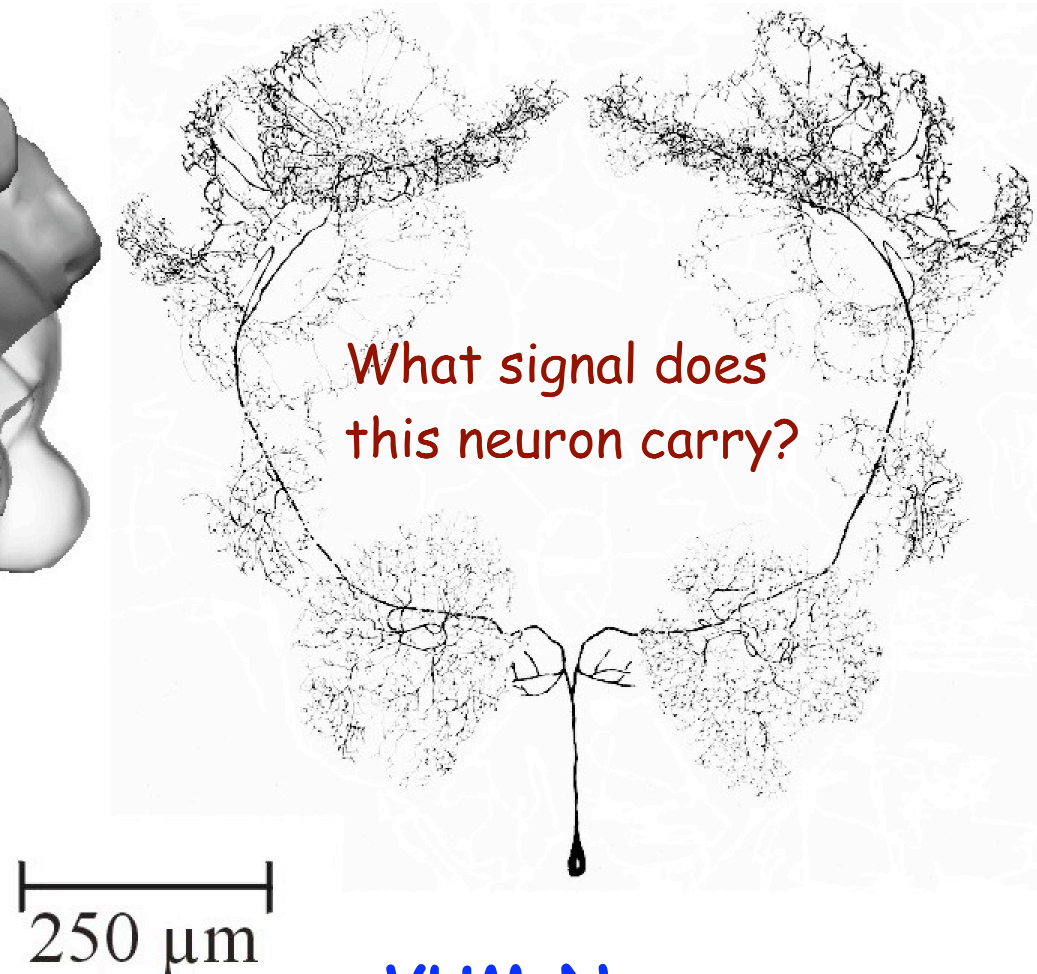
The goal of intelligence is to maximize the cumulative sum of a single received number:
“reward” = pleasure - pain

Artificial Intelligence = reward maximization

Brain reward systems



Honeybee Brain



VUM Neuron

Hammer, Menzel

Value

Value systems are hedonism with foresight

We value situations according to how much reward we expect will follow them

All efficient methods for solving sequential decision problems determine (learn or compute) "value functions" as an intermediate step

Value systems are a *means* to reward, yet we *care more* about values than rewards

Pleasure ≠ good

“Even enjoying yourself you call evil whenever it leads to the loss of a pleasure greater than its own, or lays up pains that outweigh its pleasures. ... Isn't it the same when we turn back to pain? To suffer pain you call good when it either rids us of greater pains than its own or leads to pleasures that outweigh them.”

–Plato, Protagoras

- Reward

  r : States $\rightarrow$ Pr($\Re$)   e.g., $r_t \in \Re$

- Policies

  π : States $\rightarrow$ Pr(Actions) π^* is optimal

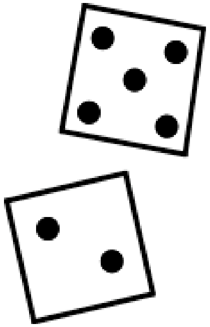
- Value Functions

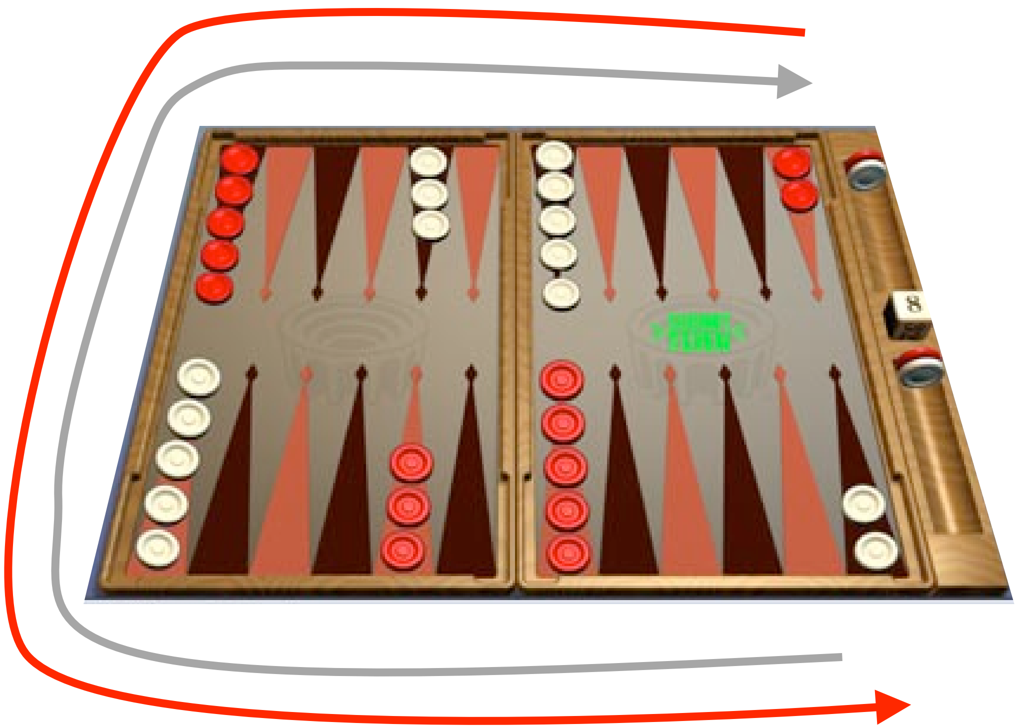
  V^π : States $\rightarrow \Re$

  $$V^\pi(s) = E_\pi\left\{\sum_{k=1}^{\infty} \gamma^{k-1} r_{t+k}\right\} \text{ given that } s_t = s$$

discount factor ≈1 but <1

Backgammon





STATES: configurations of the playing board ($\approx 10^{20}$)

ACTIONS: moves

REWARDS: win: +1

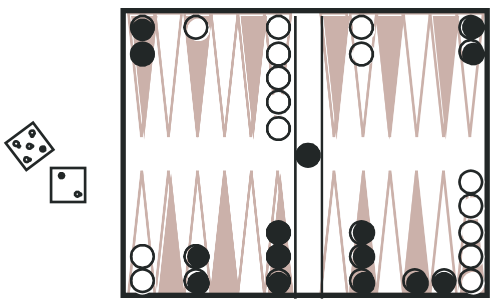
lose: –1

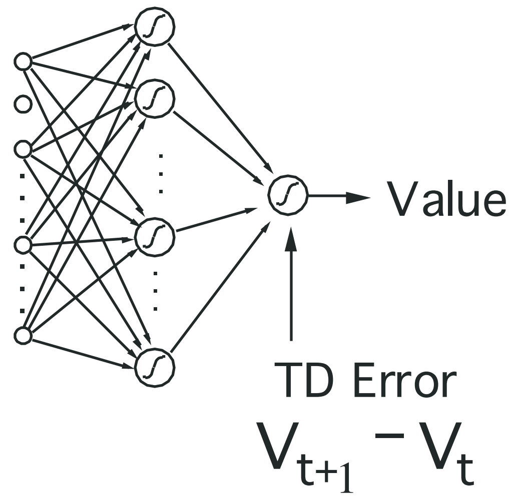
else: 0

a "big" game

TD-Gammon

Tesauro, 1992-1995





Action selection
by 2-3 ply search

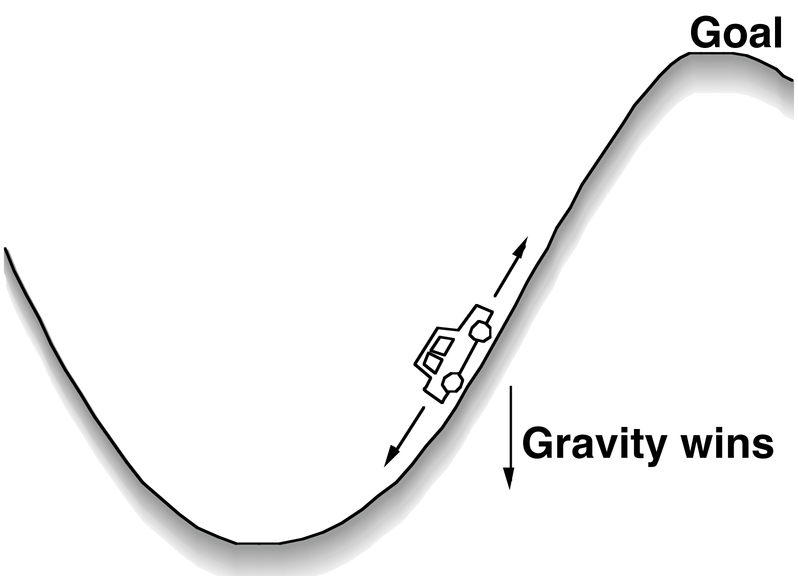
Start with a random Network

Play millions of games against itself

Learn a value function from this simulated experience

Six weeks later it's the best player of backgammon in the world

The Mountain Car Problem



SITUATIONS: car's position and velocity

ACTIONS: three thrusts: forward, reverse, none

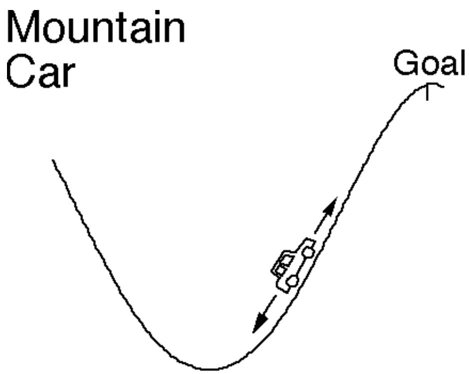
REWARDS: always –1 until car reaches the goal

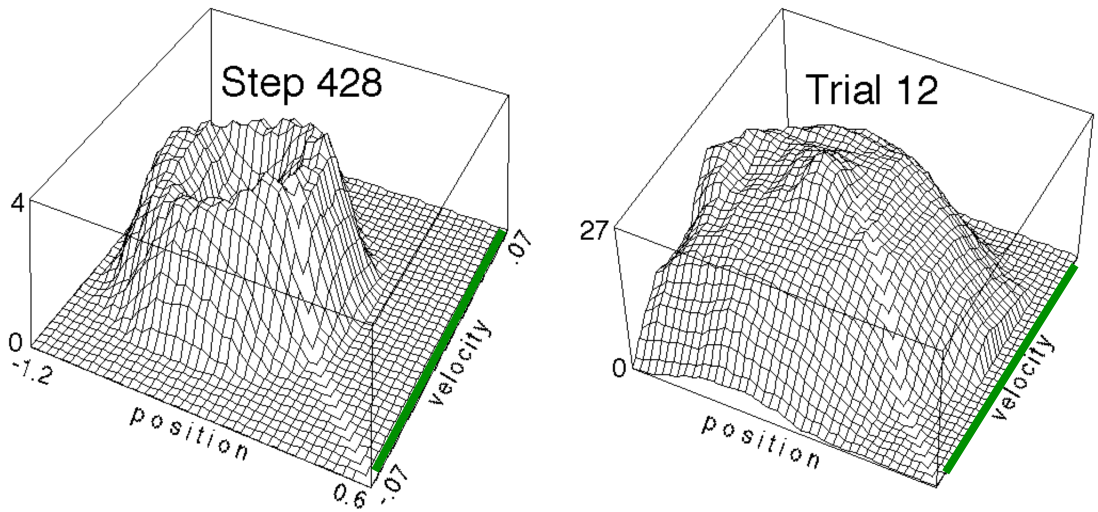
No Discounting

Minimum-Time-to-Goal Problem

Moore, 1990

Value Functions Learned while solving the Mountain Car problem

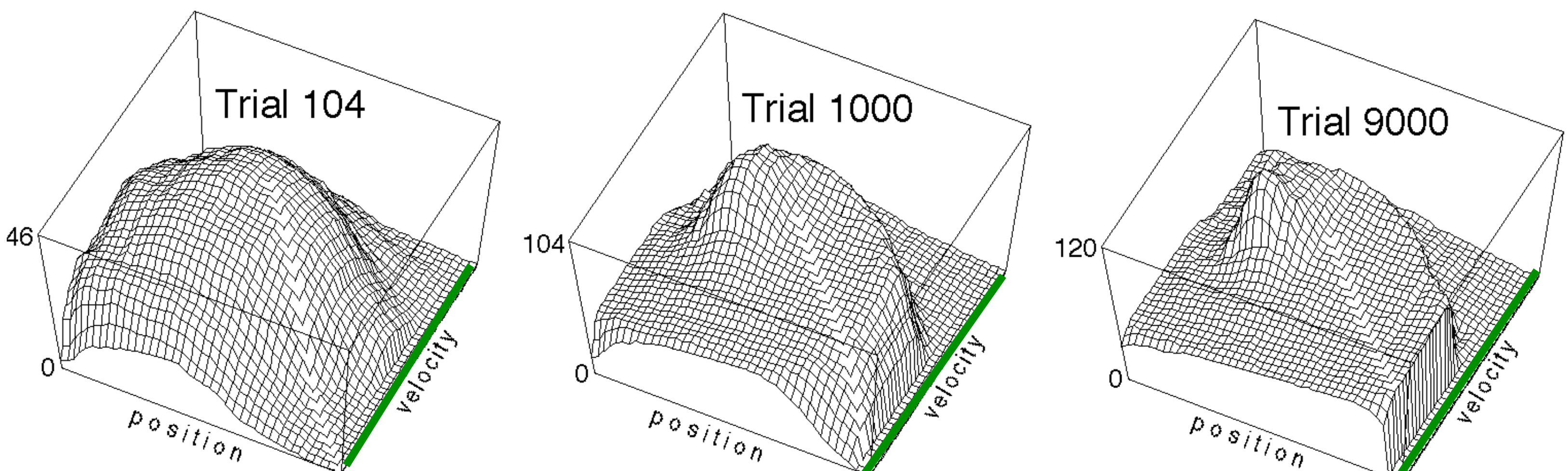




Goal region

Minimize Time-to-Goal

Value = estimated time to goal



TD error

- Reward

  r : States $\rightarrow$ Pr($\Re$) e.g., $r_t \in \Re$

- Policies

  π : States $\rightarrow$ Pr(Actions) π^* is optimal

- Value Functions

  V^{π} : States $\rightarrow \Re$

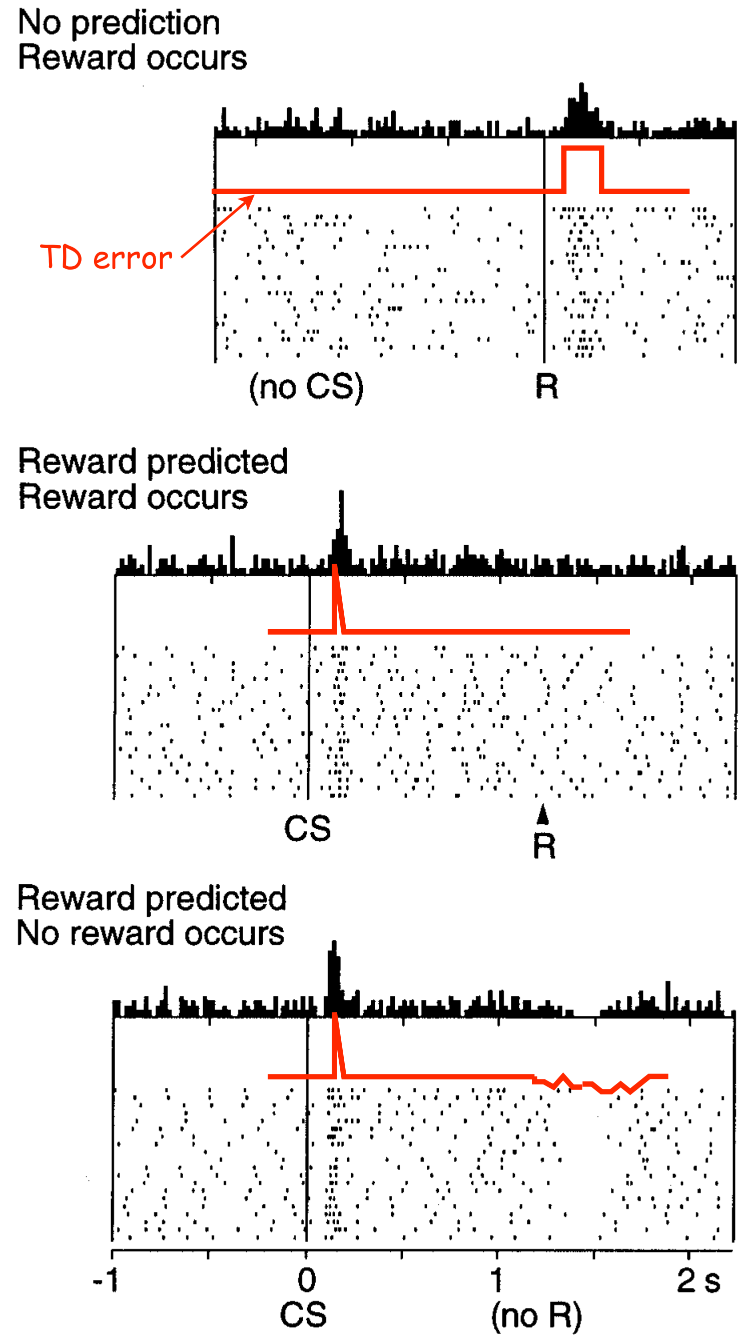
  $$V^{\pi}(s) = E_{\pi}\left\{\sum_{k=1}^{\infty} \gamma^{k-1} r_{t+k}\right\} \text{ given that } s_t = s$$

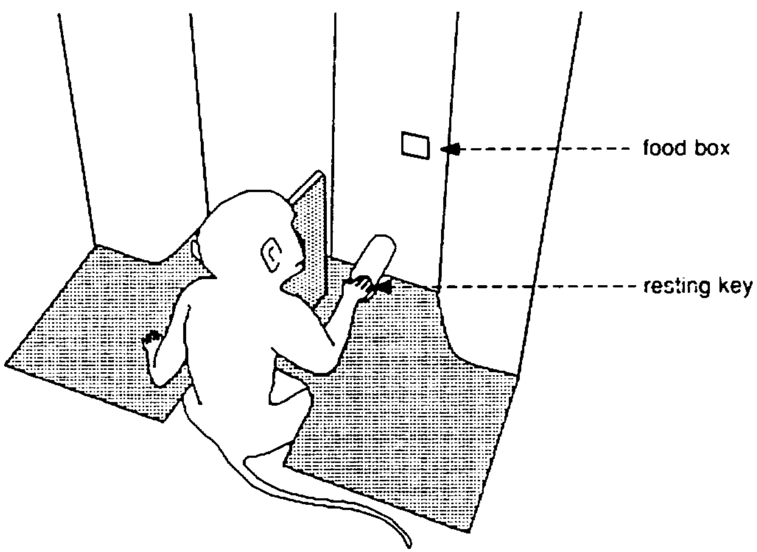
discount factor ≈1 but <1

$$
\begin{aligned}
V_t &= E\left\{\sum_{k=1}^{\infty} \gamma^{k-1} r_{t+k}\right\} \\
&= E\left\{r_{t+1} + \gamma \sum_{k=1}^{\infty} \gamma^{k-1} r_{t+1+k}\right\} \\
&\approx r_{t+1} + \gamma V_{t+1}
\end{aligned}
$$

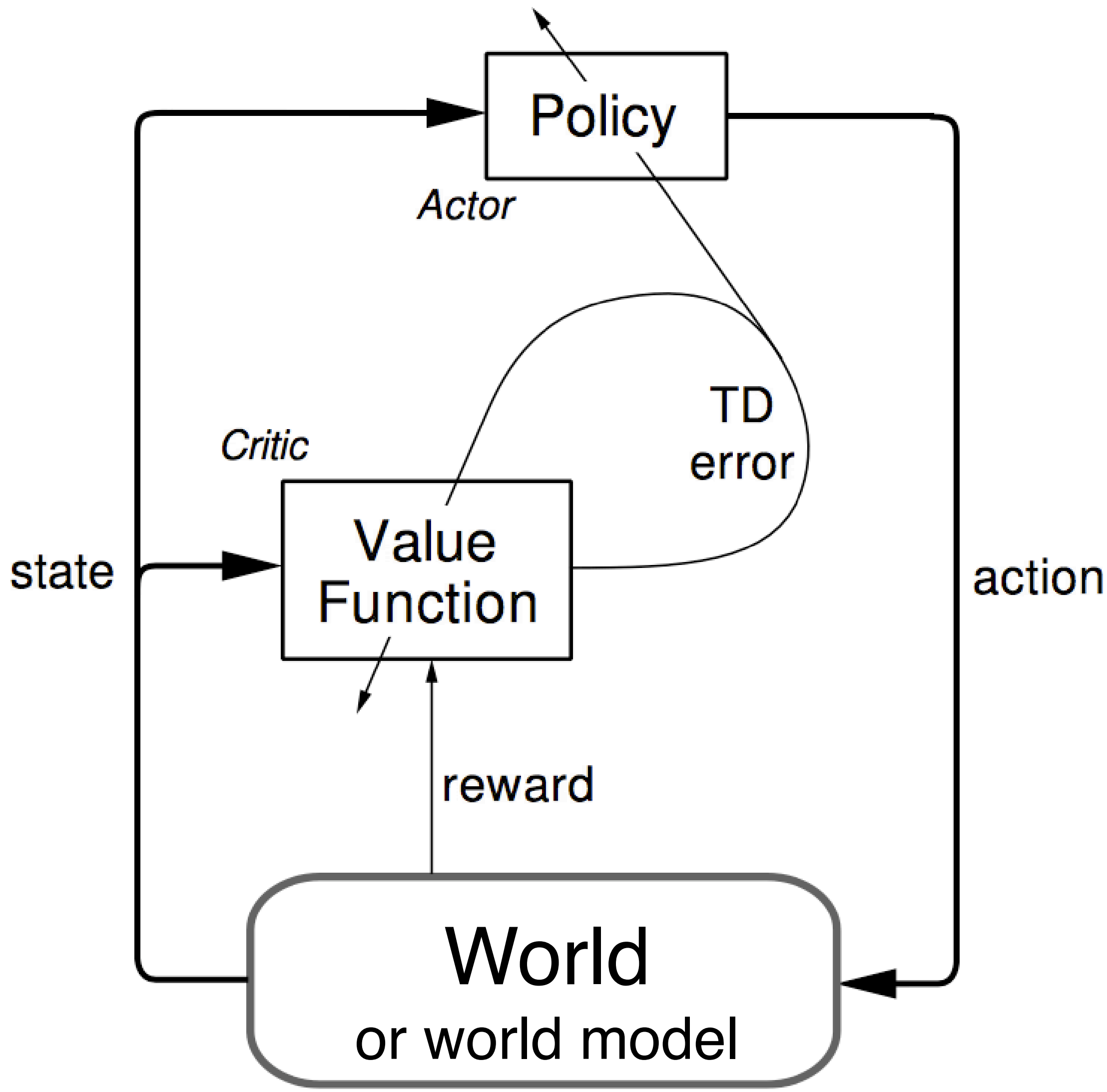
$$
TD\,error_t = r_{t+1} + \gamma V_{t+1} - V_t
$$

Brain reward systems seem to signal TD error





Wolfram Schultz, et al.



Policy
Actor
Critic
TD
error
Value
Function
state
action
reward
World
or world model

World models

“Autonomous helicopter flight via Reinforcement Learning”

Ng (Stanford), Kim, Jordan, & Sastry (UC Berkeley) 2004

Reason as RL over Imagined Experience

1. Learn a predictive model of the world's dynamics

   transition probabilities, expected immediate rewards

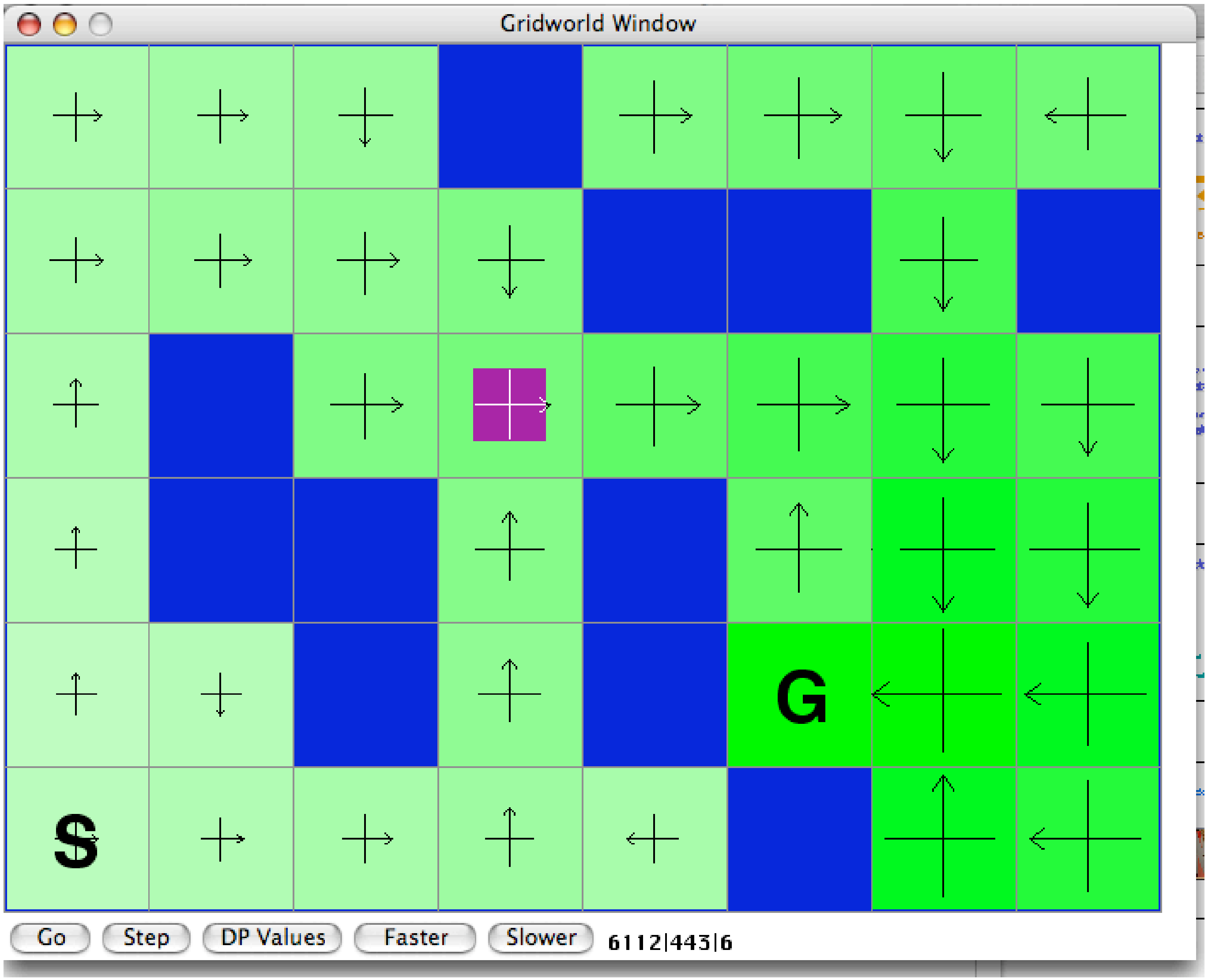
2. Use model to generate imaginary experiences

   internal thought trials, mental simulation (Craik, 1943)

3. Apply RL as if experience had really happened

   vicarious trial and error (Tolman, 1932)

GridWorld Example



Intro to RL

with a robot bias

- policy
- reward
- value
- TD error
- world model
- generalized policy iteration

Policy Iteration

$$\pi_0 \rightarrow V^{\pi_0} \rightarrow \pi_1 \rightarrow V^{\pi_1} \rightarrow \cdots \pi^* \rightarrow V^* \rightarrow \pi^*$$

policy evaluation

policy improvement
"greedification"

Improvement is monotonic

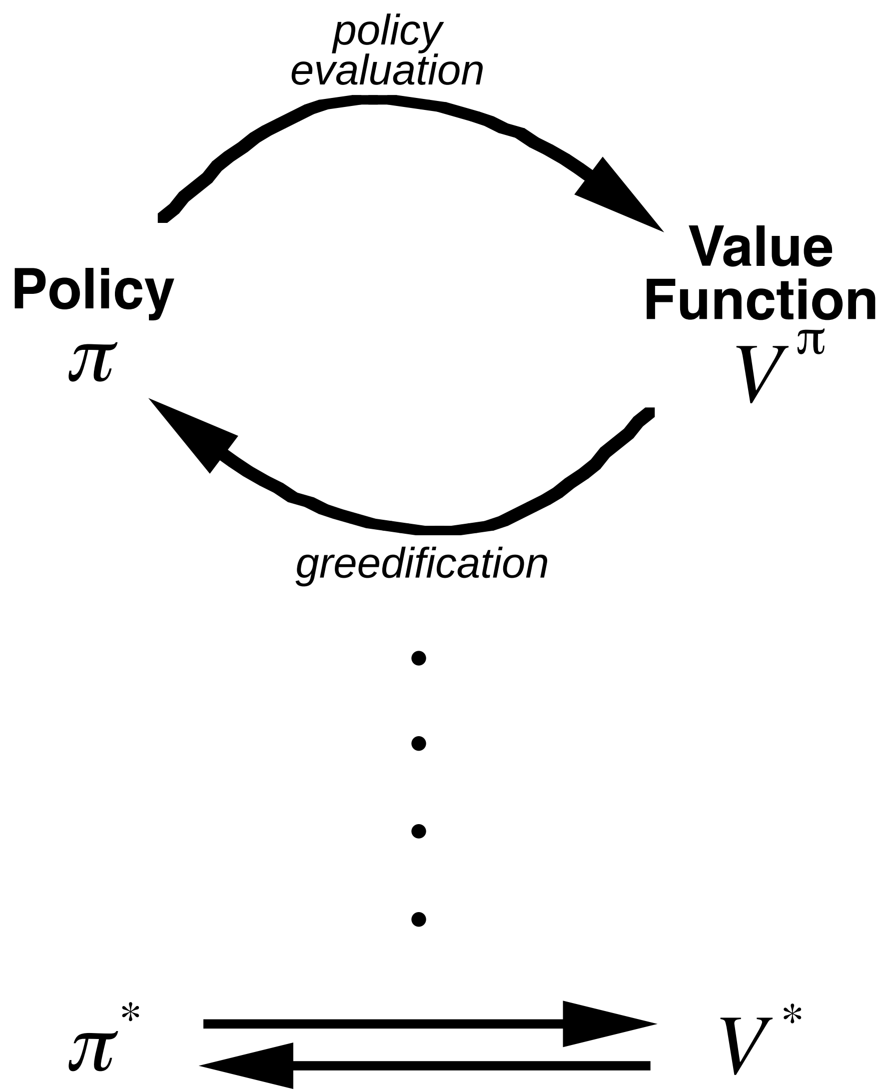
Converges is a finite number of steps

<u>Generalized</u> Policy Iteration:

Intermix the two steps more finely,

state by state, action by action, sample by sample

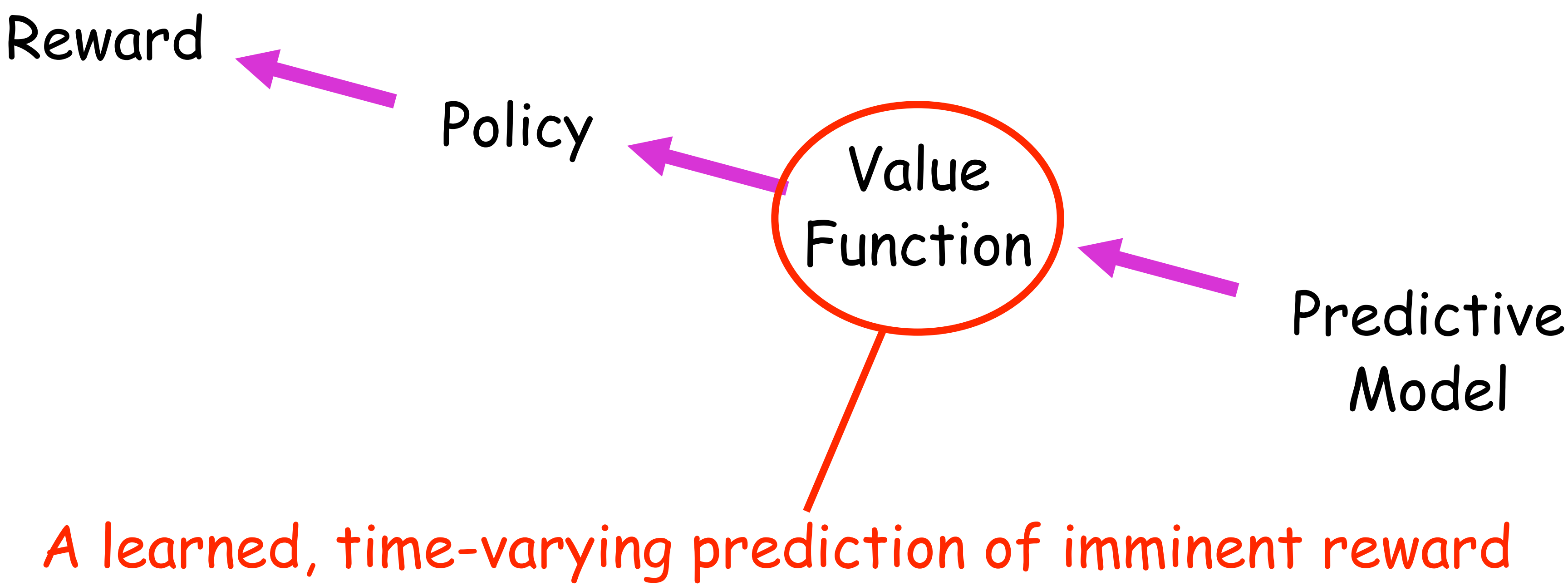
Generalized Policy Iteration



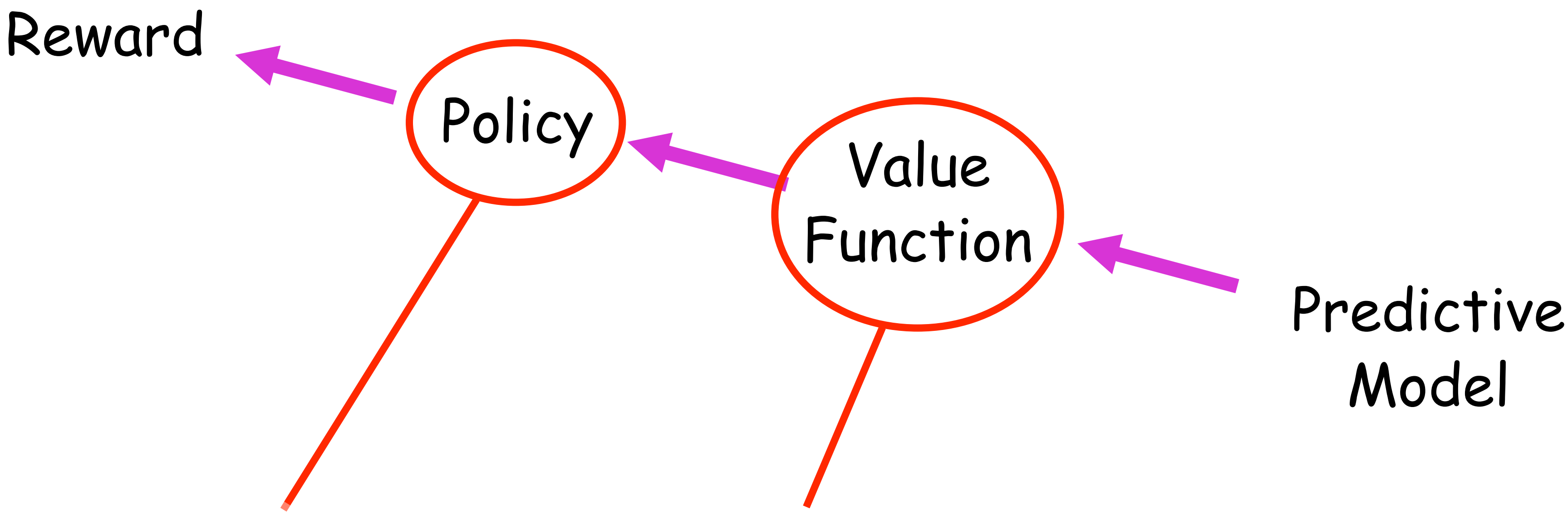
Summary:
RL's Computational Theory of Mind

Reward $\leftarrow$ Policy $\leftarrow$ Value Function $\leftarrow$ Predictive Model

Summary:
RL's Computational Theory of Mind

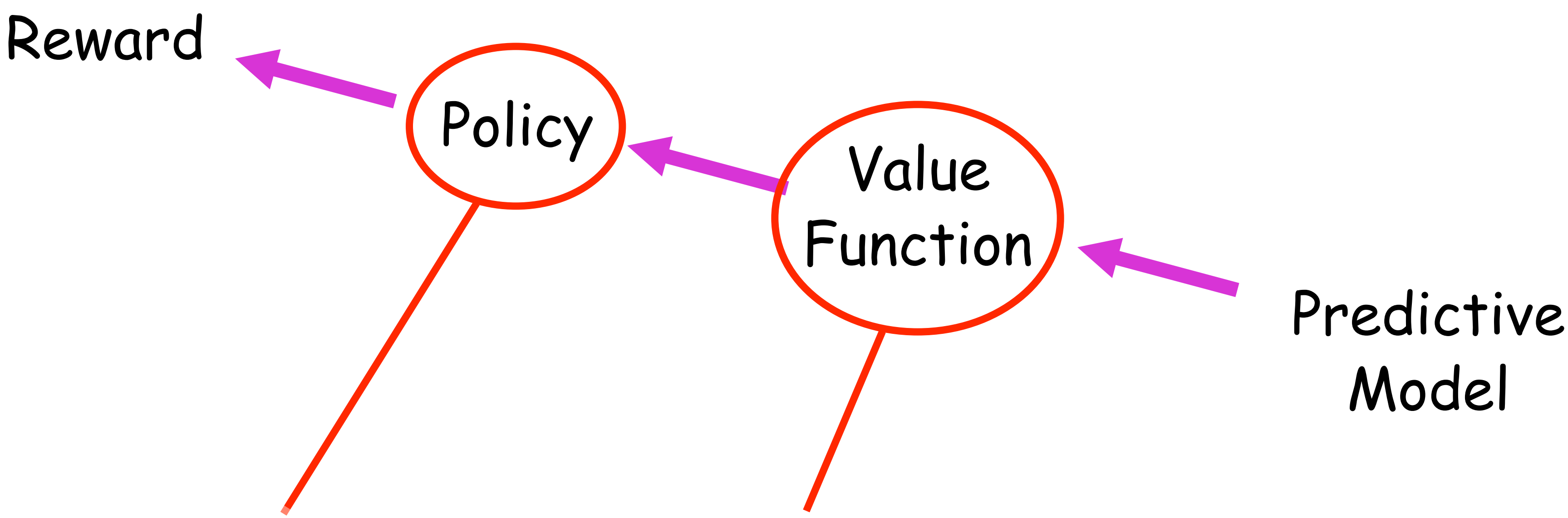


Summary:
RL's Computational Theory of Mind



A learned, time-varying prediction of imminent reward
Key to all efficient methods for finding optimal policies

Summary:
RL's Computational Theory of Mind



A learned, time-varying prediction of imminent reward
Key to all efficient methods for finding optimal policies

This has nothing to do with either biology or computers

Summary:
RL's Computational Theory of Mind

Reward

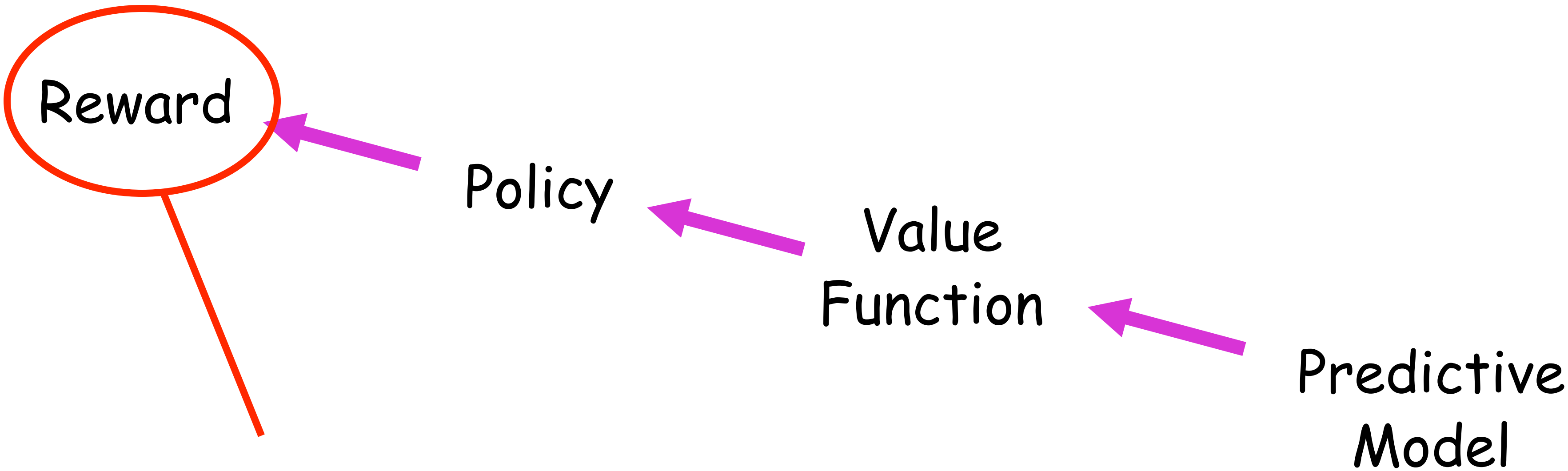
Policy

Value Function

Predictive Model

It's all created from the scalar reward signal

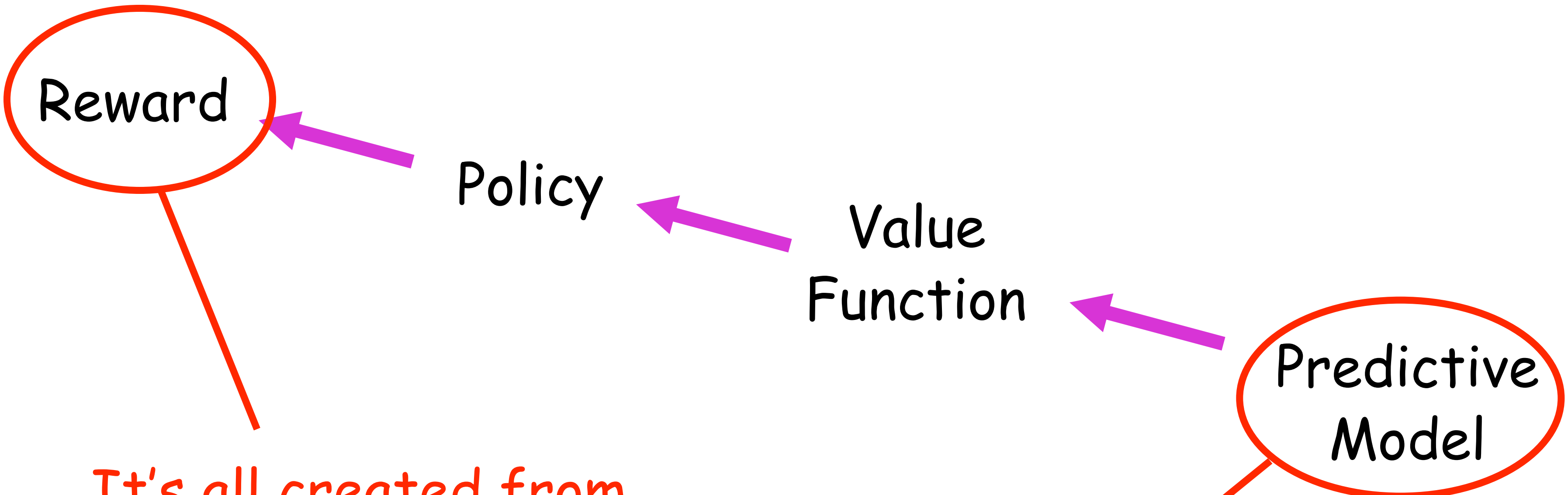
Summary:
RL's Computational Theory of Mind



It's all created from
the scalar reward signal

together with the causal structure of the world

Summary:
RL's Computational Theory of Mind

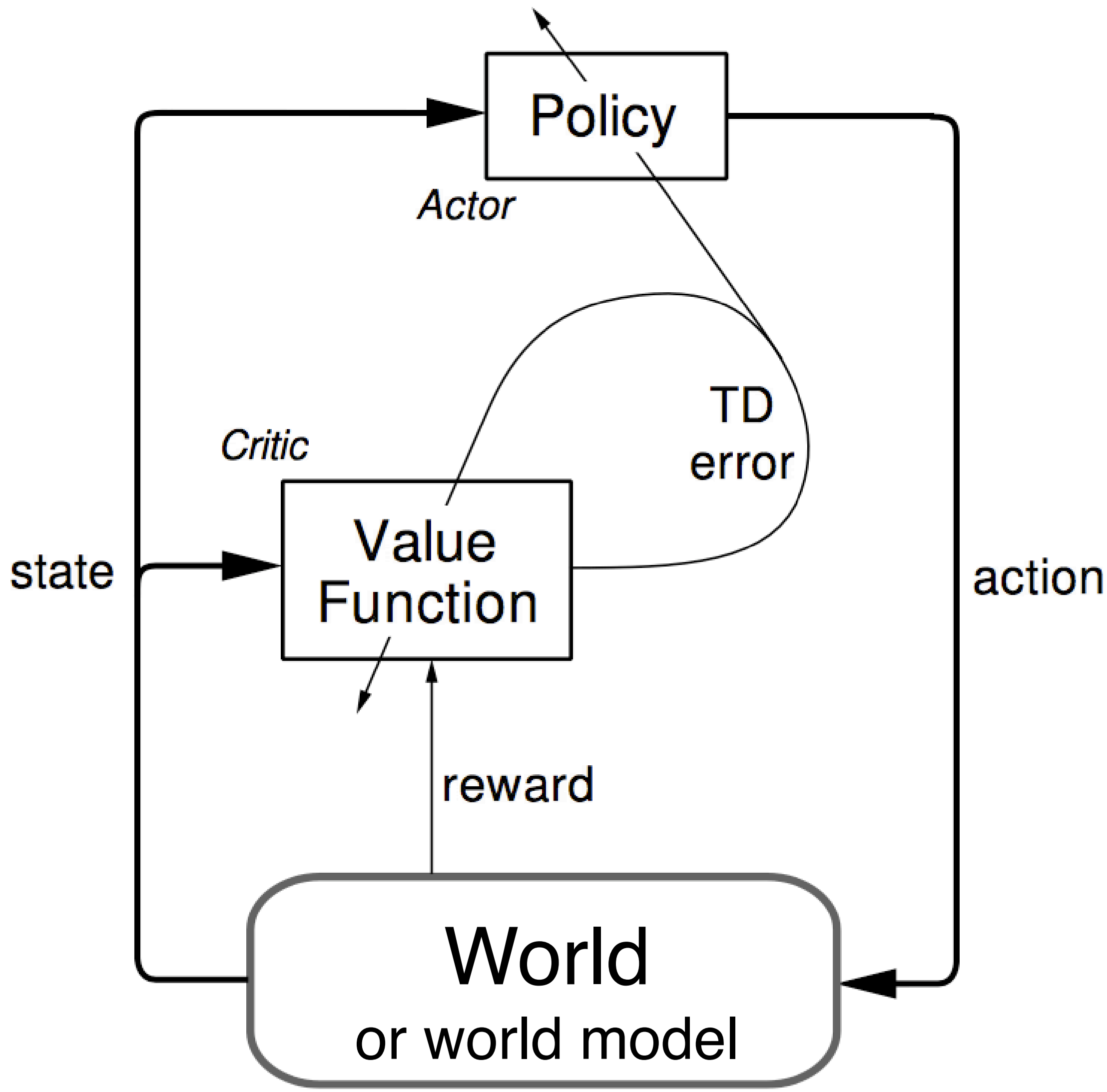


It's all created from
the scalar reward signal

together with the causal structure of the world

additions for next year

- sarsa
- how it might actually applied to something they might do


Policy
Actor
TD
error
Critic
Value
Function
state
action
reward
World
or world model

Current work

Knowledge is subjective

An individual's knowledge consists of predictions about its future (low-level) sensations and actions

Everything else (objects, houses, people, love) must be constructed from sensation and action

Subjective/Predictive Knowledge

- “John is in the coffee room”
- “My car in is the South parking lot”
- What we know about geography, navigation
- What we know about how an object looks, rotates
- What we know about how objects can be used
- Recognition strategies for objects and letters
- “The portrait of Washington on the dollar in the wallet in my other pants in the laundry, has a mustache on it”

RoboCup soccer keepaway

Stone, Sutton & Kuhlmann, 2005