Examples and Videos of Markov Decision Processes (MDPs) and Reinforcement Learning
Artificial Intelligence is interaction to achieve a goal

- Environment
- Agent
- State
- Action
- Reward

- Complete agent
- Temporally situated
- Continual learning & planning
- Object is to affect environment
- Environment stochastic & uncertain
States, Actions, and Rewards
Hajime Kimura's RL Robots

Before

Backward

After

New Robot, Same algorithm
Devilsticking

Finnegan Southey  
University of Alberta

Stefan Schaal & Chris Atkeson  
Univ. of Southern California  
“Model-based Reinforcement Learning of Devilsticking”
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
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 = Immediate Reward
≠ good = Long-term Reward

“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
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

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

Moore, 1990
Value Functions Learned while solving the Mountain Car problem

Minimize Time-to-Goal

Value = estimated time to goal
Temporal-difference (TD) error

Do things seem to be getting better or worse, in terms of long-term reward, at this instant in time?
Brain reward systems

What signal does this neuron carry?

Honeybee Brain

VUM Neuron

Hammer, Menzel
Brain reward systems seem to signal TD error

Wolfram Schultz, et al.
World models
the actor-critic reinforcement learning architecture
“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
Summary: RL’s Computational Theory of Mind

- **Policy**: A learned, time-varying prediction of imminent reward
- **Value Function**: Key to all efficient methods for finding optimal policies
- **Predictive Model**: 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.