Reinforcement Learning:
Time for Action!
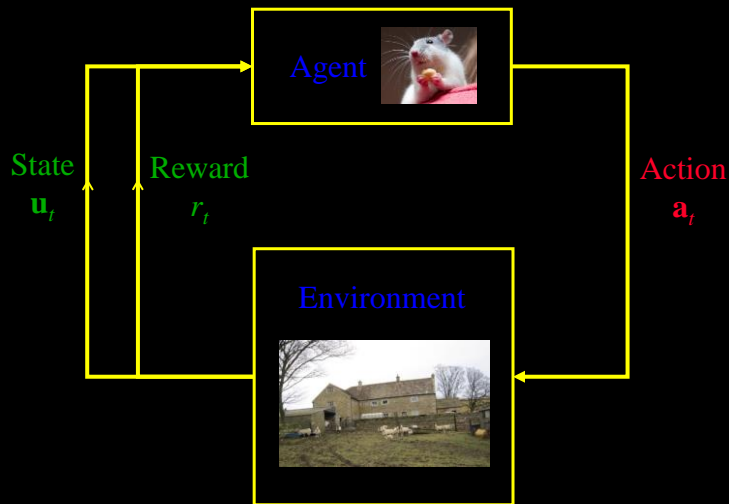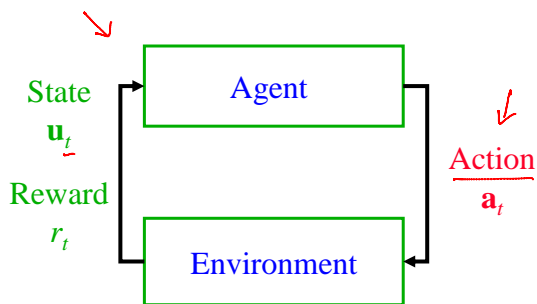
Agent

State
$\mathbf{u}_t$

Reward
r_t

Action
$\mathbf{a}_t$

Environment

Image Source: Wikimedia Commons

The Problem

State
$\mathbf{u}_t$

Reward
r_t
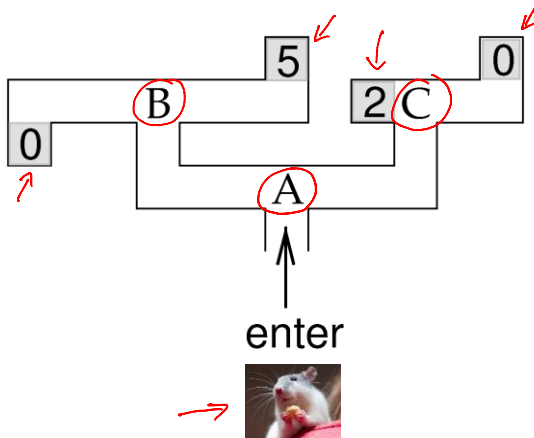
Agent

Action
$\mathbf{a}_t$

Environment

Learn a state-to-action mapping or "policy":

$$\pi(\mathbf{u}) = \mathbf{a}$$

which maximizes the expected total future reward:

$$\left\langle \sum_{\tau=0}^{T-t} r(t+\tau) \right\rangle_{trials}$$
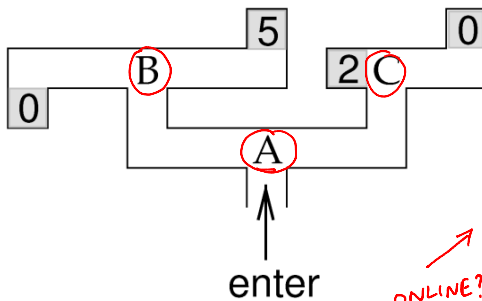
2

Example: Rat in a barn

States = locations A, B, or C

Actions= L (go left) or R (go right)

If the rat chooses L or R *at random* (random "policy"), what is the expected reward (or "*value*") v for each state?

Policy Evaluation

enter

ONLINE?

For random policy:

$$v(B) = \frac{1}{2} \cdot 0 + \frac{1}{2} \cdot 5 = 2.5$$

$$v(C) = \frac{1}{2} \cdot 2 + \frac{1}{2} \cdot 0 = 1$$

$$v(A) = \frac{1}{2} \cdot v(B) + \frac{1}{2} \cdot v(C) = 1.75$$

Let value of state u
$v(u)$ = weight $w(u)$

Can learn value of states using TD learning:

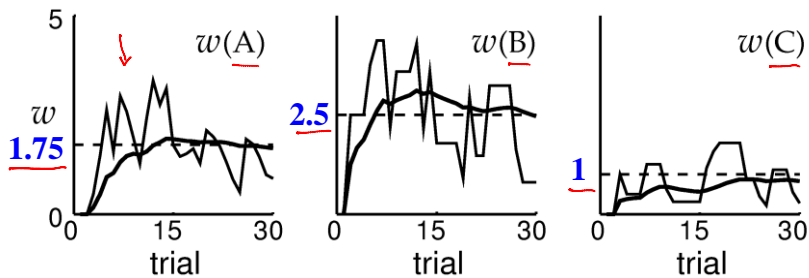$$w(u) \leftarrow w(u) + \varepsilon \left[r(u) + v(u') - v(u) \right]$$

(Location, action) $\Rightarrow$ new location i.e., $(u,a) \Rightarrow u'$

reward   reward   reward

TD Learning of Values for Random Policy

$w(A)$ $\quad$ $w(B)$ $\quad$ $w(C)$

w

1.75

2.5

1

5

0

0 $\quad$ 15 $\quad$ 30

trial

0 $\quad$ 15 $\quad$ 30
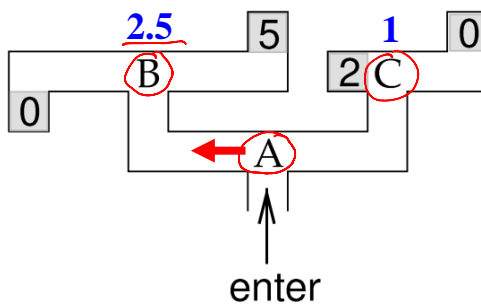
trial

0 $\quad$ 15 $\quad$ 30

trial

(For all three, $\varepsilon = 0.5$)

Once I know the values, I can pick the action that leads to the higher valued state!

5

Selecting Actions based on Values

2.5 $\quad$ 5 $\quad$ **1** $\quad$ 0

B $\qquad$ 2 C

0

A

enter

Values act as surrogate immediate rewards → Locally optimal choice leads to globally optimal policy for "Markov" environments (Related to *Dynamic Programming*)

6

Putting it all together:
Actor-Critic Learning

✦ Two separate components: Actor (selects action and maintains policy) and Critic (maintains value of each state)

1. Critic Learning ("Policy Evaluation"):
   Value of state $u = v(u) = w(u)$
   $$w(u) \leftarrow w(u) + \varepsilon\,[r(u) + v(u') - v(u)] \qquad \text{(same as TD rule)}$$

2. Actor Learning ("Policy Improvement"):
   $$P(a;u) = \frac{\exp(\beta Q_a(u))}{\sum_b \exp(\beta Q_b(u))} \qquad \text{← SOFTMAX}$$

   Probabilistically select an action a at state u   ← EXPLORE
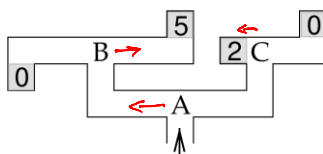
   For all actions a':
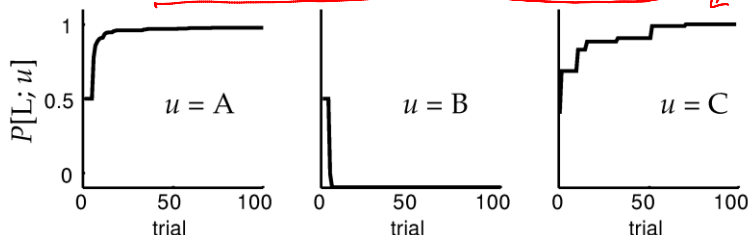 $$Q_{a'}(u) \leftarrow Q_{a'}(u) + \varepsilon[r(u) + v(u') - v(u)](\delta_{aa'} - P(a';u))$$

3. Repeat 1 and 2

7

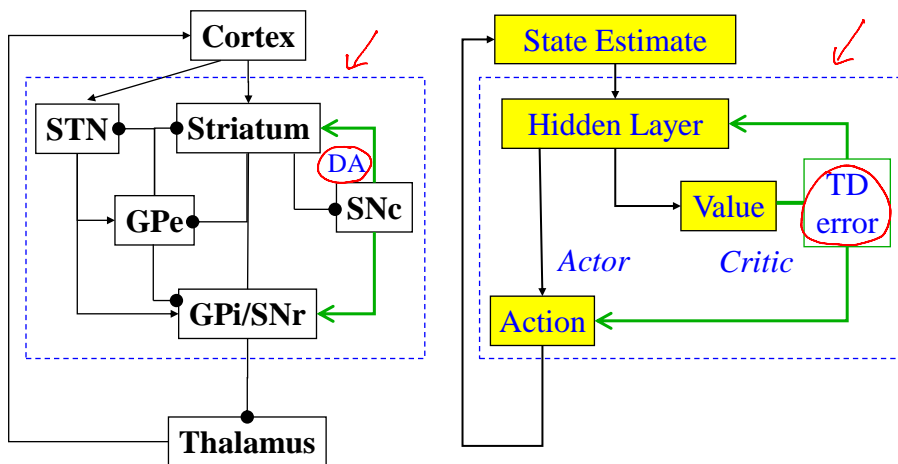## Actor-Critic Learning in our Barn Example

Probability of going Left at each location

8

Possible Implementation of the Actor-Critic Model in the Basal Ganglia

9

(See Supplementary Materials for references)

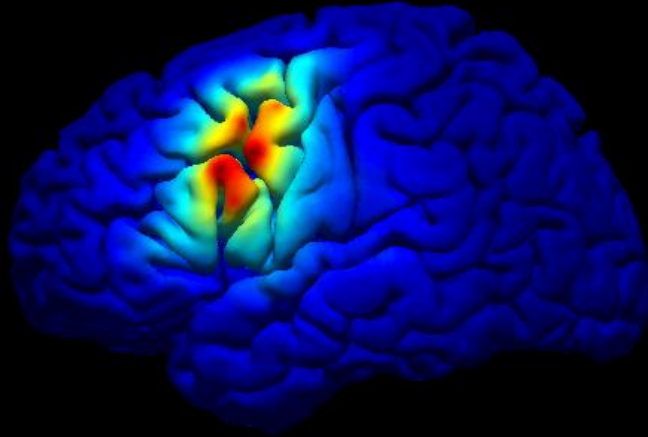**Reinforcement learning has been applied to many real-world problems!**

Example:

Autonomous Helicopter Flight

(learned from human demonstrations)

10

(Videos and papers at: http://heli.stanford.edu/)

Computational Neuroscience

Rajesh P. N. Rao

Adrienne Fairhall

University of Washington, Seattle, USA