# What you have learned from bandits

- The need to tradeoff exploitation and exploration, e.g., by an $\varepsilon$-greedy policy
- The difference between a sample, an estimate, and a true expected value

$$R_t, \quad Q_t(a), \quad q_*(a)$$

- The difference between the greedy action and the optimal action
- A learning rule. How learning can be seen as computing an average
  - The role of the step size $\alpha$ (how it can be too big, or too small, or "1/n")
- A complete example of mathematically formalized goal seeking (intelligence) —both problem and solution methods

# Defining “Intelligent Systems”

# Defining “System”

- A thing
  - with some recognizable identity over time (need not be physical)
  - usually with some inputs and outputs
  - may have state (not a function)
  - sometimes with a goal/purpose

# What is intelligence?
# What is a mind?

Question: What is the right definition of intelligence?

Answer: The computational part of the ability to:

A. improve over time with experience

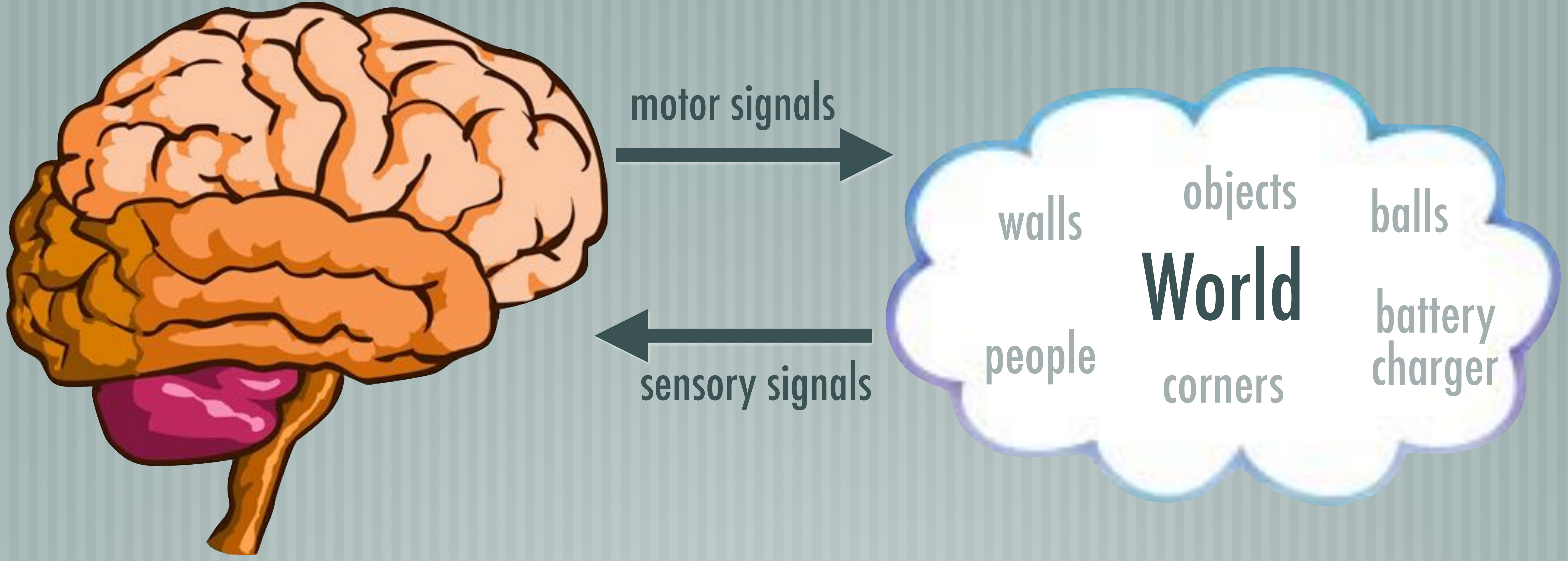
B. use language, communicate and cooperate with other intelligent agents

C. achieve goals

D. predict and control your input signals

E. There is no such thing as a right definition

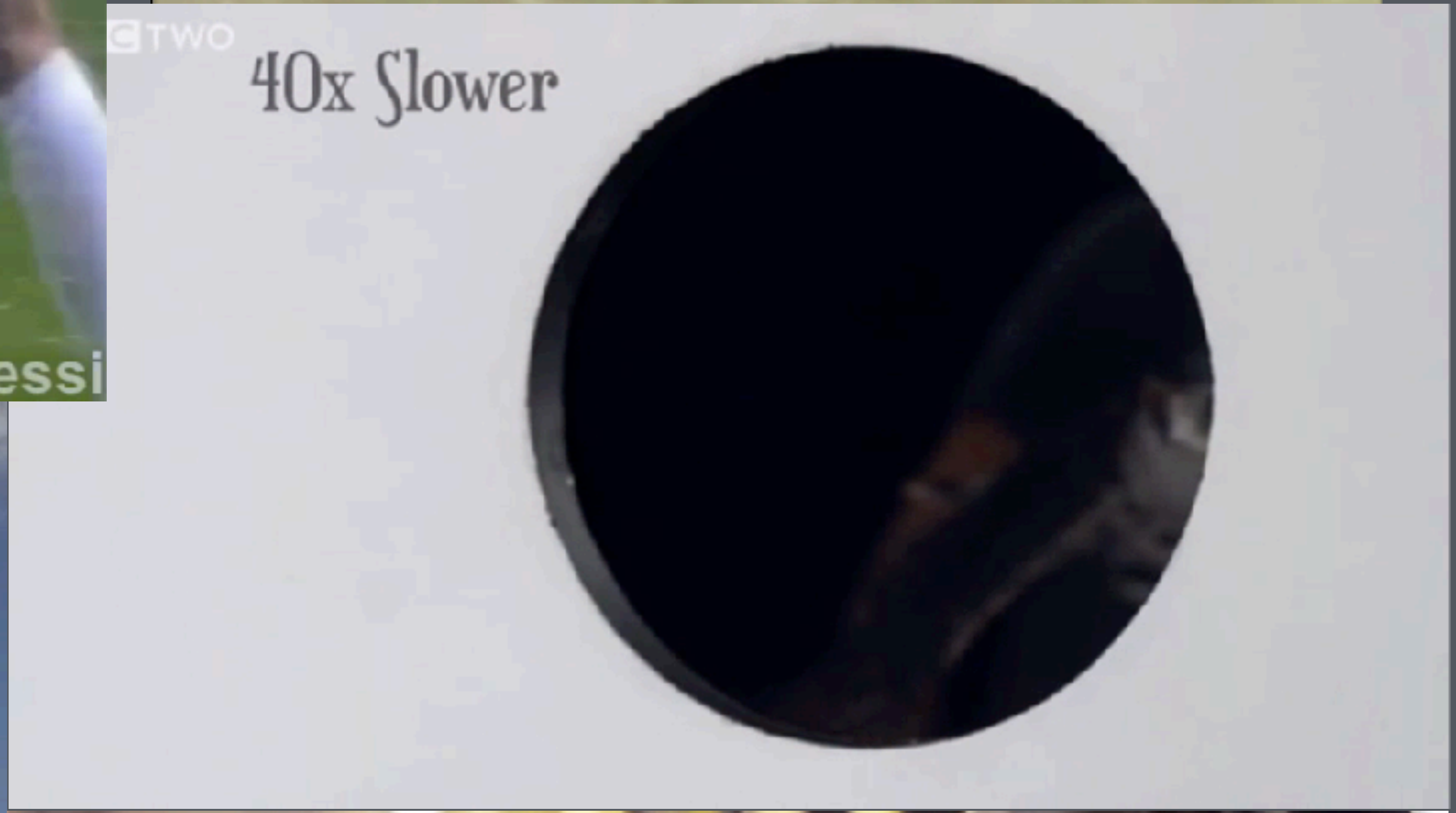
# Minds are sensori-motor information processors

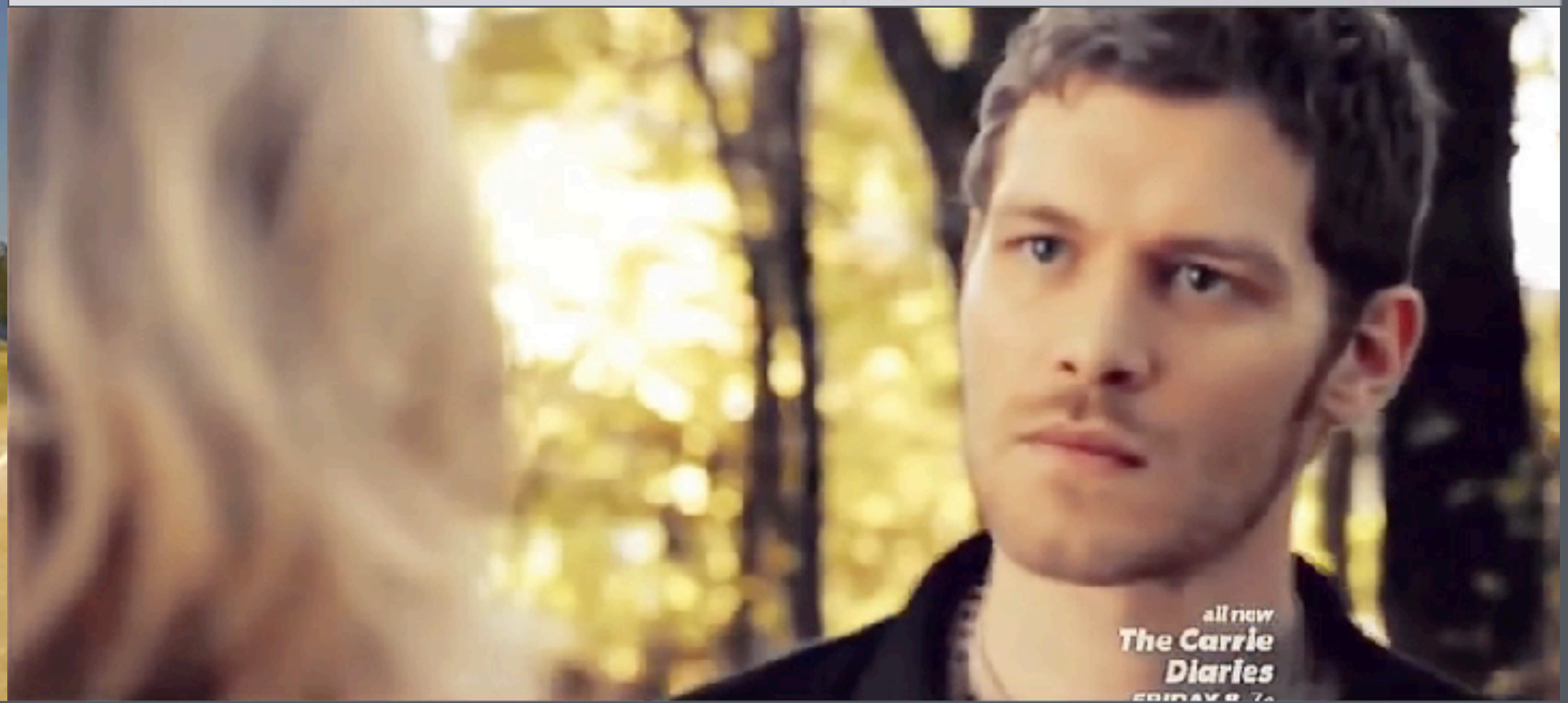



the mind's job is to predict and control its sensory signals

the mind's first responsibility is
real-time sensorimotor information processing

- Perception, action, & anticipation
  - as fast and reactive as possible

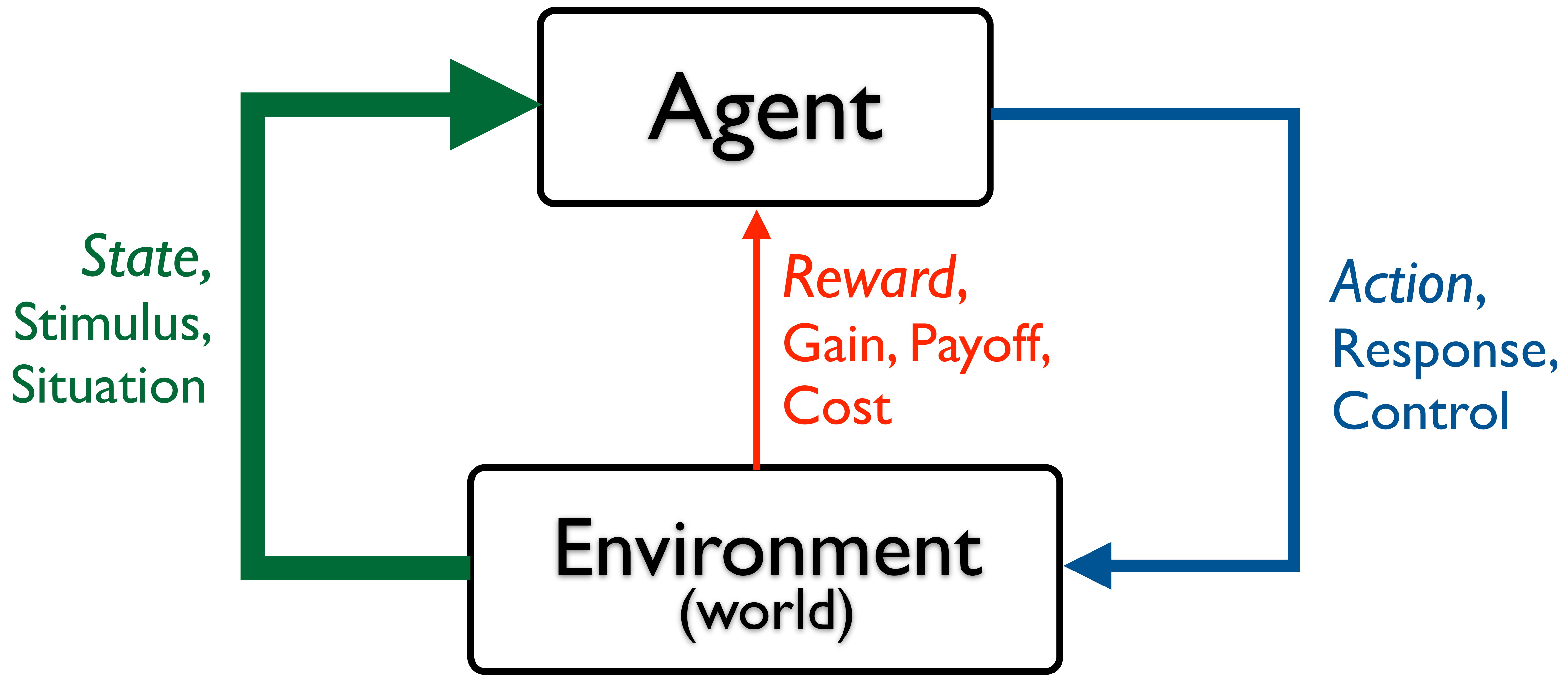
# Artificial Intelligence (definition)

- the science and technology of information processing systems with goals
  - that is, of information processing systems that observers tend to find it useful to think about in terms of goals
    - that is, in terms of outcomes rather than in terms of mechanisms

# An RL perspective on the problem of intelligence

- There are many ways to frame the problem of intelligence, and many possible solution techniques
  - logic, Bayes nets, unsupervised learning, etc…
- In this course we will frame the problem as one of maximizing reward—a reinforcement learning problem
  - Rather than covering different formalisms
- We will discuss learning value functions, finding optimal policies, planning, and function approximation in the context of solving the problem of AI
- We will explore the utility and limitations of a RL approach: what is well worked out, and what are the key open issues

# The RL Interface



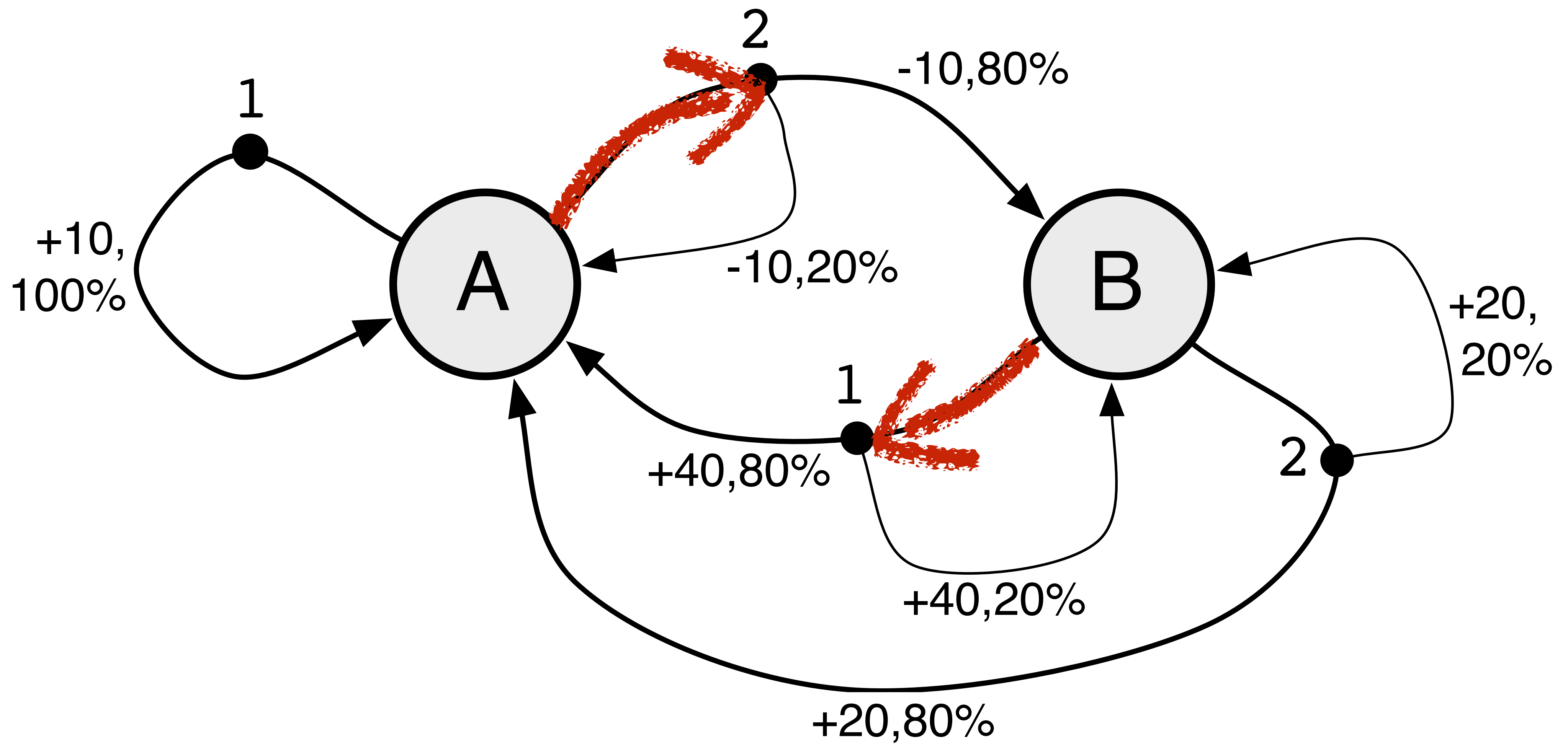

- Environment may be unknown, nonlinear, stochastic and complex
- Agent learns a policy mapping states to actions
  - Seeking to maximize its cumulative reward in the long run

# You are the reinforcement learner! (interactive demo)

Optimal policy
(deterministic)

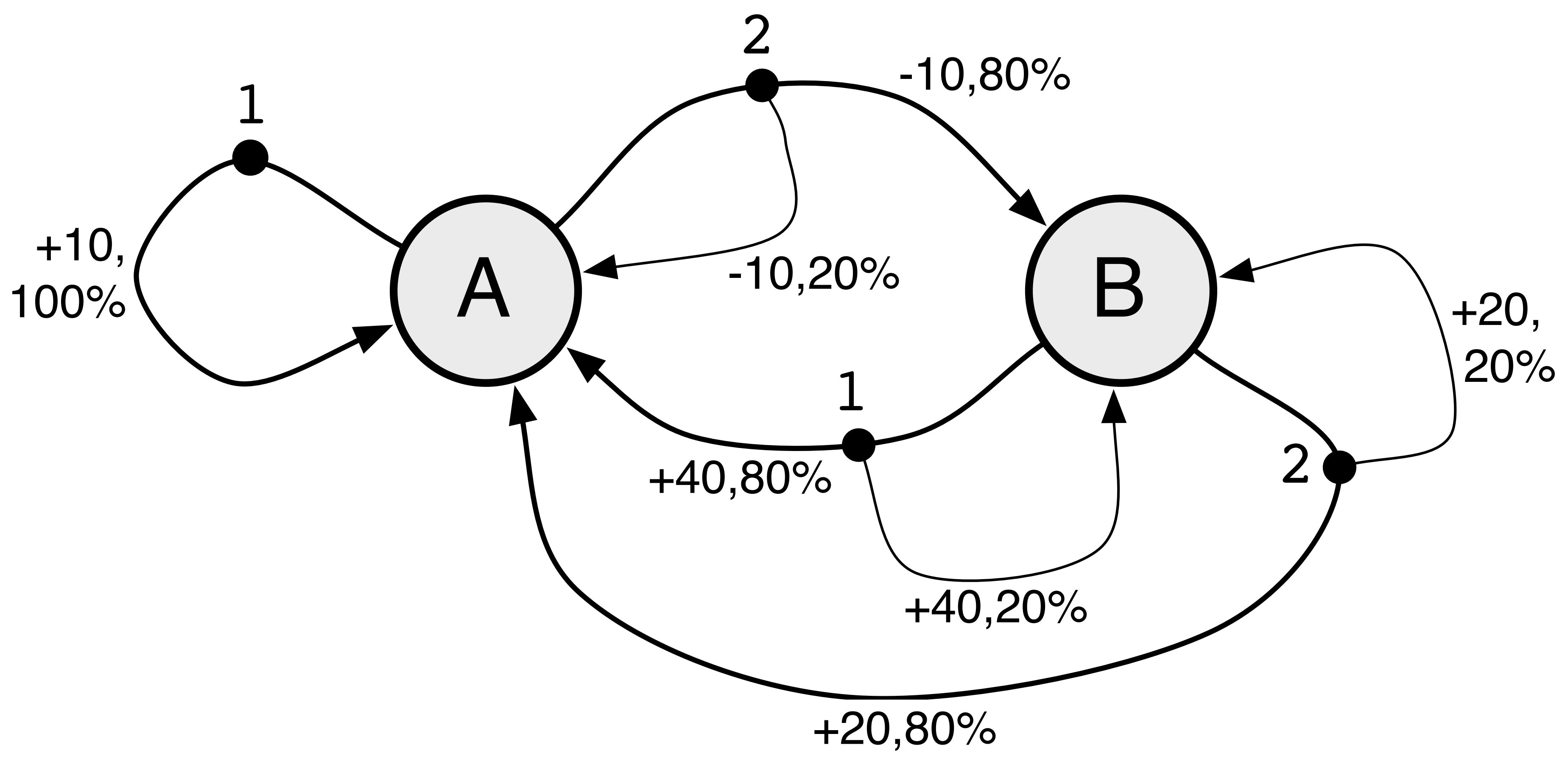
| State | Action |
|---|---|
| A $\longrightarrow$ | 2 |
| B $\longrightarrow$ | 1 |

True model of the world



# The Environment: A Finite Markov Decision Process (MDP)

- Discrete time $t = 1, 2, 3, \ldots$
- A finite set of states
- A finite set of actions
- A finite set of rewards




- Life is a trajectory:

$$\ldots S_t, A_t, R_{t+1}, S_{t+1}, A_{t+1}, R_{t+2}, S_{t+2}, \ldots$$

- With arbitrary Markov (stochastic, state-dependent) dynamics:

$$p(r, s'|s, a) = Prob\Big[R_{t+1} = r, S_{t+1} = s' \;\Big|\; S_t = s, A_t = a\Big]$$

# Policies

e.g.

| State | Action |
|---|---|
| A $\longrightarrow$ | 2 |
| B $\longrightarrow$ | 1 |

- Deterministic policy

$$a = \pi(s)$$

- An agent following a policy

$$A_t = \pi(S_t)$$

The number of deterministic policies is *exponential* in the *number of states*

- Informally the agent's goal is to choose each action so as to maximize the discounted sum of future rewards,

to choose each $A_t$ to maximize $R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + \cdots$

- We are searching for a policy

"gamma", the discount rate $\in [0,1)$

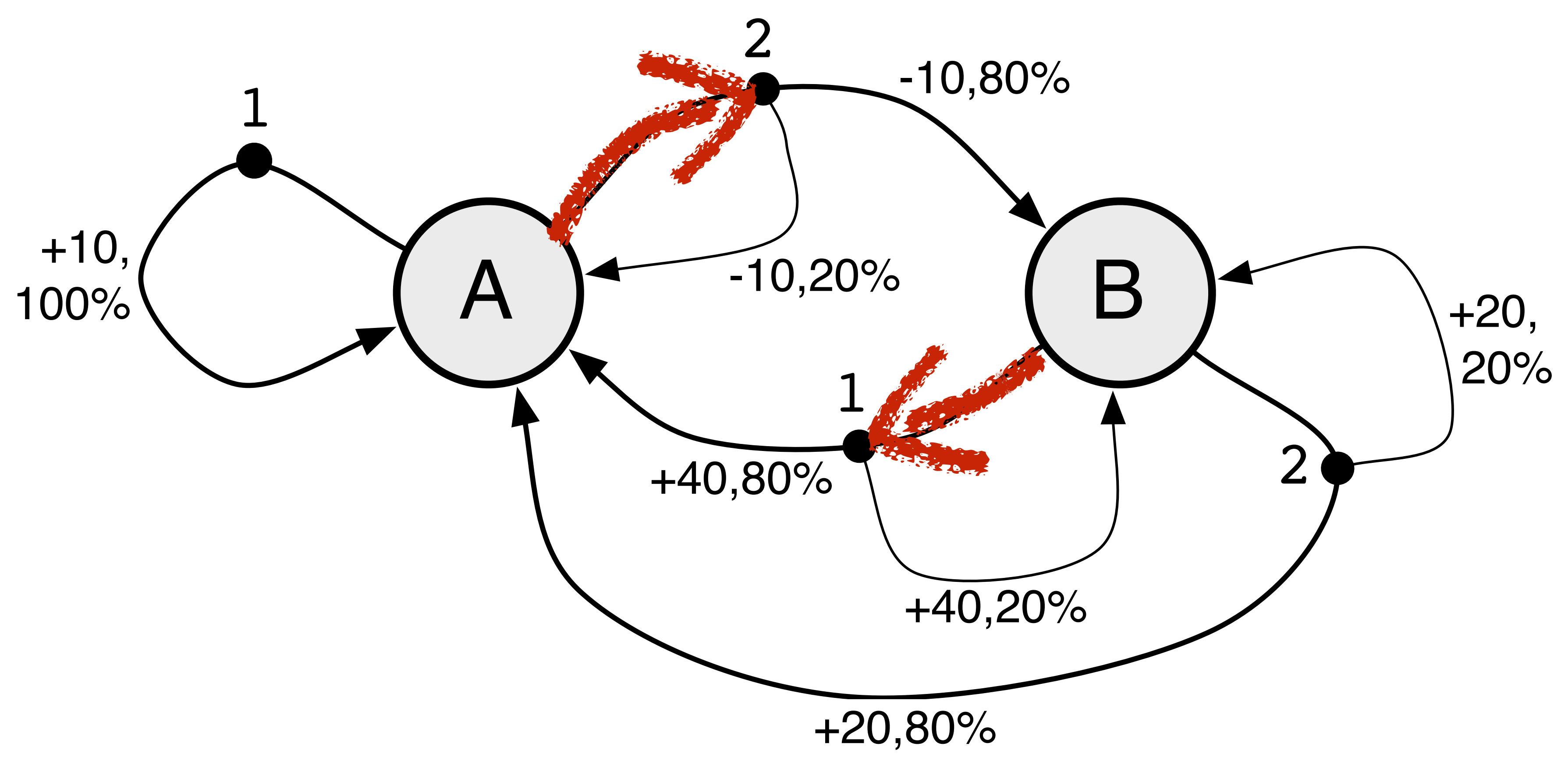
# Action-value functions

- An action-value function says how good it is to be in a state, take an action, and thereafter follow a policy:

$$q_\pi(s, a) = \mathbb{E}\Big[R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + \cdots \;\Big|\; S_t = s, A_t = a, A_{t+1:\infty} \sim \pi\Big]$$

Action-value function
for the optimal policy and $\gamma$=0.9

| State | Action | Value |
|---|---|---|
| A | 1 | 130.39 |
| A | 2 | 133.77 |
| B | 1 | 166.23 |
| B | 2 | 146.23 |



# Optimal policies

- A policy $\pi_*$ is optimal iff it maximizes the action-value function:

$$q_{\pi_*}(s,a) = \max_{\pi} q_{\pi}(s,a) = q_*(s,a)$$

- Thus all optimal policies share the same optimal value function
- Given the optimal value function, it is easy to act optimally:

$$\pi_*(s) = \arg\max_{a} q_*(s,a)$$

"greedification"

- We say that the optimal policy is greedy with respect to the optimal value function
- There is always at least one deterministic optimal policy

# Summary from first principles

- Intelligence is all about achieving *goals*
- Goals can be formulated as maximizing *reward*
  - e.g. expected cumulative discounted reward over time
- We maximize reward by finding and following an optimal policy $\pi_*$
- To find $\pi_*$ we need to first find the optimal value function $q_*$
- To find $q_*$ we need to repeatedly find the value function $q_\pi$ for a policy that is our current best guess at the optimal policy
- Thus, intelligence is all about estimating the value of the current policy!