# KI Brain Imaging in Neuroscience

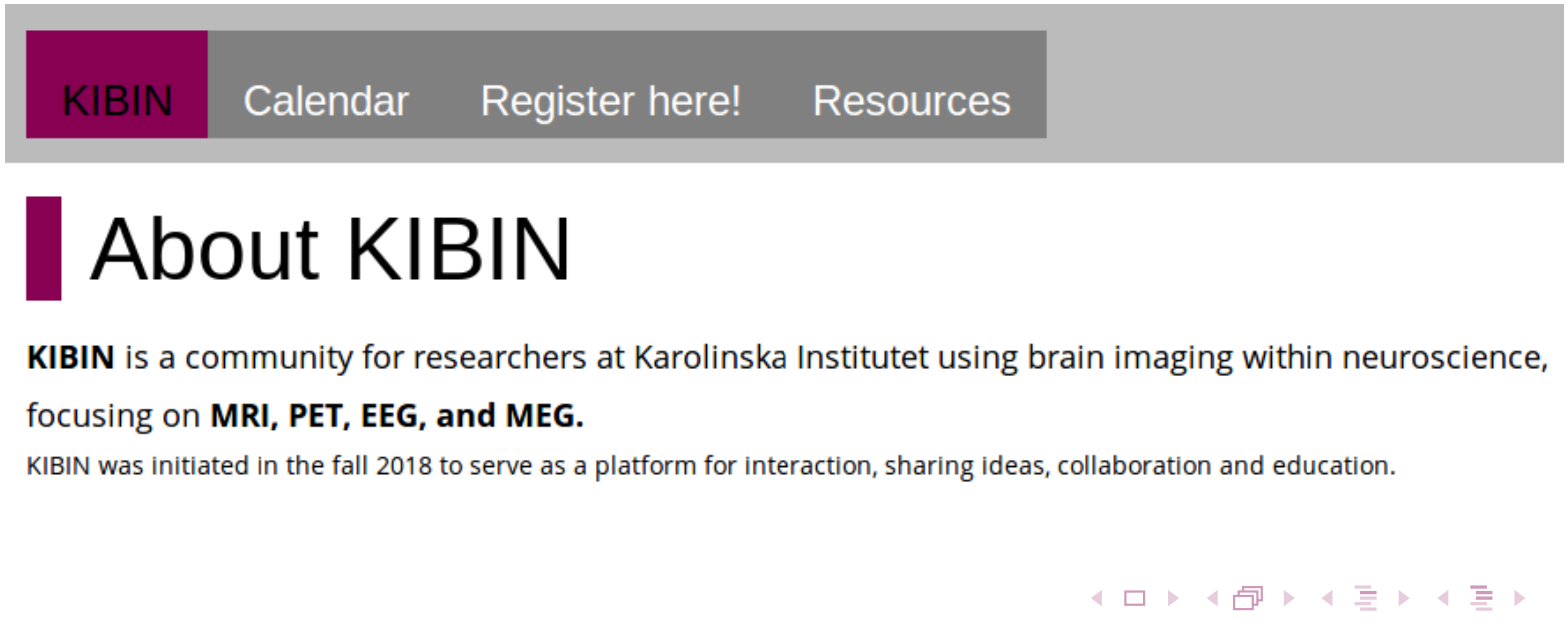
A community promoting interaction, sharing ideas, collaboration and education.

- Create a platform to share activities and resources.
- Provide a mailing list for announcements.
- Organize activities centered on methods issues.

# KI Brain Imaging in Neuroscience

- Web-page: www.kibin.se
- Email-list: KI-BIN@EMAILLIST.KI.SE

  To join:
  - register at: http://kibin.se/register.html
  - send an email to: register@kibin.se
- Calendar of activities: http://kibin.se/calendar.html
- Twitter: https://twitter.com/KarolinskaKIBIN



# Reinforcement learning in neuroscience - I

Rita Almeida

October 2018


Karolinska Institutet

# Reinforcement learning

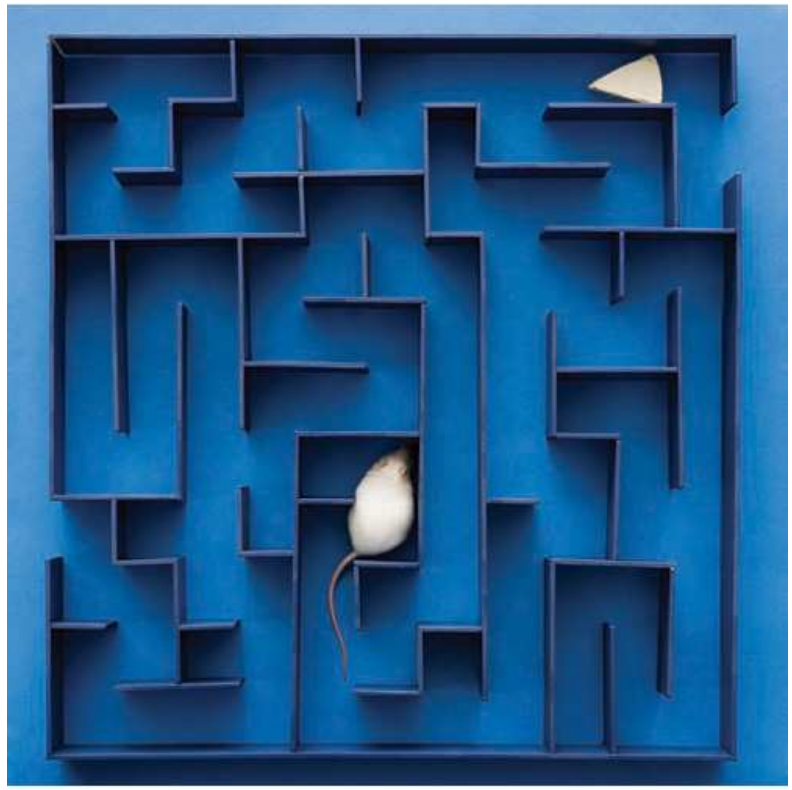
Simple definition:

- Learning by trial-and-error what to do in a given situation in order to maximize total reward and minimize total punishment.

- Goal-directed learning from interactions.

# Example

- Finding a short path in a maze.



# Example

- Learning what stimulus is associated with higher reward.

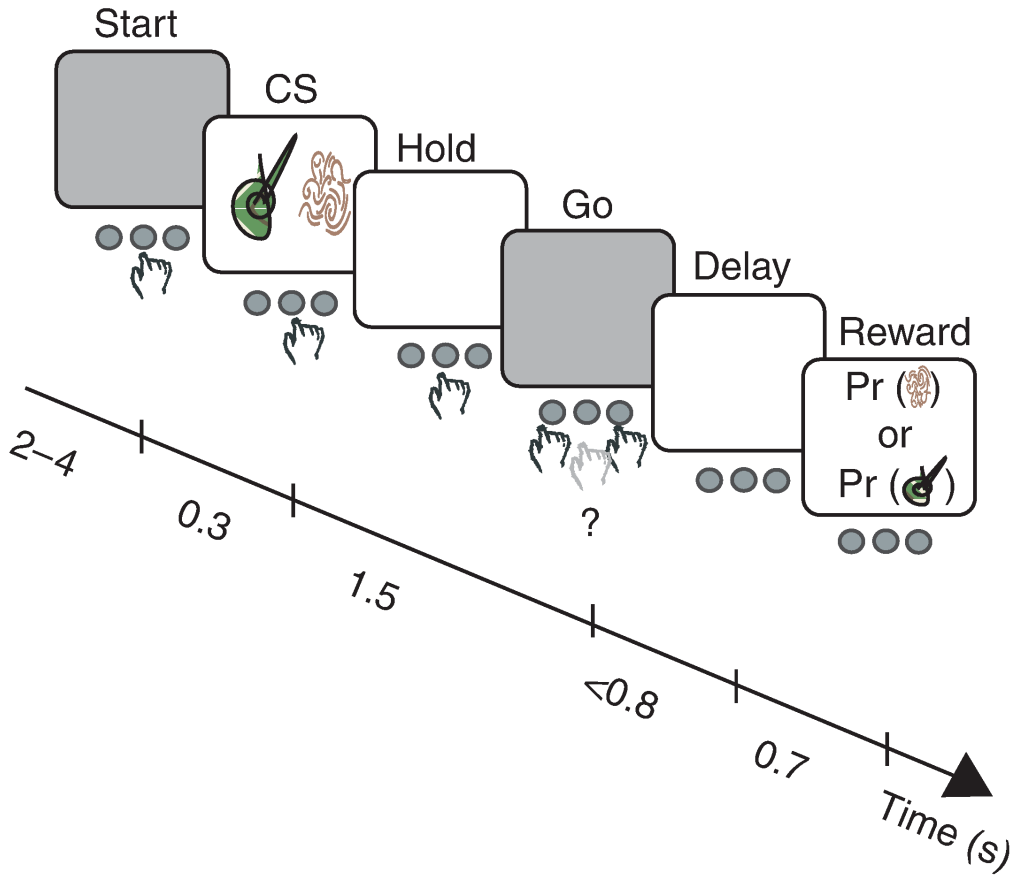



Figure adapted from Morris et al. 2005.

# Example

- Learning what sequence of stimulus choice is associated with higher reward.

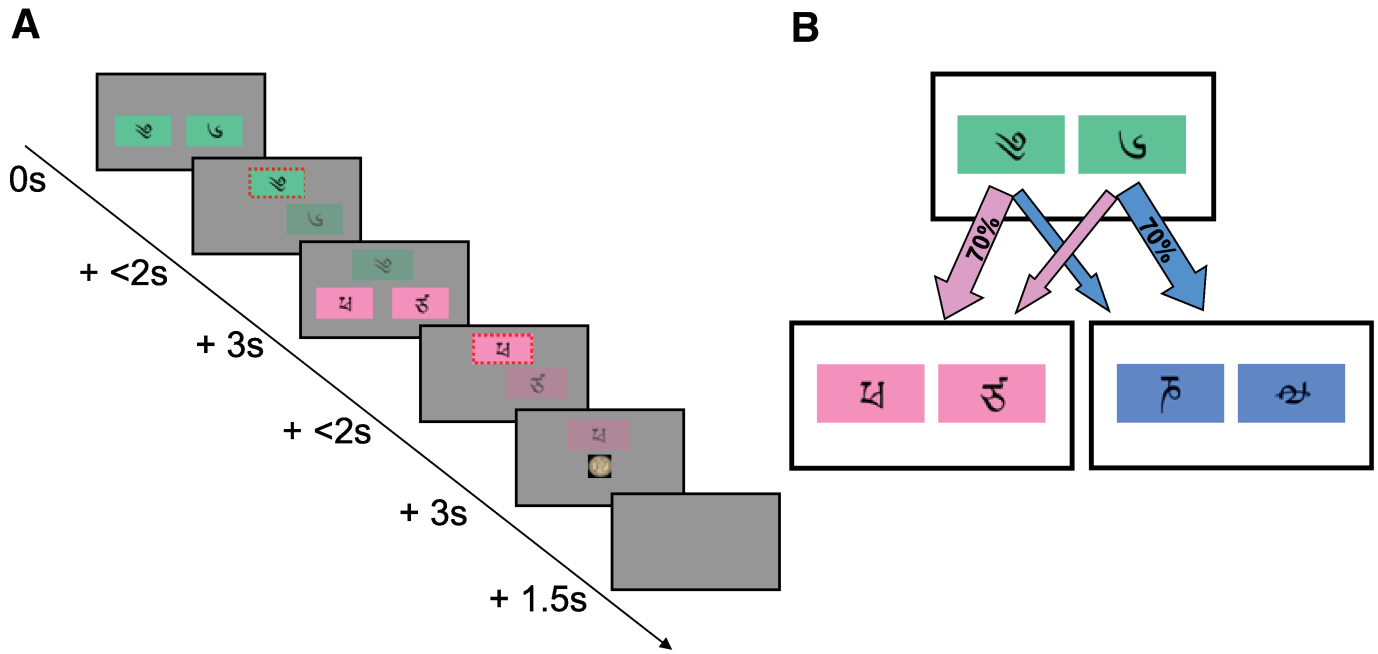
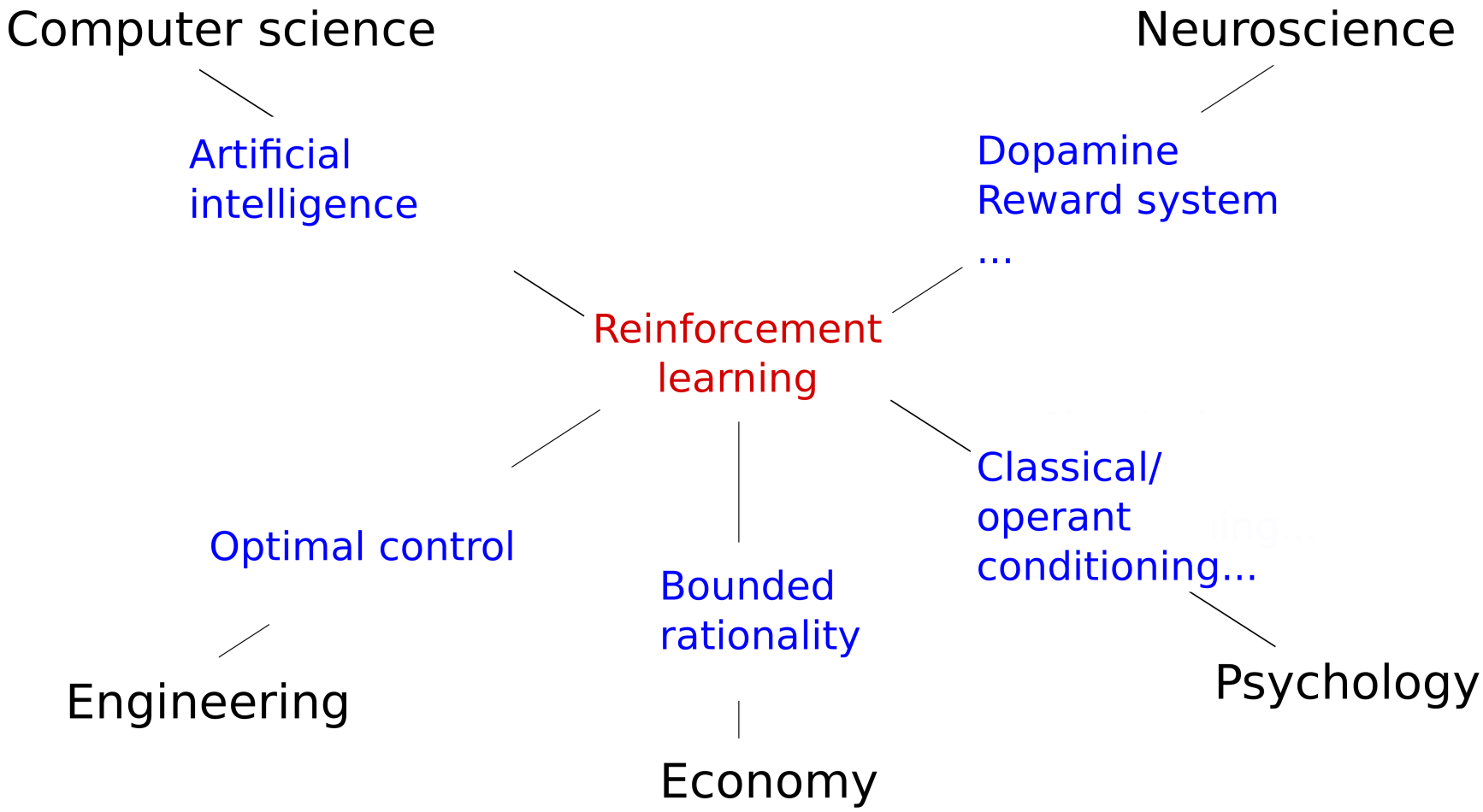
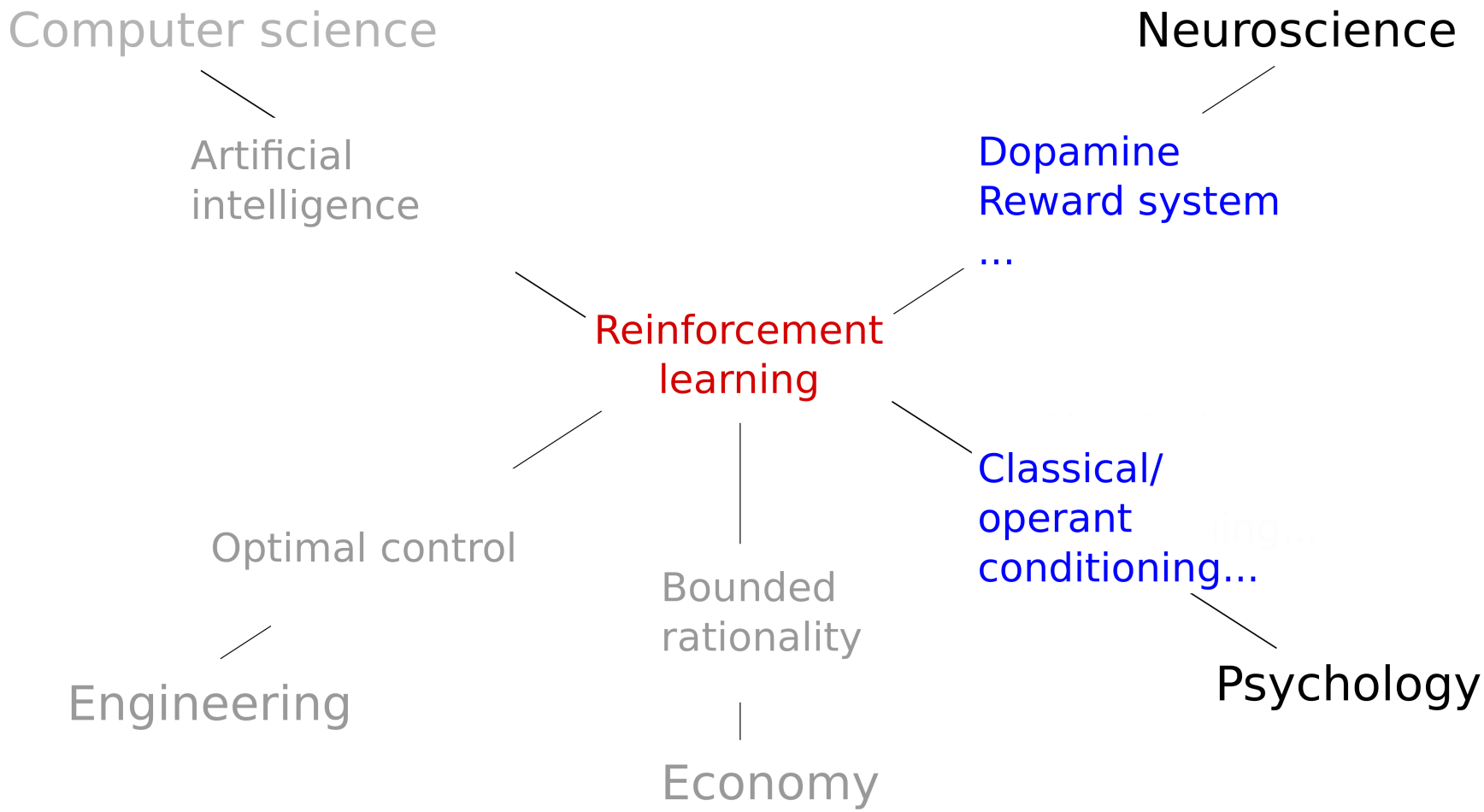



Figure adapted from Daw et al. 2011.

# Reinforcement learning in different fields



# Reinforcement learning in neurosience and psychology



# Learning

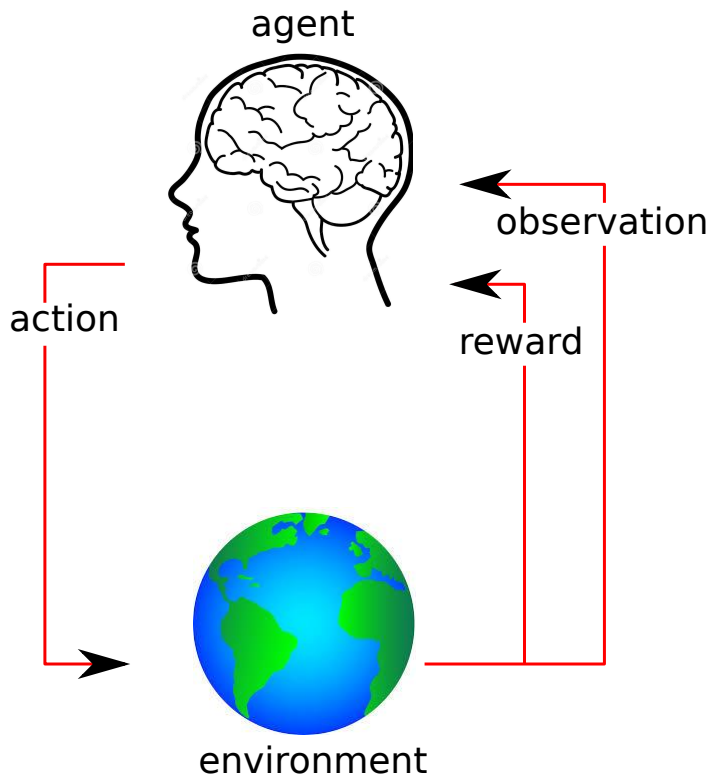
Neuroscience of learning:

- Mechanisms:
  - Activity dependent synaptic plasticity
  - Long term potentiation and depression (LTP, LTD)
  - Hebbian learning
  - Dopamine
  - Acetylcholine...
- Structures and circuits:
  - Basal ganglia...
- Cognitive neuroscience / psychology:
  - Implicit versus explicit
  - Associative versus non-associative
  - Classical and operant conditioning...

# Key aspects of RL

Learning problem where an agent interacts with the environment to achieve a goal.

- Sensation: observe the state of the environment
- Action: Take actions that affect the state of the environment
- Goal: Relating to reward



# Classical conditioning

## Pavlovian paradigm:

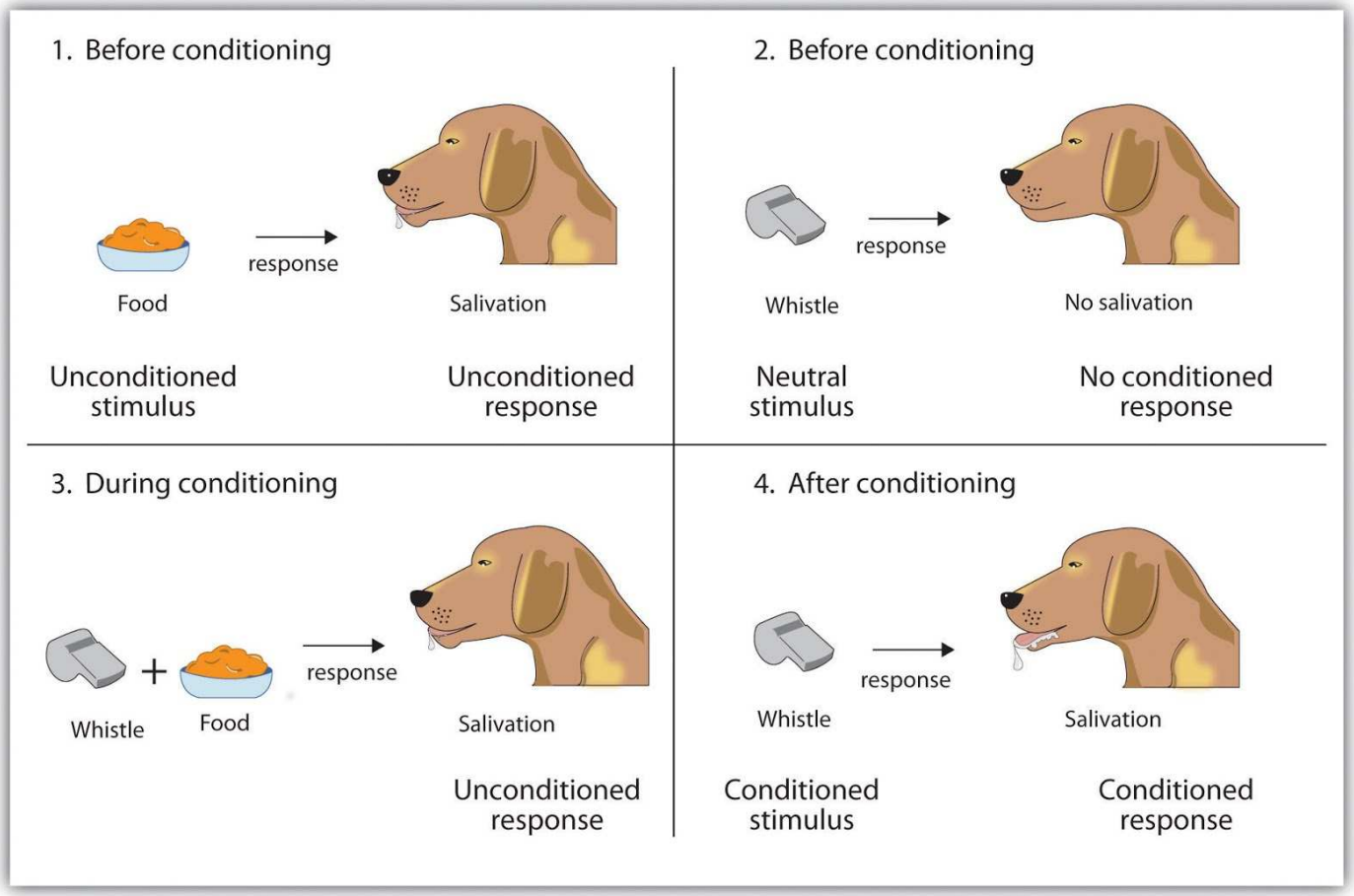


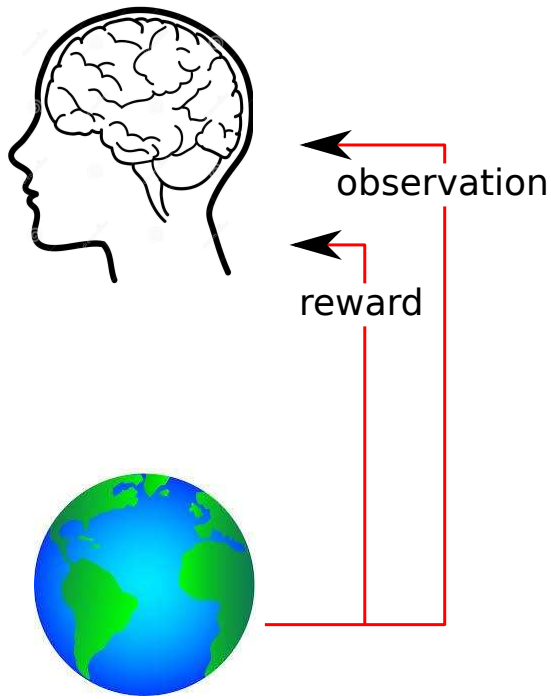

Figure from http://catalog.flatworldknowledge.com

- Does it relate with reinforcement learning?...

# Classical conditioning

- How does it relate with reinforcement learning?

- learning of values of stimuli through experience
- learning based on reward
- learning without instruction
- no acting!



# Rescorla-Wagner model of learning

- Learning happens when events are not predicted.

  Change in value is proportional to difference between actual and predicted outcome (prediction error).

For the Pavlovian paradigm:

$$V_{new}(S) = V_{old}(S) + \eta(r - V_{old}(S))$$

$$V(S) \leftarrow V(S) + \eta \quad \delta$$

- $S$ the conditioned stimuli - sound
- $r$ the unconditioned stimulus - food
- $\eta$ learning rate

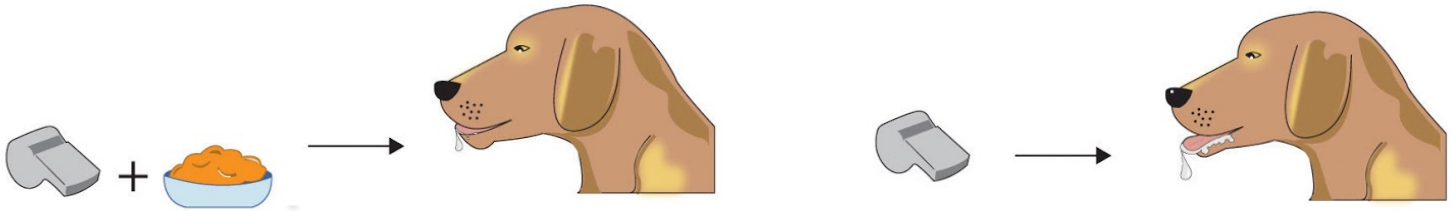
- Predictions due to different stimuli are summed linearly.

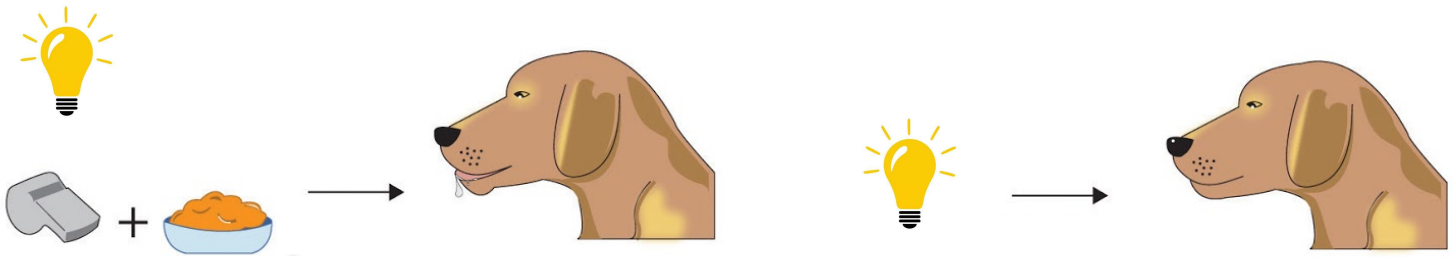
$$V_t(S_i) = V_{t-1}(S_i) + \eta(r - \sum_j V_{t-1}(S_j))$$

# Rescorla-Wagner explains other types of conditioning: Blocking

$$S_1 + r \quad \Longrightarrow \quad S_1 \to r'$$

$$S_1 + S_2 + r \quad \Longrightarrow \quad S_2 \to .'$$
$$S_1 \to r'$$

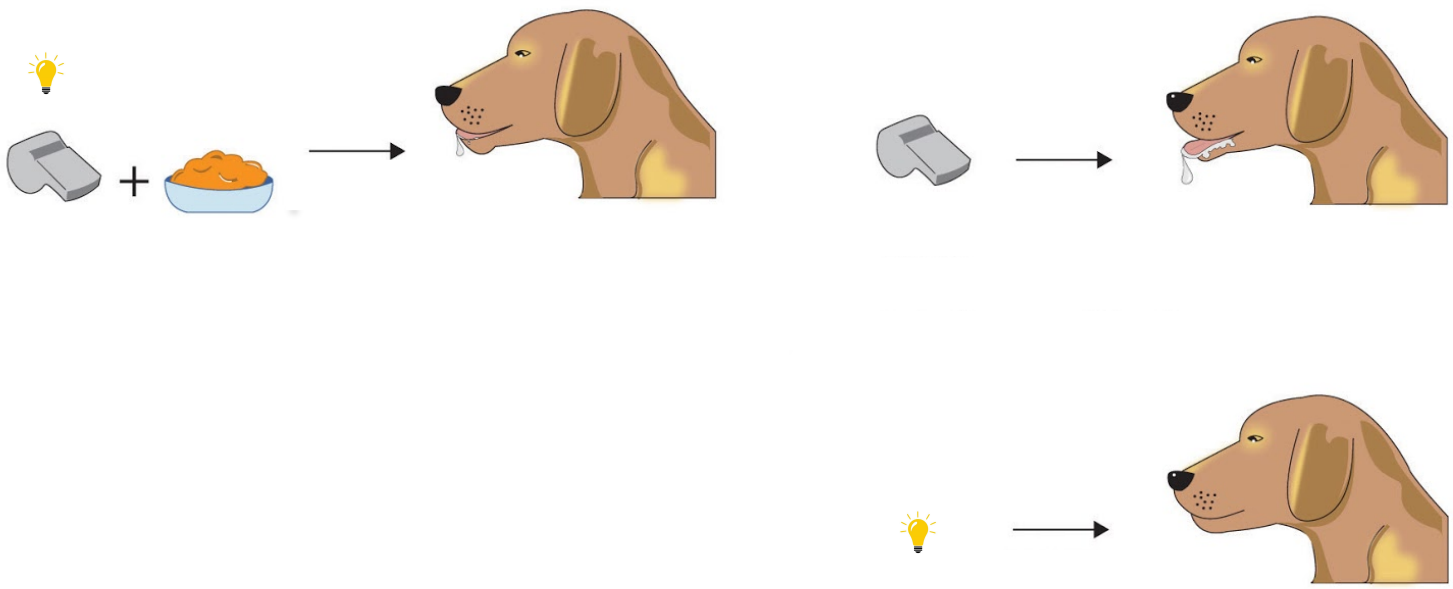




# Rescorla-Wagner explains overshadowing

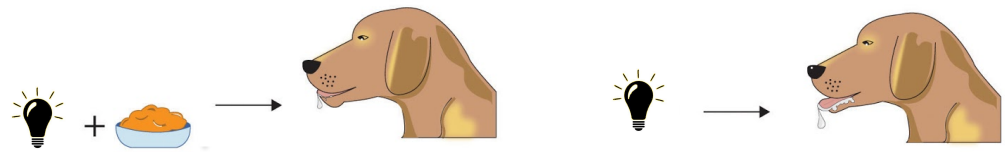
$$S_1 + S_2 + r \implies S_1 \to \alpha_1 r'$$

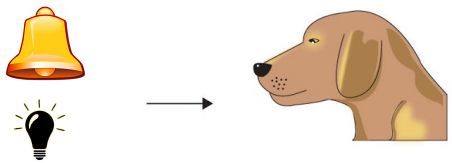
$$\implies S_2 \to \alpha_2 r'$$

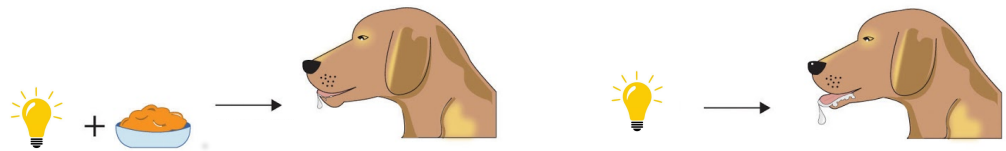


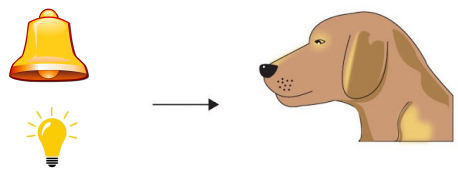
# Rescorla-Wagner explains inhibitory conditioning

$$S_1 + r \quad \Longrightarrow \quad S_1 \to r'$$
$$S_1 + S_x \to .$$
$$S_2 + r \quad \Longrightarrow \quad S_2 \to r' \quad \& \quad S_2 + S_x \to .'$$









# Rescorla-Wagner model of learning

The Rescorla-Wagner model explains other types of conditioning:

- blocking
- overshadowing
- inhibitory

The Rescorla-Wagner model also has shortcomings:

- does not account for the sensitivity of conditioning to temporal contingencies,
- it does not explain second order conditioning ($S_1 \to r \quad S_2 \to S_1 \to r \quad S_2 \to r'$),
- it does not explain extinction of inhibition

# Temporal difference learning

Idea of RL: Maximize total reward, not only immediate reward.

- Predict all future reward.
- Total reward depends on a sequence of choices.

Value of a sate $S$ is the expected future reward. Given $S_t = s$ :

$$V(s) = E[r_{t+1} + \gamma r_{t+2} + \gamma^2 r_{t+3} + ... | S_t = s]$$

where $\gamma \leq 1$ discounts the effect of rewards distant in time and $t$ is time or step within a trial.

And in a recursive form:

$$V(s) = E[r_t | S_t = s] + \gamma E[V(s_{t+1}) | S_t = s].$$

# Temporal difference learning

Then, a prediction error can be defined as:

$$\delta = E[r_t|S_t] + \gamma E[V(S_{t+1})|S_t] - V(S_t).$$

If $\delta$ is estimated it can be used in the update rule:

$$NewEstimate \leftarrow OldEstimate + LearningRate[Target - OldEstimate]$$

Applied to the value:

$$V(s) \leftarrow V(s) + \eta \quad \delta$$

# Temporal difference learning

Problem: to calculate $E[\quad]$ one needs $P(r|s_t = s)$ and $P(s_{t+1}|s_t = s)$.

- This knowledge is usually not available.
- Information can be accumulated by sampling.

The current values $r_{t+1}$ and $V_t(s_{t+1})$ can be used:

$$V(s_t) \leftarrow V(s_t) + \eta(r_{t+1} + \gamma V(s_{t+1}) - V(s_t))$$

And the temporal difference prediction error is:

$$\delta = r_{t+1} + \gamma V(s_{t+1}) - V(s_t)$$

- Over time total expected values of events can be learned even in stochastic environments with unknown dynamics.

# Operant or instrumental conditioning

Introducing actions

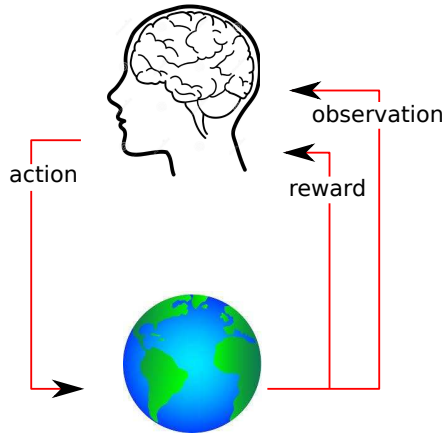



Figure from Scientific American.

Animals will behave in order to get reward.

# Reinforcement learning

- Learning by trial-and-error which action to choose.
- Learning values of actions for a given state.
- Policy for behavior.

# Elements of RL

Policy: mapping between perceived states and actions.

- Can be stochastic.

Reward function: mapping between state (or state-action) to a value summarizing the desirability of that state.

- Agent wants to maximize the reward.
- Directly given by environment.
- Can be stochastic.

Value function: mapping between state to the long term desirability of that state.

- Takes into account total amount of reward the agent expects to accumulate from that state on.
- Action choices are made based on value.
- Value is estimated from all observations the agent does.

Model of the environment: (optional) a model of the behavior of the environment.

# Action selection: exploration vs exploitation

Learning requires a trade-off between:

- Exploitation - agent exploits what is known and chooses actions with greatest estimated values (greedy actions).
- Exploration - agent explores actions in order to make better future selections. Assures that:
  - the agent does not get stuck with a good but not optimal action.
  - the expected reward is properly estimated in a stochastic task.
  - the agent adapts when the task is non-stationary.

# Example - *n*-armed bandit problem

- Repeated plays - action selections.
- Each play *n* possible actions.
- After an action a reward is received.
- Each action is associated with a stationary probabilistic reward.
- Aim is to achieve maximal reward over a number of plays.

# Estimating action-values $Q(a)$

- Sample-average method: estimating $Q(a)$ as the average reward previously received when choosing $a$.

  If at step $t$ action $a$ has been chosen $k_a$ times:

$$Q_t(a) = \frac{r_1 + r_2 + \cdots + r_{k_a}}{k_a}$$

- Incremental method: updating the estimated $Q(a)$ after each step.

  If $Q_k$ is the average of the first $k$ rewards

$$Q_{k+1} = Q_k + \frac{1}{k+1}(r_{k+1} - Q_k)$$

# Update rule

$$Q_{k+1} = Q_k + \frac{1}{k+1}(r_{k+1} - Q_k)$$

Generic update rule:

$$NewEstimate \leftarrow OldEstimate + LearningRate[Target - OldEstimate]$$

- Similar rules are common in RL.
- $[Target - OldEstimate]$: error of the estimate
- *LearningRate*: size of update

# Softmax action selection

- Idea: probability of choosing an action is proportional to its estimated value

Action $a$ is chosen on play $t$ with probability:

$$\frac{e^{Q_t(a)/\beta}}{\sum_{b=1}^{n} e^{Q_t(b)/\beta}}$$

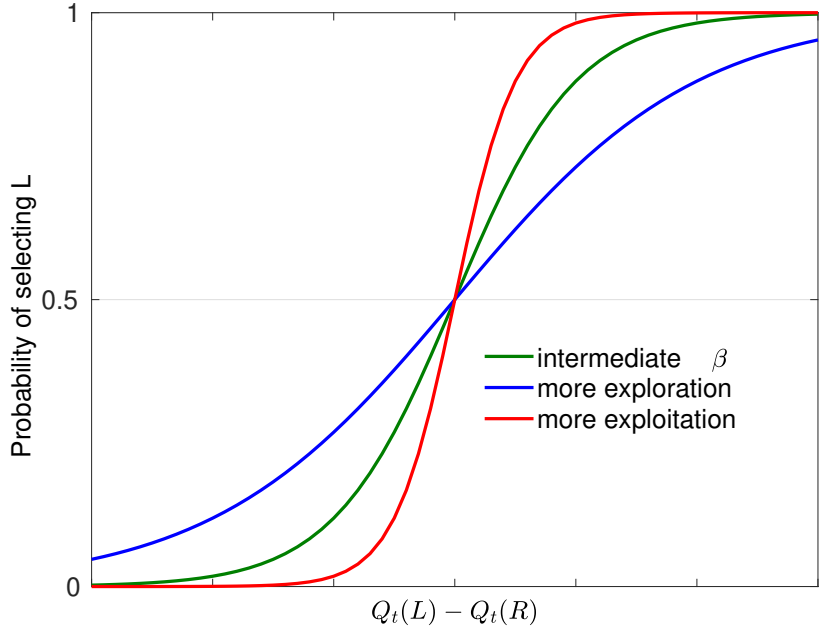
- $\beta \geq 0$ is the temperature
- large $\beta$: all actions are chosen with similar probabilities
- $\beta \to 0$: greedy action selection

# 2-armed bandit

Example: Subject chooses between left or right (L or R).

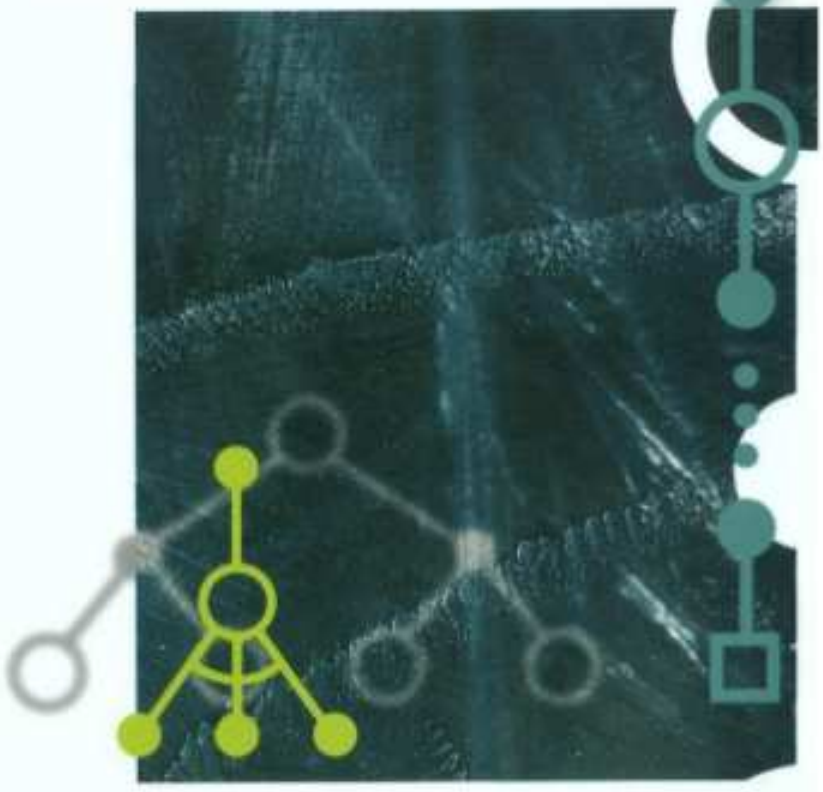
$$P(L)_t = \frac{e^{Q_t(L)/\beta}}{e^{Q_t(L)/\beta} + e^{Q_t(R)/\beta}} \qquad P(R)_t = \frac{e^{Q_t(R)/\beta}}{e^{Q_t(L)/\beta} + e^{Q_t(R)/\beta}}$$

$$P(L)_t = \frac{1}{1 + e^{-(Q_t(L) - Q_t(R))/\beta}}$$



# Reinforcement Learning

An Introduction



Richard S. Sutton and Andrew G. Barto