Action learning

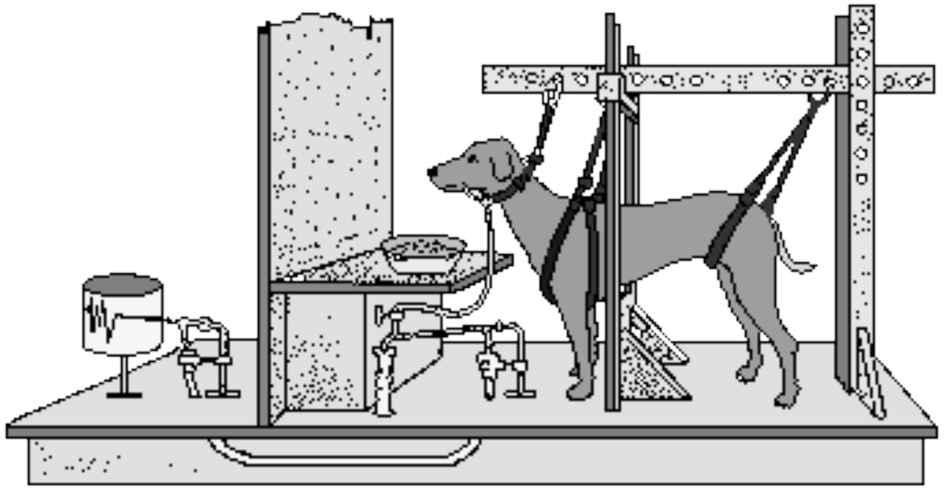
Instrumental conditioning

David G. Nagy - neural modelling - 2023

MAX PLANCK GESELLSCHAFT

EBERHARD KARLS UNIVERSITÄT TÜBINGEN

classical conditioning



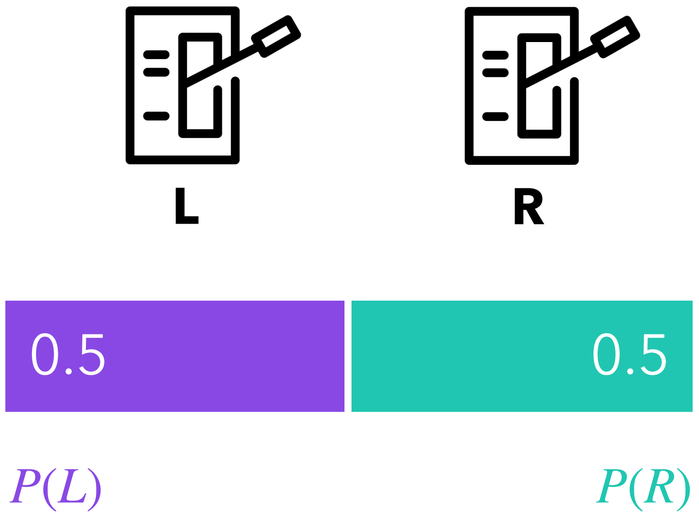
instrumental conditioning

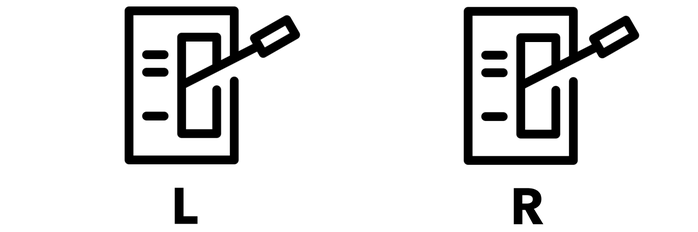
instrumental conditioning

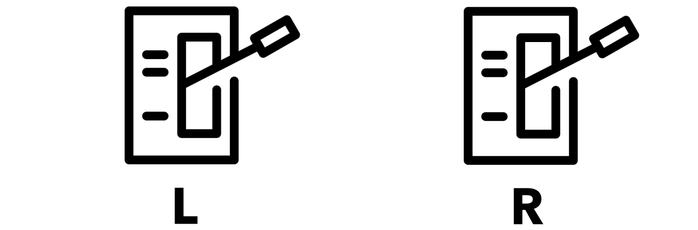
- immediate choice
- evolutionary programming
- sequential decisions
- vigour

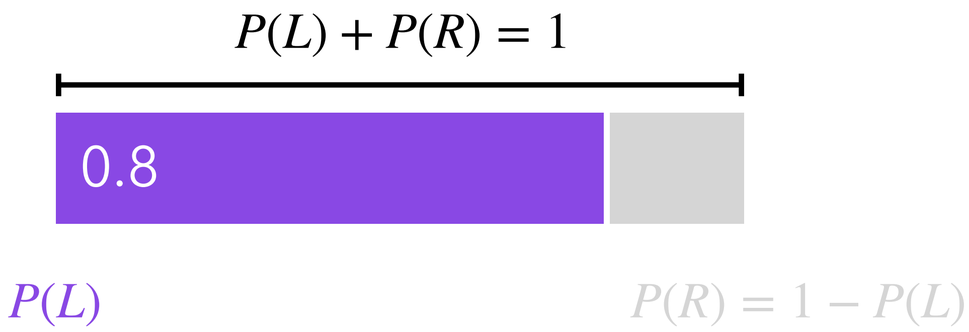
instrumental conditioning

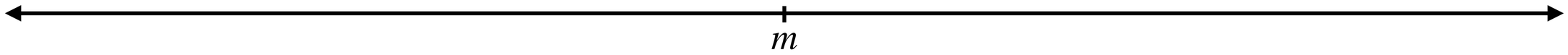
- **immediate choice**
- evolutionary programming
- sequential decisions
- vigour


L
R
0.5
0.5
$P(L)$
$P(R)$


L
R
1
0
$P(L)$
$P(R)$


L
R
0.8
0.2
$P(L)$
$P(R)$


$P(L) + P(R) = 1$
0.8
$P(L)$
$P(R) = 1 - P(L)$



$P(L)$ $P(R) = 1 - P(L)$

$P(L \mid m)$

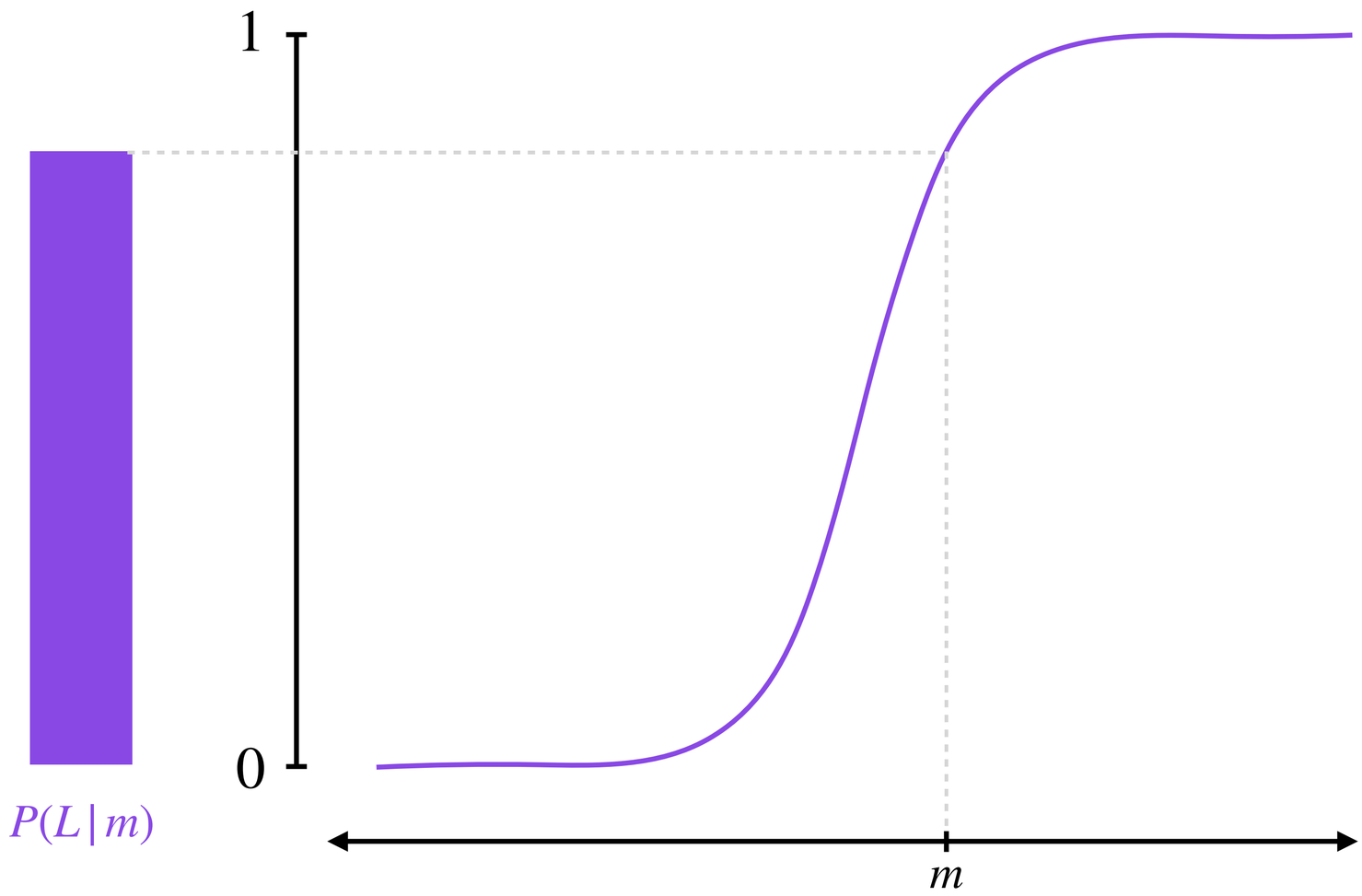
$P(R \mid m) = 1 - P(L \mid m)$

$P(L \mid m)$

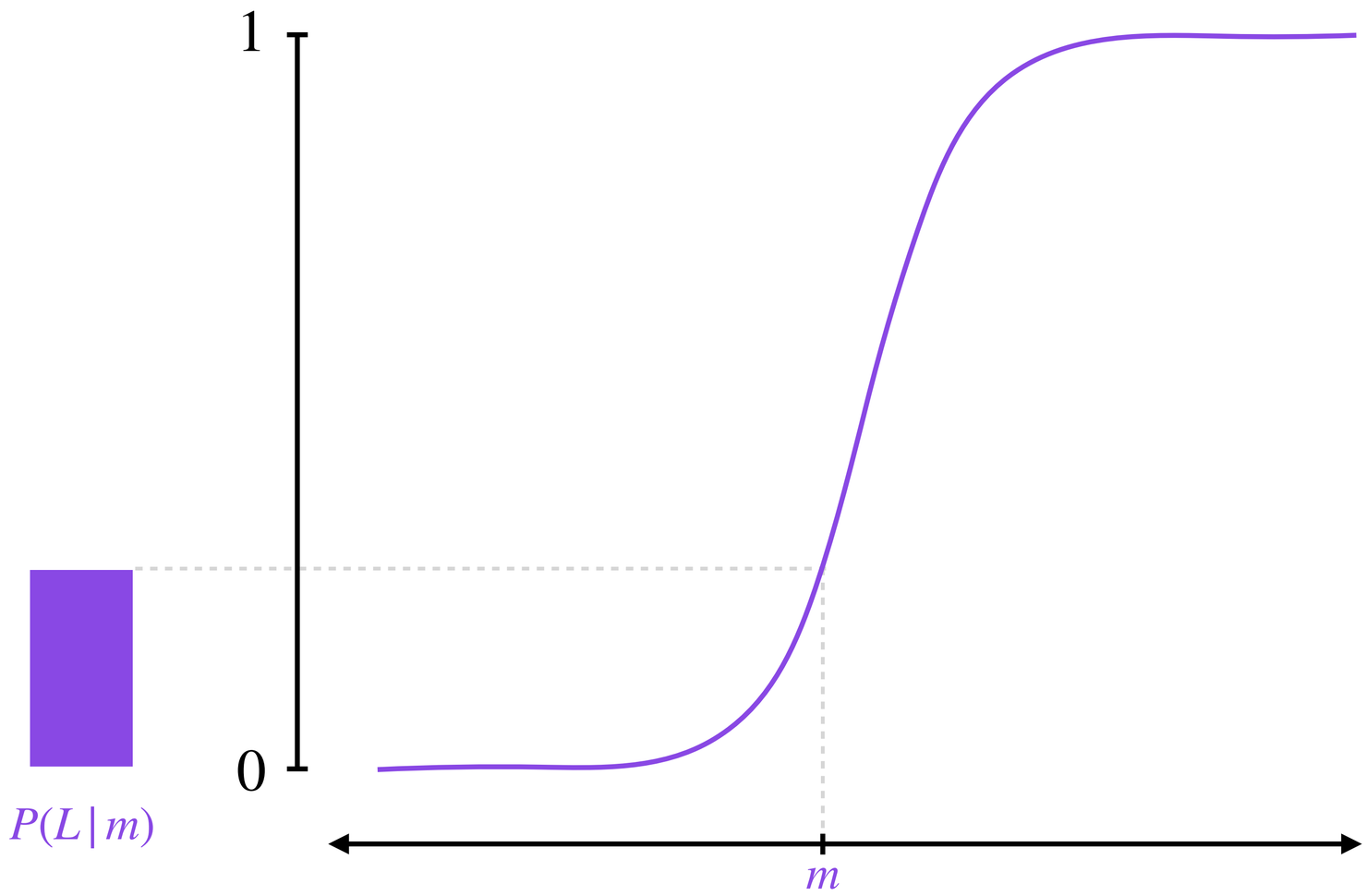
$P(R \mid m) = 1 - P(L \mid m)$

$P(L \mid m)$

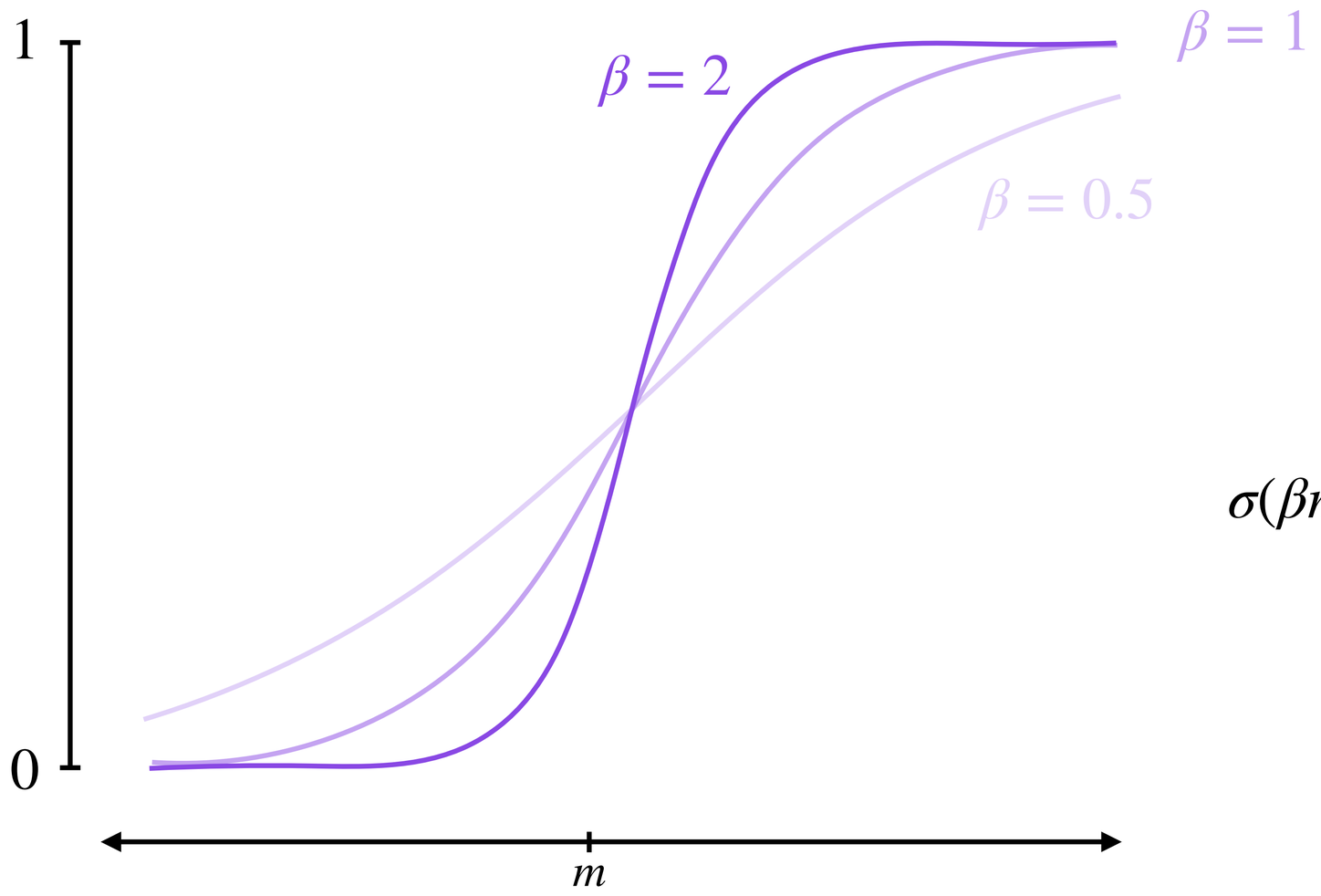
$P(R \mid m) = 1 - P(L \mid m)$



$$\sigma(m) = \frac{e^m}{1 + e^m}$$

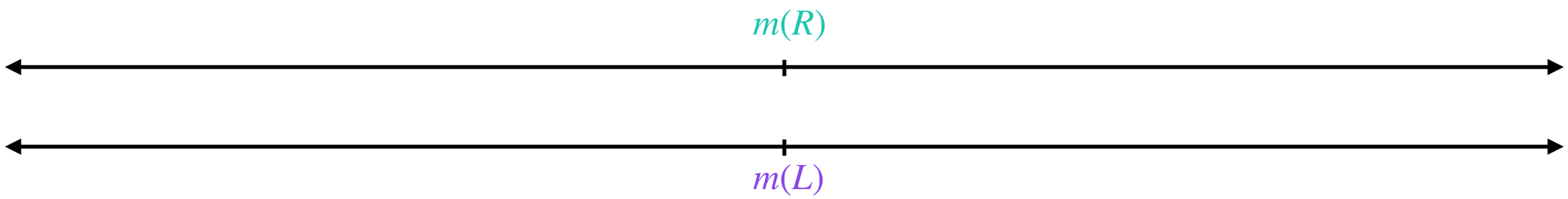


$$\sigma(m) = \frac{e^m}{1 + e^m}$$



$$\sigma(\beta m) = \frac{e^{\beta m}}{1 + e^{\beta m}}$$

exploration

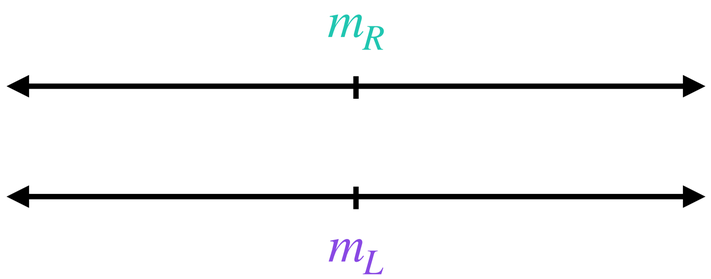


$$P(L) = \frac{e^{\beta m_L}}{e^{\beta m_R} + e^{\beta m_L}}$$

$$P(R) = \frac{e^{\beta m_R}}{e^{\beta m_R} + e^{\beta m_L}}$$

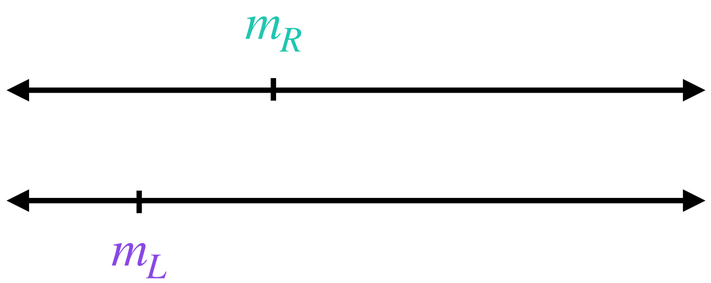
$$\text{softmax}(\mathbf{m})_i = \frac{e^{\beta m_i}}{\sum_{j=1}^{N} e^{\beta m_j}}$$

how should we set m?



indirect actor

$$m_R = \mathbb{E}[r \,|\, a = R] = \langle r_R \rangle$$

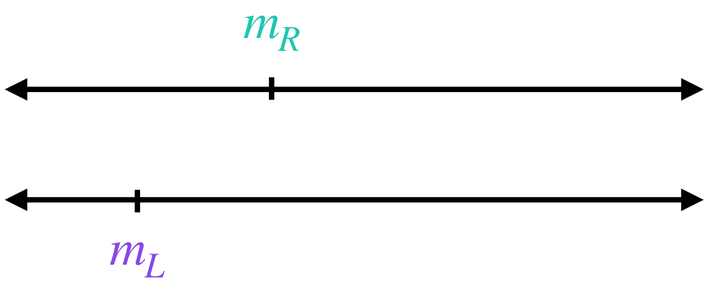


$$m_L = \mathbb{E}[r \,|\, a = L] = \langle r_L \rangle$$

indirect actor

$Q(a = R)$

$m_R = \mathbb{E}[r \mid a = R] = \langle r_R \rangle$



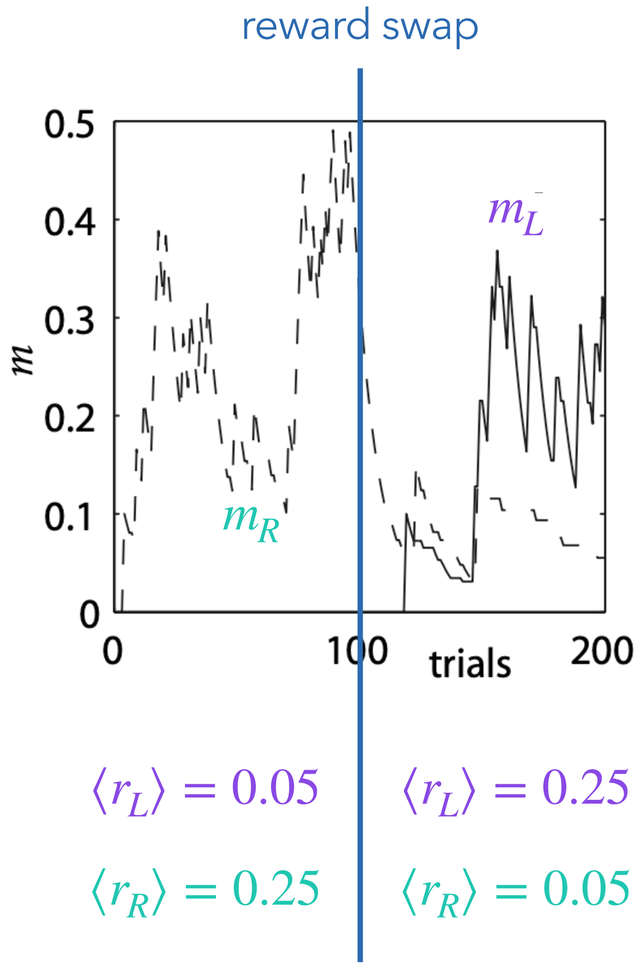
$m_L = \mathbb{E}[r \mid a = L] = \langle r_L \rangle$

update rule:

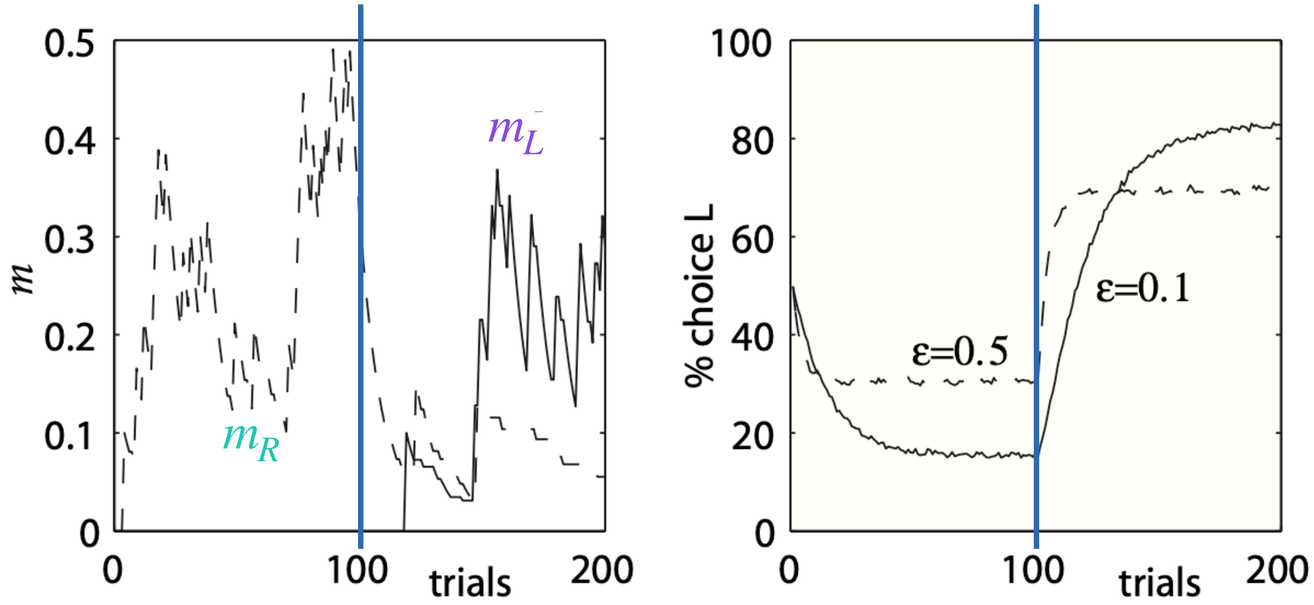
$$m_L \to m_L + \epsilon(r_L - m_L)$$

$$m_R \to m_R + \epsilon(r_R - m_R)$$

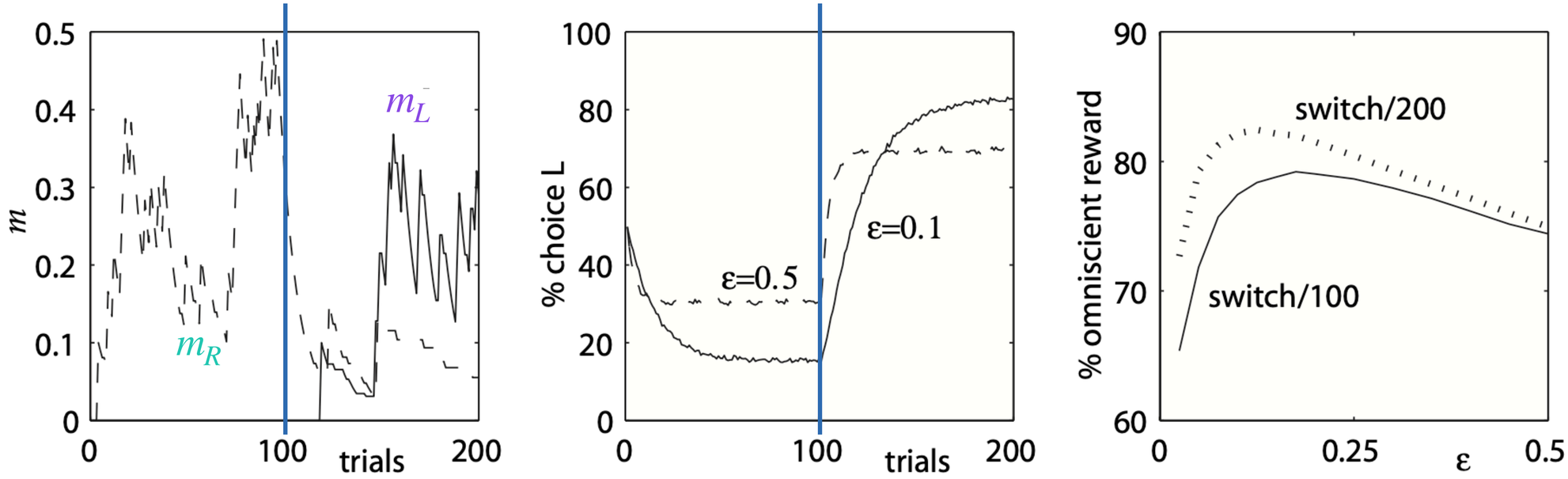
indirect actor



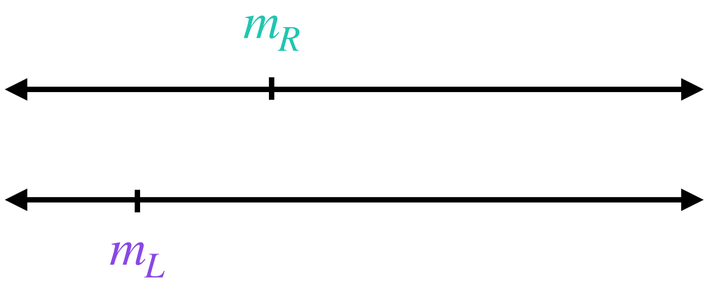
indirect actor



indirect actor



direct actor



$$\mathbb{E}[r \,|\, \mathbf{m}] = \langle r \rangle$$

$$= \mathbb{E}_{a \sim P(a|\mathbf{m})}[r_a]$$

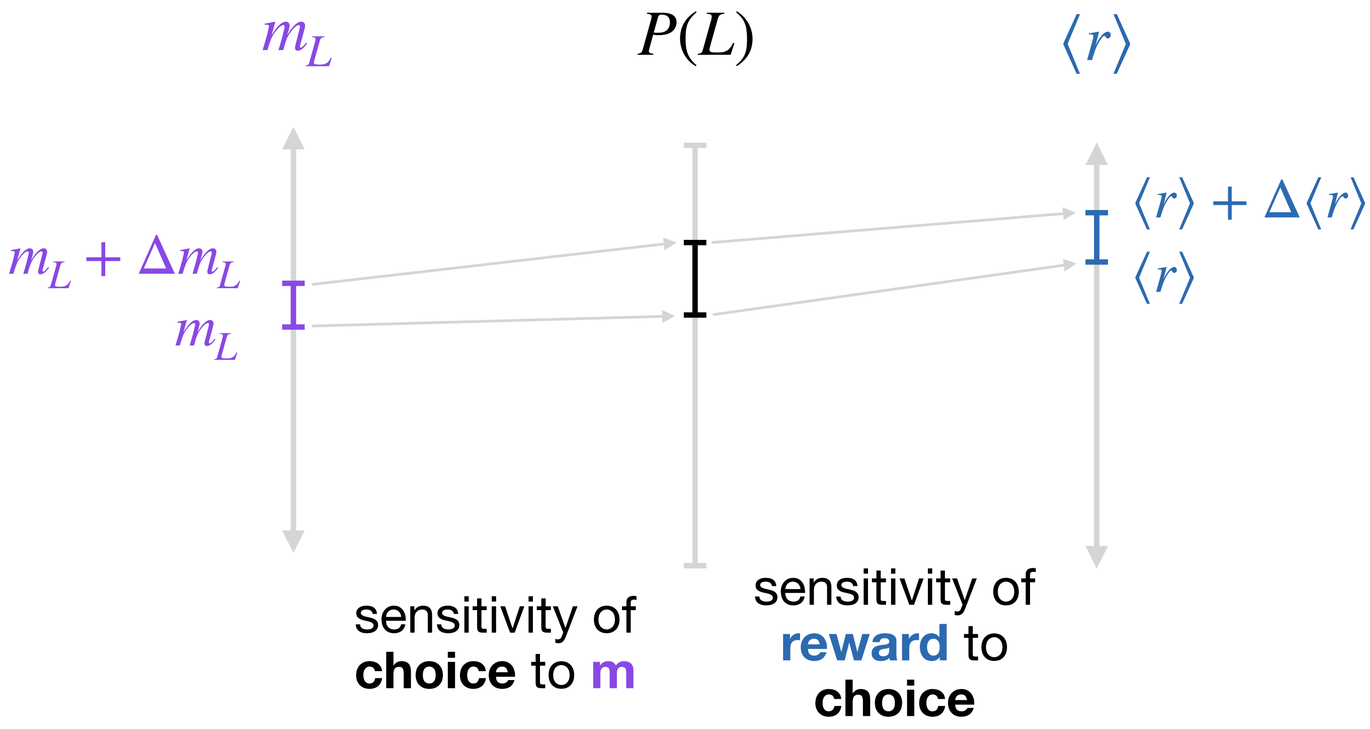
$$= P(L)\langle r_L \rangle + P(R)\langle r_R \rangle$$

update rule:

$$m_L \to m_L + \epsilon \, \Delta m_L$$

$$\mathbb{E}[\Delta m_L] \propto \frac{\partial \langle r \rangle}{\partial m_L}$$

direct actor



$$\frac{\partial P(L)}{\partial m_L} = \beta P(L)(1 - P(L))$$

$$\frac{\partial P(R)}{\partial m_L} = -\beta P(R)P(L)$$

$$\frac{\partial \langle r \rangle}{\partial m_L} = \beta P(L)(\langle r_L \rangle - \langle r \rangle)$$

direct actor

for update rule:

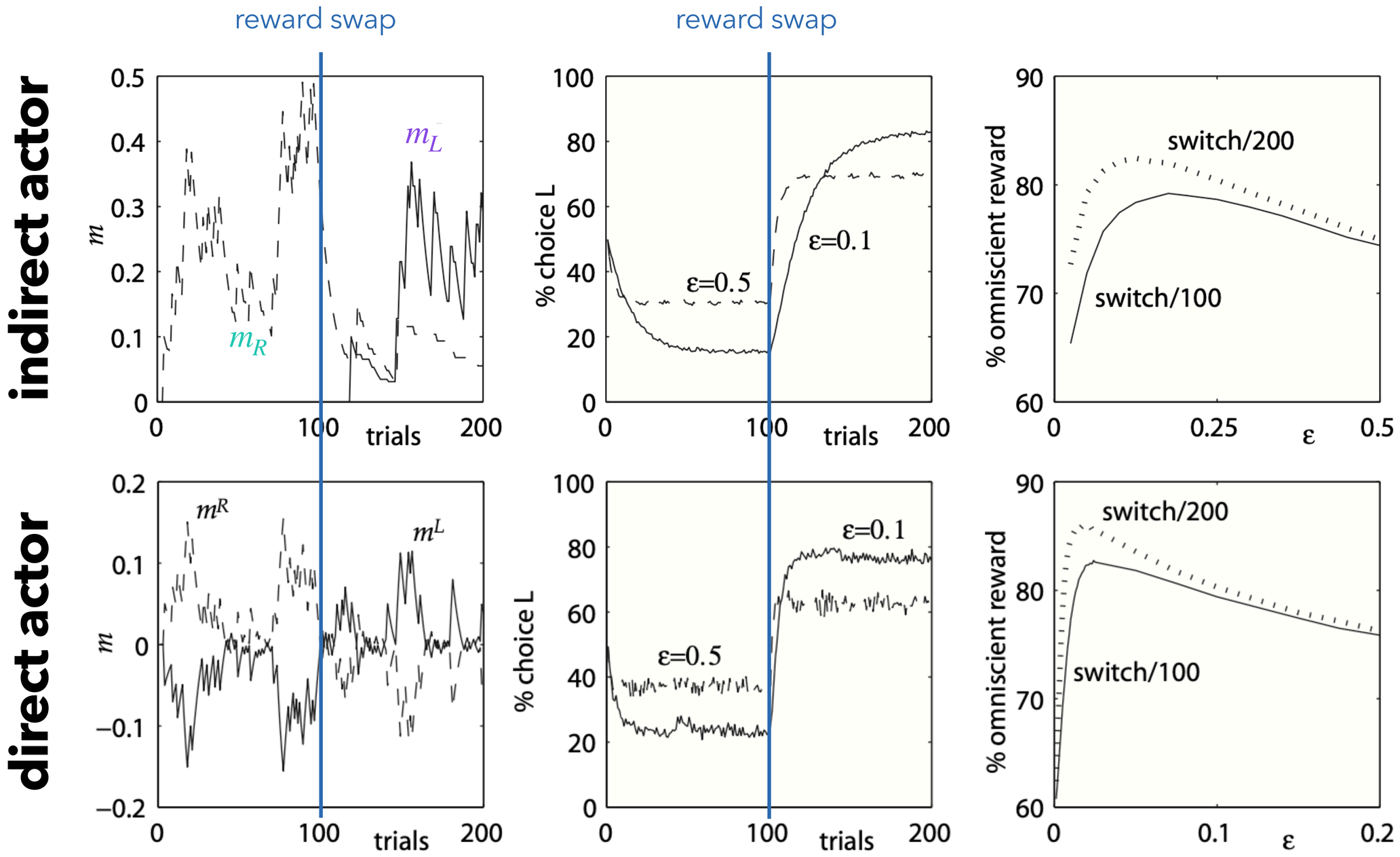
$$m_L \to m_L + \epsilon\, \Delta m_L$$

$$\mathbb{E}[\Delta m_L] \propto \frac{\partial \langle r \rangle}{\partial m_L} \quad \longleftarrow \quad \frac{\partial \langle r \rangle}{\partial m_L} = \beta P(L)(\langle r_L \rangle - \langle r \rangle)$$

$$m_L \to m_L + \epsilon(r_L - v) \;\mid\; a = L$$

decay + multiple actions

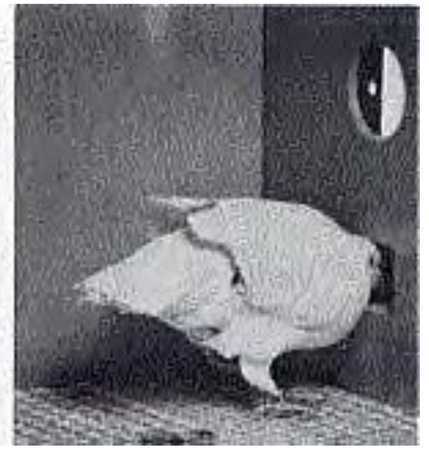
$$m_b \to (1-\epsilon) m_L + \epsilon \delta_{ab}(r_a - v)$$


indirect actor
direct actor
reward swap
reward swap
m_L
m_R
m^R
m^L
m
trials
% choice L
ε=0.5
ε=0.1
% omniscient reward
switch/200
switch/100
ε

instrumental conditioning

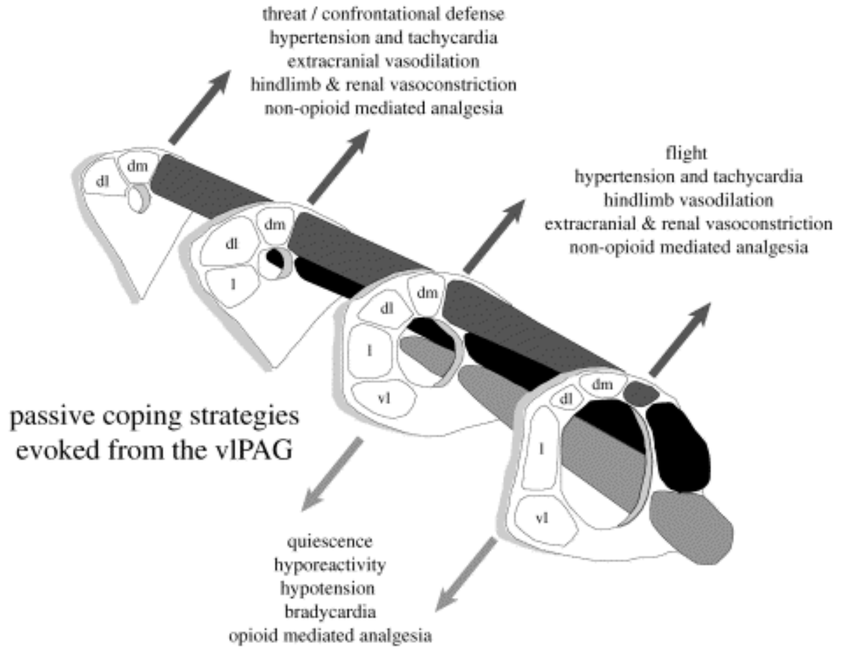
- immediate choice
- **evolutionary programming**
- sequential decisions
- vigour

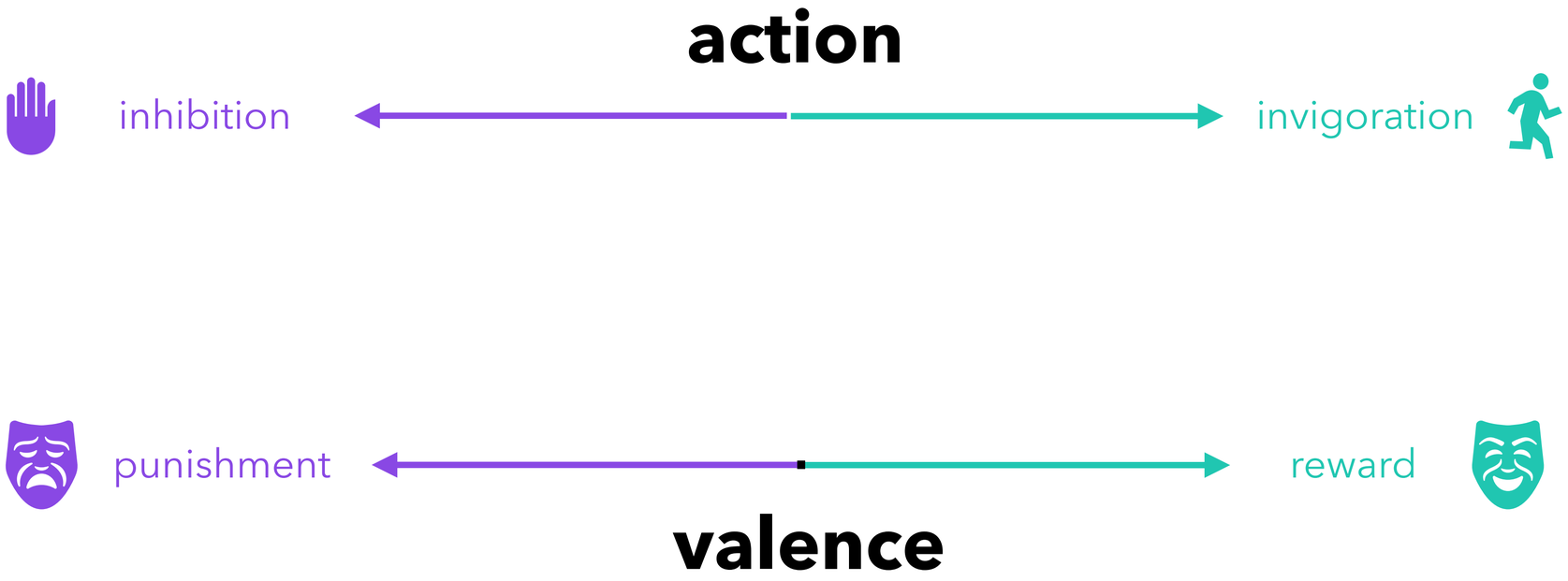
The Skinnerian Pigeon

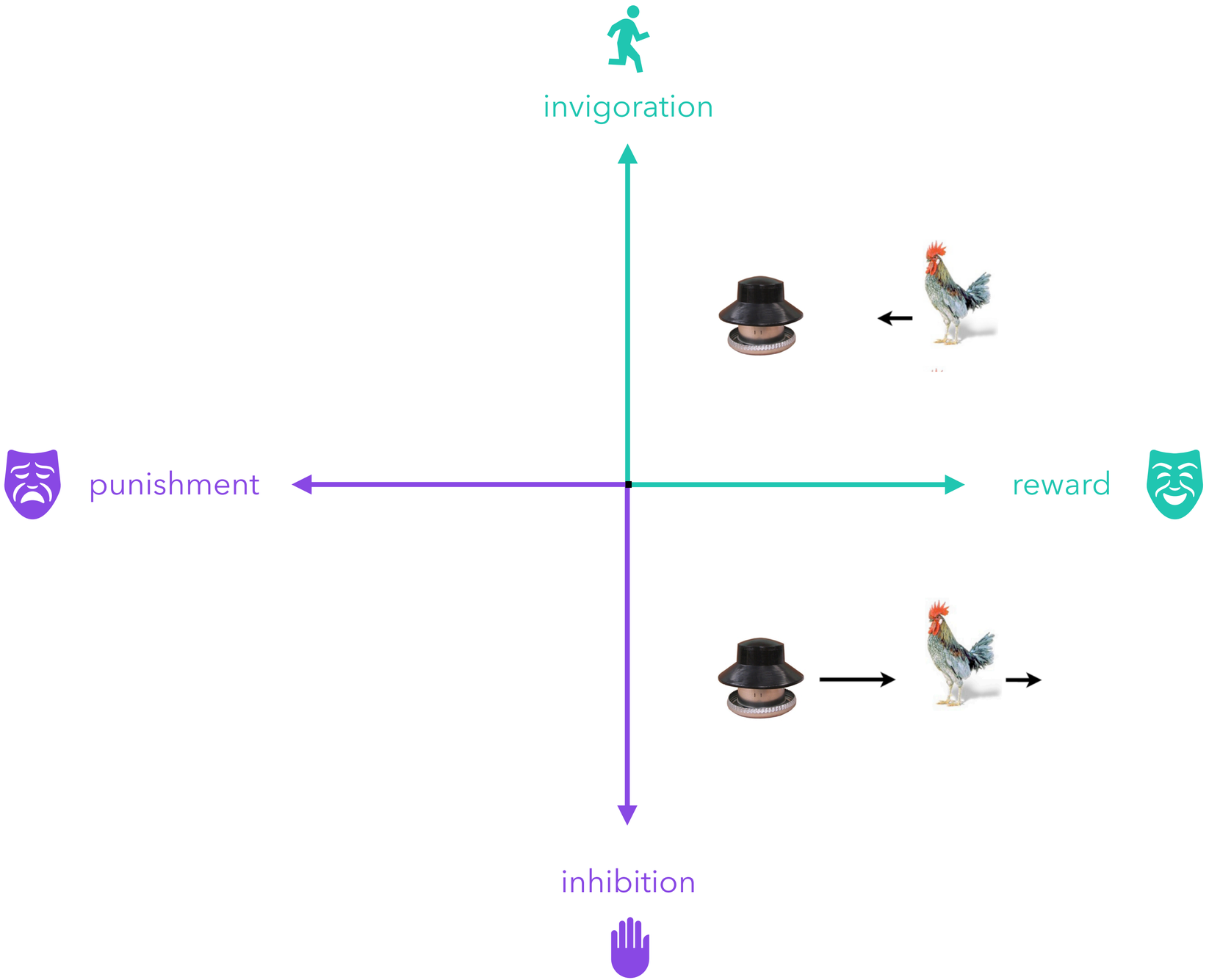


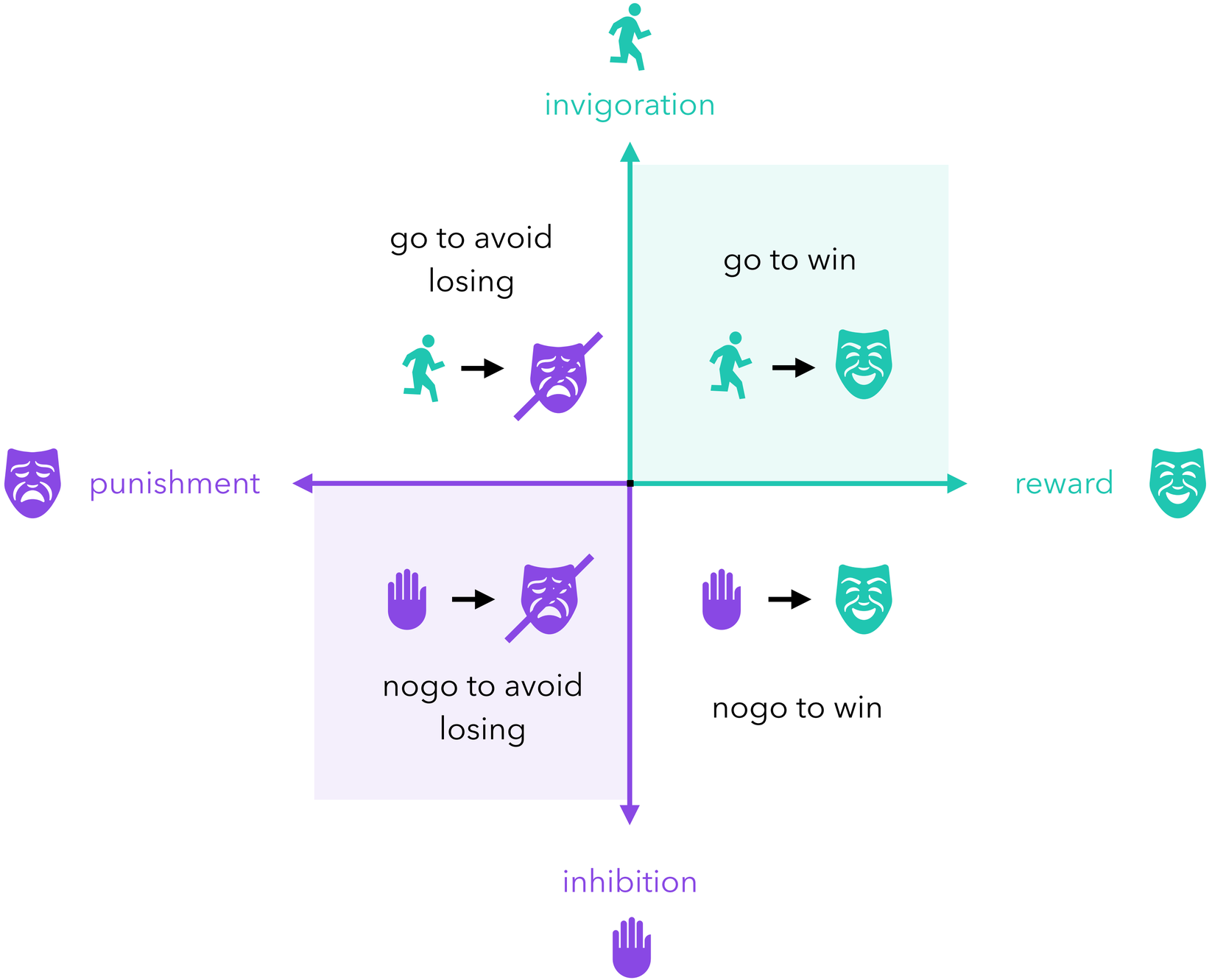
The Hapless Pigeon

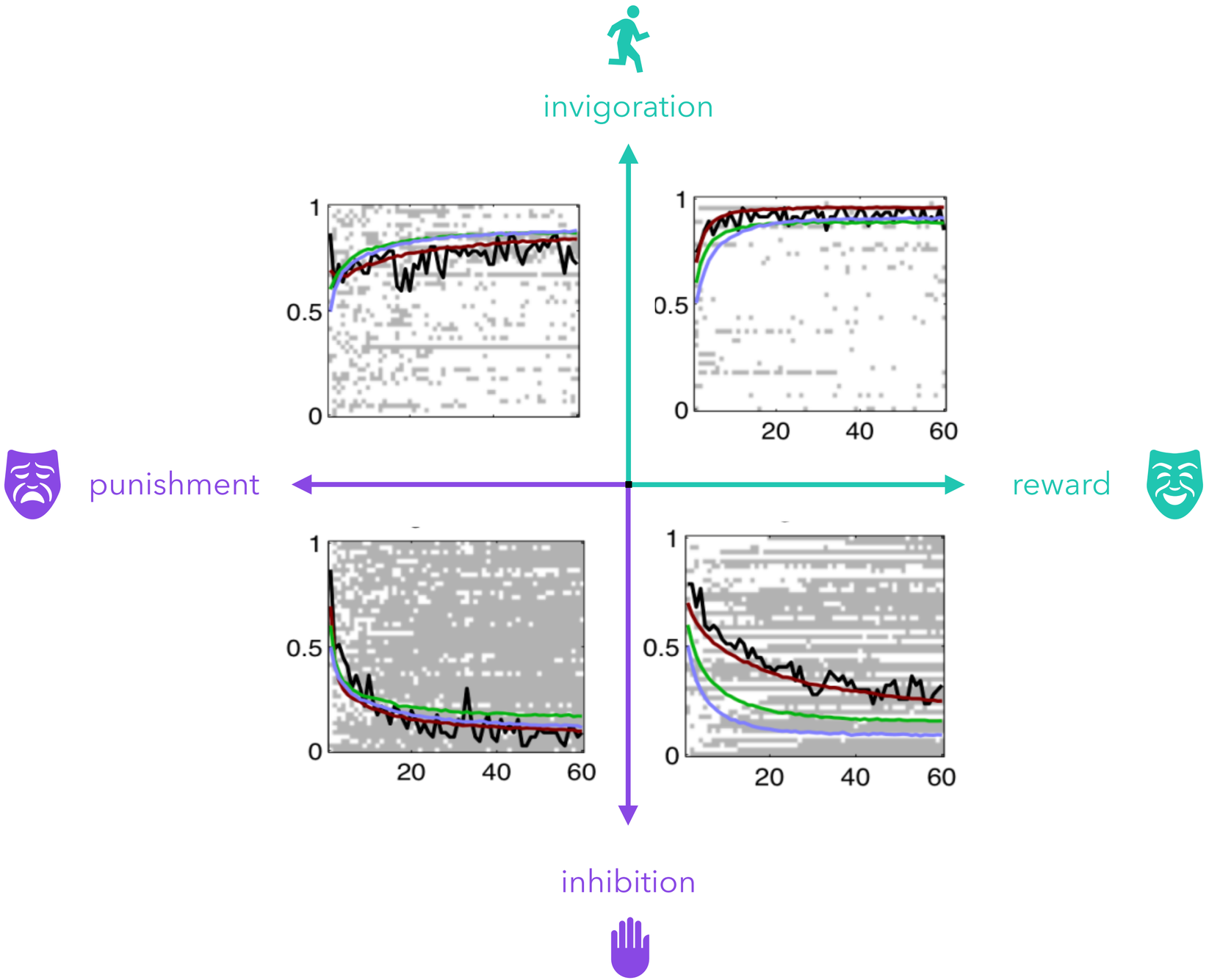
active coping strategies evoked
from the lPAG and the dlPAG

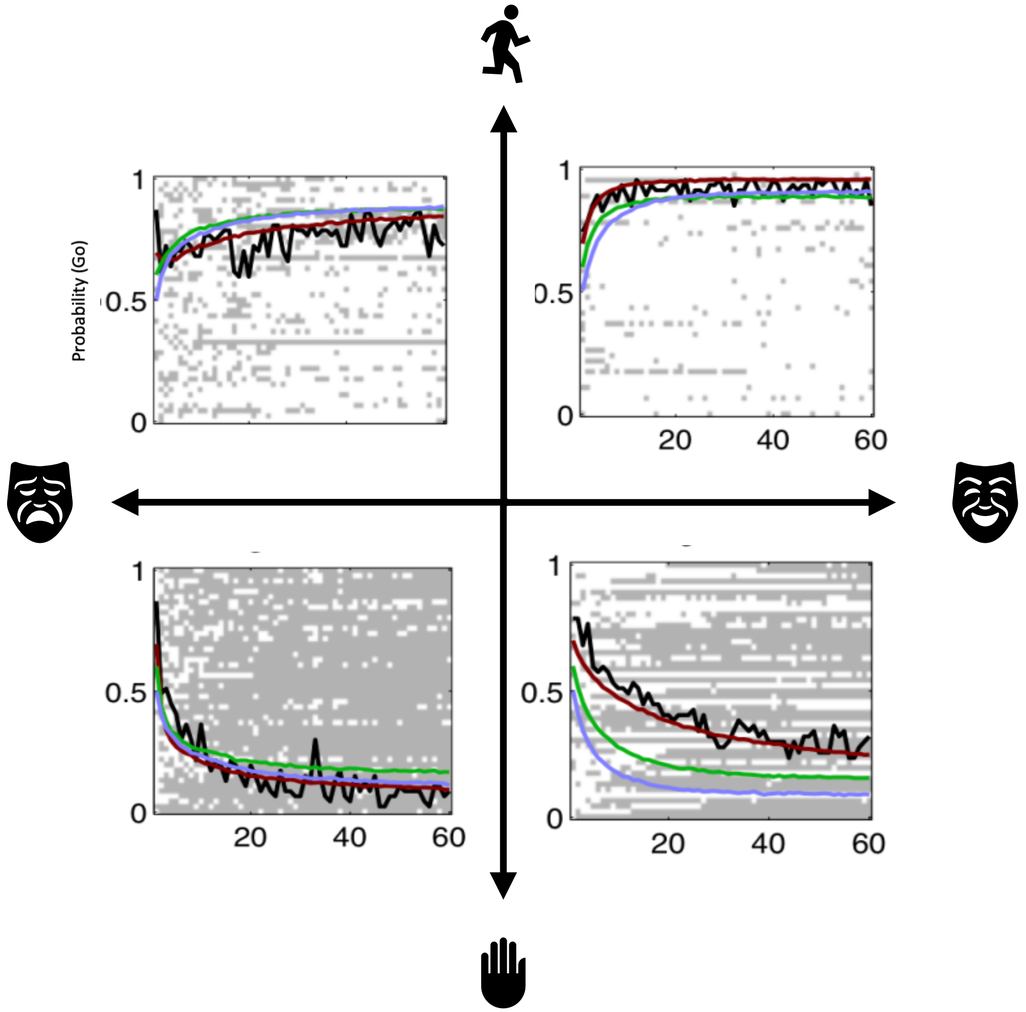



action
inhibition
invigoration
punishment
reward
valence


invigoration
punishment
reward
inhibition

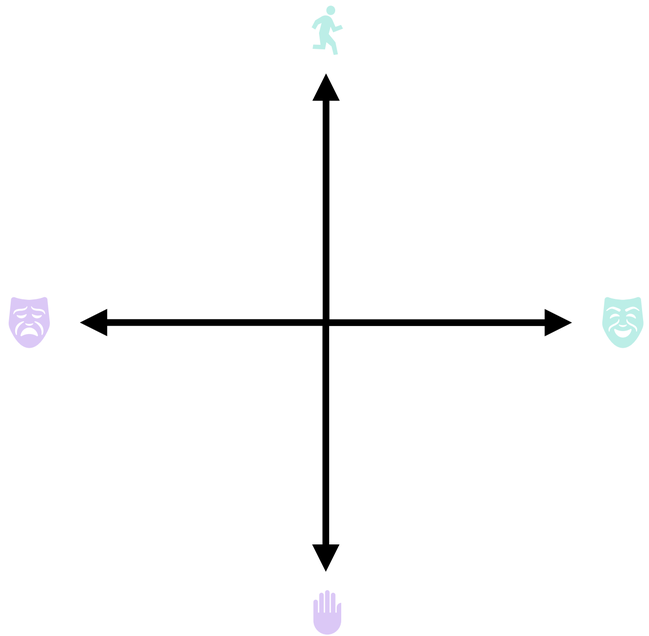

invigoration
go to avoid losing
go to win
punishment
reward
nogo to avoid losing
nogo to win
inhibition


invigoration
punishment
reward
inhibition
1
0.5
0
20
40
60

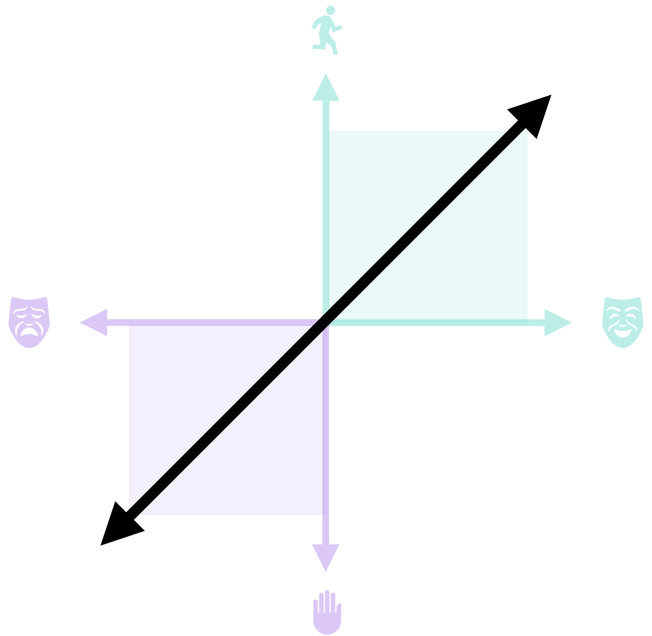


- data
- inst = RW + noise
- inst + bias
- inst + bias + Pavlovian
  - go+$\omega V(s_t)$

(Guitart-Masip et al, 2011; 2012)



instrumental control



Pavlovian control

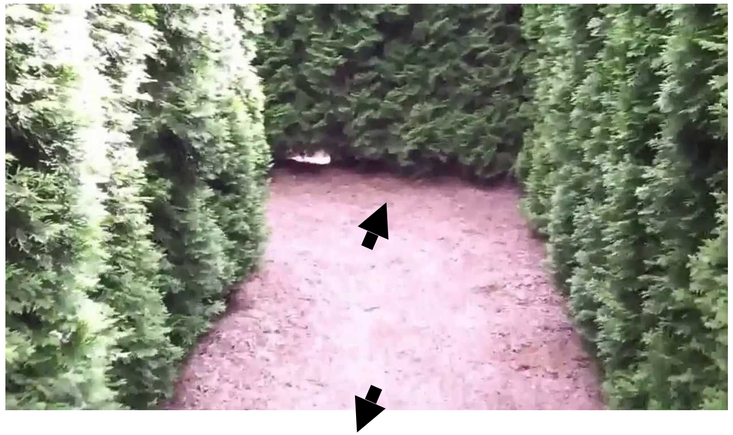
(Dayan et al., 2006)

instrumental conditioning

- immediate choice
- evolutionary programming
- **sequential decisions**
- vigour

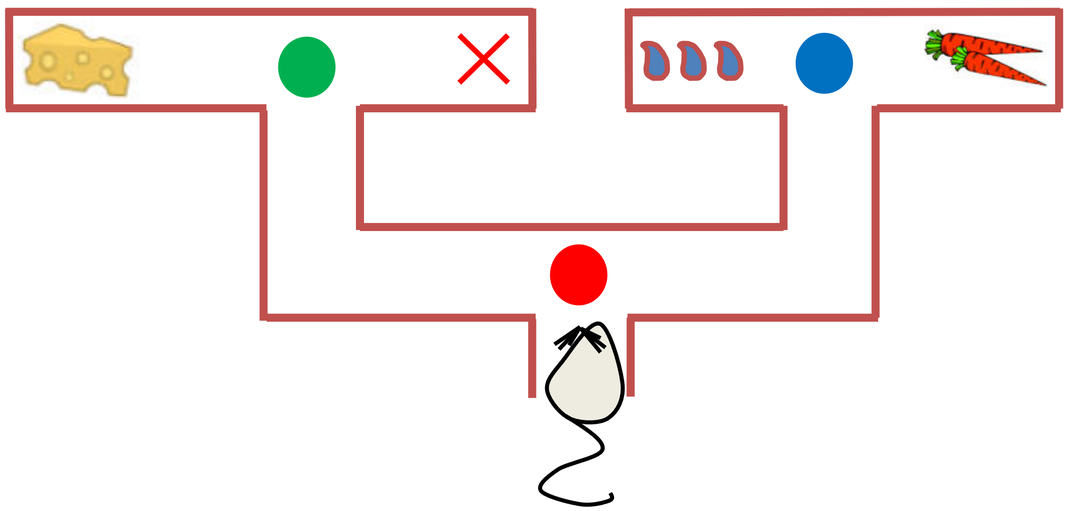
The Problem of Time

Humboldt, Saskatchewan

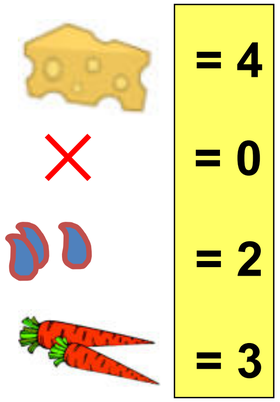


‘ethologically’:
maximize long term reward

A Simpler Example



Hunger

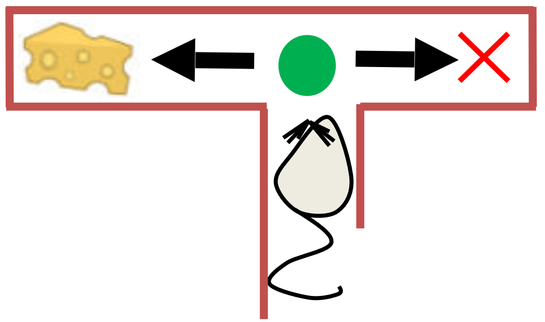


= 4

= 0

= 2

= 3



action reward

4

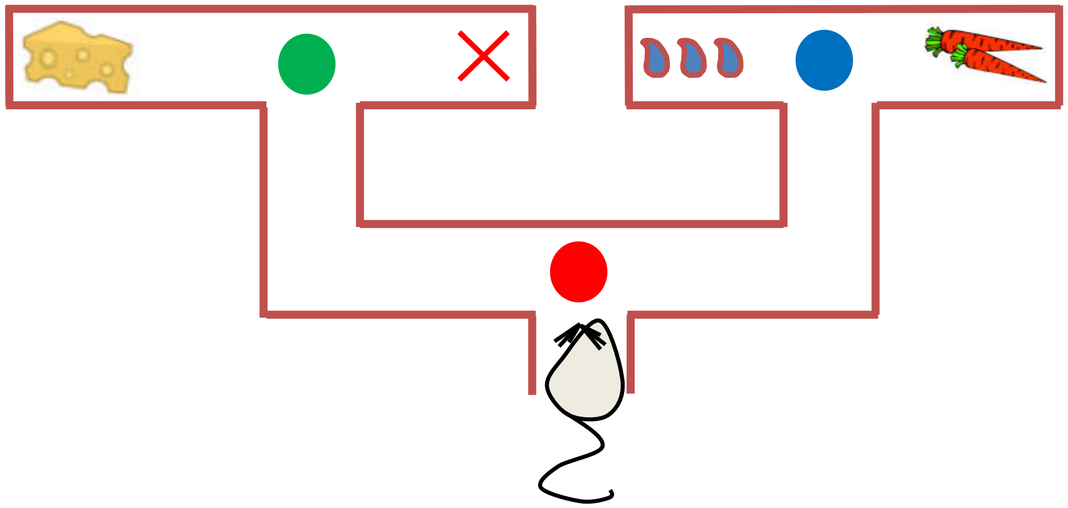
0

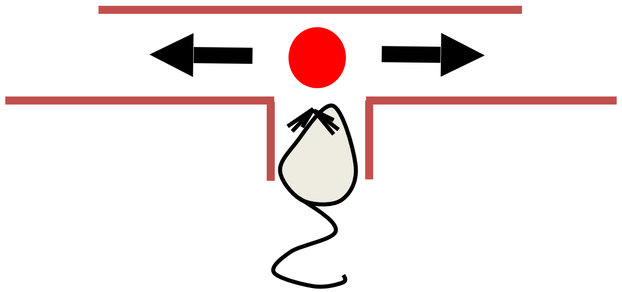
so do:

requires a signal reporting the reward:

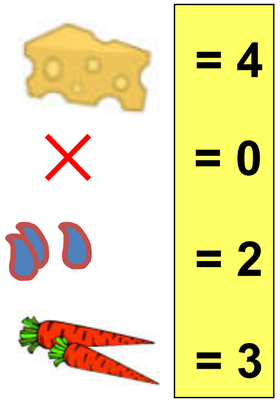
dopamine

A Simpler Example





Hunger



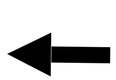
= 4

= 0

= 2

= 3

action reward

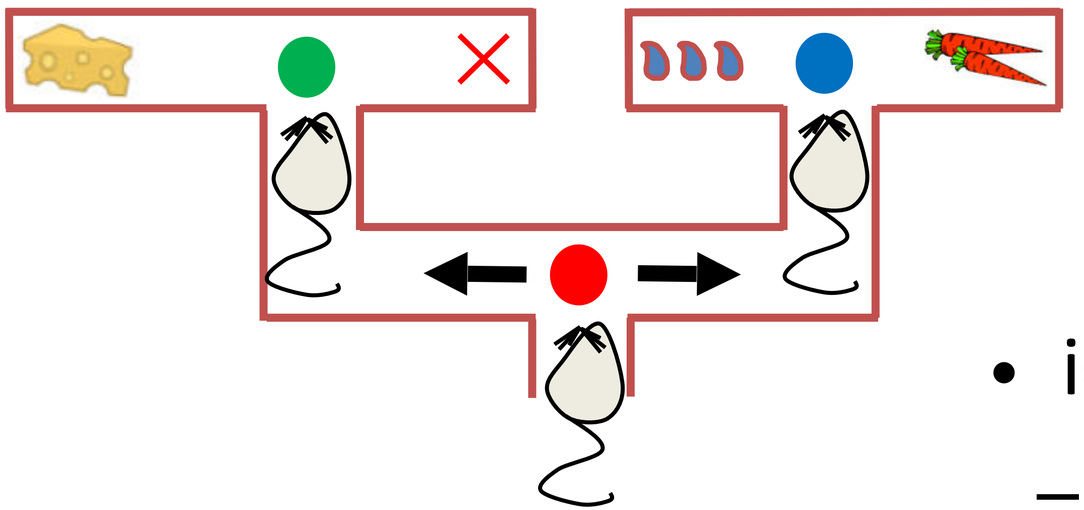


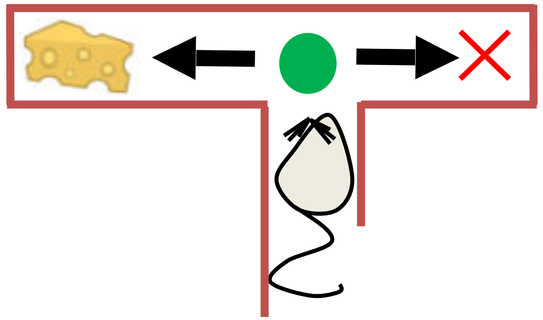
0

0

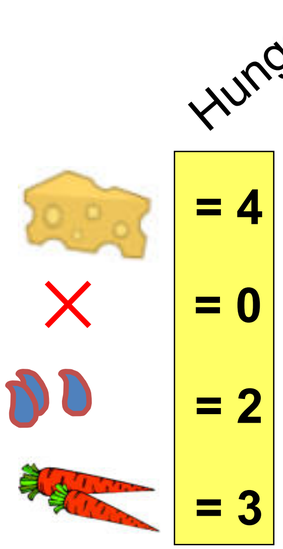
so do: **?**

Two Ideas





- idea a: learn *values* V for *states*
  - r( ) = 4 is *immediately* valuable
  - V( ) = 4 is *secondarily* valuable

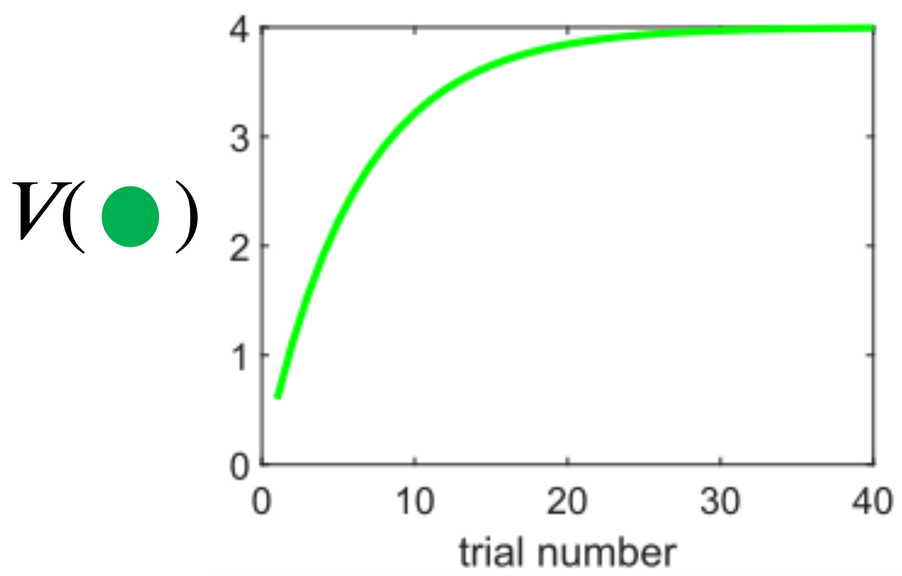


Idea a: learn values V for states

- prediction: $V(\bullet)$ 

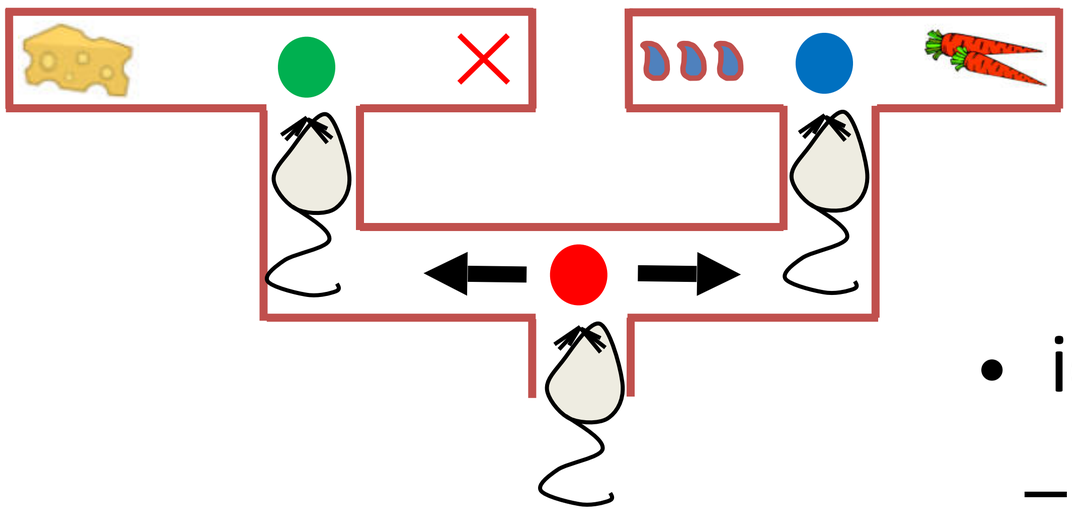
- prediction error: $\delta = r(\text{cheese}) - V(\bullet)$

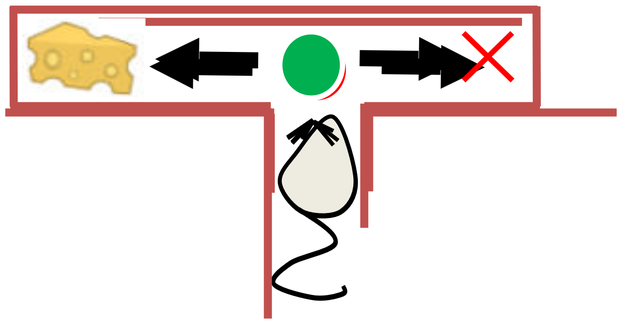
- learning rule: $V(\bullet) \rightarrow V(\bullet) + \alpha \times \delta$

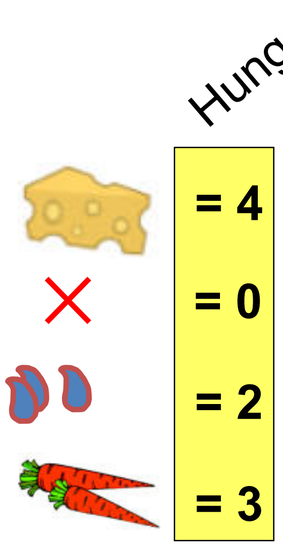


learning rate

Two Ideas



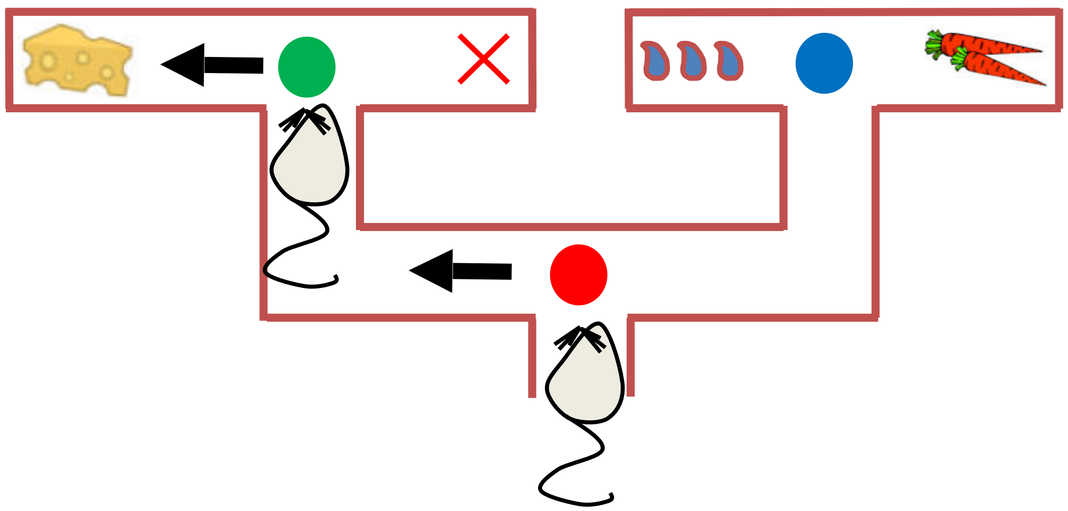


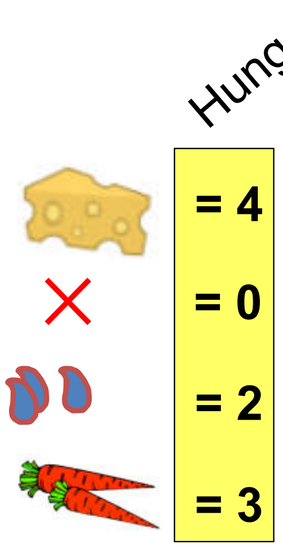


- idea a: learn *values* V for *states*
  - r( ) = 4 is *immediately* valuable
  - V( ) = 4 is *secondarily* valuable
- idea b: use values as *surrogate* feedback
  - ← : V( ) = 4
  - → : V( ) = 3

so do: ←

Idea b: Use values as surrogate feedback





$$V(\bullet) = 4$$

$$V(t) = r(t)$$

$$\delta(t) = r(t) - V(t)$$

prediction error

$$V(\bullet) = 0 + V(\bullet)$$

$$V(t) = r(t) + V(t+1)$$

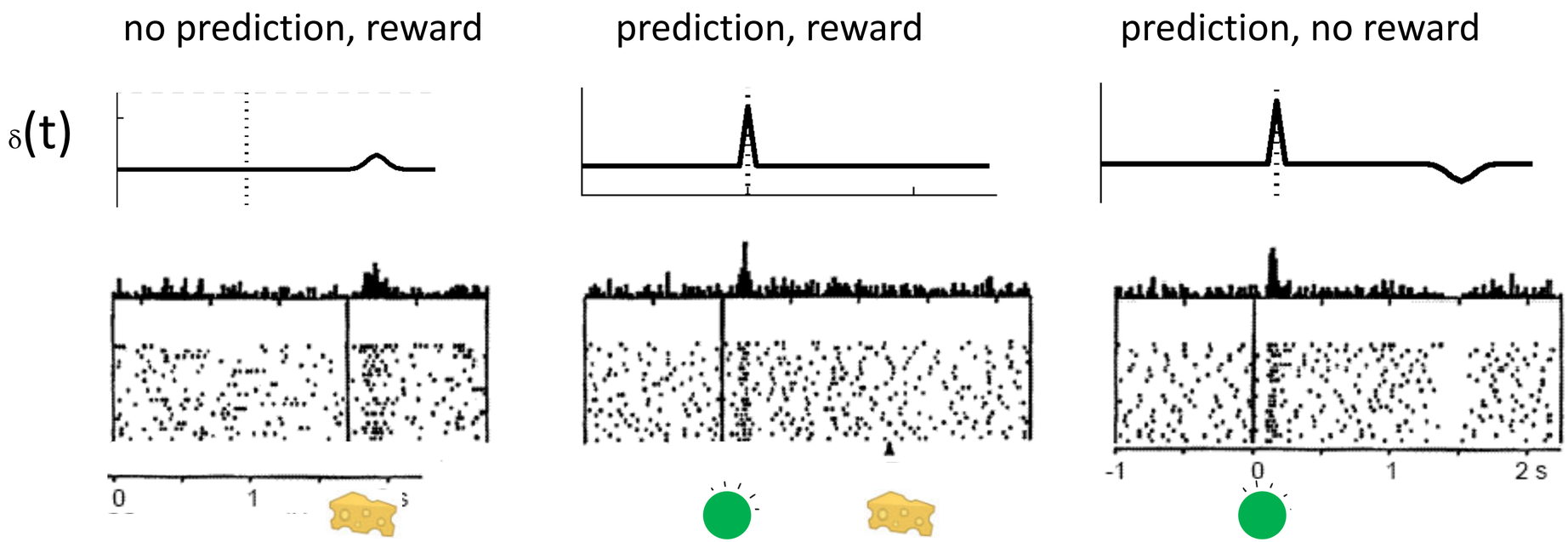
$$\delta(t) = r(t) + V(t+1) - V(t)$$

Sutton: temporal difference prediction error

Bellman: asynchronous dynamic programming

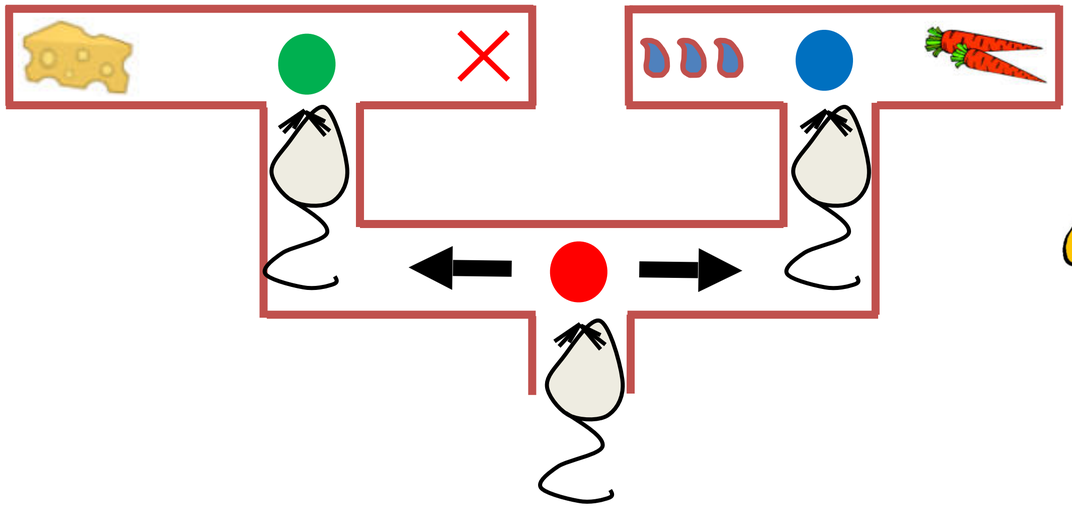
Prediction, Prediction Errors and Dopamine

$$\delta(t) = r(t) + V(t+1) - V(t)$$

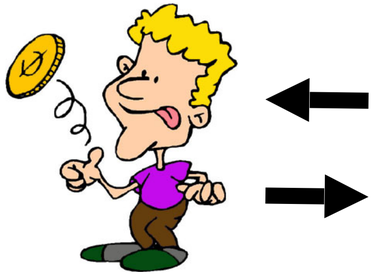


Schultz *et al*

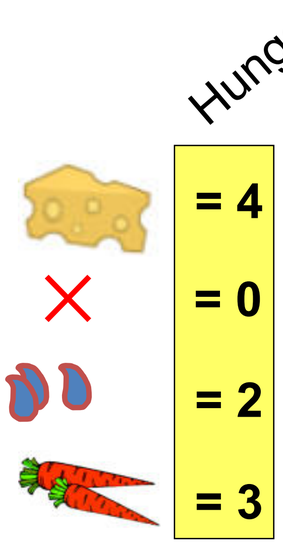
Idea 3: Prediction for Control



If: V( ) = 4 V( ) = 3



V( ) = 3.5



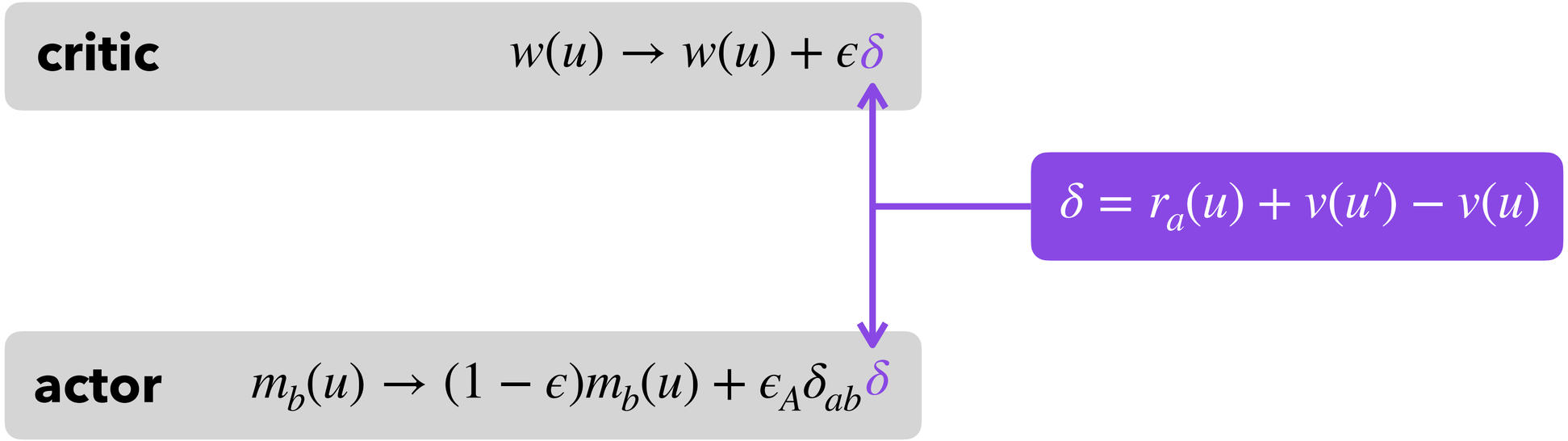
$$\delta(t) = r(t) + V(t+1) - V(t)$$

$$\leftarrow \quad \delta(t) = 0 + 4 - 3.5 = +0.5$$

$$\rightarrow \quad \delta(t) = 0 + 3 - 3.5 = -0.5$$

so increase the probability of ←

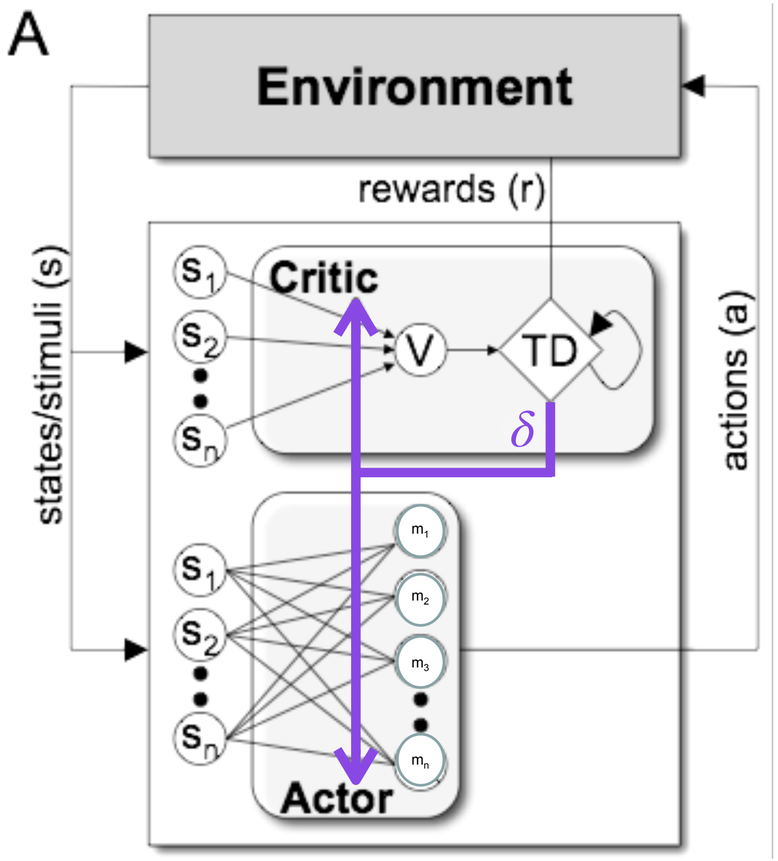
actor-critic

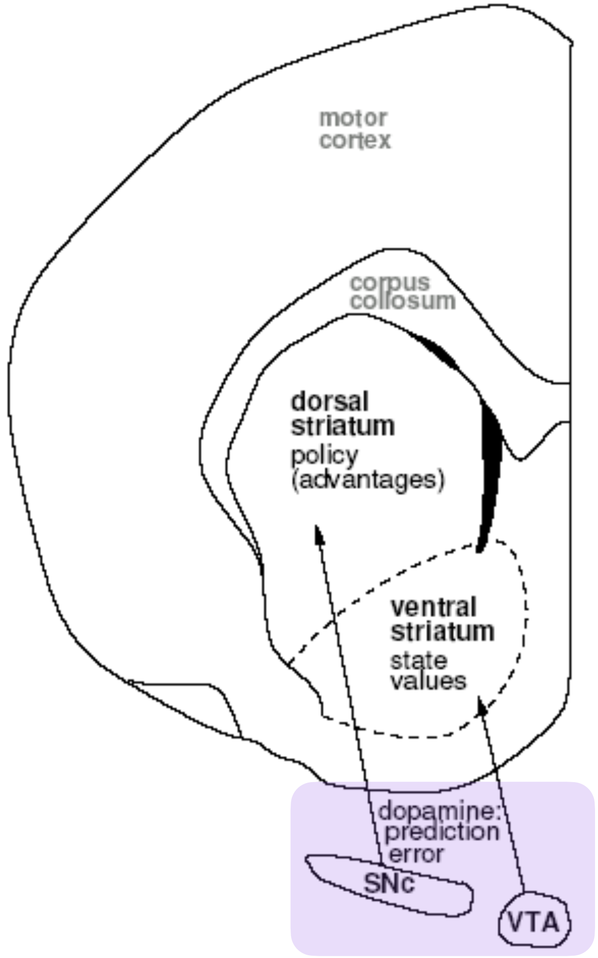


$$m_b \to (1-\epsilon)m_L + \epsilon\delta_{ab}(r_a - v)$$

(direct actor)

actor-critic





dopamine signals to both motivational & motor striatum appear, surprisingly the same

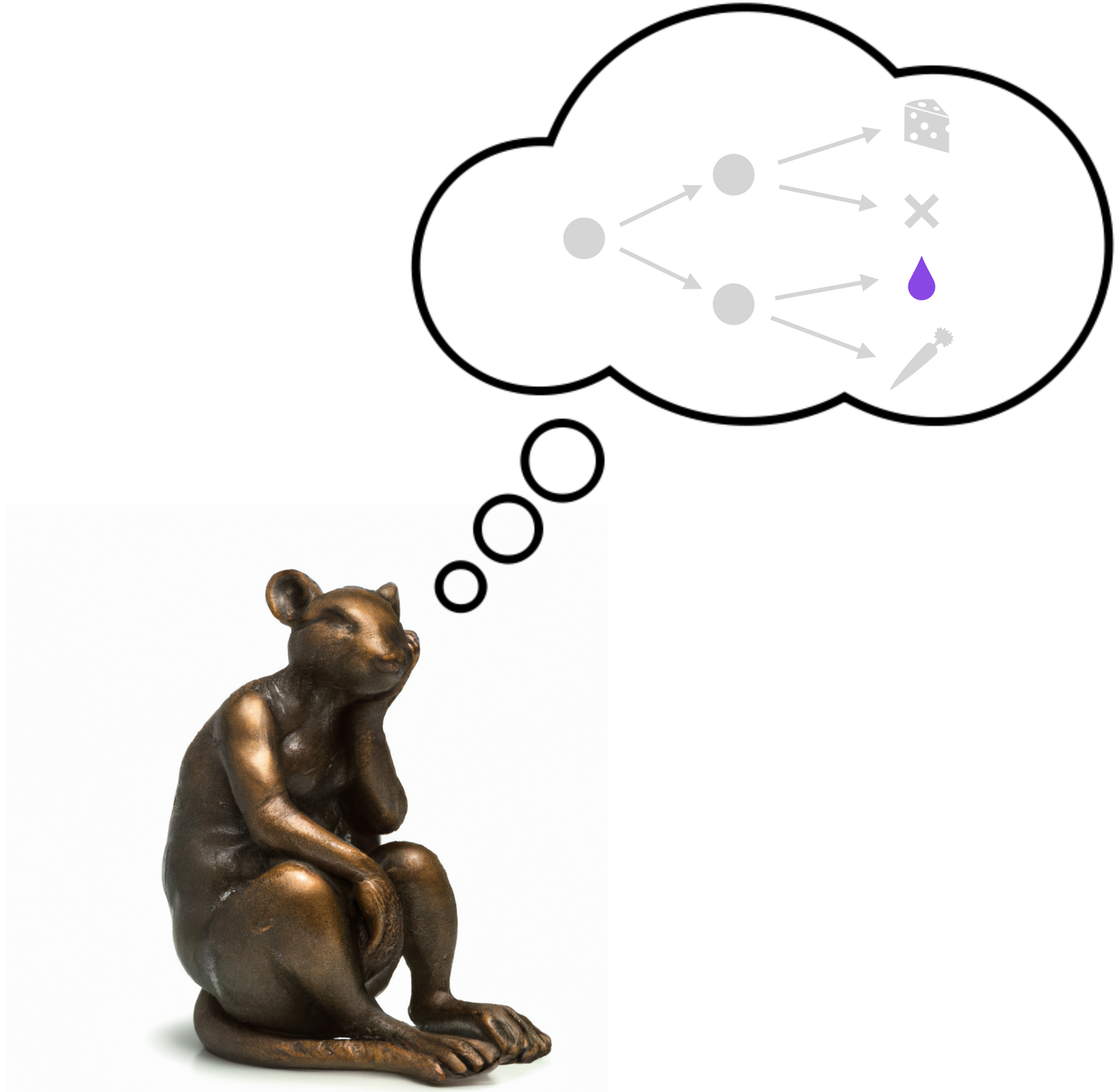
suggestion: training both values & policies

actor-critic

- direct actor
    - immediate choice version of policy gradient
- policy gradient for sequential decisions
    - REINFORCE algorithm, likelihood ratio PG, MC estimate
        - can be used without baseline, worse convergence
        - with baseline = V(s), we also have to learn value
            - learn V (via e.g. TD rule): critic
            - learn policy: actor

model based RL

model based RL



model based RL

instrumental conditioning

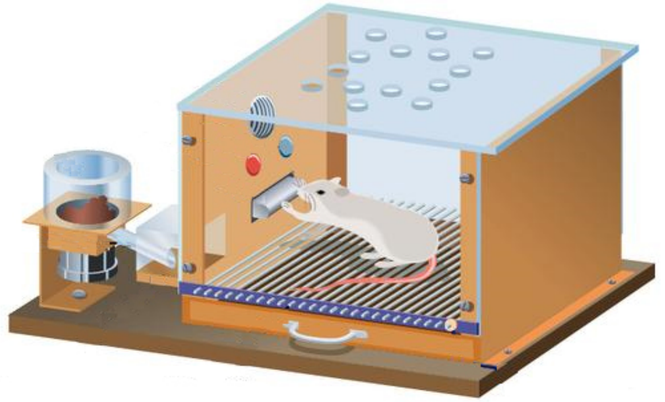
- immediate choice
- evolutionary programming
- sequential decisions
- **vigour**

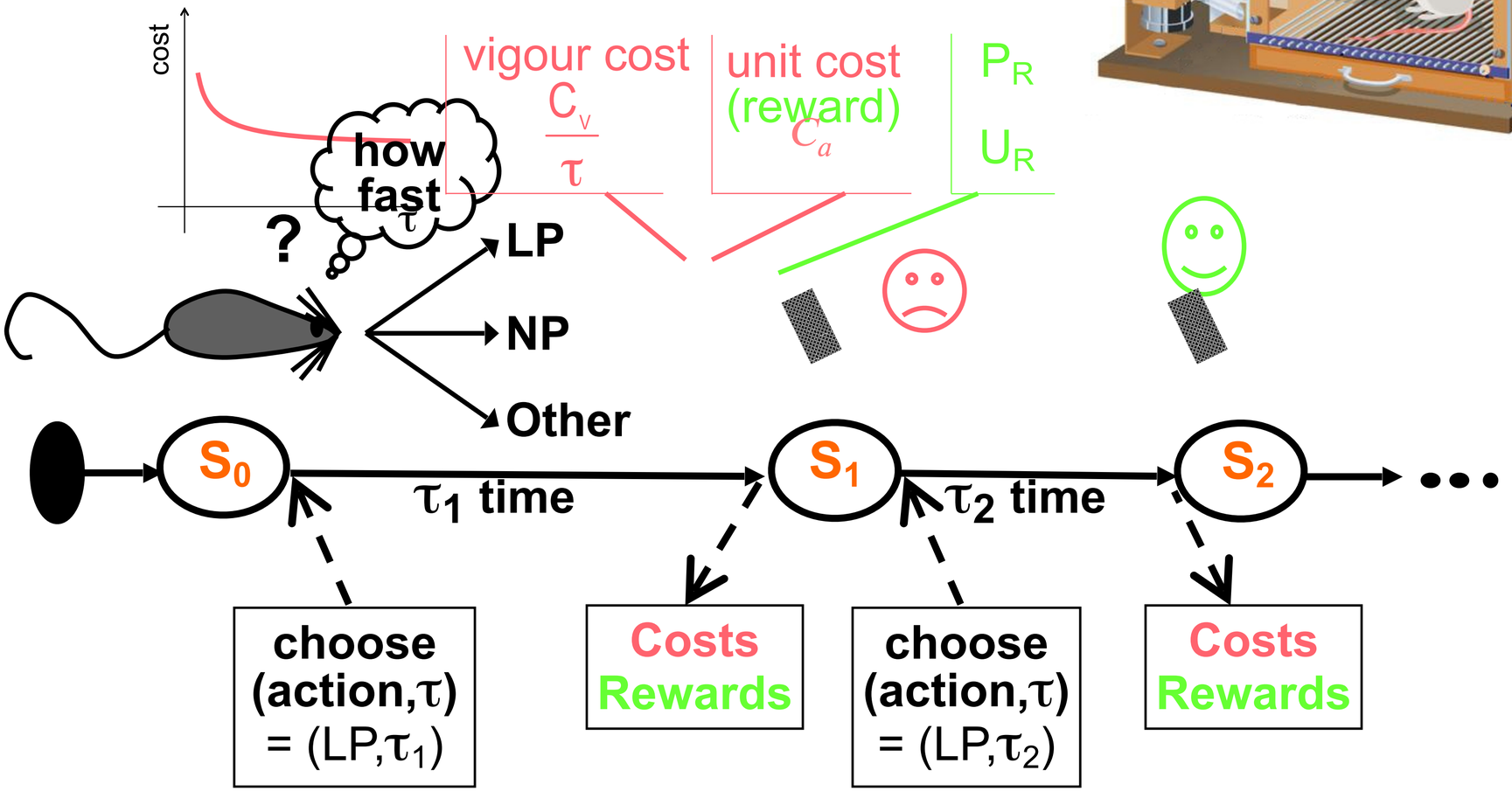
Vigour

- Two components to choice:
  - what:
    - lever pressing
    - direction to run
    - meal to choose
  - when/how fast/how vigorous
    - free operant tasks

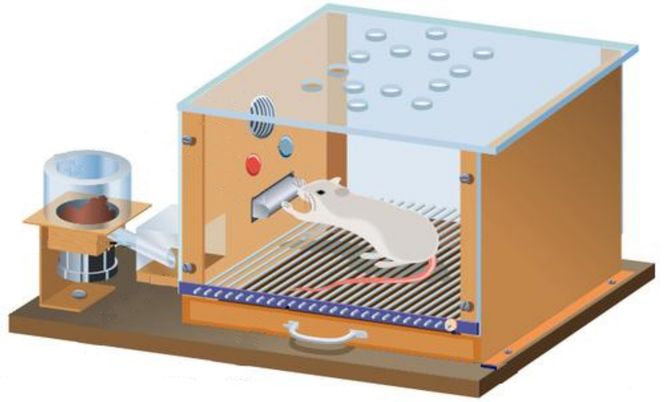
- real-valued dynamic programming

The model

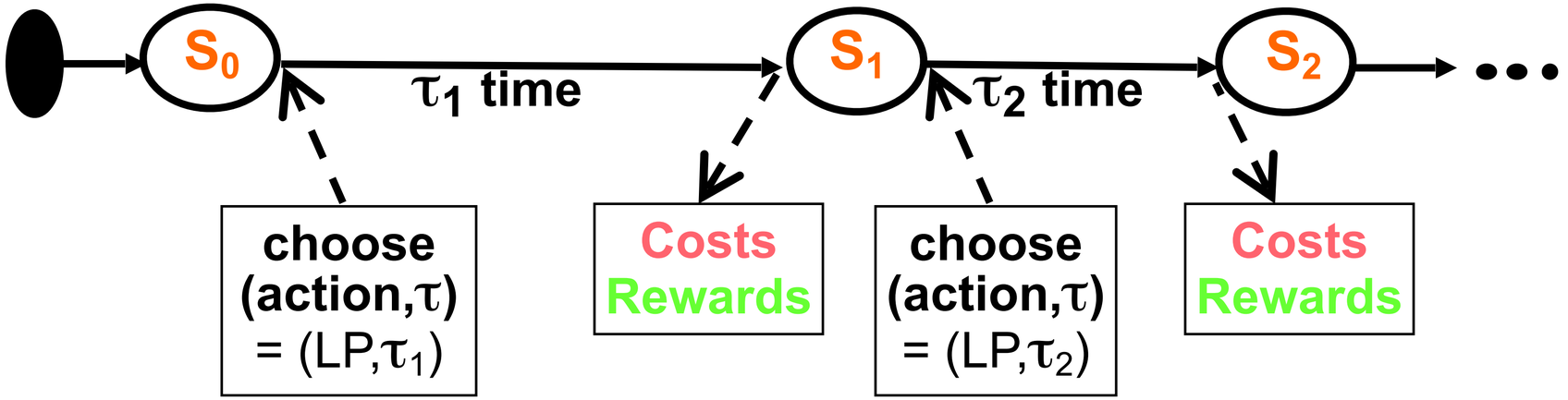



cost
how fast
τ
?
vigour cost
C_v
τ
unit cost (reward)
C_a
P_R
U_R
LP
NP
Other
S_0
S_1
S_2
τ_1 time
τ_2 time
choose (action,τ) = (LP,τ_1)
Costs Rewards
choose (action,τ) = (LP,τ_2)
Costs Rewards

The model



Goal: Choose actions and latencies to maximize
the *average rate of return* (rewards minus costs per time)



Average Reward Cost/benefit Tradeoffs

1. Which action to take?

⇒ Choose action with largest expected reward minus cost

2. How fast to perform it?

- slow → less costly (vigour cost)

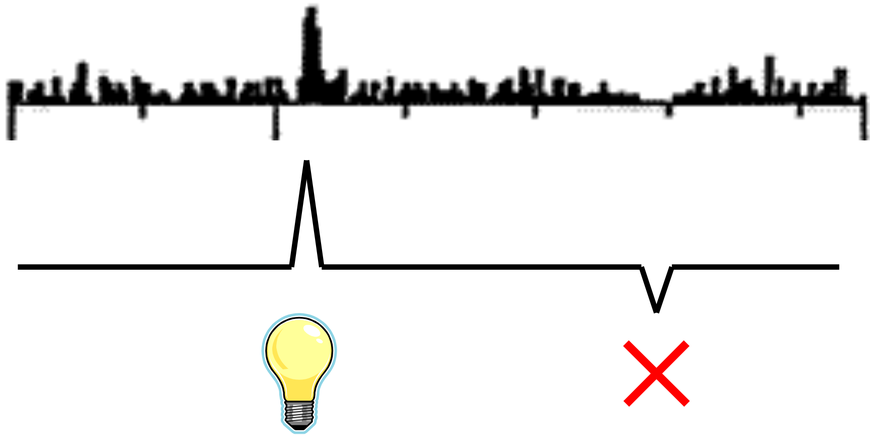
- slow → delays (all) rewards
- net rate of rewards = cost of delay (opportunity cost of time)

⇒ Choose rate that balances vigour and opportunity costs

explains faster (irrelevant) actions under hunger, etc

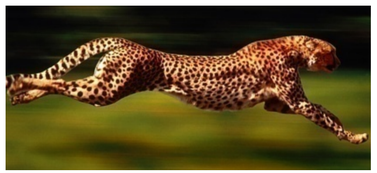
Relation to Dopamine

Phasic dopamine firing = reward prediction error

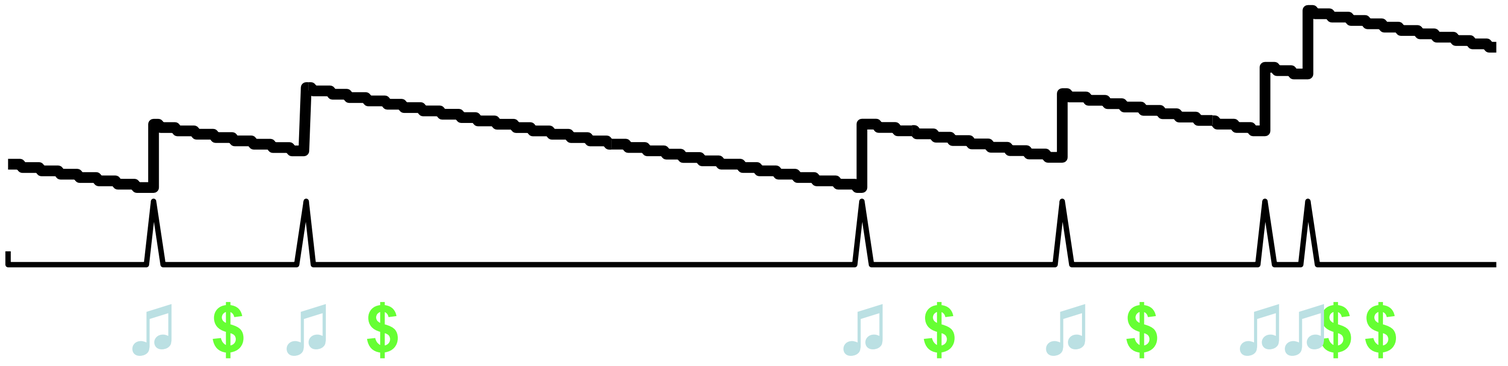


What about tonic dopamine?

less ⟵ ⟶ **more**



Tonic dopamine hypothesis

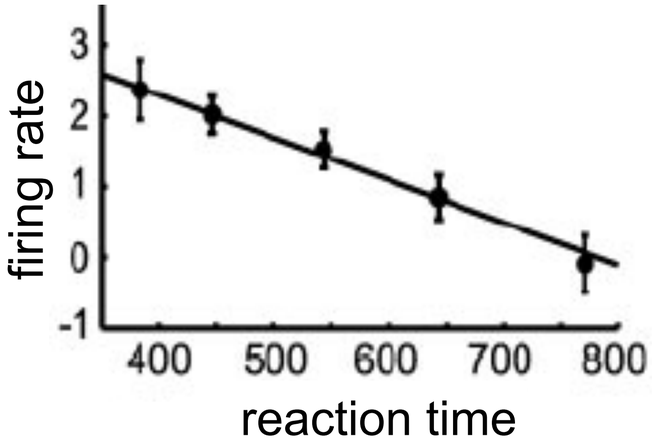


$$\sum \delta_t = \sum [r_{t+1} + V(s_{t+1}) - V(s_t)]$$

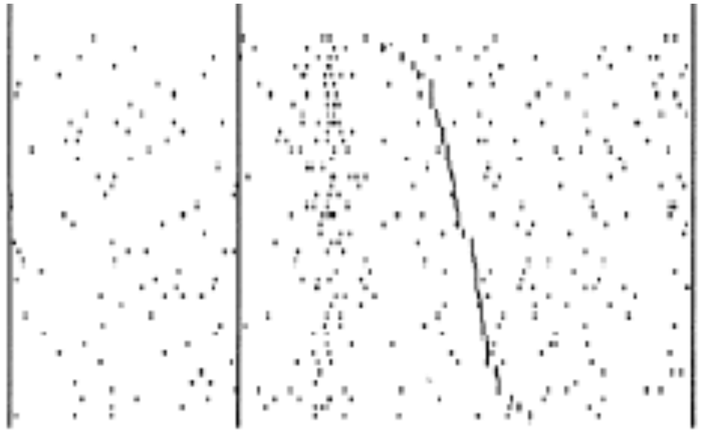
$$= \sum [r_{t+1}] + V(s_T) - V(s_0)$$

so the average is the same
as the average total reward,
the timing is just moved around

…also explains effects of phasic dopamine on response times



Satoh and Kimura 2003



Ljungberg, Apicella and Schultz 1992

+ direct evidence from Berke: pro and contra

Classical/Instrumental Conditioning

prediction: of important events
control: in the light of those predictions

- Ethology
  - optimality
  - appropriateness
- Psychology
  - classical/operant conditioning

- Computation
  - dynamic progr.
  - Kalman filtering
- Algorithm
  - TD/delta rules
  - simple weights

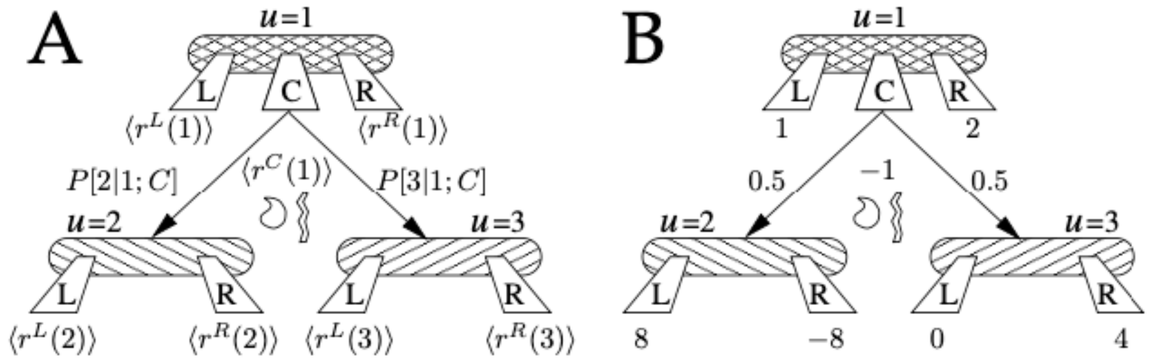
- Neurobiology

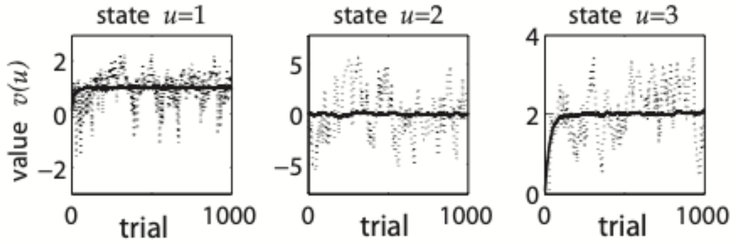
neuromodulators; amygdala; OFC
nucleus accumbens; dorsal striatum

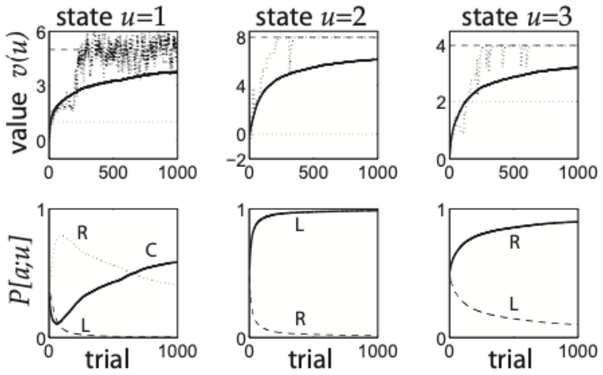
homework

- reproduce figures 9.8 (policy evaluation) and 9.9 (actor-critic) from the chapter for the environment of figure 9.7

- try actor critic for a maze task







homework

- Tutorial by Noémi Éltető tomorrow at DZNE Lecture Room