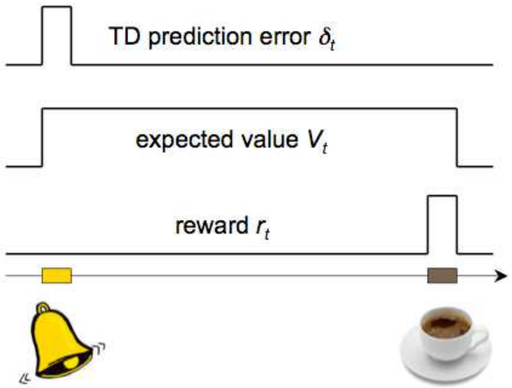
Summary of part I: prediction and RL



Prediction is important for action selection

- The problem: prediction of future reward
- The algorithm: temporal difference learning
- Neural implementation: dopamine dependent learning in BG

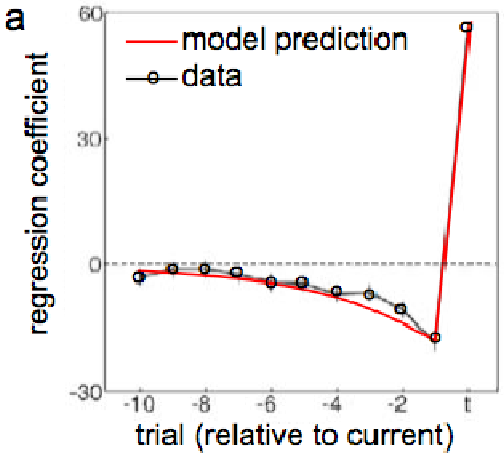
⇒ A precise computational model of learning allows one to look in the brain for "hidden variables" postulated by the model

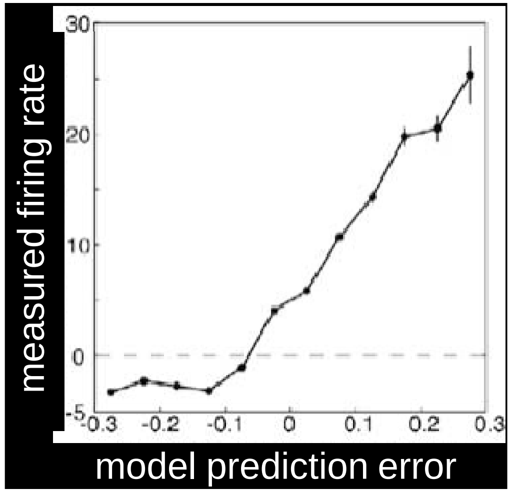
⇒ Precise (normative!) theory for generation of dopamine firing patterns

⇒ Explains anticipatory dopaminergic responding, second order conditioning

⇒ Compelling account for the role of dopamine in classical conditioning: prediction error acts as signal driving learning in prediction areas

prediction error hypothesis of dopamine





at end of trial: $\delta_t = r_t - V_t$ (just like R-W)

$$V_t = \eta \sum_{i=1}^{t} (1-\eta)^{t-i} r_i$$

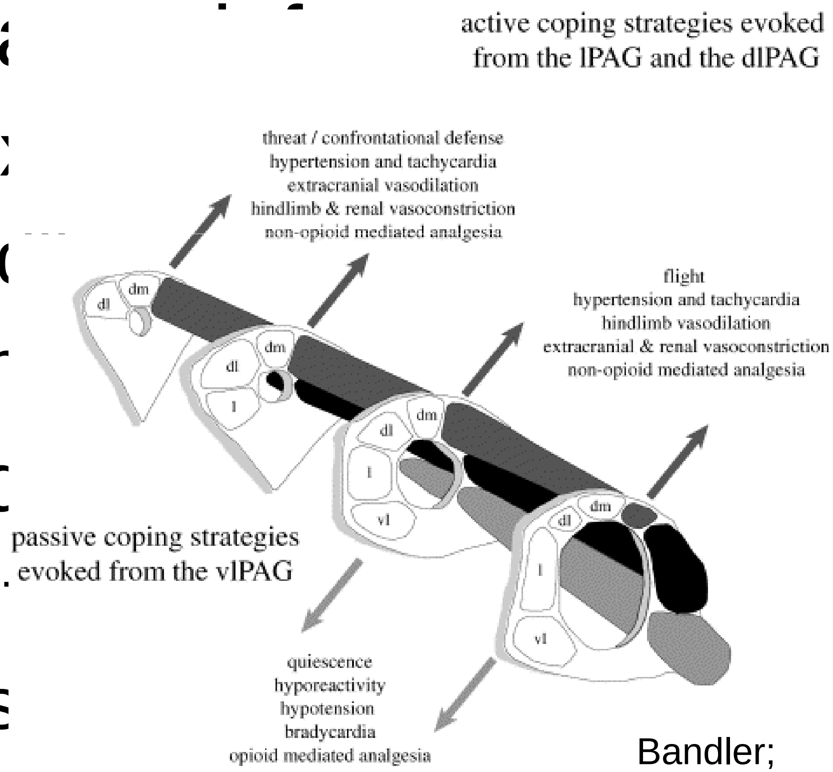
Bayer & Glimcher (2005)

Global plan

- Reinforcement learning I:
  - prediction
  - classical conditioning
  - dopamine
- Reinforcement learning II:
  - dynamic programming; action selection
  - Pavlovian misbehaviour
  - vigor
- Chapter 9 of Theoretical Neuroscience

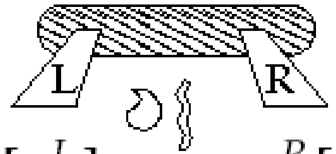
Action Selection

- Evolutionary specification
- Immedi
  - leg fle
  - Thorn
  - pigeo
- Delaye
  - these
  - mazes
  - chess

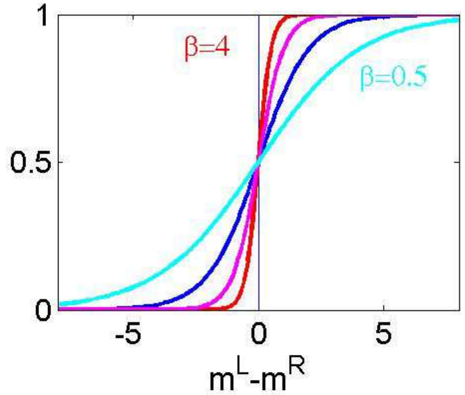


Bandler;
Blanchard

Immediate Reinforcement



$p^L[r^L]$ $p^R[r^R]$



- stochastic policy:

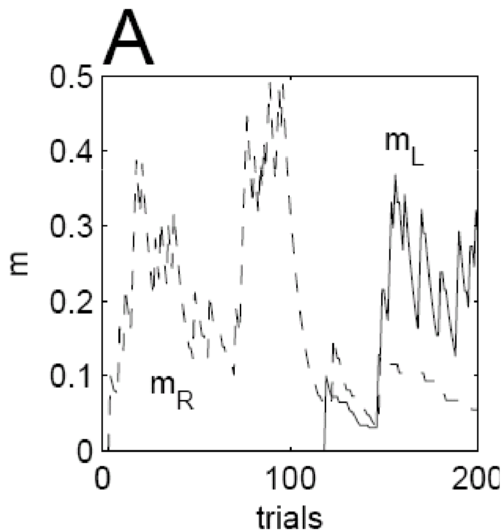
$$P[L] = \frac{\exp(\beta m^L)}{\exp(\beta m^L) + \exp(\beta m^R)} = \sigma(\beta(m^L - m^R))$$

- based on action values: $m^L; m^R$

Indirect Actor

use RW rule:

$$m^L \to m^L + \epsilon\delta \quad \text{with} \quad \delta = r^L - m^L$$



$$\left\langle r^L \right\rangle_{p^L} = 0.05; \left\langle r^R \right\rangle_{p^R} = 0.25$$

switch every 100 trials

Direct Actor

$$E(\mathbf{m}) = P[L]\left\langle r^L \right\rangle + P[R]\left\langle r^R \right\rangle$$

$$\frac{\partial P[L]}{\partial m^L} = \beta P[L]P[R] \qquad \frac{\partial P[R]}{\partial m^R} = -\beta P[L]P[R]$$

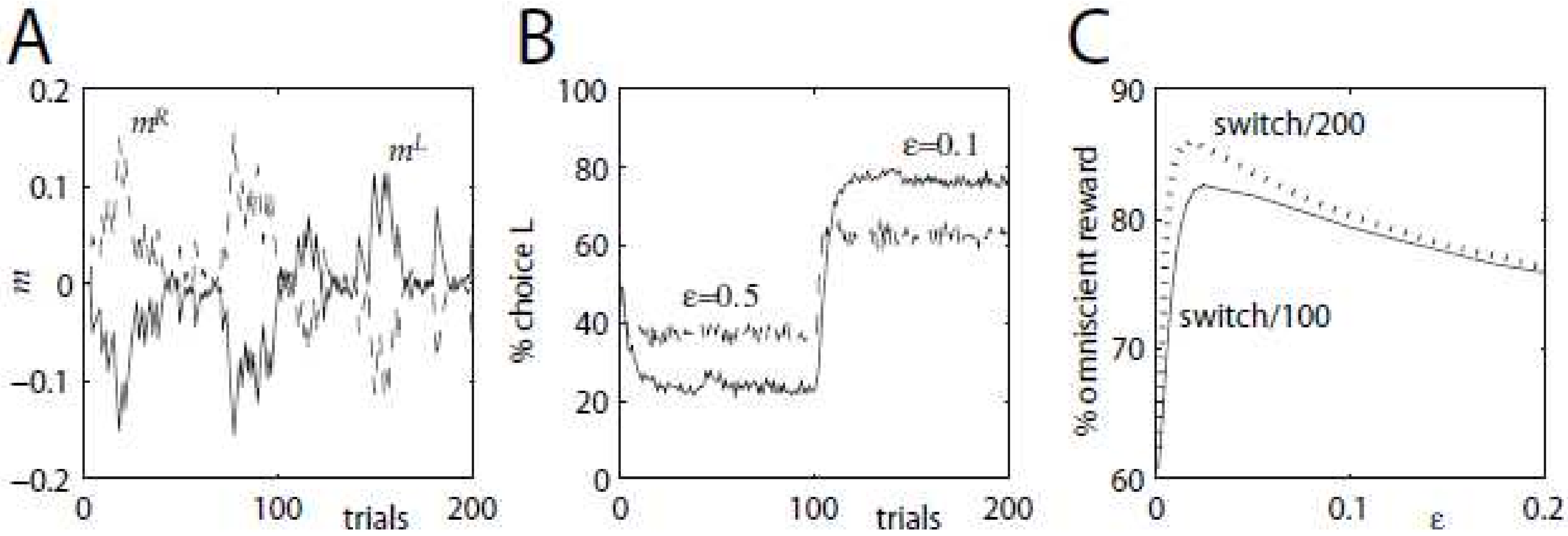
$$\frac{\partial E(\mathbf{m})}{\partial m^L} = \beta P[L]\left(\left\langle r^L \right\rangle - \left(P[L]\left\langle r^L \right\rangle + P[R]\left\langle r^R \right\rangle\right)\right)$$

$$\frac{\partial E(\mathbf{m})}{\partial m^L} = \beta P[L]\left(\left\langle r^L \right\rangle - E(\mathbf{m})\right)$$

$$\frac{\partial E(\mathbf{m})}{\partial m^L} \approx \beta\left(r^L - E(\mathbf{m})\right) \qquad \text{if L is chosen}$$

$$m^L - m^R \rightarrow (m^L - m^R) + \varepsilon(r^a - E(\mathbf{m}))(L - R)$$

Direct Actor



Could we Tell?

- correlate past rewards, actions with present choice
- indirect actor (separate clocks):

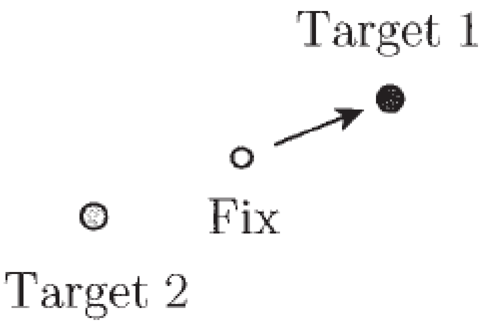
$$\log \frac{P_S[L]}{P_S[R]} = \beta(m_S^L - m_S^R) = \beta\epsilon\left(\sum_i (1-\epsilon)^i r_i^L - \sum_i (1-\epsilon)^i r_i^R\right)$$

- direct actor (single clock):

$$\log \frac{P_{K+1}[L]}{P_{K+1}[R]} = \beta(m_{K+1}^L - m_{K+1}^R) \quad (1$$

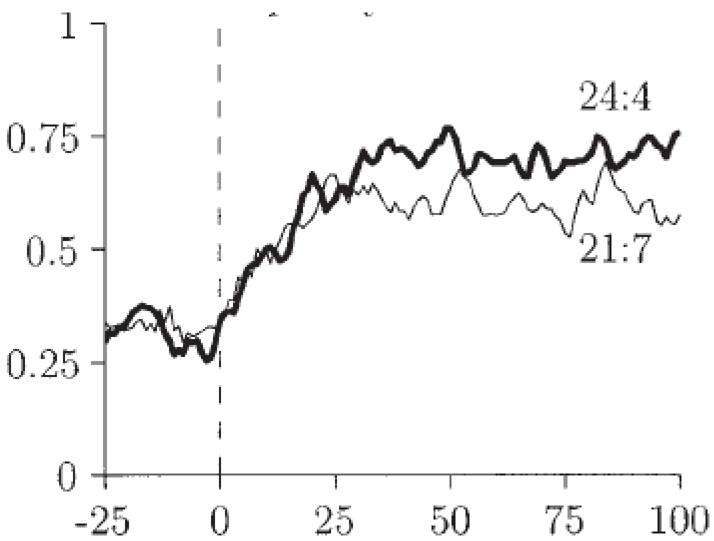
$$= \beta\epsilon \sum_{k=0} (1-\epsilon)^k r_{K-k}^a (L_{K-k} - R_{K-k}) - \beta\epsilon \sum_{k=0} (1-\epsilon)^k v_{K-k} (L_{K-k} - R_{K-k})$$

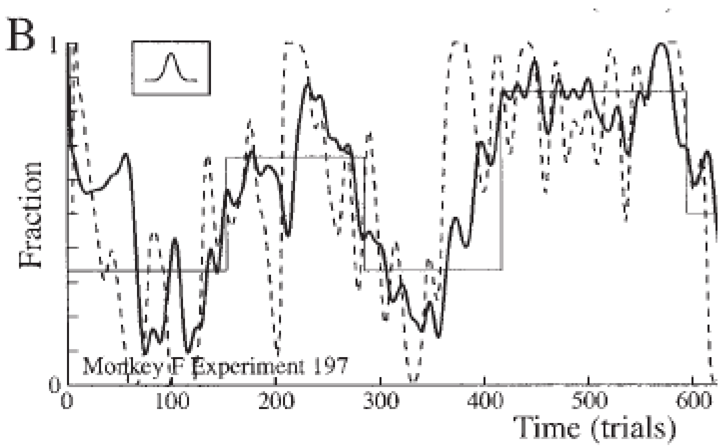
Matching: Concurrent VI-VI



Monkey H

Arming probabilities	Magnitudes	Number blocks
0.24/0.04	0.35/0.35	16
0.21/0.07	0.35/0.35	12
0.07/0.21	0.35/0.35	11
0.04/0.24	0.35/0.35	19
0.24/0.04	0.4/0.4	5
0.21/0.07	0.4/0.4	8
0.07/0.21	0.4/0.4	8
0.04/0.24	0.4/0.4	4
0.24/0.04	0.45/0.45	7
0.21/0.07	0.45/0.45	10
0.07/0.21	0.45/0.45	5
0.04/0.24	0.45/0.45	10

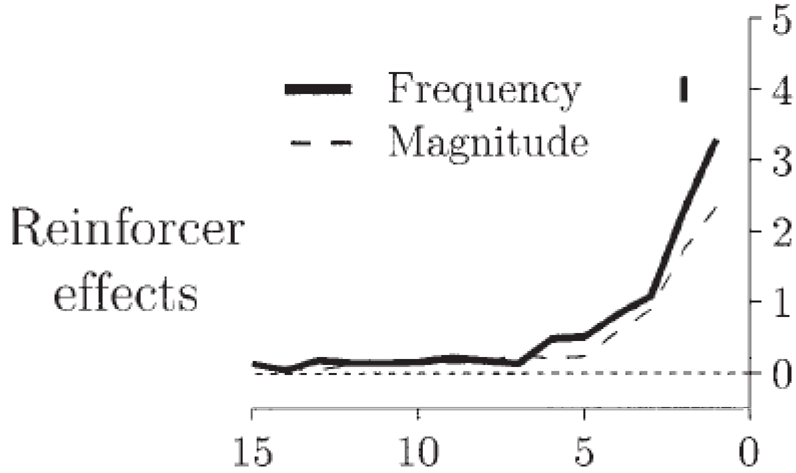




Lau, Glimcher, Corrado, Sugrue, Newsome

Matching

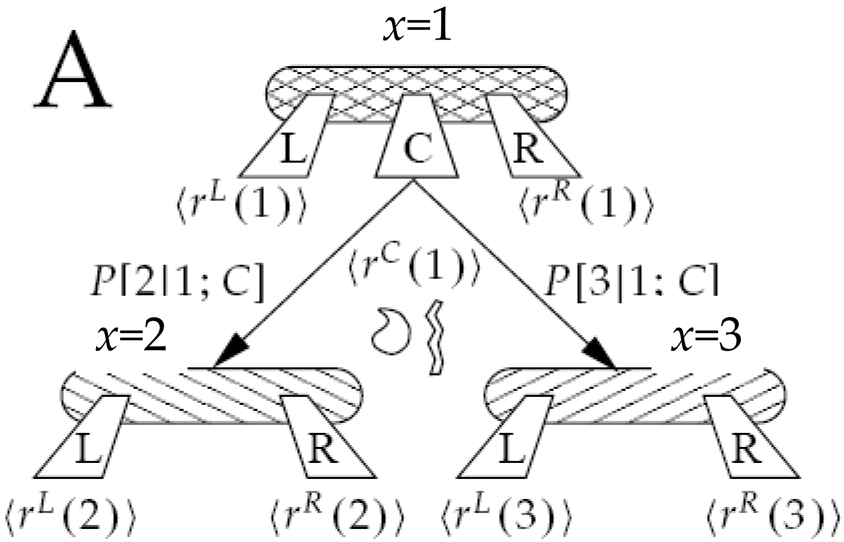
$$\log\left(\frac{p_{R,i}}{p_{L,i}}\right) = \sum_{j=1} \alpha_j \left(r_{R,i-j} - r_{L,i-j}\right)$$



- income not return
- approximately exponential in r

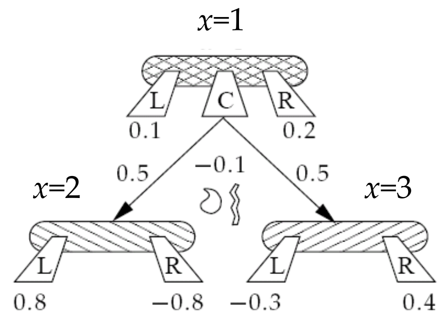
- alternation choice kernel

Action at a (Temporal) Distance



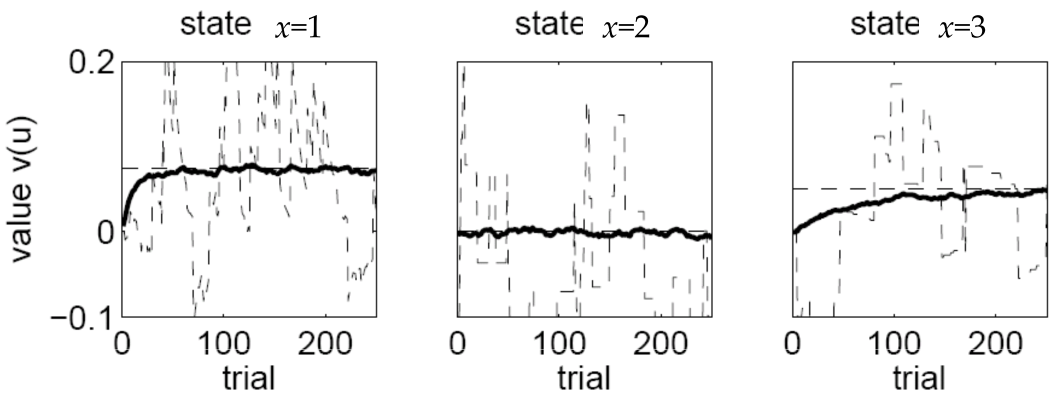
- learning an appropriate action at x=1:
  - depends on the actions at x=2 and x=3
  - gains no immediate feedback
- idea: use prediction as surrogate feedback

Action Selection

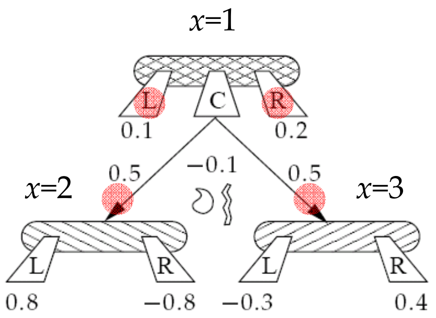


start with policy: $P[L;x] = \sigma(m^L(x) - m^R(x))$

evaluate it: $V(1), V(2), V(3)$



improve it:



$$\delta(t) = \quad r_t + \quad V(x_{t+1}) \quad - V(x_t)$$

0.025

-0.175

-0.125

0.125

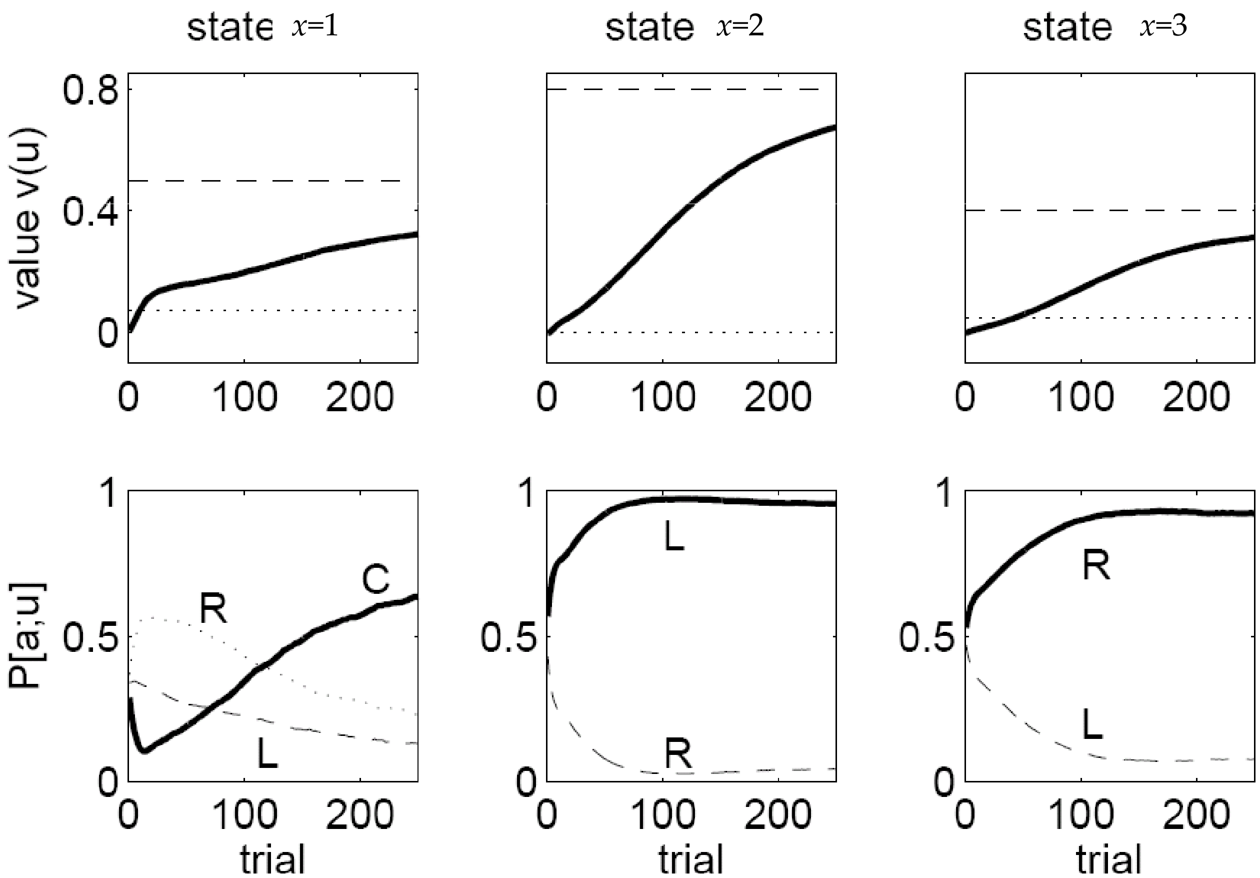
 thus choose R more frequently than L;C

$$\Delta m_* \; \alpha \; \delta$$

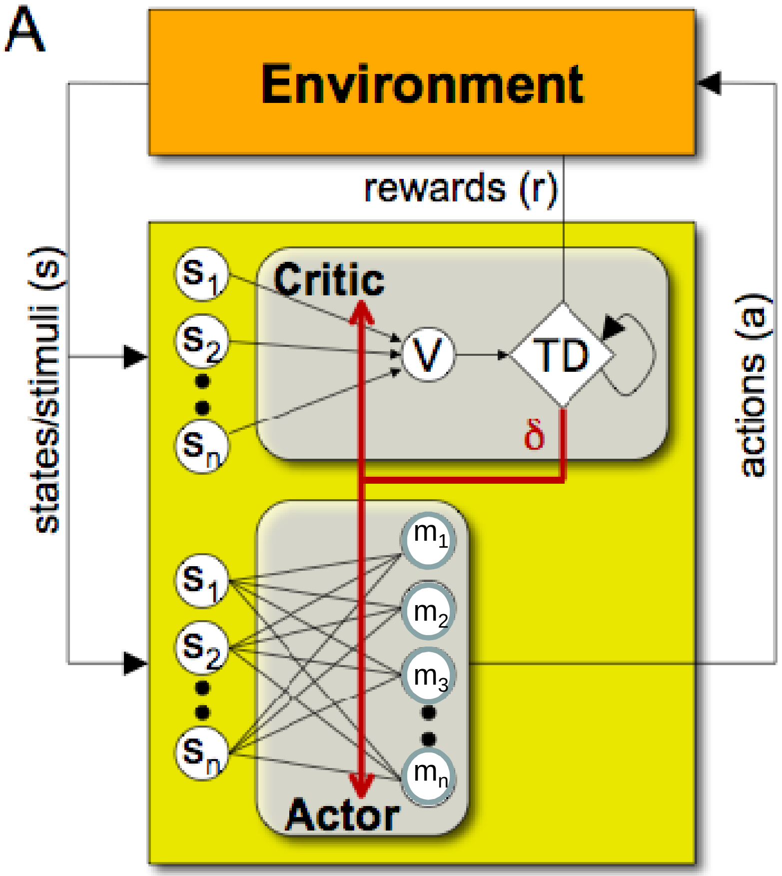
Policy

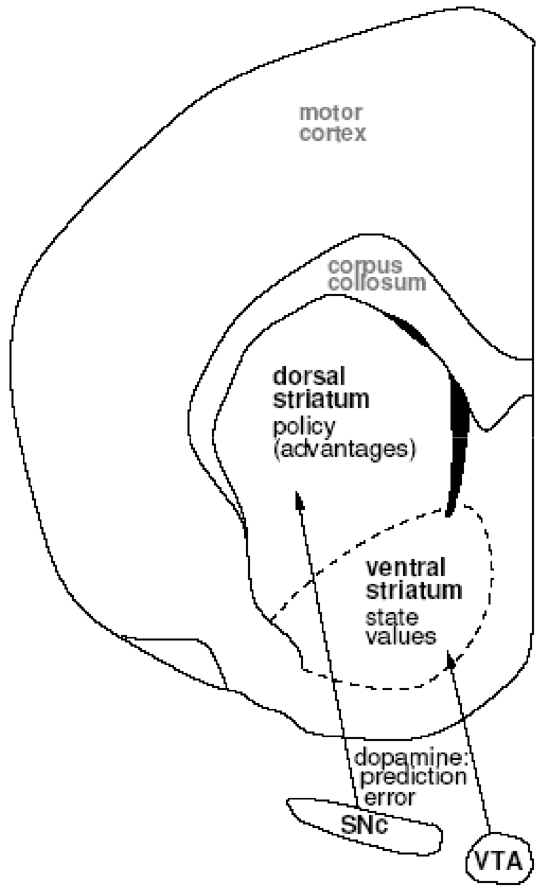
$\delta > 0$ if

- value is too pessimistic $\Rightarrow \Delta v$
- action is better than average $\Rightarrow \Delta P$



actor/critic





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

suggestion: training both values & policies

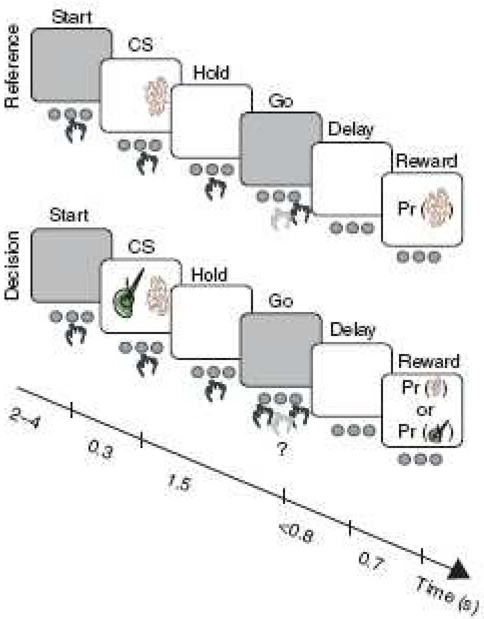
Formally: Dynamic Programming

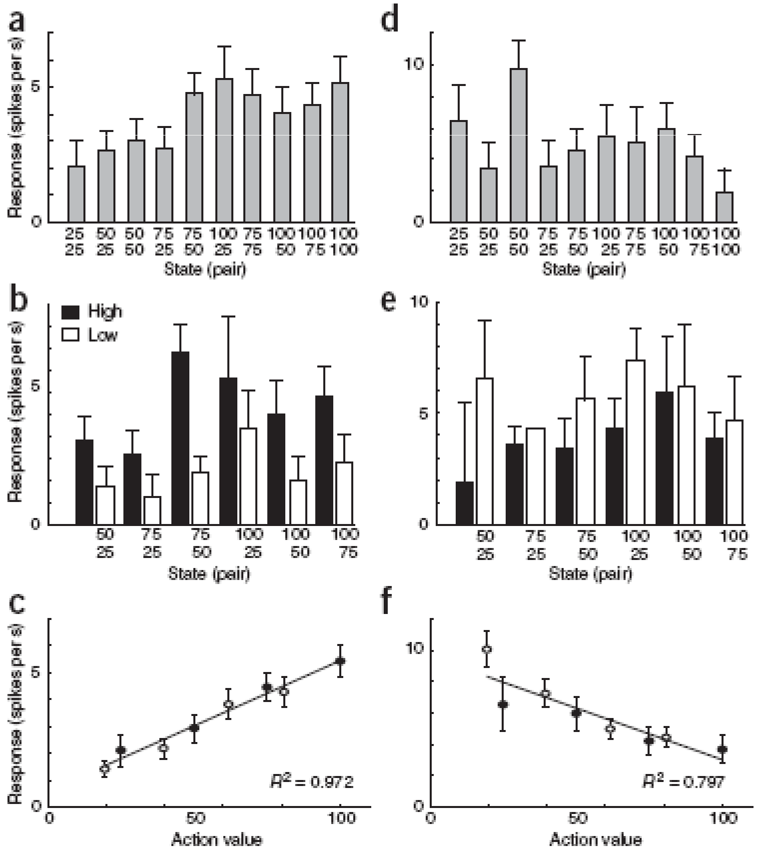
- $V^*(x_t) = \max_u\{E[r(x_t,u) + V^*(x_{t+1})]\}$
- $\qquad\qquad = \max_u\left\{E\left[r(x_t,u) + \sum_y P(y|x_t,u)\, V^*(y)\right]\right\}$
- $Q^*(x,u) = E[r(x,u) + \sum_y P(y|x,u)\, V^*(y)]$
- $V^*(y) \quad = \max_{u'}\{Q^*(y,u')\}$
- policy iteration:
  - $V^\pi(x) = \sum_u \pi(u|x) \left\{E\left[r(x_t,u) + \sum_y P(y|x_t,u)\, V^\pi(y)\right]\right\}$
  - $\pi'(x) = \mathrm{argmax}_u\{Q^\pi(x,u)\}$
- value iteration
  - $V^{n+1}(x) = \max_u\left\{E\left[r(x_t,u) + \sum_y P(y|x_t,u)\, V^{n+1}(y)\right]\right\}$

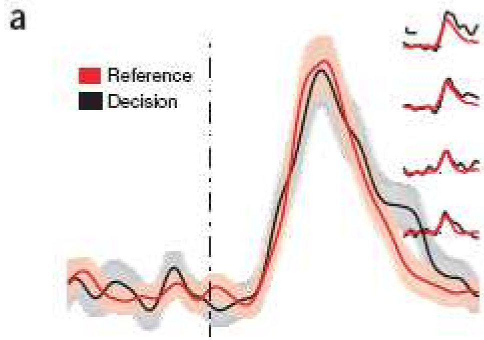
Variants: SARSA

$$Q^*(1,C) = E\left[r_t + V^*(x_{t+1}) \mid x_t = 1, u_t = C\right]$$

$$Q(1,C) \rightarrow Q(1,C) + \varepsilon\left(r_t + Q(2,u^{actual}) - Q(1,C)\right)$$





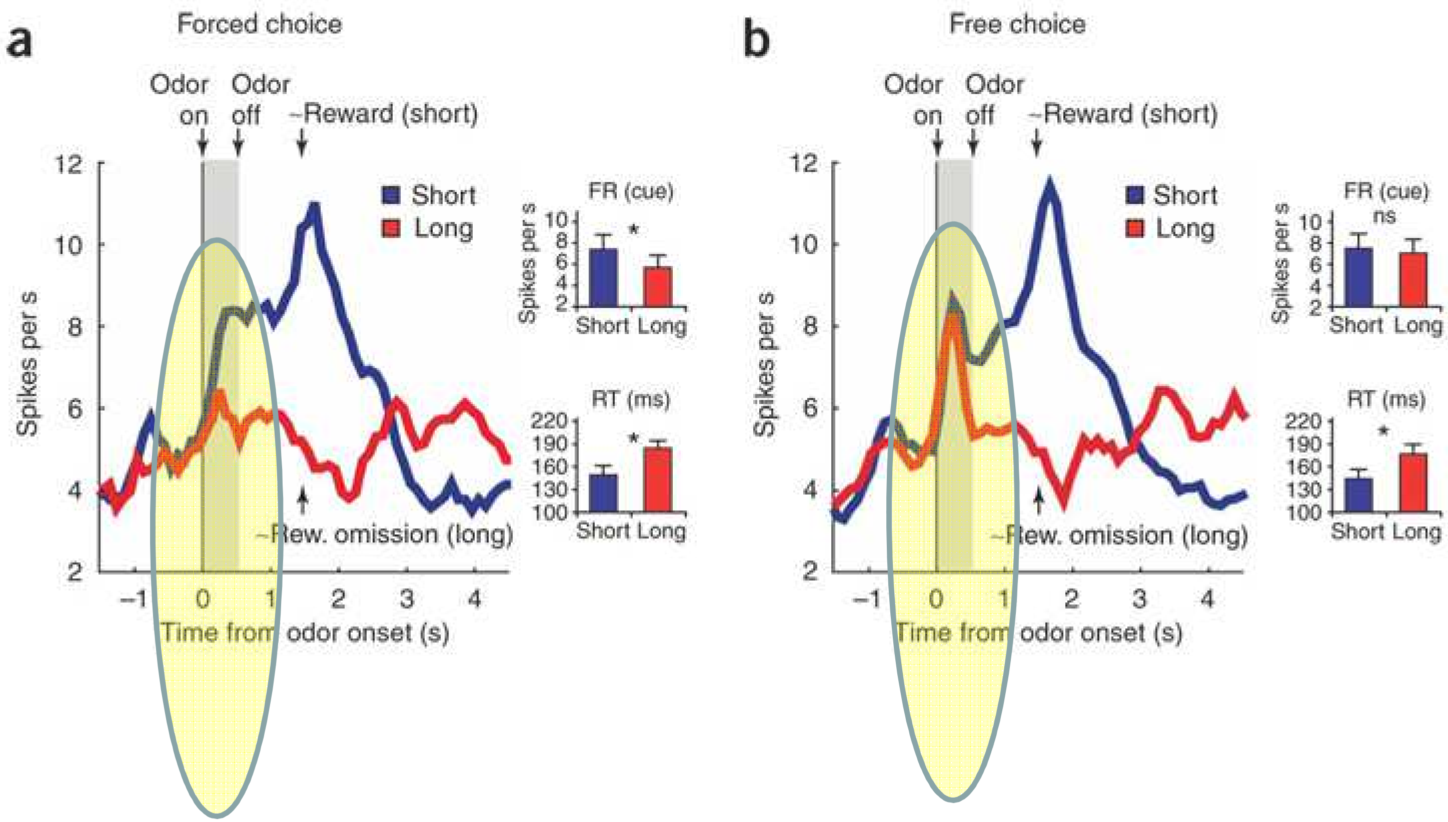


Morris et al, 2006

Variants: Q learning

$$Q^*(1,C) = E\left[r_t + V^*(x_{t+1}) \mid x_t = 1, u_t = C\right]$$

$$Q(1,C) \to Q(1,C) + \varepsilon\left(r_t + \max_u Q(2,u) - Q(1,C)\right)$$



Roesch et al, 2007

Summary

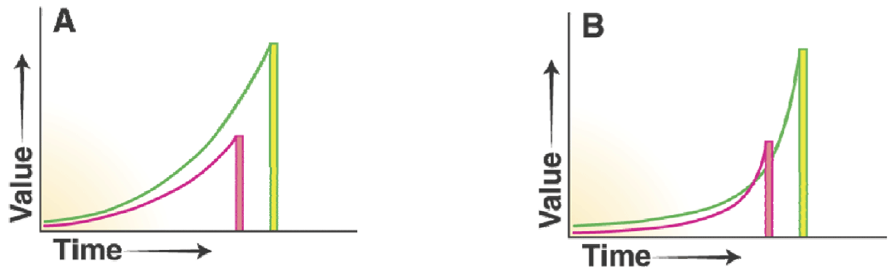
- prediction learning
  - Bellman evaluation
- actor-critic
  - asynchronous policy iteration
- indirect method (Q learning)
  - asynchronous value iteration

$$V^*(1) = E\left[r_t + V^*(x_{t+1}) \mid x_t = 1\right]$$

$$Q^*(1, C) = E\left[r_t + V^*(x_{t+1}) \mid x_t = 1, u_t = C\right]$$

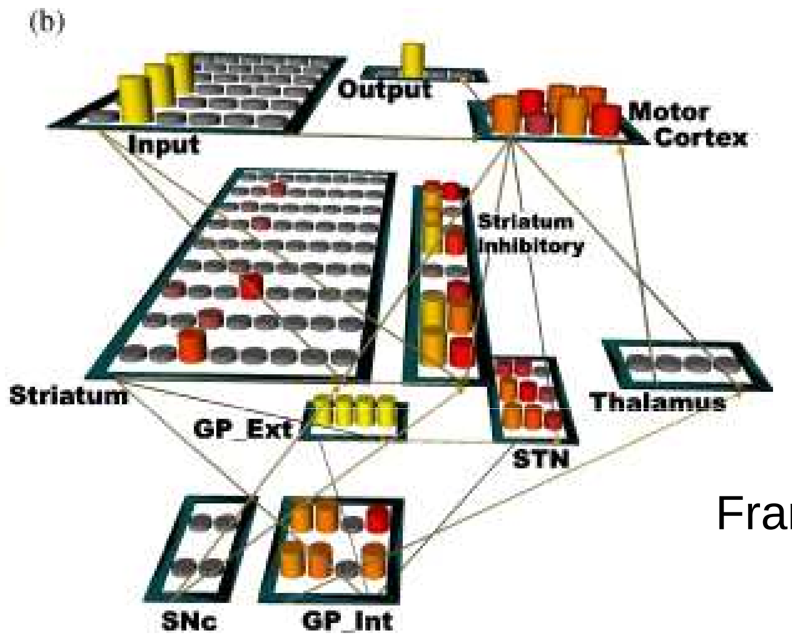
Impulsivity & Hyperbolic Discounting

- humans (and animals) show impulsivity in:
  - diets
  - addiction
  - spending, …
- intertemporal conflict between short and long term choices
- often explained via hyperbolic discount functions



- alternative is Pavlovian imperative to an immediate reinforcer
- framing, trolley dilemmas, etc

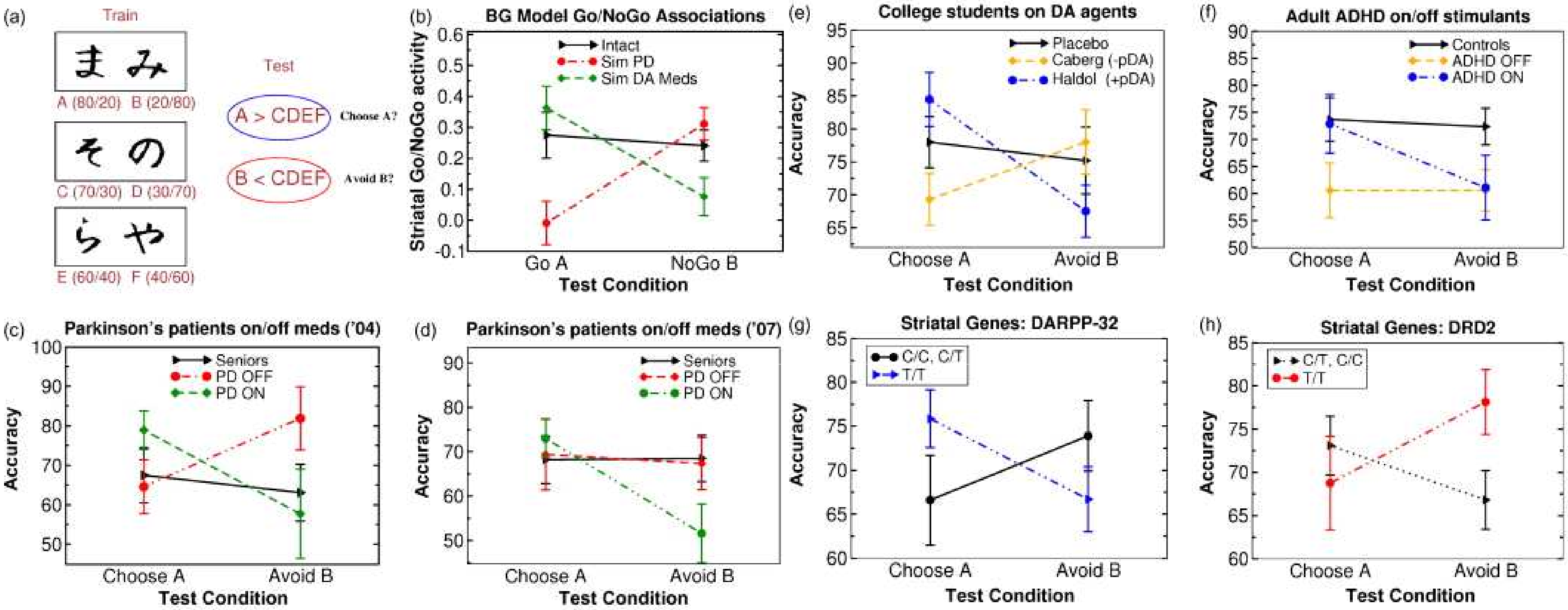
Direct/Indirect Pathways



Frank

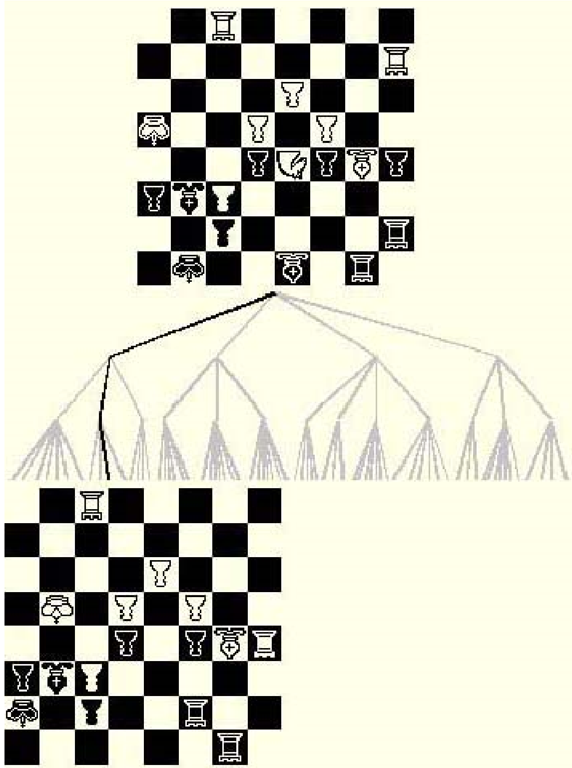
- direct: D1: GO; learn from DA increase
- indirect: D2: noGO; learn from DA decrease
- hyperdirect (STN) delay actions given strongly attractive choices

Frank



- DARPP-32: D1 effect
- DRD2: D2 effect

Three Decision Makers

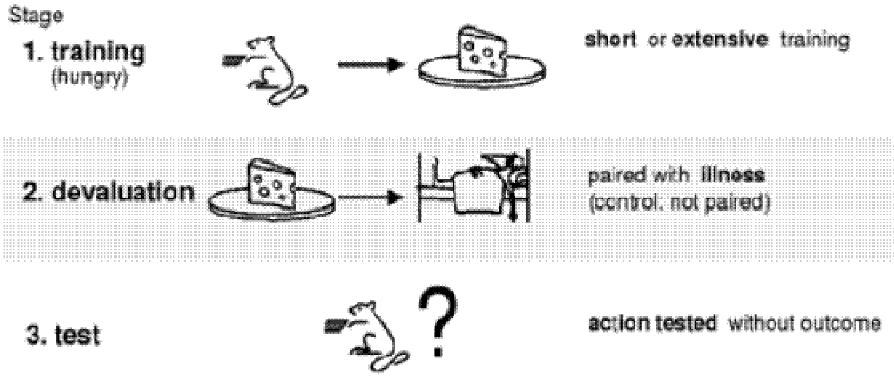


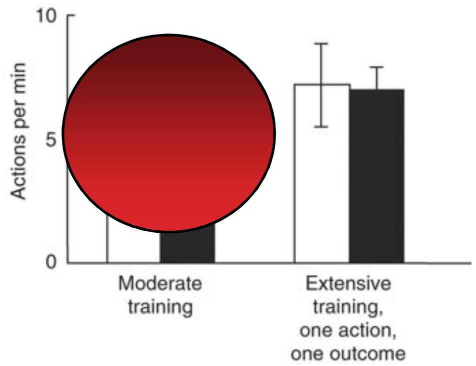
- tree search
- position evaluation
- situation memory

Multiple Systems in RL

- model-based RL
  - build a forward model of the task, outcomes
  - search in the forward model (online DP)
    - optimal use of information
    - computationally ruinous
- cached-based RL
  - learn Q values, which summarize future worth
    - computationally trivial
    - bootstrap-based; so statistically inefficient
- learn both – select according to uncertainty

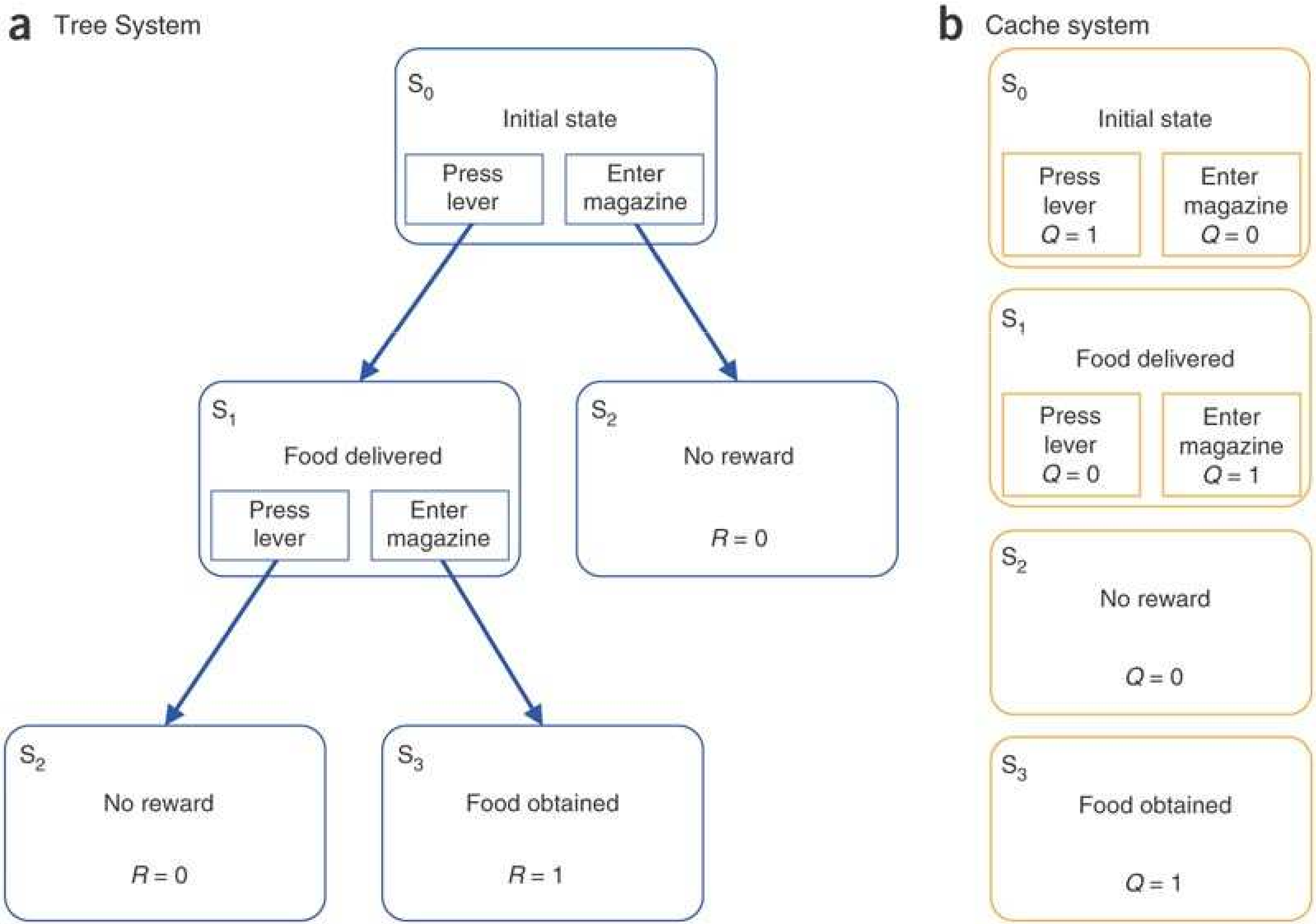
Animal Canary



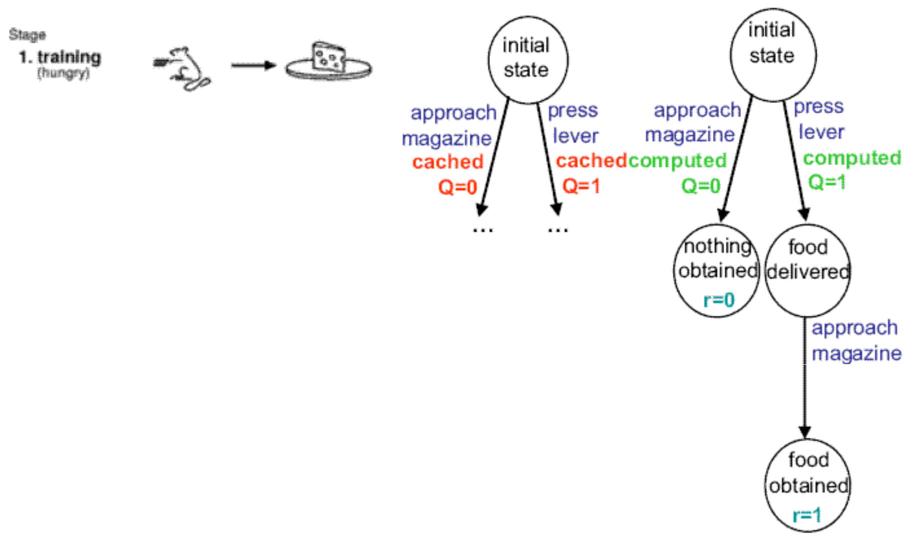


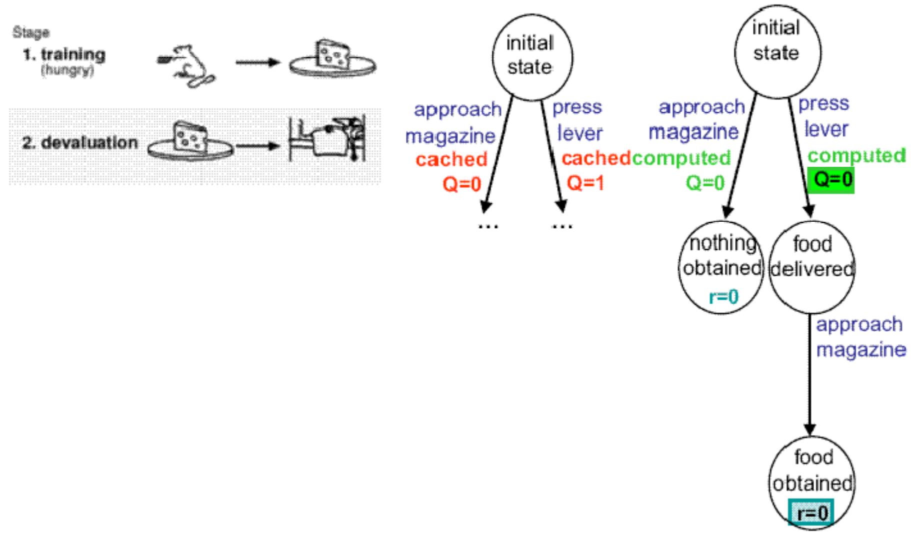
- OFC; dlPFC; dorsomedial striatum; BLA?
- dosolateral striatum, amygdala

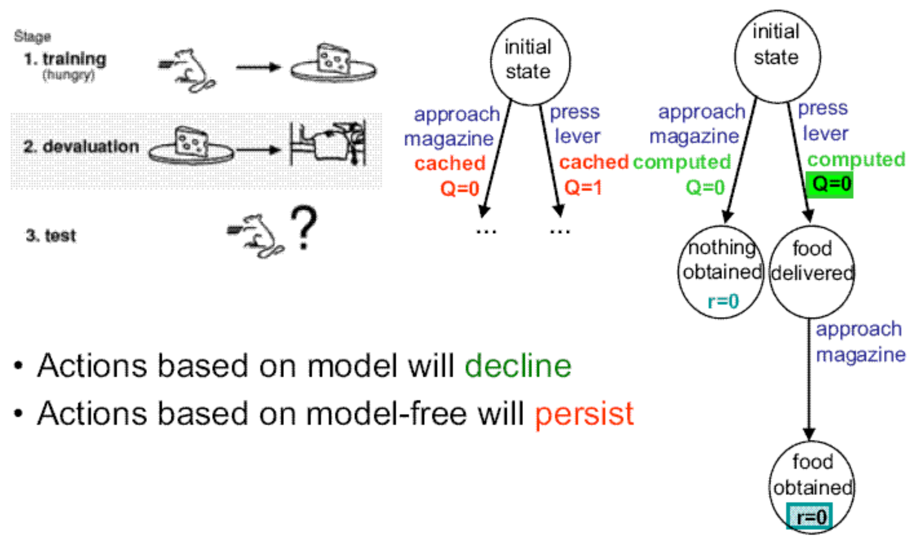
Two Systems:


a Tree System
S0
Initial state
Press lever
Enter magazine
S1
Food delivered
Press lever
Enter magazine
S2
No reward
R = 0
S2
No reward
R = 0
S3
Food obtained
R = 1
b Cache system
S0
Initial state
Press lever Q = 1
Enter magazine Q = 0
S1
Food delivered
Press lever Q = 0
Enter magazine Q = 1
S2
No reward
Q = 0
S3
Food obtained
Q = 1

Behavioural Effects

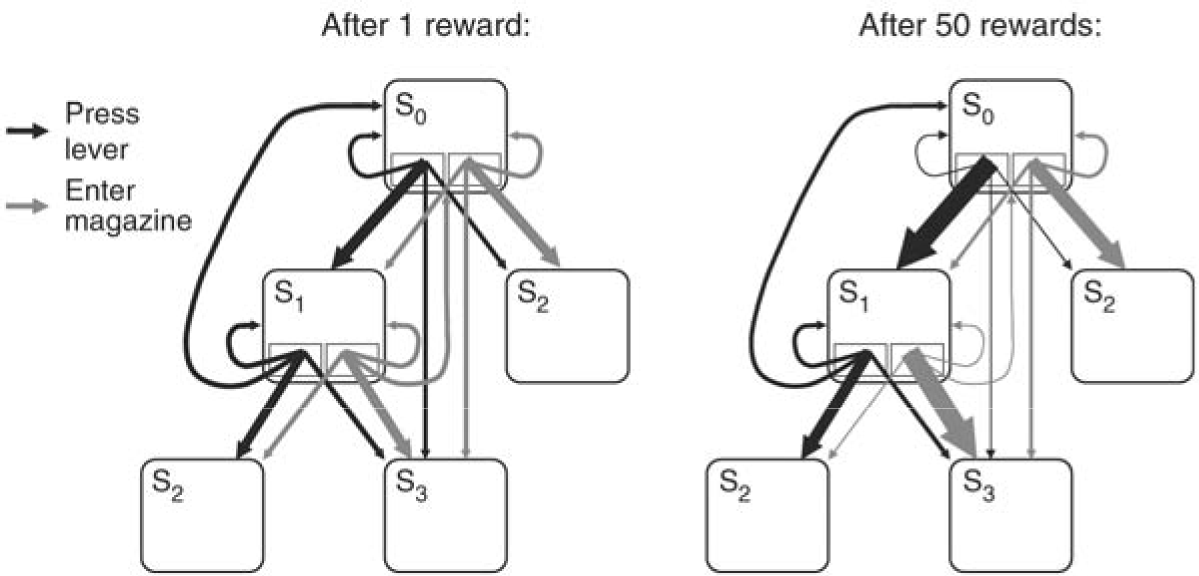


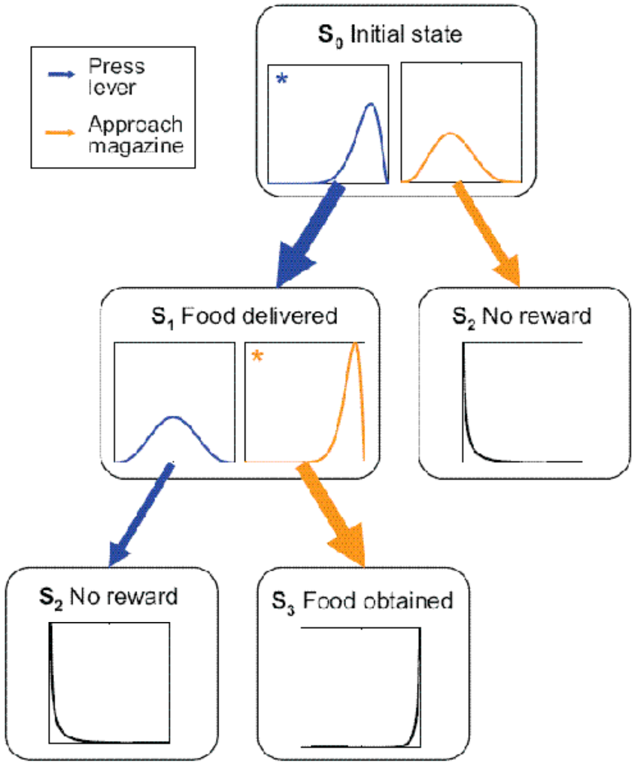




- Actions based on model will decline
- Actions based on model-free will persist

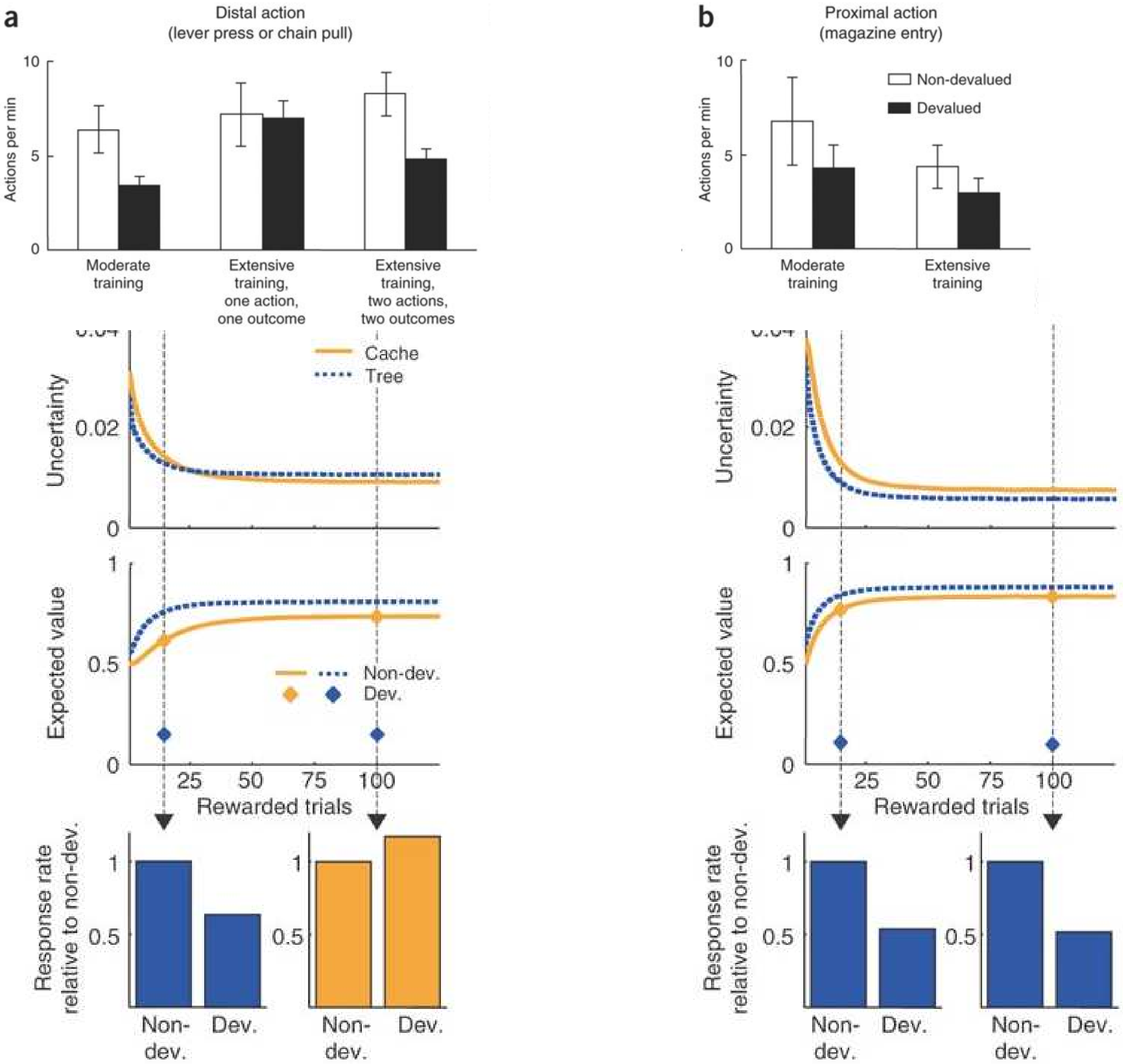
Effects of Learning





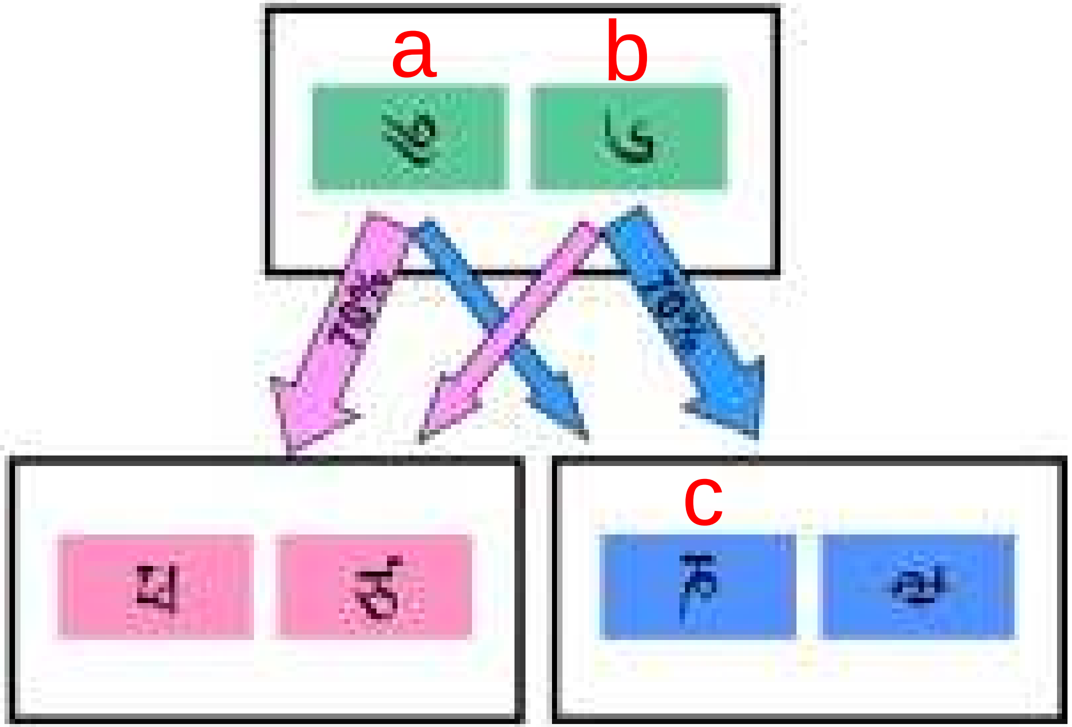
- distributional value iteration
  - (Bayesian Q learning)
- fixed additional uncertainty per step

One Outcome



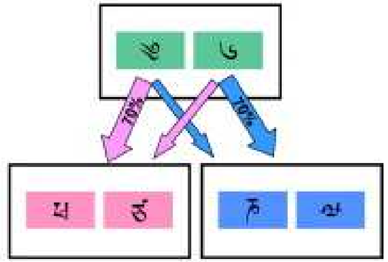
shallow tree implies goal-directed control wins

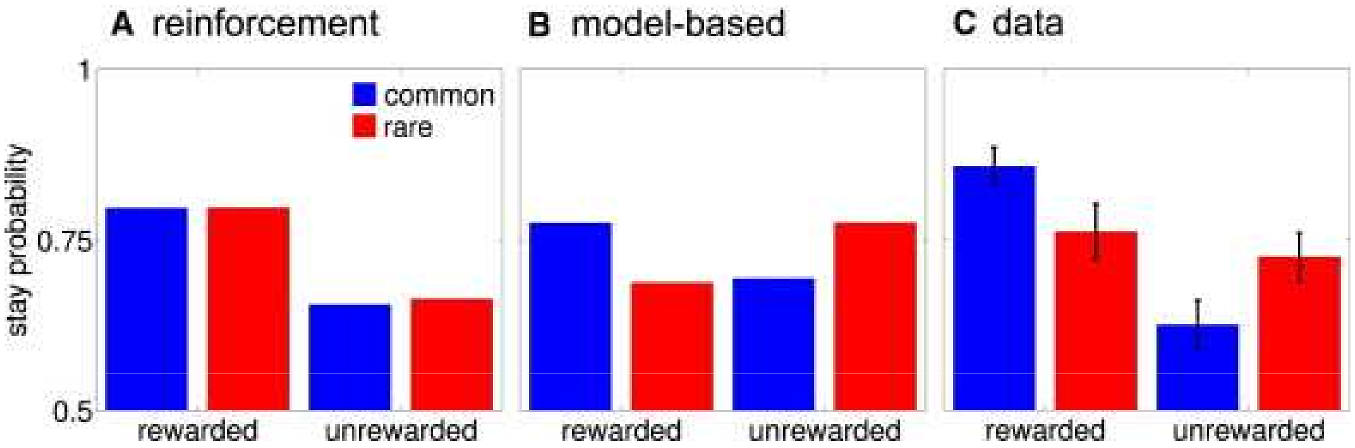
Human Canary...



- if $a \rightarrow c$ and $c \rightarrow$ **£££** , then do more of a or b?
  - MB: b
  - MF: a (or even no effect)

Behaviour



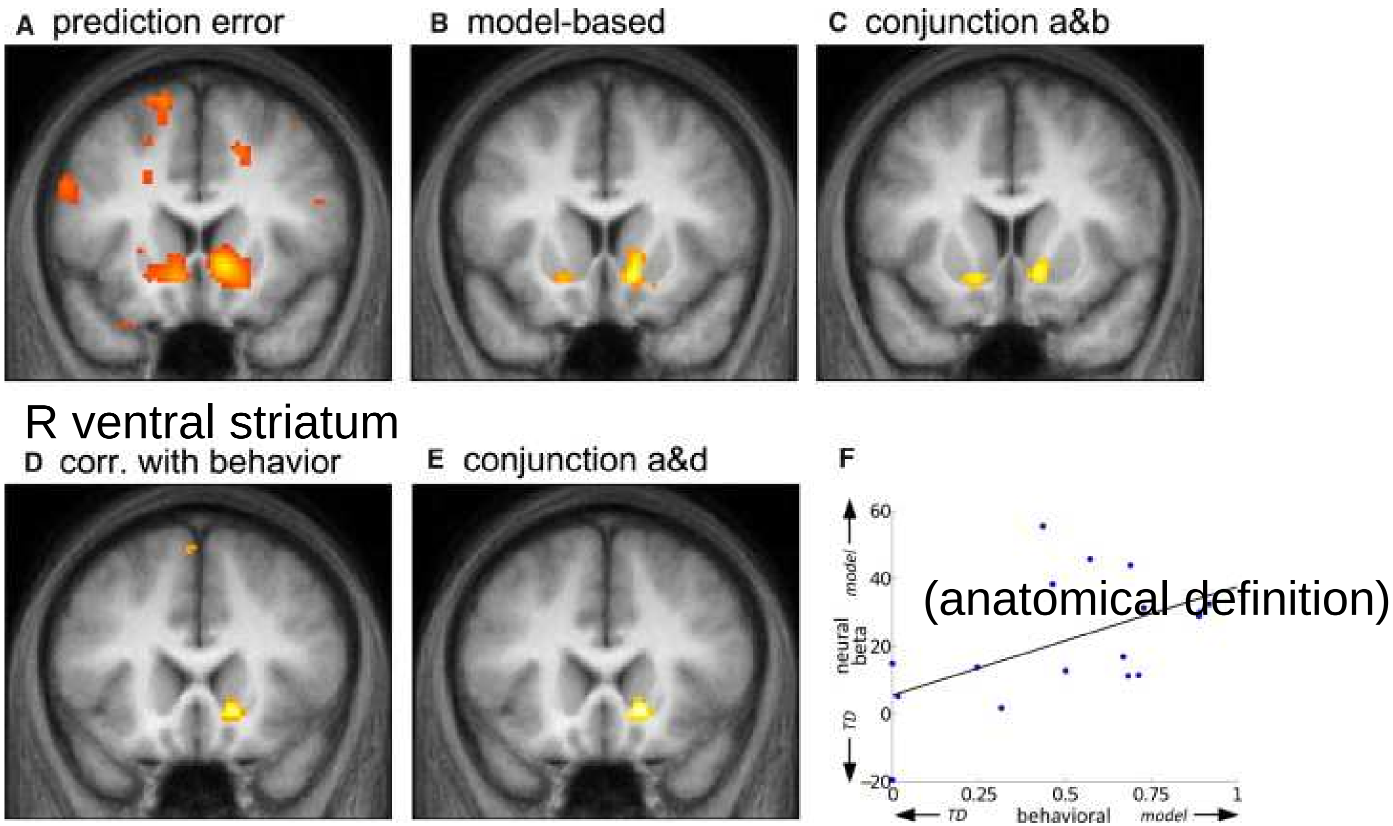


- action values depend on both systems:

$$Q_{tot}(x,u) = Q_{MF}(x,u) + \beta Q_{MB}(x,u)$$

- expect that β will vary by subject (but be fixed)

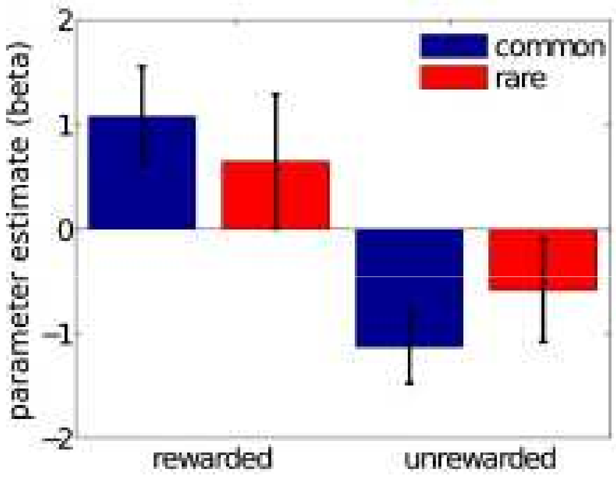
Neural Prediction Errors (1→2)

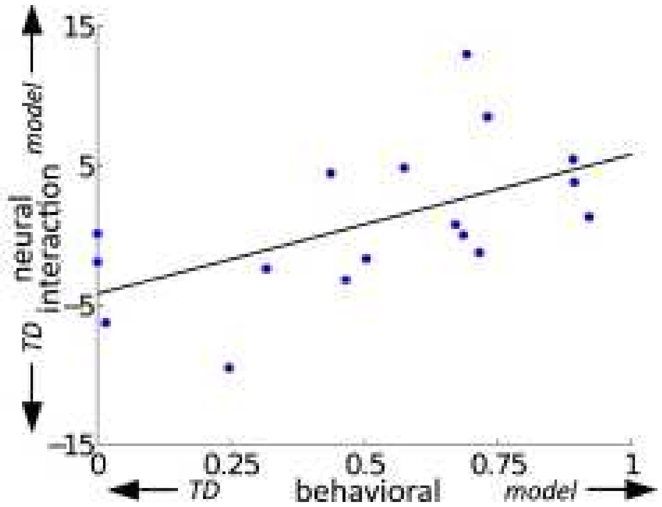


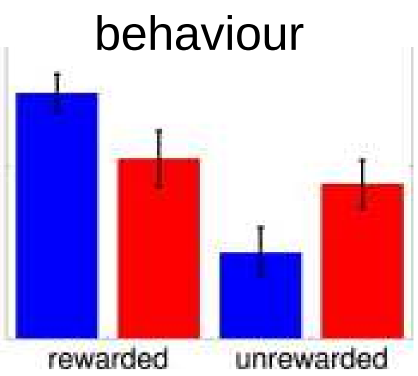
- note that MB RL does not use this prediction error – training signal?

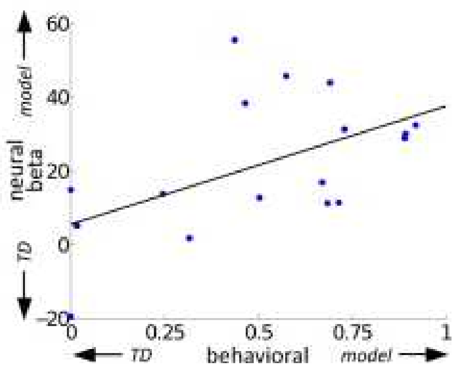
Neural Prediction Errors (1)

- right nucleus accumbens









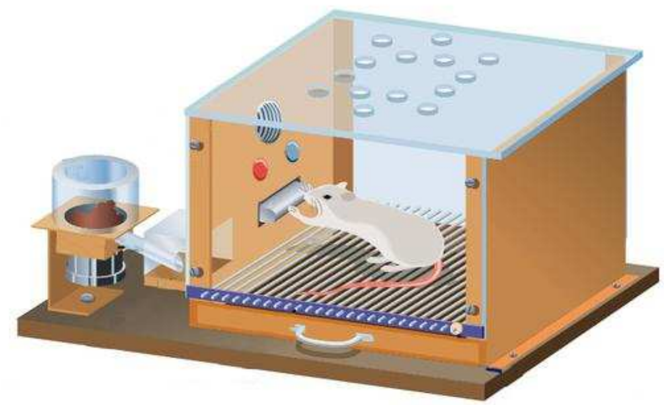
1-2, not 1

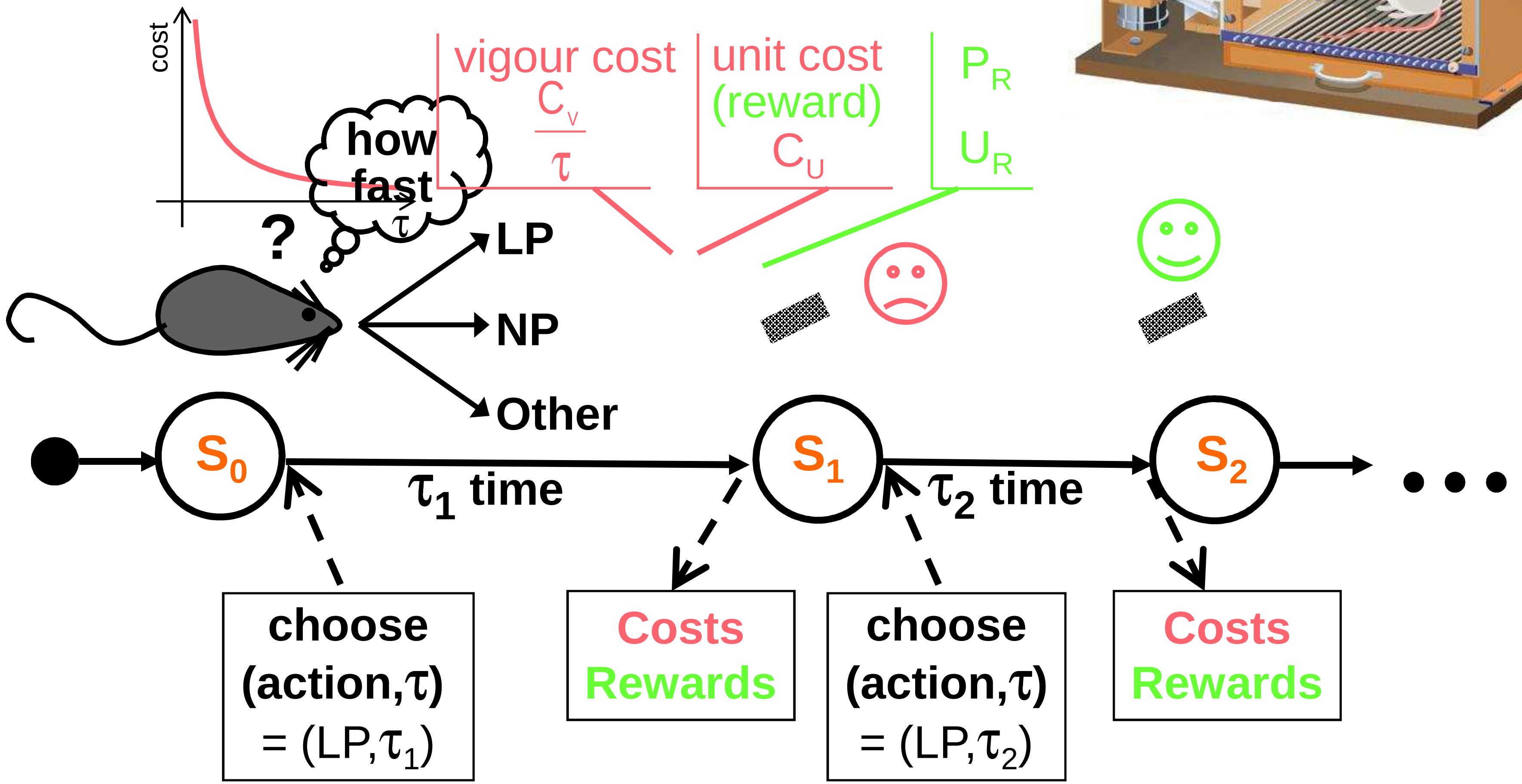
Vigour

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

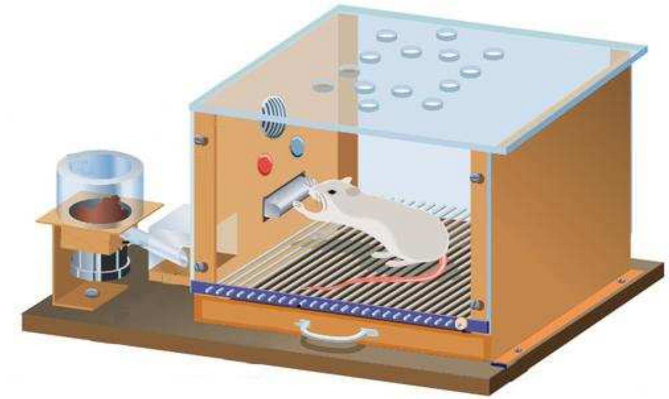
- real-valued DP

The model

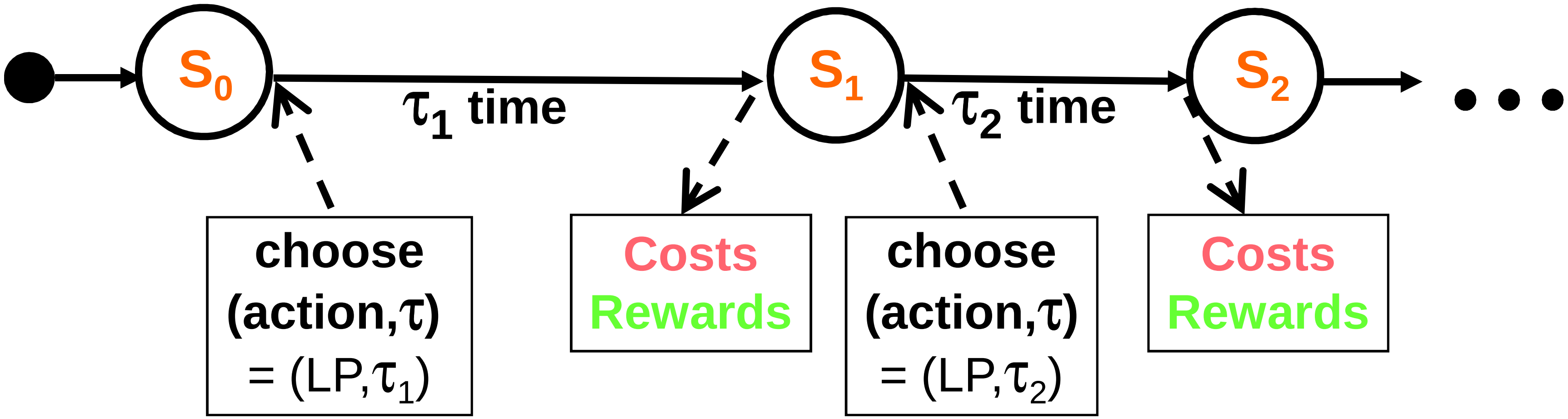



cost
τ
how fast
?
LP
NP
Other
vigour cost
$\frac{C_v}{\tau}$
unit cost
(reward)
C_U
P_R
U_R
S_0
τ_1 time
S_1
τ_2 time
S_2
choose
(action,τ)
= (LP,τ_1)
Costs
Rewards
choose
(action,τ)
= (LP,τ_2)
Costs
Rewards

The model



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



Average Reward RL

Compute differential values of actions

Differential value of taking action L with latency τ when in state x

$\bar{\rho}$ = average rewards minus costs, per unit time

$$Q_{L,\tau}(x) \quad = \quad \text{Rewards} \quad - \quad \underset{C_u + C_v/\tau}{\text{Costs}} \quad + \quad \underset{V(x')}{\text{Future Returns}} \quad - \quad \tau\rho$$

- steady state behavior (not learning dynamics)

(Extension of Schwartz 1993)

Average Reward Cost/benefit Tradeoffs

1. Which action to take?

$\Rightarrow$ Choose action with largest expected reward minus cost

2. How fast to perform it?

• slow $\rightarrow$ less costly (vigour cost)	• slow $\rightarrow$ delays (all) rewards • net rate of rewards = cost of delay (opportunity cost of time)

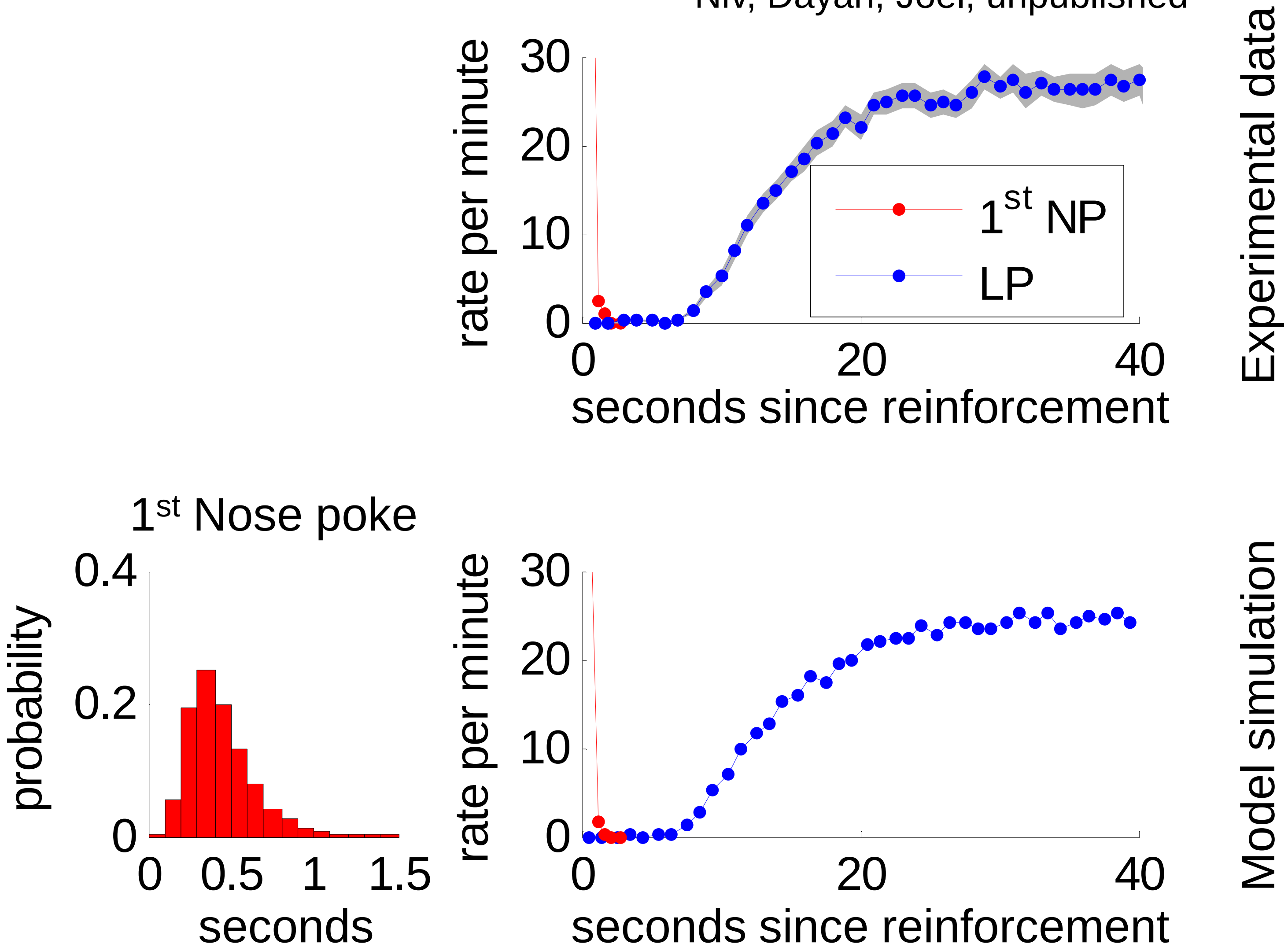
$\Rightarrow$ Choose rate that balances vigour and opportunity costs

explains faster (irrelevant) actions under hunger, etc

masochism

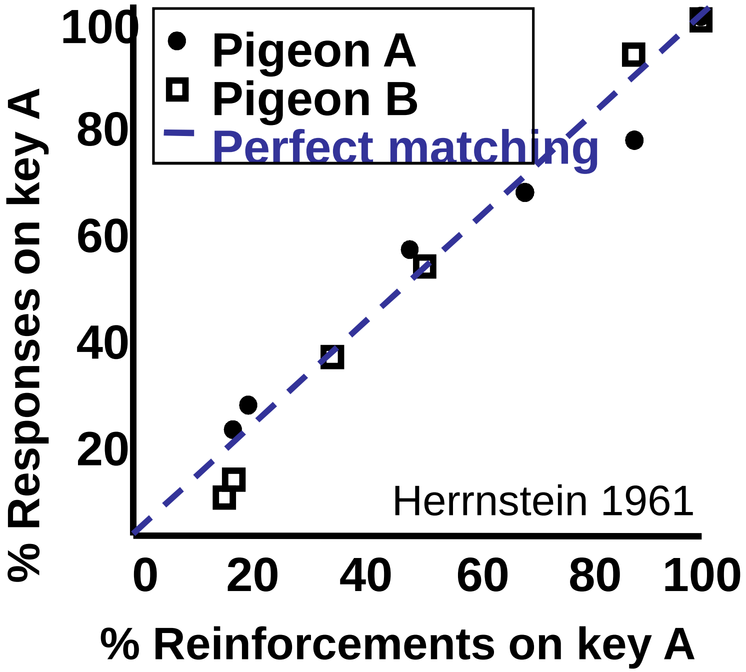
Optimal response rates

Niv, Dayan, Joel, unpublished

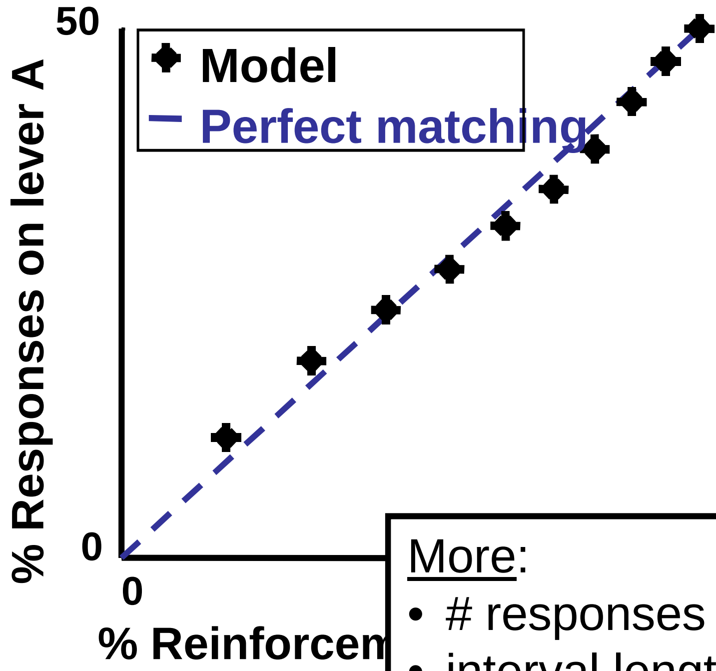


Optimal response rates

Experimental data



Model simulation



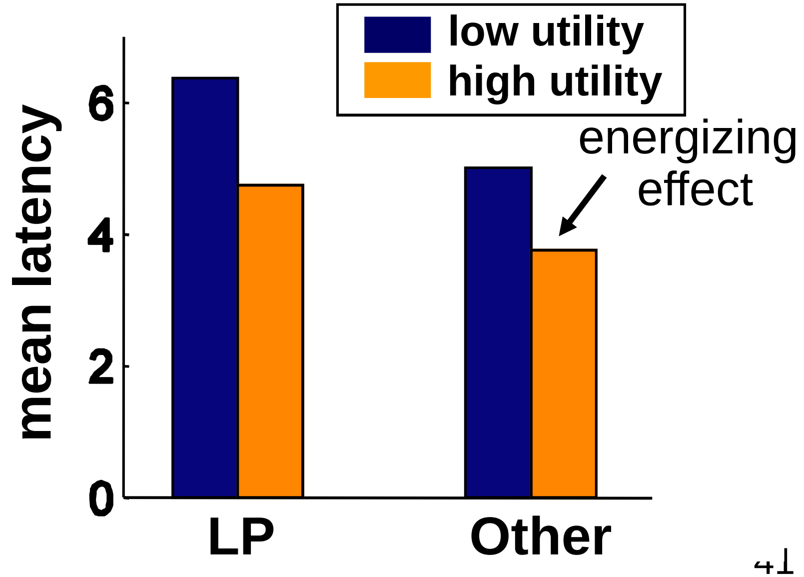
<u>More</u>:
- # responses
- interval length
- amount of reward
- ratio vs. interval
- breaking point
- temporal structure
- etc.

Effects of motivation (in the model)

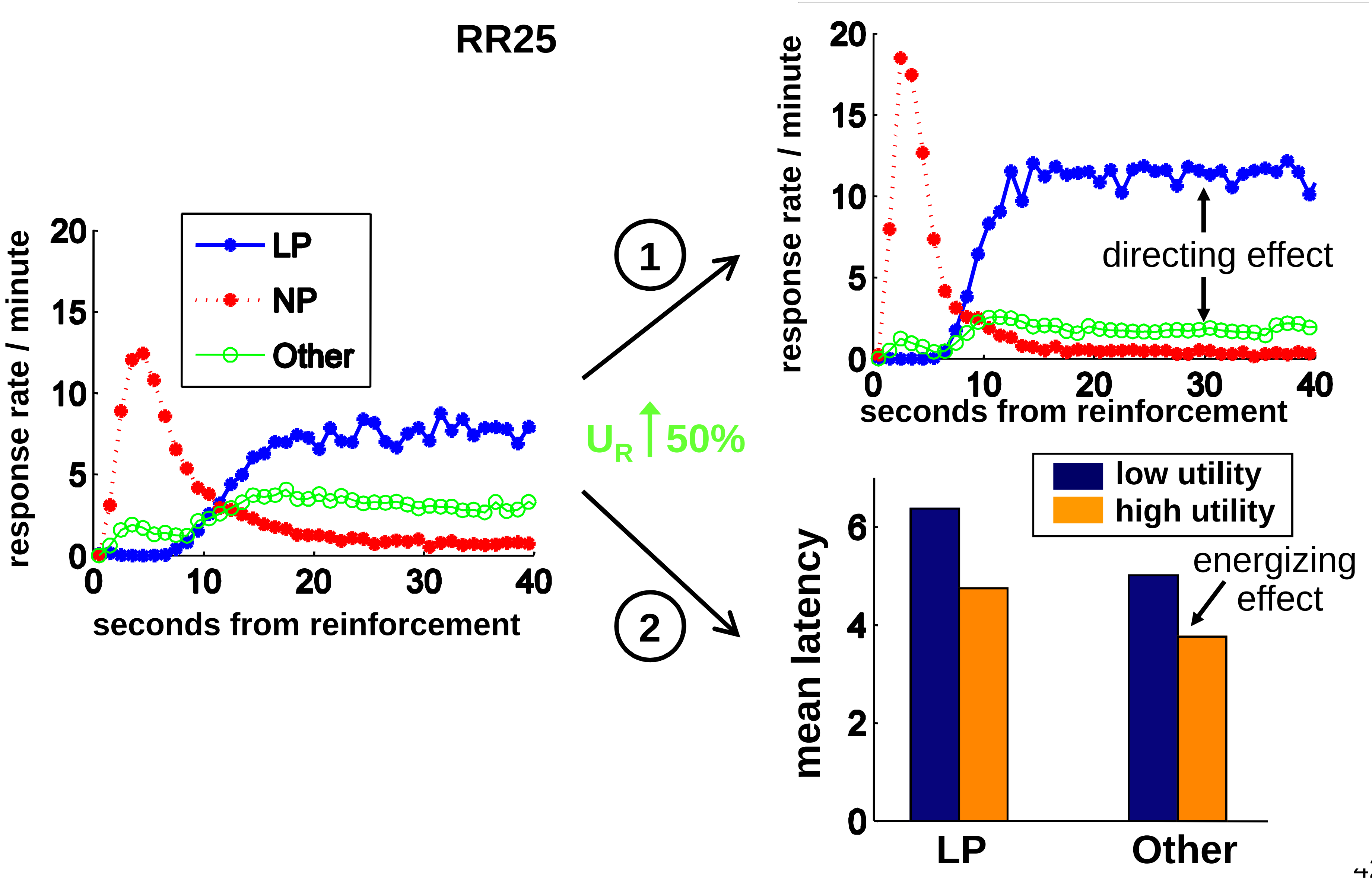
RR25

$$Q(x,u,\tau) = p_r R - C_u - \frac{C_v}{\tau} + V(x') - \tau \cdot \overline{R}$$

$$\frac{\partial Q(x,u,\tau)}{\partial \tau} = \frac{C_v}{\tau^2} - \overline{R} = 0 \quad \Longrightarrow \quad \boxed{\tau_{opt} = \sqrt{\frac{C_v}{\overline{R}_{opt}}}}$$

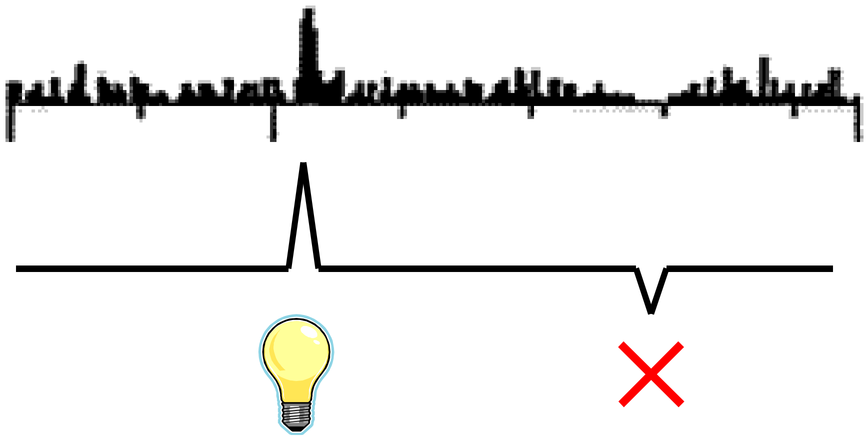


Effects of motivation (in the model)

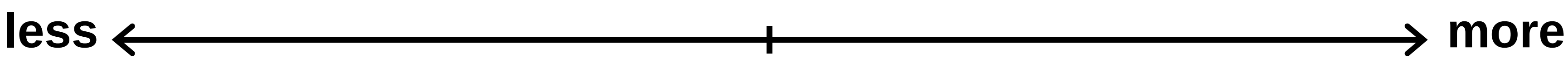


Relation to Dopamine

Phasic dopamine firing = reward prediction error

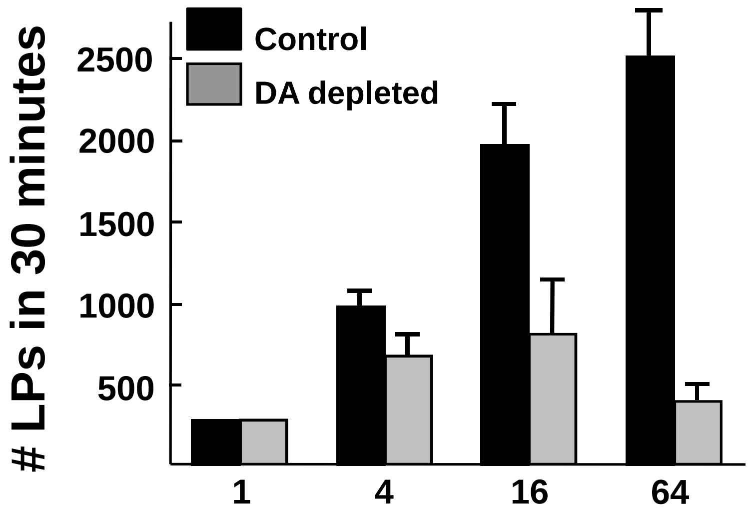


What about tonic dopamine?

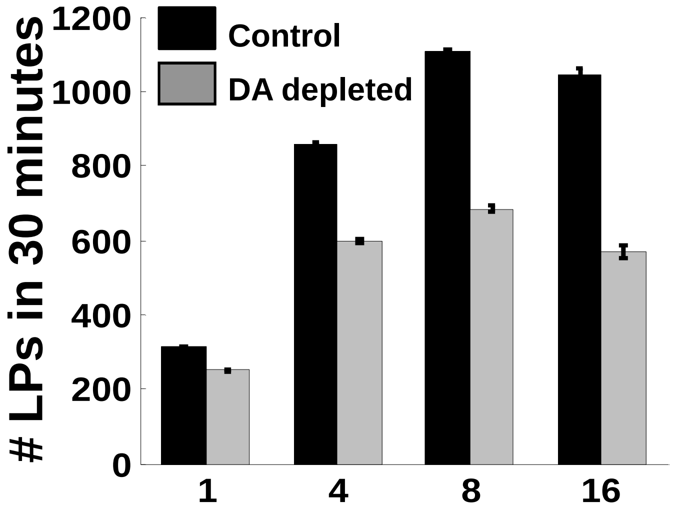


Tonic dopamine = Average reward rate

1. explains pharmacological manipulations
2. dopamine control of vigour through BG pathways



Aberman and Salamone 1999

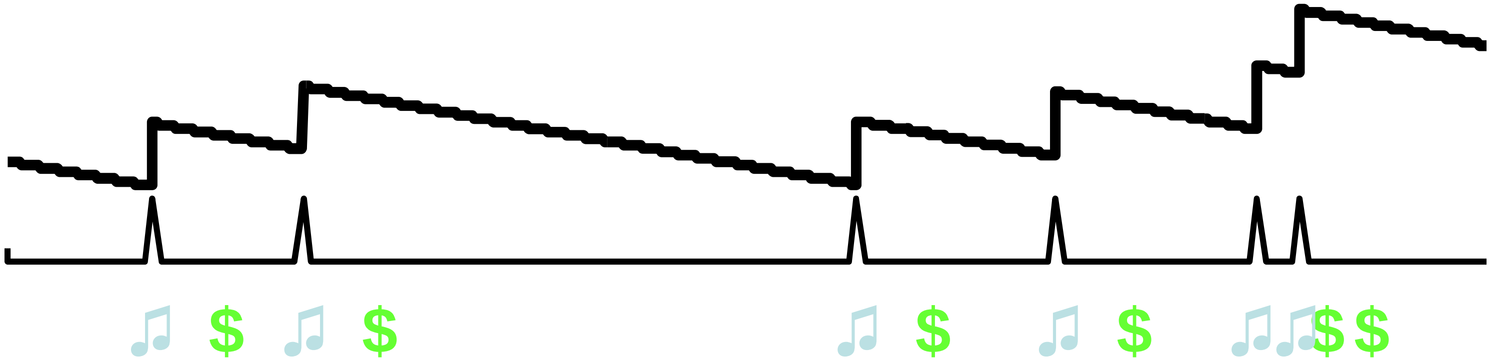


Model simulation

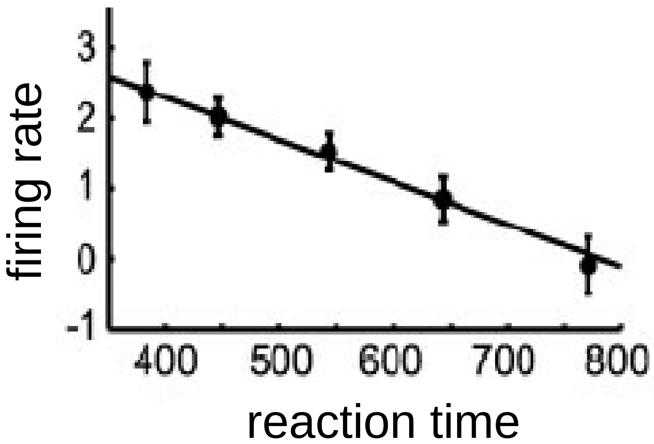
- eating time confound
- context/state dependence (motivation & drugs?)
- less switching=perseveration

NB. phasic signal RPE for choice/value learning

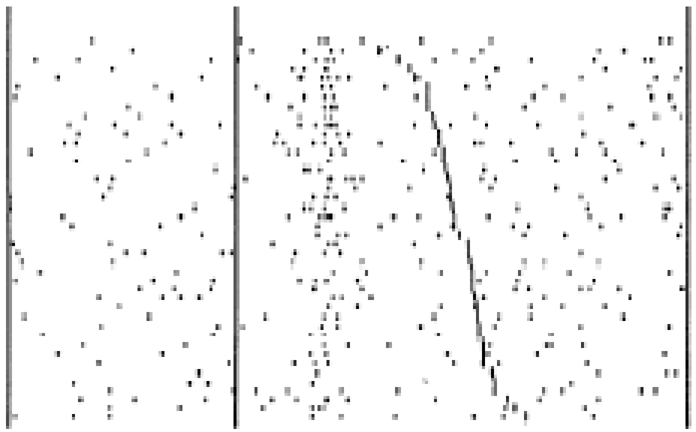
Tonic dopamine hypothesis



…also explains effects of phasic dopamine on response times



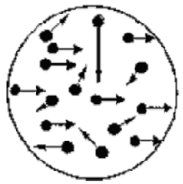
Satoh and Kimura 2003



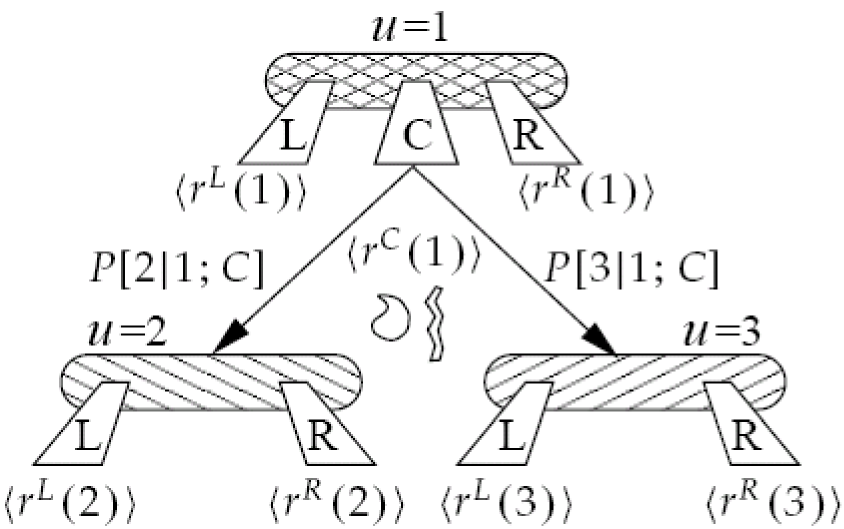
Ljungberg, Apicella and Schultz 1992

Sensory Decisions as Optimal Stopping

- consider listening to:



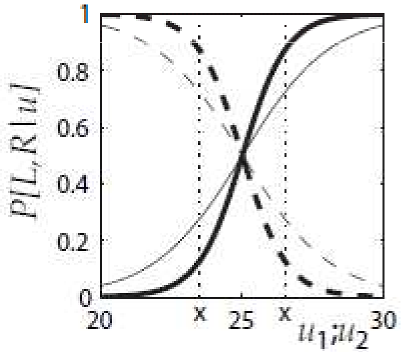
- decision: choose, or sample



Optimal Stopping

- equivalent of state u=1 is $u_1 = n_1$

$$P[L|u_1] = \sigma\left(d'\frac{n_1 - n_{\text{ave}}}{\sigma_n}\right) \text{ for } n_{\text{ave}} = \frac{1}{2}(n^L + n^R)$$



$$\sigma = 2.5$$
$$\left\langle r^C \right\rangle = -0.1$$

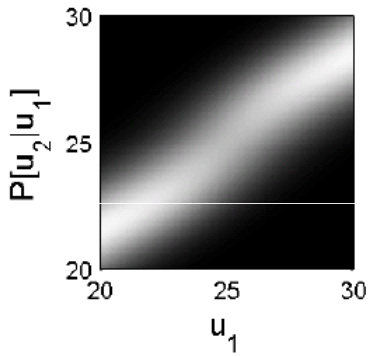
- and states u=2,3 is $u_2 = \frac{1}{2}\left(n_1 + n_2\right)$

$$P[L|u_2] = \sigma(2d'(u_2 - n_{\text{ave}})/\sigma_n)$$

Transition Probabilities

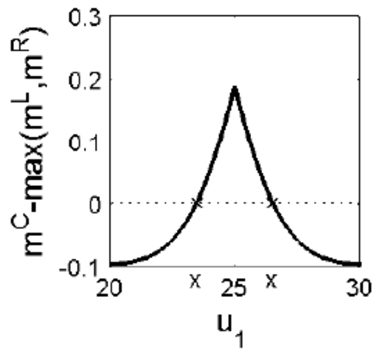
$$p[u_2|u_1;C] = P[L|u_1]p[u_2|L,u_1] + P[R|u_1]p[u_2|R,u_1]$$

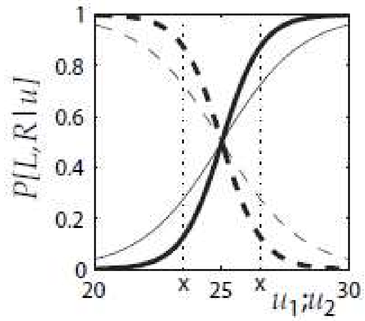
$$= P[L|u_1]\mathcal{N}(2u_2 - u_1; n^L, \sigma_n^2) + P[R|u_1]\mathcal{N}(2u_2 - u_1; n^R, \sigma_n^2)$$



$$m^C(u_1) = \langle r^C(u_1) + v(u_2)\rangle_{u_1}$$

$$= r^C(u_1) + \int_{u_2} du_2\ p[u_2|u_1;C]\max\{P[L|u_2], P[R|u_2]\}$$





Computational Neuromodulation

- dopamine
  - phasic: prediction error for reward
  - tonic: average reward (vigour)
- serotonin
  - phasic: prediction error for punishment?
- acetylcholine:
  - expected uncertainty?
- norepinephrine
  - unexpected uncertainty; neural interrupt?

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

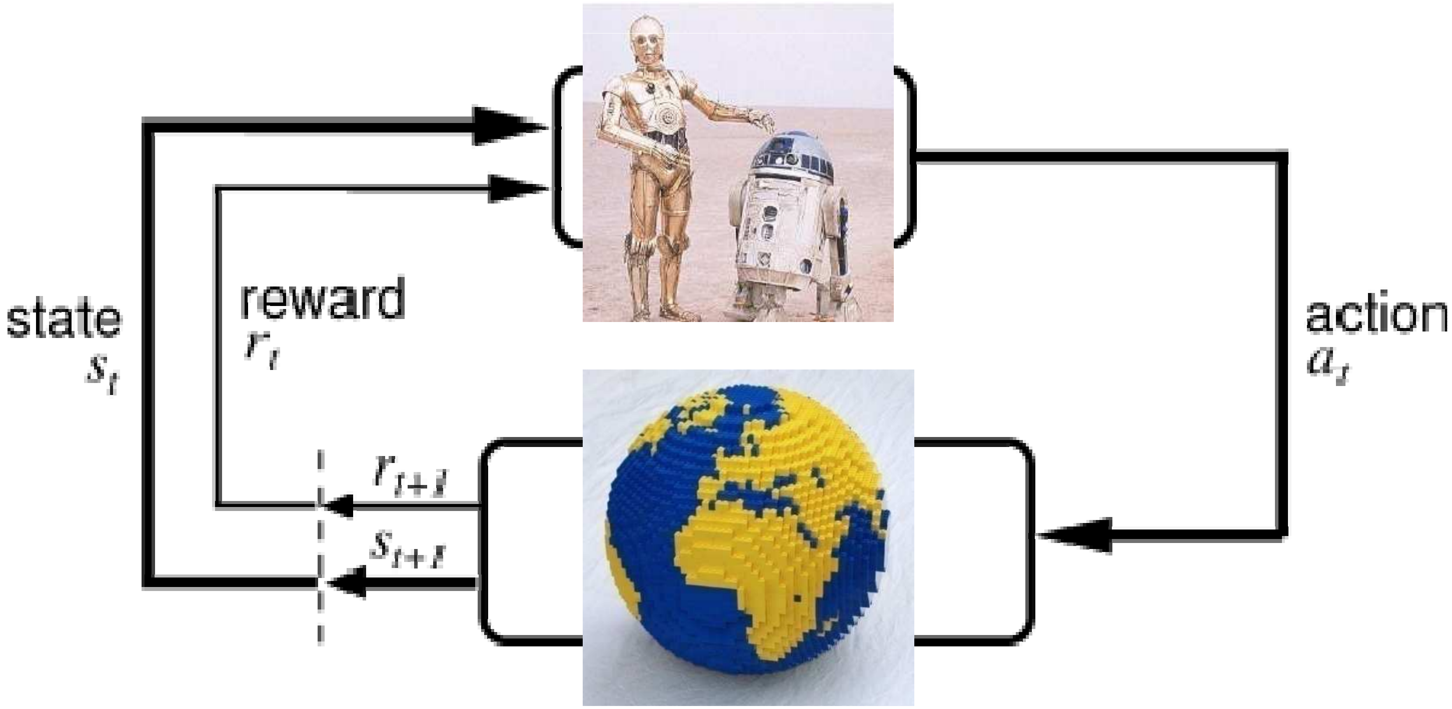
neuromodulators; amygdala; OFC
nucleus accumbens; dorsal striatum

Markov Decision Process

class of stylized tasks with

states, actions & rewards

- at each timestep *t* the world takes on state s_t and delivers reward r_t, and the agent chooses an action a_t



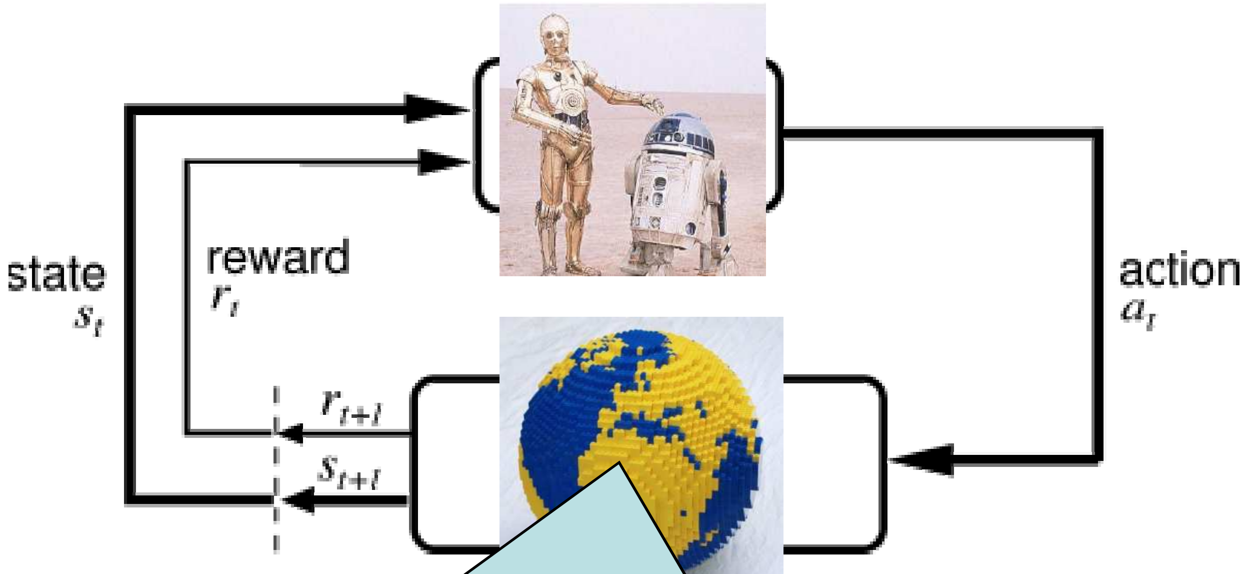
Markov Decision Process

World: You are in state 34.
Your immediate reward is 3. You have 3 actions.
Robot: I'll take action 2.

World: You are in state 77.
Your immediate reward is -7. You have 2 actions.
Robot: I'll take action 1.

World: You're in state 34 (again).
Your immediate reward is 3. You have 3 actions.

Markov Decision Process



Stochastic process defined by:

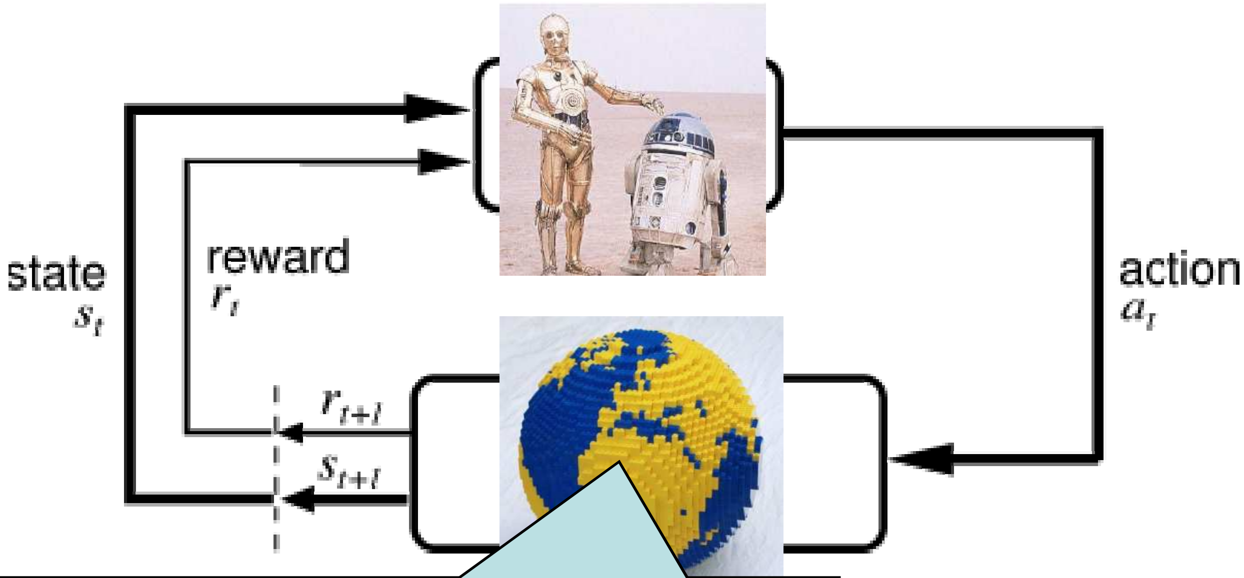
- –reward function:

  $r_t \sim P(r_t \mid s_t)$

- –transition function:

  $s_t \sim P(s_{t+1} \mid s_t, a_t)$

Markov Decision Process



Stochastic process defined by:

- reward function:

  $r_t \sim P(r_t \mid s_t)$

- transition function:

  $s_t \sim P(s_{t+1} \mid s_t, a_t)$

Markov property

- future conditionally independent of past, given s_t

The optimal policy

Definition: a policy such that at every state, its expected value is better than (or equal to) that of all other policies

Theorem: For every MDP there exists (at least) one deterministic optimal policy.

→ by the way, why is the optimal policy just a mapping from states to actions? couldn't you earn more reward by choosing a different action depending on last 2 states?

Pavlovian & Instrumental Conditioning

- Pavlovian
  - learning values and predictions
  - using TD error
- Instrumental
  - learning actions:
    - by reinforcement (leg flexion)
    - by (TD) critic
  - (actually different forms: goal directed & habitual)

Pavlovian-Instrumental Interactions

- synergistic
  - conditioned reinforcement
  - Pavlovian-instrumental transfer
    - Pavlovian cue predicts the instrumental outcome
    - behavioural inhibition to avoid aversive outcomes
- neutral
  - Pavlovian-instrumental transfer
    - Pavlovian cue predicts outcome with same motivational valence
- opponent
  - Pavlovian-instrumental transfer
    - Pavlovian cue predicts opposite motivational valence
  - **negative automaintenance**

-ve Automaintenance in Autoshaping

- simple choice task
  - N: nogo gives reward r=1
  - G: go gives reward r=0
- learn three quantities
  - average value
  - *Q* value for N
  - *Q* value for G

$$v(t+1) = v(t) + \eta(r(t) - v(t))$$

$$q_{\mathsf{N}}(t+1) = q_{\mathsf{N}}(t) + \eta\left(r(t) - q_{\mathsf{N}}(t)\right)$$

$$q_{\mathsf{G}}(t+1) = q_{\mathsf{G}}(t) + \eta\left(r(t) - q_{\mathsf{G}}(t)\right)$$

- instrumental propensity is

$$p(a(t) = \mathsf{N}) = \frac{e^{\mu(q_{\mathsf{N}}(t) - v(t))}}{e^{\mu(q_{\mathsf{N}}(t) - v(t))} + e^{\mu(q_{\mathsf{G}}(t) - v(t))}}$$

$$= \sigma\left(\mu(q_{\mathsf{N}}(t) - q_{\mathsf{G}}(t))\right)$$

-ve Automaintenance in Autoshaping

- Pavlovian action
  - assert: Pavlovian impetus towards G is *v(t)*
  - weight Pavlovian and instrumental advantages by ω – competitive reliability of Pavlov
- new propensities

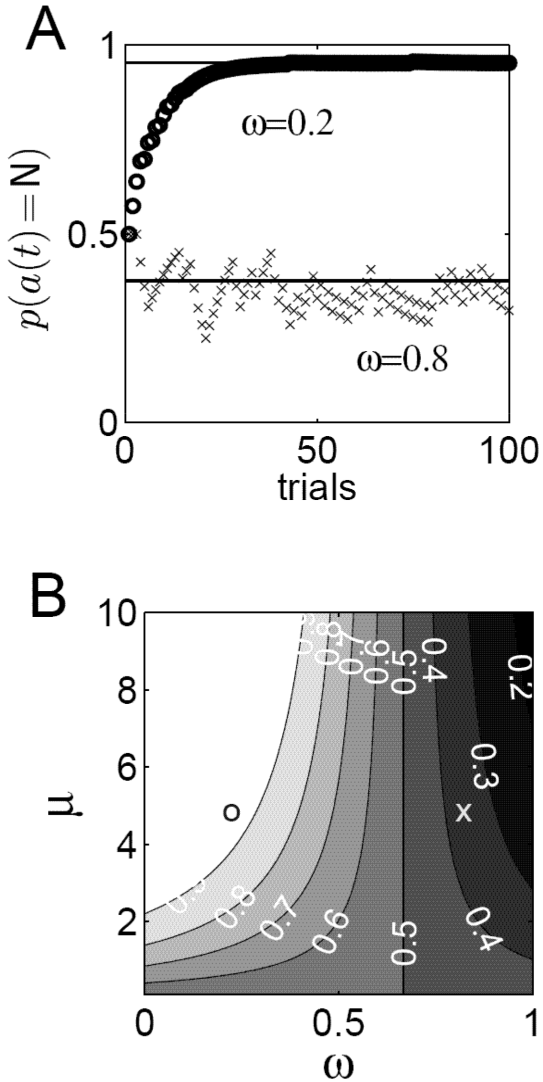
$$\mathsf{N}: \quad q_{\mathsf{N}}(t) - v(t) \quad \Rightarrow \quad (1-\omega)(q_{\mathsf{N}}(t) - v(t))$$

$$\mathsf{G}: \quad q_{\mathsf{G}}(t) - v(t) \quad \Rightarrow \quad (1-\omega)(q_{\mathsf{G}}(t) - v(t)) + \omega v(t)$$

- new action choice

$$p(a(t) = \mathsf{N}) = \sigma\left(\mu\left((1-\omega)(q_{\mathsf{N}}(t) - q_{\mathsf{G}}(t)) - \omega v(t)\right)\right)$$

-ve Automaintenance in Autoshaping



- basic –ve automaintenance effect (μ=5)
- lines are theoretical asymptotes
- equilibrium probabilities of action