Reinforcement Learning

Peter Dayan
GCNU

Dayan & Abbott
Theoretical Neuroscience:
Computational and Mathematical Modeling of Neural Systems.
MIT Press, 2001.

Reinforcement Learning

Computational analysis of prediction and action selection

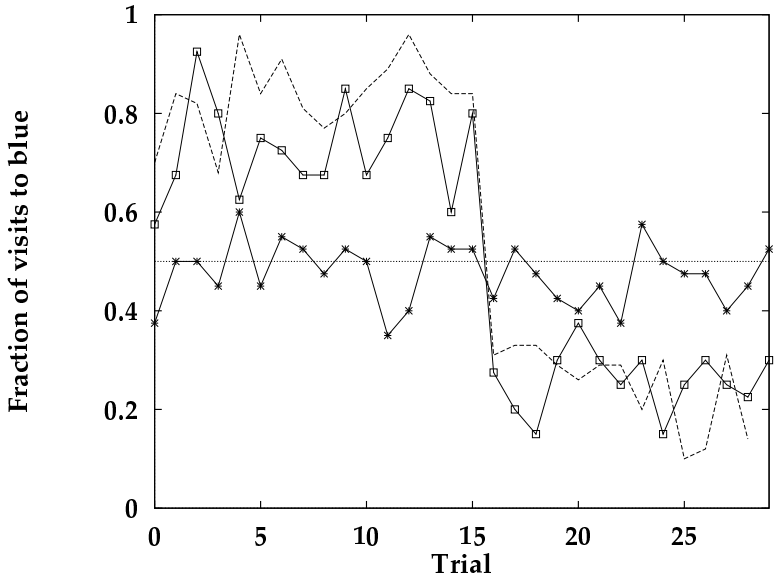
Ethology	**Computation**
optimality	dynamic programming
appropriateness	Kalman filtering
Psychology	**Algorithm**
classical/operant	TD/delta rules
conditioning	simple weights
Neurobiology	**Implementation**
'neuromodulatory' systems	
striatum	

Data

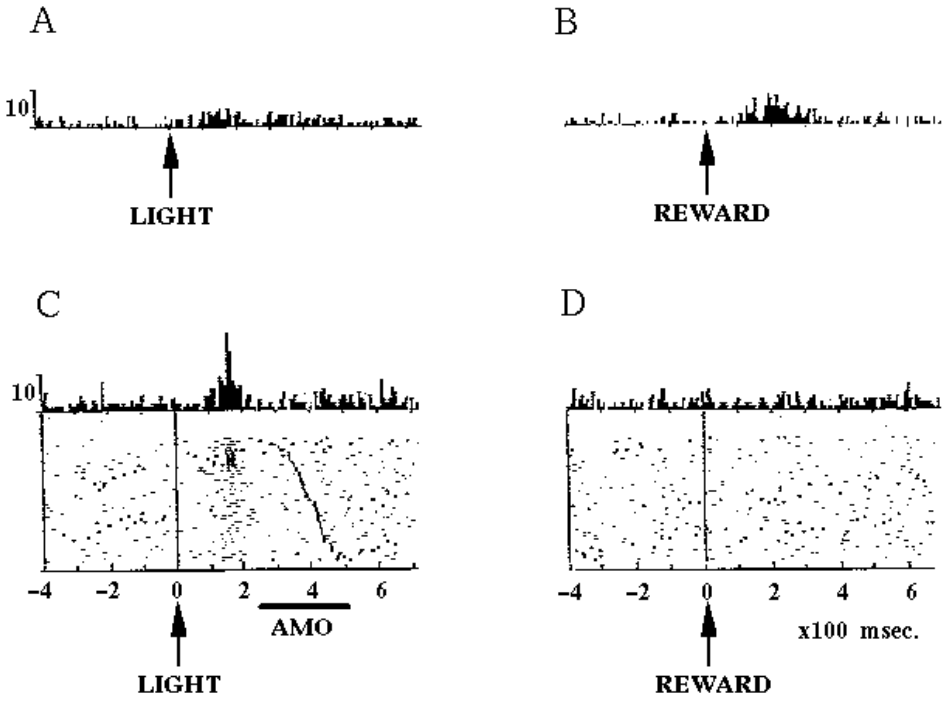
Four results in classical conditioning:

Name	Set 1	Set 2	Test
Pavlovian		$u_1 \rightarrow$ food	$u_1 \rightarrow +$
Inhibitory		$u_1 \rightarrow$ food $u_1 + u_2 \rightarrow \cdot$	$u_2 \rightarrow —$
Blocking	$u_1 \rightarrow$ food	$u_1 + u_2 \rightarrow$ food	$u_2 \rightarrow \cdot$
Secondary	$u_1 \rightarrow$ food	$u_2 \rightarrow u_1$	$u_2 \rightarrow +$

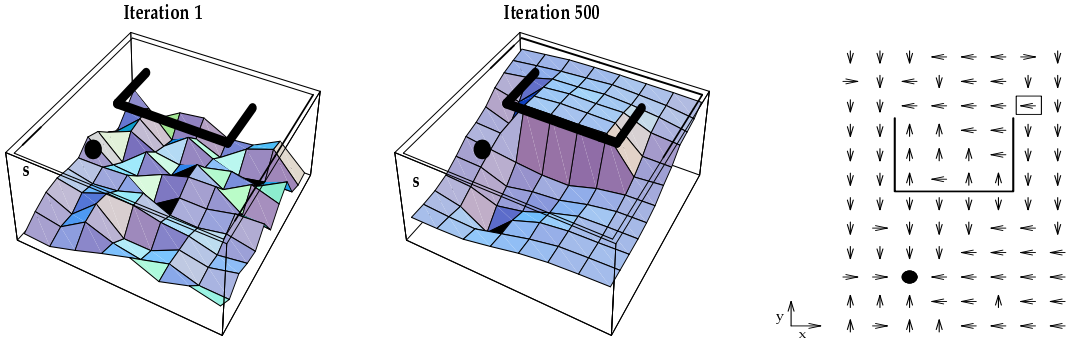
Instrumental conditioning in the bumble bee:



Activity in dopaminergic neurons:



Path-finding in a maze:



Rescorla-Wagner

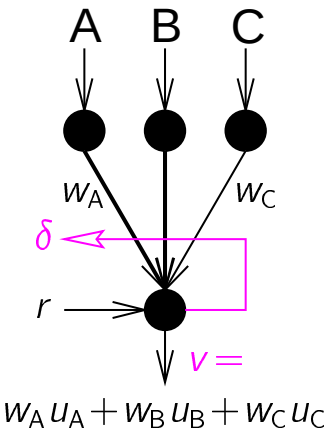
Paradigm	Set 1	Set 2		Result	
Pavlovian		$u \rightarrow r$		$u \rightarrow \hat{r}$	
Extinction	$u \rightarrow r$	$u \rightarrow \cdot$		$u \rightarrow \hat{\cdot}$	
Partial		$u \rightarrow r$	$u \rightarrow \cdot$	$u \rightarrow \alpha \hat{r}$	
Blocking	$u_1 \rightarrow r$	$u_1 + u_2 \rightarrow r$		$u_1 \rightarrow \hat{r}$	$u_2 \rightarrow \hat{\cdot}$
Inhibitory		$u_1 + u_2 \rightarrow \cdot$	$u_1 \rightarrow r$	$u_1 \rightarrow \hat{r}$	$u_2 \rightarrow -\hat{r}$
Overshadow		$u_1 + u_2 \rightarrow r$		$u_1 \rightarrow \alpha_1 \hat{r}$	$u_2 \rightarrow \alpha_2 \hat{r}$

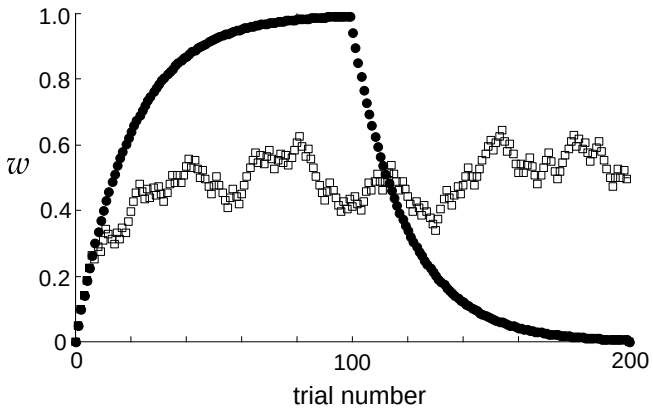
are nicely summarised by the delta rule – minimising $E(\mathbf{w}) = \langle (r - v)^2 \rangle$

$$v = \mathbf{w} \cdot \mathbf{u}$$

$$\delta = r - v$$

$$\mathbf{w} \rightarrow \mathbf{w} + \epsilon \delta \mathbf{u}$$





post-synaptic explaining away

Secondary Conditioning

Paradigm	Set 1	Set 2	Result
Secondary	$u_1 \rightarrow r$	$u_2 \rightarrow u_1$	$u_2 \rightarrow$ 'r'

fails under RW:

- no reward (in set 2)
- $\Rightarrow$ *inhibitory* learning for u_2

this actually happens – but only after positive learning

ignored *time*

traces allow synapses to change *after* they are active

$$\bar{\mathbf{u}} \rightarrow (1-\lambda)\mathbf{u} + \lambda\bar{\mathbf{u}}$$

$$\mathbf{w} \rightarrow \mathbf{w} + \epsilon\delta\bar{\mathbf{u}}$$

allows learning about future, but doesn't affect δ

temporal difference need to change $E(\mathbf{w})$

TD (Sutton)

Idea expressed as a goal for prediction is:

$$v(t) \sim \left\langle \sum_{s \geq t} r(s) \right\rangle \qquad v(t) \sim \left\langle \sum_{s \geq t} r(s) \gamma^{s-t} \right\rangle$$

within (t) a single trial, or

$$v(t) \sim \left\langle \lim_{T \to \infty} \frac{1}{T-t} \sum_{s=t}^{T} r(s) \right\rangle$$

this is a very radical departure.

Can't do stochastic gradient descent directly:

$$\delta = \sum_{s \geq t} r(s) - v(t)?$$

so learn by boostrapping:

$$v(t) = \langle r(t) \rangle + \left\langle \gamma \sum_{s \geq t+1} r(s) \gamma^{s-(t+1)} \right\rangle$$

$$= \langle r(t) \rangle + \gamma v(t+1)$$

and use

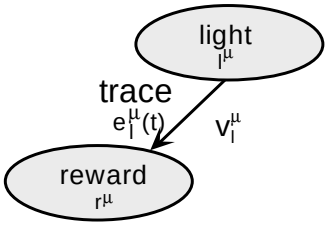
$$\delta(t) = (r(t) + \gamma v(t+1)) - v(t)$$

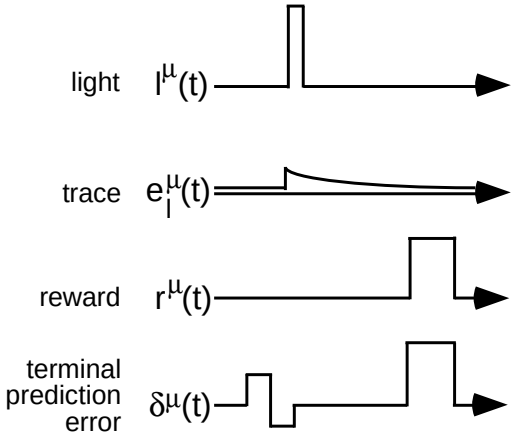
as a surrogate stochastic prediction error

TD Learning

$$v(t) = \mathbf{w} \cdot \mathbf{u}(t)$$
$$\delta(t) = r(t) + \gamma v(t+1) - v(t)$$
$$\mathbf{w} \rightarrow \mathbf{w} + \epsilon \delta(t) \mathbf{u}(t)$$

- mildly non-instantaneous – have to wait to get derivative
- under *trace* conditioning, still need to couple with a way of bridging time:
  - stimulus traces

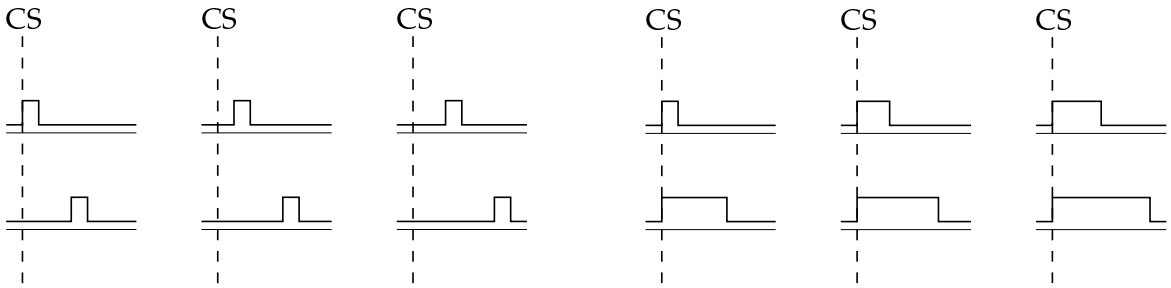


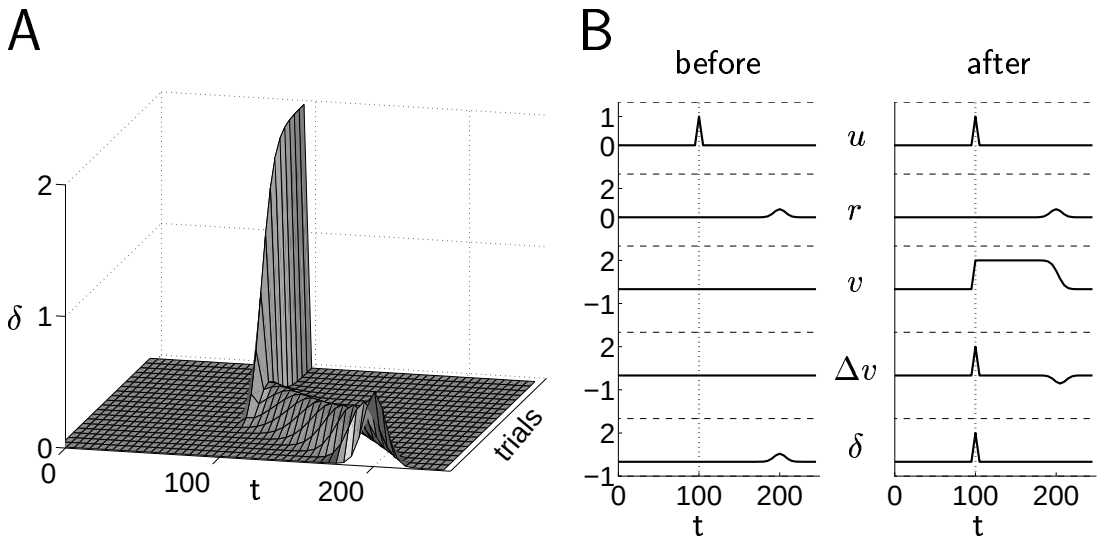


  - filter prediction model – expand $\mathbf{u}(t)$ to include *past* stimuli – allows precise predictions of the timing of rewards

TD Learning

Use $v(t) = \sum_{\tau=0}^{t} w(\tau)u(t-\tau)$ or

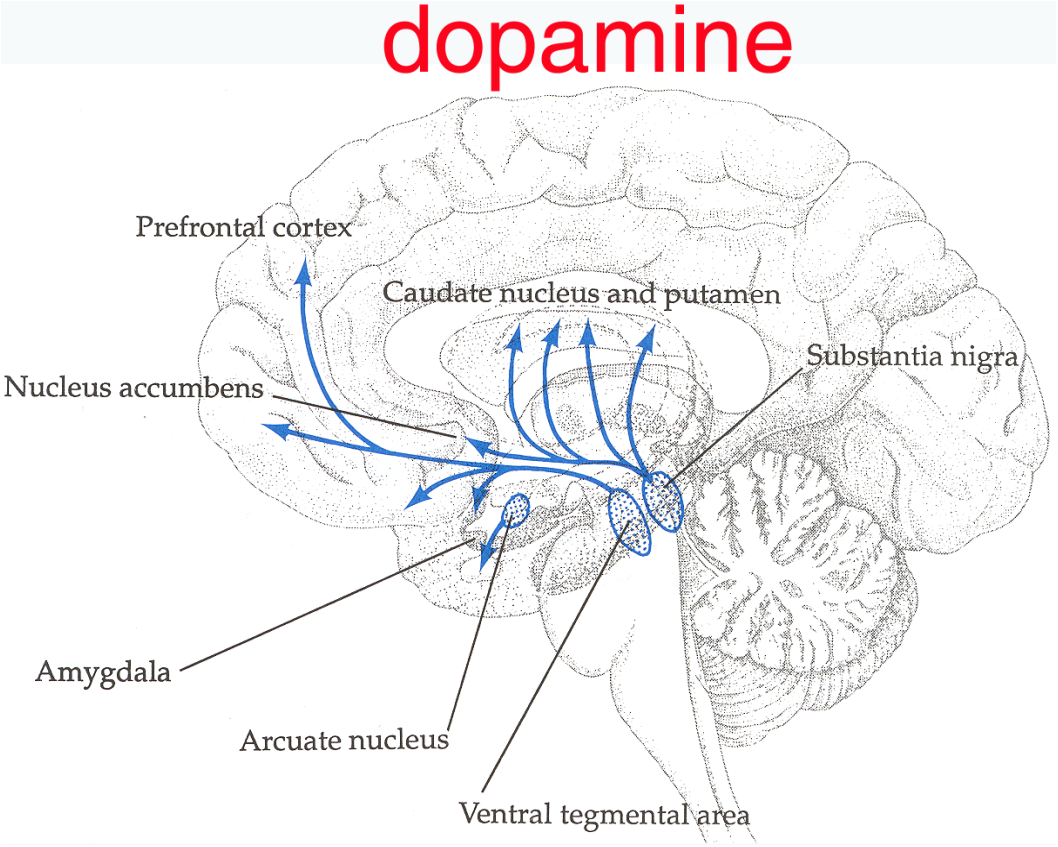




substrate for secondary conditioning

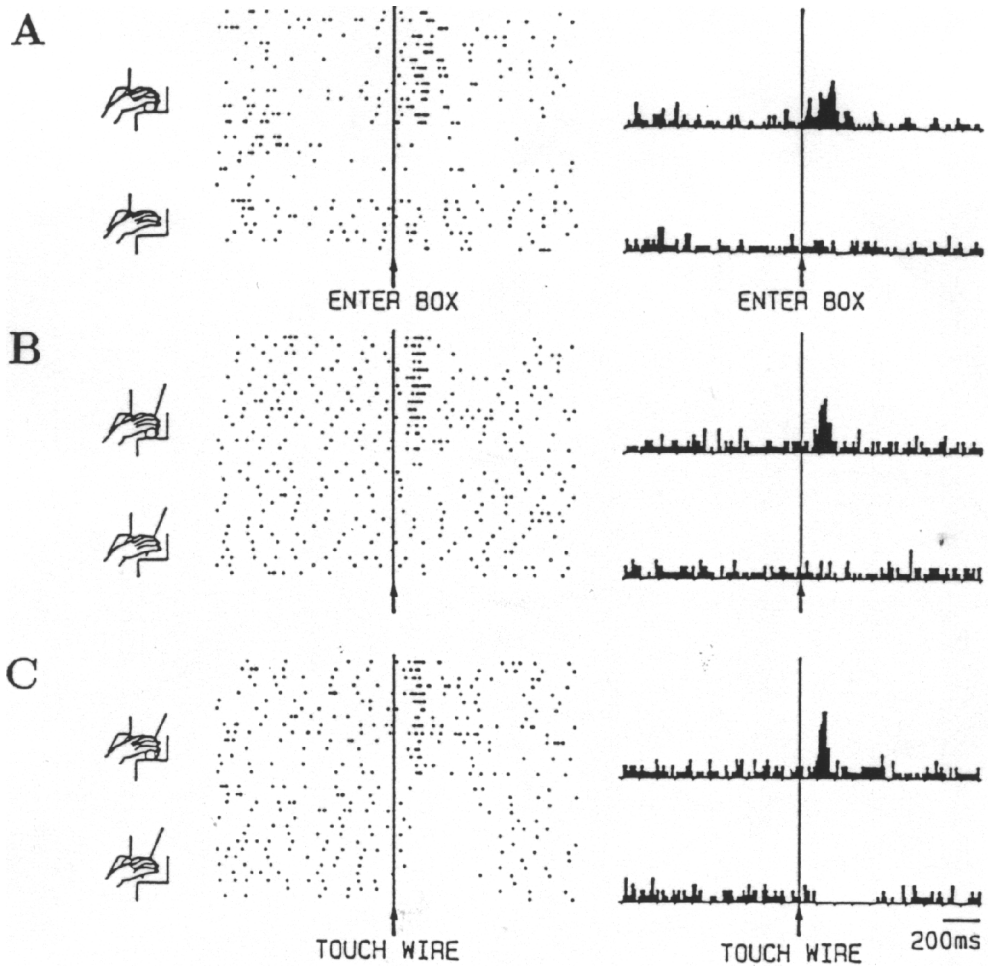
where should we look in the brain?

The Primate Dopamine System



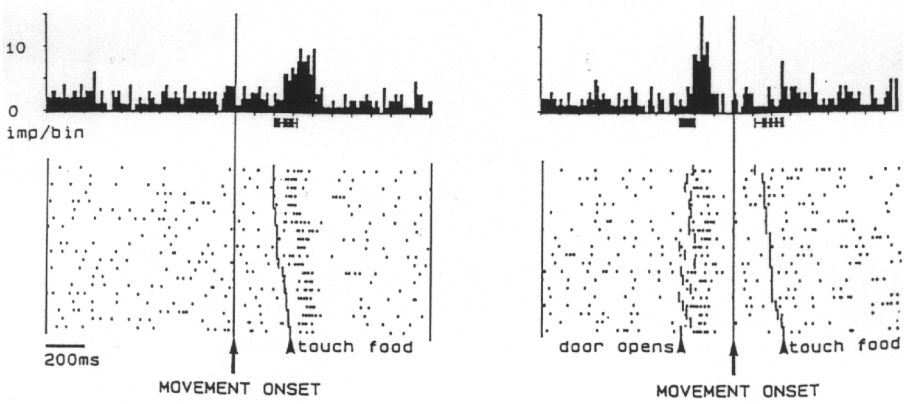
- self-stimulation, addiction, antagonists
- dialysis measurements
- association with limbic system
- association with the striatum

Reward Responses

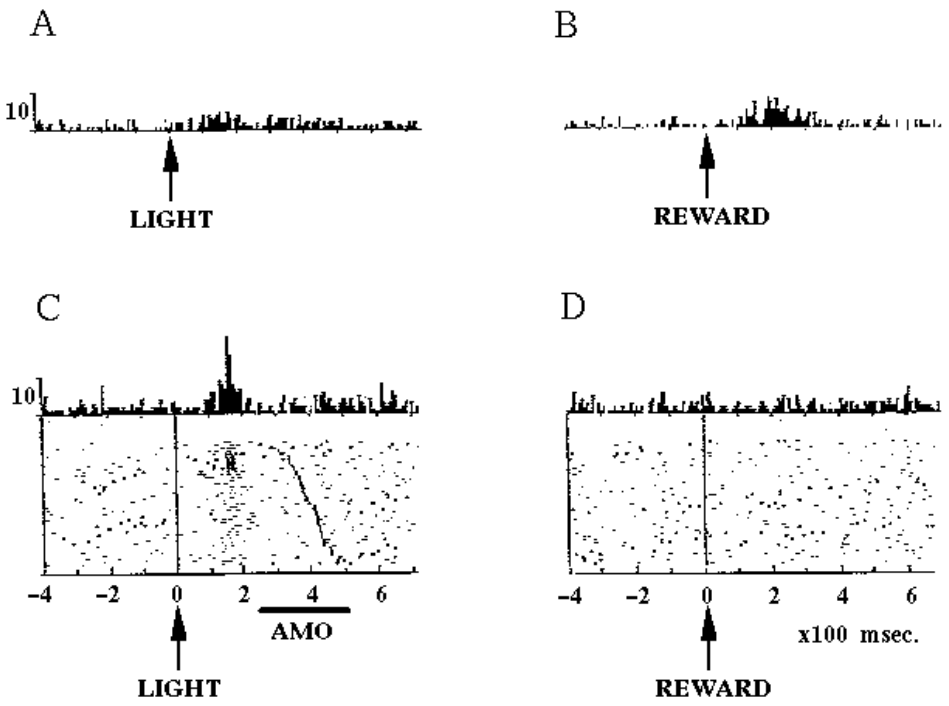


Romo, R & Schultz, W (1990). Dopamine neurons of the monkey midbrain: Contingencies of responses to active touch during self-initiated arm movements. J Neurophys, 63, 592-606, fig 8.

Conditioned Responses

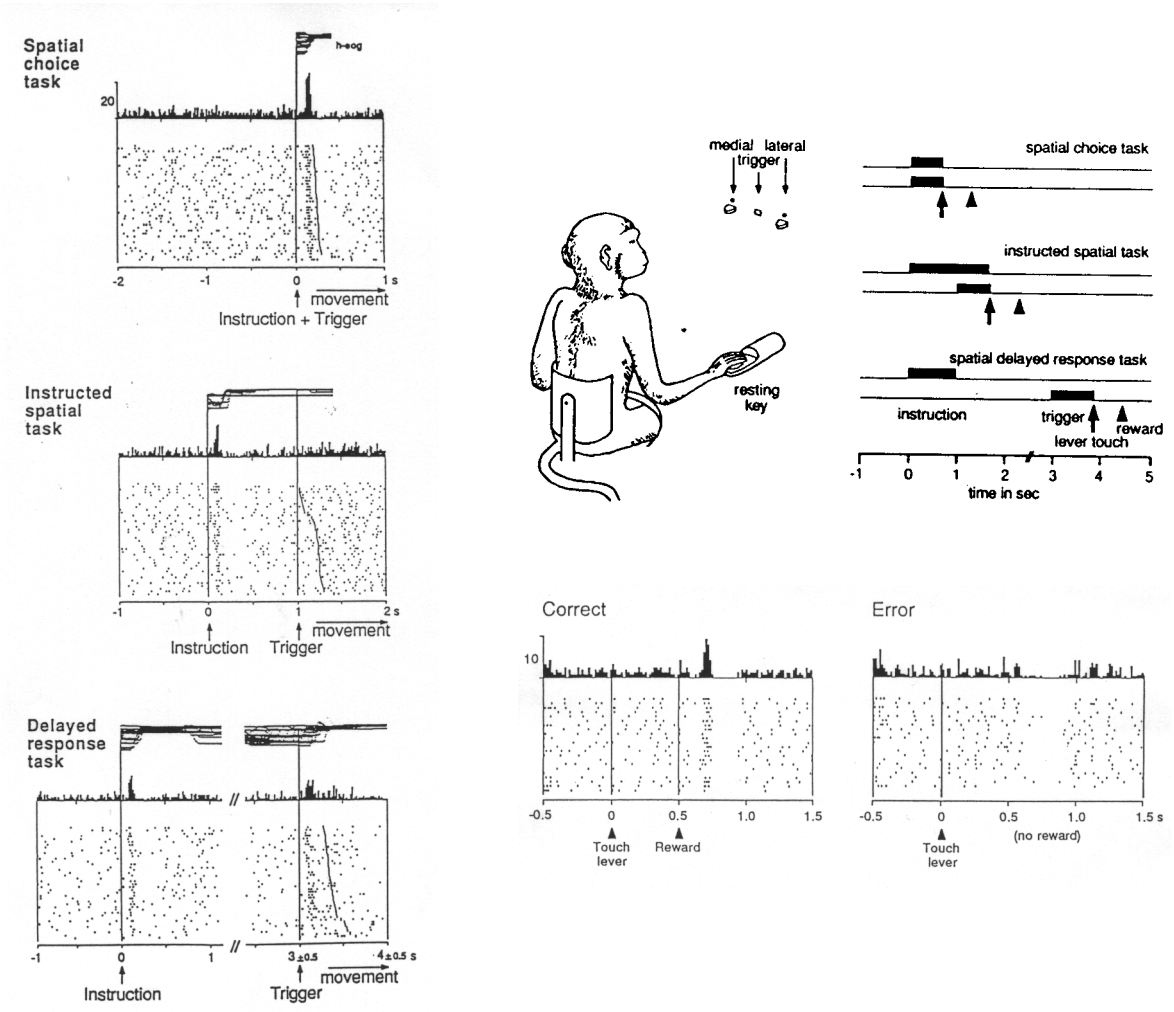


ibid, fig 11.



Schultz, W (1992). Activity of dopamine neurons in the behaving primate. SINS, 4, 129-138, fig 3.

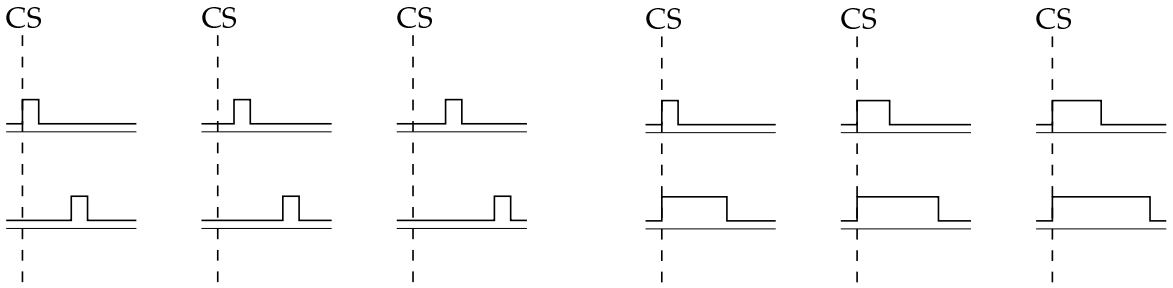
Conditioned Responses

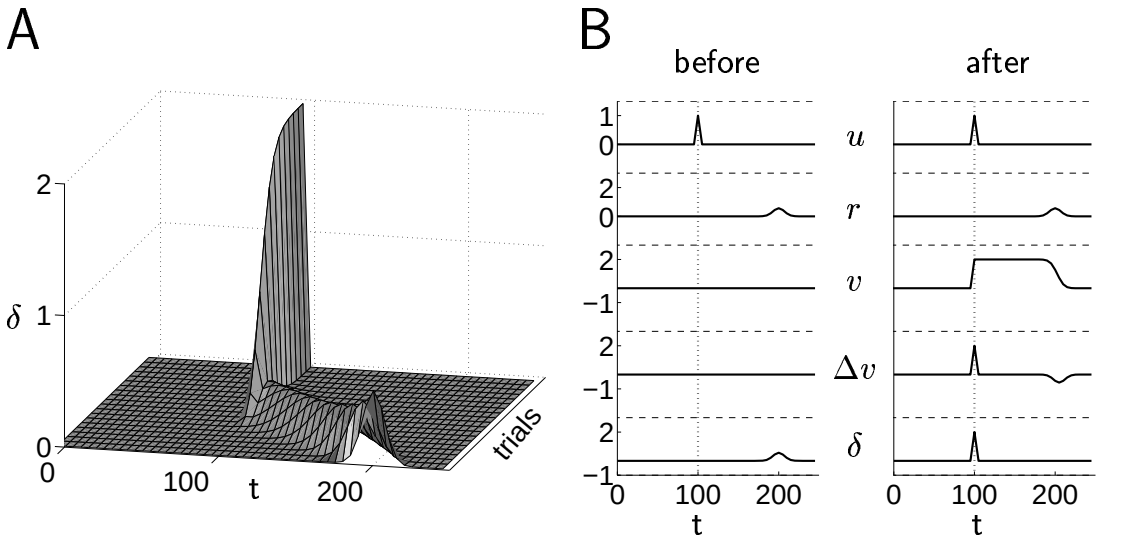


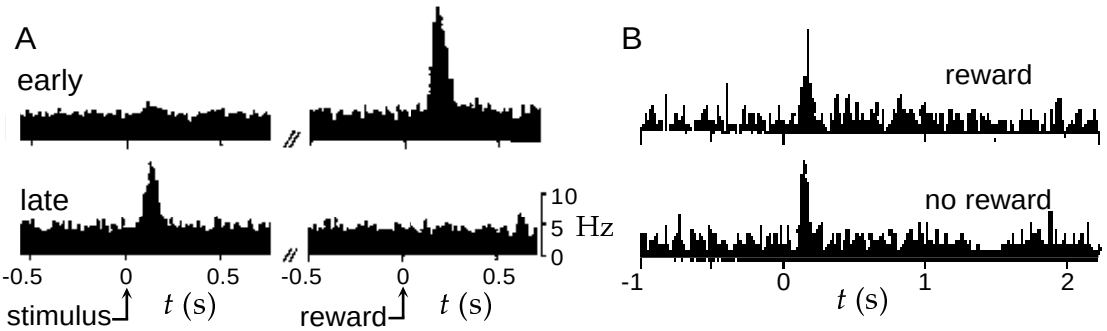
Schultz, W, Apicella, P & Ljungberg, T (1993). Responses of monkey dopamine neurons to reward and conditioned stimuli during successive steps of learning a delayed response task. J Neuroscience, 13, 900-913, figs 4,8.

TD Learning

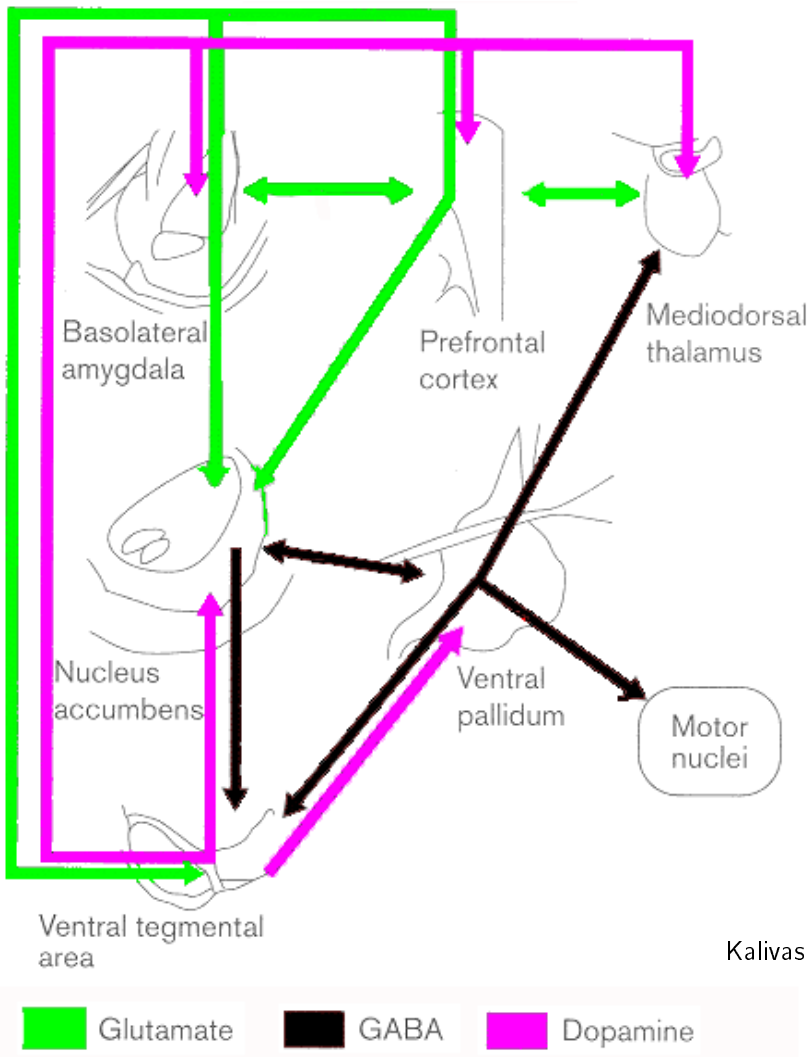
Use $v(t) = \sum_{\tau=0}^{t} w(\tau)u(t-\tau)$ or







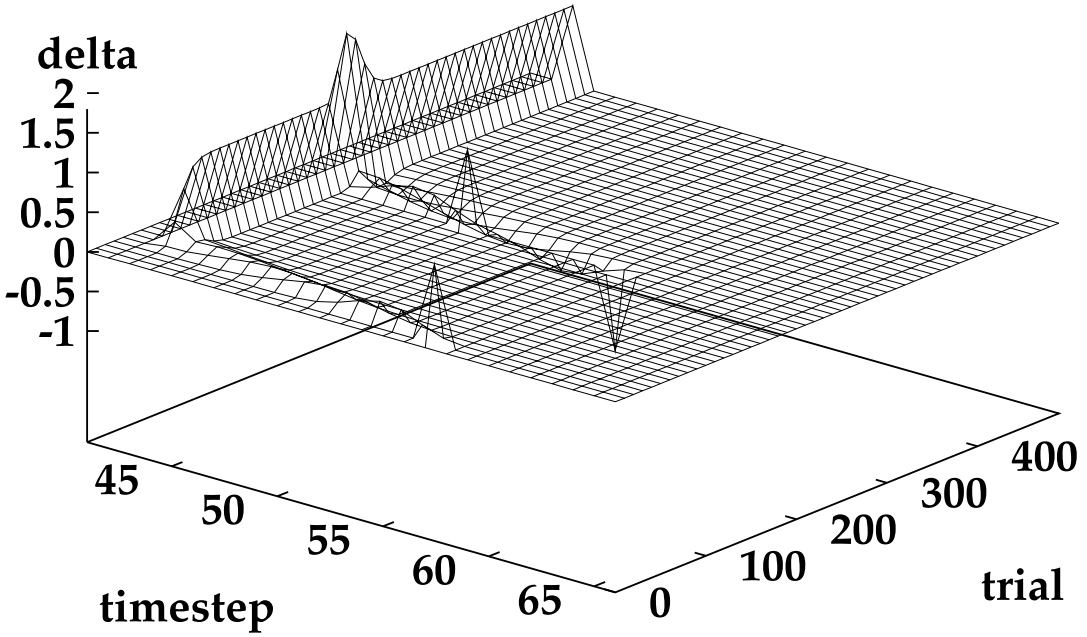
Prediction Pathway



Kalivas & Nakamura, 1999

- secondary values in BLA and orbitofrontal cortex
- stimulus competition in nucleus accumbens?
- ample reaffirmations (*eg* Waelti *et al* 2001)

Manipulation of the Timing



Static Action Choice

In *operant* conditioning, actions affects rewards. Eg bee:

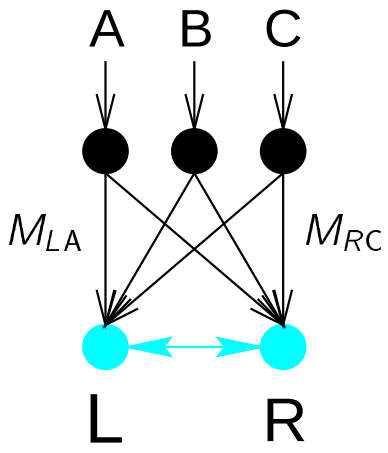
b nectar volume $r_\mathrm{b} \sim p[r_\mathrm{b}]$

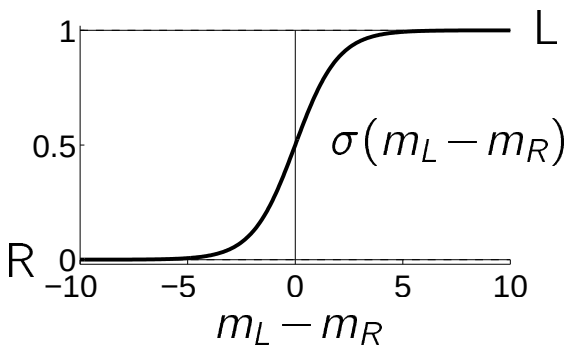
y nectar volume $r_\mathrm{y} \sim p[r_\mathrm{y}]$

Determine a *policy* $P[\mathrm{b}], P[\mathrm{y}]$ to optimise rewards.

Use action values **m**

$$P[\mathrm{b}] = \frac{\exp(\beta m_\mathrm{b})}{\exp(\beta m_\mathrm{b}) + \exp(\beta m_\mathrm{y})}$$



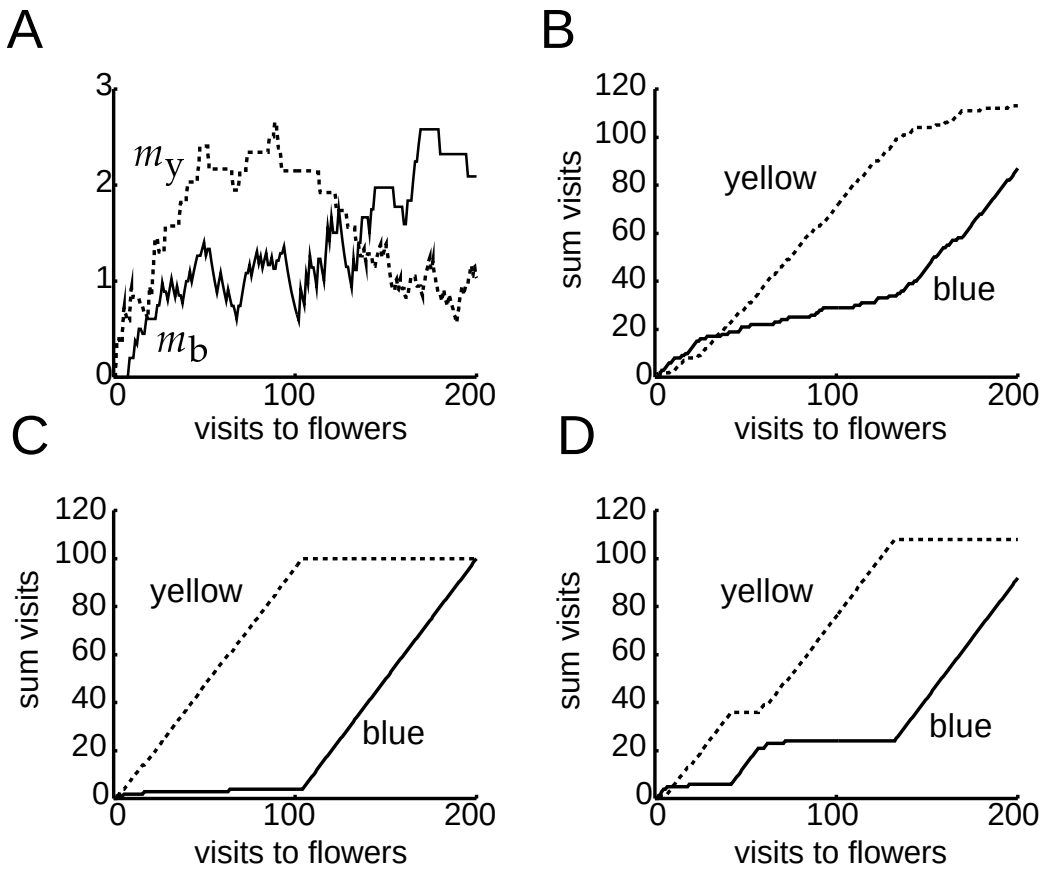


Indirect Actor

Use RW/stochastic delta rule

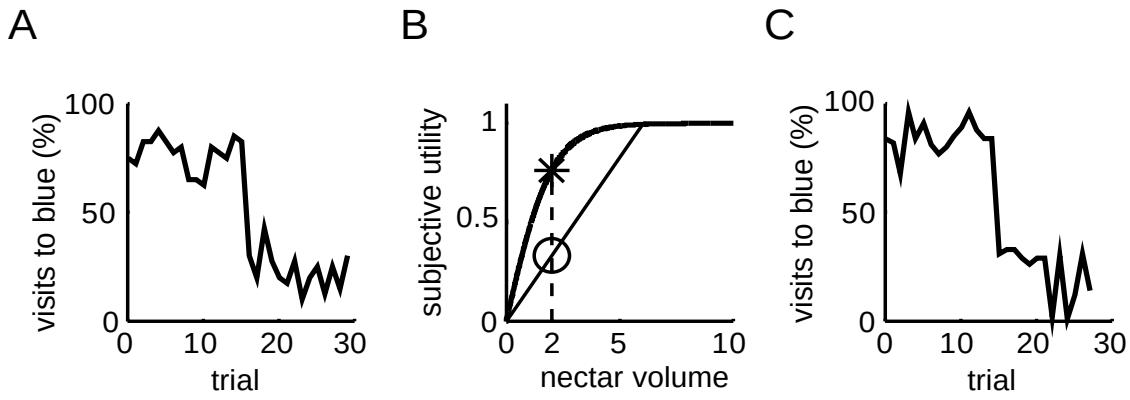
$$m_{\rm b} \to m_{\rm b} + \epsilon\delta \quad \text{with} \quad \delta = r_{\rm b} - m_{\rm b}\,,$$

given that the bee lands on a yellow flower:



$\langle r_{\rm b}\rangle = 1, \langle r_{\rm y}\rangle = 2$ for 100, then swap. B: $\beta = 1$; C,D: $\beta = 50$

Comparative Performance



- risk aversion ($\frac{1}{3}6\mu$l; $\frac{2}{3}0\mu$l) $\neq 2\mu$l
- explain by saturating utility
- or variance aversion
- rapid switching (ϵ, β)
- real bumble bees more variable (geographical effects)

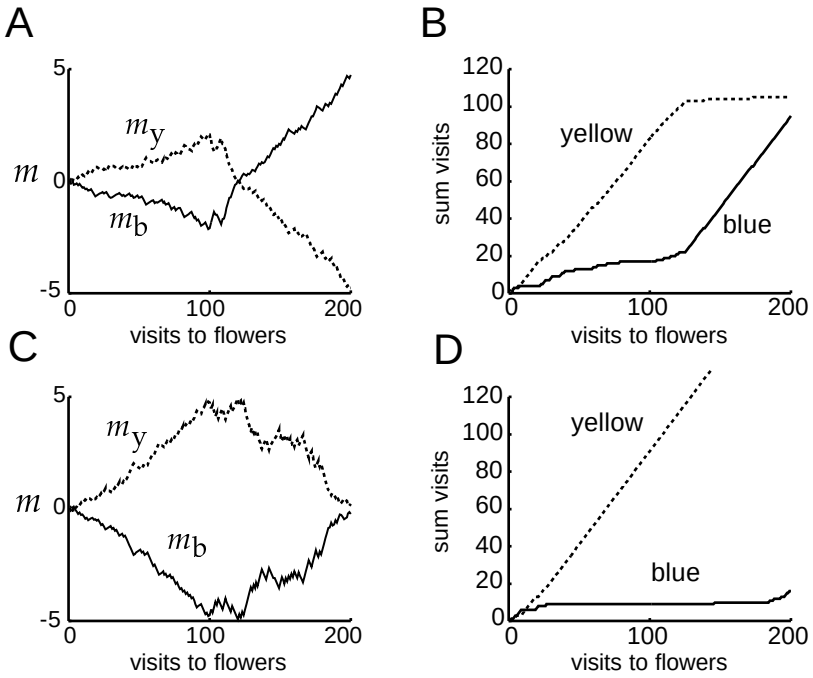
Direct Actor

$$E(\mathbf{m}) = P[\mathrm{b}]\langle r_\mathrm{b}\rangle + P[\mathrm{y}]\langle r_\mathrm{y}\rangle$$

and perform stochastic gradient ascent directly. For m_b:

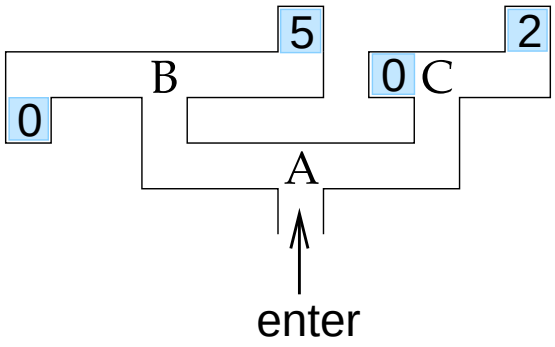
$$\begin{array}{rcll} m_\mathrm{b} & \rightarrow & m_\mathrm{b} + \epsilon(1 - P[\mathrm{b}])(r_\mathrm{b} - \overline{r}) & \text{if b is selected} \\ m_\mathrm{b} & \rightarrow & m_\mathrm{b} - \epsilon P[\mathrm{b}]\ (r_\mathrm{y} - \overline{r}) & \text{if y is selected} \end{array}$$

- $\overline{r}$ is *arbitrary – eg* $E(\mathbf{m})$
- stochastic learning automaton rule
- automatic saturation
- changes m_b even if given yellow landing



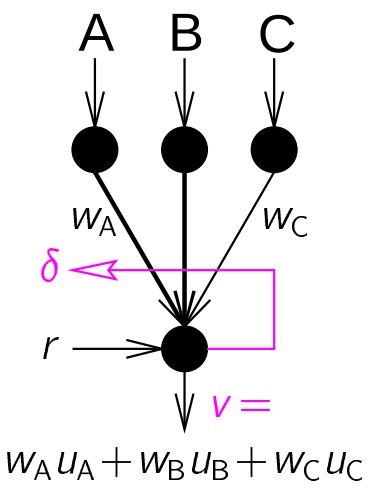
Deutschian Policy Improvement

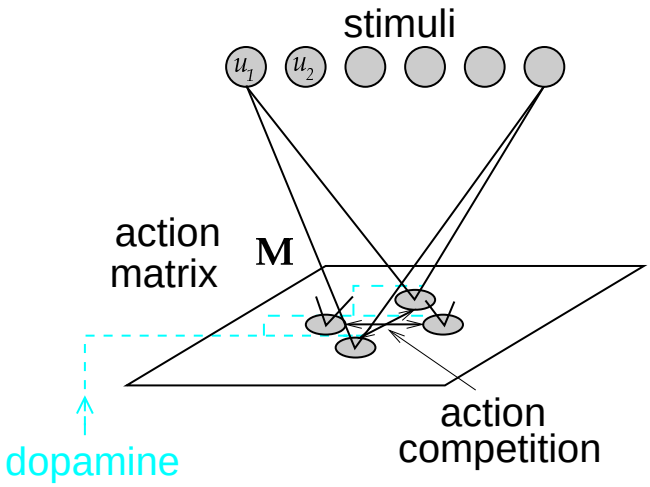
How to choose action at A in:



Choose a policy; use the same DA signal to

- *evaluate* A,B,C given the policy
- *improve* the policy





Formalism

Markov decision problems:

- stimuli signal *states* u
- states lead to other states in a Markovian manner and controlled by actions $\mathcal{P}_{uu'}(a)$
- states and actions are associated with rewards $r(u, a)$
- goal: maximise discounted long term rewards

$$\max_{a(t)} \sum_t \gamma^t r(u(t), a(t))$$

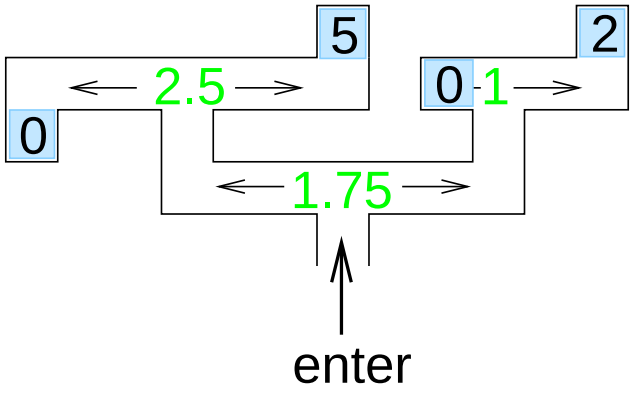
dynamic programming learns *values* for states based on *policies.*

policy iteration:

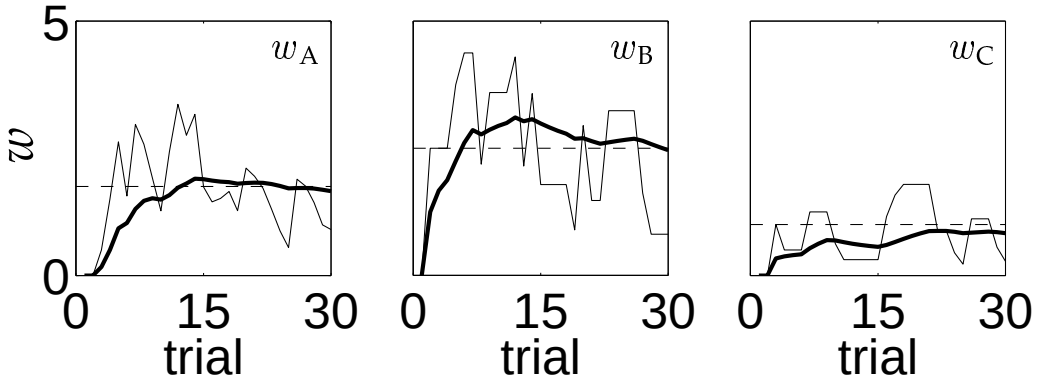
- classical: learn values as a function of stimuli
- instrumental: improve policies using values

Policy Evaluation

Given a random policy:



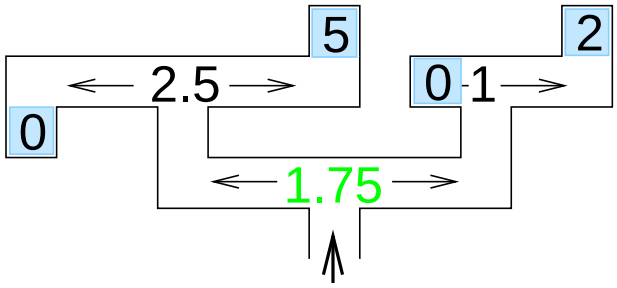
values are just $\langle \sum_{s \geq t} r(s) \rangle$:



learned until $\langle \delta(t) \rangle = 0$

Policy Improvement

once $\langle \delta(t) \rangle = 0$, fluctuations quantify action quality:

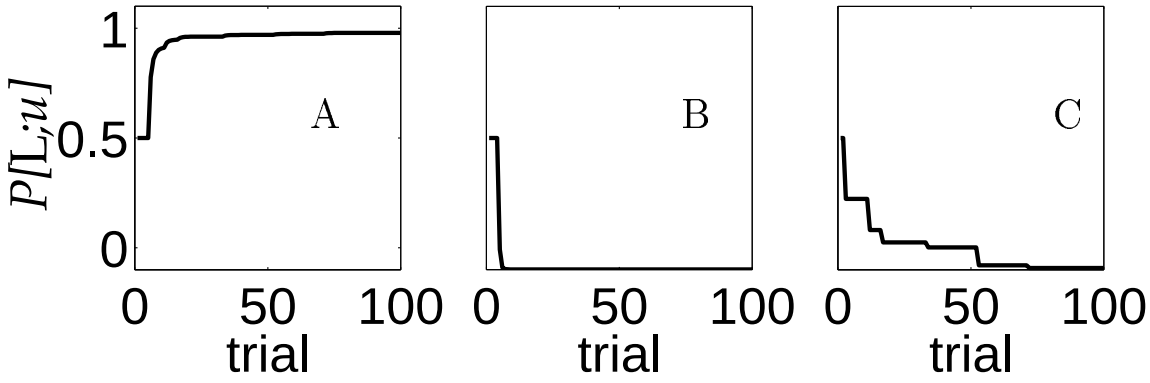


L $\quad \delta(t) = 0 + 2.5 - 1.75 = \quad 0.75$

R $\quad \delta(t) = 0 + 1 \quad - 1.75 = -0.75$

thus choose L more frequently than R.

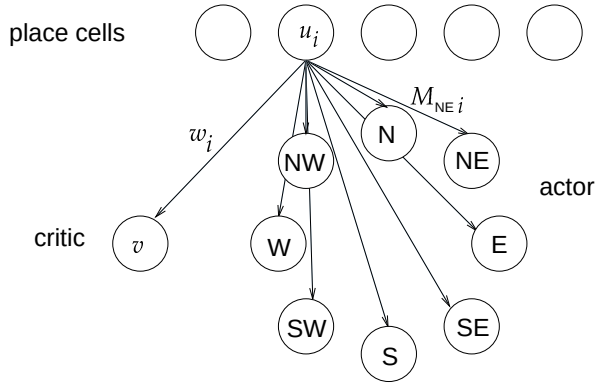
$\delta(t)$ **substitutes for** $r - \bar{r}$

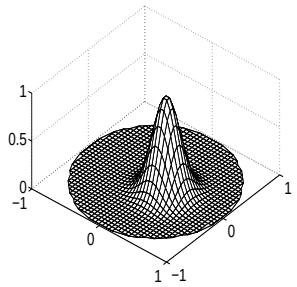


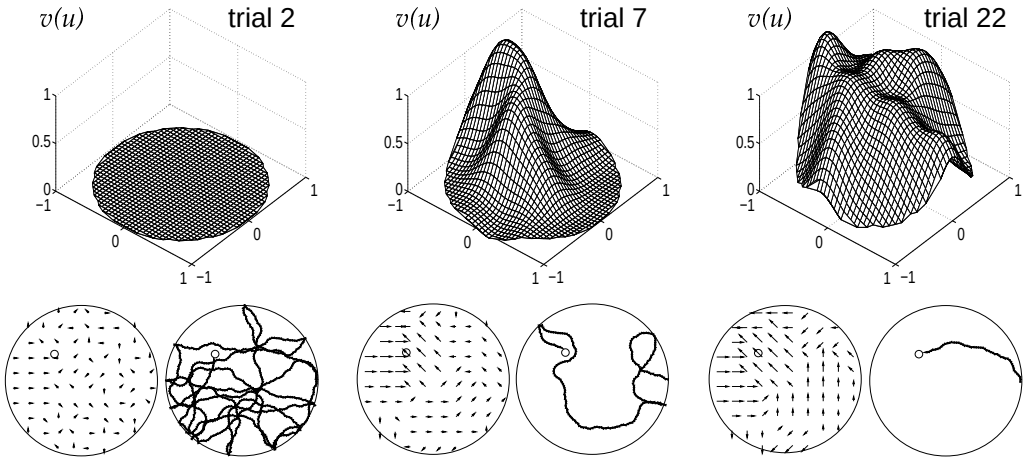
$\delta(t) > 0$ if

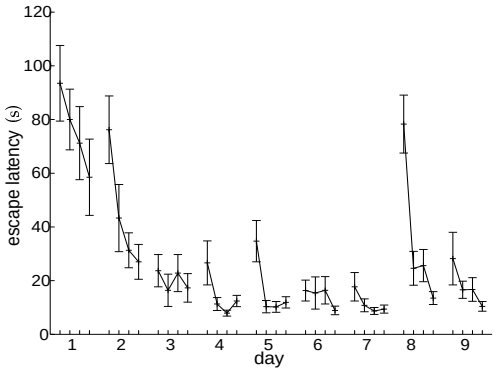
- value is too pessimistic $\Rightarrow \Delta v$
- action is better than average $\Rightarrow \Delta \pi$

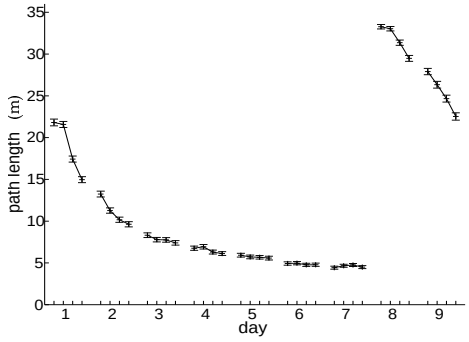
Water-maze navigation (FMD)











Value Iteration in DP

- start with value function $V_n(u)$
- generate a new value function using Bellman equation:

$$V_{n+1}(u) = \max_a \left[r(u,a) + \gamma \sum_{u'} \mathcal{P}_{uu'}(a) V_n(u') \right]$$

- since $V^*(u)$ satisfies the above equation, we can show that:

$$|V_{n+1}(u) - V^*(u)| \leq \gamma \max_u |V_n(u) - V^*(u)|$$

- so can use contraction mapping theory in the L_∞ norm to show convergence

note that the values are not associated with any policy prior to convergence

$\mathcal{Q}$-learning

Similar problems as for policy iteration – Watkins suggested learning $\mathcal{Q}(u, a)$ functions based on interaction with the world:

$$\Delta\mathcal{Q}(u(t), a(t)) \propto r(t) + \gamma V(u(t+1)) - \mathcal{Q}(u(t), a(t))$$

where $V(u(t+1)) = \max_b Q(u(t+1), b)$.

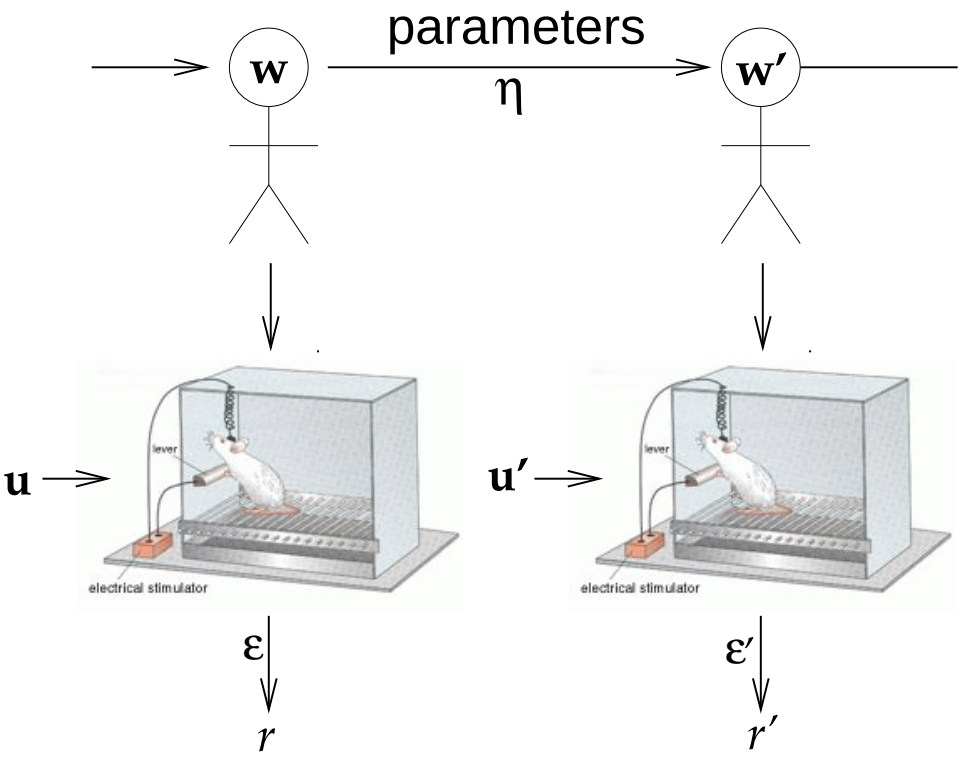
Watkins showed that this makes $\mathcal{Q}$ converge appropriately to give the optimal state-action value function.

Can represent $\mathcal{Q}$ – but still have to perform the maximisation explicitly.

Proofs of convergence of $\mathcal{Q}$-learning are easier, *cf:*

$$\begin{aligned} V^*(u) &= \max_a \{\mathcal{E}\left[r(u, a) + \gamma V^*(u')\right]\} \\ Q^*(u, a) &= r(u, a) + \gamma\mathcal{E}\left[\max_b \{Q^*(u', b)\}\right] \end{aligned}$$

Attention



uncertainty P&H; CEN + SNc

$$\mathbf{w} \sim N[\hat{\mathbf{w}}, \Sigma]$$

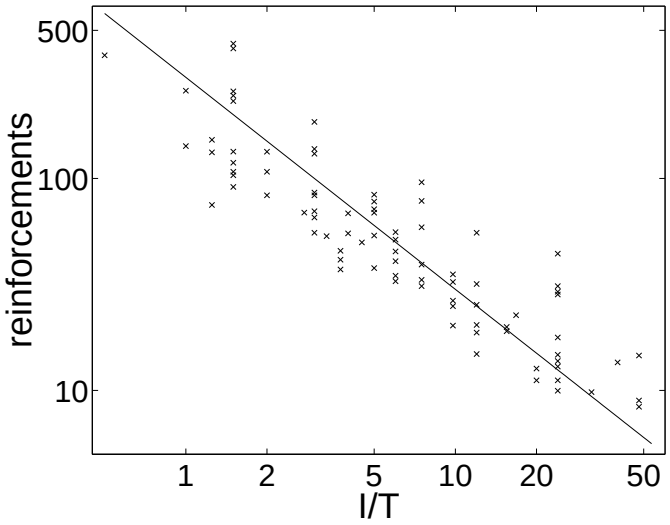
where Σ determines associabilities, backwards blocking, *etc*

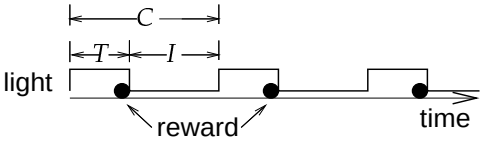
unreliability P&H; comparator models; NAcc

$$prediction = \sum_a \alpha_a w_a u_a$$

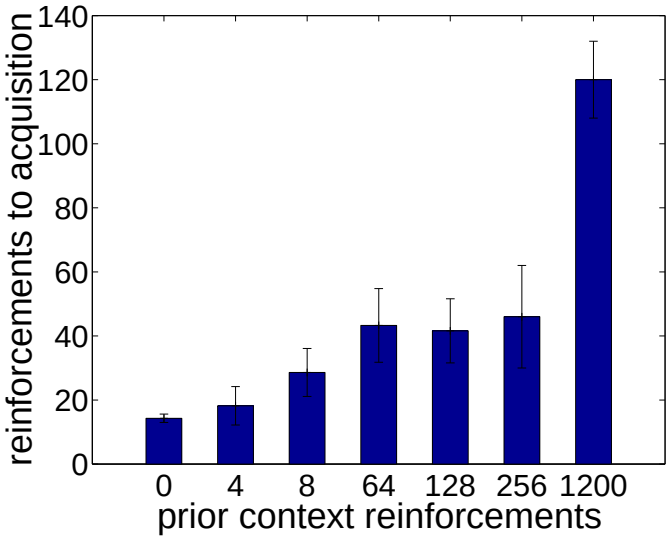
α_a controls bond between u_a and r

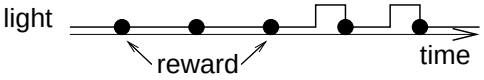
Attention





Gallistel & Gibbon, 2000





Balsam & Schwartz, 1981

- unreliability competition between light, context
- uncertainty dictates fast learning (SET)

But

What happened to S-R psychology?

actions competition between simultaneous choices

habits single sensible action, but is it worth executing?

Descartes

The passions:

- dispose the soul to desire those things which nature tells us are of use,
- and to persist in this desire,
- and also bring about that same **agitation of the spirits** which customarily causes them to dispose the body to movement which serves for the carrying into effect of these things.

Descartes, R (1649) Passions of the soul. In Haldane & Ross *The Philosophical Works of Descartes.* CUP.

Evidence

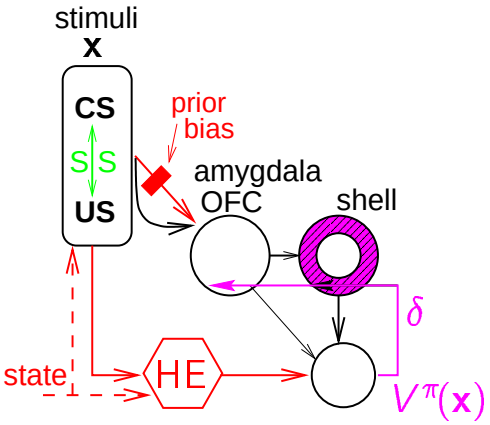
1. appetitive Pavlovian instrumental transfer (Estes) under shell-based dopaminergic control
2. instant, intrinsic drive preference under specific deprivation states
3. incentive learning (Dickinson, Balleine) – complete distinction between instrumental and classical motivation?
4. transfer from 'action' to 'habit'

Suggests a new model – using advantages:

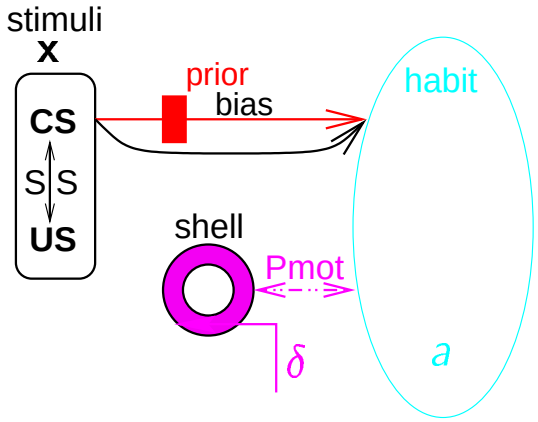
$$A^{\pi}(\mathbf{u}, a) = Q^{\pi}(\mathbf{u}, a) - V(\mathbf{u})$$

'how much better is a than average at $\mathbf{u}$?'

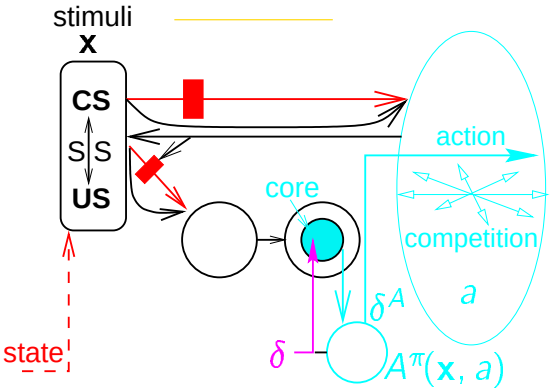
New Model



- hard-wired evaluator
- competes with learned route
- stimulus substitution



- Pavlovian motivation via δ
- URs and CRs



- Policy advantages
- $\delta^A = \delta - A^\pi(\mathbf{u}, a)$
- $A^\pi(\mathbf{u}, a) \to 0$ 'birth of a habit'

Conclusions

Computational analysis of role of diffuse scalar systems in organisation and function:

Ethology	**Computation**
optimality	dynamic programming
appropriateness	Kalman filtering
Psychology	**Algorithm**
classical/operant	TD/delta rules
conditioning	simple weights
Neurobiology	**Implementation**
'neuromodulatory' systems	

Open Issues

- opponent process – 5HT *versus* dopamine
- location of the weights w_i (amygdala, prefrontal cortex, nucleus accumbens)
- persistent up- and down- regulation during drug and disease states
- effects of surprise – latent inhibition, overshadowing, *etc* and dependence on the hippocampus
- link to experimental data on 3-term learning rules in the striatum and the effect of DA on storage in pFC
- quantitative accounts for rates of acquisition
- model-based, *indirect*, methods involving latent learning *versus* model-free. *direct* methods – theory and practice

Open Theoretical Issues

- hierarchical methods for solving *many* tasks in the *same* environment:
  - *options:* make choices at a coarser temporal grain
  - *fragments:* split the world up at a coarser spatial grain, *eg* rooms in a maze, *etc*
  - *feudality:* 'subroutines'
- representational learning – integration of reinforcement learning with unsupervised learning
- *partial observability* and approximations to exact inference for POMDPs
- optimal trade-offs between *exploration* versus *exploitation*

References

Barto, AG, Sutton, RS & Anderson, CW (1983) Neuronlike elements that can solve difficult learning problems. *IEEE Transactions on Systems, Man, and Cybernetics* **13**:834-846.

Bertsekas, DP & Tsitsiklis, JN (1996) *Neuro-Dynamic Programming*. Belmont, MA:Athena Scientific.

Chapter 9 of Dayan, P & Abbott, LF (2001). *Theoretical Neuroscience.* Cambridge, MA: MIT Press. `people.brandeis.edu/~abbott/book/`

Dickinson, A (1980) *Contemporary Animal Learning Theory*. Cambridge: Cambridge University Press.

Gallistel, CR (1990) *The Organization of Learning*. Cambridge, MA:MIT Press.

Grossberg, S, editor (1988) *Neural Networks and Natural Intelligence*. Cambridge, MA:MIT Press.

Montague, PR, Dayan, P & Sejnowski, TJ (1996) A framework for mesencephalic dopamine systems based on predictive Hebbian learning. *Journal of Neuroscience* **16**:1936-1947.

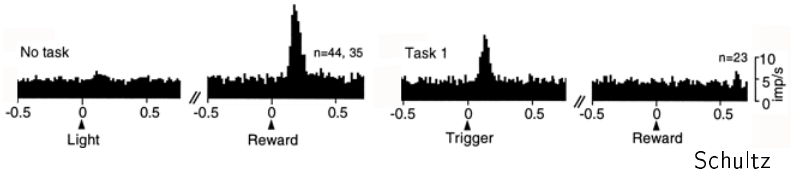
Schultz, W, Dayan, P & Montague, PR (1997) A neural substrate of prediction and reward. *Science* **275**:1593-1599.

Sutton, RS & Barto, AG (1998) *Reinforcement Learning.* Cambridge, MA: MIT Press.

Watkins, CJCH (1989) *Learning from Delayed Rewards*. PhD Thesis, University of Cambridge, Cambridge, UK.

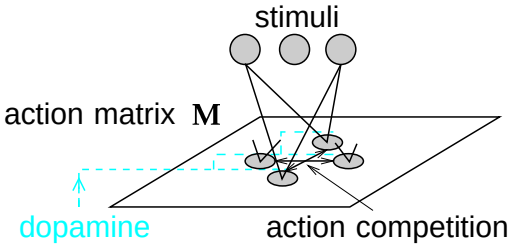
Summary

1: Prediction



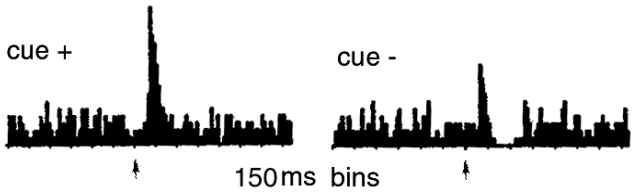
Schultz

2: Action



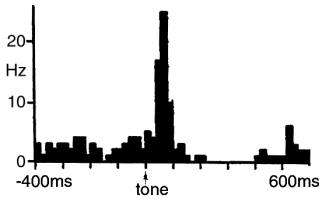
under revision

3: Generalization



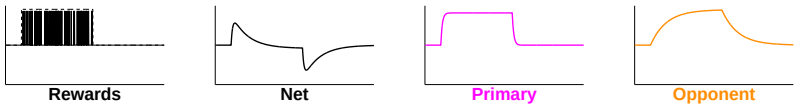
Schultz

4: Innovation



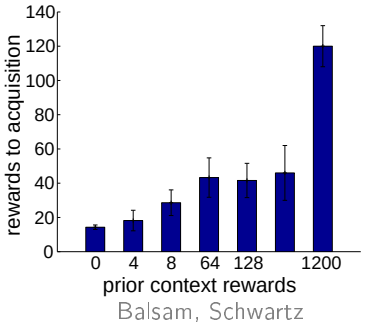
Horvitz; Jacobs

5: Aversion



Solomon & Corbit

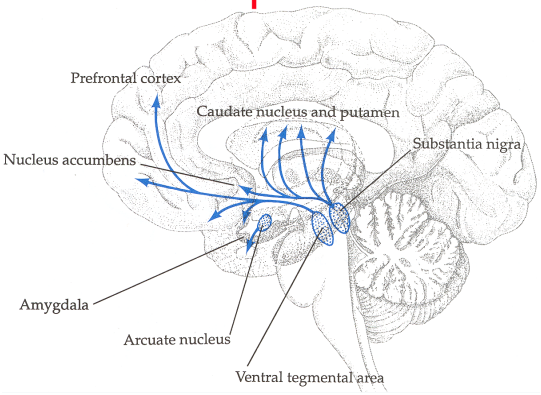
6: Attention



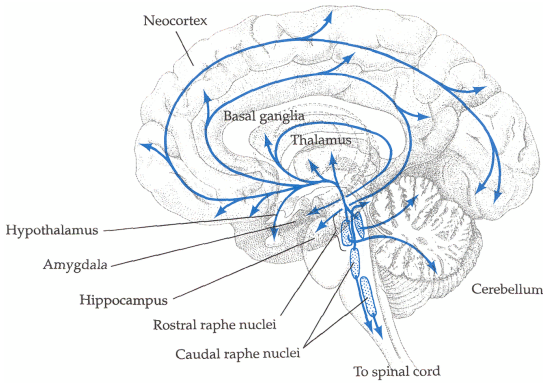
Balsam, Schwartz

Neural Substrate

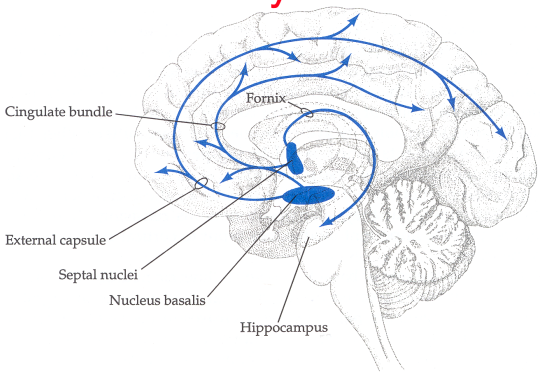
dopamine



5HT



acetylcholine



norepinephrine

