# Global plan

- Reinforcement learning I:
  - prediction
  - classical conditioning
  - dopamine
- Reinforcement learning II:
  - dynamic programming; action selection
  - Pavlovian misbehaviour
  - vigor

- Chapter 9 of Theoretical Neuroscience

(thanks to Yael Niv)

# Conditioning

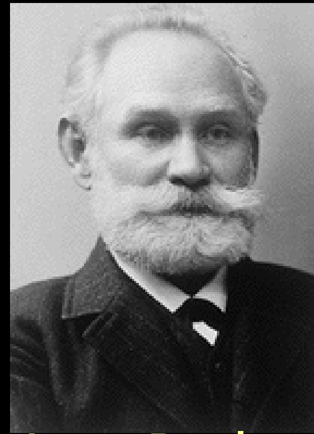
prediction: of important events
control: in the light of those predictions

- 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

# Animals learn predictions

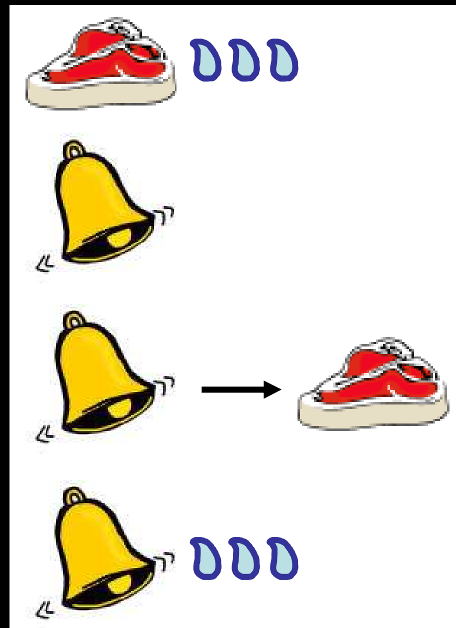


Ivan Pavlov

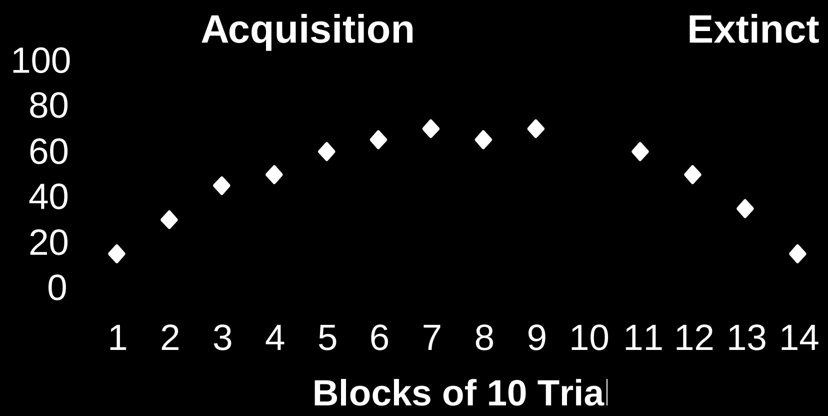
 = Unconditioned Stimulus

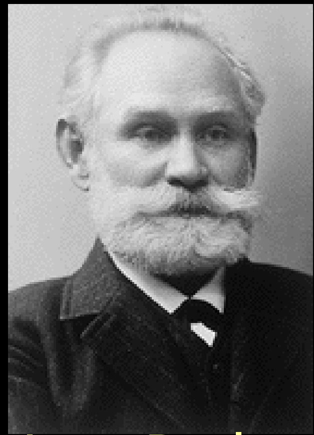
 = Conditioned Stimulus

 = Unconditioned Response (reflex);
Conditioned Response (reflex)

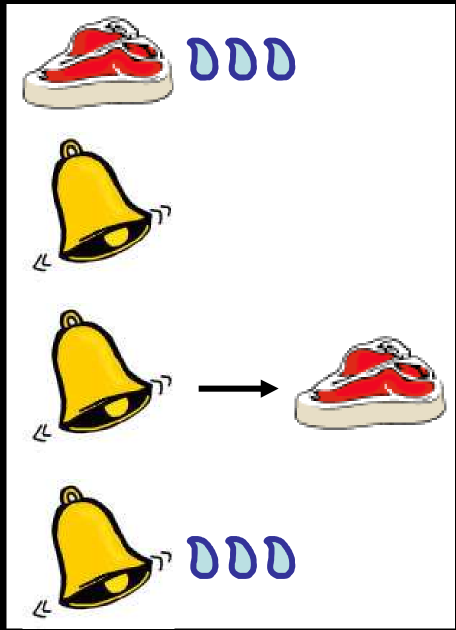


# Animals learn predictions



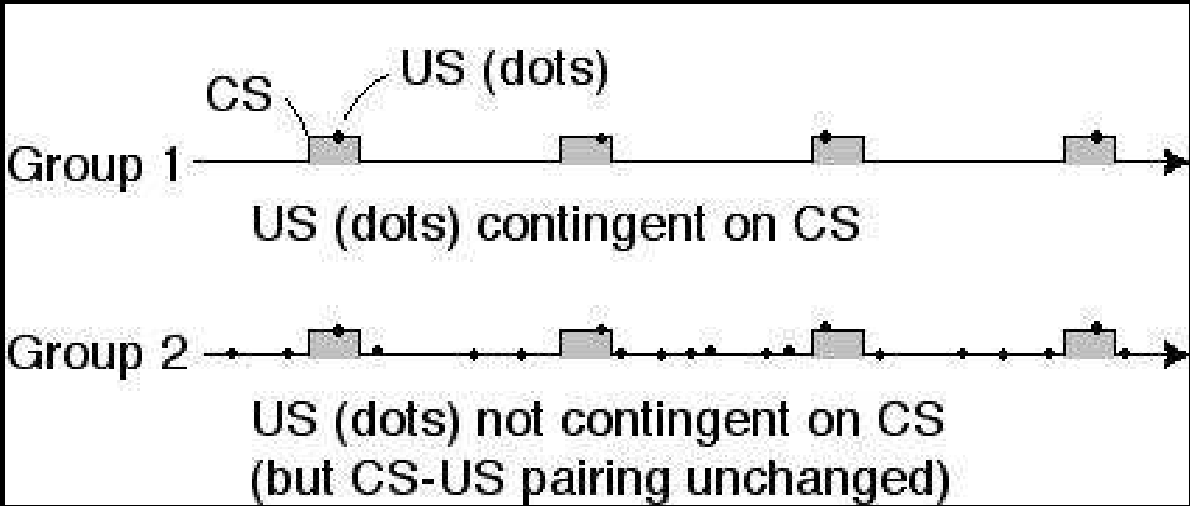
Ivan Pavlov



very general across species, stimuli, behaviors

# But do they really?

1. Rescorla's control

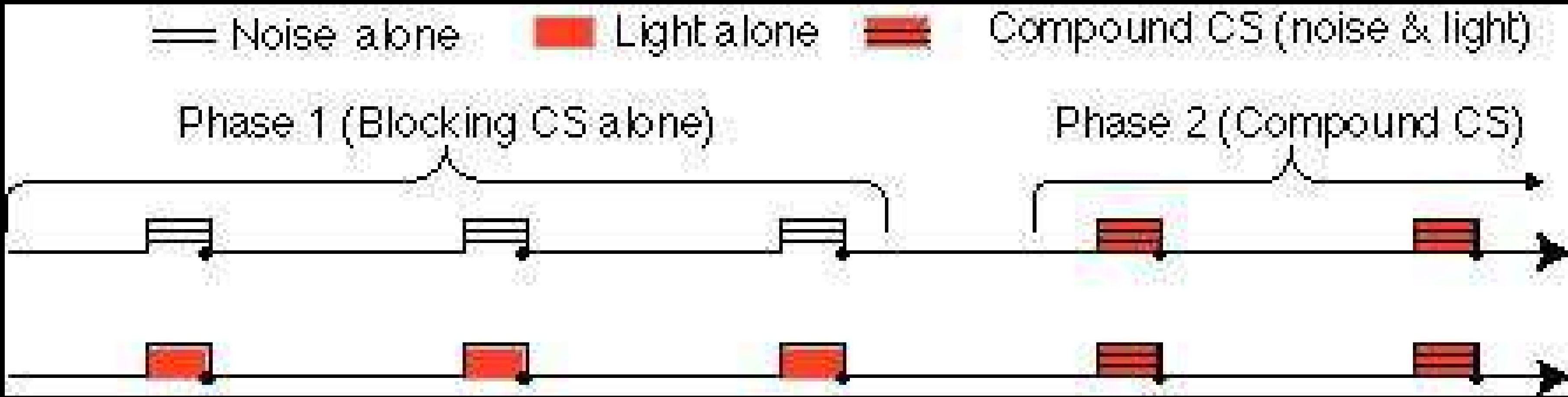



temporal contiguity is not enough - need contingency

P(food | light) > P(food | no light)

# But do they really?

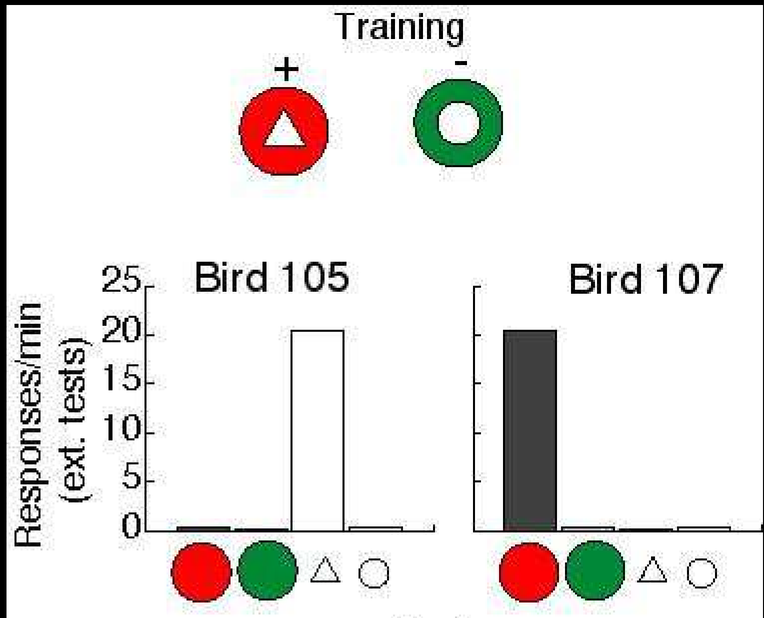
2. Kamin's blocking




contingency is not enough either… need surprise

# But do they really?

3. Reynold's overshadowing




seems like stimuli compete for learning

# Theories of prediction learning: Goals

- Explain how the CS acquires “value”
- When (under what conditions) does this happen?
- Basic phenomena: gradual learning and extinction curves
- More elaborate behavioral phenomena
- (Neural data)

P.S. Why are we looking at old-fashioned Pavlovian conditioning?

$\rightarrow$ it is the perfect uncontaminated test case for examining prediction learning on its own

# Rescorla & Wagner (1972)

error-driven learning: change in value is proportional to the difference between actual and predicted outcome

$$\Delta V_{CS_i} = \eta \left( r_{US} - \sum_j V_{CS_j} \right)$$

Assumptions:

1. learning is driven by error (formalizes notion of surprise)
2. summations of predictors is linear

A simple model - but very powerful!

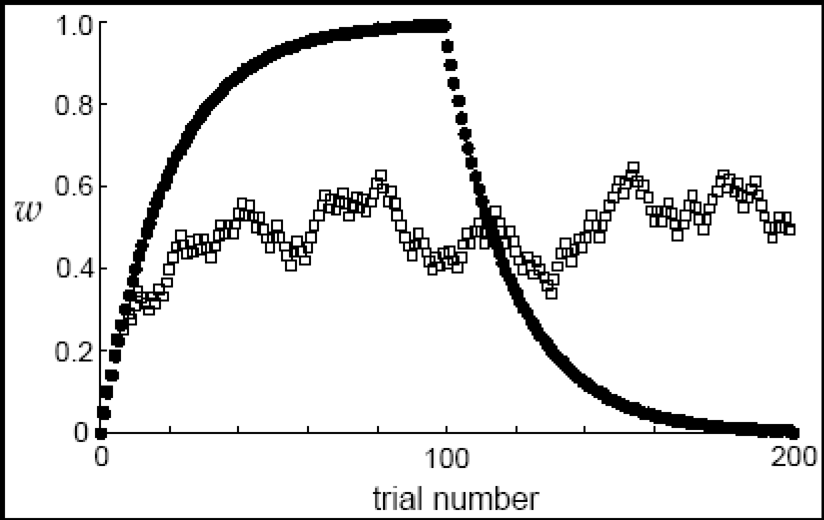
- explains: gradual acquisition & extinction, blocking, overshadowing, conditioned inhibition, and more..
- predicted overexpectation

note: US as “special stimulus”

# Rescorla-Wagner learning

$$V_{t+1} = V_t + \eta(r_t - V_t)$$

- how does this explain acquisition and extinction?
- what would V look like with 50% reinforcement? eg. 1 1 0 1 0 0 1 1 1 0 0
  - what would V be on average after learning?
  - what would the error term be on average after learning?



# Rescorla-Wagner learning

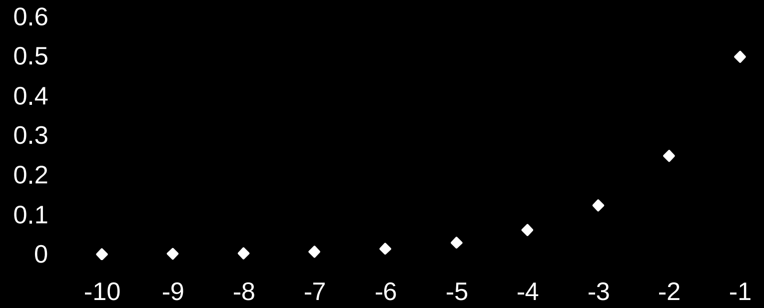
$$V_{t+1} = V_t + \eta(r_t - V_t)$$

how is the prediction on trial (t) influenced by rewards at times (t-1), (t-2), …?

$$V_{t+1} = (1-\eta)V_t + \eta r_t$$

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

the R-W rule estimates expected reward using a weighted average of past rewards




recent rewards weigh more heavily

why is this sensible?

learning rate = forgetting rate!

# Summary so far

Predictions are useful for behavior

Animals (and people) learn predictions (Pavlovian conditioning = prediction learning)

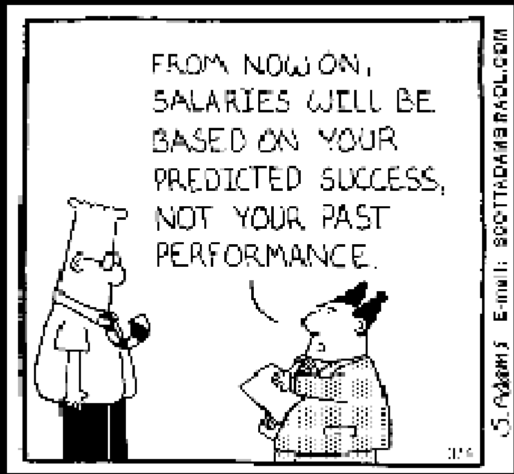
Prediction learning can be explained by an error-correcting learning rule (Rescorla-Wagner): predictions are learned from experiencing the world and comparing predictions to reality

Marr:

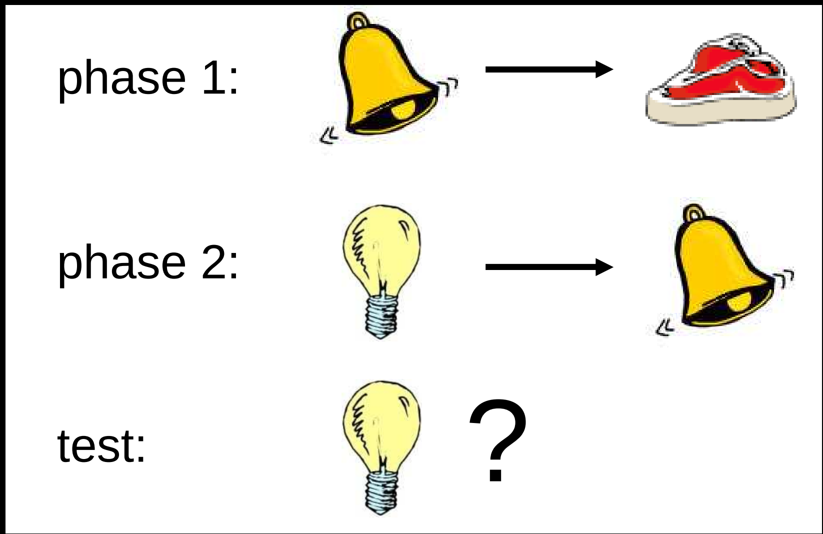
$$V = \sum_j V_{CS_j}$$

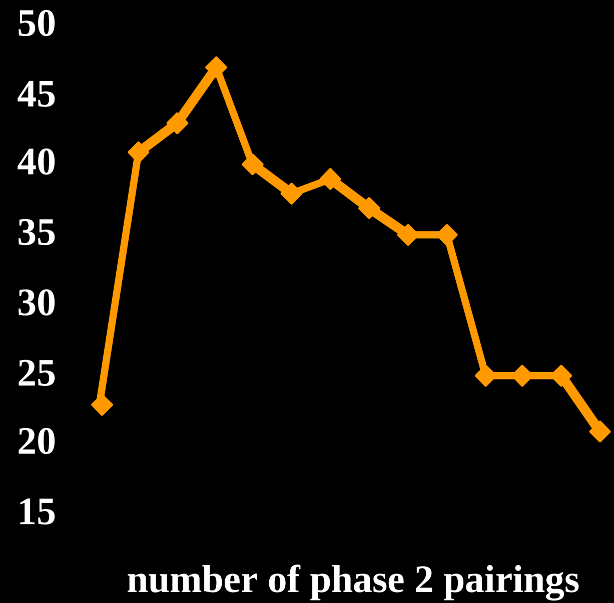
$$E = \left\langle \left( r_{US} - V \right)^2 \right\rangle$$

$$\Delta V_{CS_i} \alpha \frac{\partial E}{\partial V_{CS_i}} = \left( r_{US} - V \right) = \delta$$



# But: second order conditioning




what do you think will happen?

what would Rescorla-Wagner learning predict here?

animals learn that a predictor of a predictor is also a predictor of reward!

$\Rightarrow$ not interested solely in predicting immediate reward

# lets start over: this time from the top

Marr's 3 levels:

- The problem: optimal prediction of future reward

$$V_t = E\left[\sum_{i=t}^{T} r_i\right]$$

want to predict expected sum of future reward in a trial/episode

(N.B. here t indexes time within a trial)

- what's the obvious prediction error?

$$\delta^{\mathrm{RW}} = r - V_{CS}$$

$$\delta_t = \sum_{i=t}^{T} r_i - V_t$$

- what's the obvious problem with this?

# lets start over: this time from the top

Marr's 3 levels:

- The problem: optimal prediction of future reward

$$V_t = E\left[\sum_{i=t}^{T} r_i\right]$$

want to predict expected sum of future reward in a trial/episode

$$V_t = E\left[r_t + r_{t+1} + r_{t+2} + \ldots + r_T\right]$$

Bellman eqn for policy evaluation

# lets start over: this time from the top

Marr's 3 levels:

- The problem: optimal prediction of future reward
- The algorithm: temporal difference learning

$$V_t = E[r_t] + V_{t+1}$$

$$V_t \leftarrow (1-\eta)V_t + \eta(r_t + V_{t+1})$$

temporal difference prediction error $\delta_t$

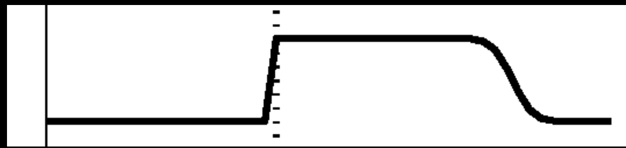
compare to: $V_{T+1} \leftarrow V_T + \eta(r_T - V_T)$

# prediction error

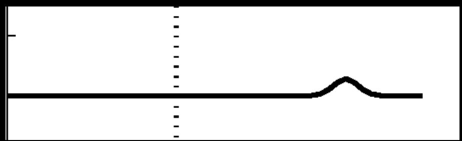
TD error $$\delta_t = r_t + V_{t+1} - V_t$$

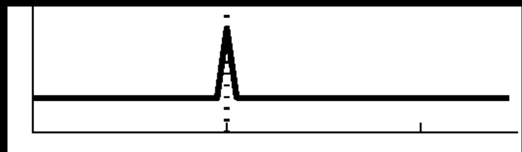
L

$V_t$



$\delta_t$





R

no prediction

prediction, reward

prediction, no reward

# Summary so far

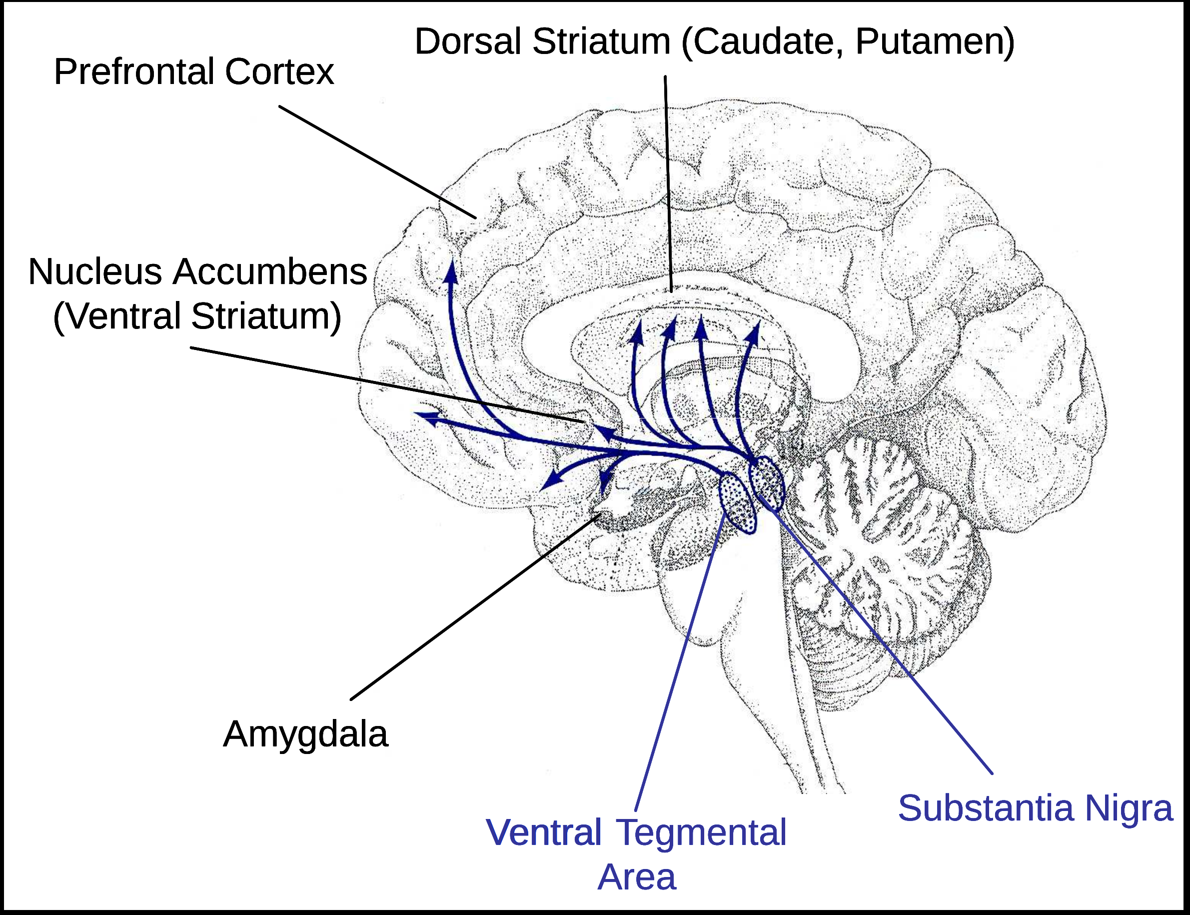
Temporal difference learning versus Rescorla-Wagner

- derived from first principles about the future
- explains everything that R-W does, and more (eg. $2^{nd}$ order conditioning)
- a generalization of R-W to real time

# Back to Marr's 3 levels

- The problem: optimal prediction of future reward
- The algorithm: temporal difference learning
- Neural implementation: does the brain use TD learning?

# Dopamine




Parkinson's Disease

→ Motor control + initiation?

Intracranial self-stimulation;

Drug addiction;

Natural rewards

→ Reward pathway?

→ Learning?

Also involved in:

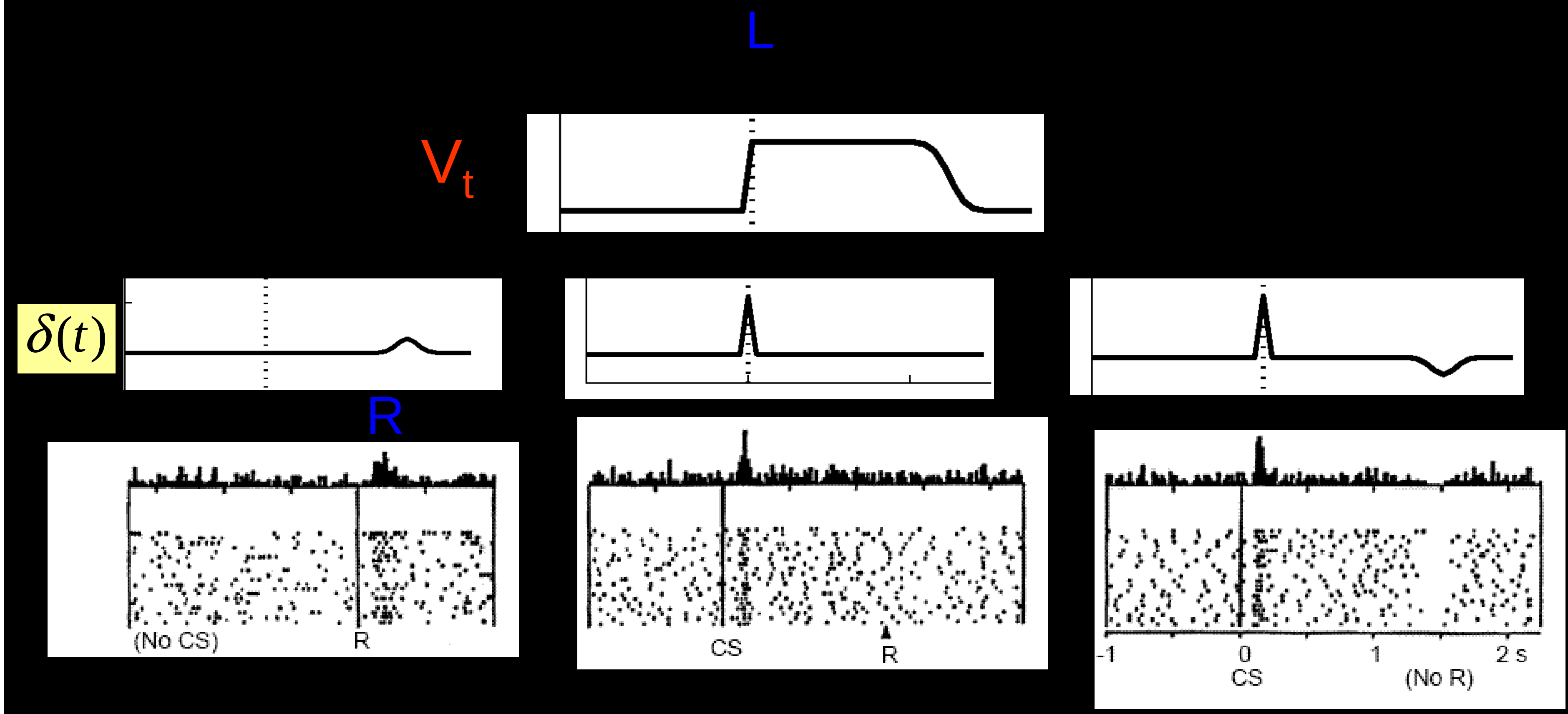
- Working memory
- Novel situations
- ADHD
- Schizophrenia
- …

# Role of dopamine: Many hypotheses

- Anhedonia hypothesis
- Prediction error (learning, action selection)
- Salience/attention
- Incentive salience
- Uncertainty
- Cost/benefit computation
- Energizing/motivating behavior

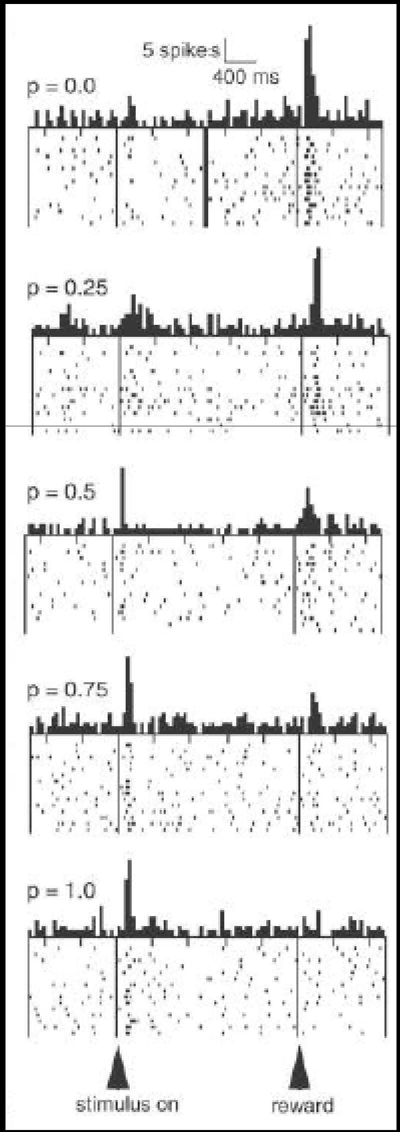
# dopamine and prediction error

TD error $$\delta_t = r_t + V_{t+1} - V_t$$




no prediction　　prediction, reward　　prediction, no reward

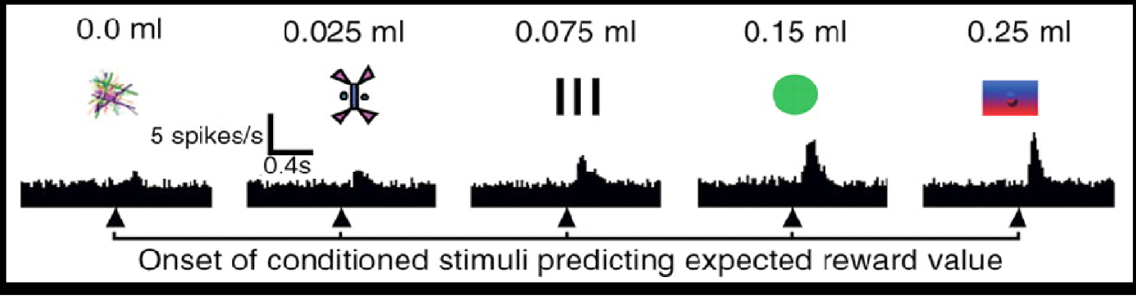
# prediction error hypothesis of dopamine



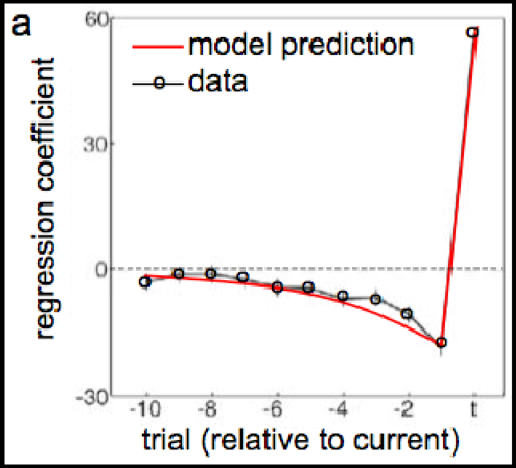

Fiorillo et al, 2003

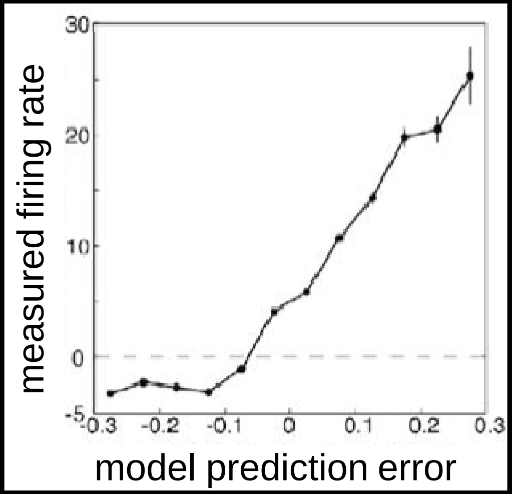
The idea: Dopamine encodes a reward prediction error

Tobler et al, 2005



# prediction error hypothesis of dopamine



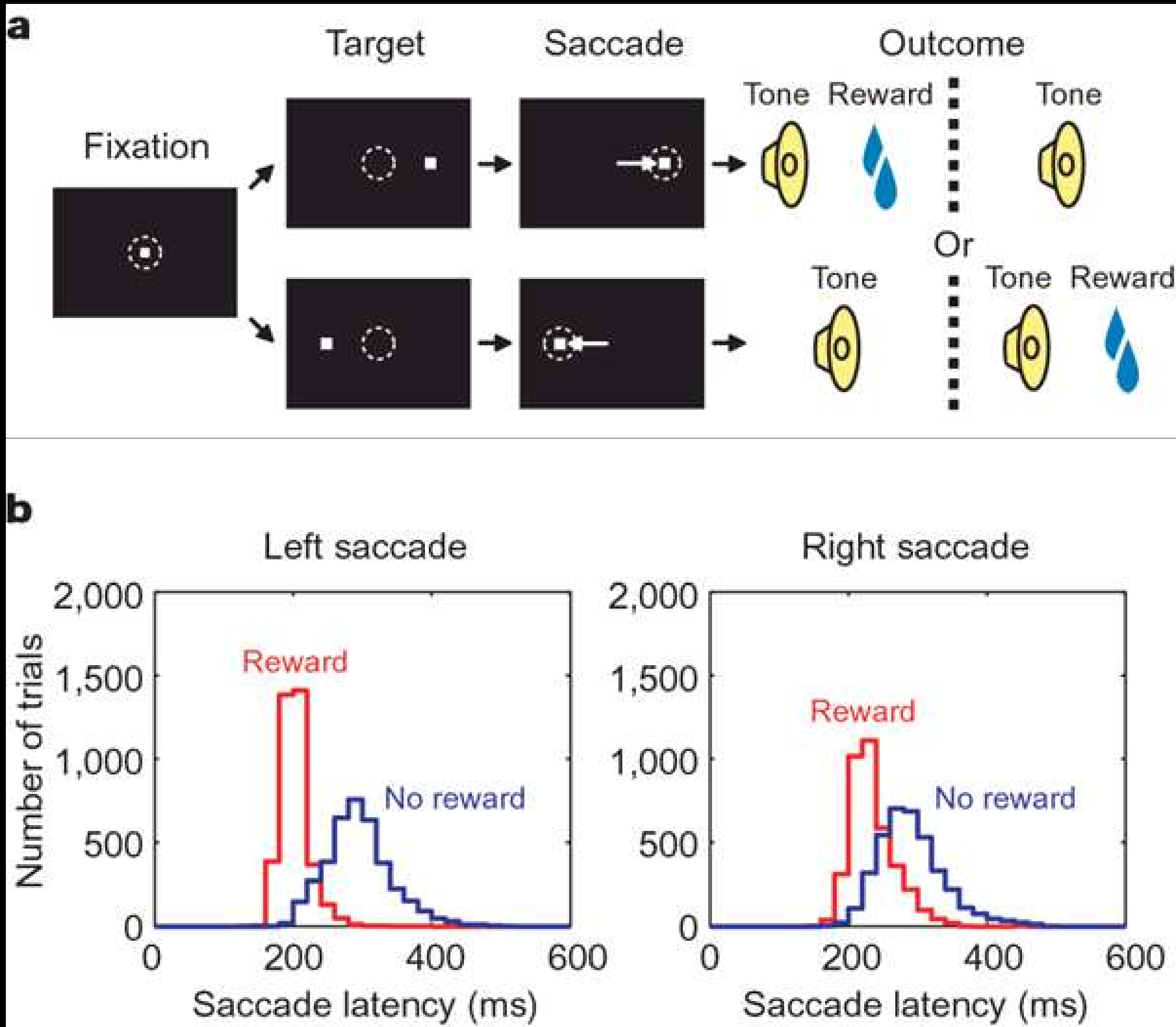




at end of trial: $\delta_t = r_t - V_t$ (just like R-W)

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

Bayer & Glimcher (2005)

# what drives the dips?

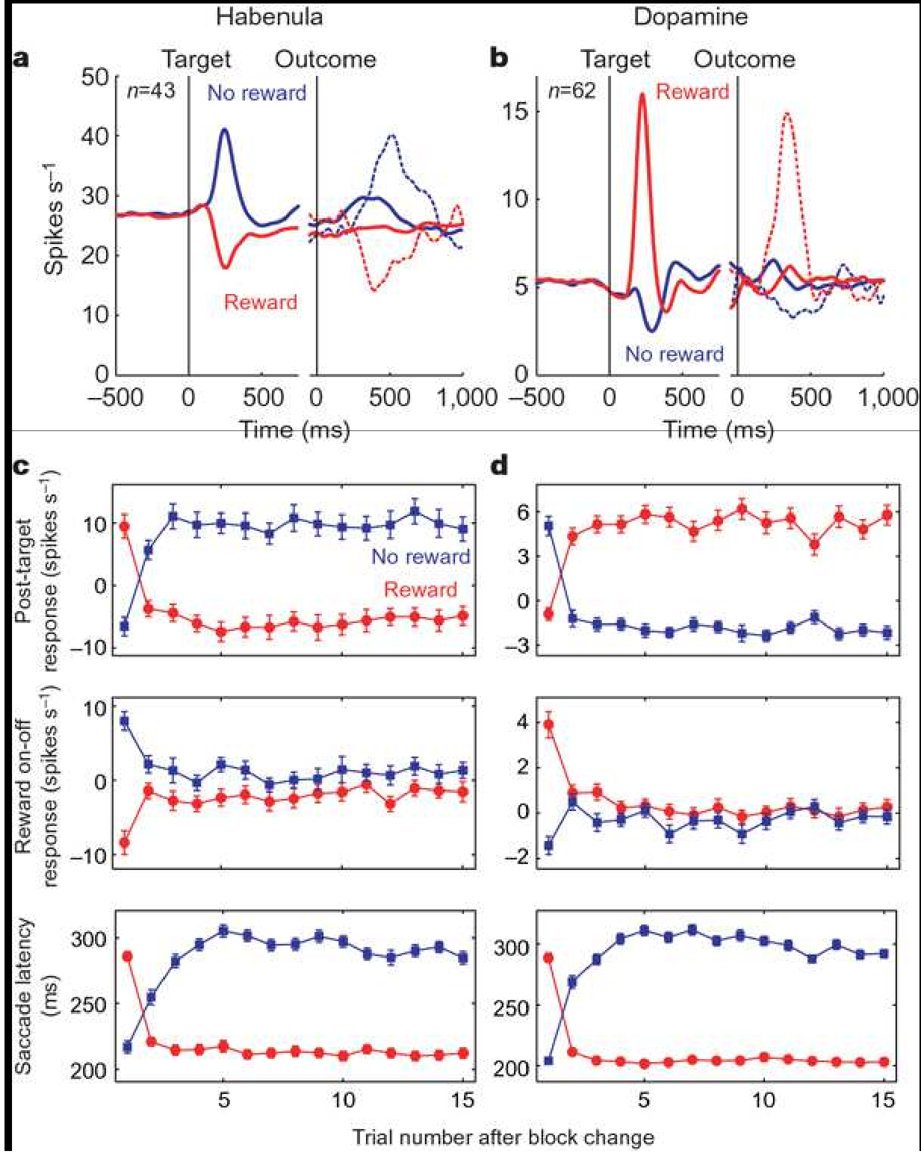


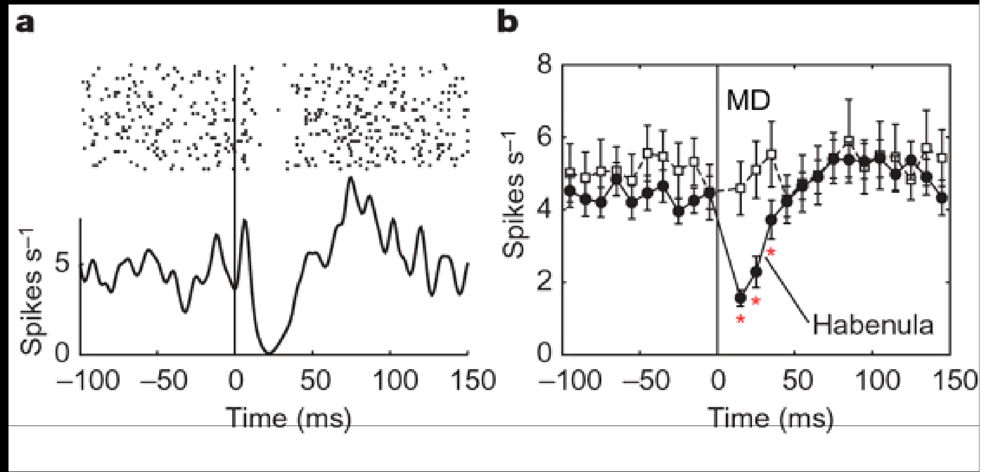

- why an effect of reward at all?
  - Pavlovian influence

Matsumoto & Hikosaka (2007)

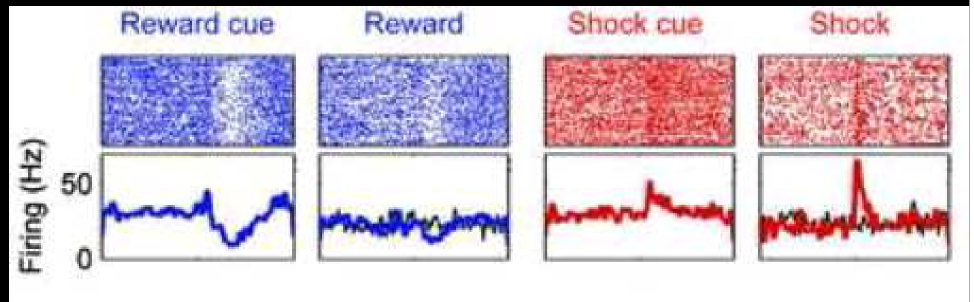
# what drives the dips?

Matsumoto & Hikosaka (2007)




- rHab -> rSTN



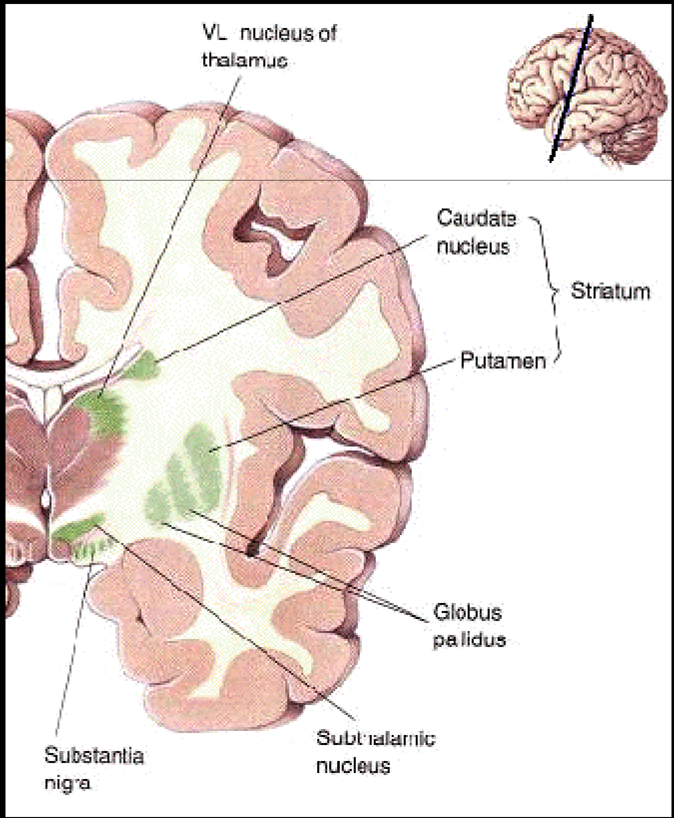

- RMTg (predicted R/S)

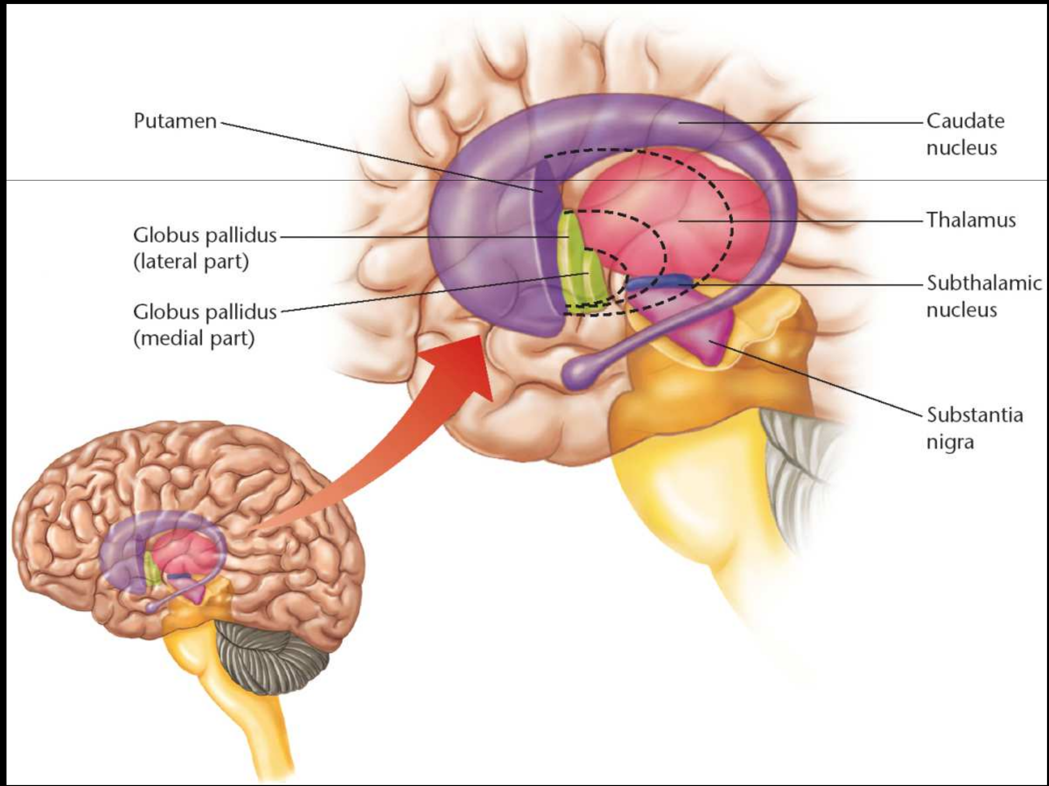
Jhou et al, 2009

# Where does dopamine project to? Basal ganglia

Several large subcortical nuclei

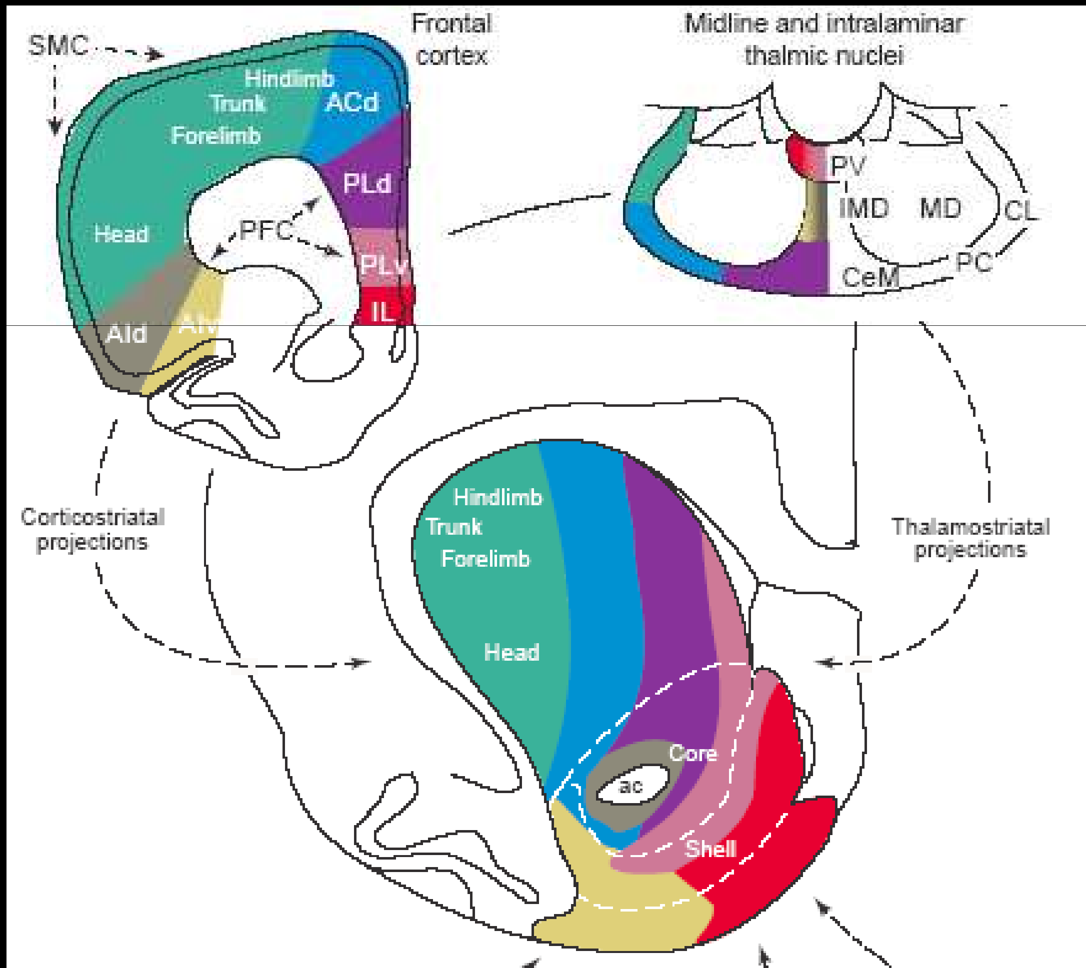
(unfortunate anatomical names follow structure rather than function, eg caudate + putamen + nucleus accumbens are all relatively similar pieces of striatum; but globus pallidus & substantia nigra each comprise two different things)



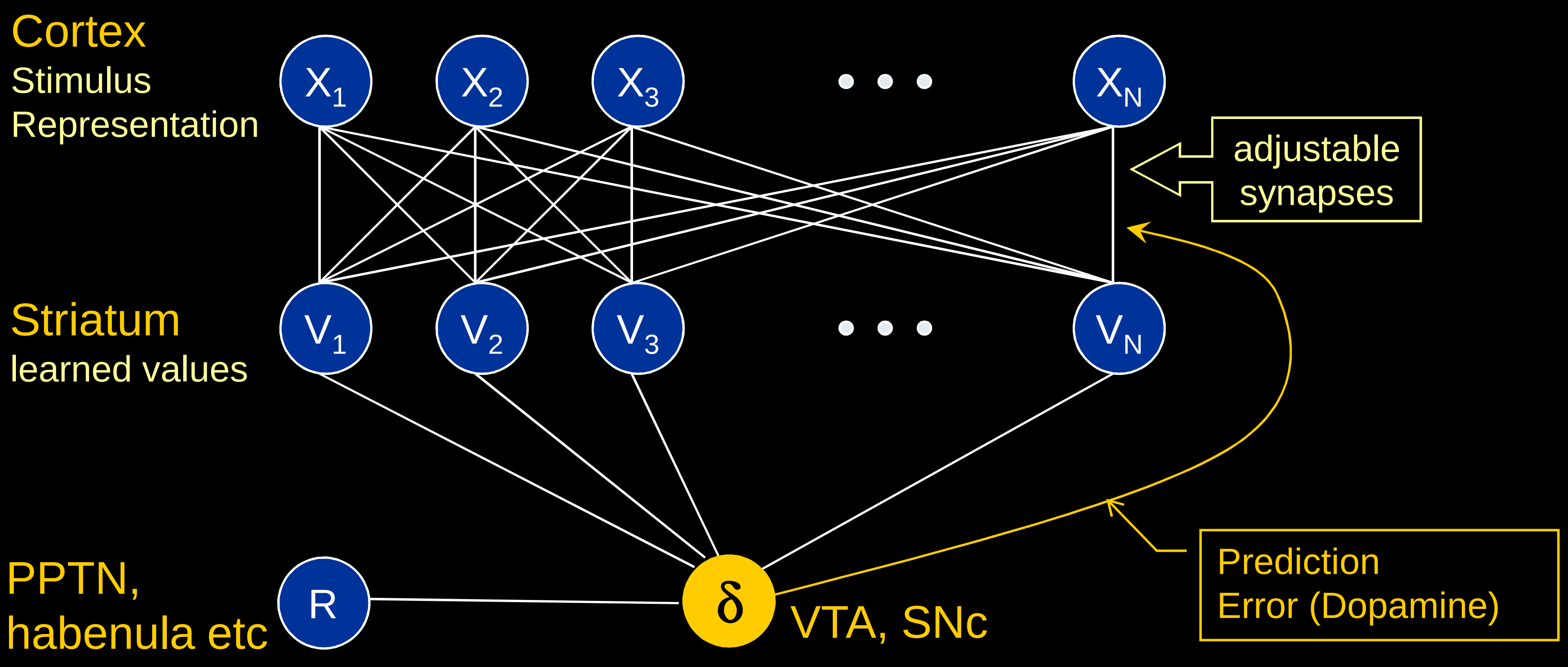
# Where does dopamine project to? Basal ganglia

inputs to BG are from all over the cortex (and topographically mapped)



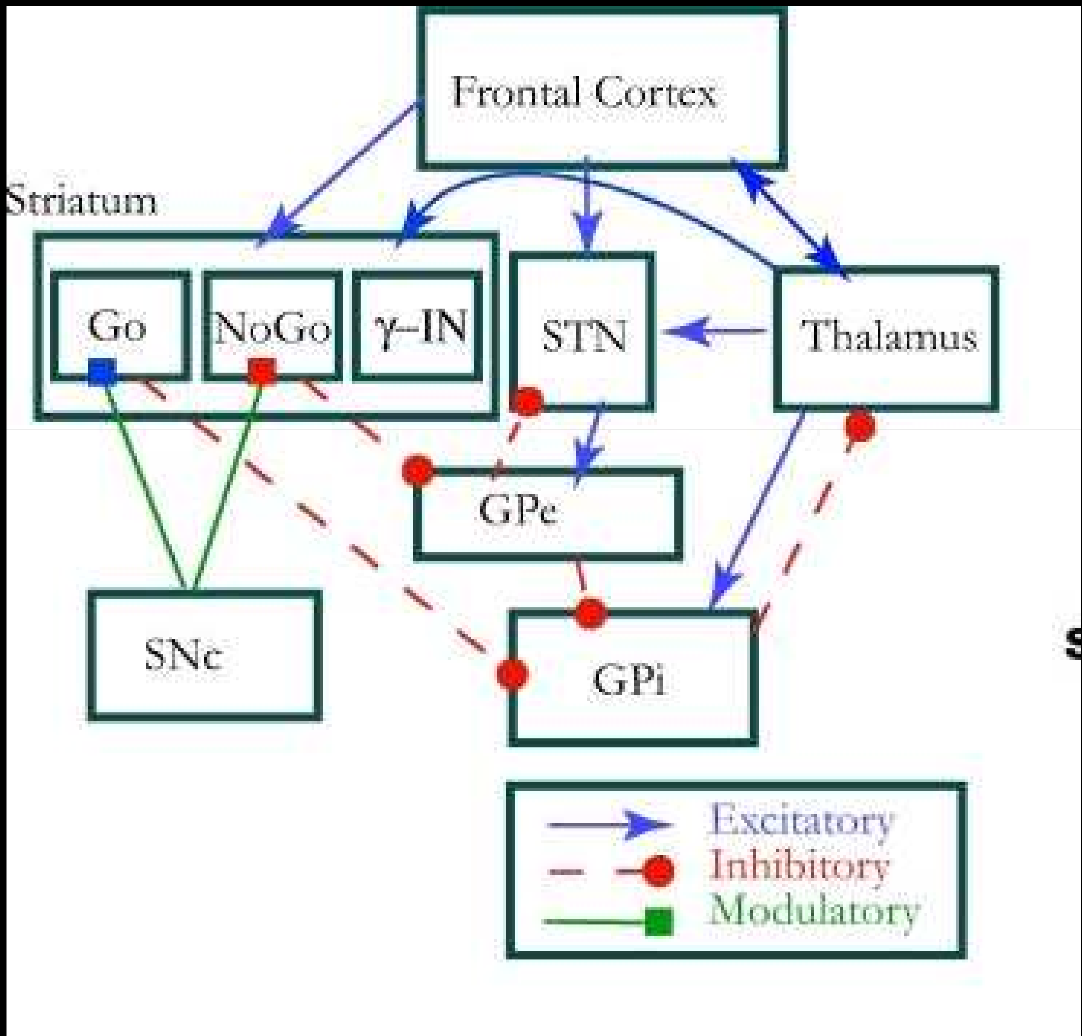

Voorn et al, 2004

# Corticostriatal synapses: 3 factor learning




but also amygdala; orbitofrontal cortex; ...

# striatal complexities

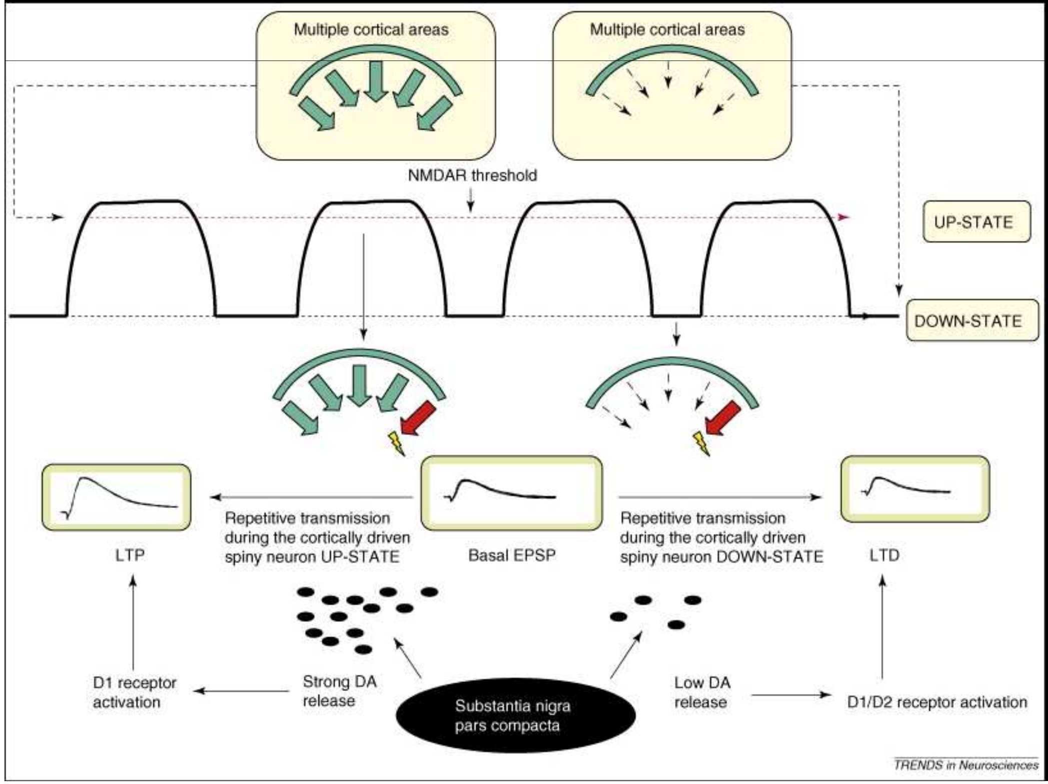


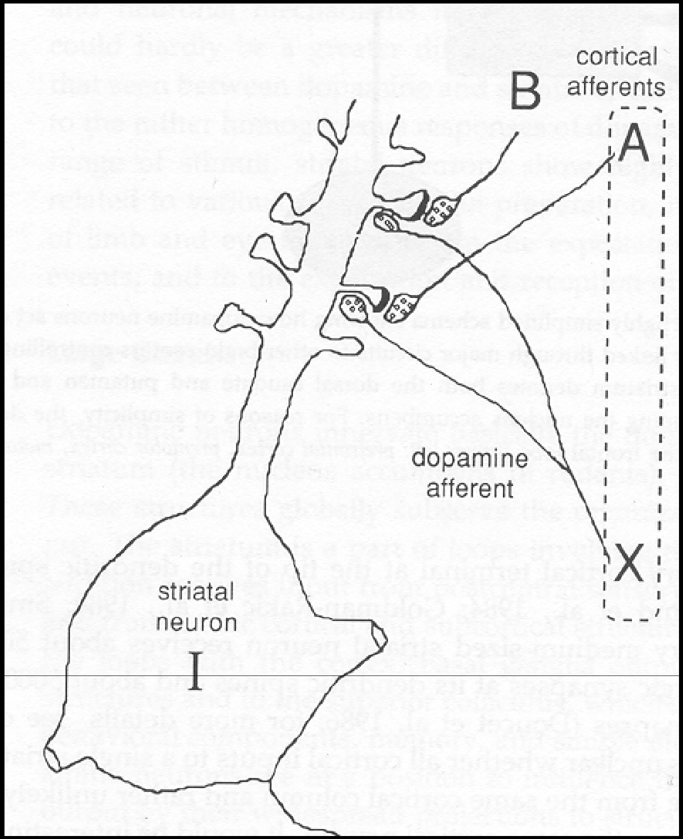

Cohen & Frank, 2009

# Dopamine and plasticity

Prediction errors are for learning…

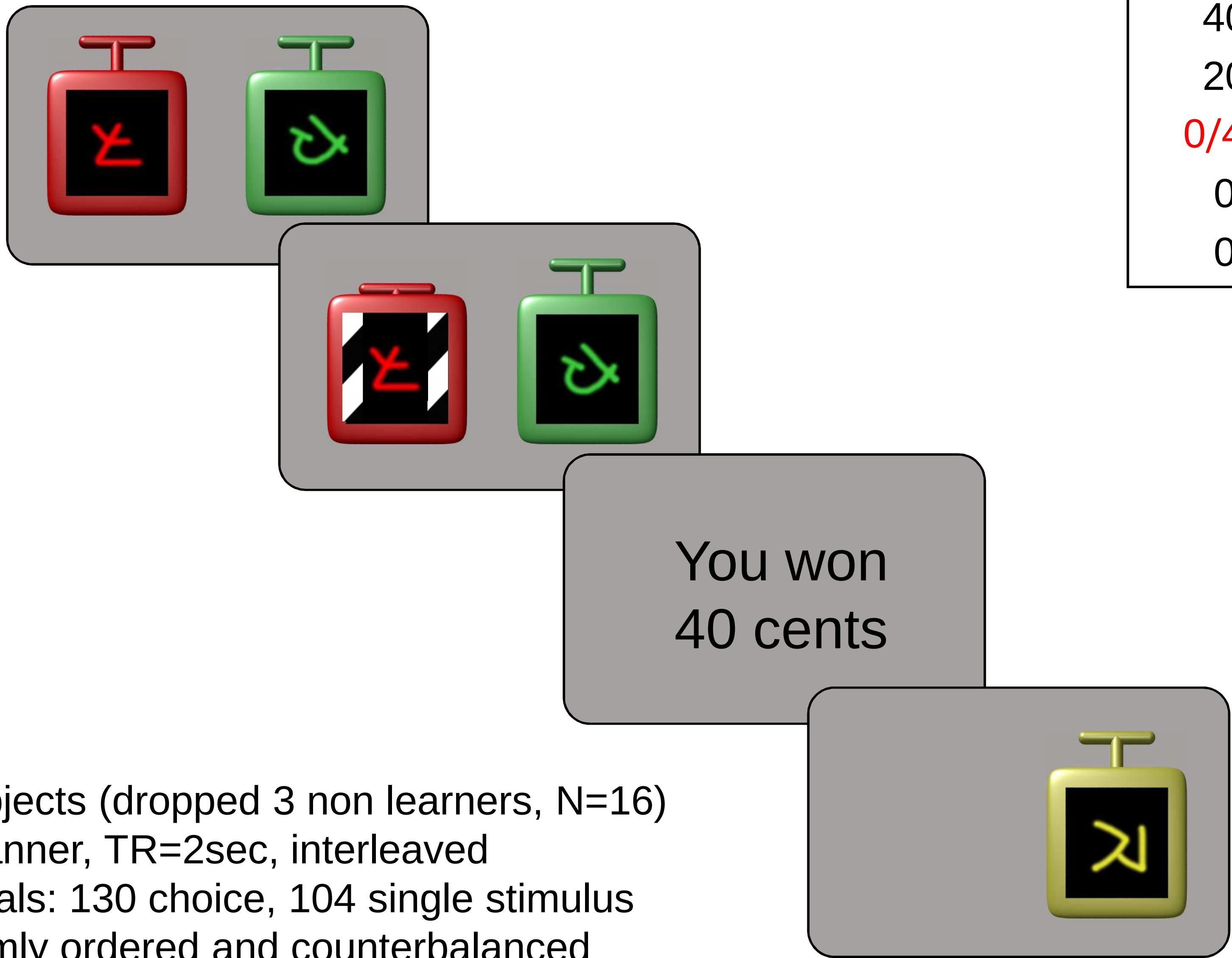
Cortico-striatal synapses show complex dopamine-dependent plasticity




Wickens et al, 1996

# Risk Experiment

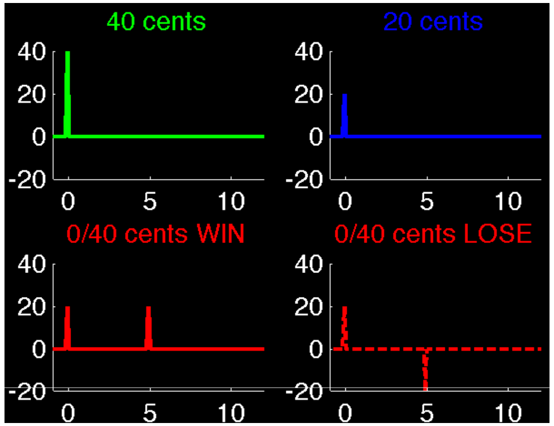


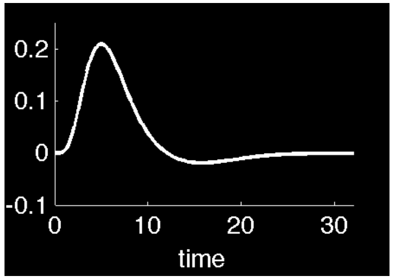

<u>5 stimuli</u>:
40¢
20¢
0/40¢
0¢
0¢

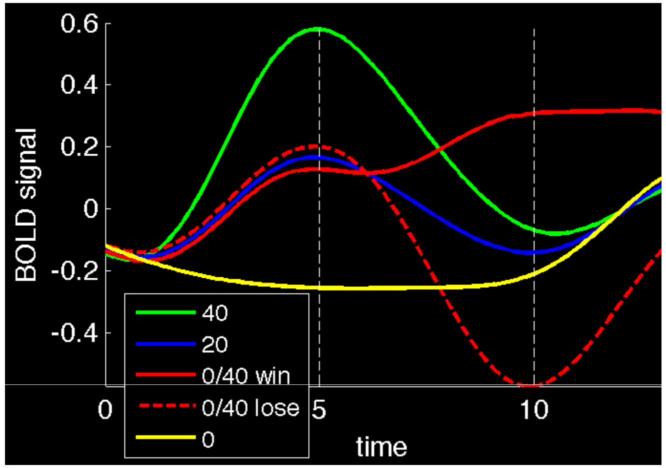
19 subjects (dropped 3 non learners, N=16)
3T scanner, TR=2sec, interleaved
234 trials: 130 choice, 104 single stimulus
randomly ordered and counterbalanced

# Neural results: Prediction Errors

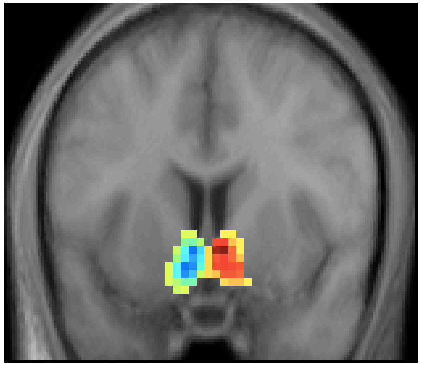
what would a prediction error look like (in BOLD)?



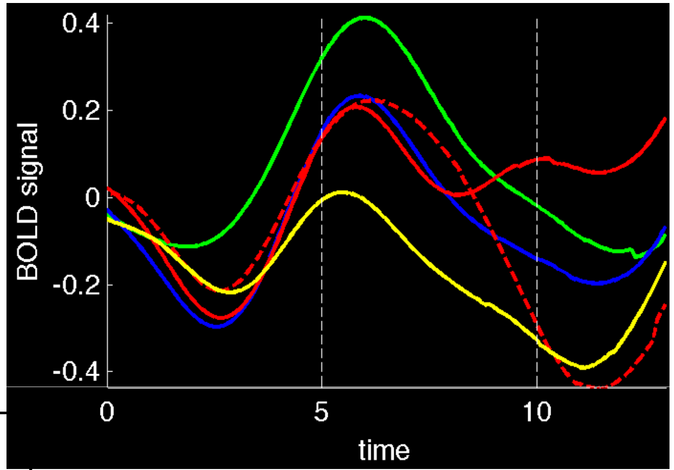



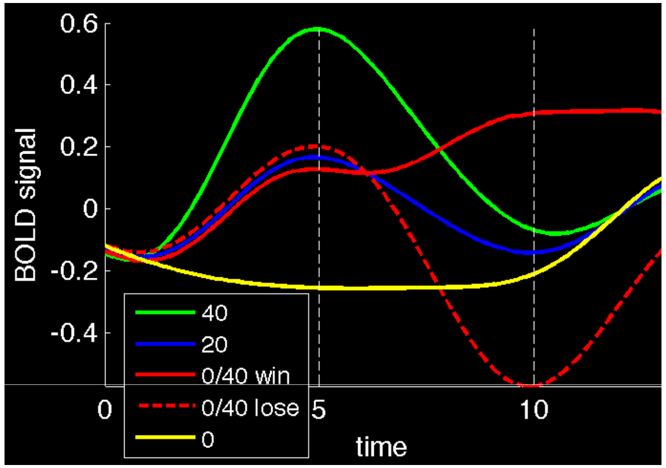
# Neural results: Prediction errors in NAC



unbiased anatomical ROI
in nucleus accumbens
(marked per subject*)




raw BOLD
(avg over all subjects)

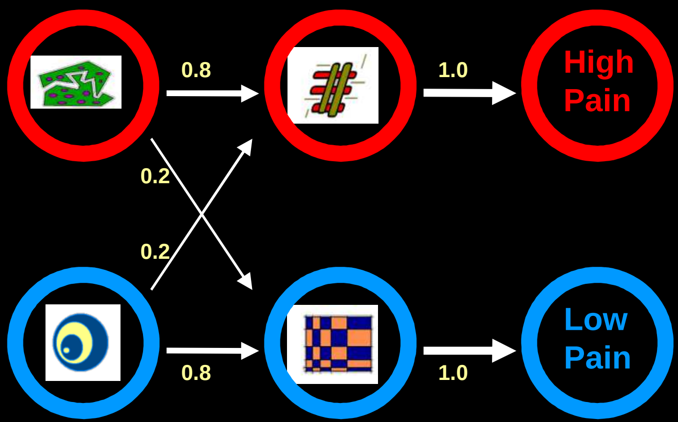


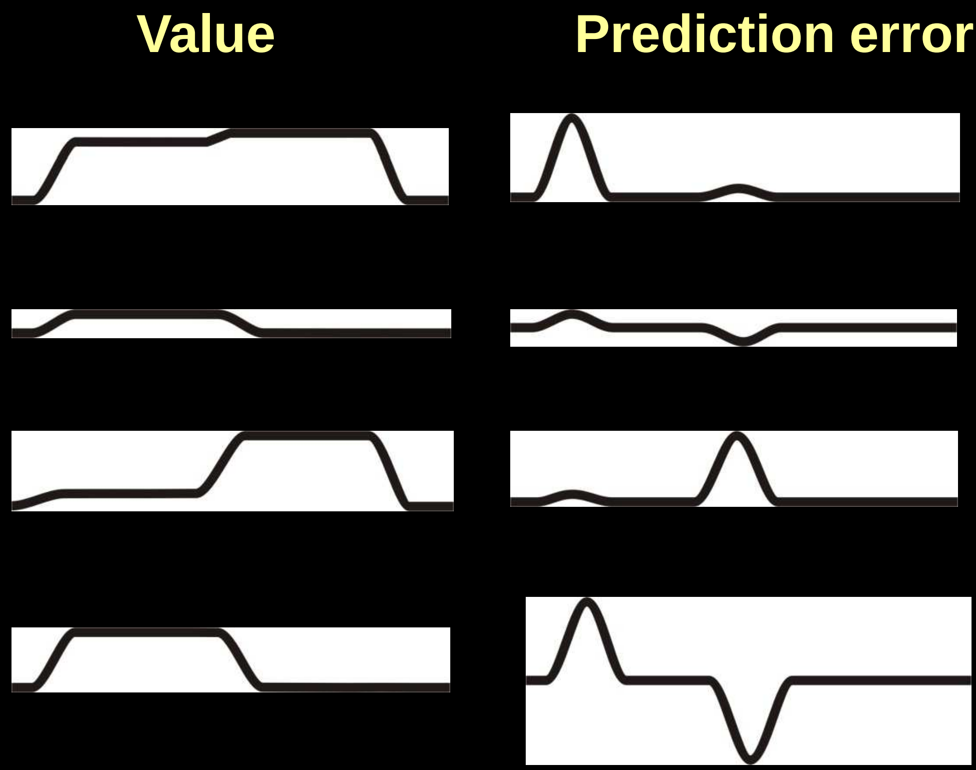

can actually decide between different neuroeconomic models of risk

* thanks to Laura deSouza

# punishment prediction error

TD error $$\delta_t = r_t + V_{t+1} - V_t$$



# punishment prediction error

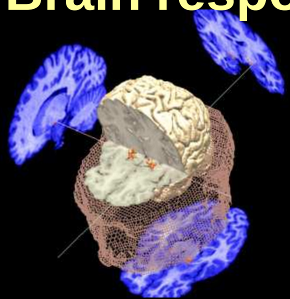
**experimental sequence…..**

A – B – HIGH   C – D – LOW   C – B – HIGH   A – B – HIGH   A – D – LOW   C – D – LOW   A – B – HIGH   A – B – HIGH   C – D – LOW   C – B – HIGH

**MR scanner**

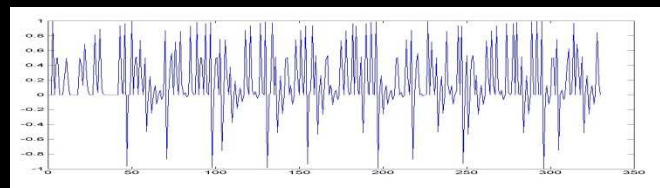
**TD model**

**Brain responses**



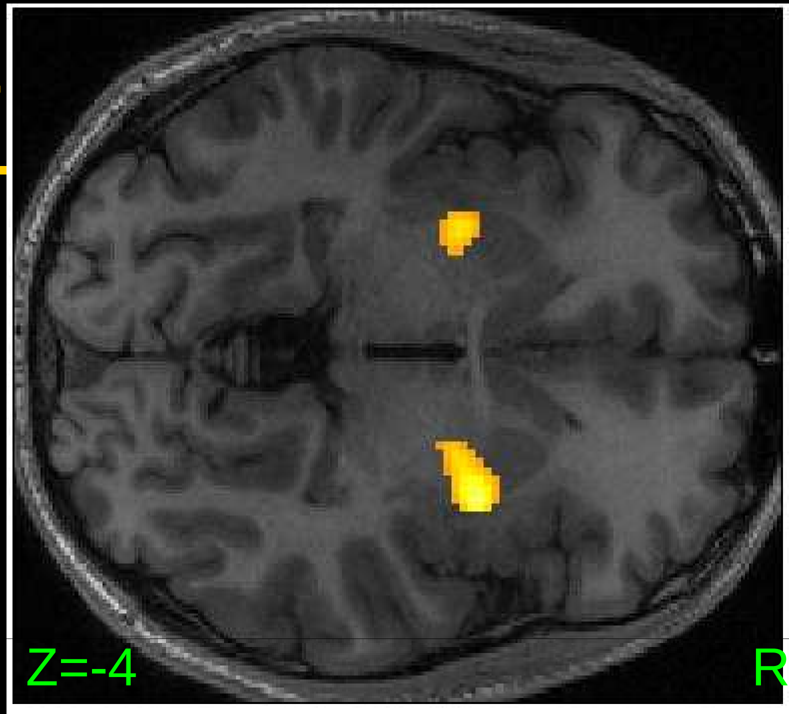
?

**Prediction error**

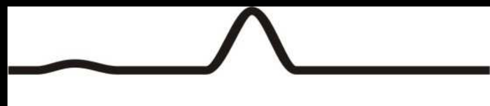


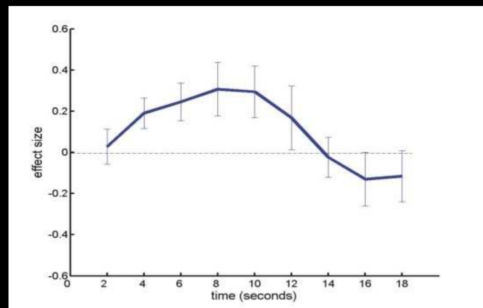
Ben Seymour; John O’Doherty

# punishment prediction error

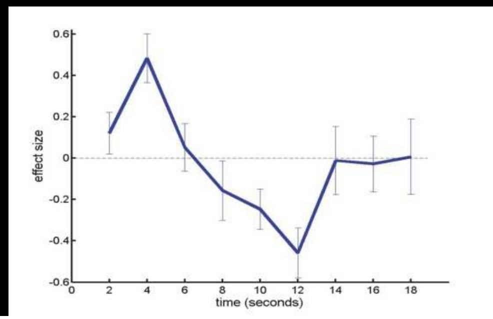



**TD prediction error**:
ventral striatum

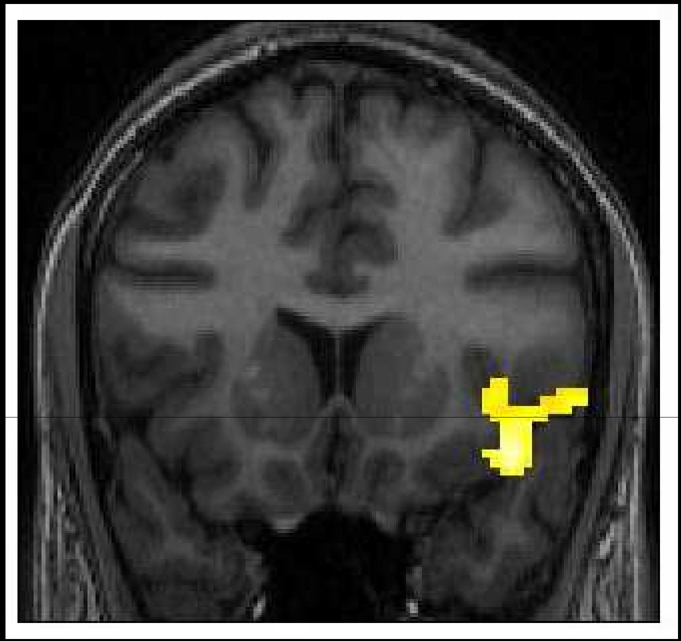





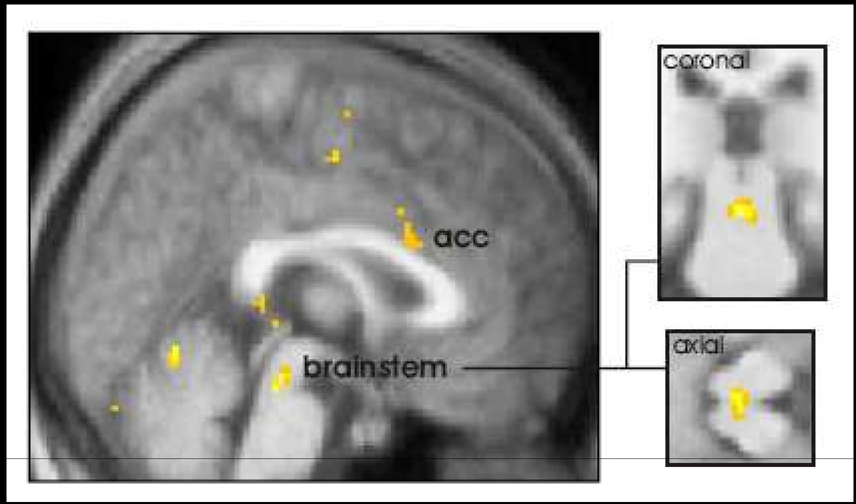




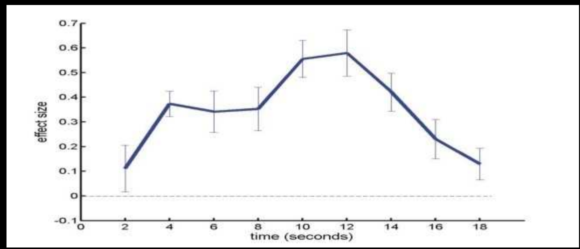
# punishment prediction



right anterior insula

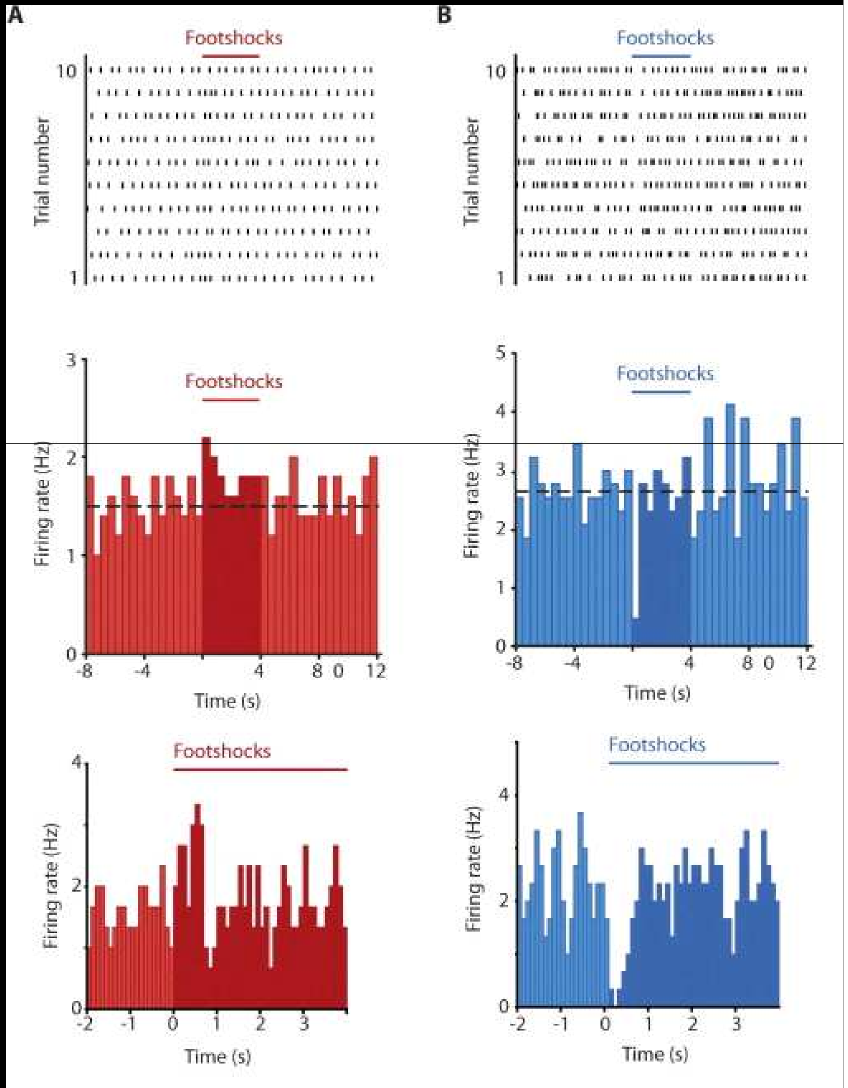



dorsal raphe (5HT)?

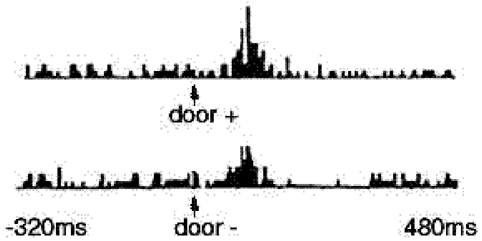
# punishment

- dips below baseline in dopamine
  – Frank: D2 receptors particularly sensitive
  – Bayer & Glimcher: length of pause related to size of negative prediction error
- but:
  – can't afford to wait that long
  – negative signal for such an important event
  – opponency a more conventional solution:
    - serotonin…



# generalization

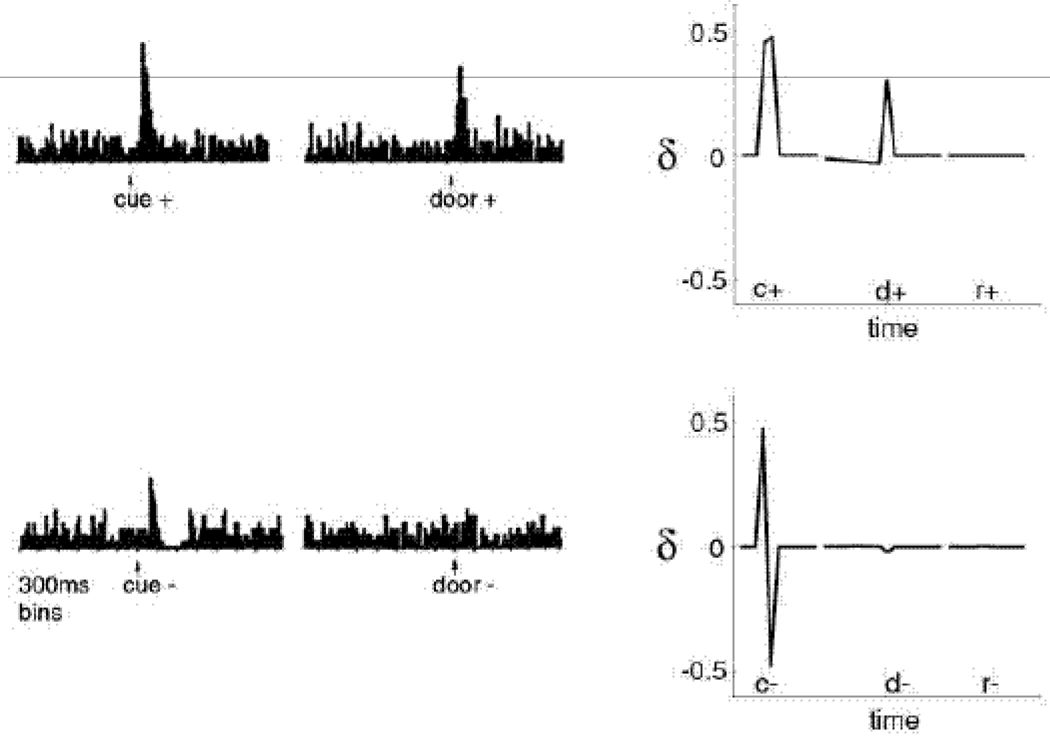
dopamine cells also respond for similar stimuli:



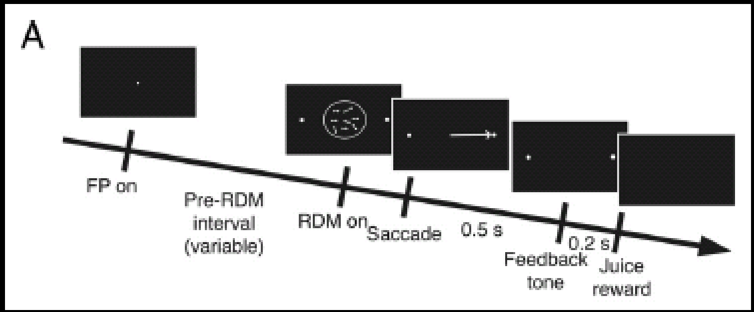
# generalization

what if there is

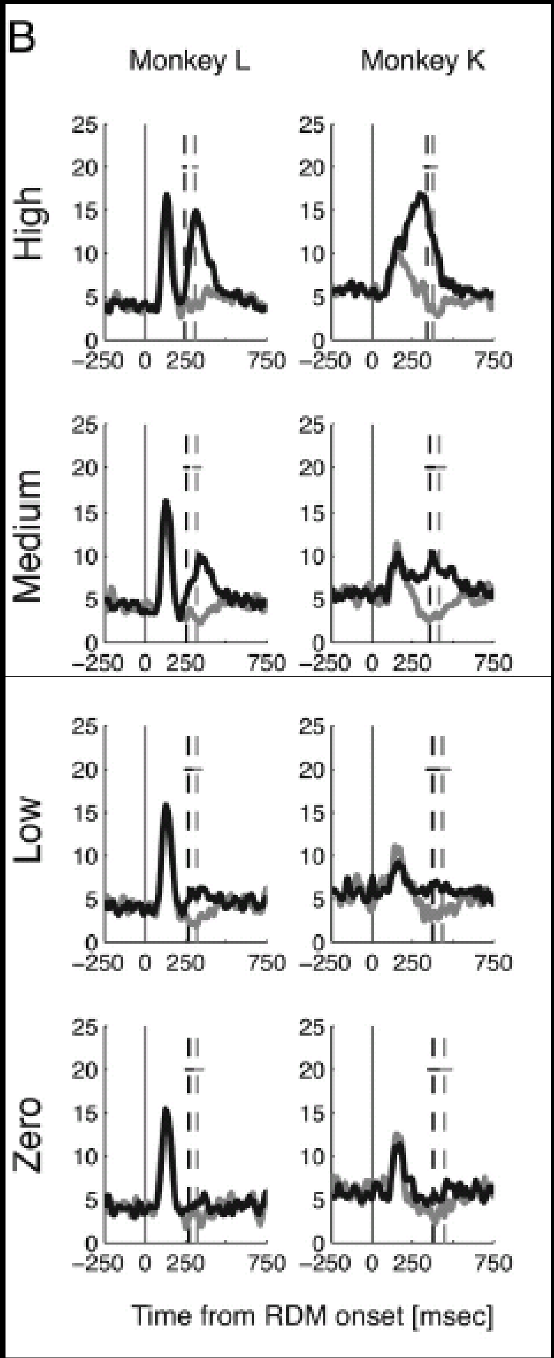
- generalizing cue before the door?
- random interval between cue and door?



# random-dot discrimination




differential reward (0.16ml; 0.38ml)

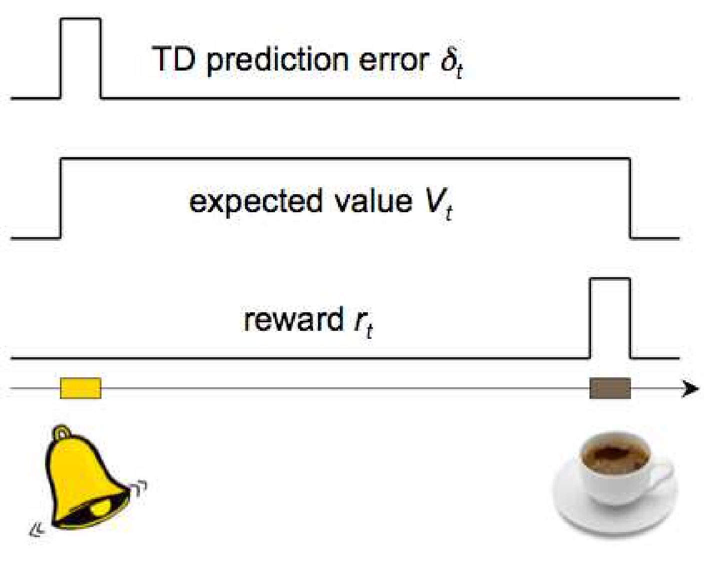



Sakagami (2010)

# other paradigms

- inhibitory conditioning
- transreinforcer blocking
- motivational sensitivities
- backwards blocking
  - Kalman filtering
- downwards unblocking
- primacy as well as recency (highlighting)
  - assumed density filtering

# Summary of this part: 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

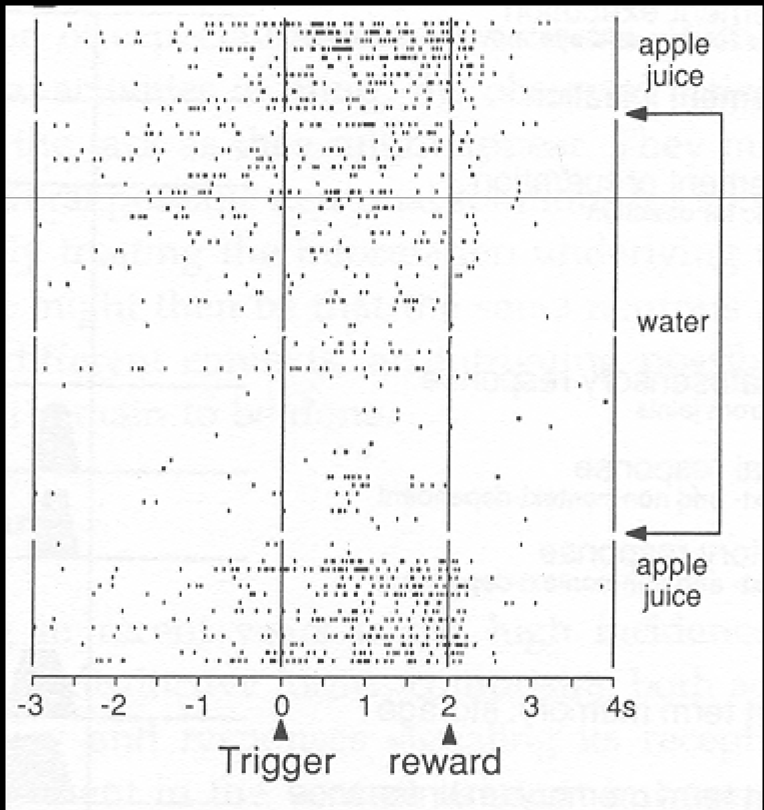
⇒ Precise (normative!) theory for generation of dopamine firing patterns

⇒ Explains anticipatory dopaminergic responding, second order conditioning

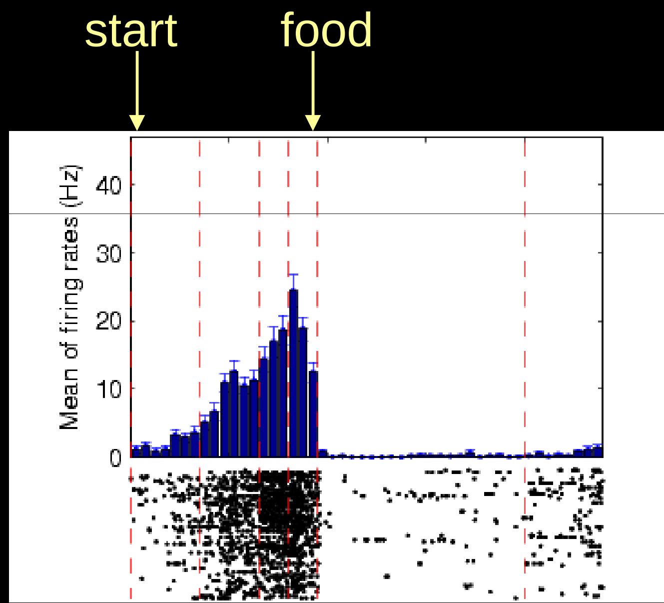
⇒ Compelling account for the role of dopamine in classical conditioning: prediction error acts as signal driving learning in prediction areas

# Striatum and learned values

Striatal neurons show ramping activity that precedes a reward (and changes with learning!)




(Schultz)




(Daw)

# Phasic dopamine also responds to…

- Novel stimuli
- Especially salient (attention grabbing) stimuli
- Aversive stimuli (??)

- Reinforcers and appetitive stimuli induce approach behavior and learning, but also have attention functions (elicit orienting response) and disrupt ongoing behaviour.

→ Perhaps DA reports salience of stimuli (to attract attention; switching) and not a prediction error? (Horvitz, Redgrave)