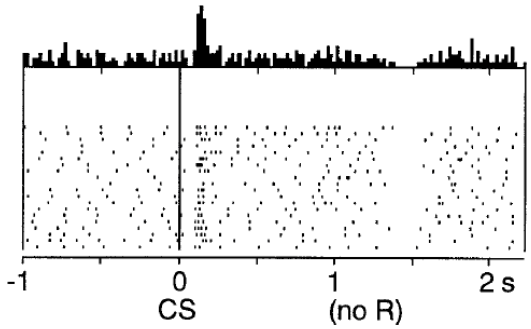
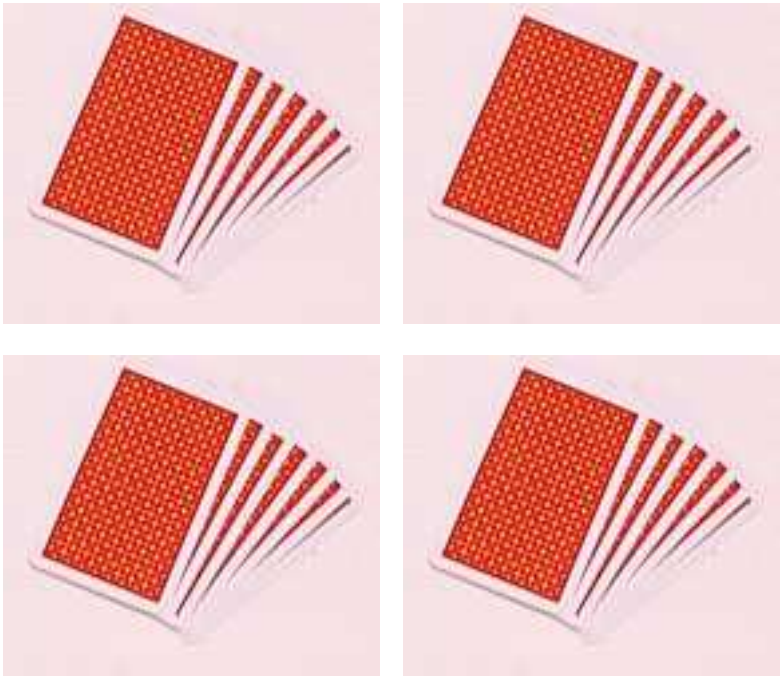


Dopamine & reinforcement learning

outline

- learning behavior
- basal ganglia & dopamine
- responses & interpretation



repeated trial-and-error decision makling

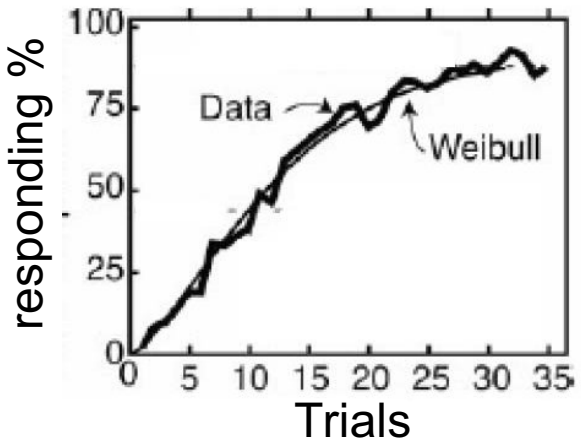
strategy:

1. Predict the outcomes
2. Choose the best
3. Learn from experience to improve predictions

Classical conditioning

- Pair stimulus (bell, light)
- …with significant event (food, shock)
- Measure anticipatory behavior (salivation, freezing)

Rescorla-Wagner (72) model



"error-driven" learning:

minimize discrepancy between received reward *r* and predicted reward *V*

Predict: $V_t = \Sigma_i w_{i,t}$ for each presented stimulus *i*

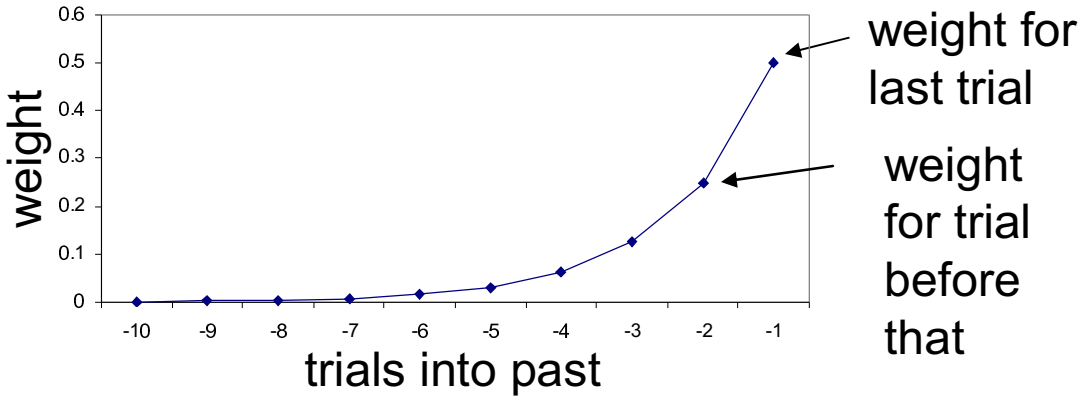
Learn: $w_{i.t+1} = w_{i,t} + \varepsilon\delta_t$; $\delta_t = (r_t - V_t)$; for each presented stimulus *I*

predicts phenomena like "blocking" – no learning without prediction error

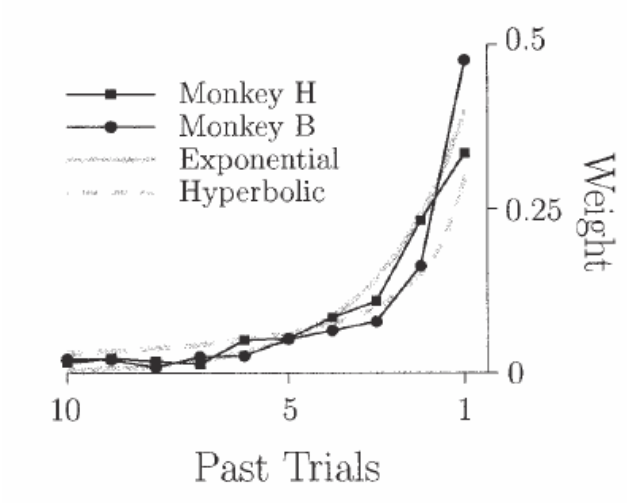
the R-W rule estimates the expected reward using an exponentially weighted average of recently received rewards:

$w_{i,t+1} = w_{i,t} + \varepsilon\delta_t$; $\delta_t = (r_t - V_t)$

$w_{i,t+1} = \varepsilon r_t + (1-\varepsilon)\, w_{i,t}$



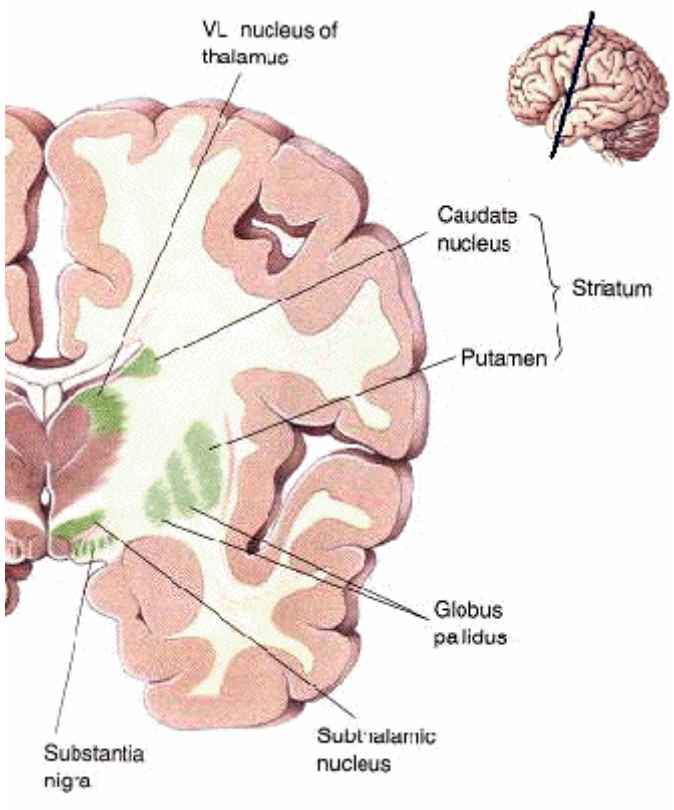
the influence of past rewards on animals' choice behavior also shows this form:

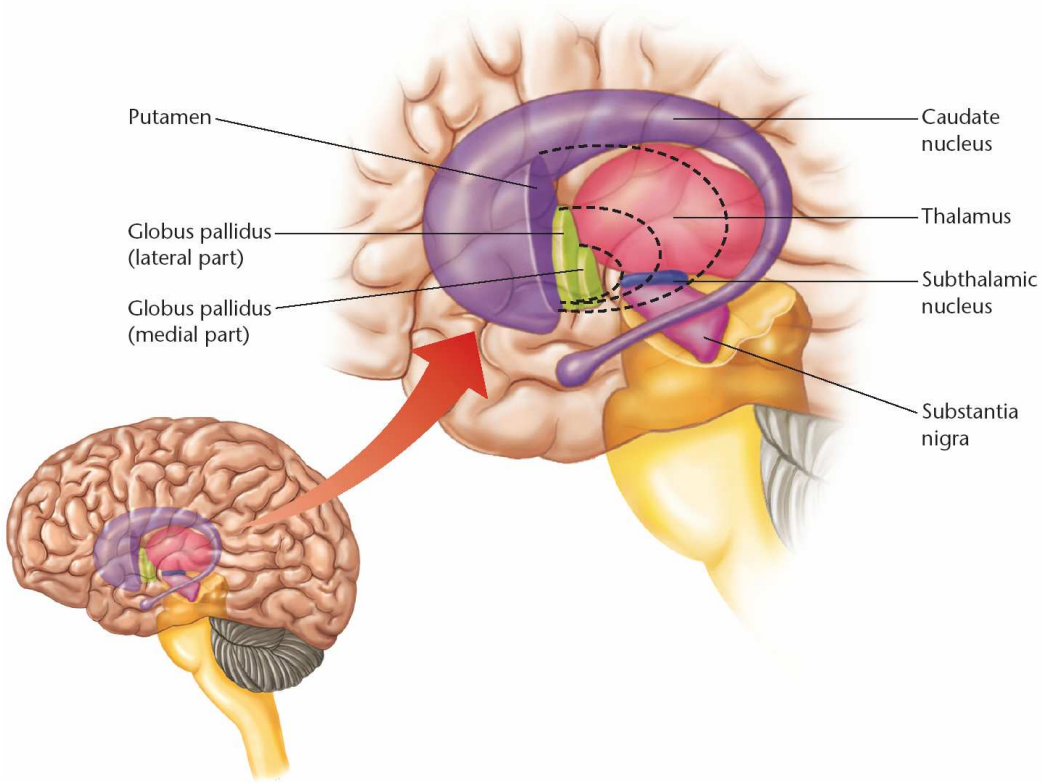


(lau & glimcher 2005)

Basal ganglia

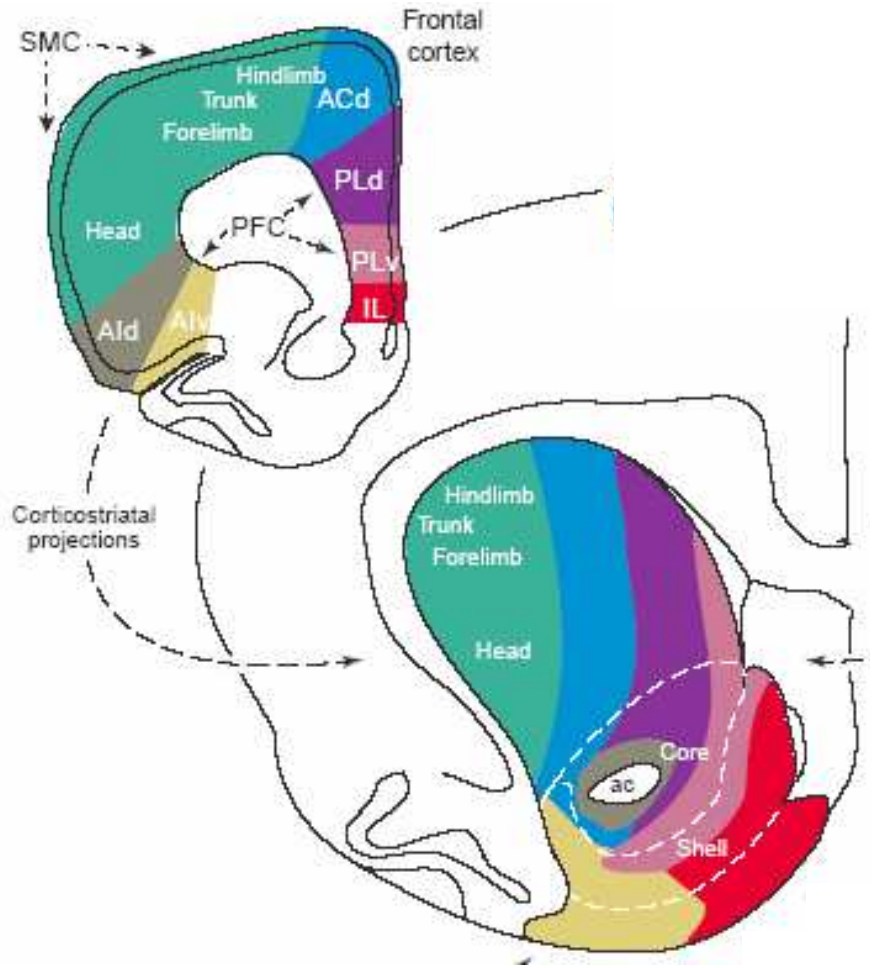
- Several large subcortical nuclei
  - unfortunate latin names follow proximity rather than function (eg caudate + putamen + nucleus accumbens are all pieces of striatum; but globus pallidus & substantia nigra each comprise two different things)





Basal ganglia input

- Projection from entire cortex (including sensory, motor, associative areas) to striatum
- Topographic

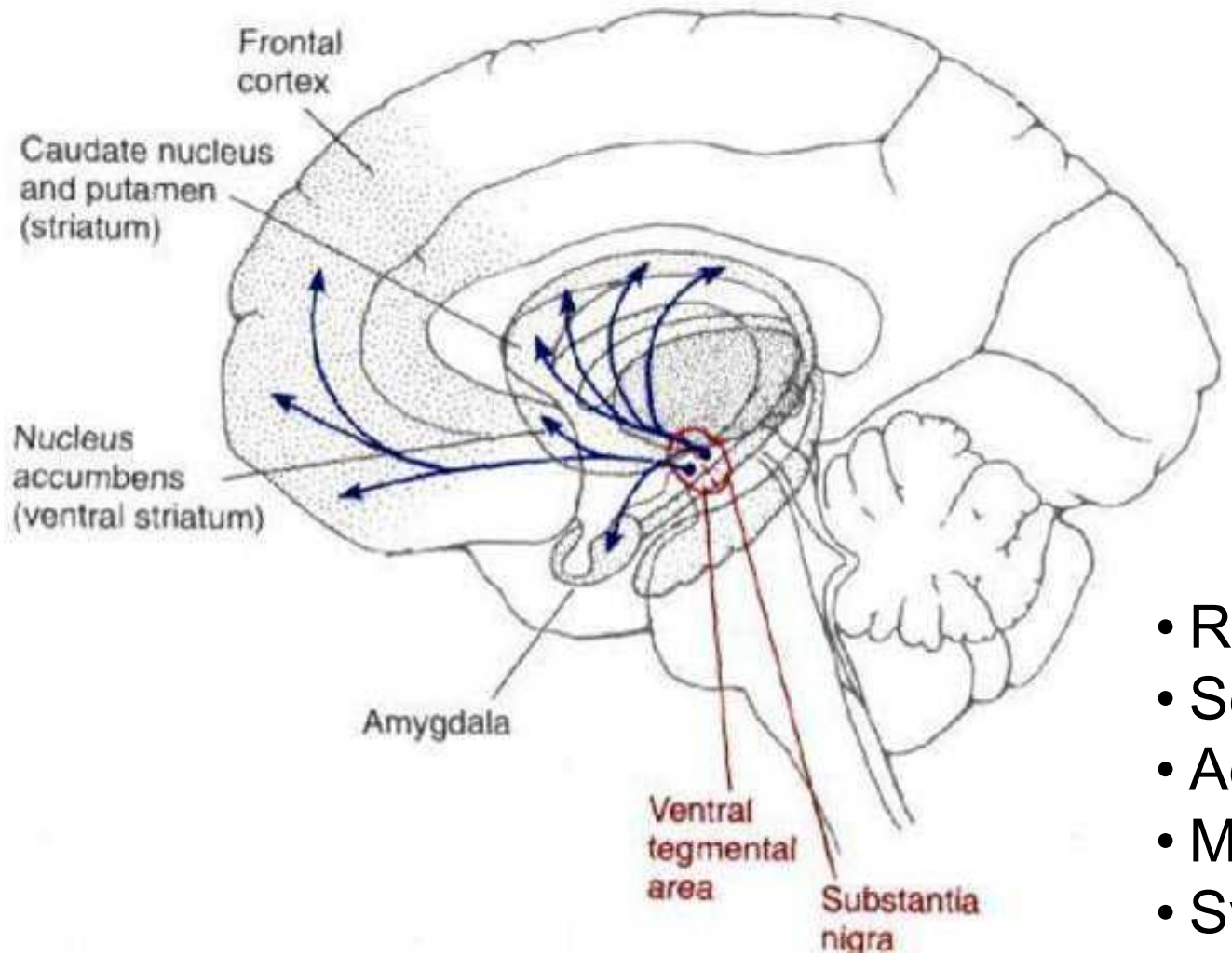


Voorn et al 2004

Basal ganglia functions

- Motor control *plus*
  - Range of motor disorders
  - But also drug abuse, reward, motivation
- Particular ideas (many overlapping)
  - Action selection or facilitation and suppression
  - Behavioral switching
  - Behavioral monitoring / regulation
  - Sequential movements
  - Internally generated movements (or stimulus-cued habits!)
  - “Limbic/motor gateway”
  - Reinforcement learning

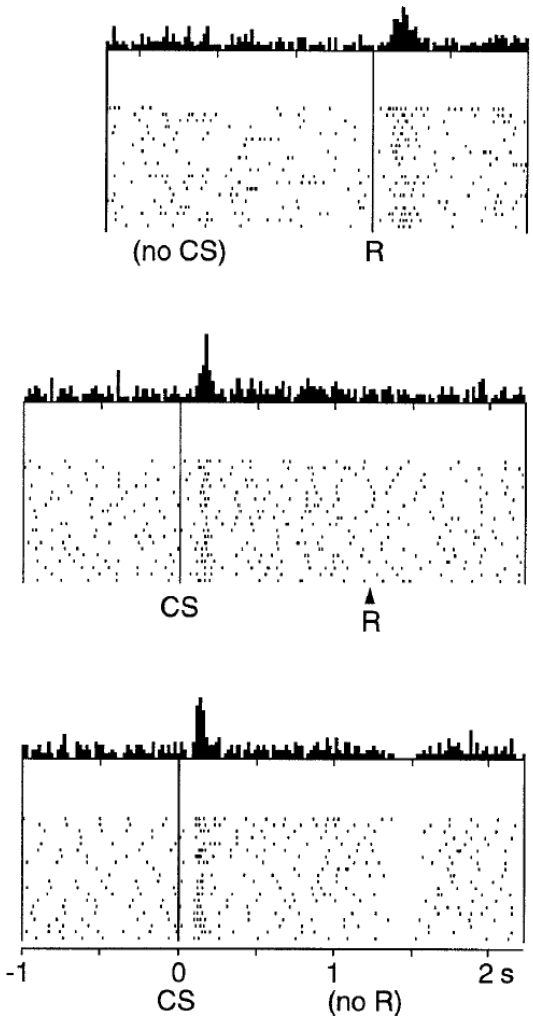
Dopamine



- Reward
- Self-stimulation
- Addiction
- Motor control
- Synaptic plasticity

(Kandel and Schwartz)

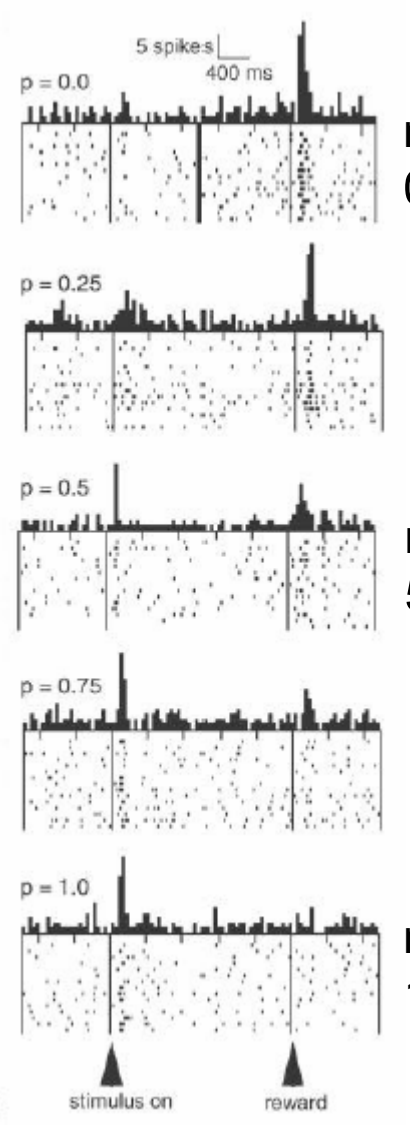
Typical dopamine responses



- Burst to unexpected reward
- Response transfers to reward predictors
- Pause at time of omitted reward

(Schultz et al. 1997)

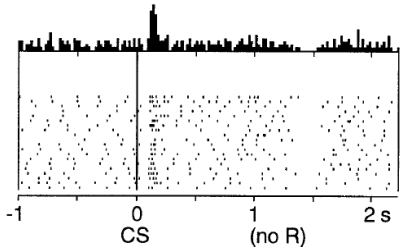
More dopamine responses



reward following
0% predictive cue

reward following
50% predictive cue

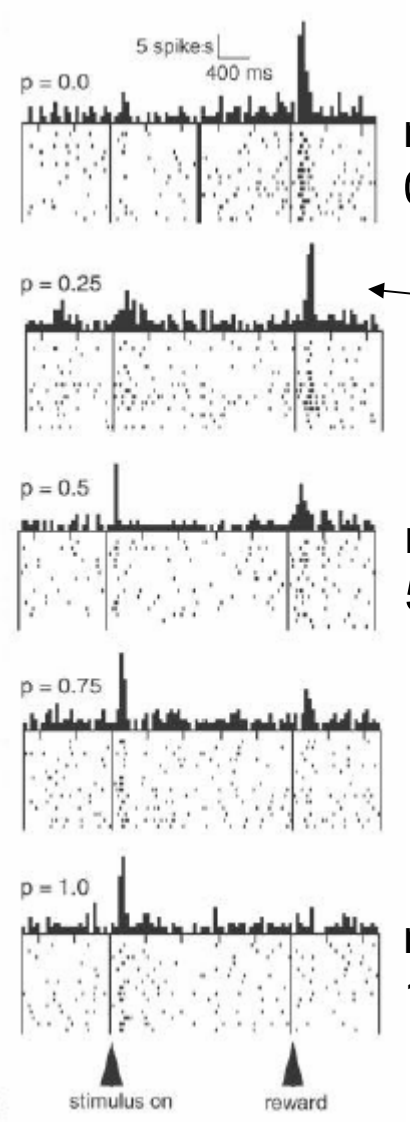
reward following
100% predictive cue



no reward following
100% predictive cue

(Fiorillo et al 2003)

More dopamine responses

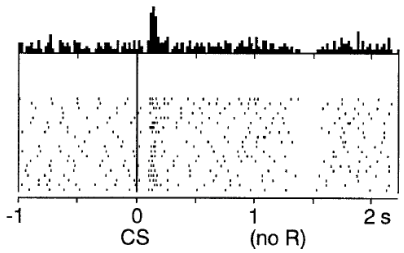


reward following
0% predictive cue

reward following
50% predictive cue

reward following
100% predictive cue

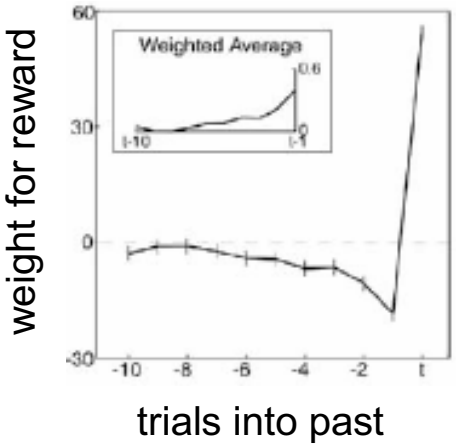
Prediction error:
$\delta_t = r_t - V_t$



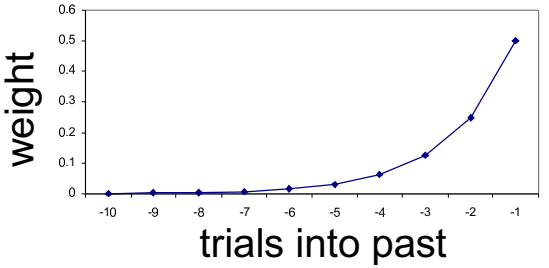
no reward following
100% predictive cue

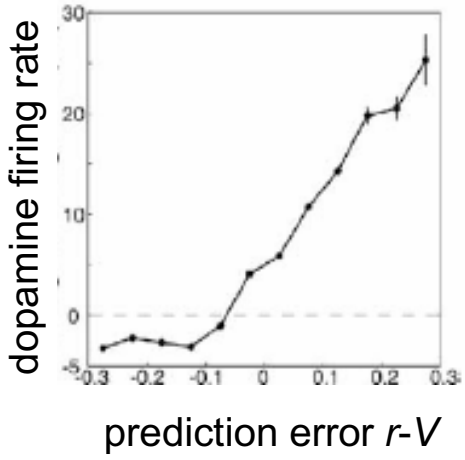
(Fiorillo et al 2003)

Prediction error

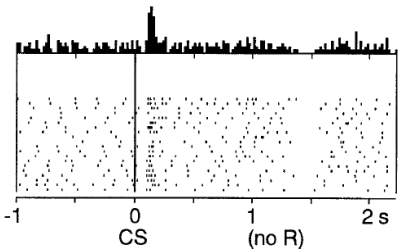


express dopamine response to reward as weighted sum of current & past rewards
→ looks like current *r* minus weighted average of past *r*s (*r* – *V*)





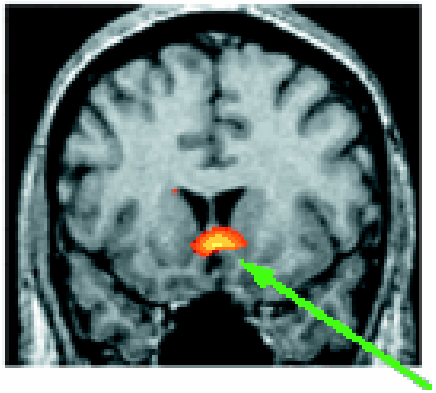
dopamine response to reward as function of prediction error *r* – *V*
→ quite linear; negative error cut off due to low baseline response



(Bayer & Glimcher 2004)

Prediction error in humans

BOLD response to reward in striatum (chief DA target) is modulated by prediction

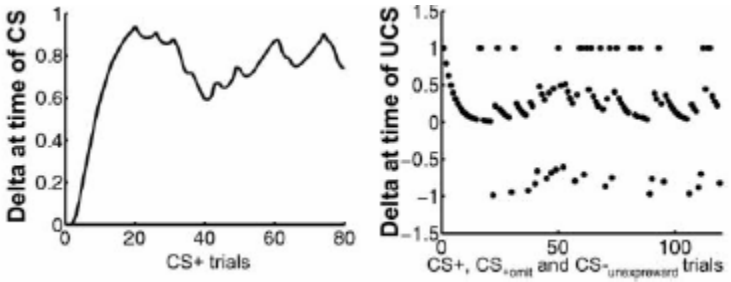


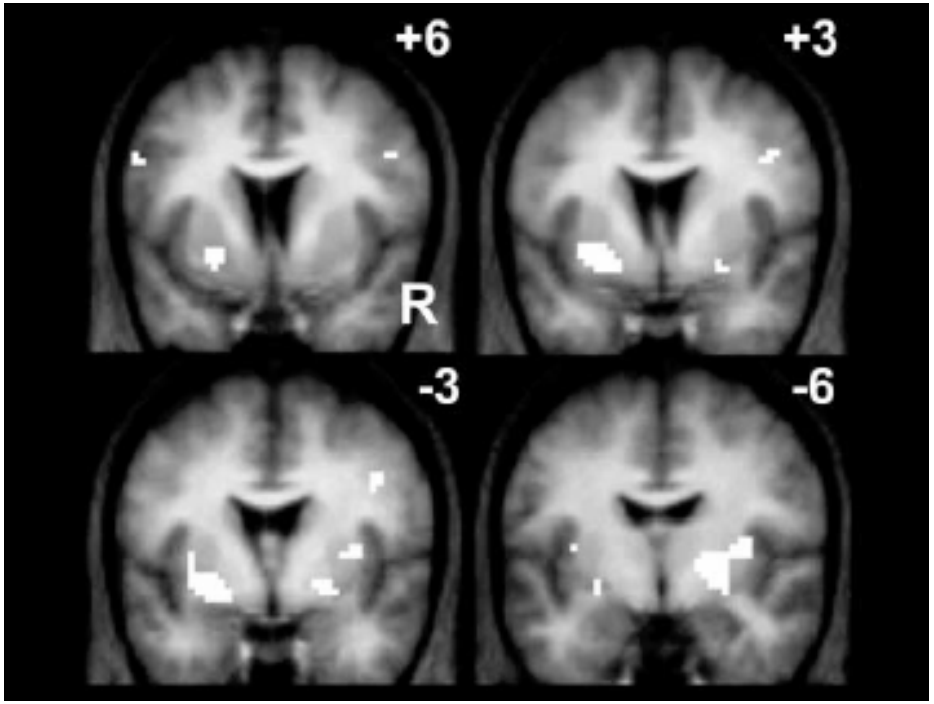
juice
unexpected - expected

Berns et al 2001

FMRI

BOLD signal in striatum correlates parametrically, trial-by-trial with prediction error

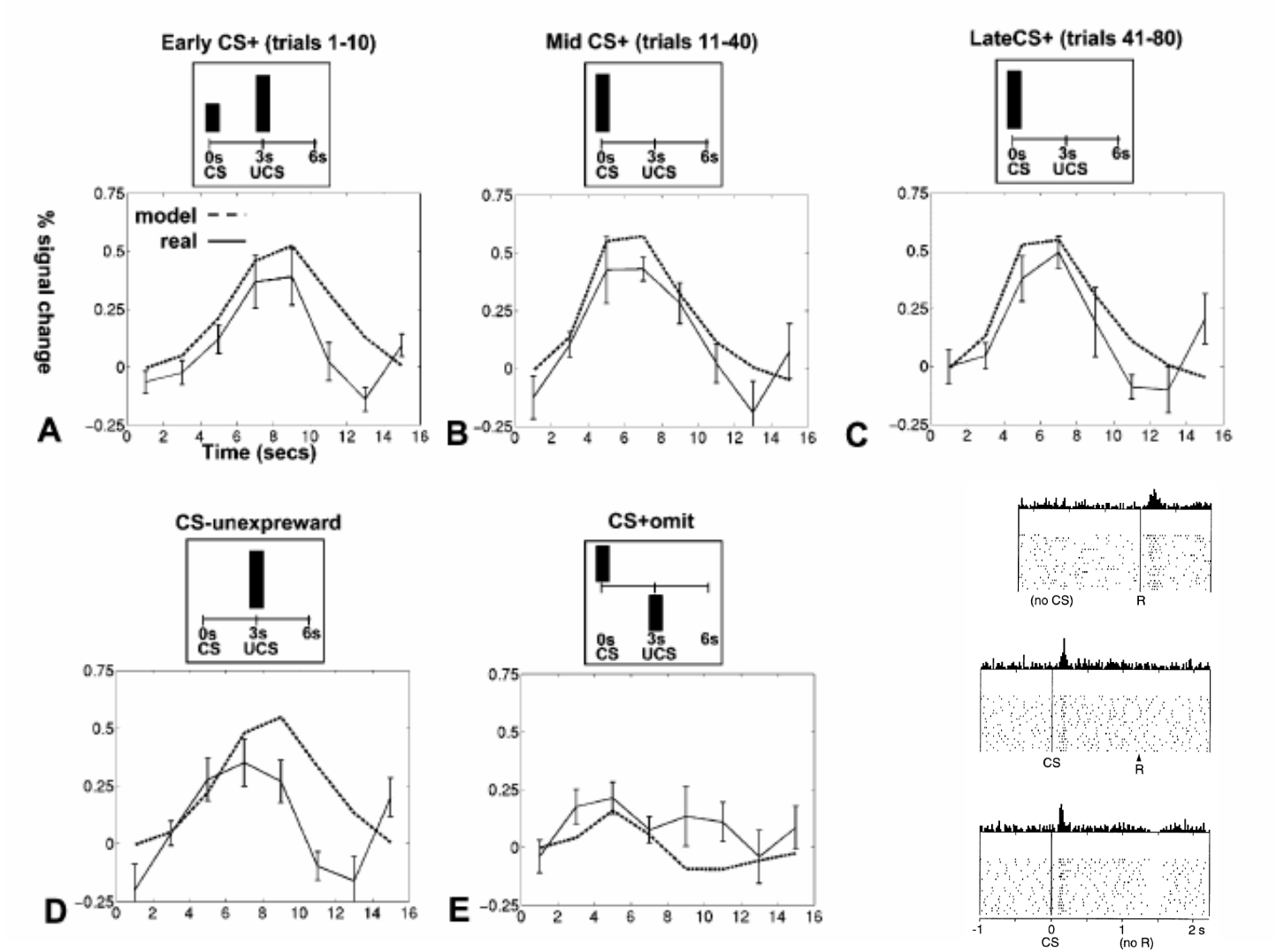




(O'Doherty et al 2003)

+ this signal modulated up & down by dopaminergic drugs (Pessiglione et al 2006)

FMRI



(O'Doherty et al 2003)

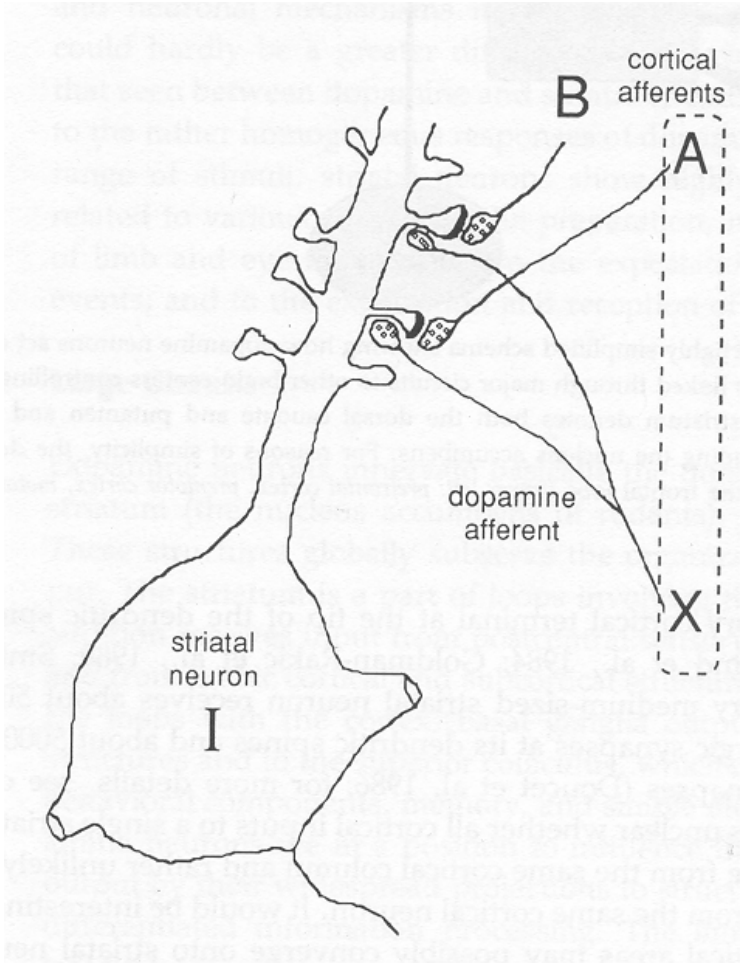
prediction error

- what should prediction error do?
  - drive learning
  - …about expected rewards
  - …to guide decision-making

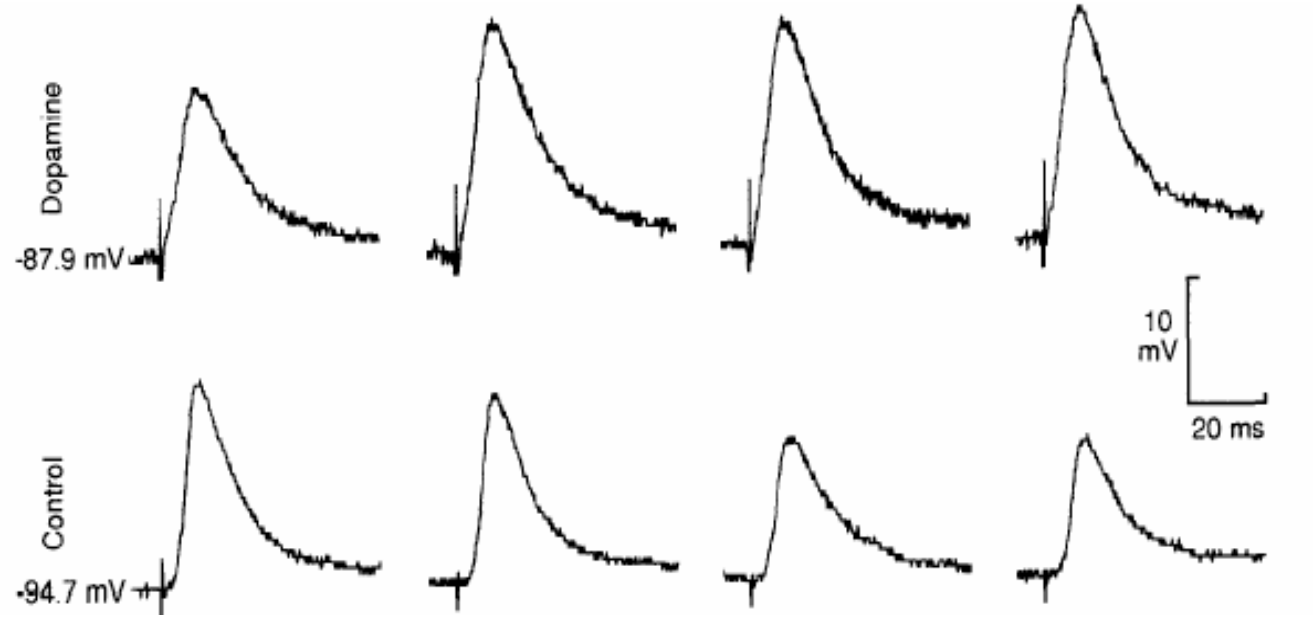
→ this fits well with the multifarious roles of dopamine & its targets

Dopamine and plasticity

- If dopamine carries a prediction error, where does learning happen?
- Potentially, the cortico-striatal synapse



DA and corticostriatal plasticity



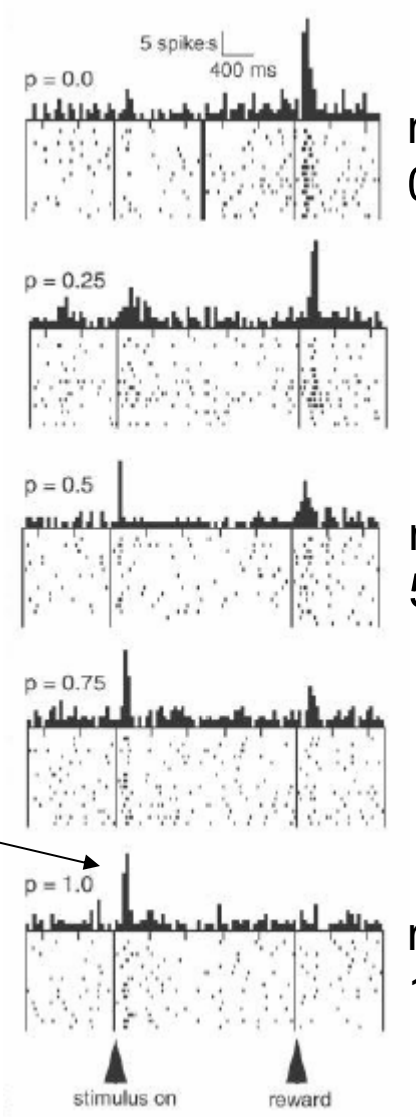
Wickens et al. 1996

Three-factor learning rule? (pre/post/dopamine)

$w_{i,t+1} = w_{i,t} + \varepsilon\delta_t$

More dopamine responses

what about these responses?



reward following 0% predictive cue

reward following 50% predictive cue

reward following 100% predictive cue

(Fiorillo et al 2003)

temporal-difference learning

Rescorla-Wagner:

Want $V_n = r_n$ ← (here *n* indexes trials, treated as units)

Use prediction error $\delta_n = r_n - V_n$

Temporal-difference learning (Sutton & Barto):

Want $V_t = r_t + r_{t+1} + r_{t+2} + r_{t+3} + \ldots$ ← (here *t* indexes time within trial)

$= r_t + V_{t+1}$ ← (clever recursive trick)

Use prediction error $\delta_t = [r_t + V_{t+1}] - V_t$

temporal difference learning

Temporal-difference learning (Sutton & Barto):

Want $V_t = r_t + r_{t+1} + r_{t+2} + r_{t+3} + \ldots$

$= r_t + V_{t+1}$

Use prediction error $\delta_t = [r_t + V_{t+1}] - V_t$

- learn to predict cumulative future rewards $r_t + r_{t+1} + \ldots$
- learn using what I predict at time $t+1$ (V_{t+1}) as stand in for all future rewards
  - so I don't have to wait forever to learn
- learn consistent predictions based on temporal difference $V_{t+1} - V_t$
  - if $V_{t+1} = V_t$, my predictions are consistent
  - if $V_{t+1} > V_t$, things got unexpectedly better
  - if $V_{t+1} < V_t$, things got unexpectedly worse

→ and these act like reward to generate prediction error and learning

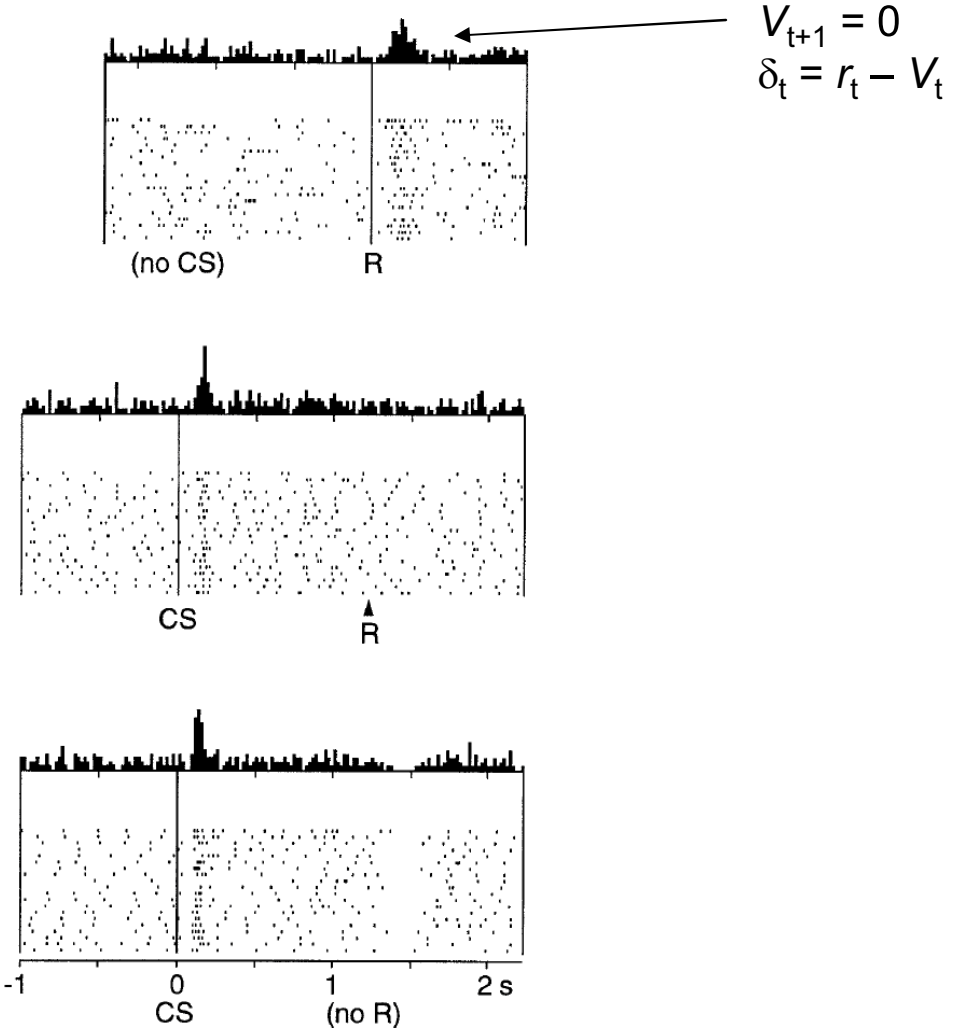
(example on board)

Second order conditioning

	Phase 1:	**Phase 2:**	**Test:**
Second order:	A → R	B → A	A? resp
			B? resp

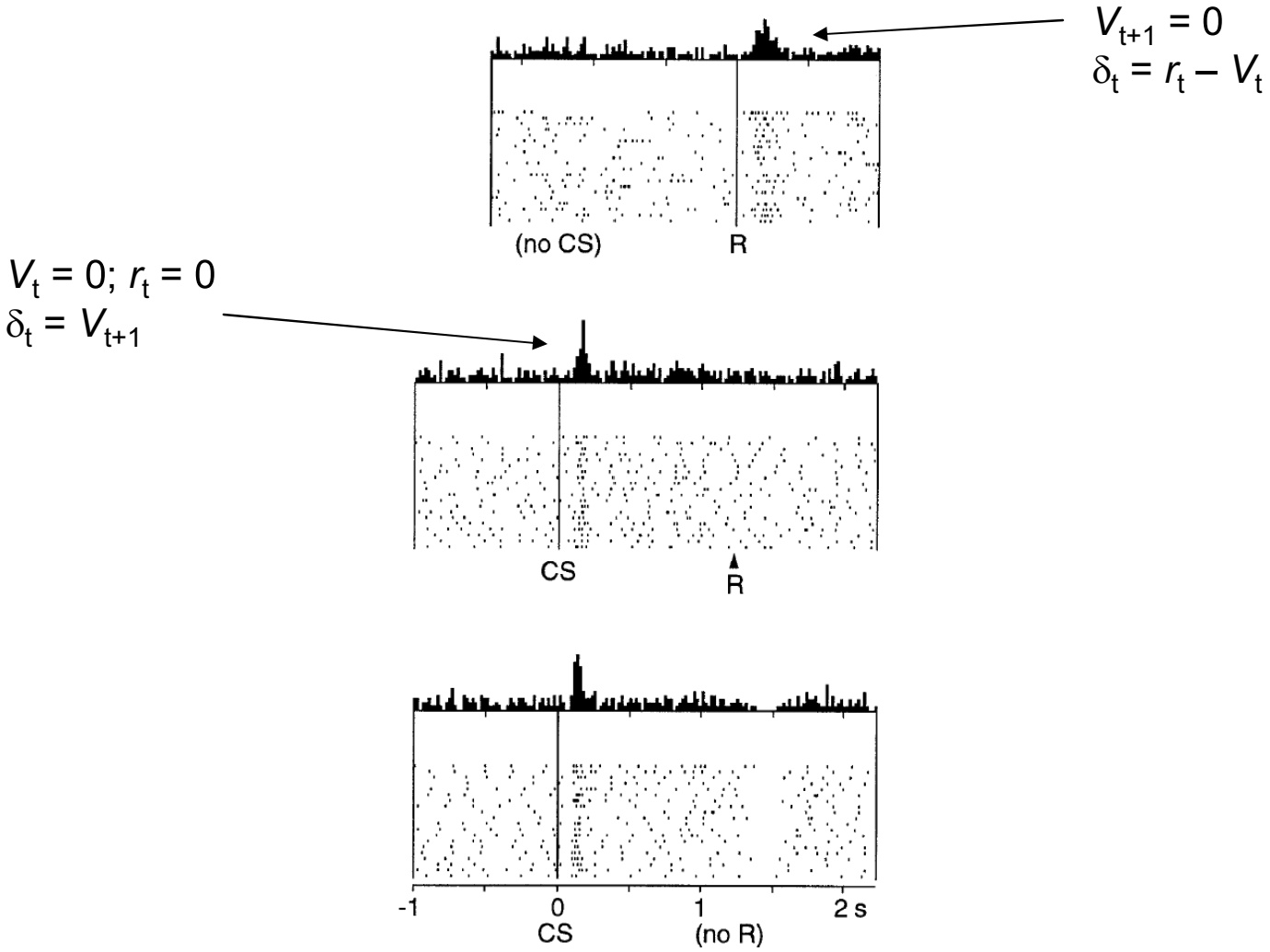
- B associated with reward even though never directly paired
- Rescorla/Wagner say B should be nothing, or negative (r_t always zero when B arrives)
- Temporal-difference learning explains this, if B precedes A
  - Positive prediction error when *A* appears
  - ie V_{t+1}-V_t positive, trains w_B
  - on board

Typical dopamine responses



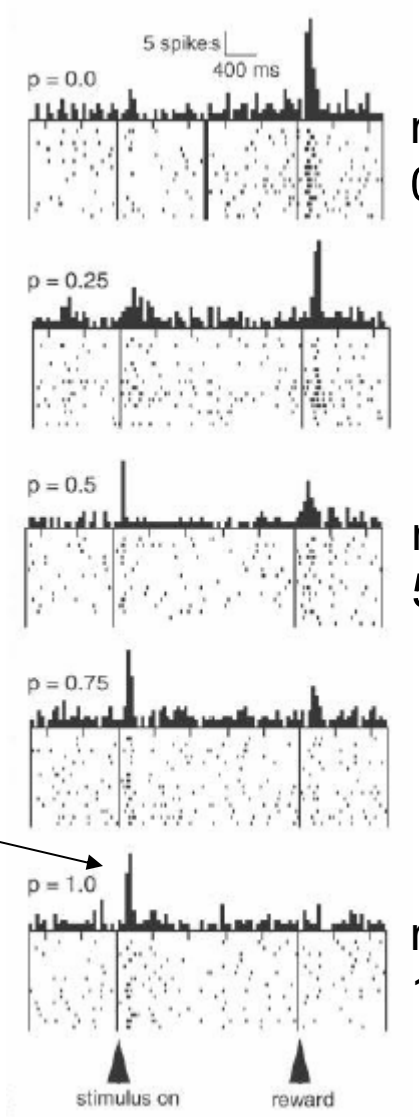
(Schultz et al. 1997)

Typical dopamine responses



(Schultz et al. 1997)

More dopamine responses



reward following
0% predictive cue

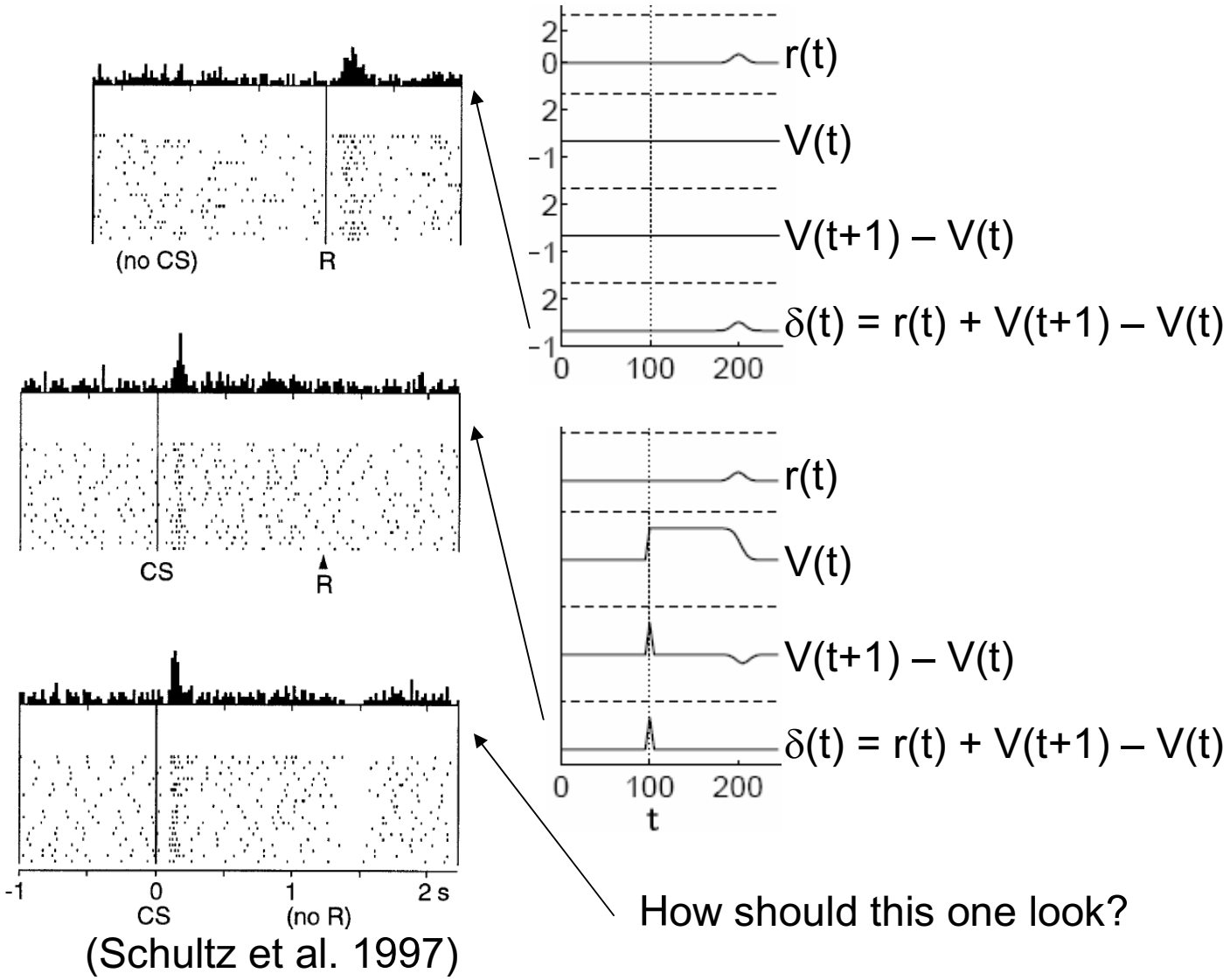
reward following
50% predictive cue

reward following
100% predictive cue

same story here

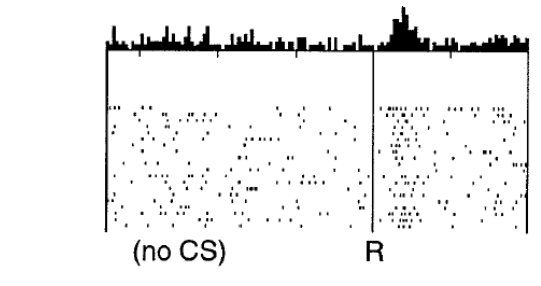
(Fiorillo et al 2003)

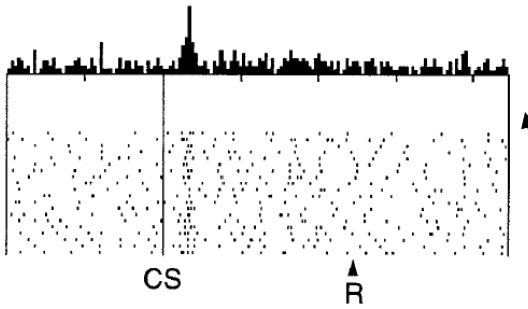
Dopamine responses interpreted

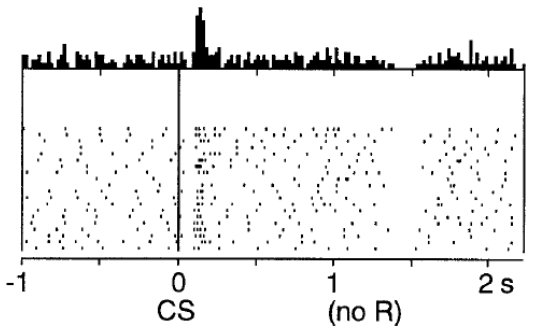


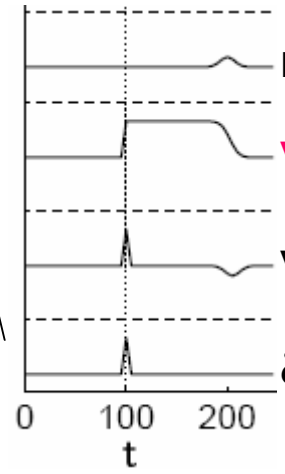
(Schultz et al. 1997)

Dopamine responses interpreted









r(t)

V(t)

V(t+1) – V(t)

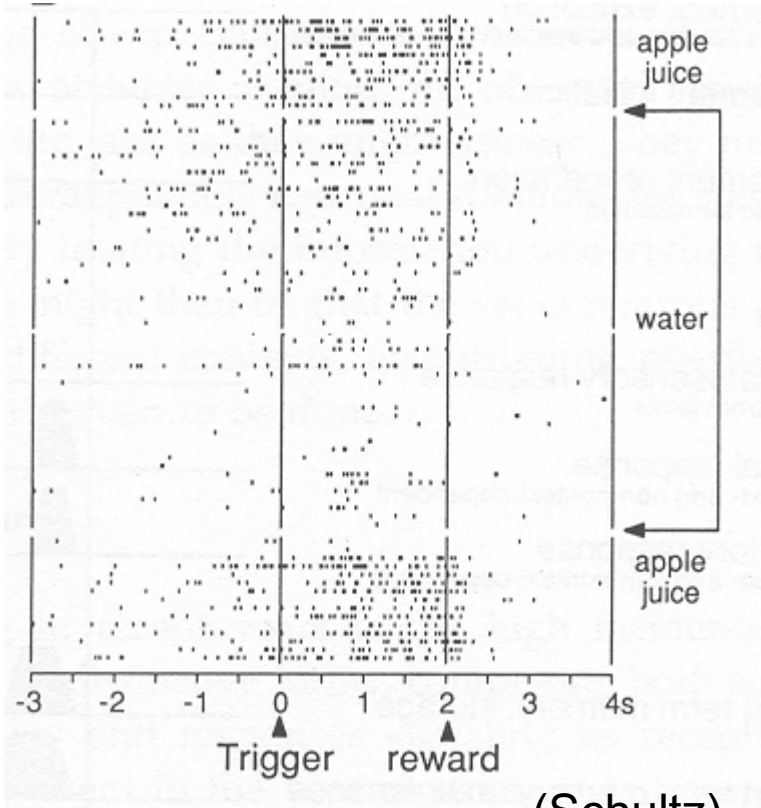
$\delta(t) = r(t) + V(t+1) - V(t)$

This is what is really supposed to be learned here. Where is it?

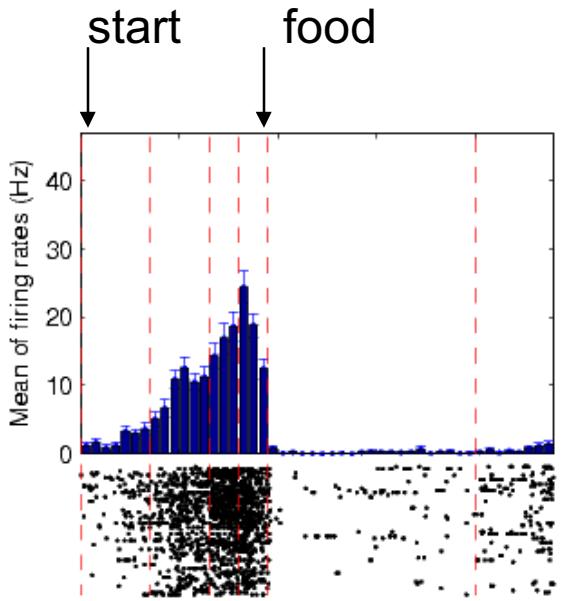
(Schultz et al. 1997)

striatum & value

- striatal neurons do show ramping activity preceding reward, which changes with learning



(Schultz)



Summary

- dopamine neurons report error in reward prediction
  - seen also in human BOLD
  - drives plasticity at striatal synapses
  - would be useful for learned decision-making

- full response explained by temporal-difference learning
  - Generalization of Rescorla-Wagner
  - learns to predict cumulative future reward
  - changes in future reward expectancy drive learning
  - this explains anticipatory dopaminergic responding, second order conditioning

- big implications for decision-making: sequential decision problems involving many future rewards