Neuroimaging of associative learning

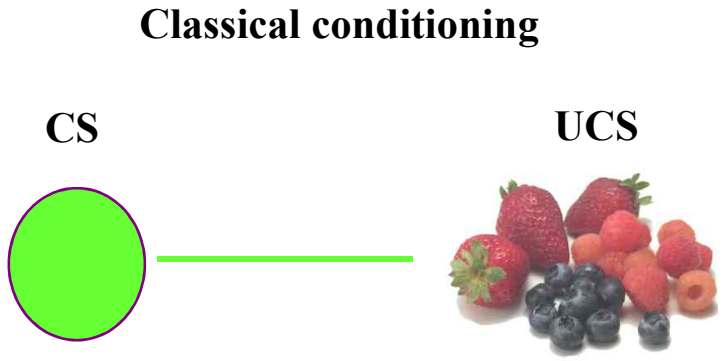
John O'Doherty
Functional Imaging Lab
Wellcome Department of Imaging Neuroscience
Institute of Neurology
Queen Square,
London

Collaborators on this project:

Peter Dayan
Ray Dolan
Karl Friston
Hugo Critchley
Ralf Deichmann

Acknowledgements to: Eric Featherstone, Peter Aston

Neuroimaging of associative learning



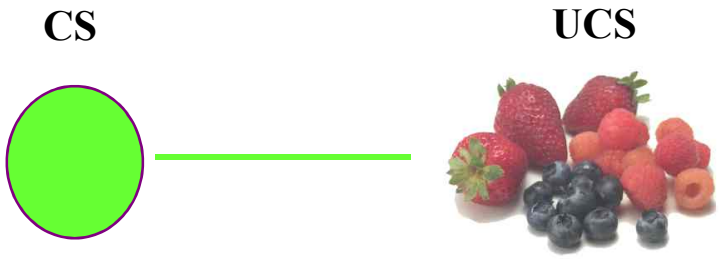
1/ *How* Pavlovian value predictions are learned

2/ *Where* this is implemented in the human brain

3/ Extend approach to instrumental conditioning

Neuroimaging of associative learning

How value predictions are learned



•**Learning is mediated by a prediction error (Rescorla and Wagner, 1972; Pearce & Hall, 1980)**

$$\delta = (r - v)$$

where r = reward received on a given trial (UCS)
v = expected reward - value of CS stimulus

Neuroimaging of associative learning

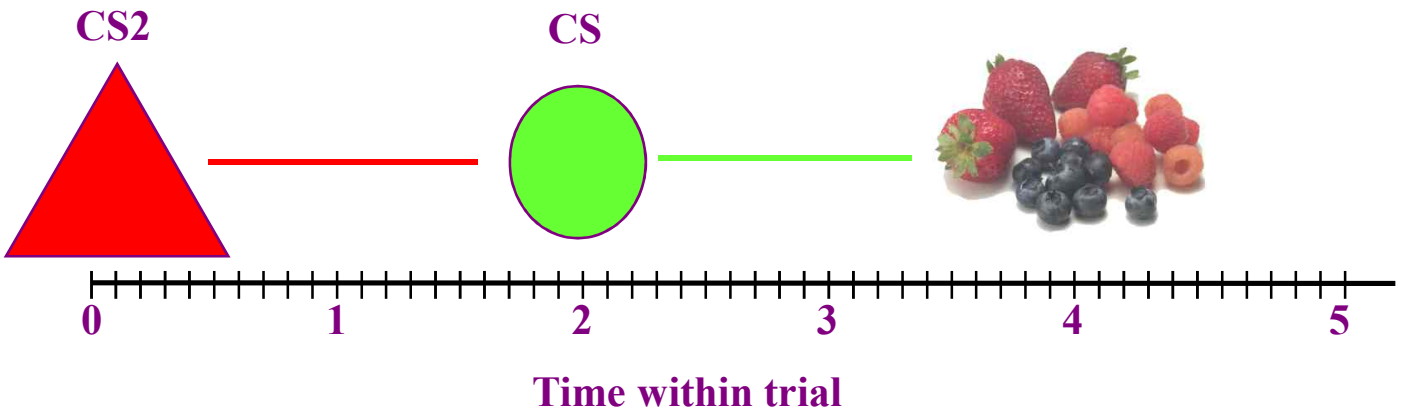
•**Temporal difference learning (TD learning):**

Differs from previous trial based theories -

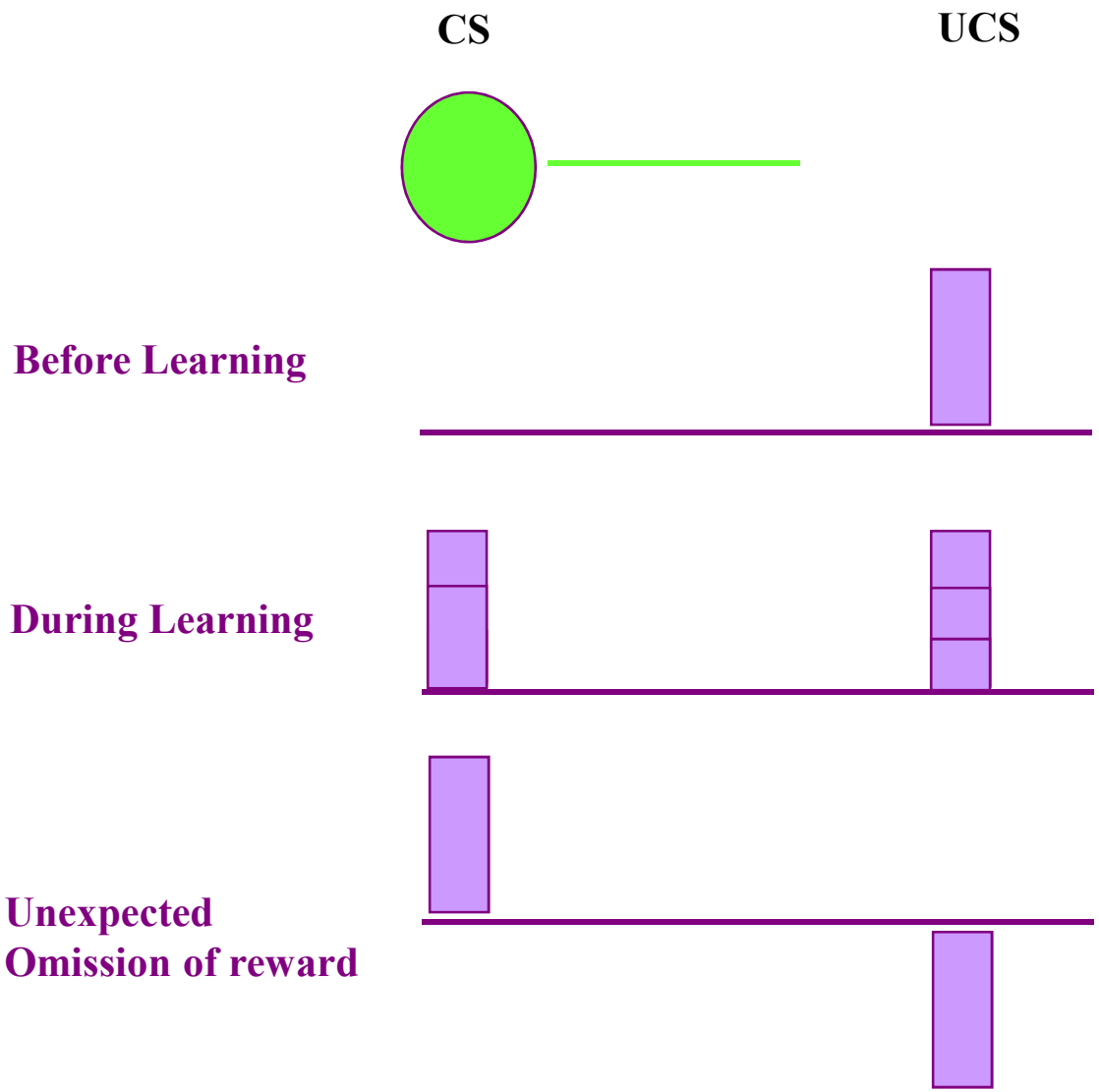
Predictions are learned about the total future reward available within a trial for each time t in which a CS is presented

(Sutton and Barto, 1989;Montague et al., 1996; Schultz, Dayan and Montague,1997)

$$\delta(t) = r(t) + \gamma \hat{V}(t+1) - \hat{V}(t)$$



Temporal difference learning model



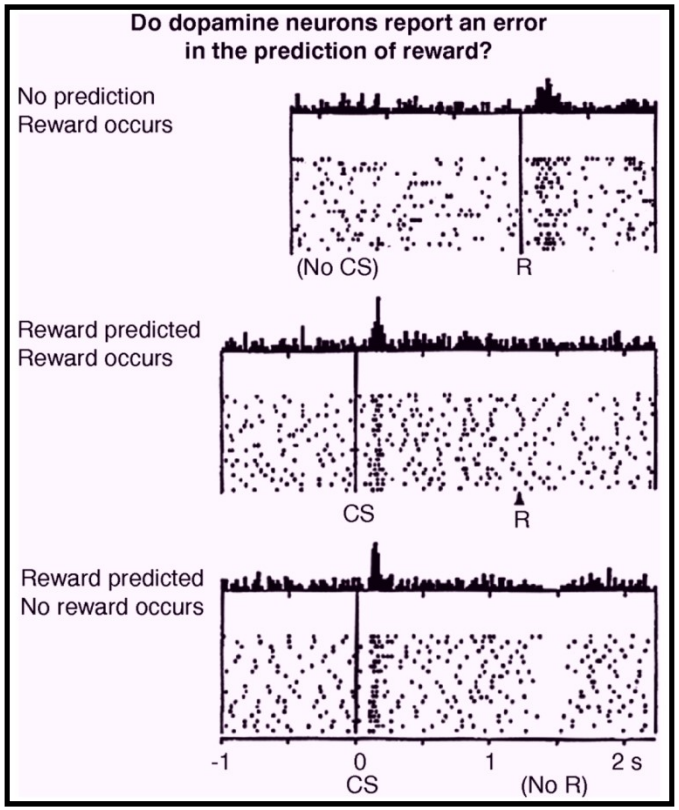
Reward Learning:
Pavlovian Appetitive Conditioning

Single unit recordings from dopamine neurons revealed that these neurons produce responses consistent with TD - learning (Schultz, 1998):

(1) Transferring their responses from the time of the presentation of the reward to the time of the presentation of the CS during learning.

(2) Decreasing firing from baseline at the time the reward was expected following an omission of expected reward.

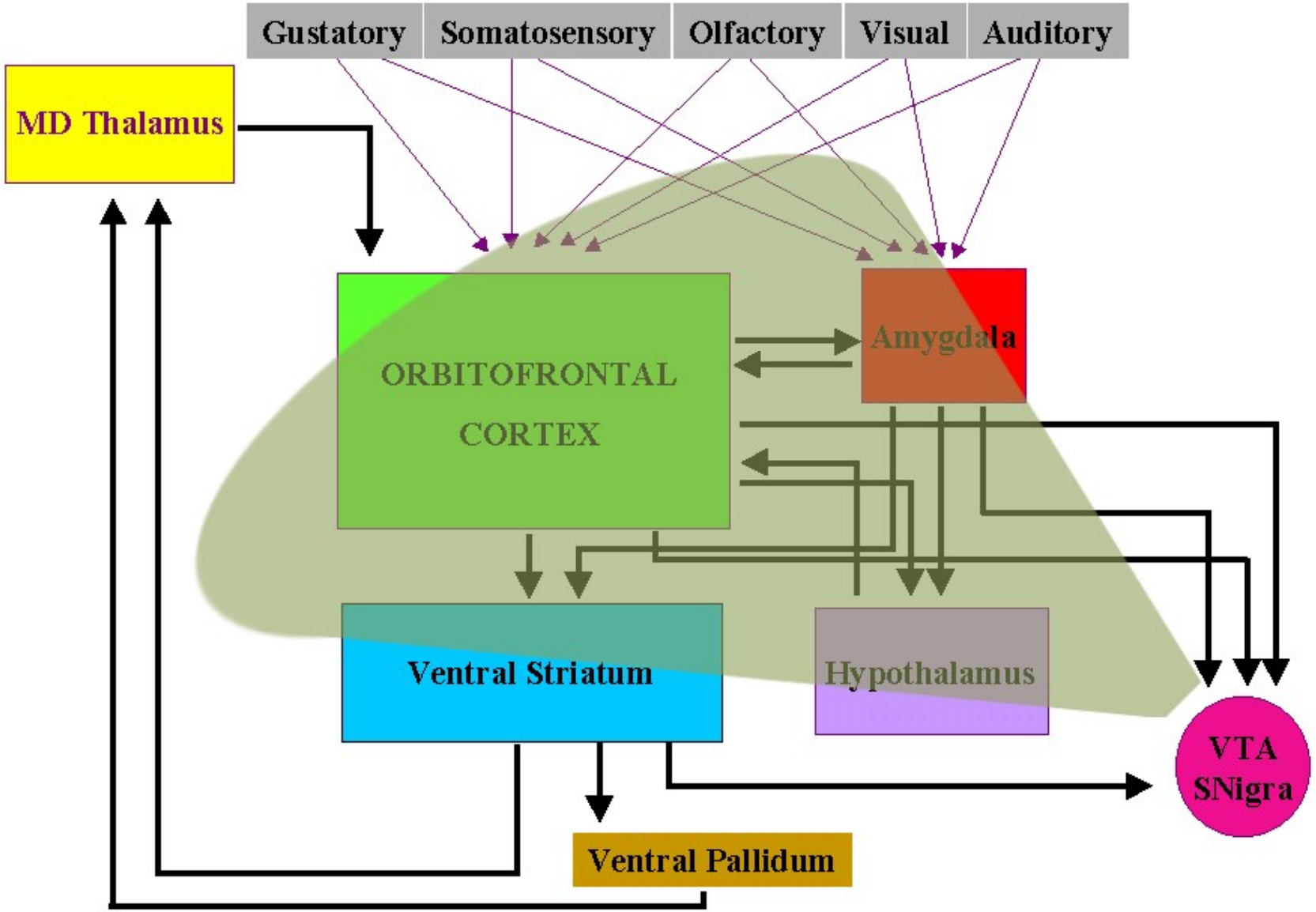
(3) Responding at the time of the reward following the unexpected delivery of reward



From Schultz, Montague and Dayan, 1997

Can we find evidence of a temporal difference prediction error in the human brain during appetitive conditioning?

Reward circuits in the brain



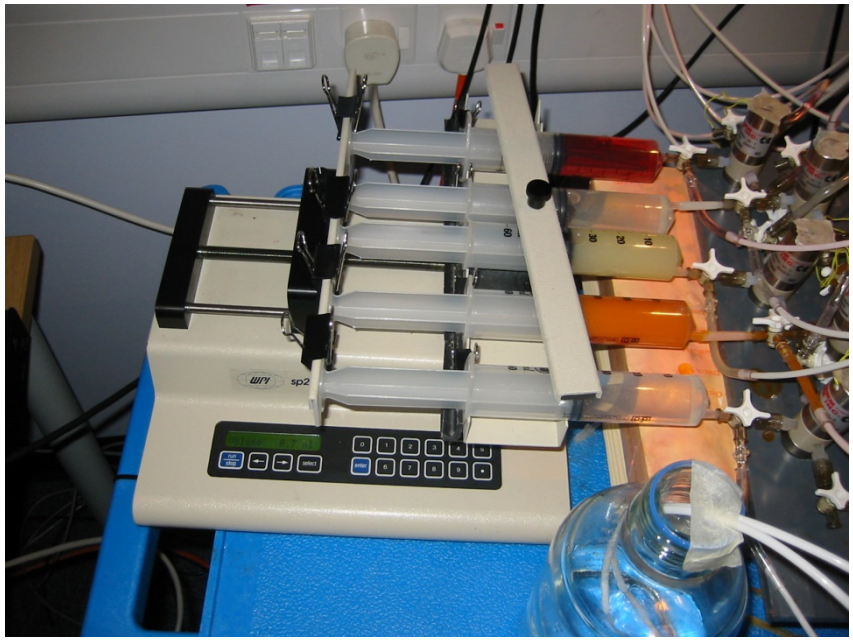
Experimental Set Up

•Scanning conducted at 2 Tesla (Siemens)

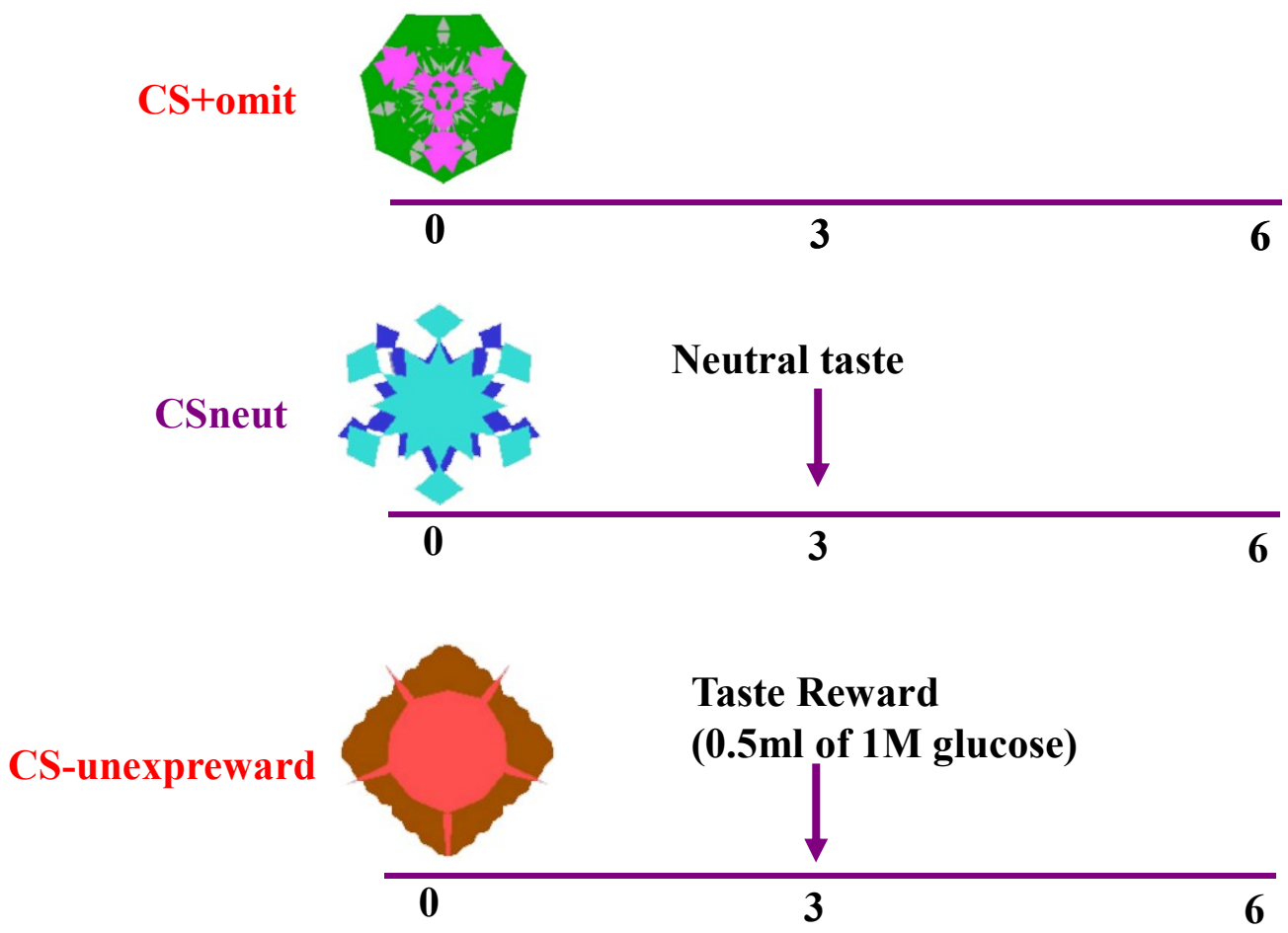
•Taste delivered using an electronic syringe pump
•positioned outside the scanner room

•On-line measurement of pupillary responses

•13 subjects participated (9 found taste pleasant at end of scanning)



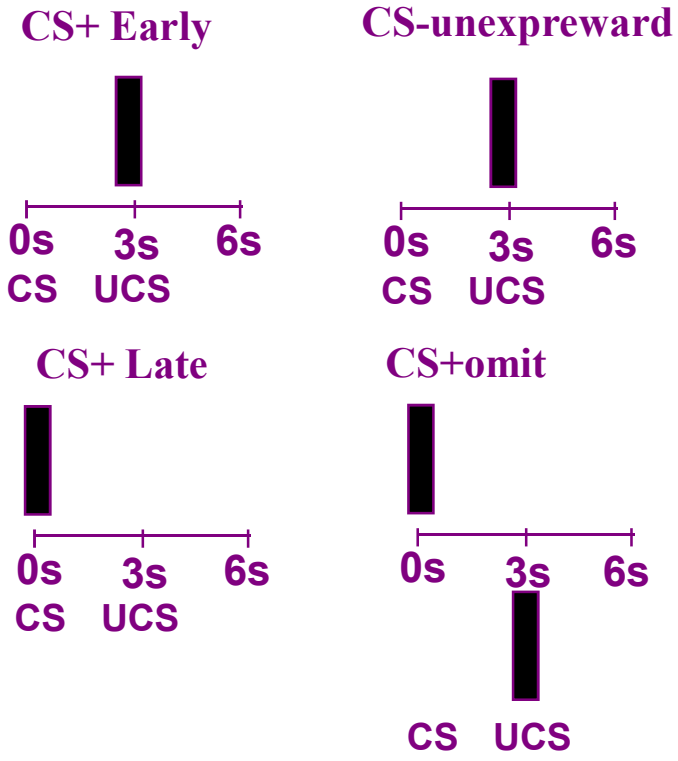
Experimental Design

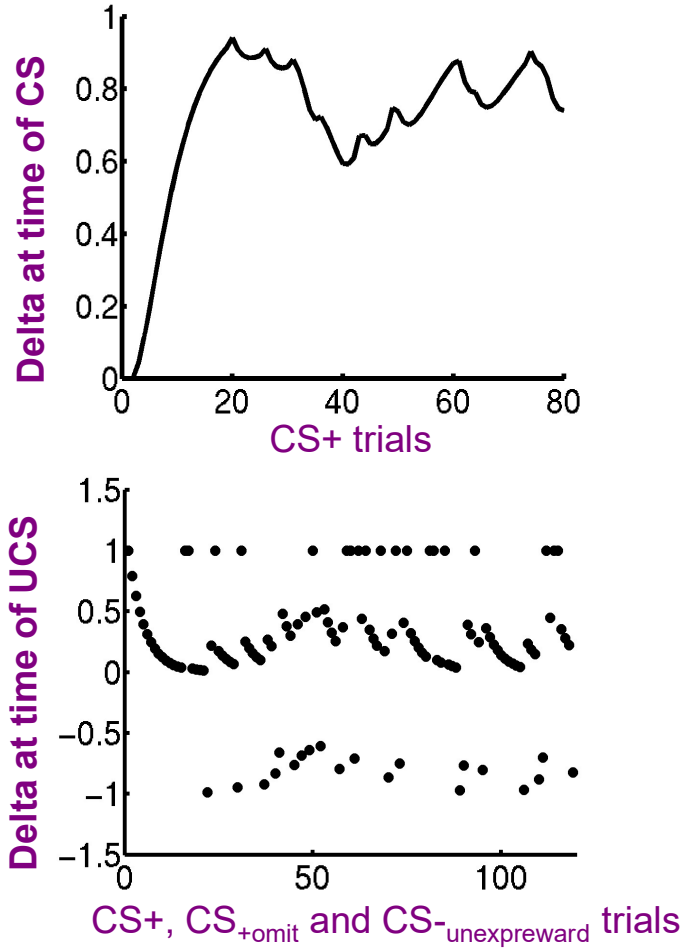


Ratio of regular to 'surprise' trials 4:1

Statistical Analysis

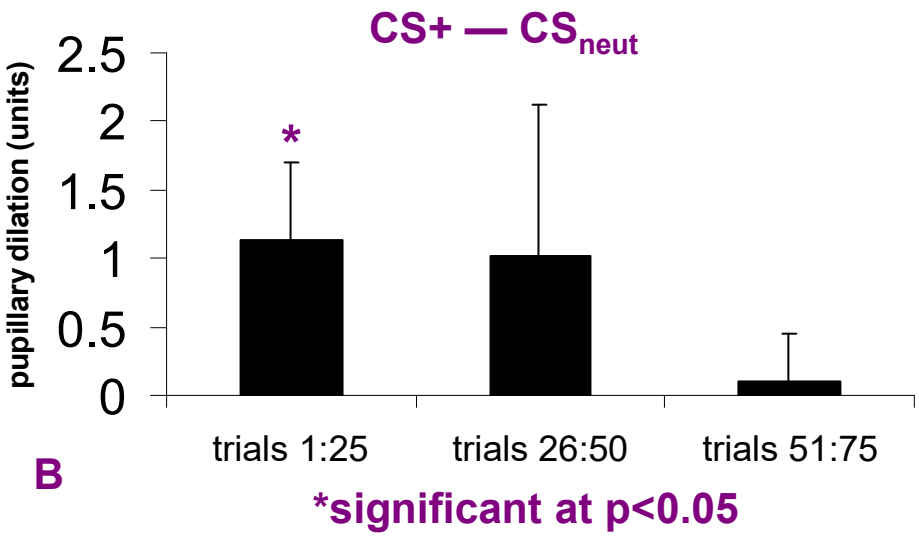
TD related δ responses across the experiment





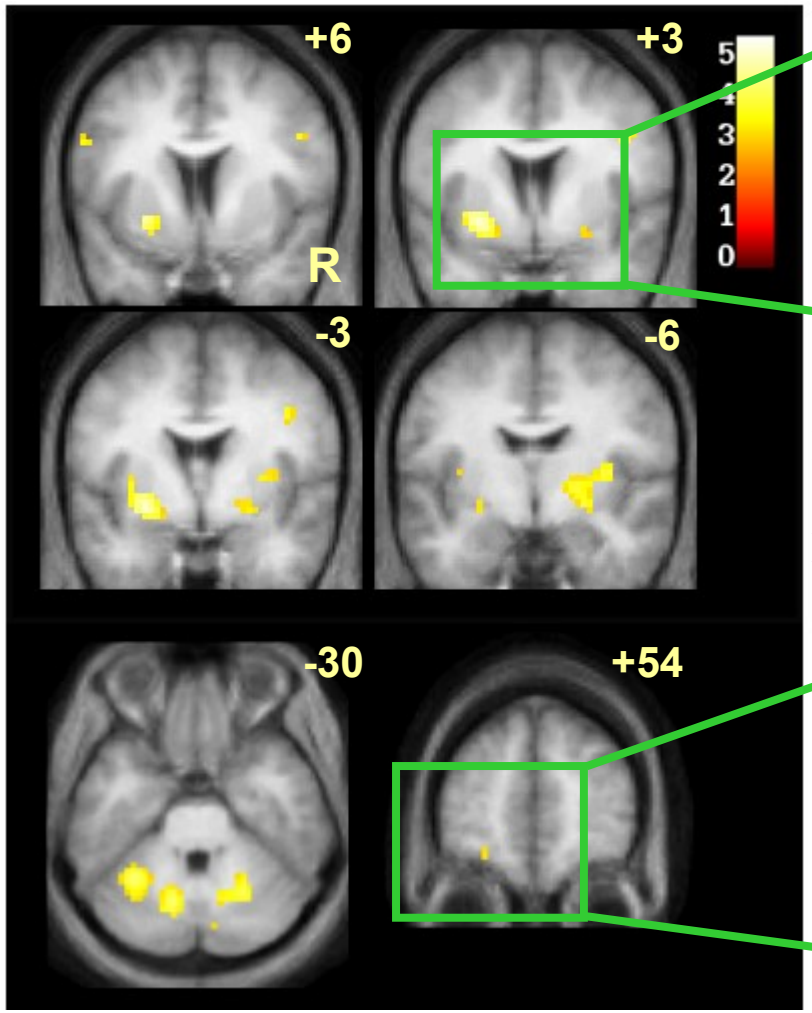
Results

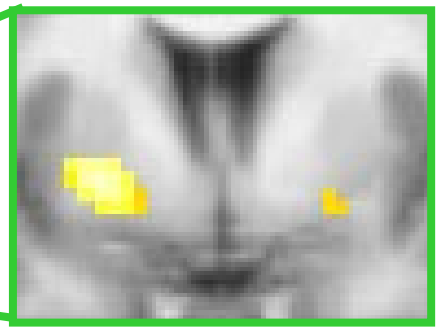
Discriminatory pupillary responses

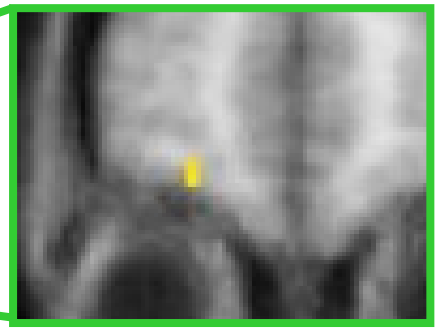


Results

Areas showing signed TD-related PE responses





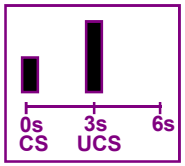


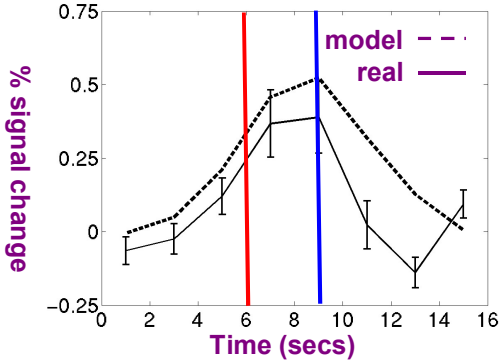
Random effects p<<0.001
p<0.05 corrected for small volume (ventral striatum)
Different learning rates fit some regions better than others (α=0.2 vs. α=0.7)

From O'Doherty, Dayan, Friston, Critchley and Dolan. **Neuron,** *2003*

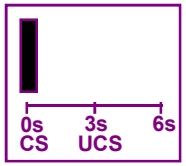
Results

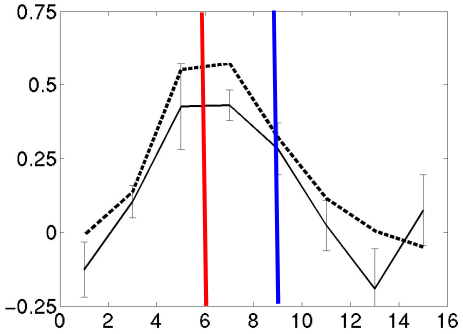
Early CS+ (trials 1-10)



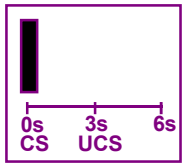


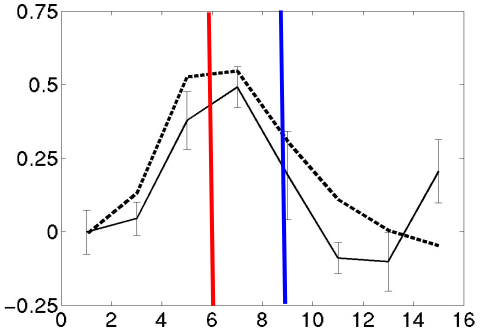
Mid CS+ (trials 11-40)



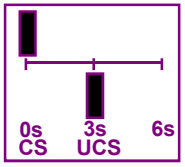


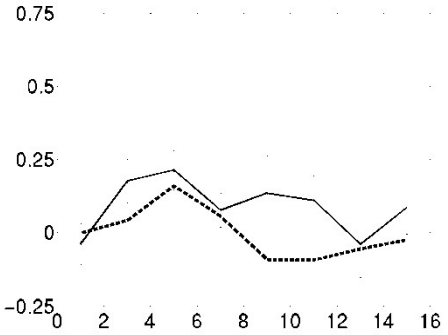
LateCS+ (trials 41-80)



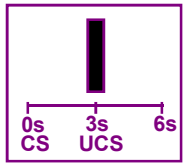


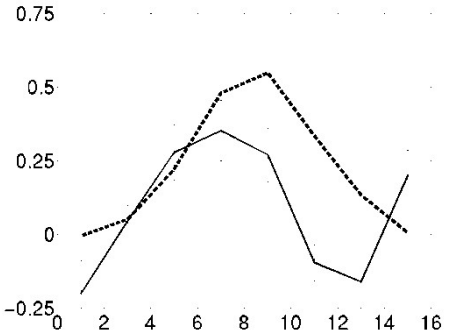
CS+omit





CS-unexpreward



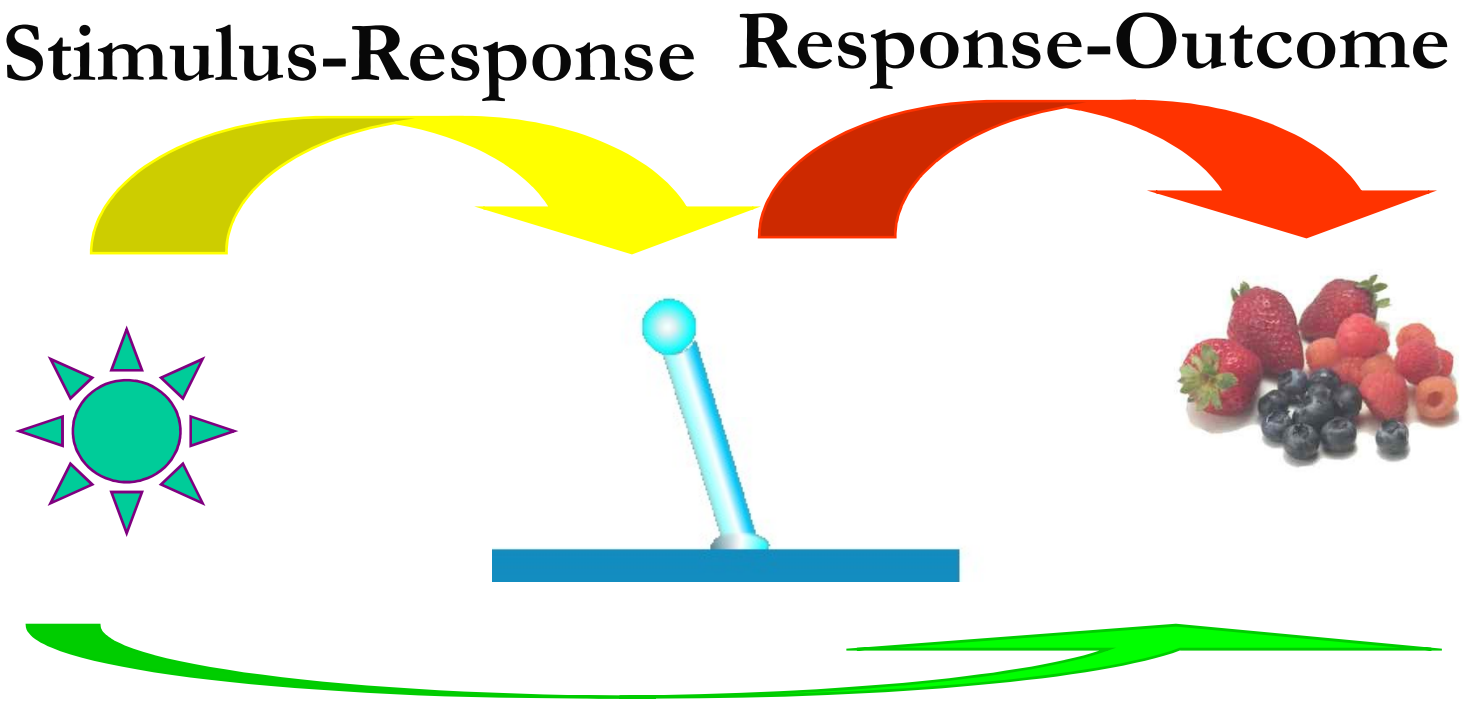


Interim Conclusions (1)

•Responses in a part of human ventral striatum and orbitofrontal cortex can be described by a theoretical learning model: temporal difference learning.

•On the basis of evidence from non-human primates, it is likely that a source of TD-learning related activity in these regions is the modulatory influence exerted by the phasic responses of dopamine neurons.

Reward Learning:
Instrumental Conditioning



Model of instrumental conditioning: Actor Critic

Actor-critic TD(0) learning

Follow policy π, in state u, take action a, with probability P[a',u] (a function of the value of that action) moving from state u to u'.

Critic:

v=wu

$\delta = r_a(u)+v(u')-v(u)$

where v(u)= value of state (averaged over all possible actions).

Update weights: w(u)=w(u)+εδ
Where ε = learning rate.

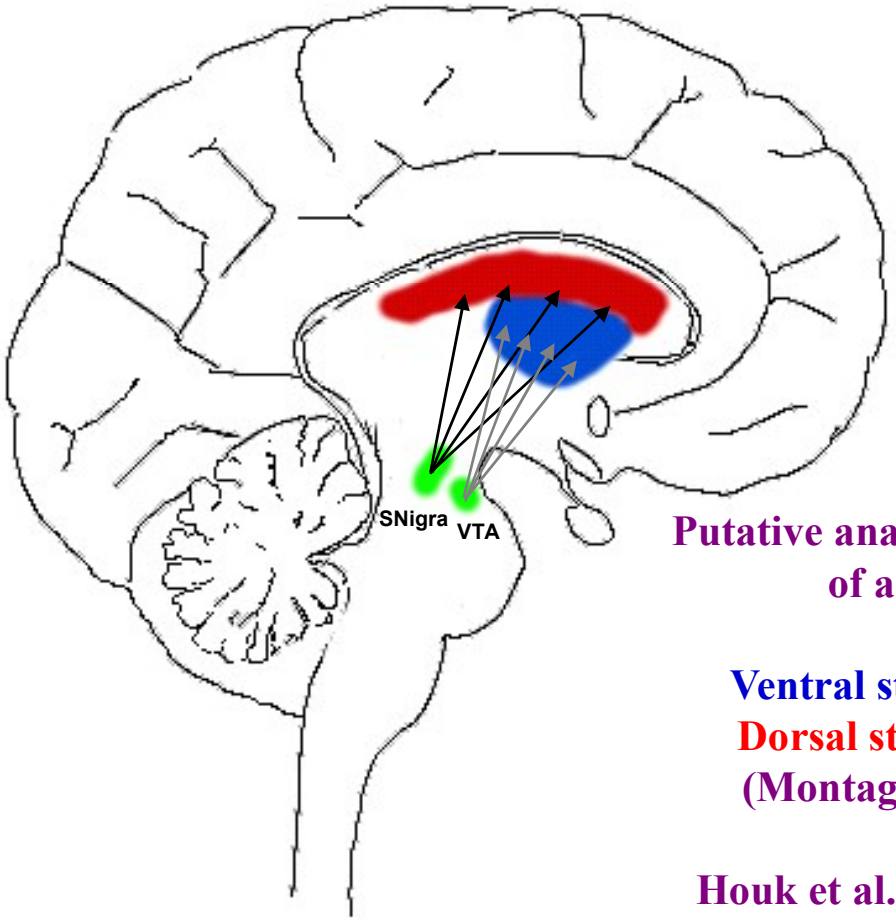
Actor:

update policy π, by changing value of state-action pairs (Q) for action a, as well as the value of all other actions a'

$Q(a',u) = Q(a',u)+ \varepsilon(\delta aa' – P[a',u])\delta$

P[a';u] = probabilty of taking action a' in state u

Dorsal vs Ventral Striatum



Putative anatomical substrates of actor-critic

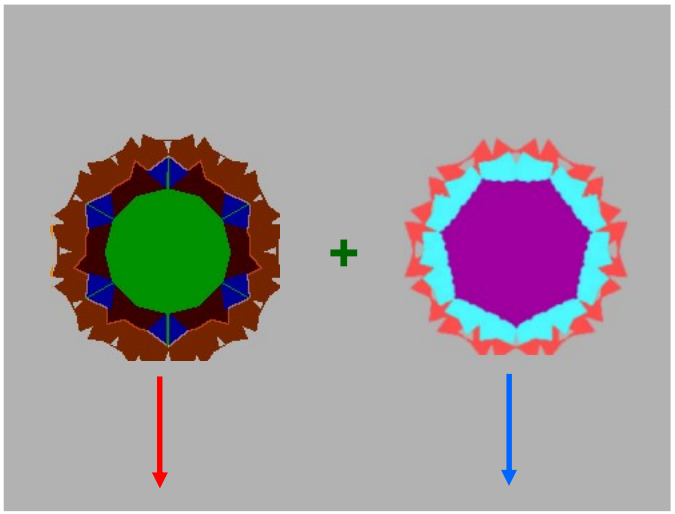
Ventral striatum = critic
Dorsal striatum = actor
(Montague et al., 1996)

Houk et al., (1995) – suggest matrix/striosome distinction

Experimental Design

Two trial types: 80 trials of each

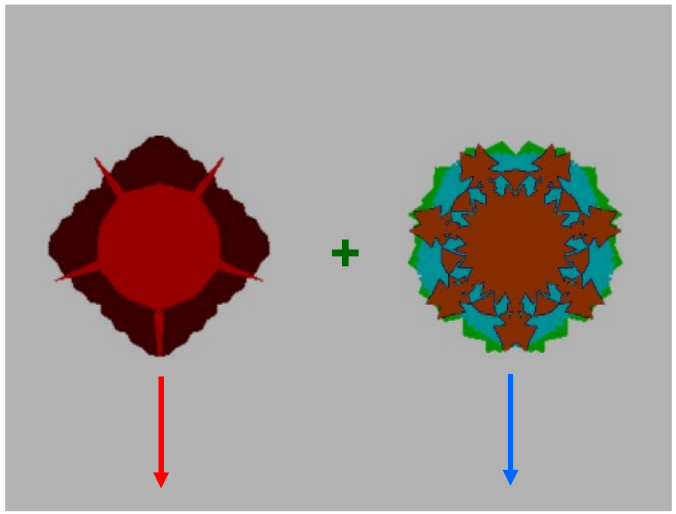
high valence



60% probability receive fruit juice

30% probability receive fruit juice

low valence



60% probability receive neutral taste

30% probability receive neutral taste

Experimental Design

•The design is split into two ‘sessions’:
Pavlovian and Instrumental (each ~15 minutes in duration).

•Order of presentation of sessions counterbalanced across subjects

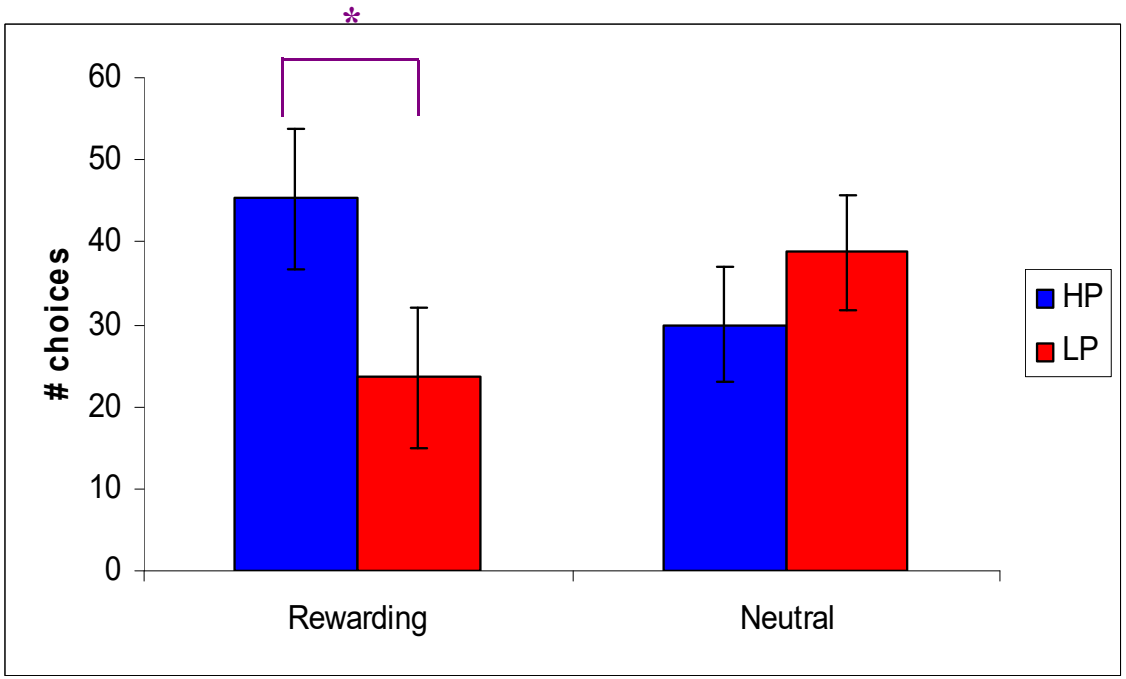
•Used two different fruit juices as the reward: peach juice and blackcurrant juice.

•To control for habituation in the pleasantness of the juices over the course of the experiment a different juice was used in the Pavlovian and Instrumental tasks for each subject. Again this was counterbalanced across subjects.

•Instrumental responses from one subjects were used as a ‘yoke’ to the Pavlovian contingencies from another subject.

Behavioral results

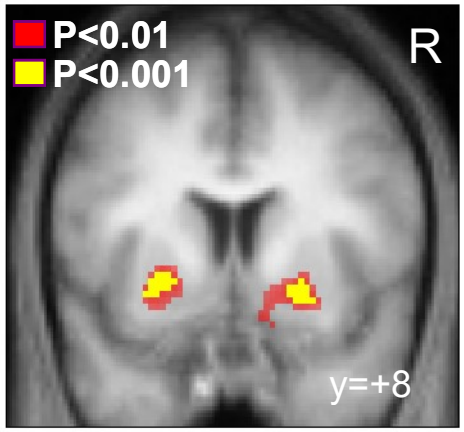
Instrumental choices



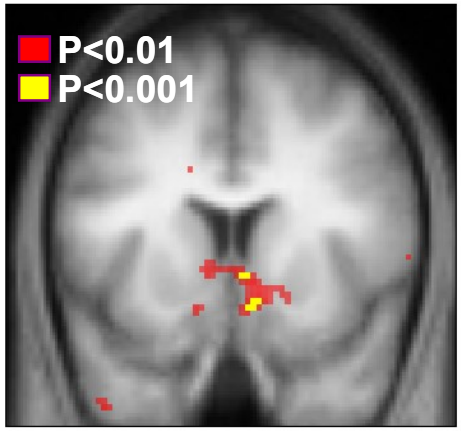
*** Significant at p<0.05**

Results: Ventral Striatum

PAVLOVIAN CONDITIONING

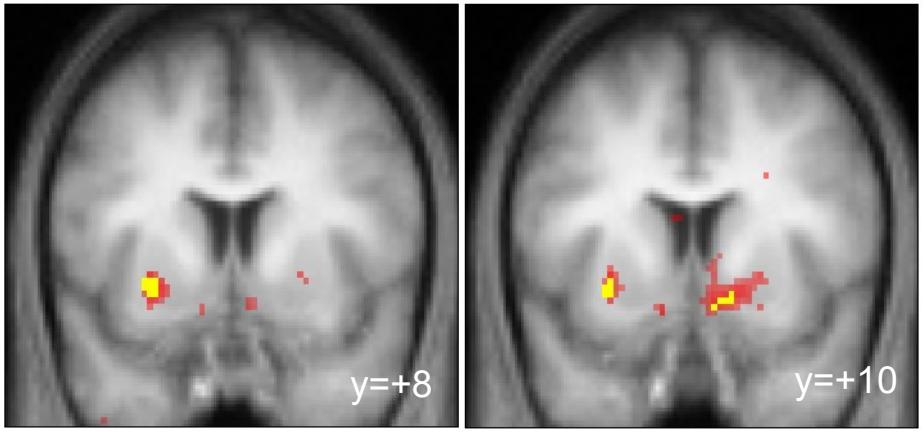


INSTRUMENTAL CONDITIONING



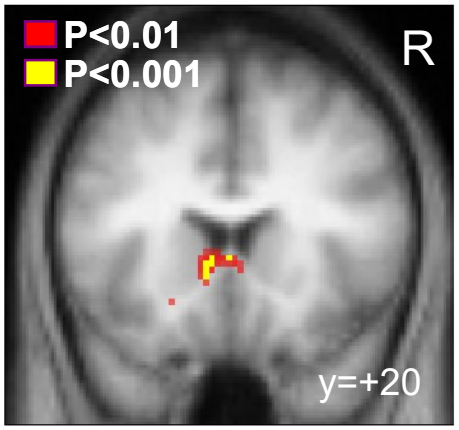
Results: Ventral Striatum

CONJUNCTION OF INSTRUMENTAL AND PAVLOVIAN CONDITIONING

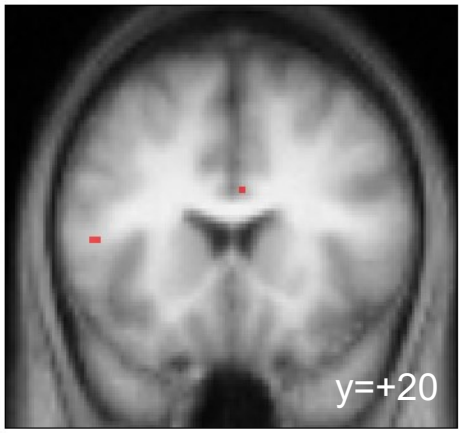


Results: Dorsal Striatum

INSTRUMENTAL CONDITIONING

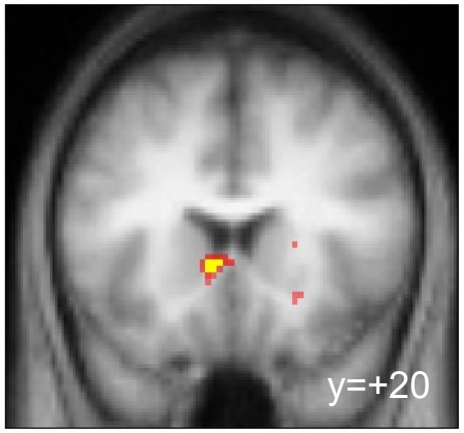


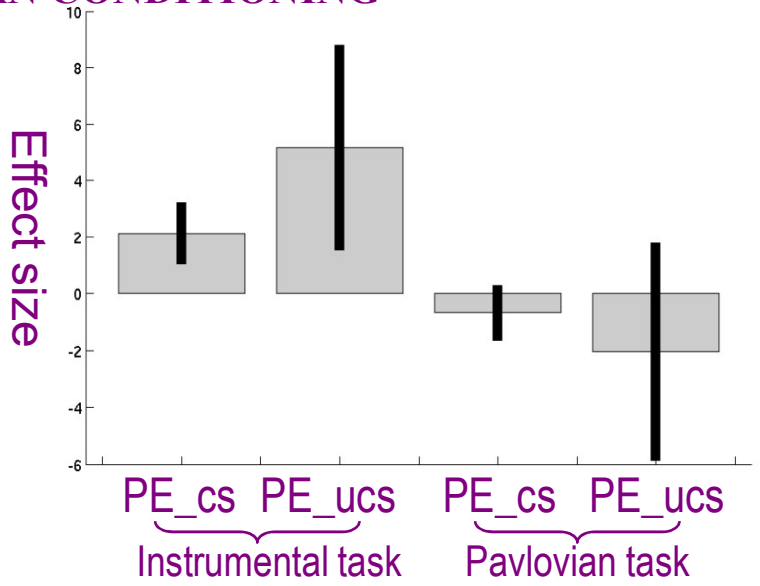
PAVLOVIAN CONDITIONING



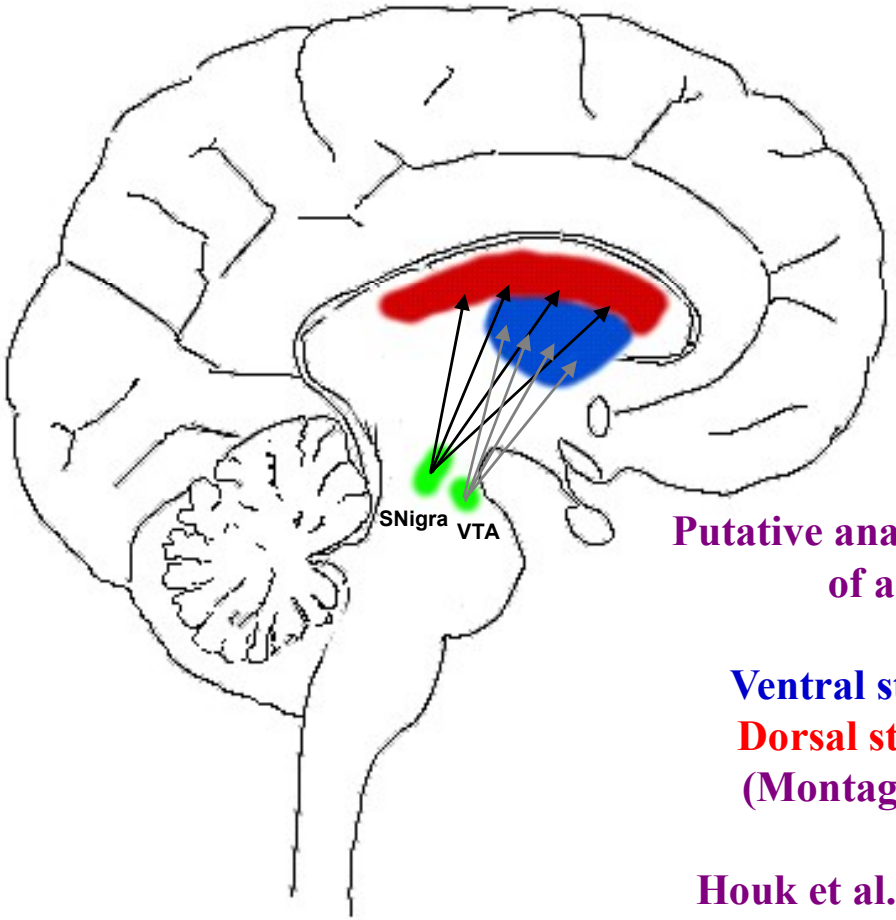
Results: Dorsal Striatum

INSTRUMENTAL – PAVLOVIAN CONDITIONING





Dorsal vs Ventral Striatum



Putative anatomical substrates of actor-critic

Ventral striatum = critic
Dorsal striatum = actor
(Montague et al., 1996)

Houk et al., (1995) – suggest matrix/striosome distinction

Conclusions (1)

- A temporal difference prediction error signal is present in a part of the human brain (ventral striatum) during appetitive conditioning.

- A putative neural substrate of the reward-related prediction error signal in the striatum is the phasic activity of afferent dopamine neurons.

- TD-related response is present in ventral striatum during Instrumental as well as Pavlovian conditioning

- Dorsal striatum has significantly enhanced responses during Instrumental relative to Pavlovian Conditioning

Conclusions (2)

- Suggests that actor-critic like process is implemented in human striatum:

  - -Ventral striatum may correspond to the critic: involved in forming predictions of future reward

  - -Dorsal striatum may correspond to the instrumental actor: may mediate stimulus-response learning

- More generally, demonstrates application of event-related fMRI to test constrained computational models of human brain function.

Neuroimaging of associative learning

John O'Doherty
Functional Imaging Lab
Wellcome Department of Imaging Neuroscience
Institute of Neurology
Queen Square,
London

Collaborators on this project:

Peter Dayan
Ray Dolan
Karl Friston
Hugo Critchley
Ralf Deichmann

Acknowledgements to: Eric Featherstone, Peter Aston