On neural correlates of reinforcement learning

the role of dopamine in planning and action

Genela Morris

Dept. of Neurobiology and Ethology

Haifa University

Overview

- Learning through reward – reinforcement learning
- Midbrain dopamine neurons and temporal difference (TD) learning
- ACh in the striatum
- Dopamine neurons and their impact on decision behaviour
- Implications for computational models of action selection

Reinforcement learning
the basics

Supervised learning –

all knowing teacher, detailed feedback

Reinforcement learning –

scalar (correct/incorrect) feedback

Unsupervised learning –

self organization

Reinforcement learning: The law of effect

"The Law of Effect is that: Of several responses made to the same situation, those which are accompanied or closely followed by satisfaction to the animal will, other things being equal, be more firmly connected with the situation, so that, when it recurs, they will be more likely to recur"

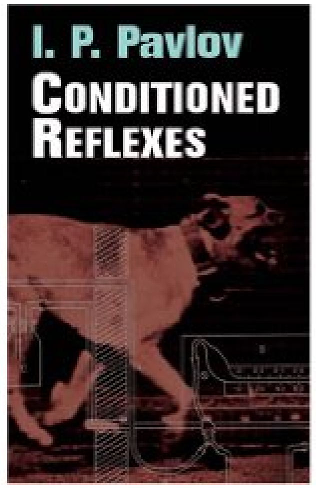
Edward Lee Thorndike (1911)

Early attempts at modeling

By associative rules

Classical conditioning

Classical conditioning



The Elements:

US: Unconditioned stimulus

UR: Unconditioned response

NS: Neutral stimulus

CS: Conditioned stimulus

- **CS1**: Conditioned stimulus 1
- **CS2**: Conditioned stimulus 2

CR: Conditioned response

Properties of classical conditioning

(Pavlov 1927)

- **Acquisition.**
- **Partial Reinforcement**
- **Generalization.**
- **Interstimulus Interval (ISI) effects.**
- **Intertrial Interval (ITI) effects.**

So far…

- A simple association (coincidence, Hebbian) model can explain the phenomenon.

  – But…

Properties of classical conditioning

(Cnt'd)

CS must RELIABLY predict US

- **Conditioned Inhibition**
- **Relative validity** (Wagner 1968).
- **Blocking** (Kamin 1968)
- …

Which simple association can't explain

Learning occurs not because two events co-occur, but because that co-occurrence is UNPREDICTED

Rescorla-Wagner rule (1972)

Learning to predict reward R given stimulus U=1

Goal: Form a prediction of the reward V of the form:

$V = \omega U$

And learn to change ω :

$\Delta \omega = \varepsilon (R - V) U$

After learning of consistent pairing: $\omega = R$

Where:

U=CS availability (0,1);

V=reward prediction:

R=reward availability (0,1) :

ω = weight of the connection between U and V

ε = learning rate

R-V = prediction error

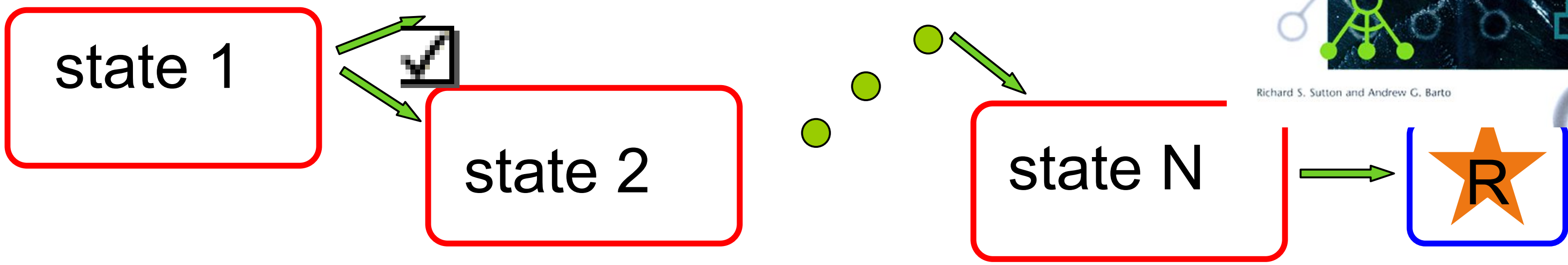
Blocking with Rescorla Wagner

- Given U1, U2 and R, after U1 has been learnt:
- $\omega 1 = R$
- $V = \underbrace{\omega 1 U1}_{R} + \underbrace{\omega 2 U2}_{0}$
- Prediction error: $R - V = 0$

  And no learning occurs for $\omega 2$

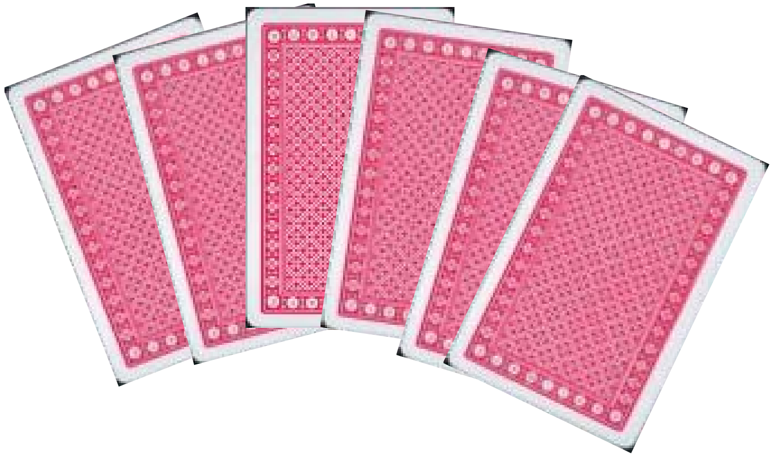
Critical problems in reinforcement learning (and in Rescorla-Wagner)

1. Temporal credit assignment



Critical problems, for control

2. Exploration/exploitation



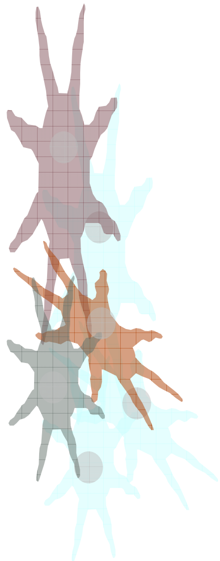
TD learning - solution for temporal credit assignment

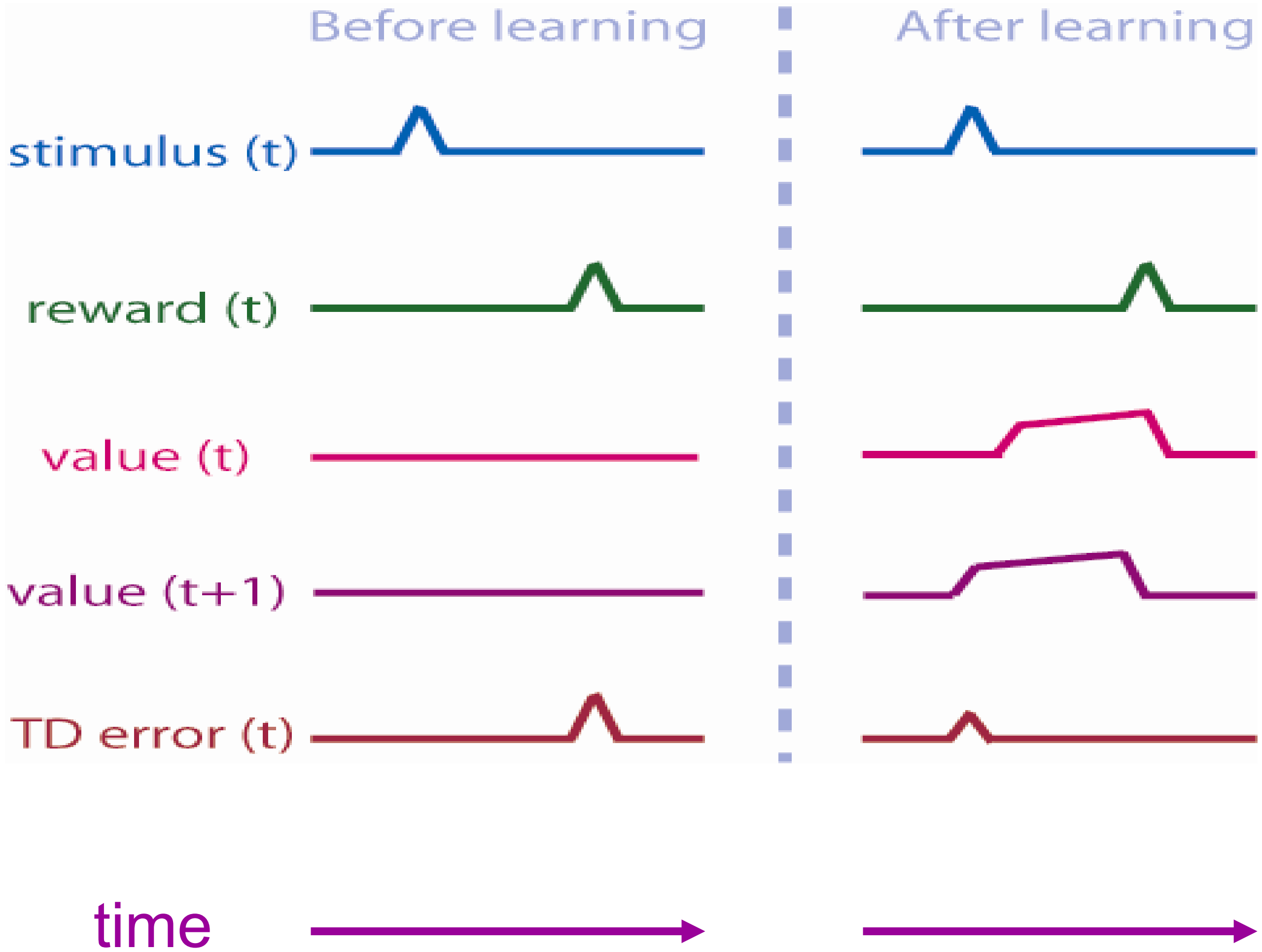
1. Estimate value of current state $(V_t = r_t + \gamma' r_{t+1} + \cdots)$: (discounted) sum of expected rewards
2. Measure 'truer' value of current state: reward at present state + estimated value of next state $(r_t + \gamma V_{t+1})$
3. TD error $\delta_t = r_t + \gamma V_{t+1} - V_t$
4. Use TD error to improve 1 $(V_t^{k+1} = V_t^k + \eta\, \delta_t)$

where: $V_{t\,=\,value}$ *of the state reached at time t in iteration k*

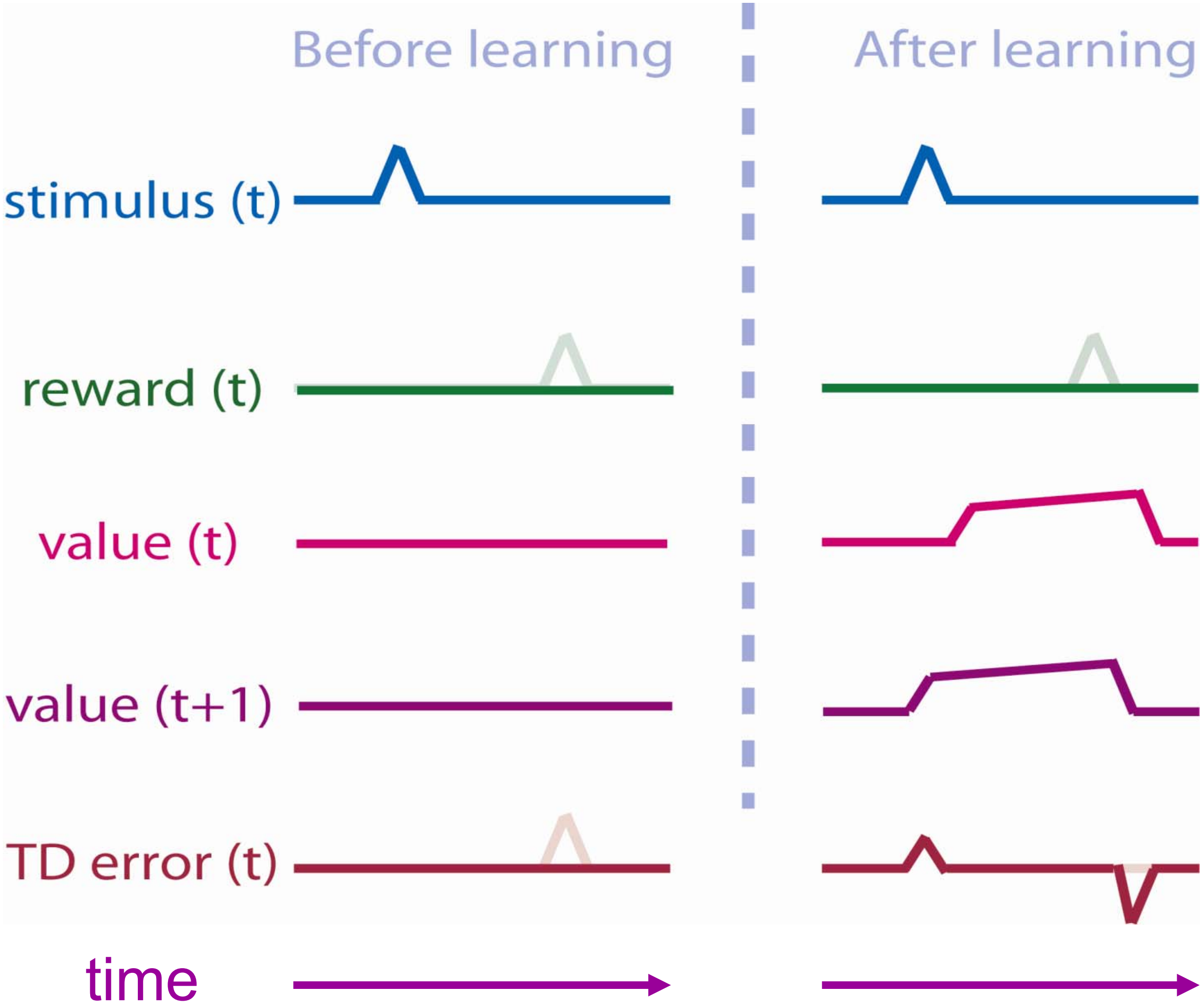
r_t *= reward given at time t;* η *= learning rate,* δ *= prediction error*

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

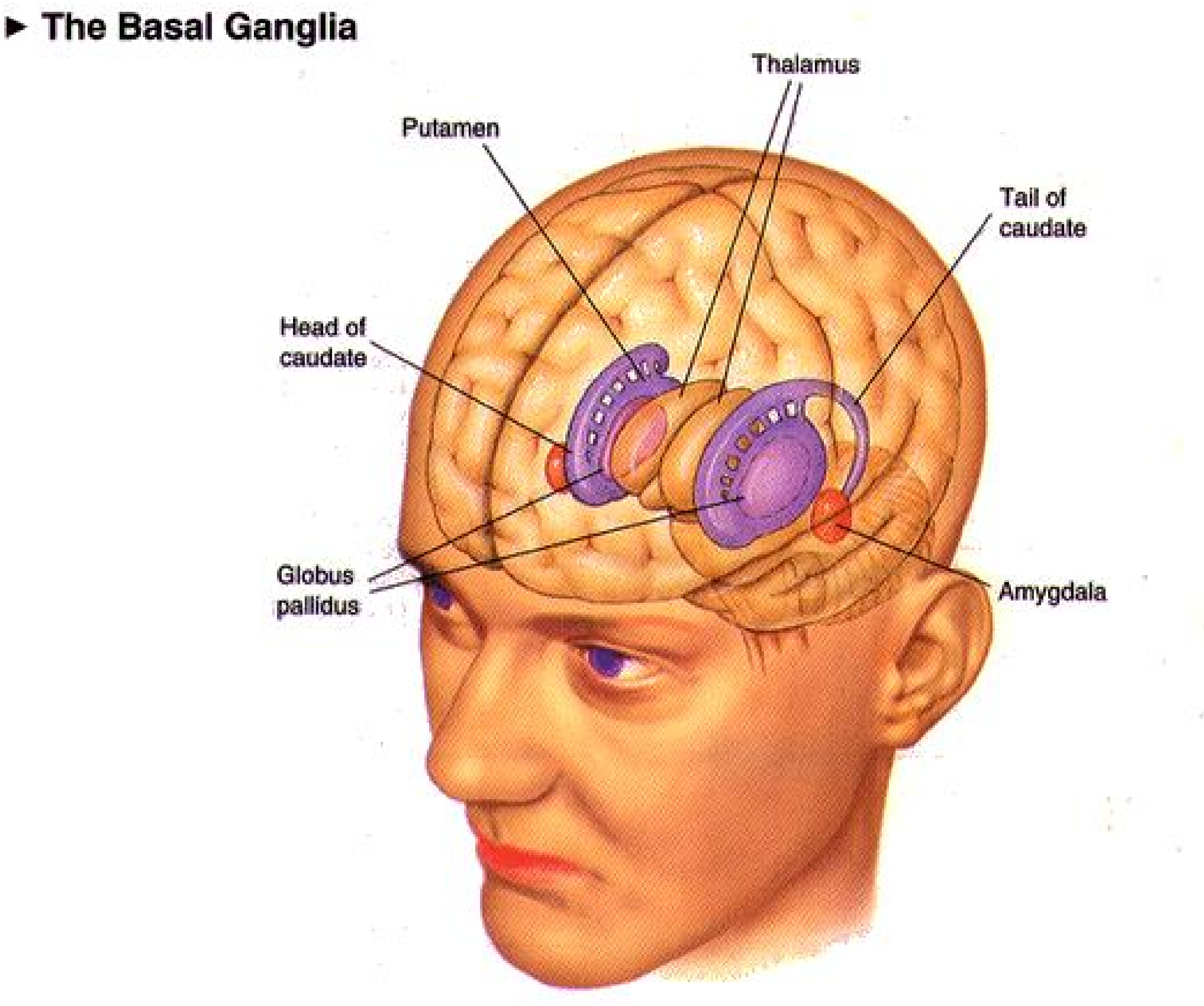




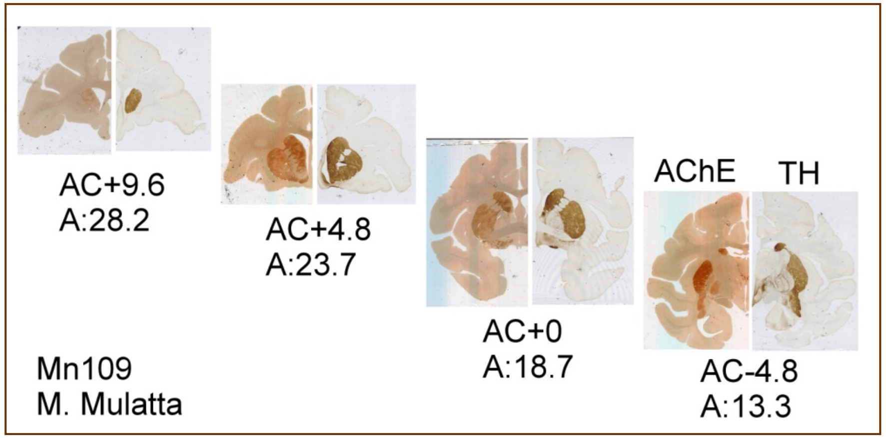
TD error: $\delta_t = \gamma V_{t+1} - V_t + r_t$



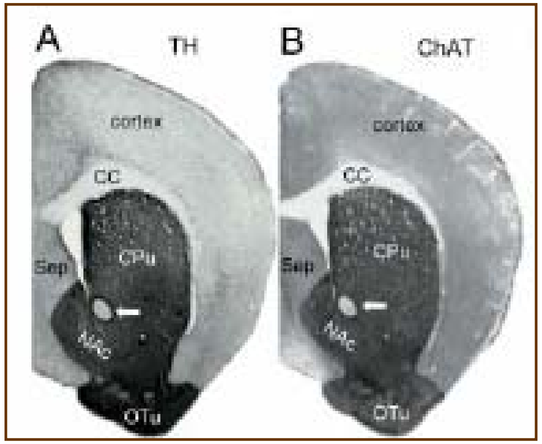
Basal ganglia - anatomy



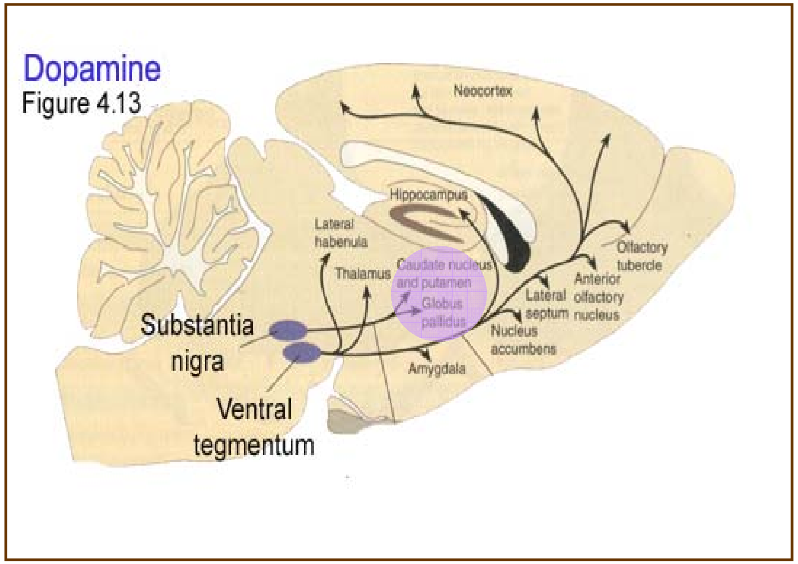
Dopamine and acetylcholine meet in the striatum

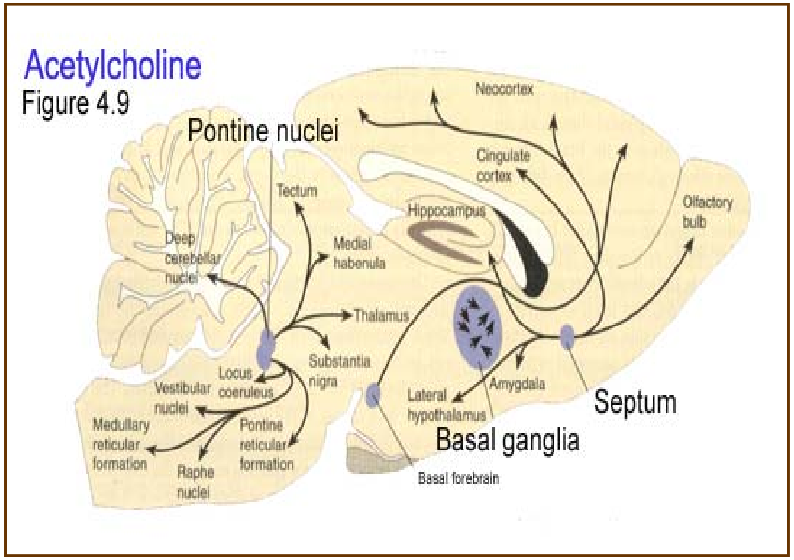


Monkey



Mouse

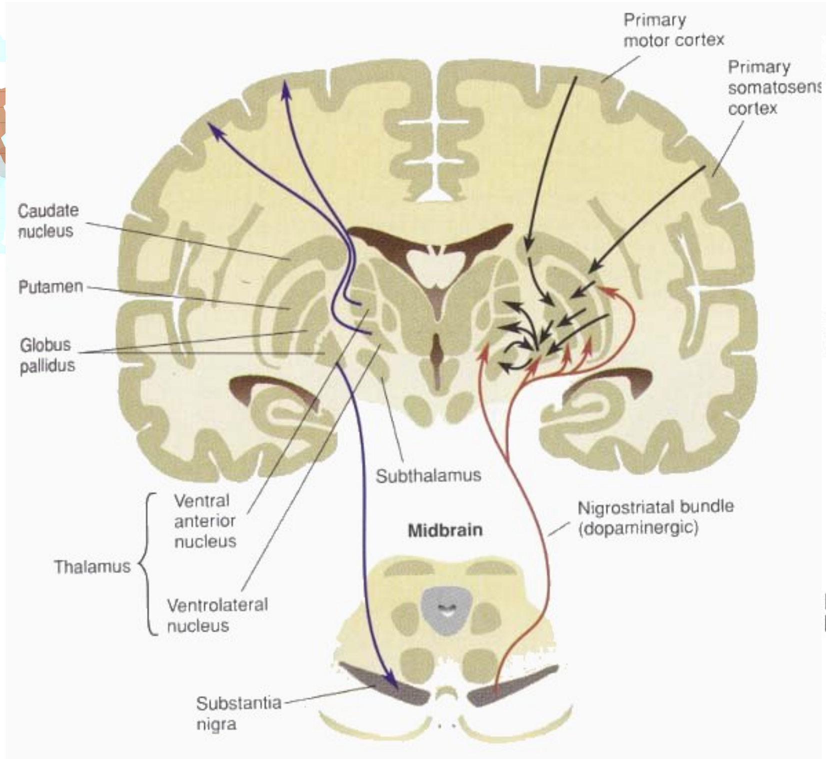


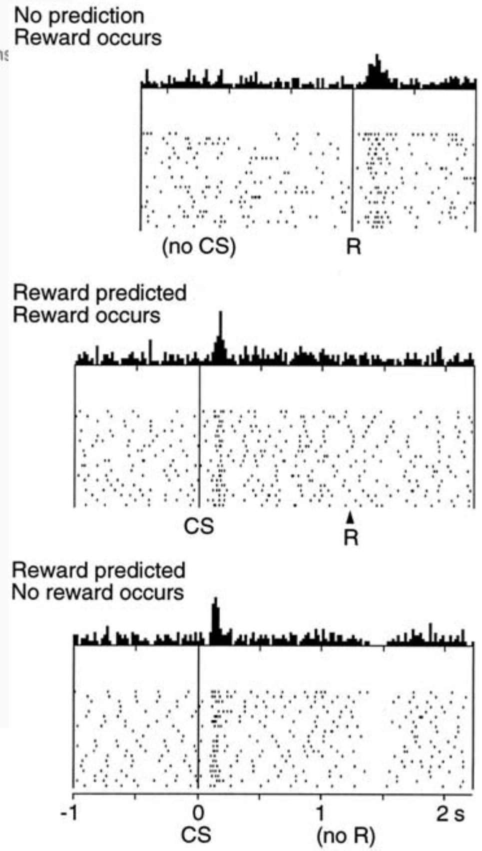


Facts to remember (1)

- Basal ganglia receive cortical input
- Basal ganglia project to frontal cortex
- Dopamine and acetylcholine localization

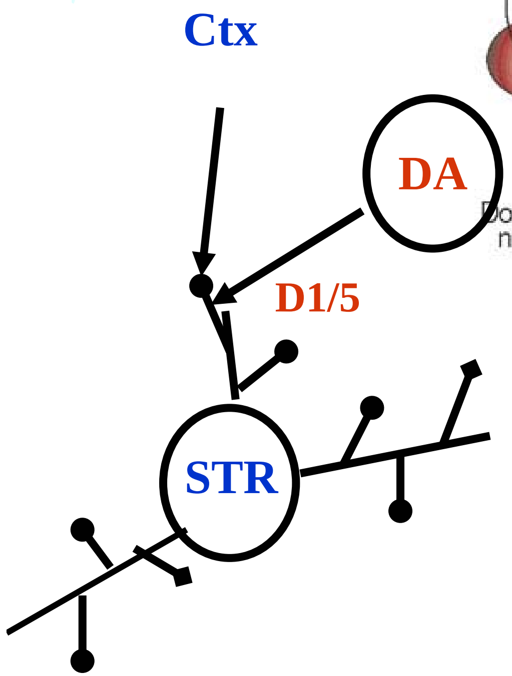
The midbrain dopamine system

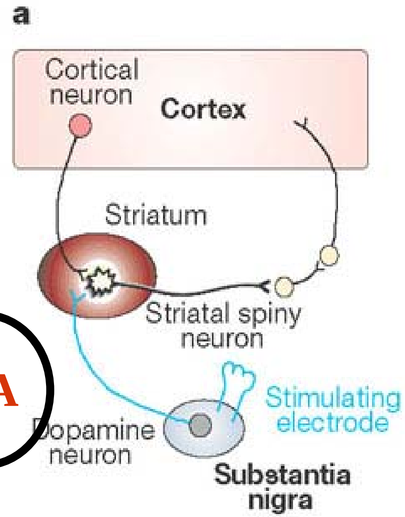


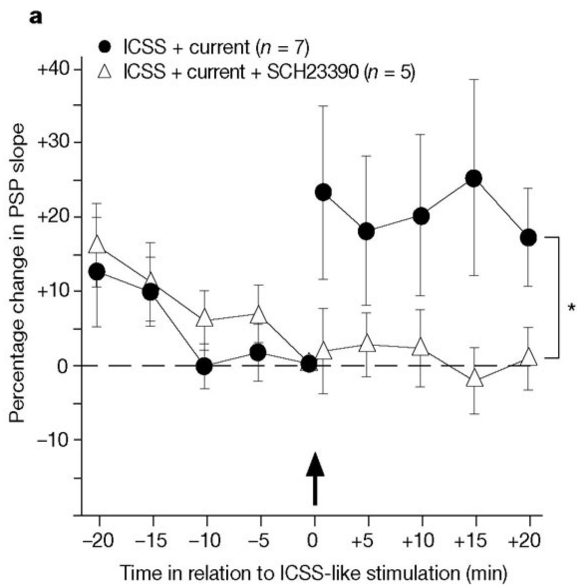


Schultz et al,
JNS 13:
900-913 ,1993

… and it can cause long term plasticity of cortico-striatal synapses

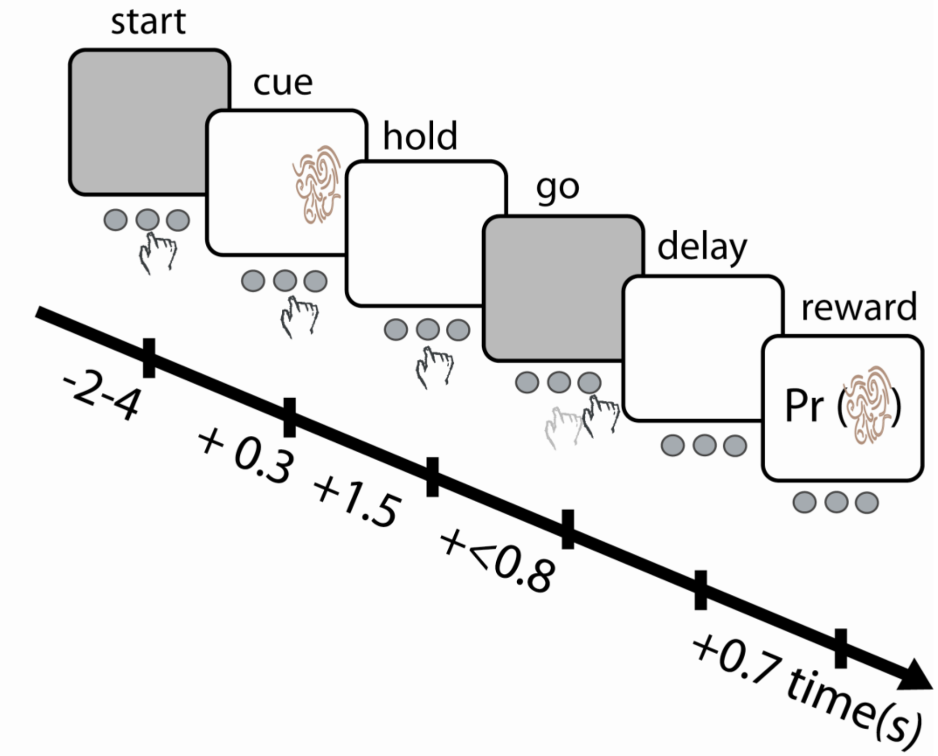




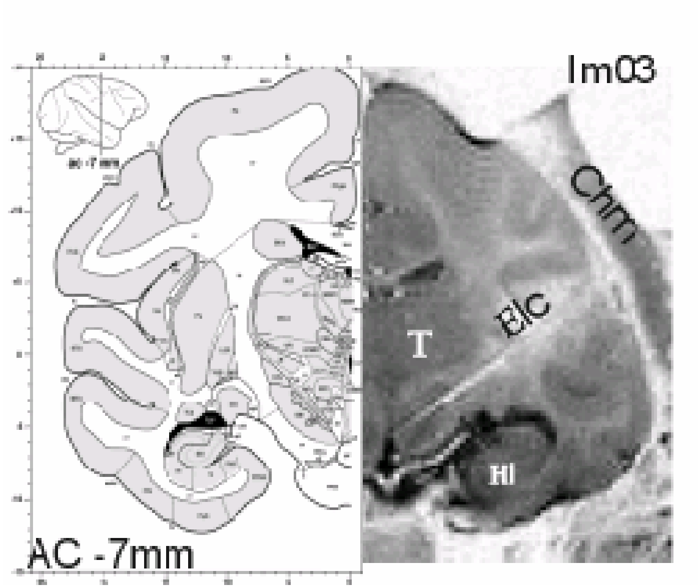


Reynolds et al, A cellular mechanism of reward-related learning Nature 413, 67 - 70 (2001)

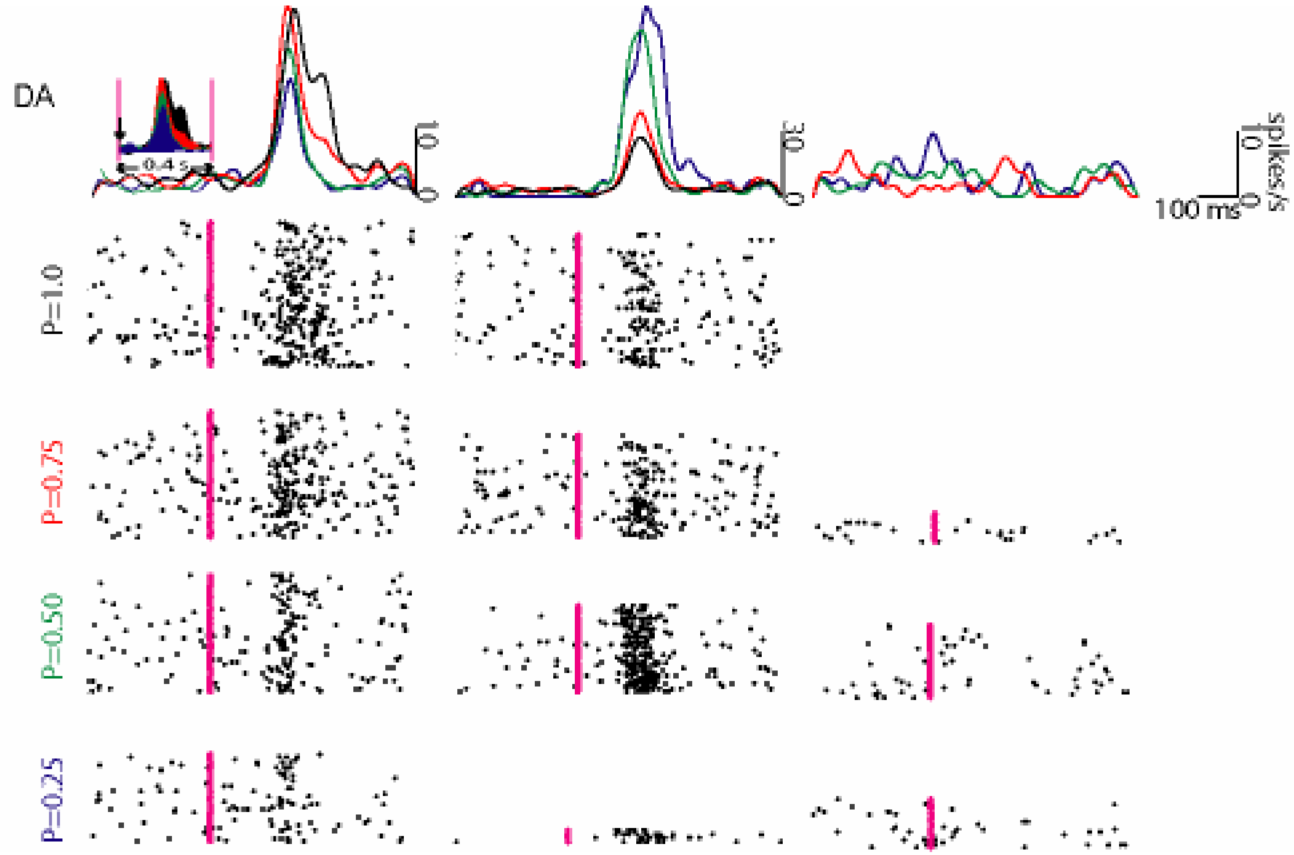
Probabilistic instrumental conditioning task



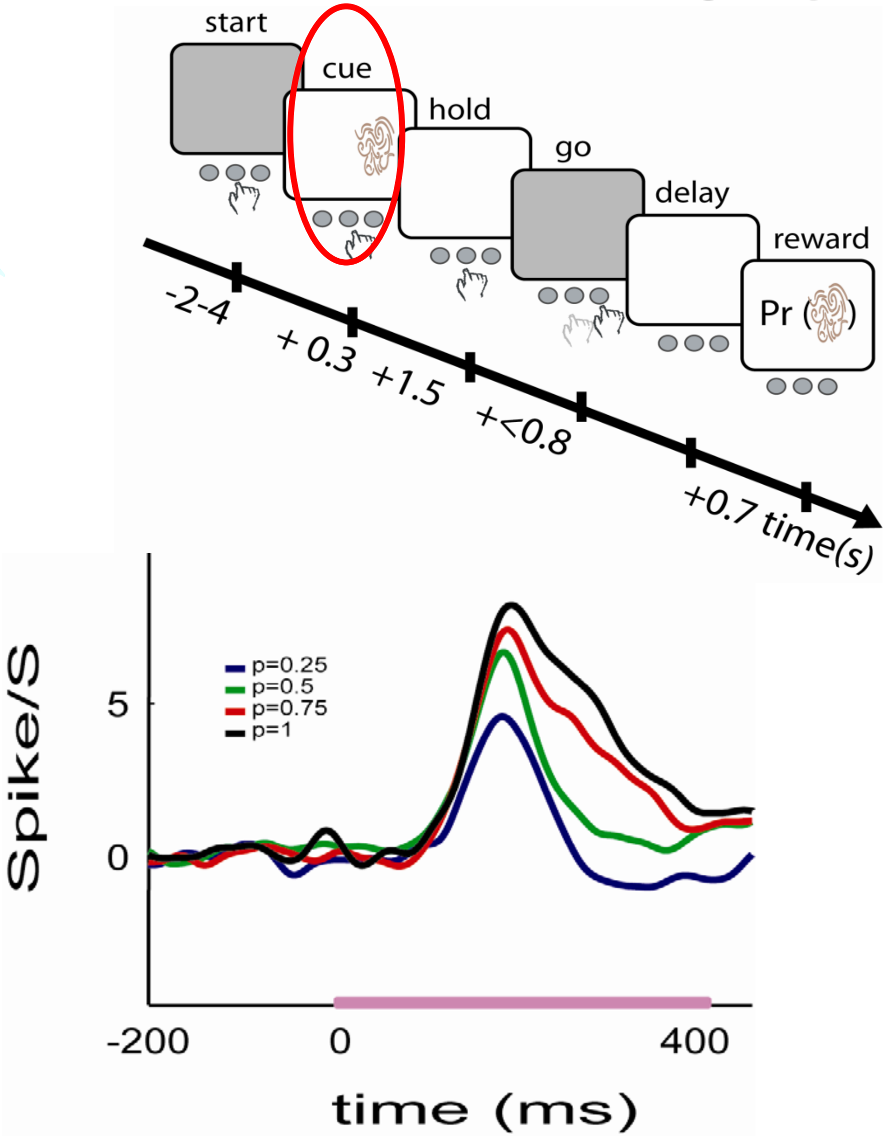
$$\delta_t = \gamma V_{t+1} - V_t + r_t$$

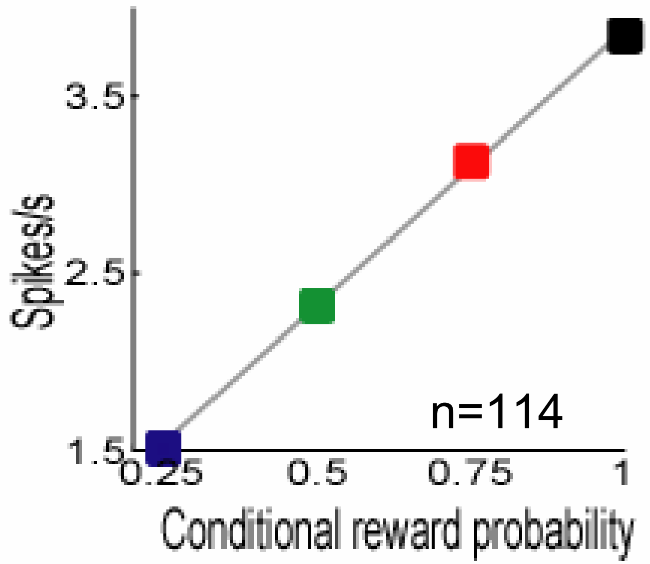


DA response

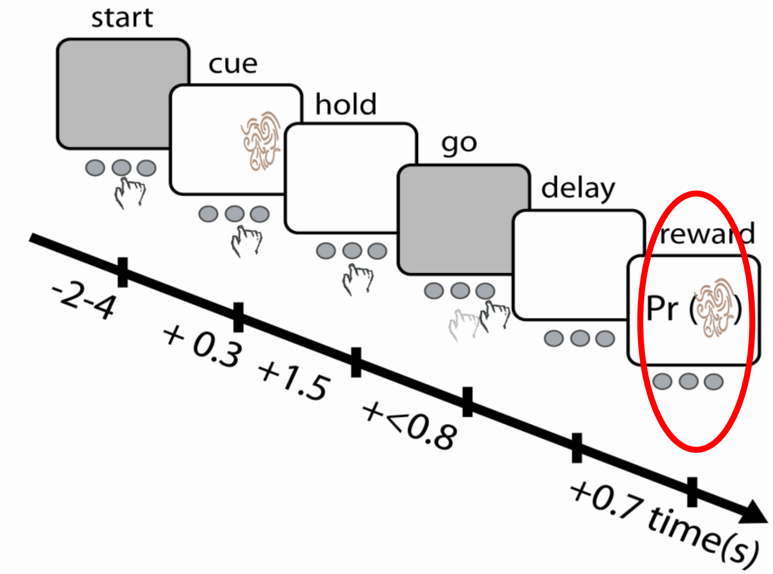


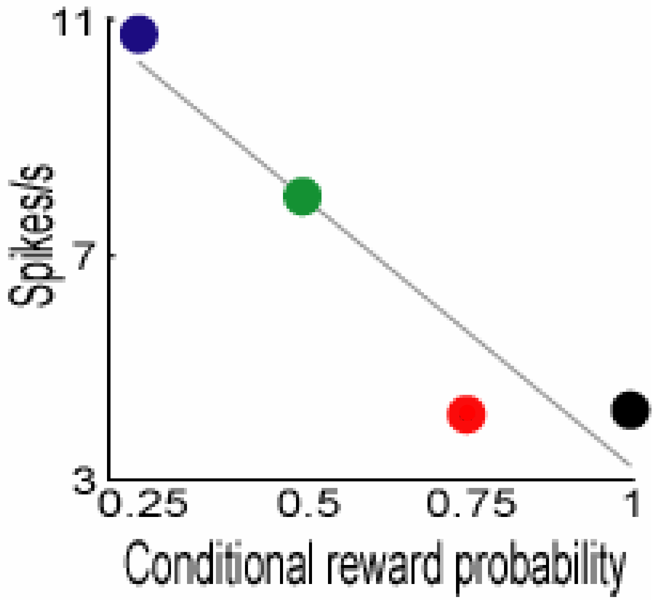
Dopamine population response-cue

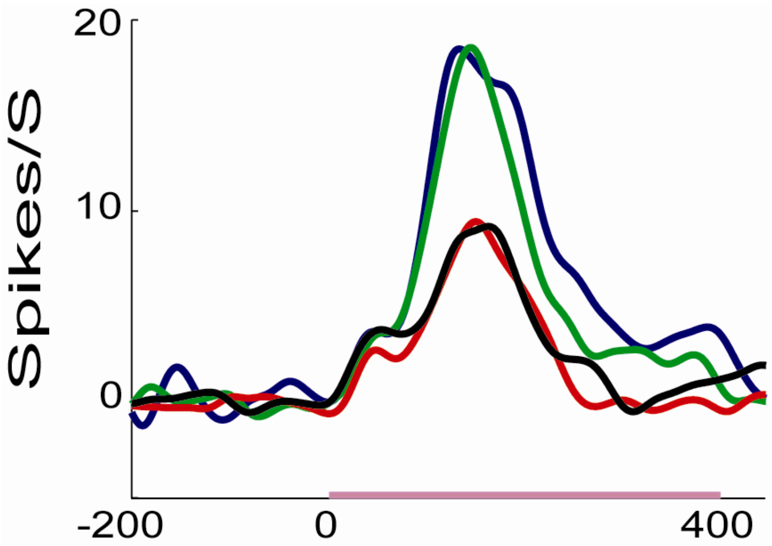




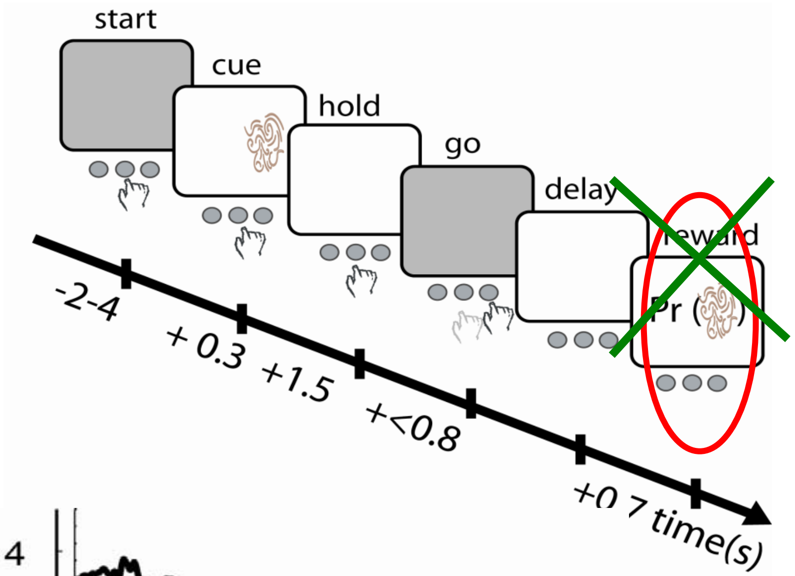
Dopamine population response-reward

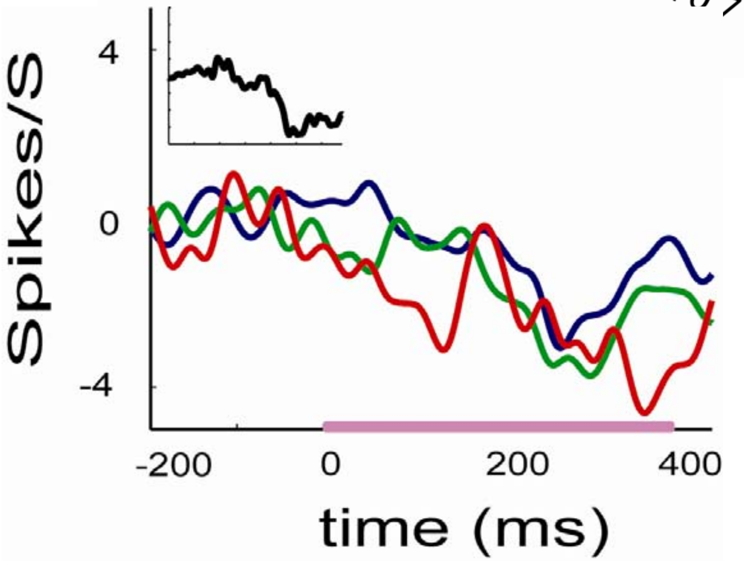


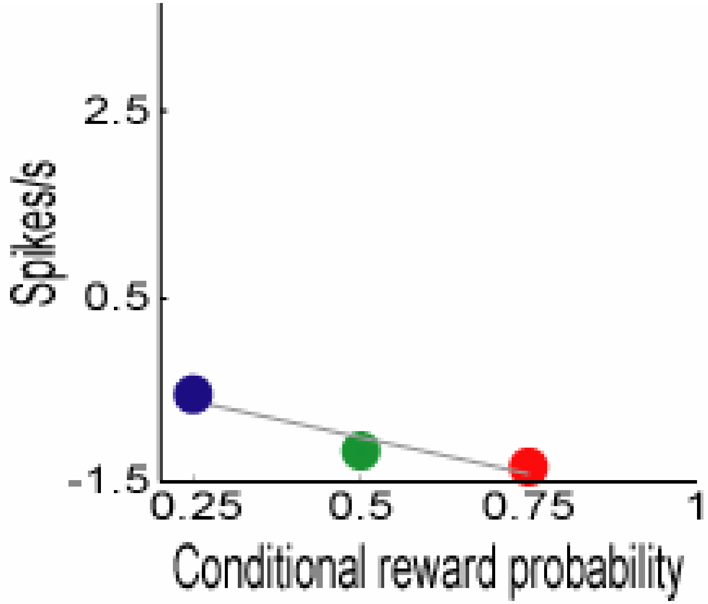




Dopamine population response – reward omission







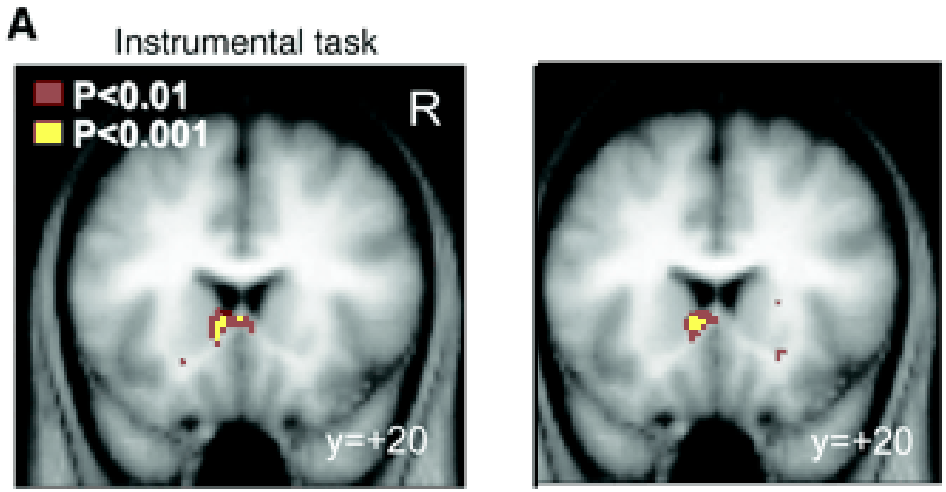
Instrumental conditioning - results

- Responses to visual cue are correlated with future reward probability
- Responses to reward are inversely correlated with reward probability
- Responses to reward omission are indifferent to reward probability
- Dopamine neurons provide an accurate TD signal (but only in the positive domain)

But how is this related to behaviour?

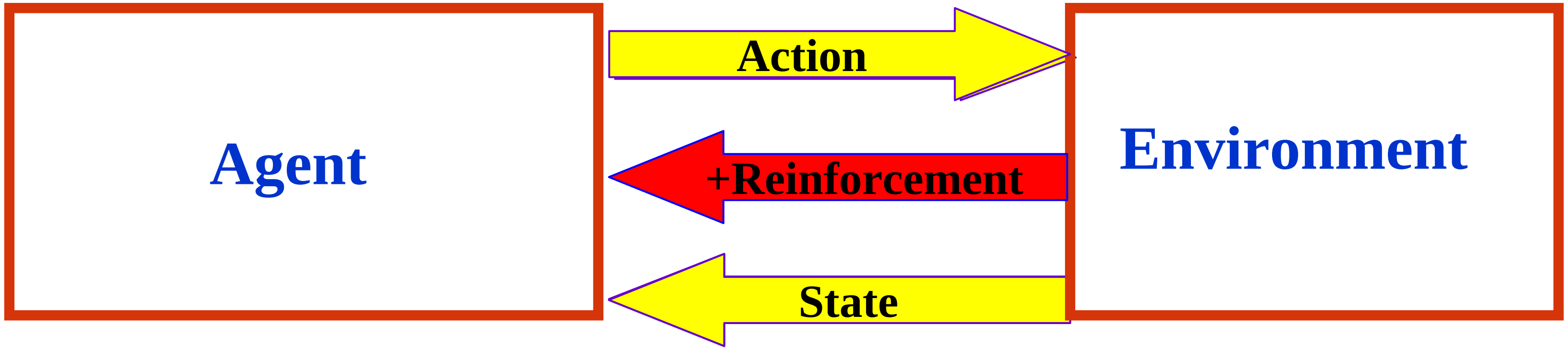
- The law of effect is all about actions
- AI agents act to maximize a goal
- …and so do people and animals

- The basal ganglia are involved in action



O'Doherty et

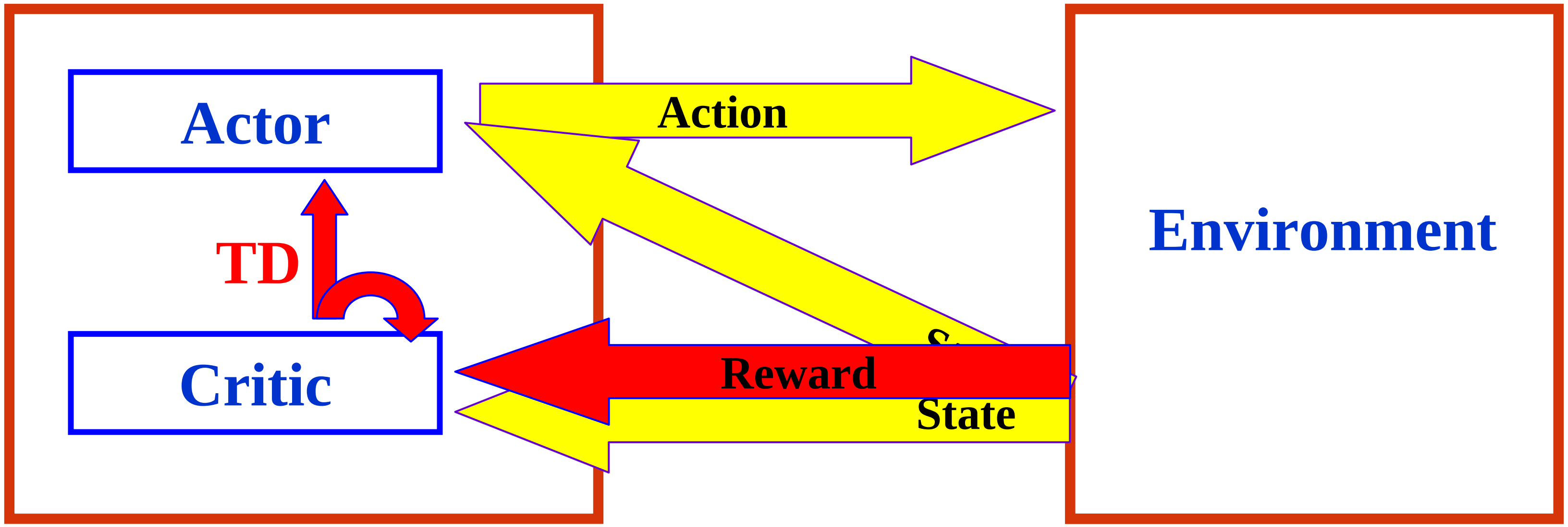
Control - Adding action



The agent has to:

- Learn to predict reinforcement *state value*
- Know the state-action-state transitions *behavioural policy*

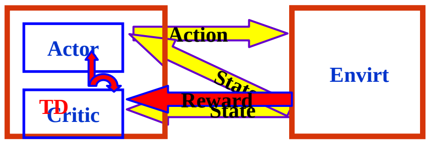
Solution 1: actor/critic networks

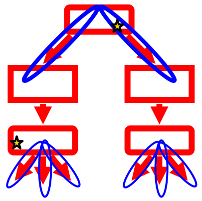


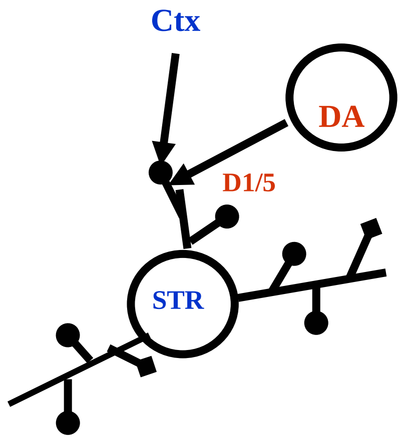
How can the dopamine signal contribute to decision behaviour?

- Long term policy-shaping effect

  through synaptic plasticity

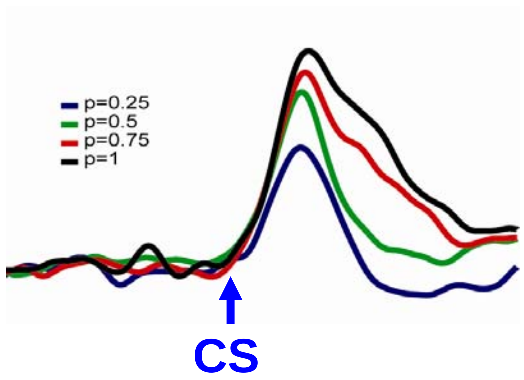


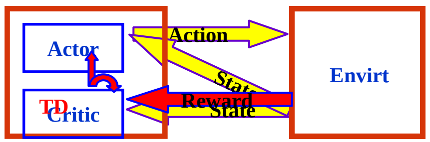




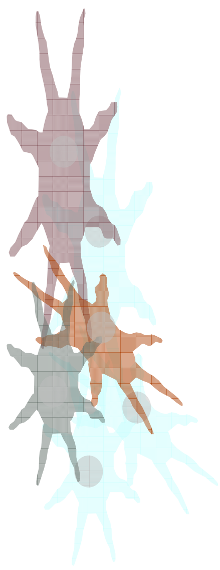
- Immediate effect on action

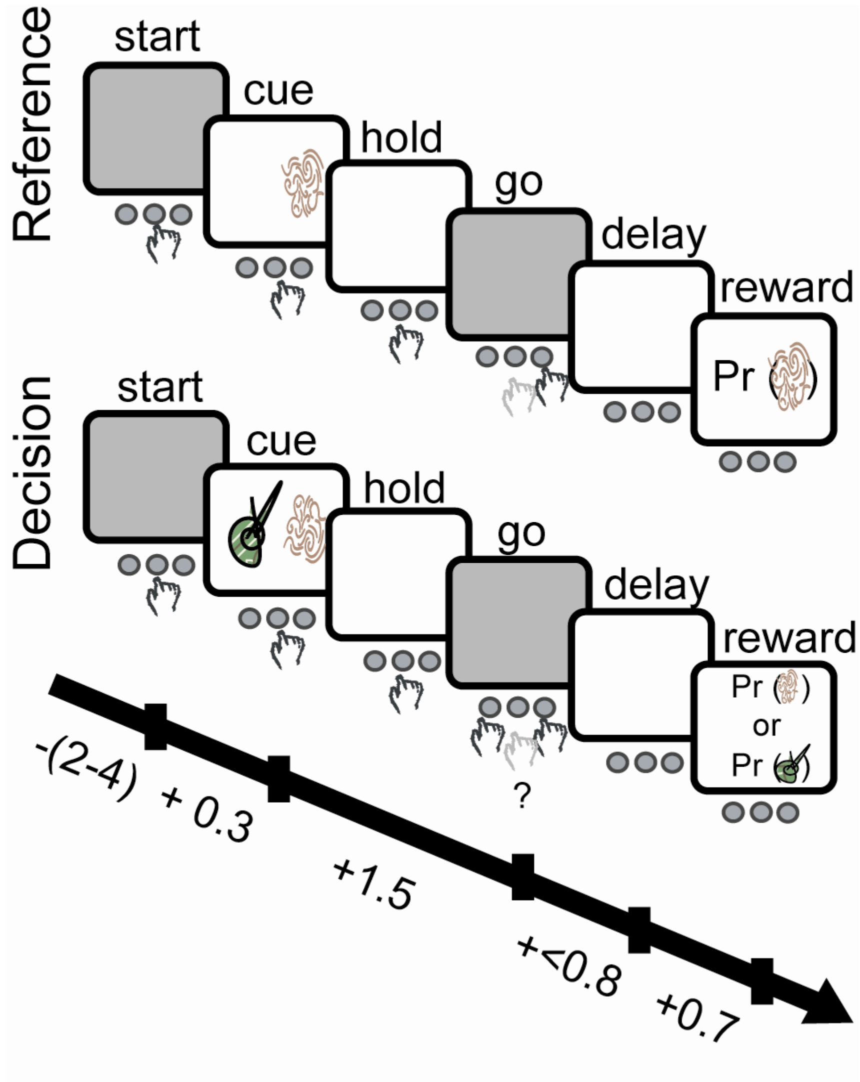
$$P_{action} = \frac{1}{1+e^{-m\delta(t)+b}}$$



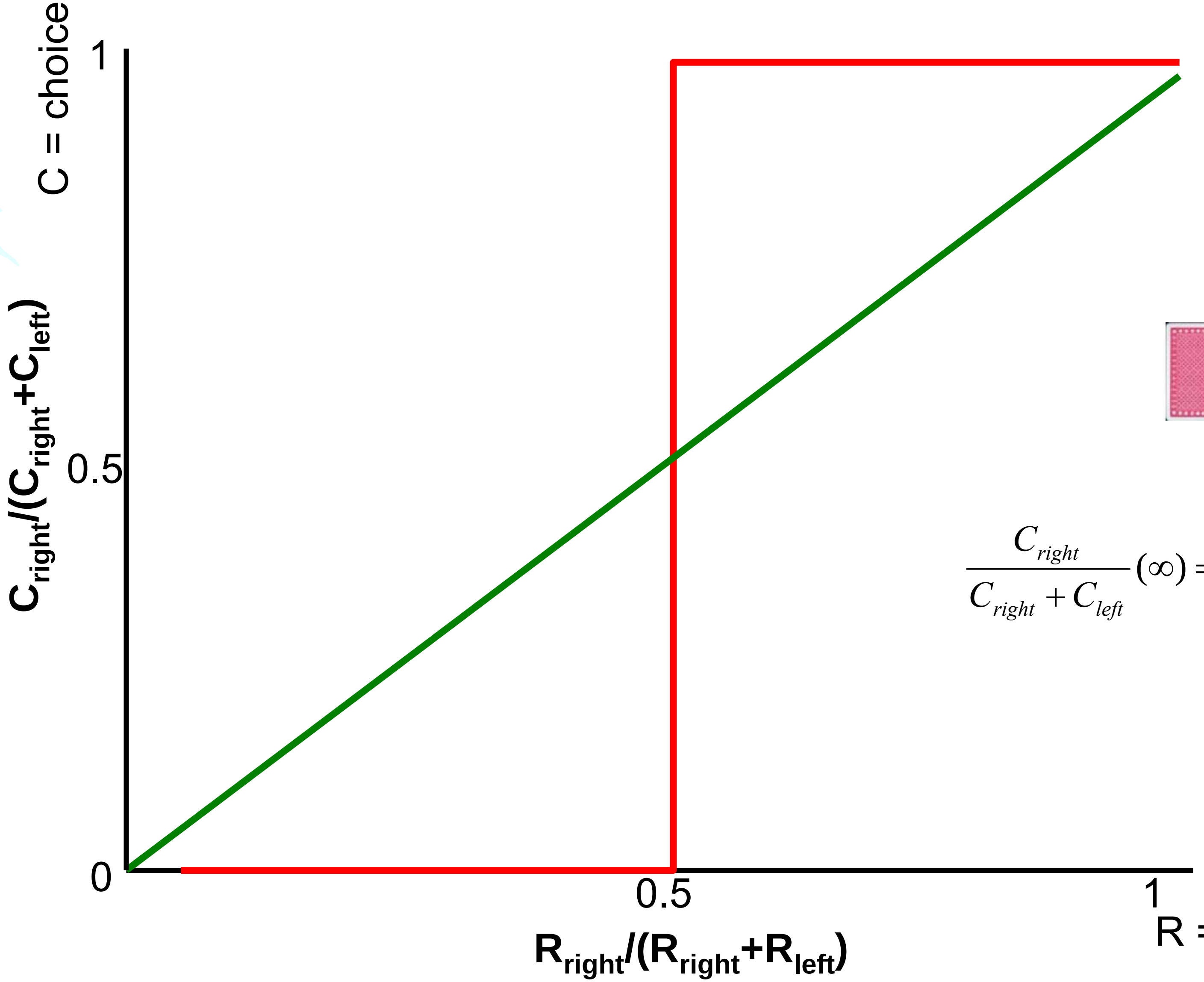


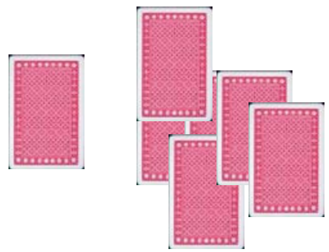
The two armed bandit task





Decision behaviour, theory and practice

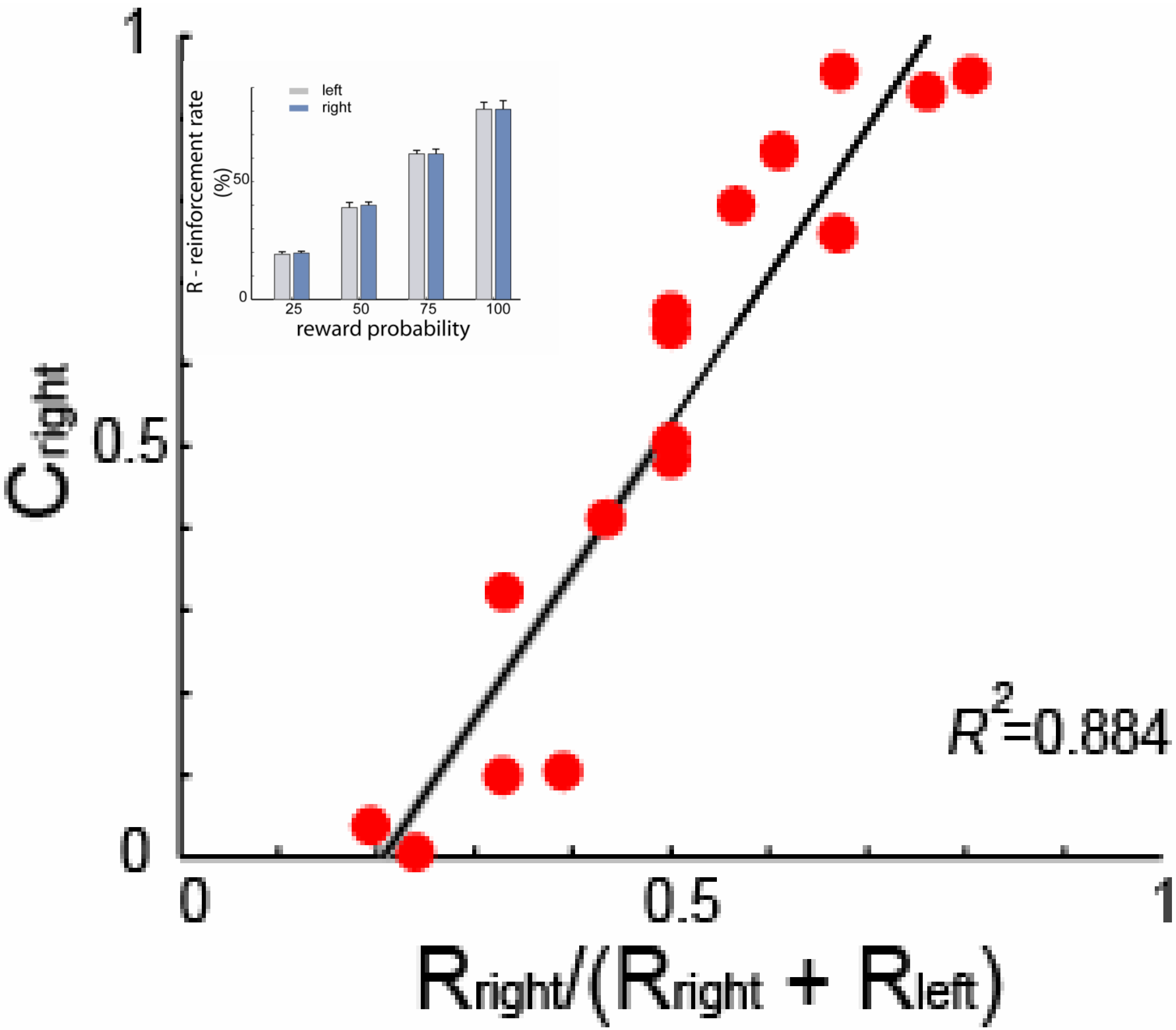




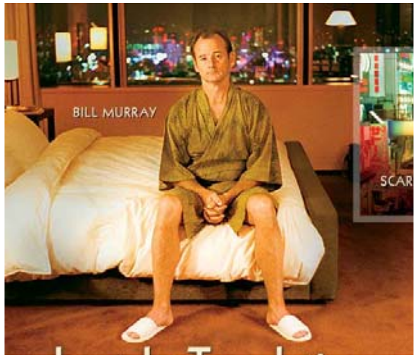
$$\frac{C_{right}}{C_{right}+C_{left}}(\infty)=\frac{R_{right}}{R_{right}+R_{left}\cdot\dfrac{\theta_{left}}{\theta_{right}}}$$

monkeys?

Monkeys' decisions: probability matching



Lost in translation?



reward → behaviour

dopamine response → plasticity in action circuits

Monkeys' decisions: shaping by dopamine