Genela Morris
Dept. of Neurobiology
Haifa University
gmorris@sci.haifa.ac.il

the role of dopamine in planning and action

ON NEURAL CORRELATES OF REINFORCEMENT LEARNING

Suggested reading

- Dayan P and Abbott LF. **Theoretical Neuroscience: Computational and Mathematical Modeling of Neural Systems**. MIT Press, Cambridge MA (2001): Ch. 9
- Barto AG & Sutton RS. **Reinforcement Learning: An introduction**. MIT Press, Cambridge MA (1988) : Ch. 3, Ch. 6 + some of Ch. 2
- Schultz W, Dayan P, Montague PR (1997), **A neural substrate of prediction and reward**, *Science* 275: 1593-1599
- Figures from research papers are referenced throughout the presentation

Reinforcement learning
the basics

Supervised learning –

all knowing teacher, detailed feedback

Reinforcement learning –

scalar (correct/incorrect) feedback

Unsupervised learning –

self organization

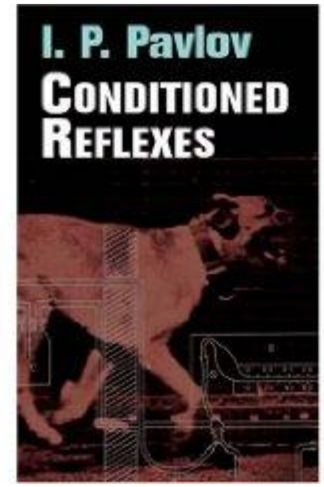
Reinforcement learning: The law of effect

“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



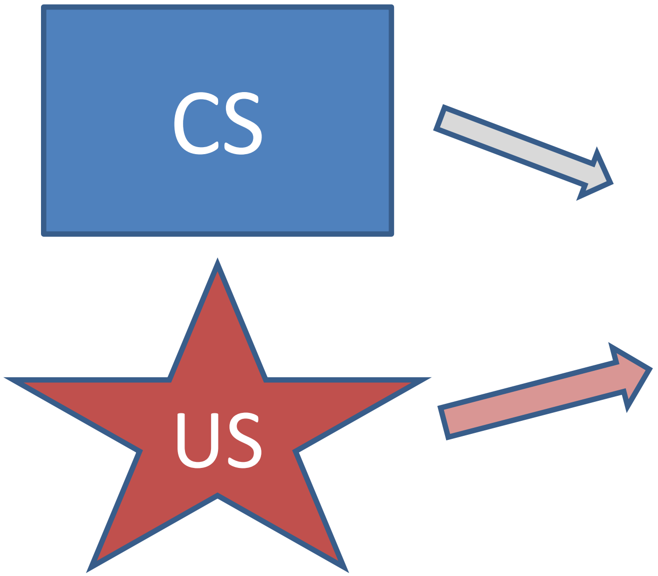
Properties of classical conditioning

(Pavlov 1927)

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

So far…

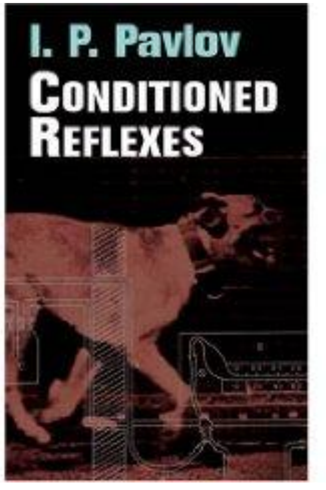
- A simple association (coincidence, Hebbian) model can explain the phenomenon.



  - But…

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

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

(Cnt'd)

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

CS must RELIABLY predict US

Which simple association can't explain

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

Rescorla-Wagner rule (1972)

Learning to predict reward R given stimulus U=1

Goal: Form a prediction V of the reward 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:
- ω1=R
- V= ω1U1+ ω2U2

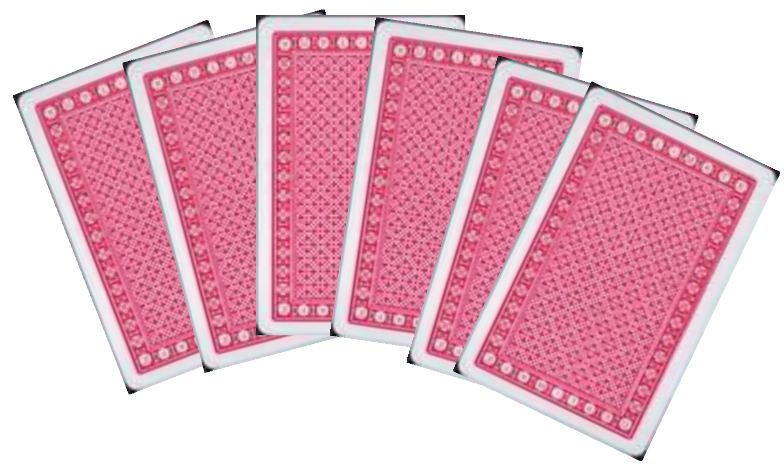
R 0

- Prediction error: R-V=0

And no learning occurs for ω2

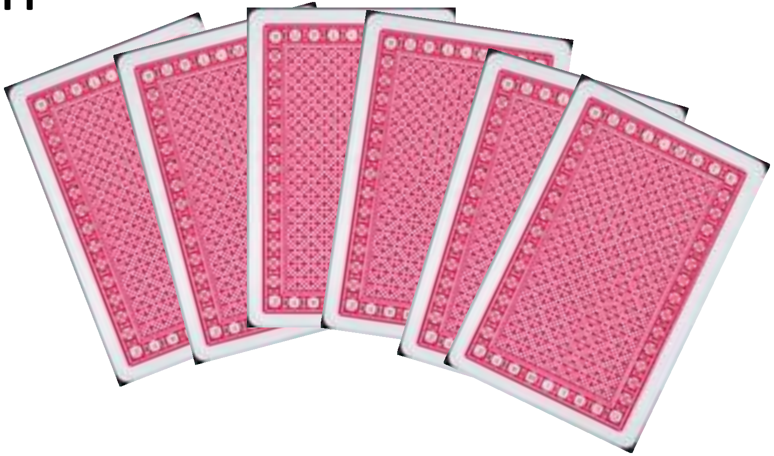
Critical problems, for control

1. Exploration/exploitation

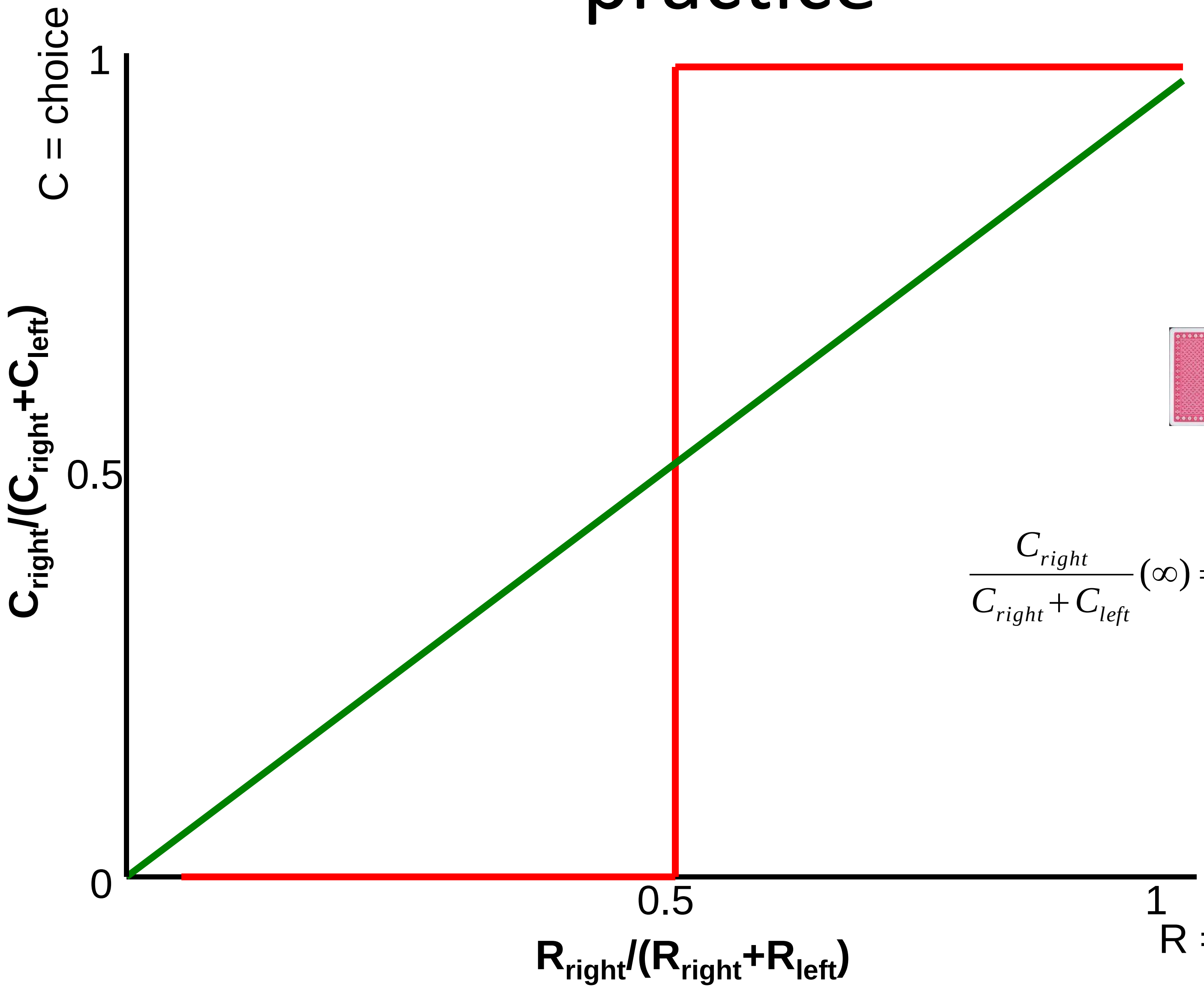


Solutions, for control

1. Variability in response policy
   1. Greedy 🡐 → Random (gambling)
   2. Based on expected return

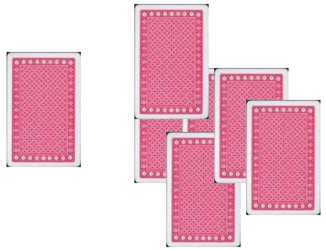


Decision behaviour, theory and practice



maximizing

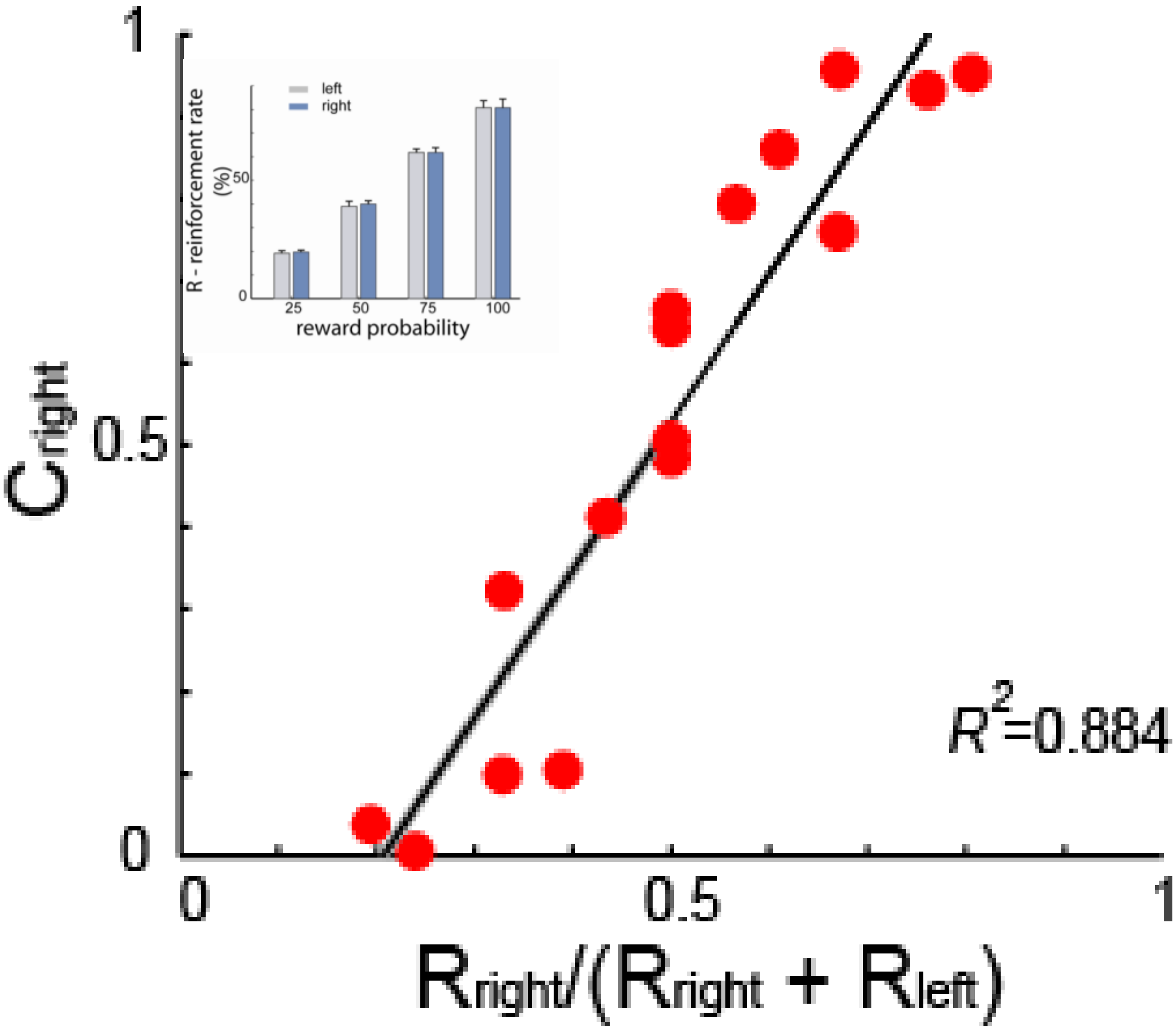
probability-matching



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

monkeys?

Monkeys' decisions: probability matching

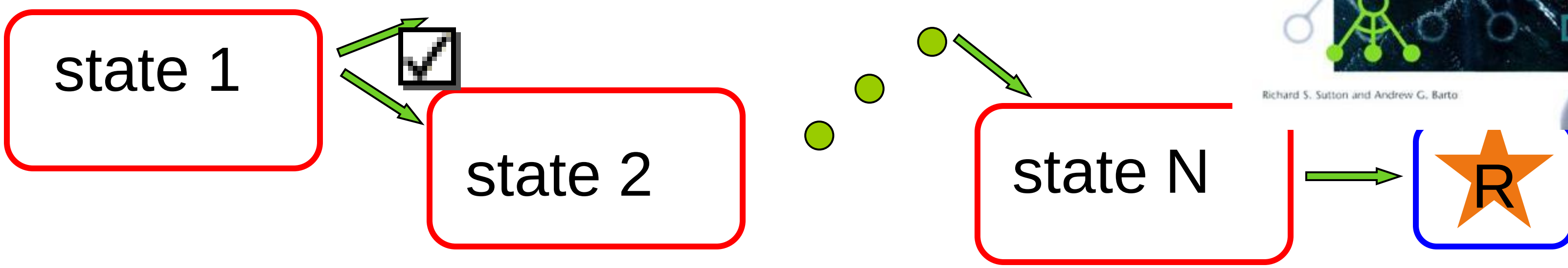


… whether optimal or not

- Actions are related to their consequences

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

2. Temporal credit assignment



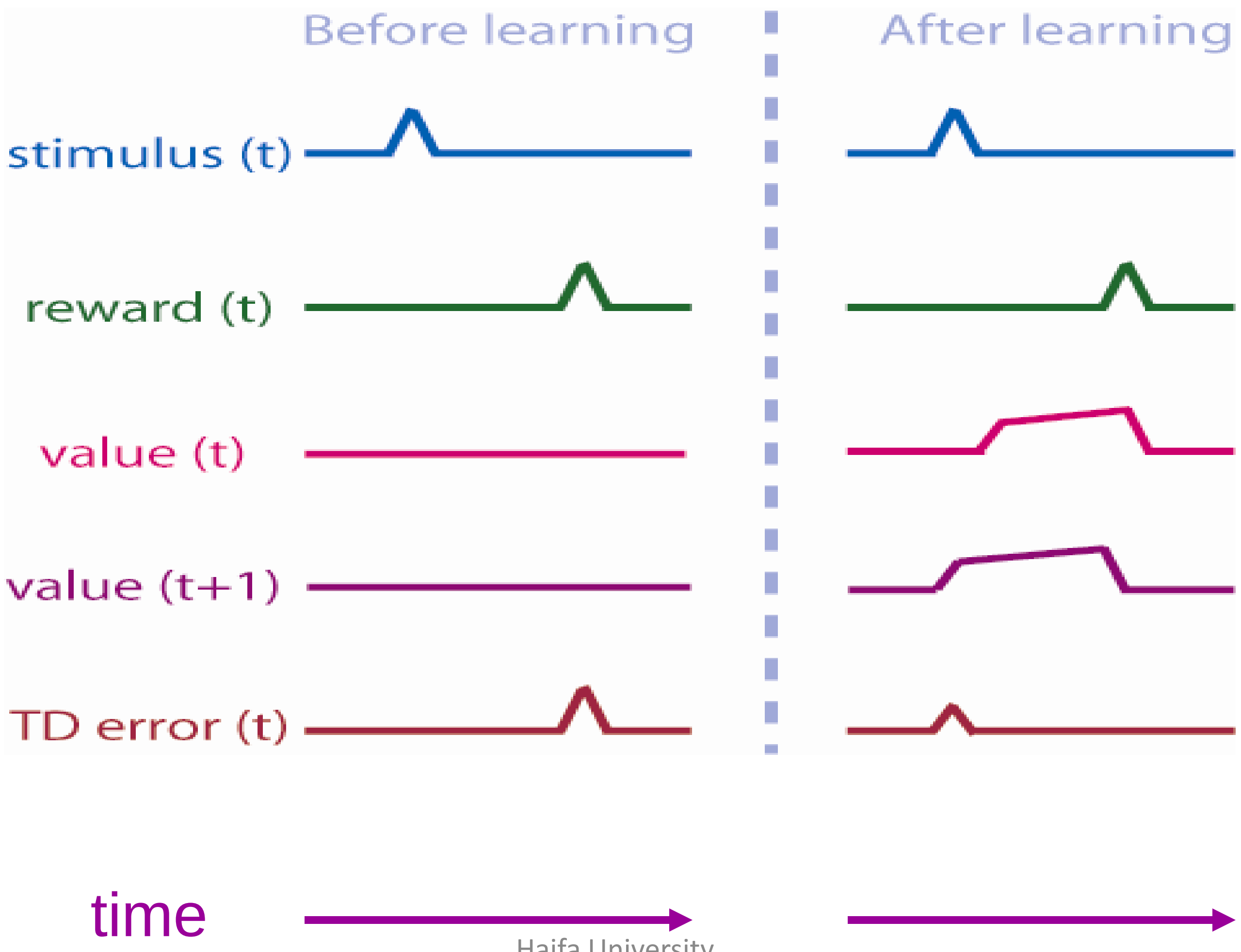
TD learning - solution for temporal credit assignment

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 $(V_t^{k+1} = V_t^k + \eta\,\delta_t)$

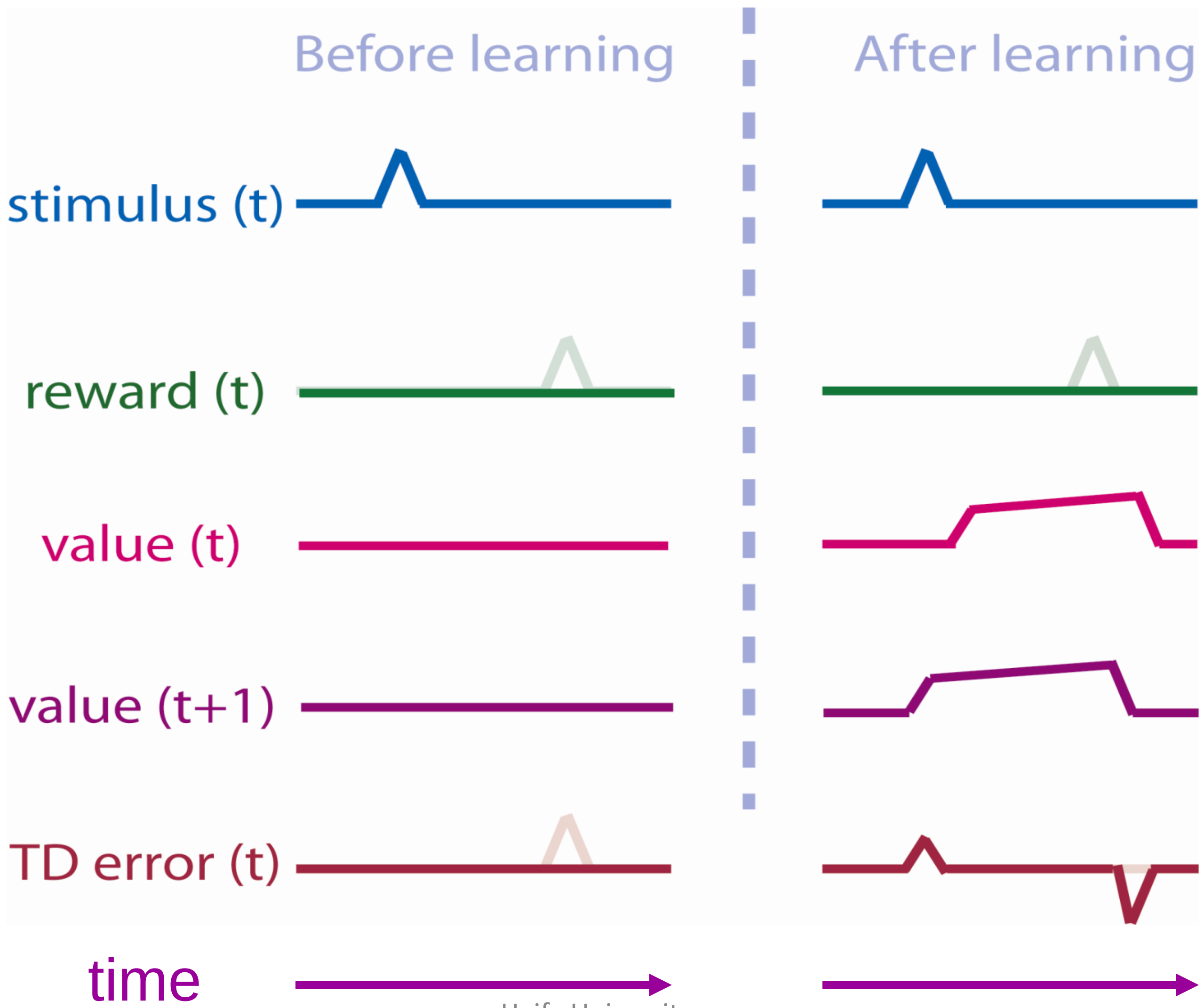
where: $V_{t\,=\,value}$ *of the state reached at time t in iteration k*

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

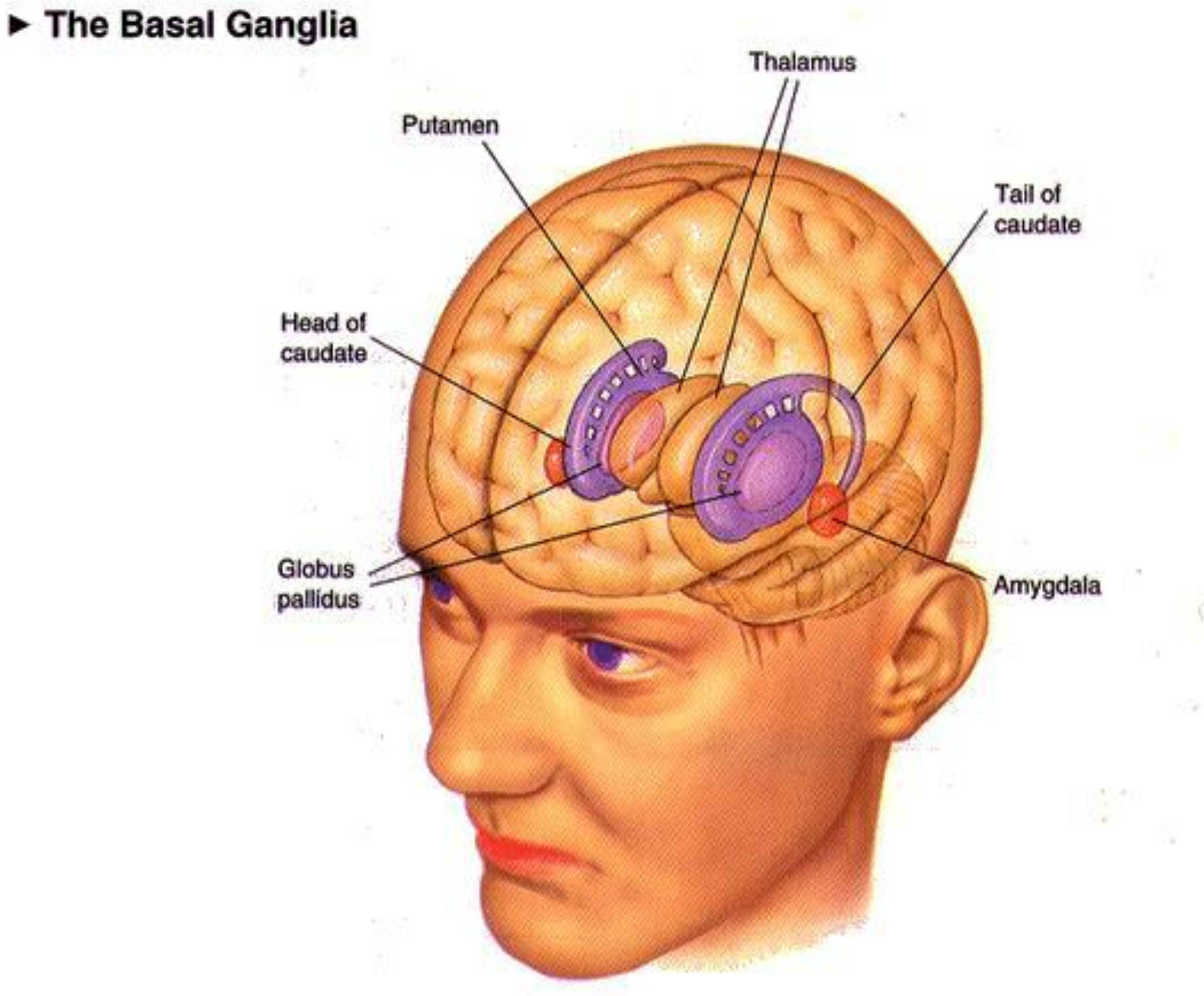
TD error: $\delta_t = r_t + \gamma V_{t+1} - V_t$



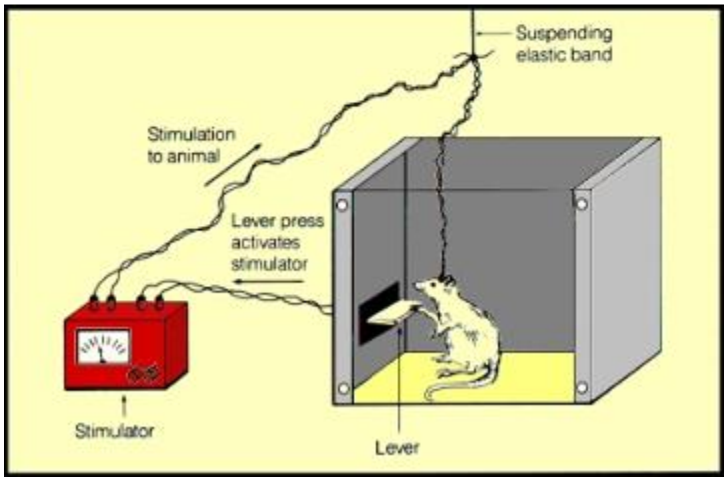
TD error: $\delta_t = \gamma V_{t+1} - V_t + r_t$



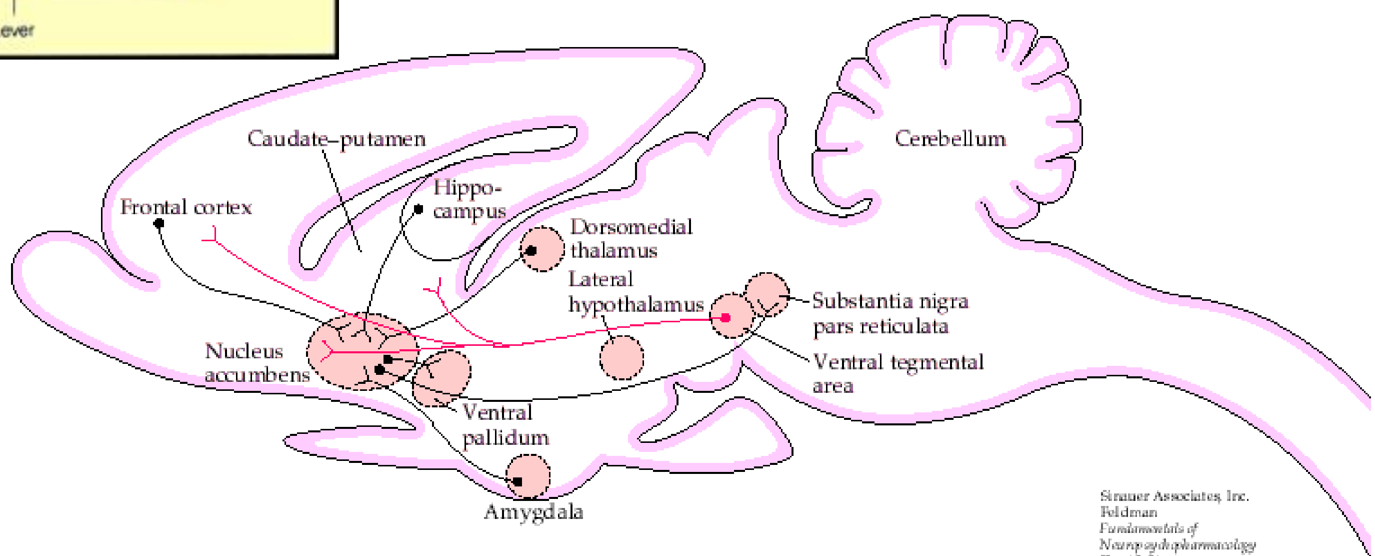
Basal ganglia - anatomy



Intracranial self stimulation

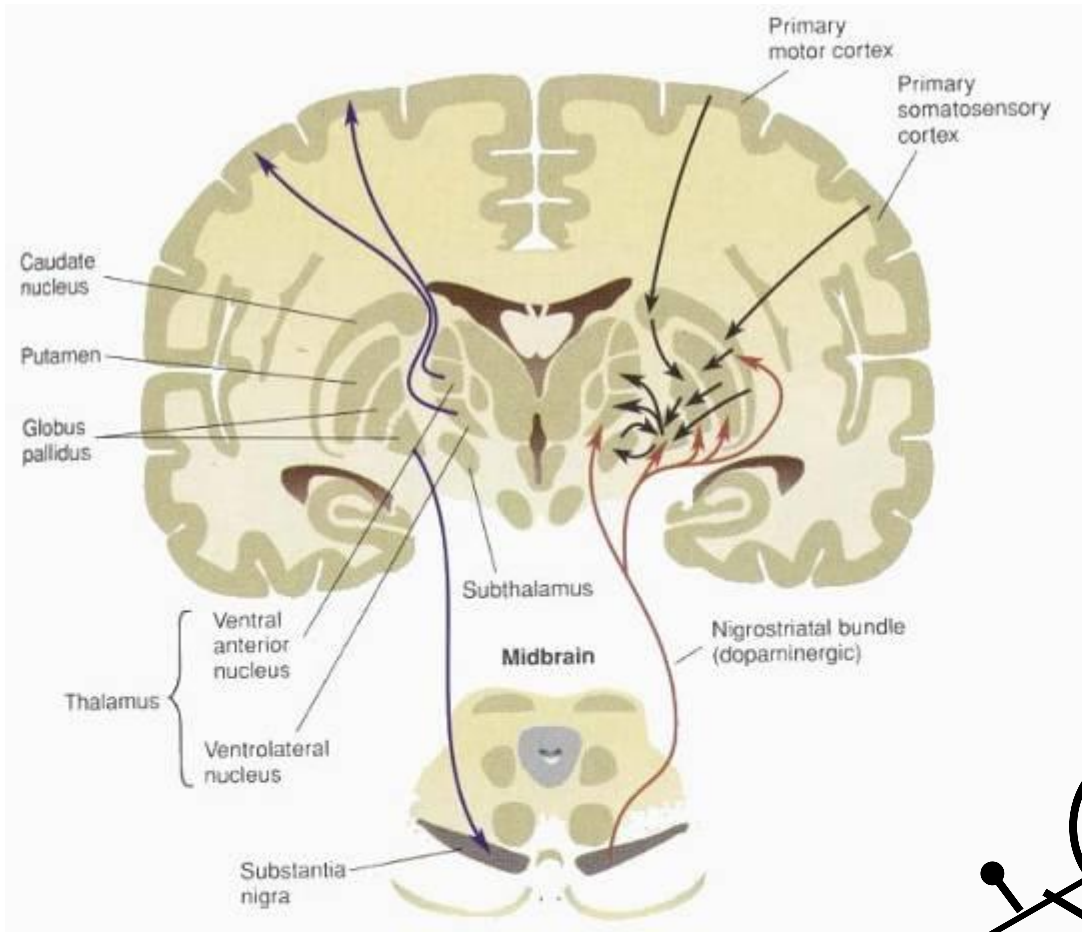


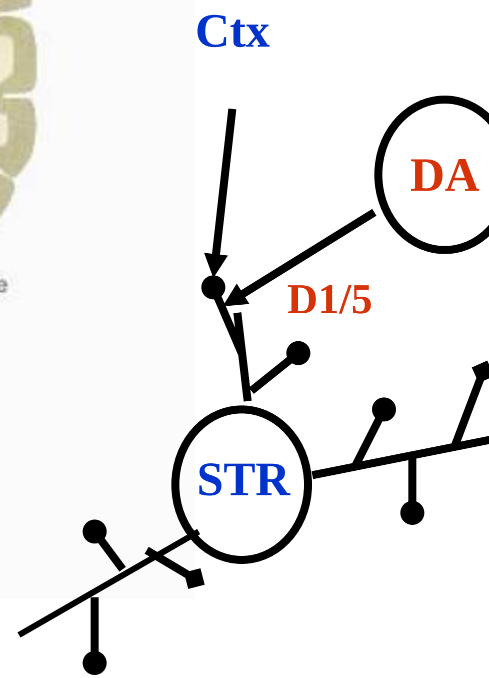
ACTIVATES REWARD CIRCUITS



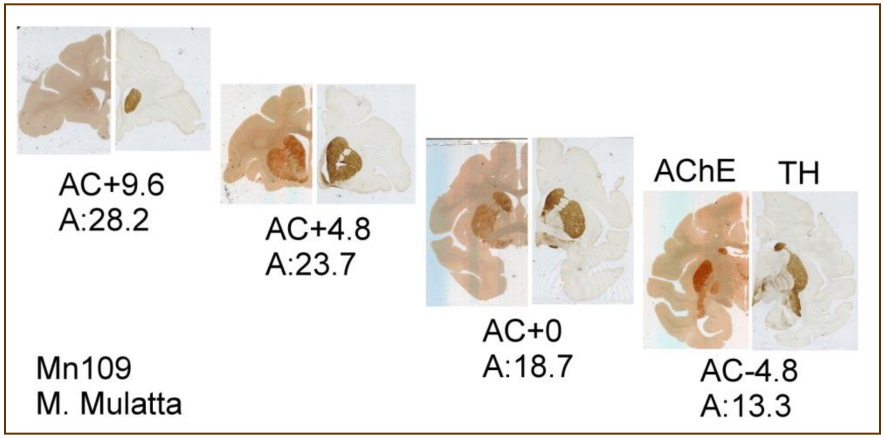
Sinauer Associates, Inc.
Feldman
Fundamentals of Neuropsychopharmacology
Fig. 13-21

The midbrain dopamine system

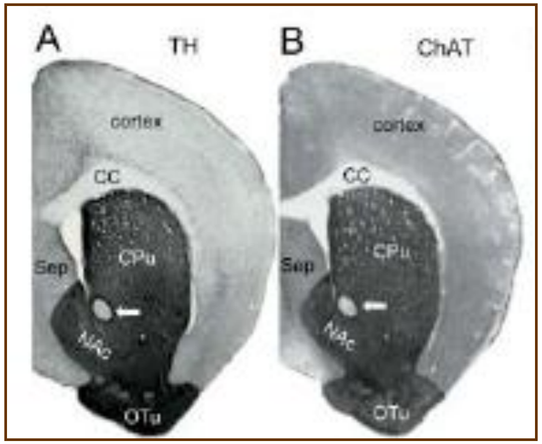




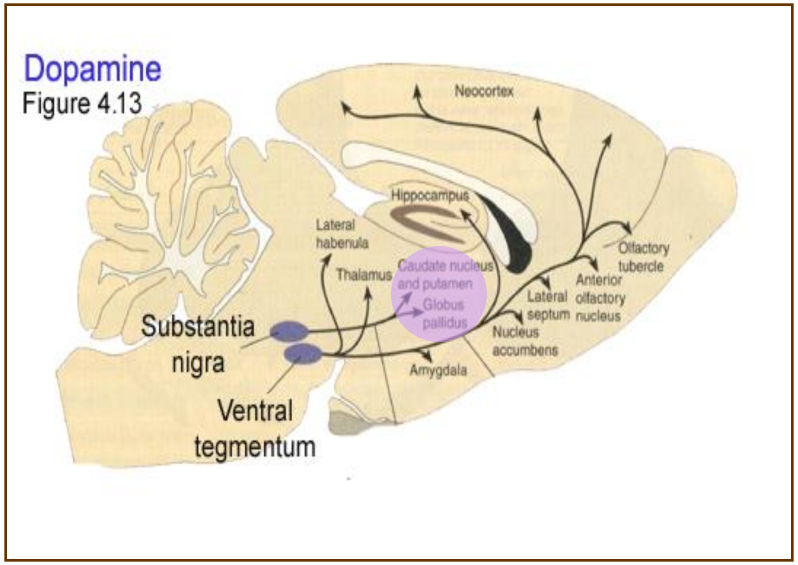
Dopamine and acetylcholine meet in the striatum

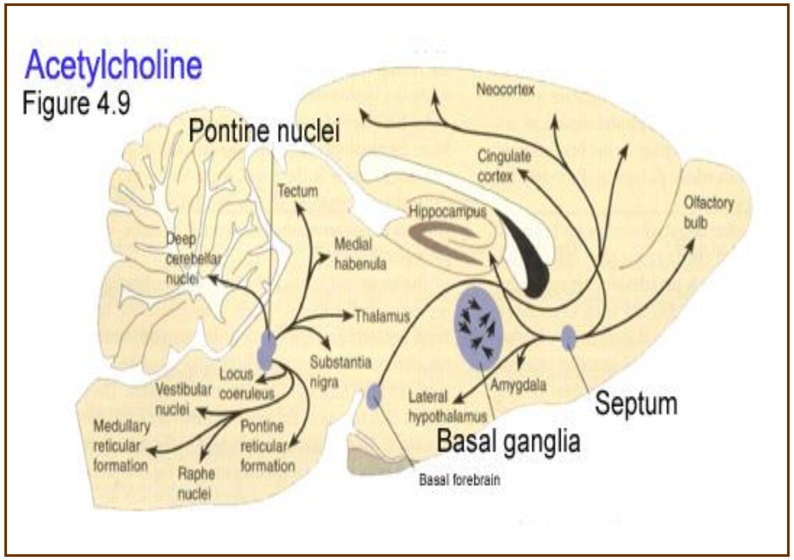


Monkey



Mouse

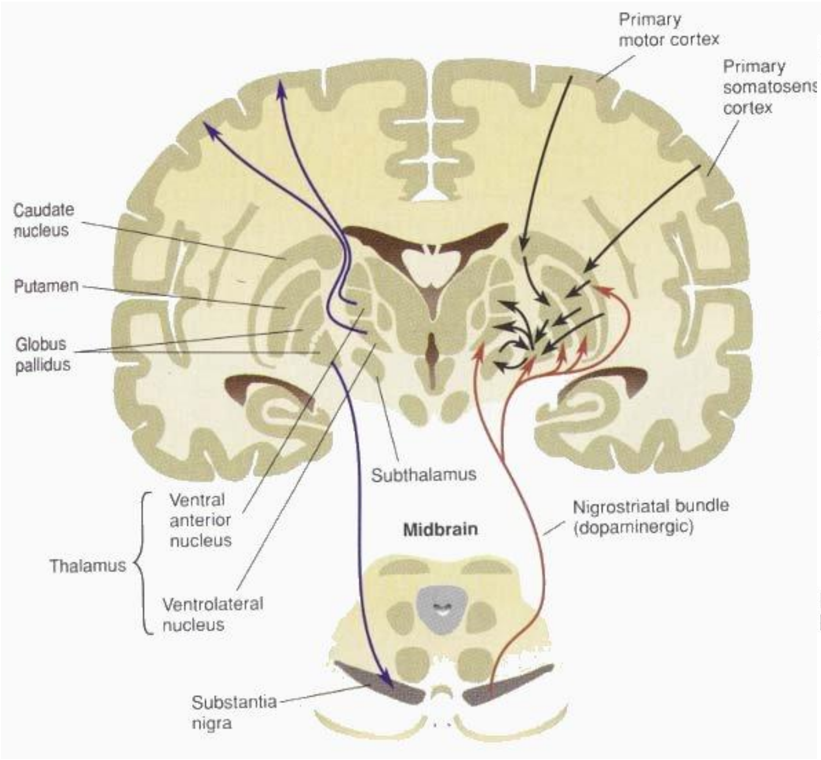


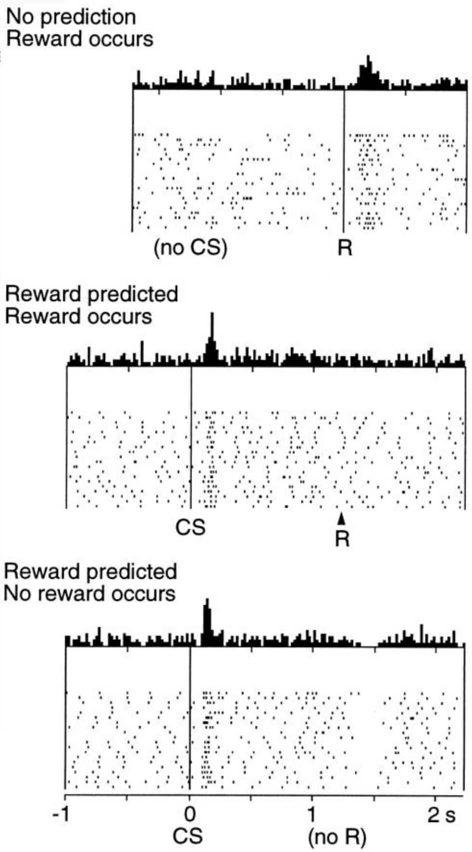


Facts to remember (1)

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

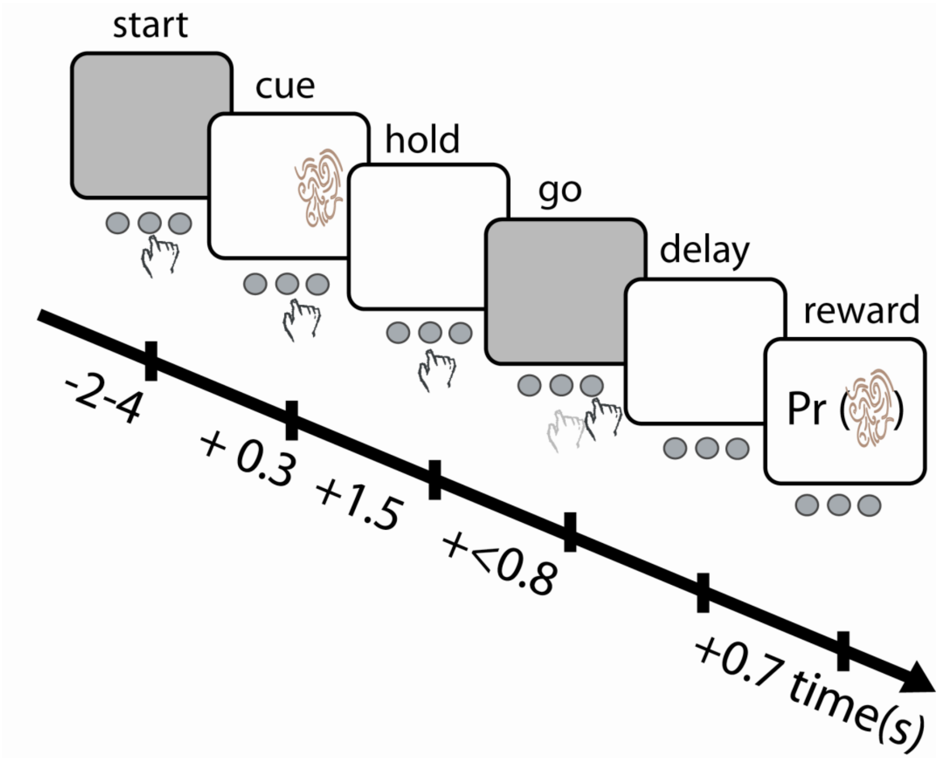
The midbrain dopamine system

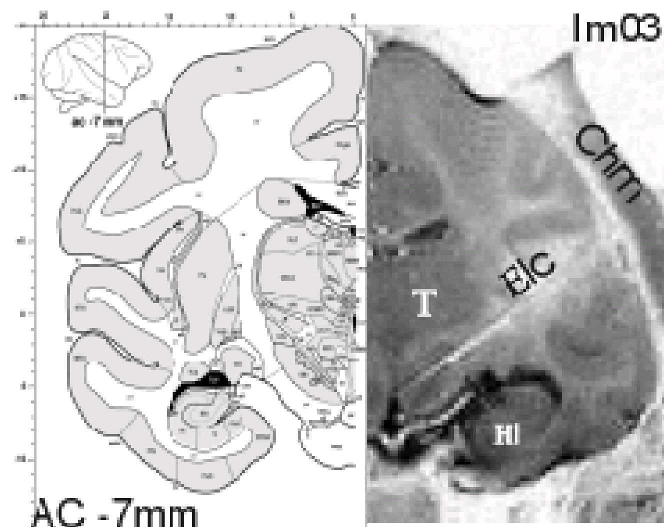




Schultz et al,
J. Neurosci 13: 900-913 ,1993

Probabilistic instrumental conditioning task

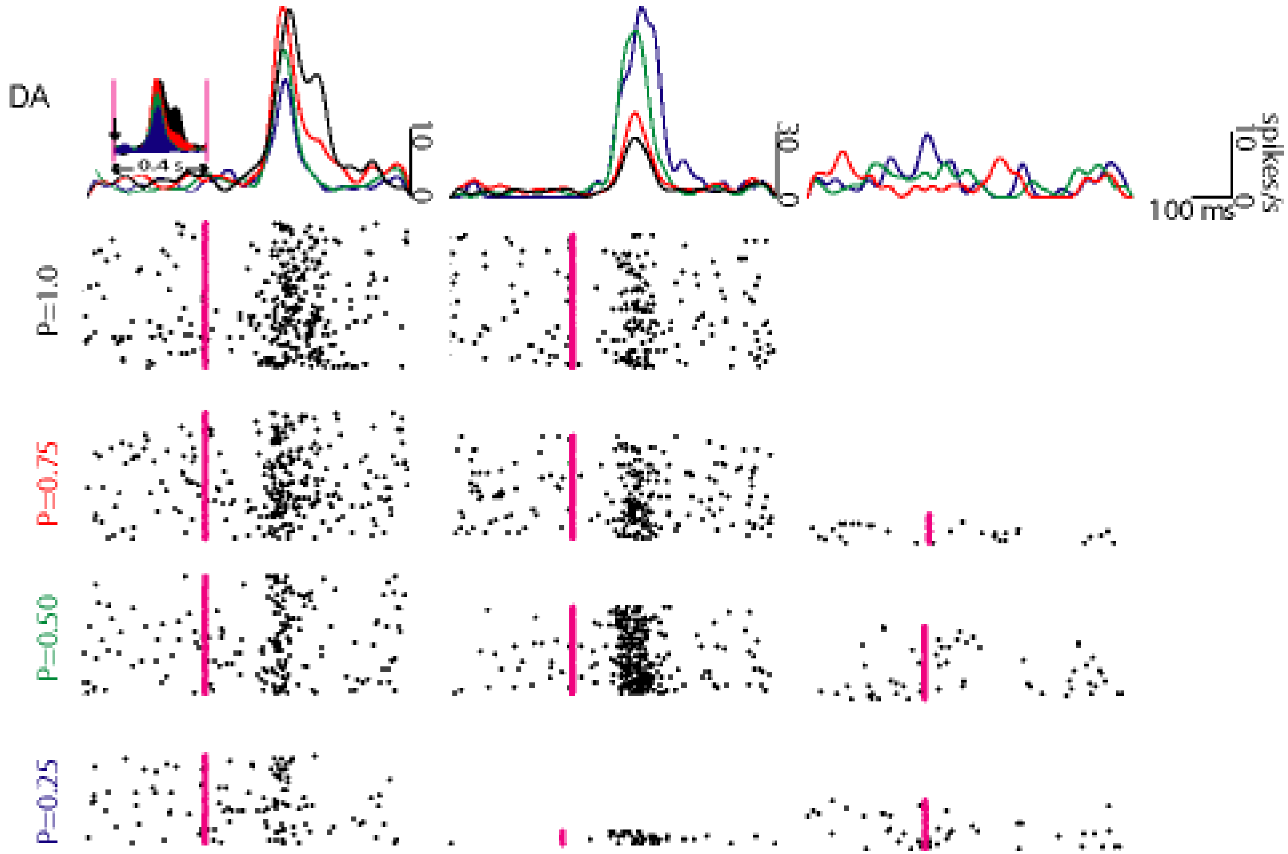




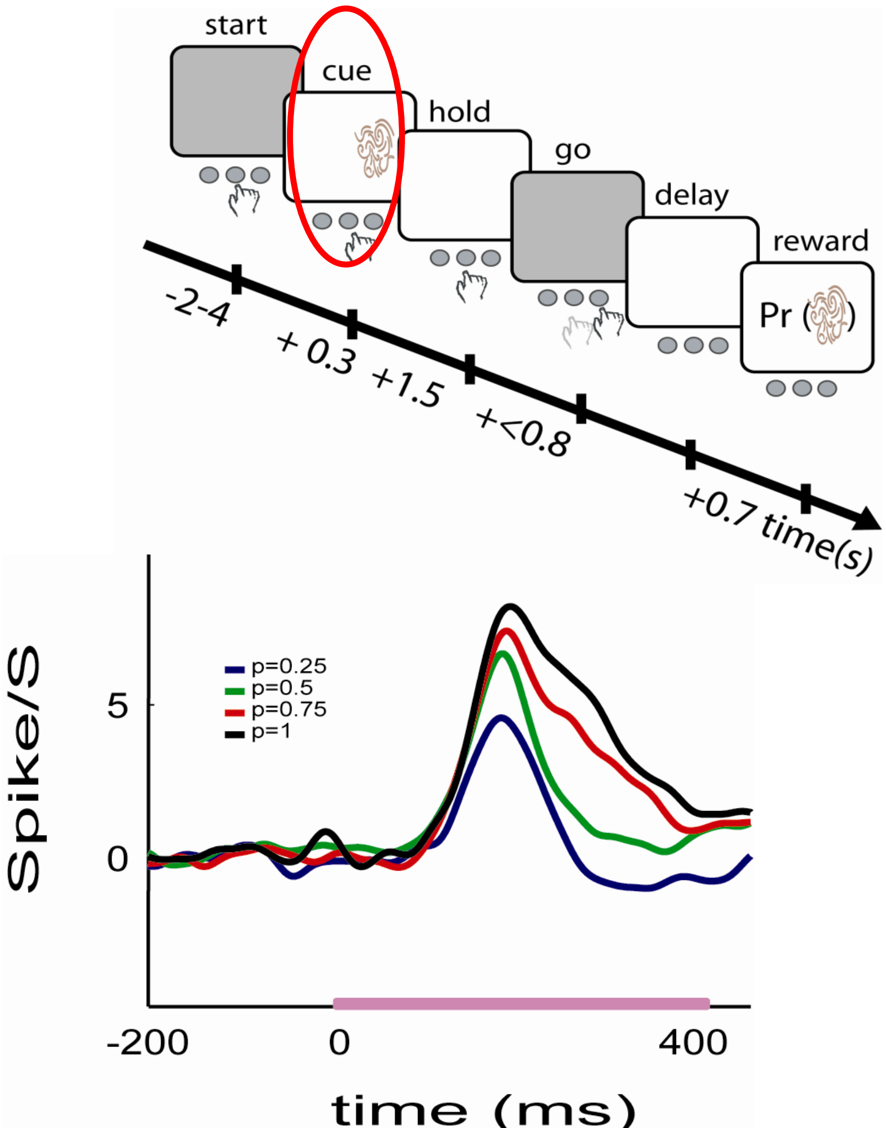
$$\delta_t = \gamma V_{t+1} - V_t + r_t$$

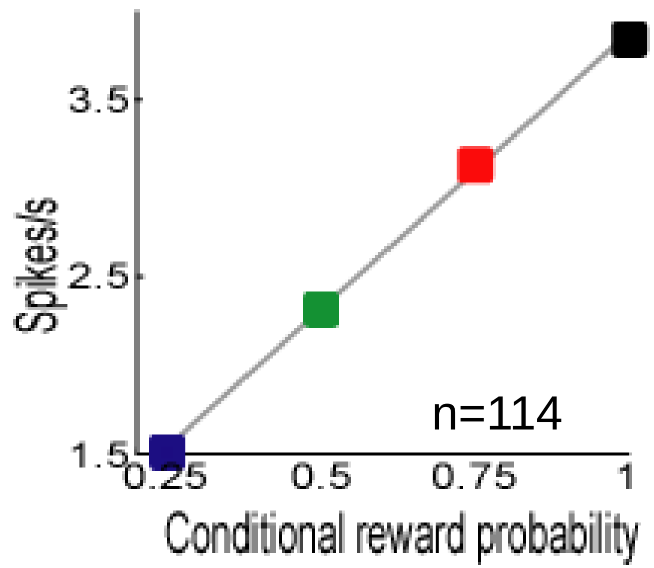
Morris et al., Neuron 43(1): 133-143, 2004

DA response

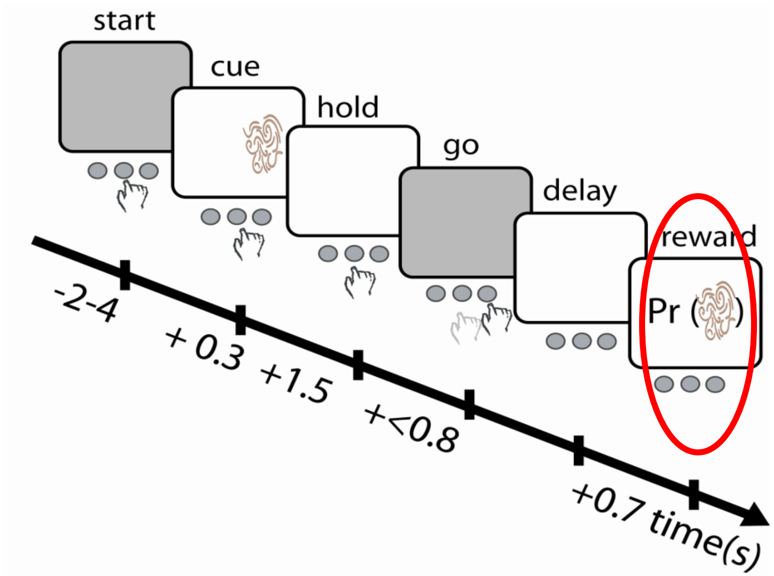


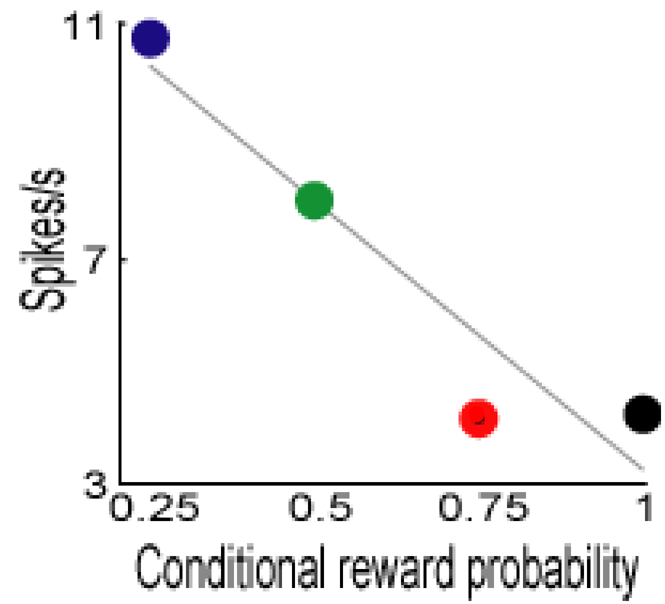
Dopamine population response-cue

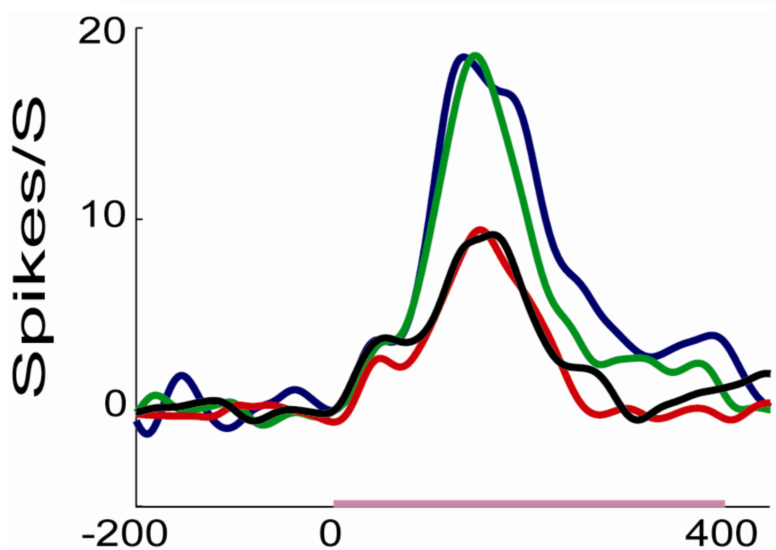




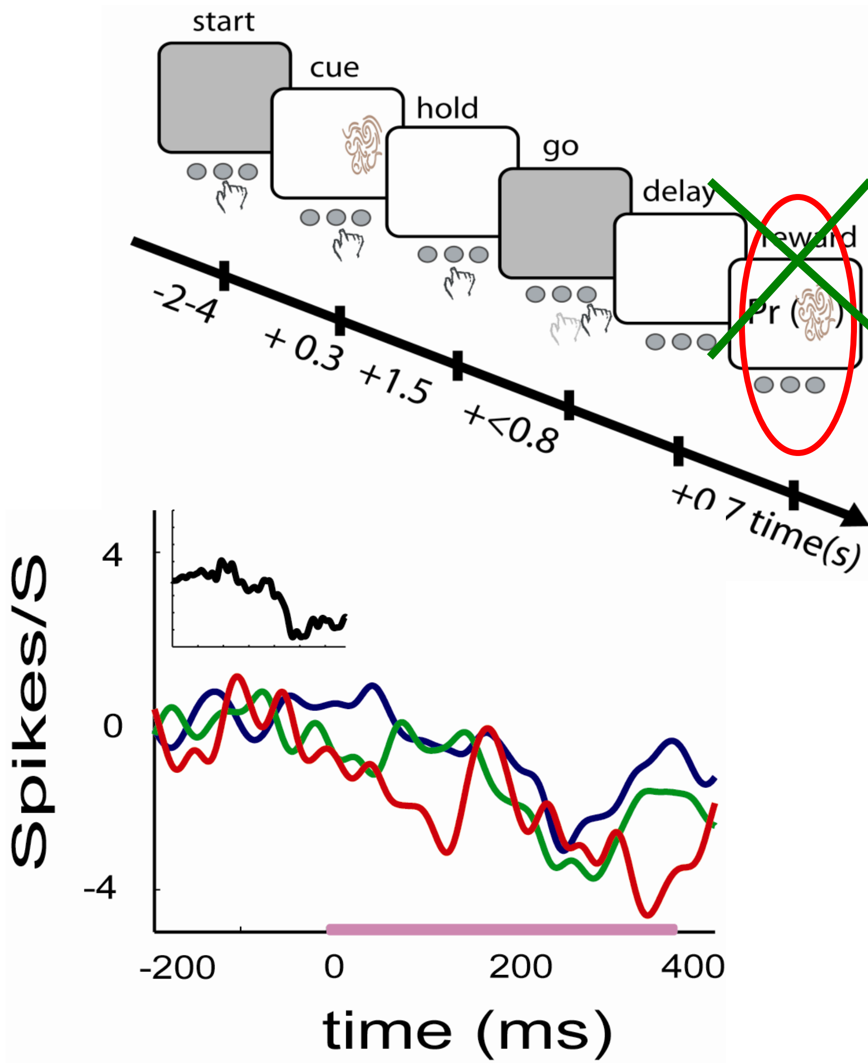
Dopamine population response-reward

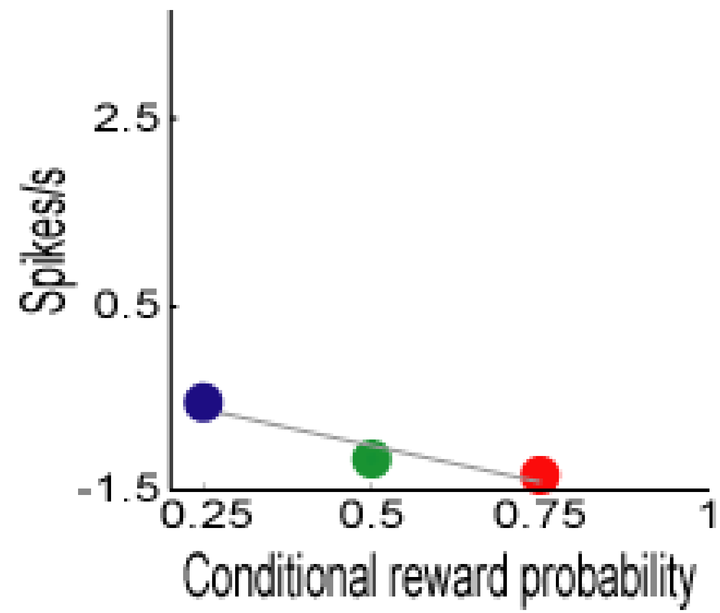






Dopamine population response – reward omission

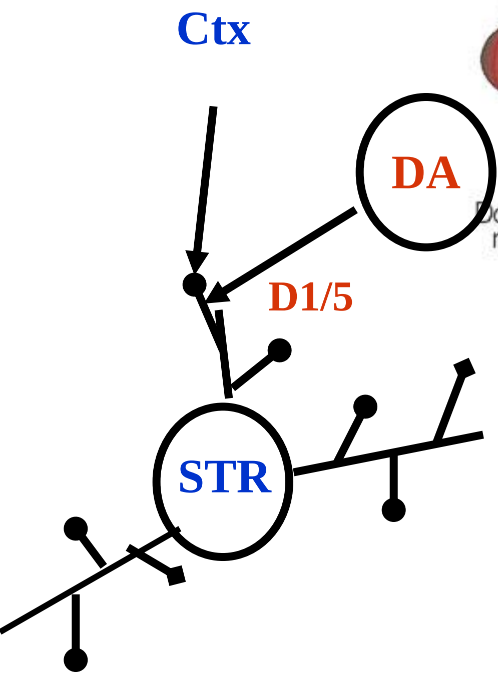


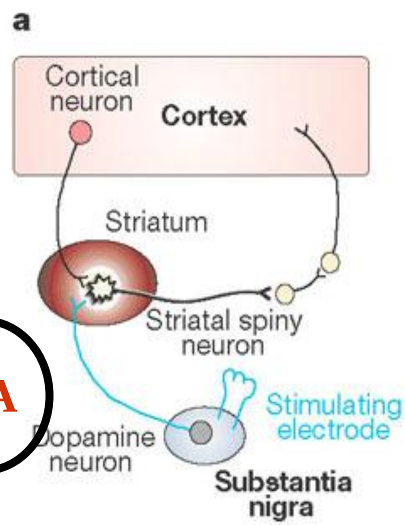


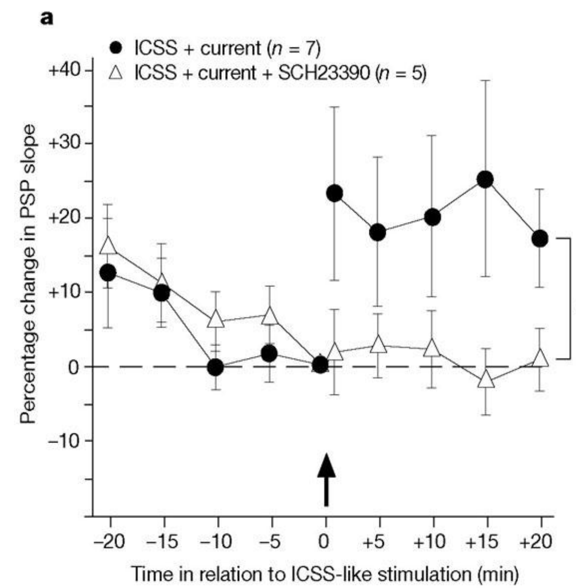
Instrumental conditioning - results

- 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)

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

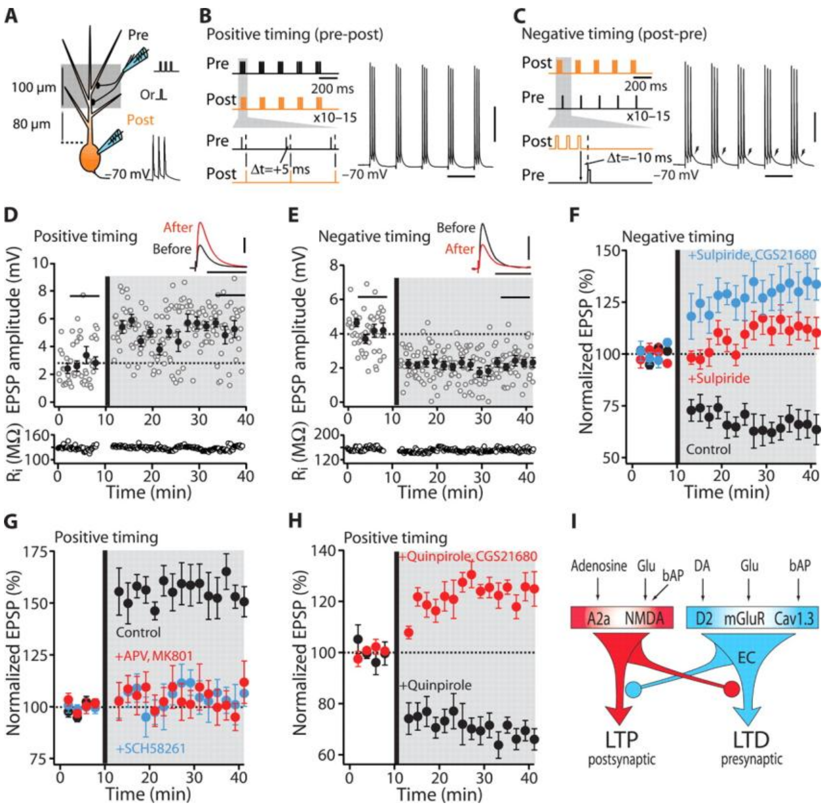


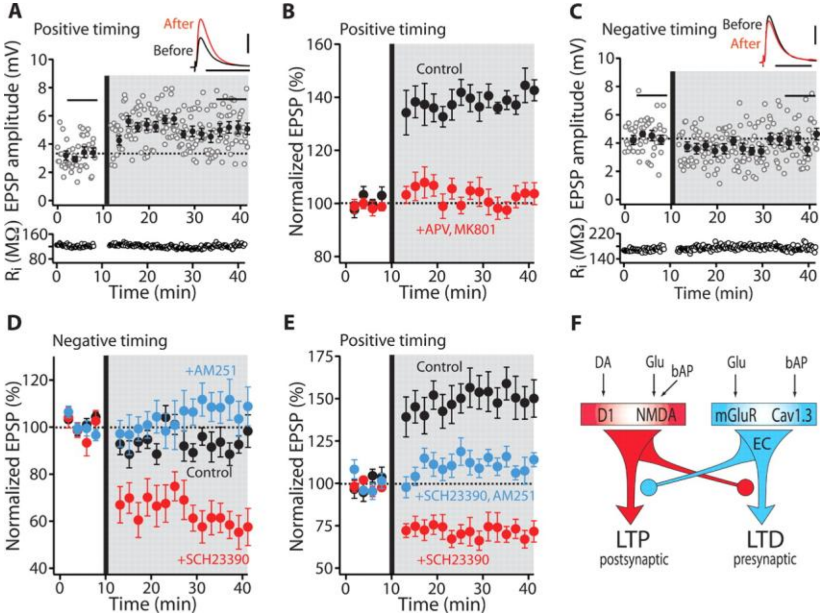




Reynolds et al, A cellular mechanism of reward-related learning Nature 413, 67 - 70 (2001)

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



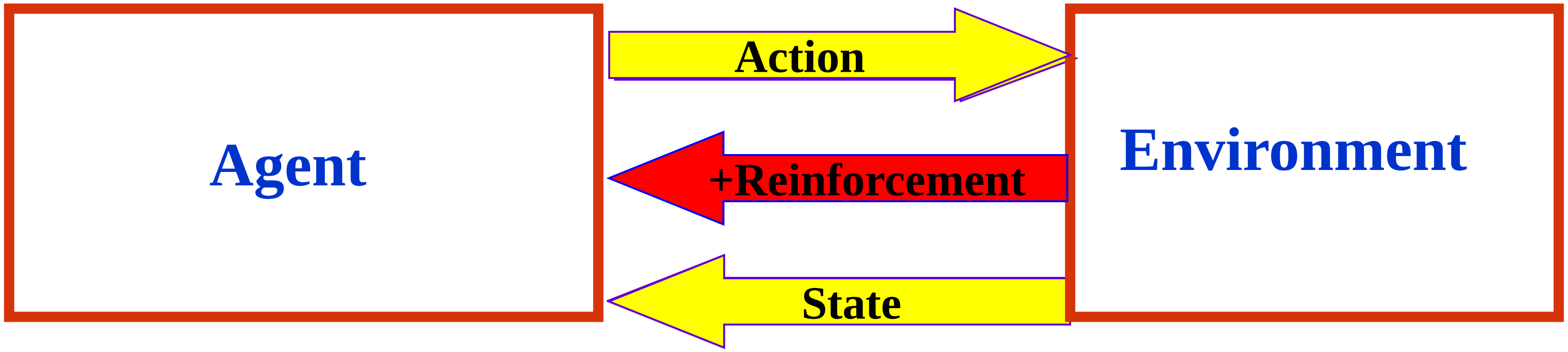


Shen et al., Science 321:848-851 2008

Facts to remember 2

- DA neurons provide a TD error signal
- To the cortico (state) striatal (action) synapses
- And DA modulates synaptic plasticity

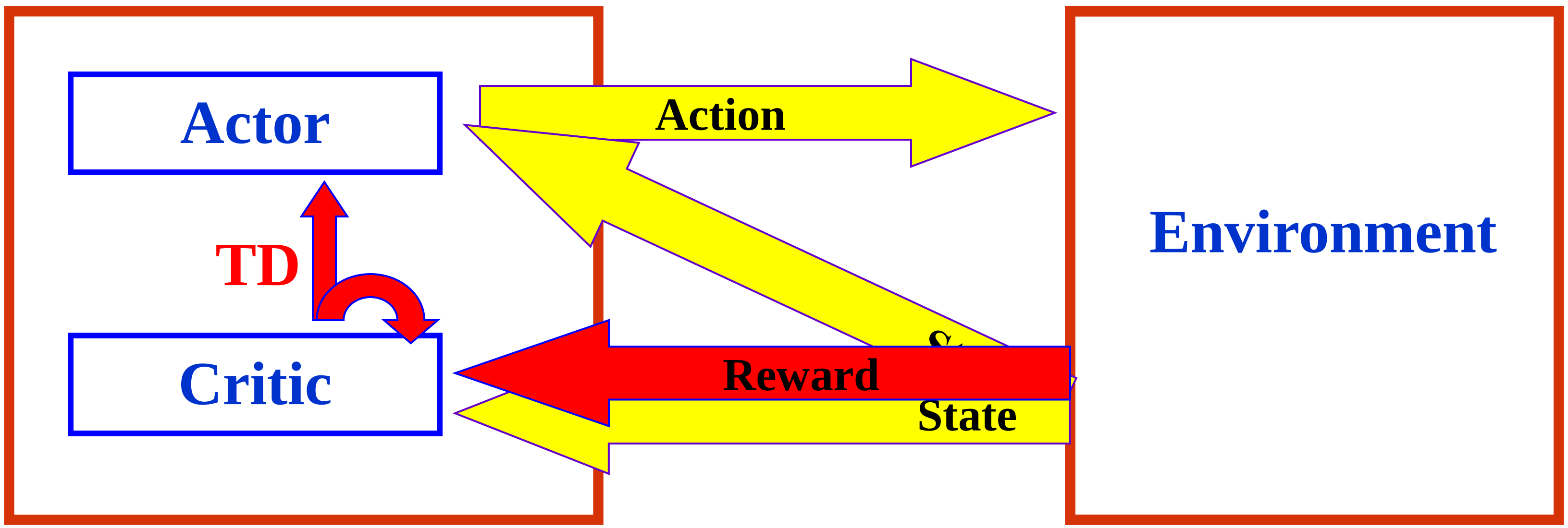
Control - Adding action



The agent has to:

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

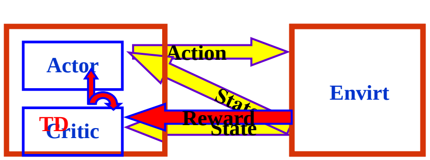
Solution 1: actor/critic networks

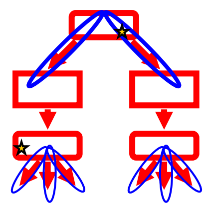


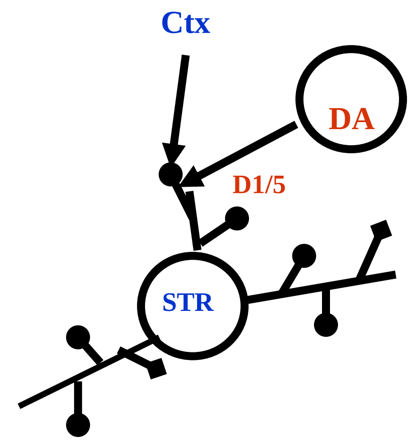
How can the dopamine signal contribute to decision behaviour?

- Long term policy-shaping effect

  through synaptic plasticity

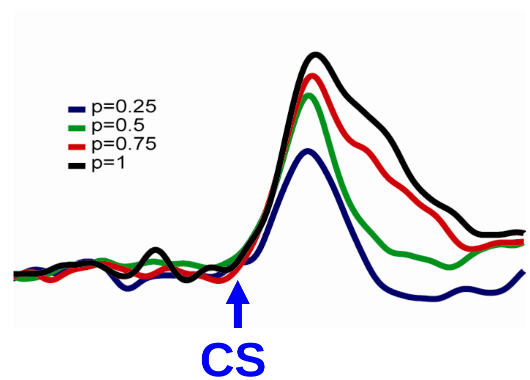


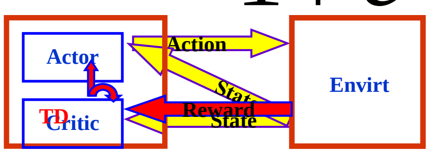




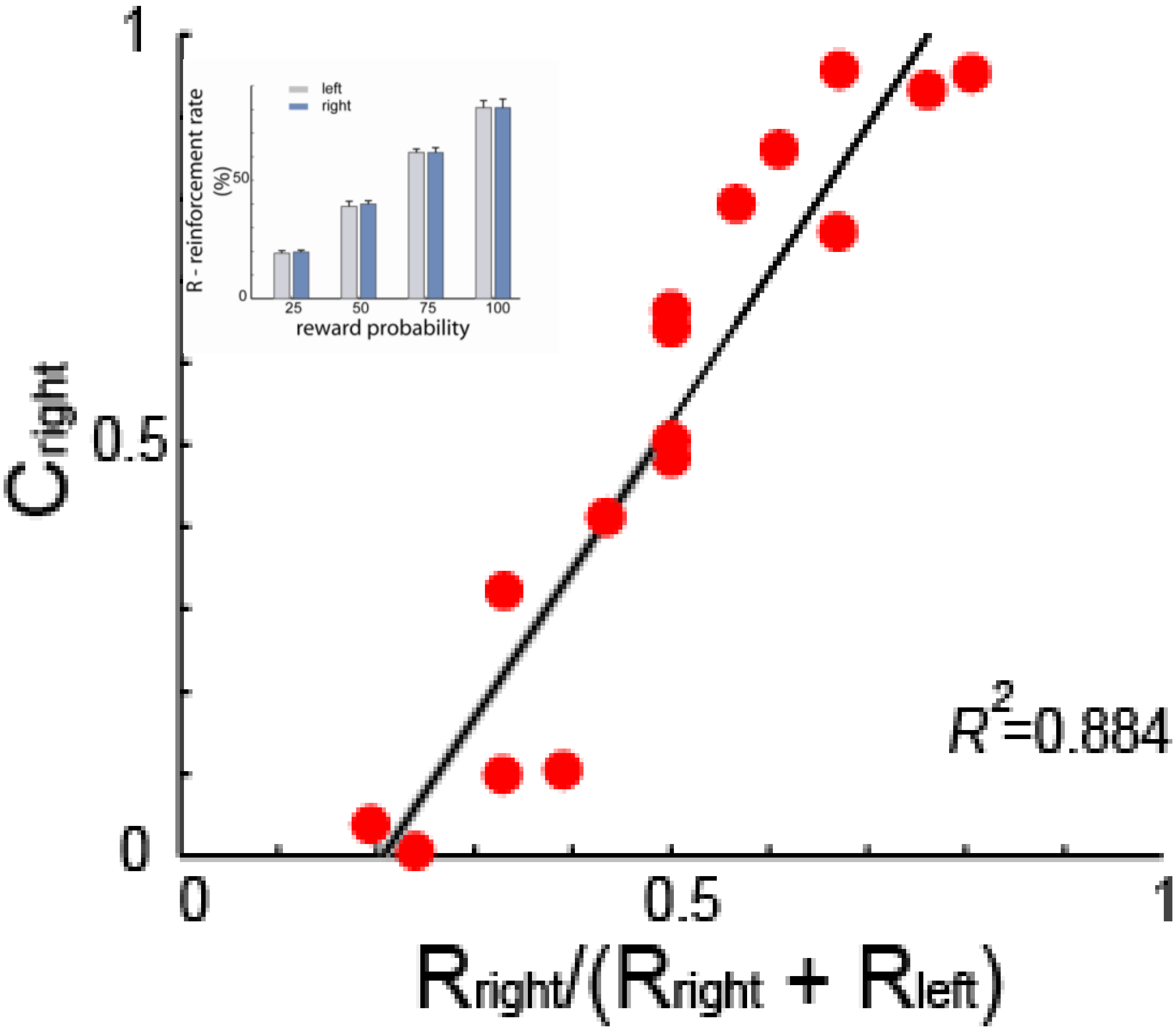
- Immediate effect on action

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

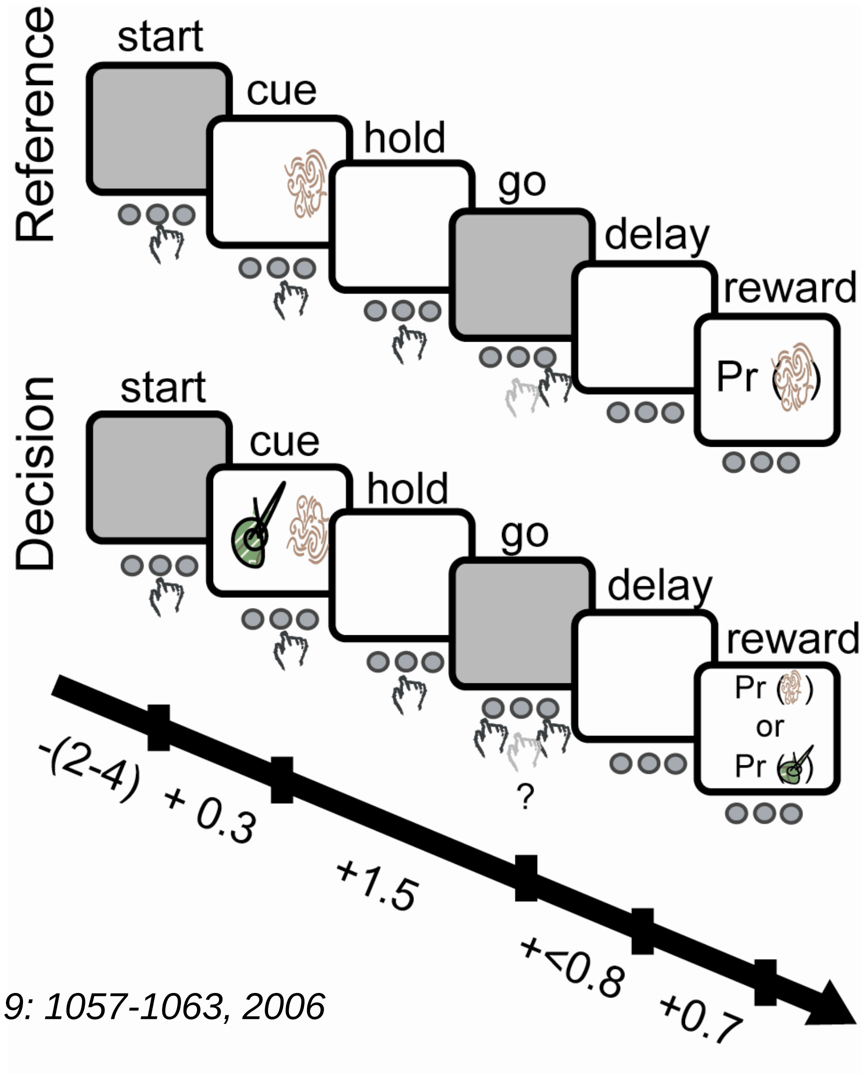




Monkeys' decisions: probability matching

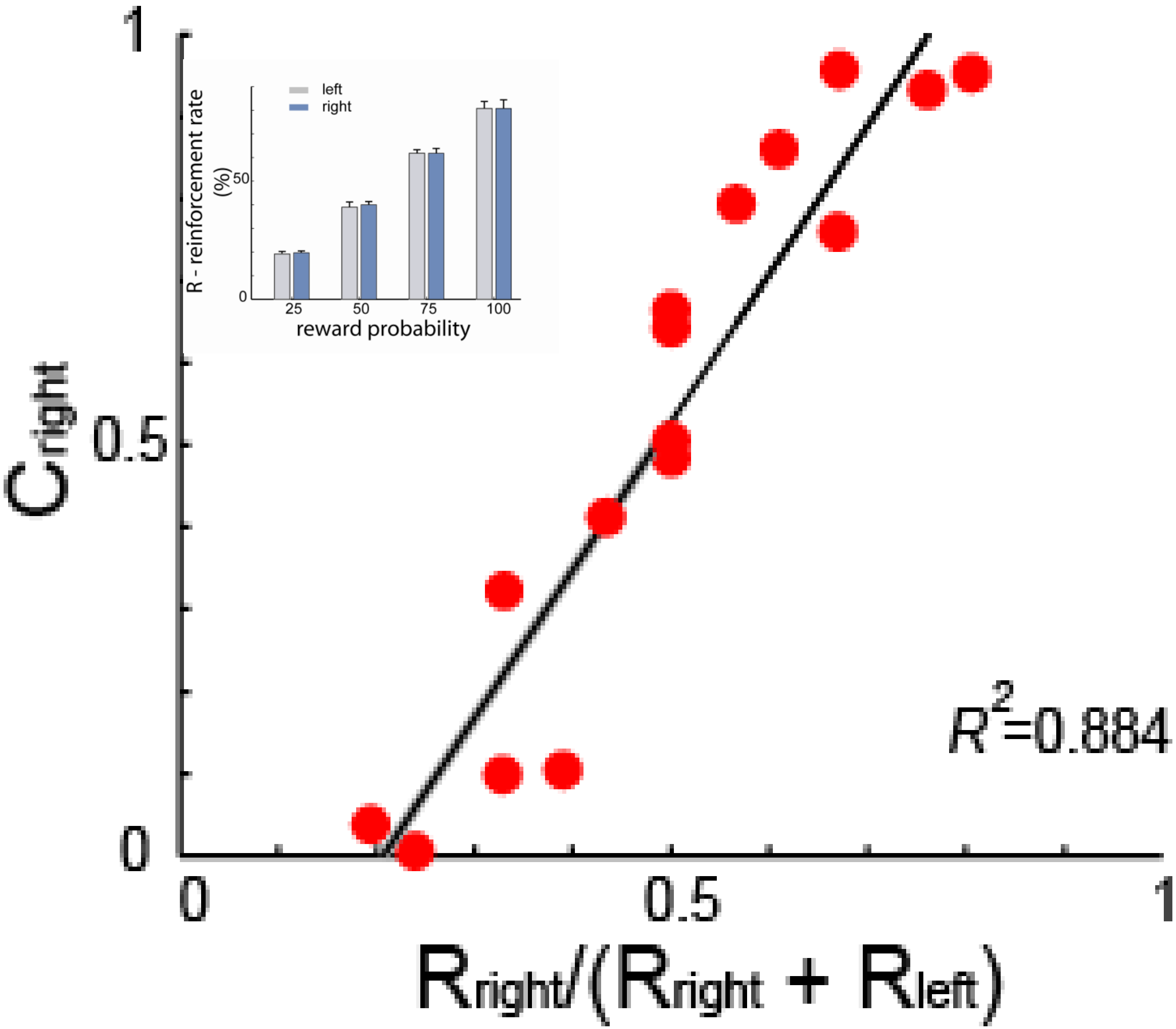


The two armed bandit task

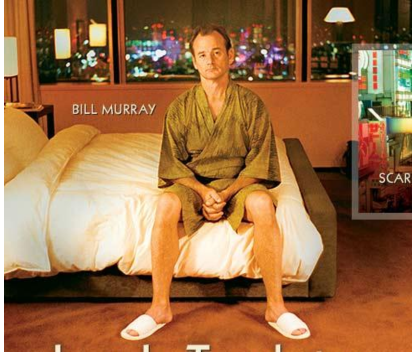


Morris et al., Nature Neurosci 9: 1057-1063, 2006

Monkeys' decisions: probability matching



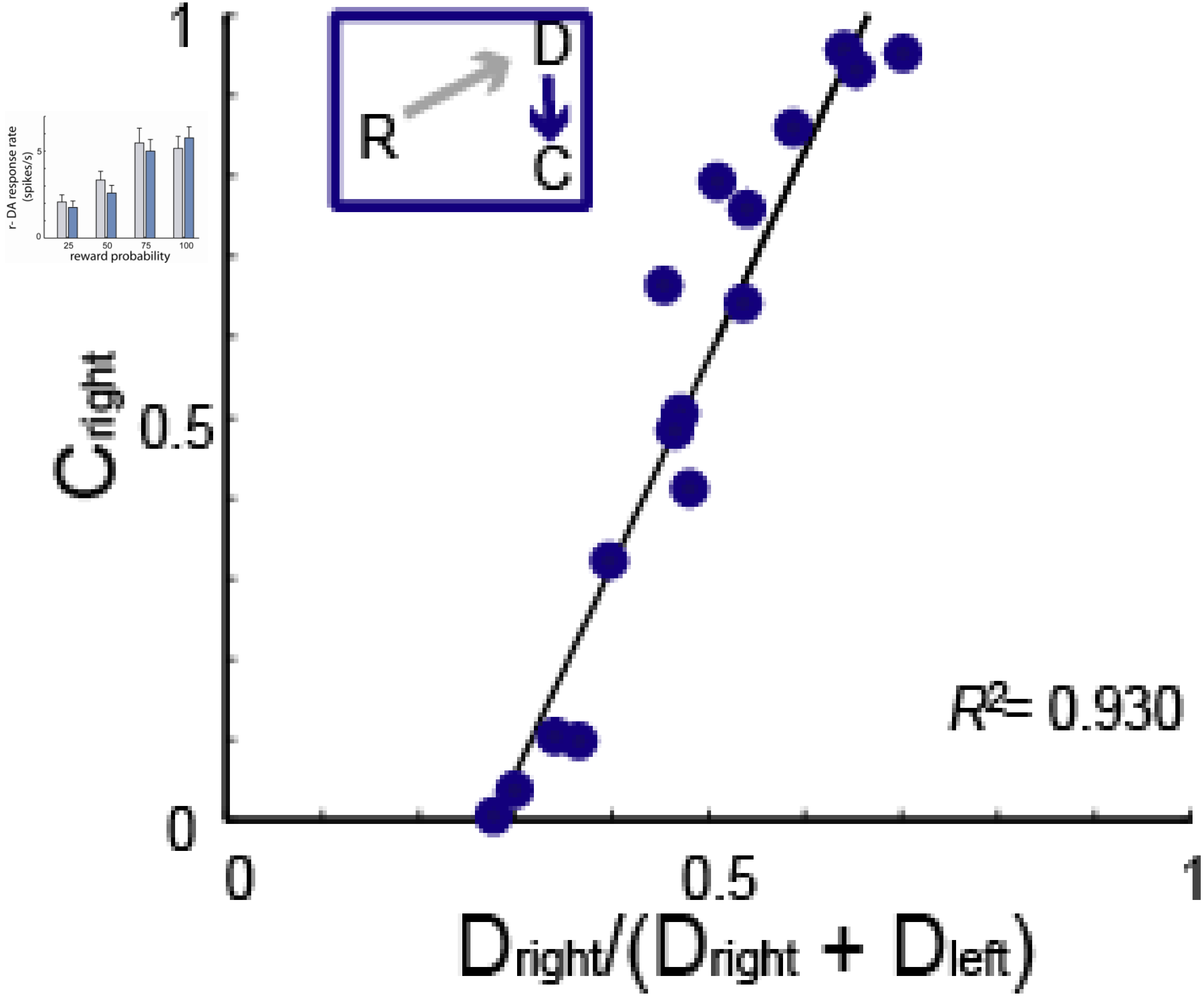
Lost in translation?



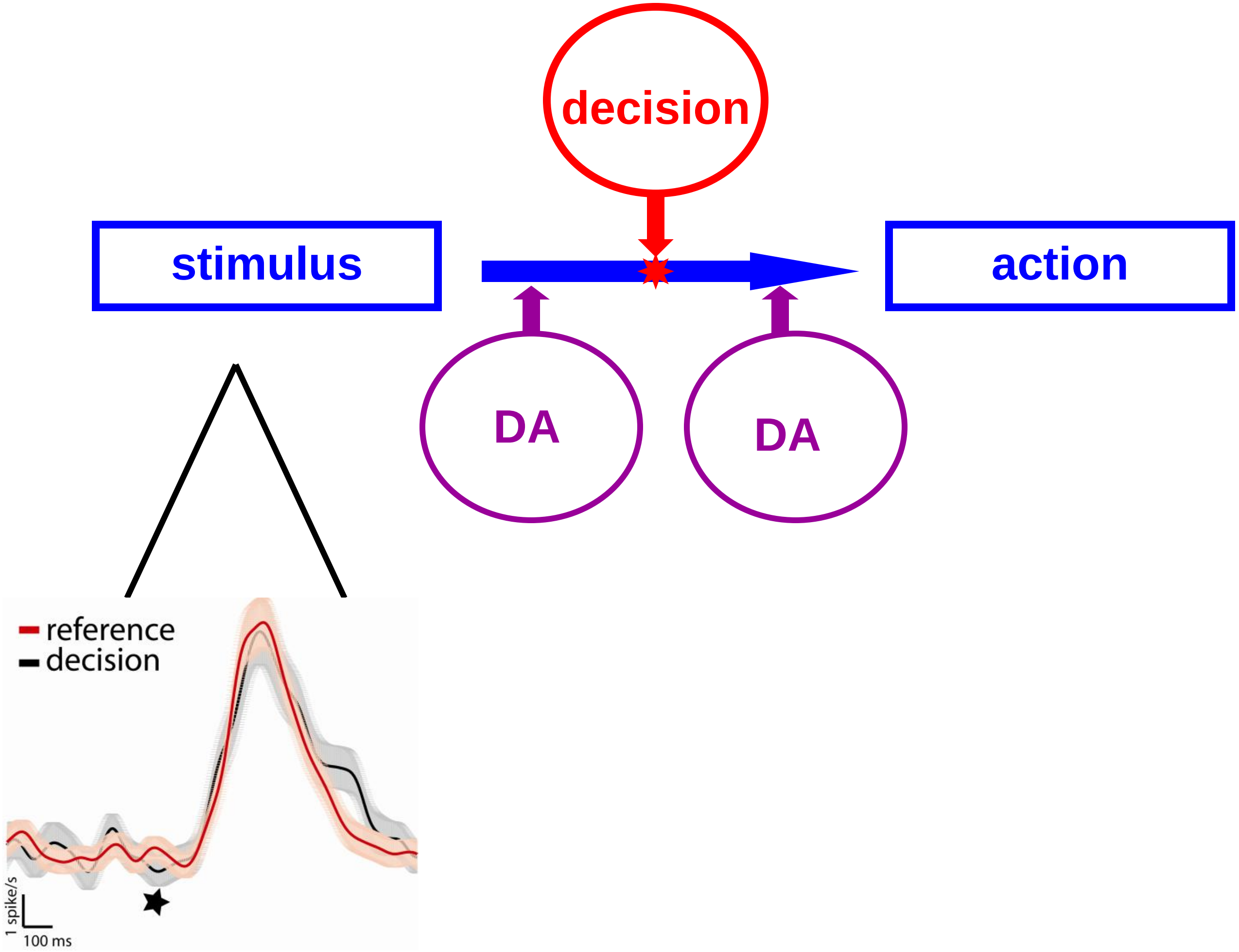
reward → behaviour

dopamine response → plasticity in action circuits

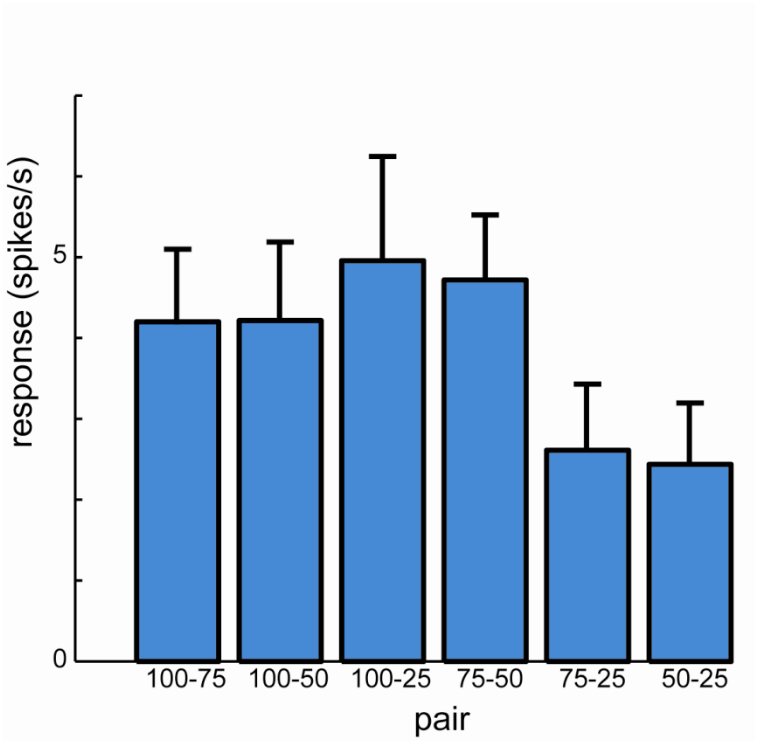
Monkeys' decisions: shaping by dopamine

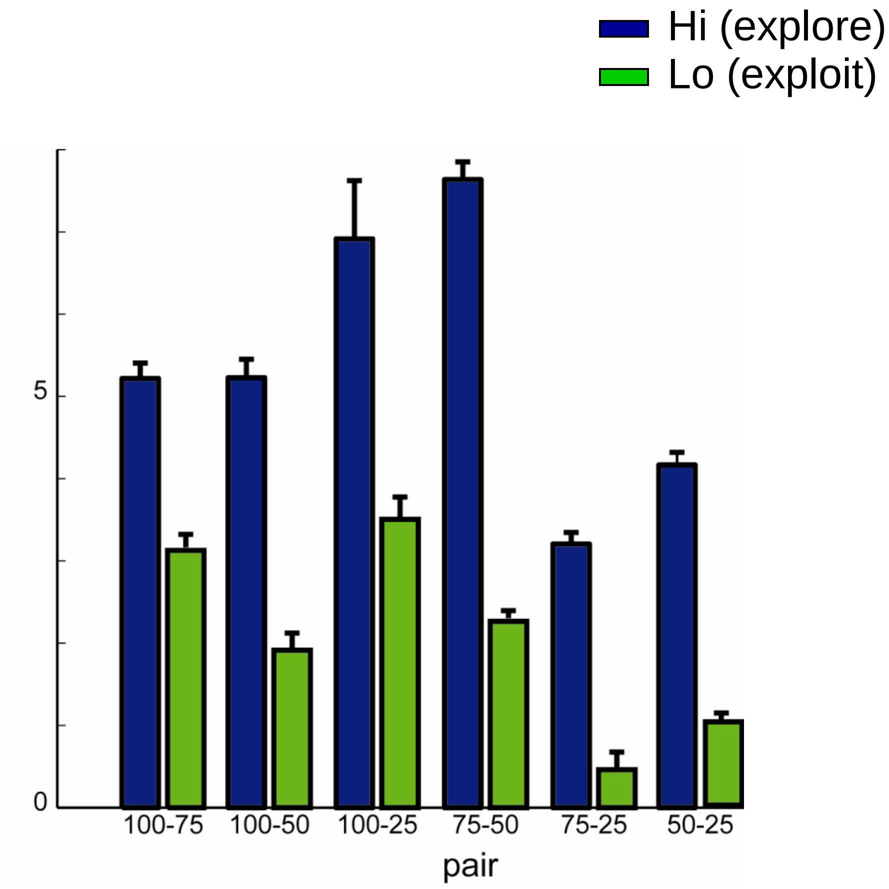


Dopamine neurons during decision



Are DA neurons aware of future choice





The learning is of state-action values