

# Dopamine & Reinforcement Learning

*Cognitive Computational Neuroscience - SOW-MKI48*



[hannekedenouden.ruhosting.nl](http://hannekedenouden.ruhosting.nl)



[h.denouden@donders.ru.nl](mailto:h.denouden@donders.ru.nl)



@HannekedenOuden



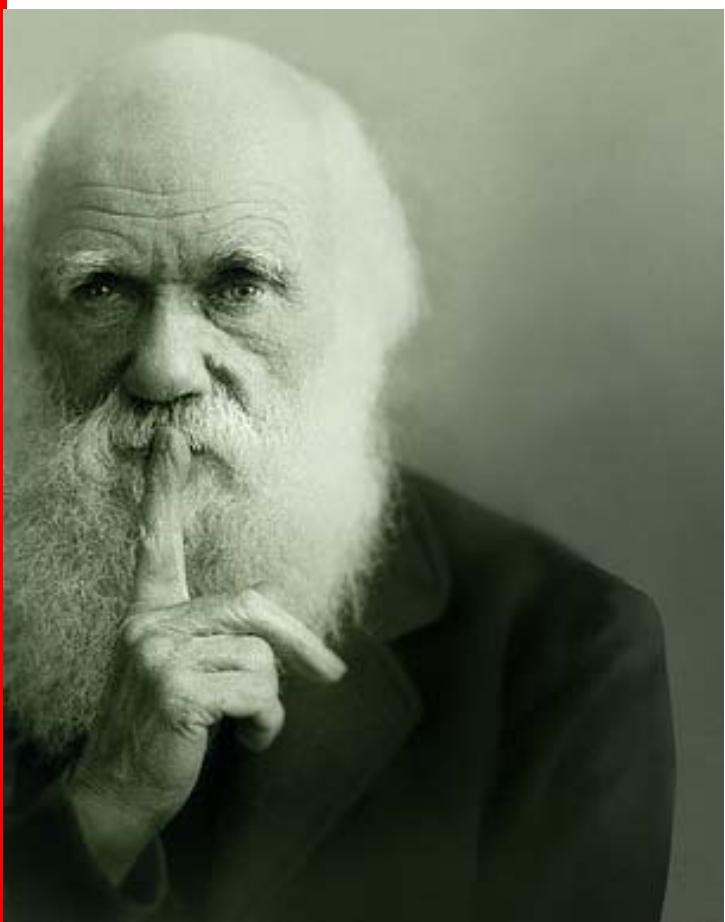
# Outline

- processing reward
- introduction to dopamine
- learning about rewards
  - error-driven learning
  - neural basis
- learning actions
  - TD learning
  - law of effect
  - second-order reinforcement
  - multiple decision systems

# What drives our behaviour?

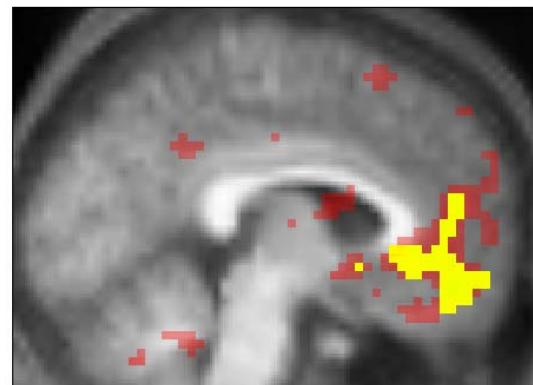
- goal: survival and procreation
- implementation: seek rewards

reward = something that, all other things being equal, **we seek to maintain, repeat, or enhance**

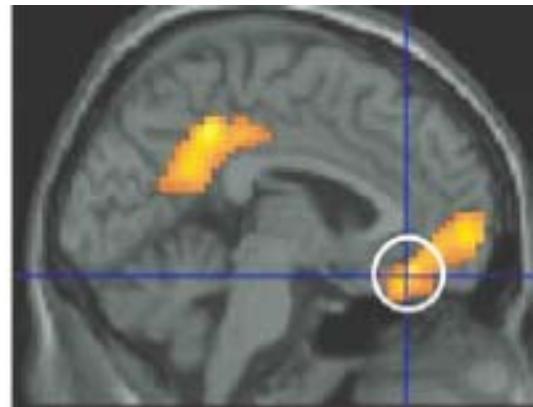


# Key Neural Reward Areas

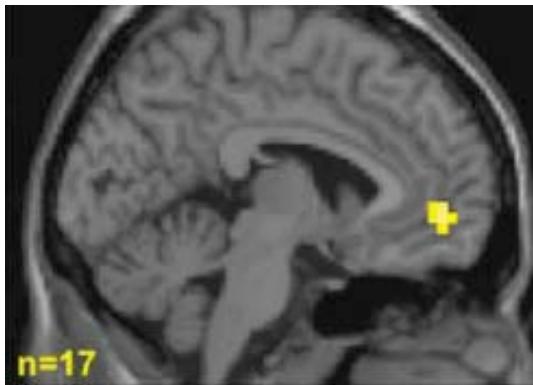
ventromedial prefrontal /  
orbitofrontacortex



Money  
Daw et al 2006

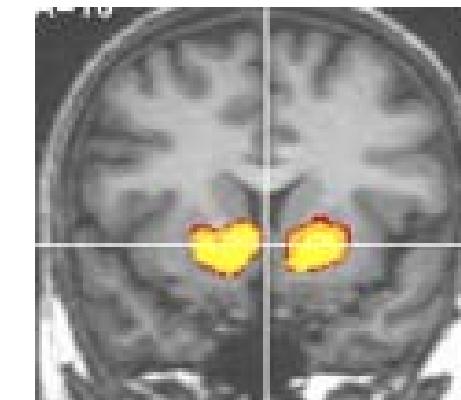


Attractive faces  
O'Doherty et al 2003

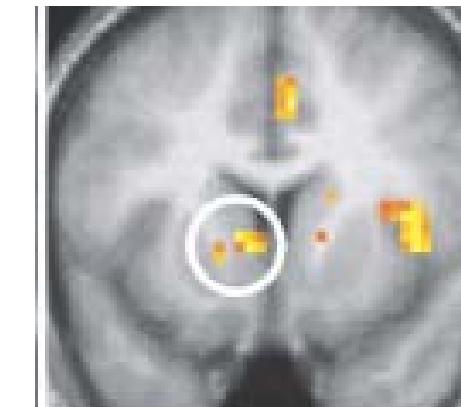


Liking of Coke vs. Pepsi  
McClure et al. 2004

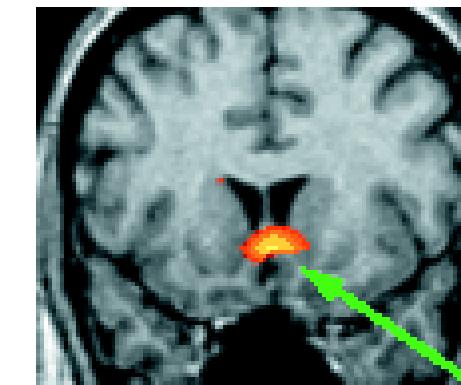
striatum



Money  
Kuhnen & Knutson 2005



Food odors  
Gottfreid et al 2003



Juice  
Berns et al 2001

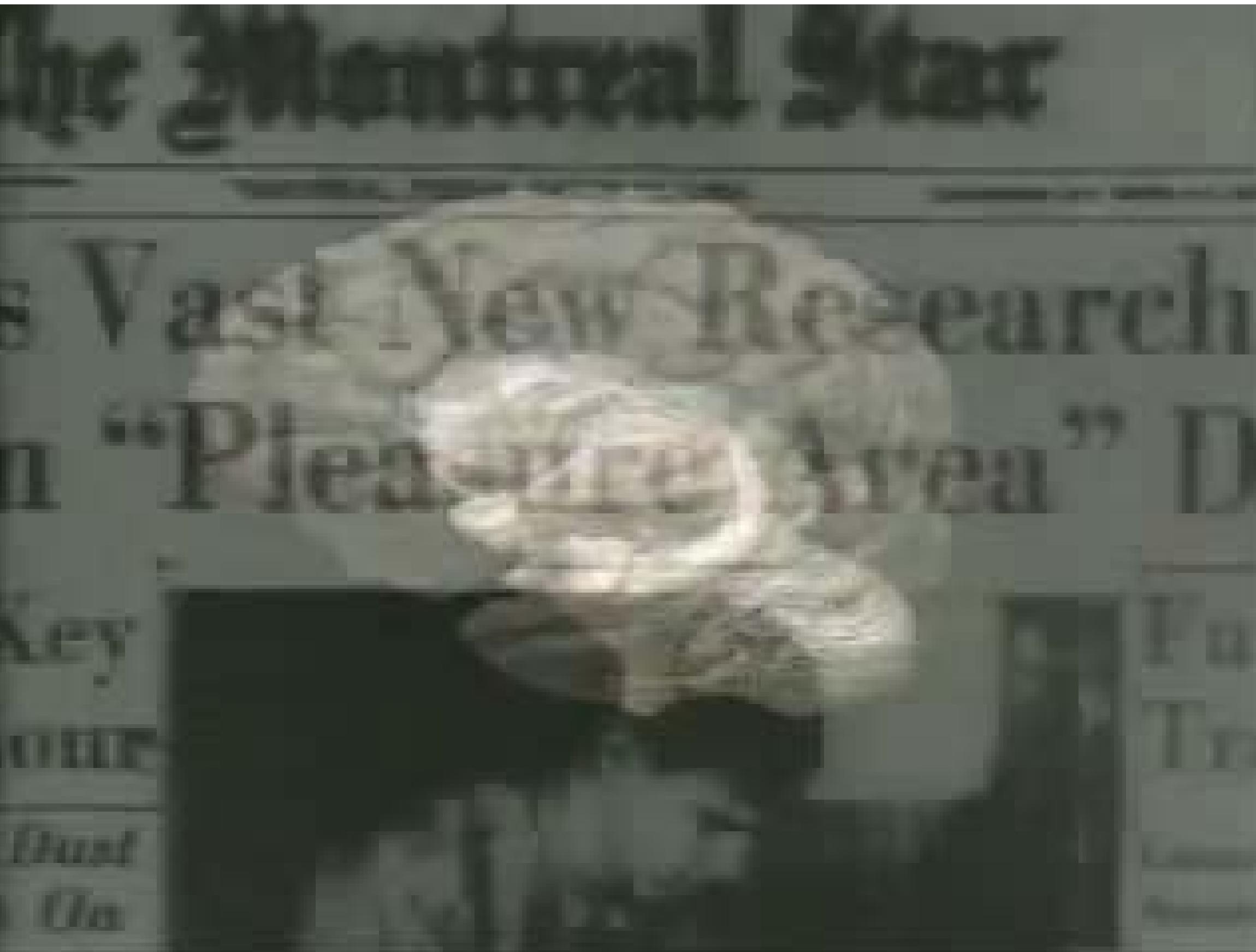
commonality of responding across reinforcers suggests generalised appetitive function





What is the neural implementation of this  
generalised appetitive function?

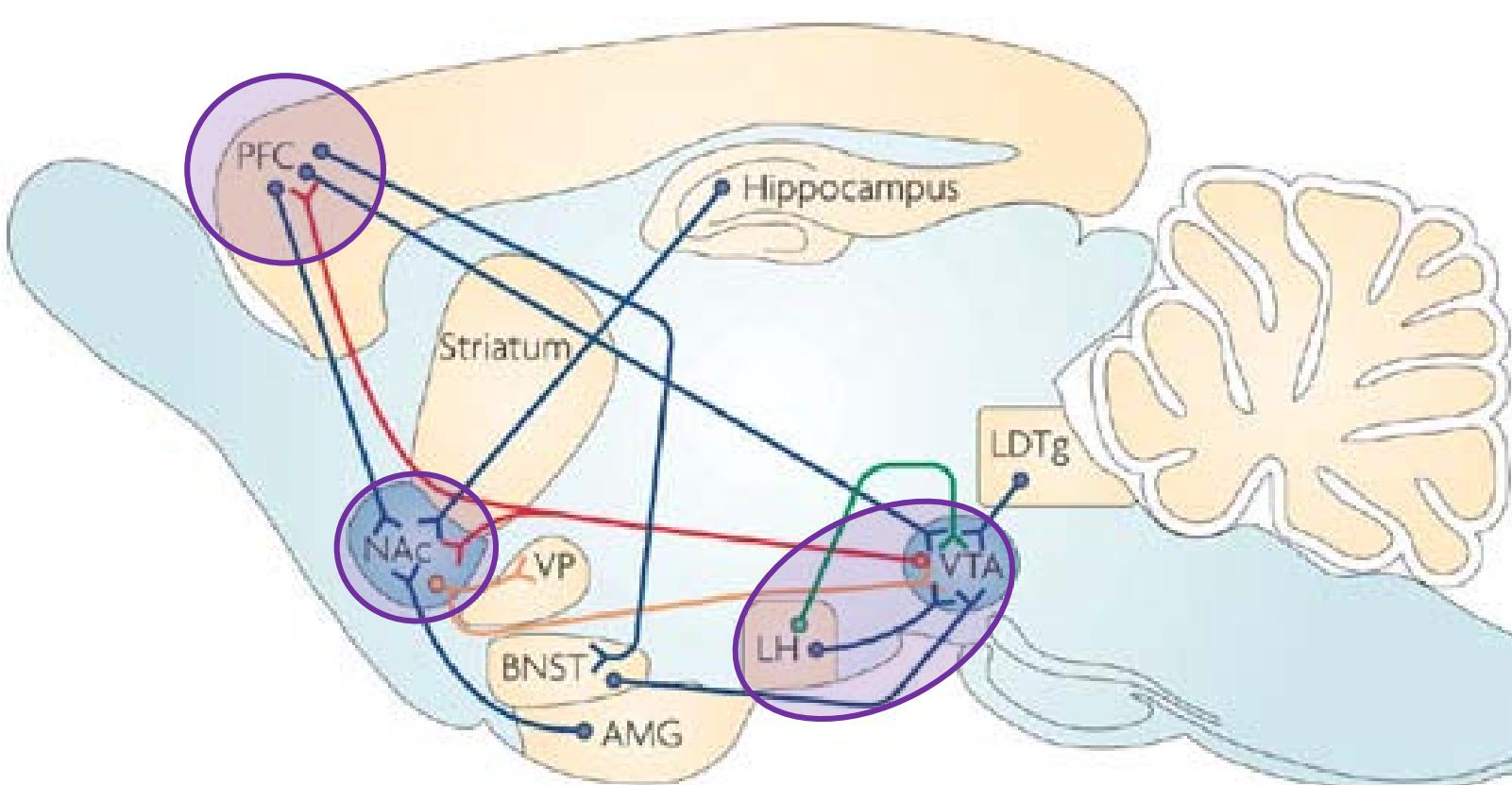
# Olds & Milner (1954)



# Brain stimulation rewards

Olds & Milner 1954:  
lateral hypothalamus

Other BSR sites:  
prefrontal cortex  
ventral striatum (NAc)  
ventral tegmentum



Rats pay high price to receive electrical stimulation (electric shocks, forgo food)

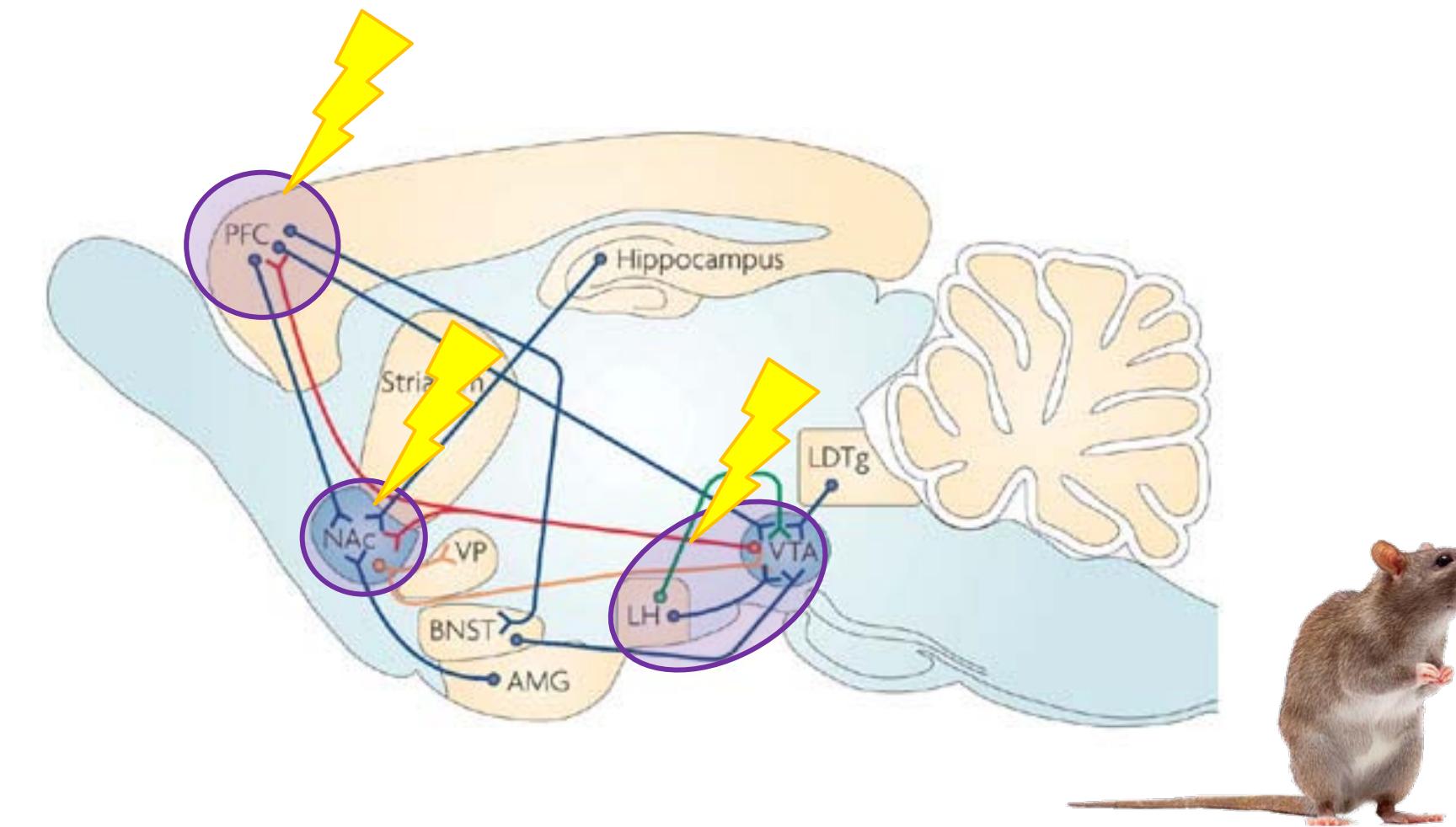
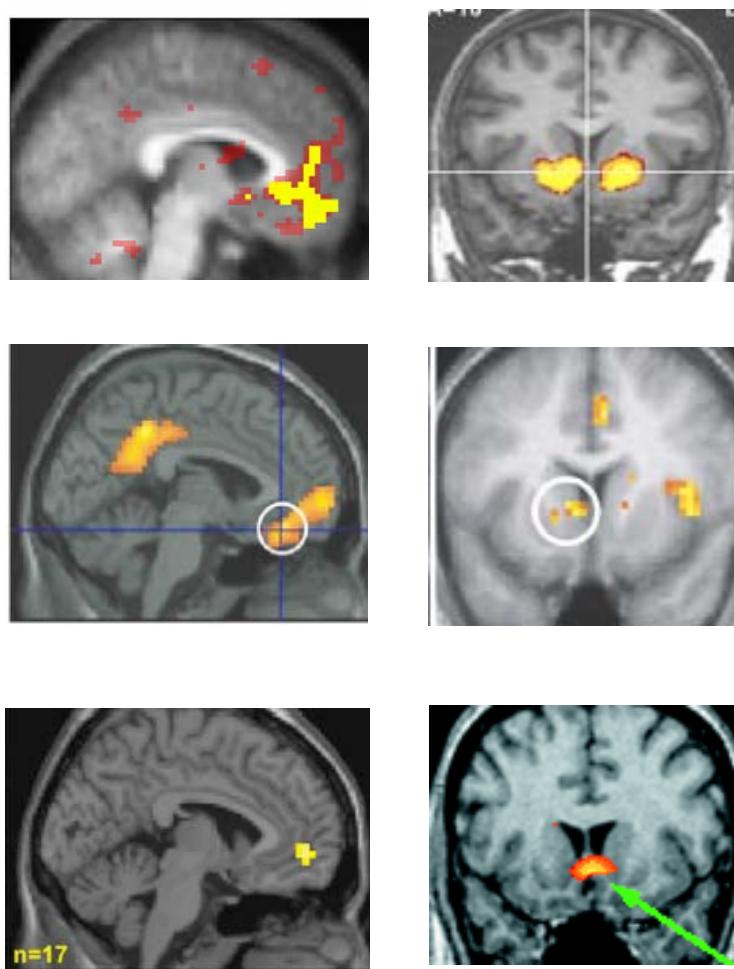
Stimulation serves as reinforcer even stronger than natural rewards

Since rats **seek to repeat** the stimulation, BSR sites were theorised as “reward centres”

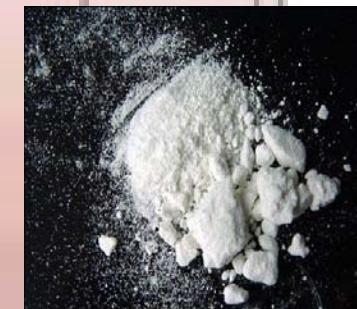
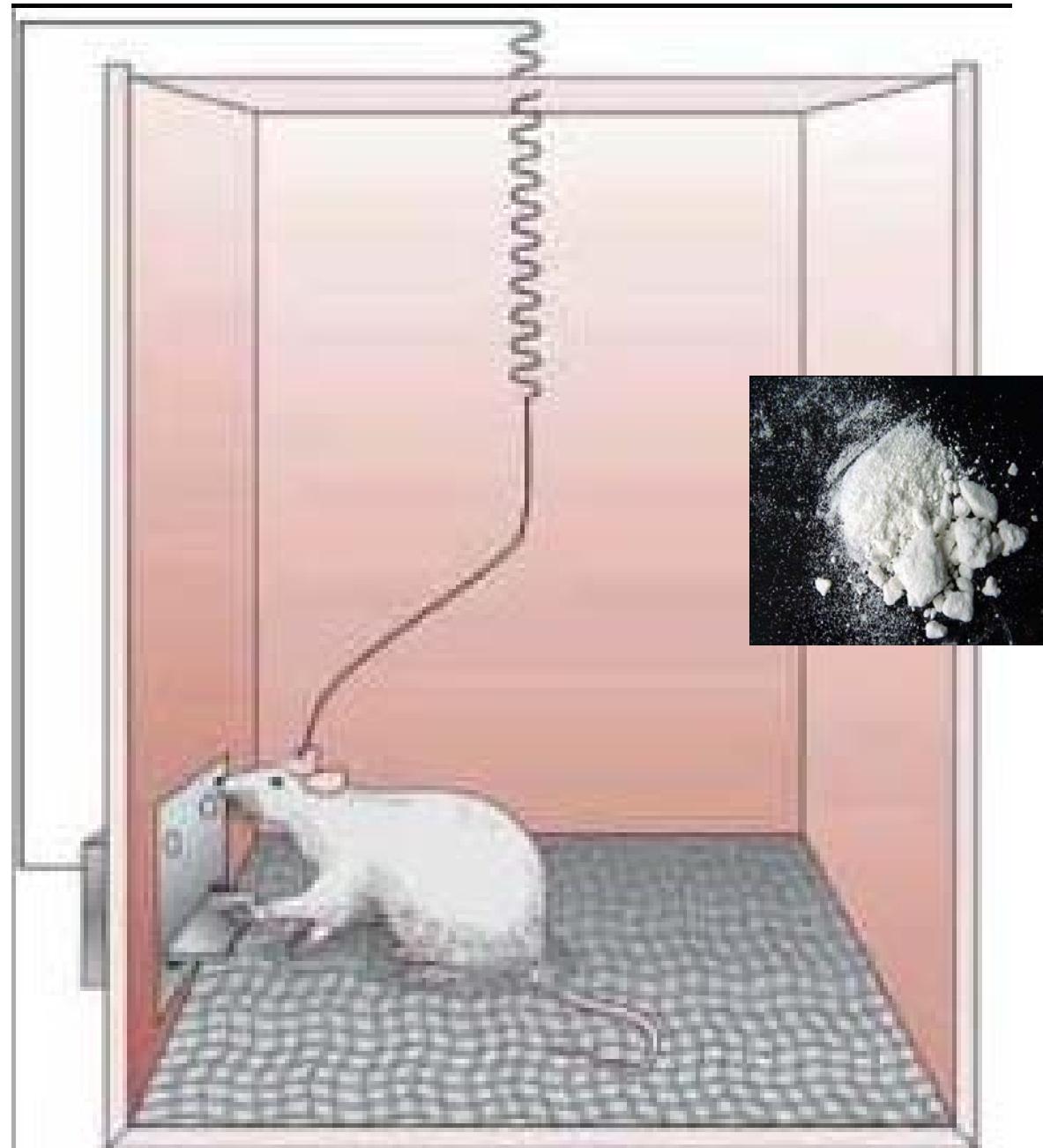
# Brain 'reward centres'

Ventral striatum and ventromedial prefrontal cortex

- In humans, are active when processing rewards
- In rodents, stimulation acts as a reward



# Neurochemical basis of brain stimulation rewards



drug injection is like BSR  
stimulation

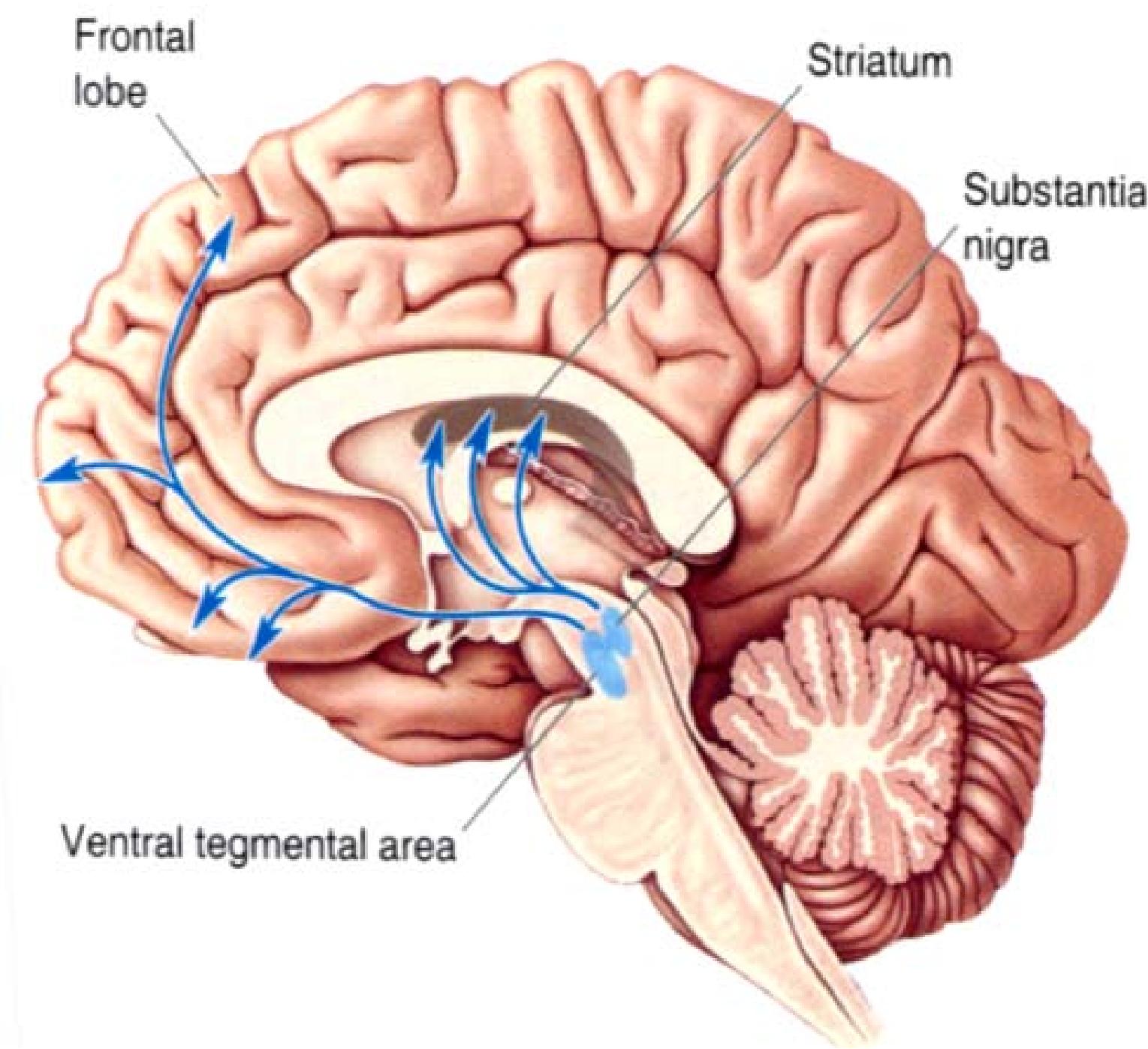
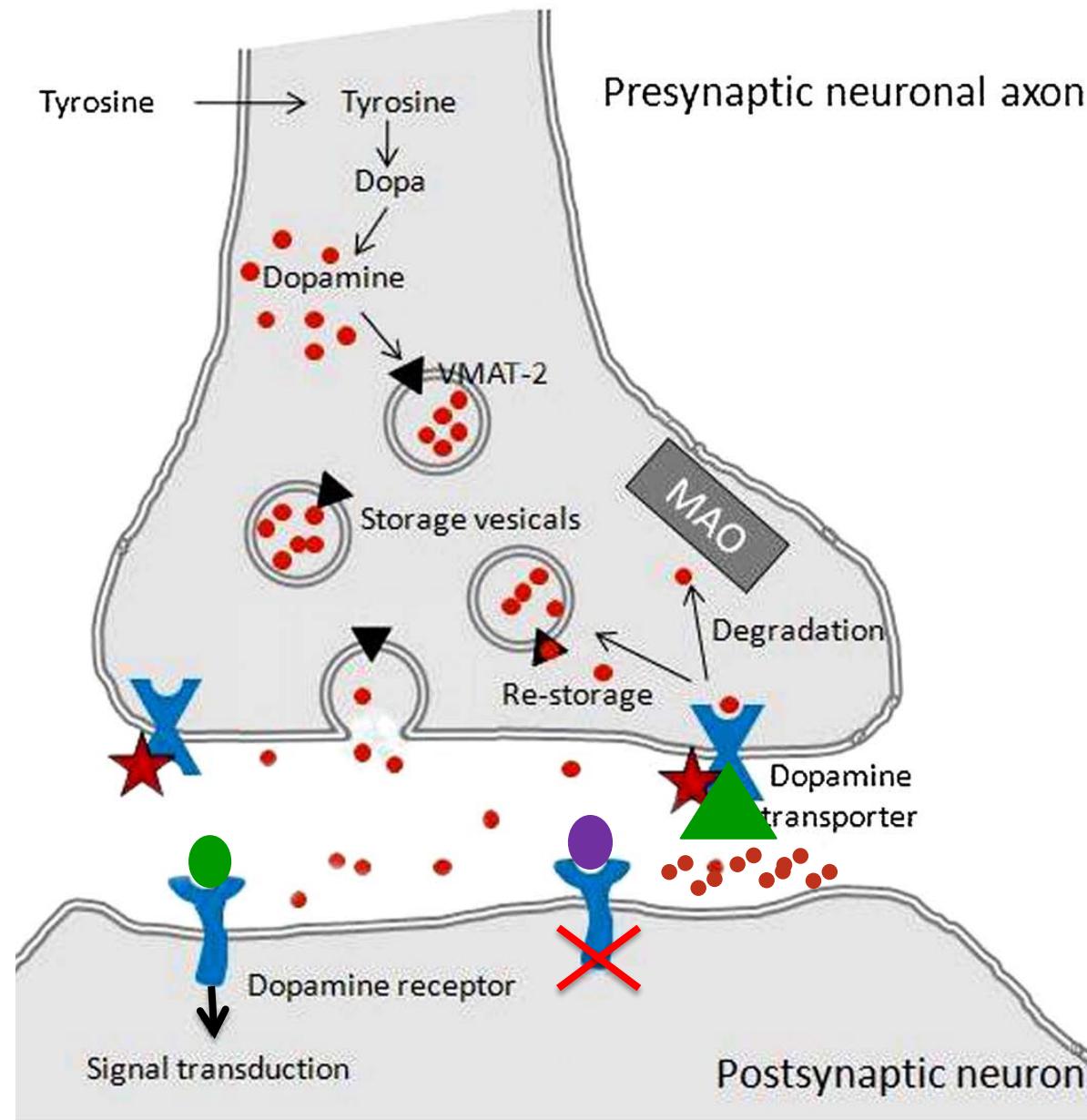
which drug?

# biology of dihydroxyphenylalanine (dopamine)

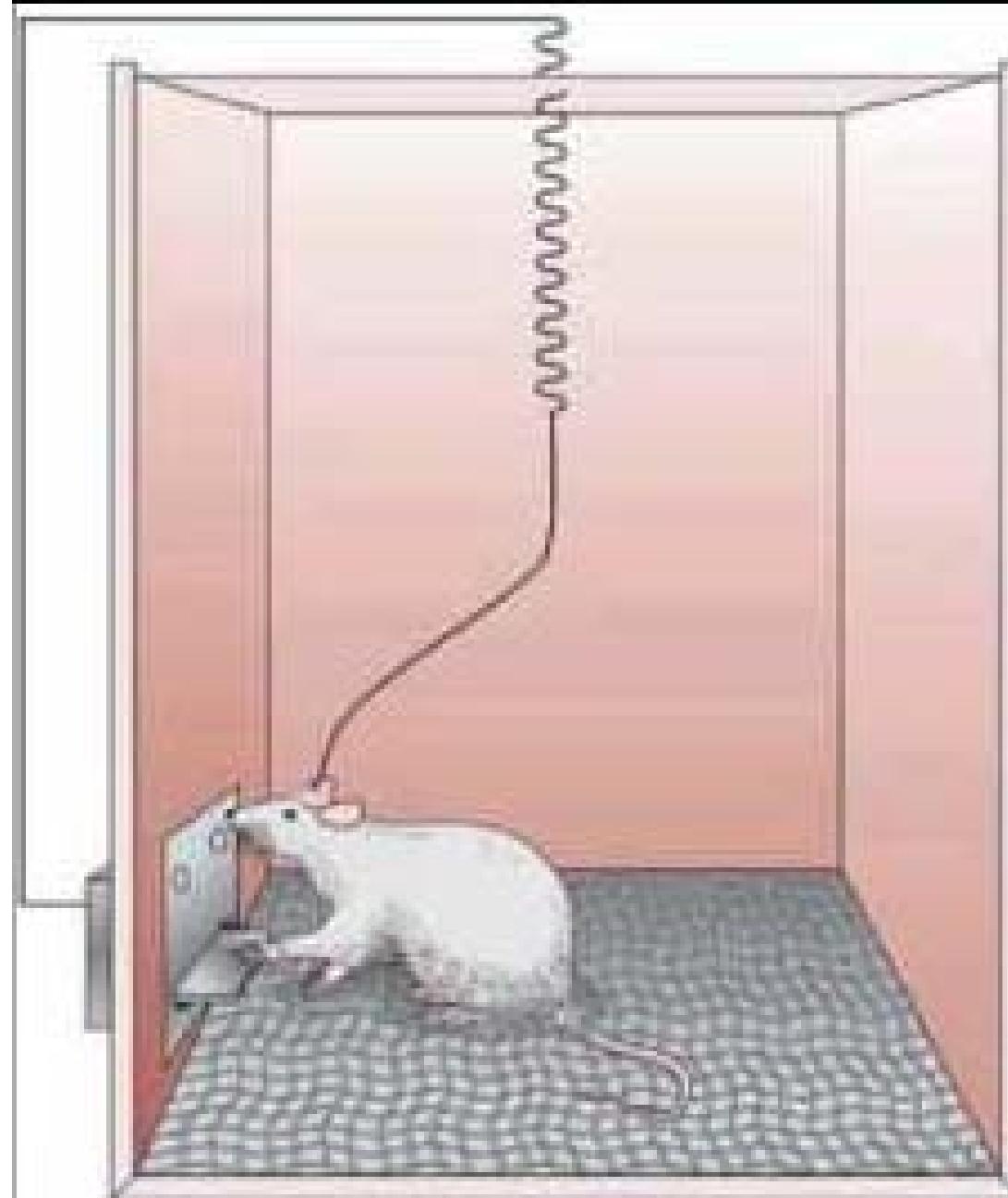
Agonist

Antagonist

Reuptake blockade



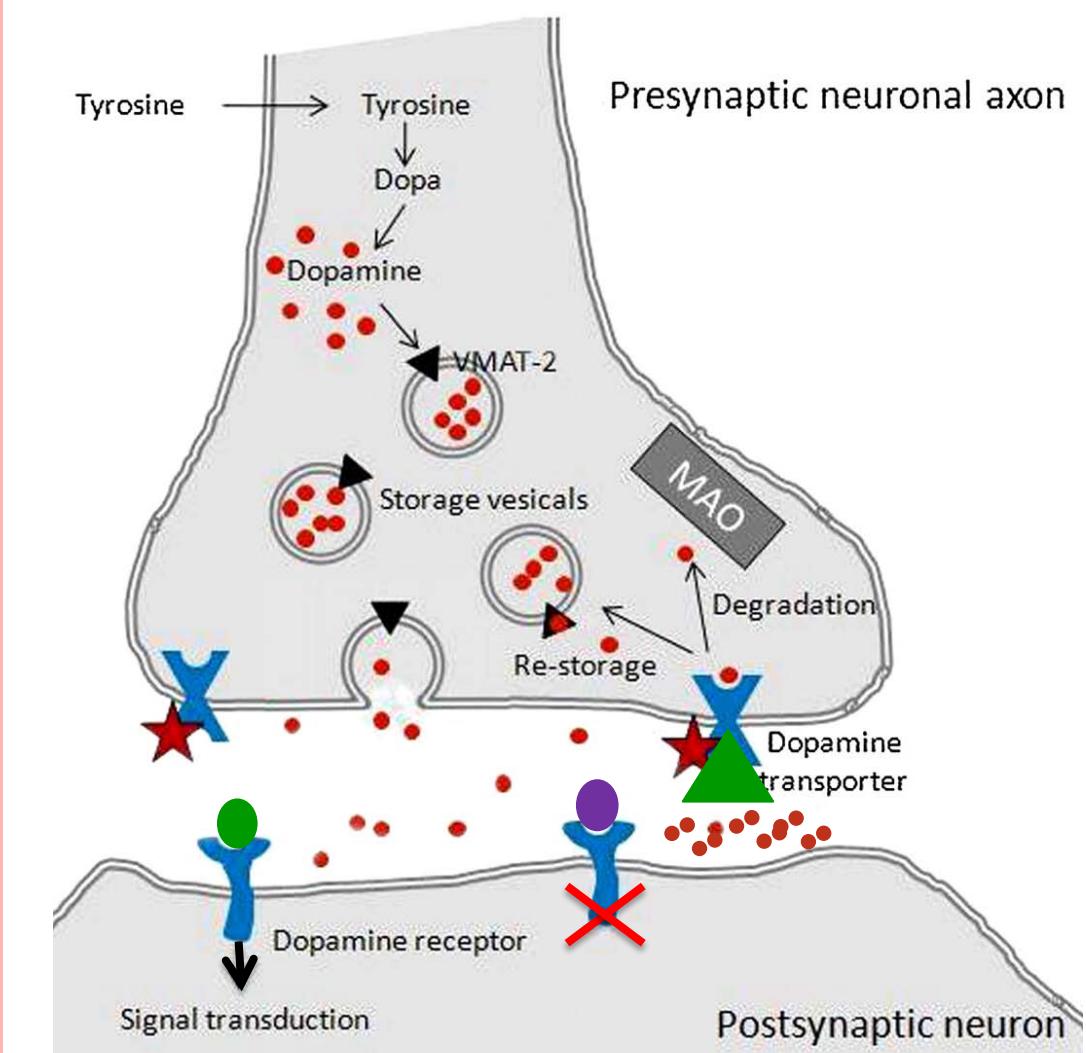
# Dopamine as 'reward neurotransmitter'?



drug injection is like BSR stimulation

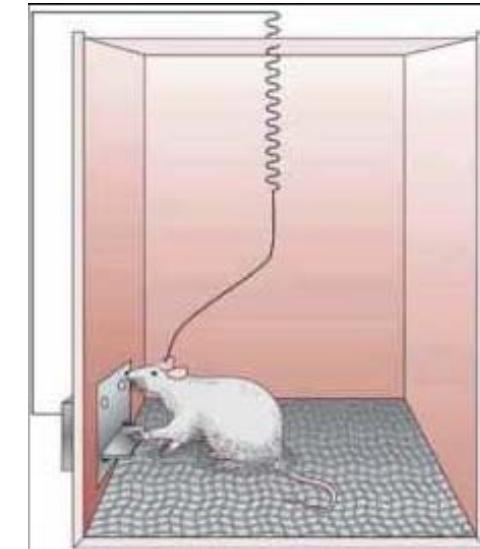
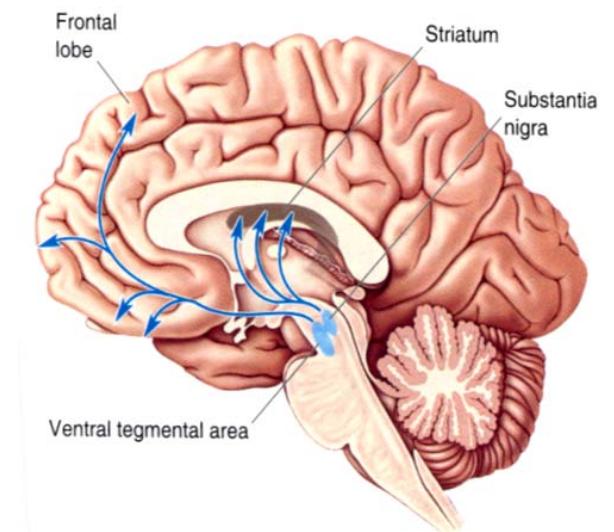
rats work hard to press a lever which injects  
**dopamine agonists & transporter blockers**  
(amphetamine, cocaine)  
into the BSR sites

**dopamine antagonists**  
stop the rat pressing the lever when stimulated



# Dopamine = (neural basis of) reward

- dopamine project to reward centres
- neural self-administration of dopamine is rewarding
- electric self-stimulation to reward centres is blocked by DA antagonists
- dopaminergic drugs are addictive
- dopaminergic medication in PD patients associated with addiction



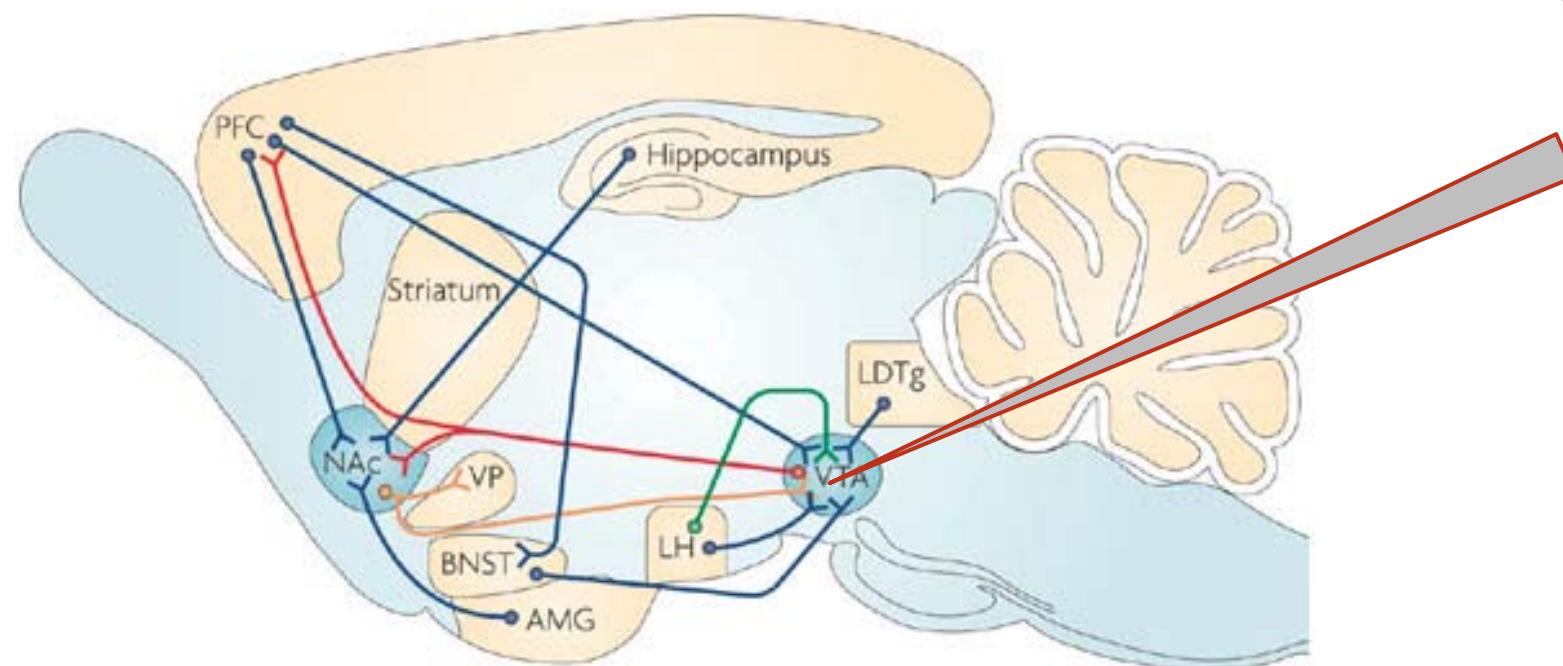
# Pavlovian conditioning



dopamine ≠ reward?

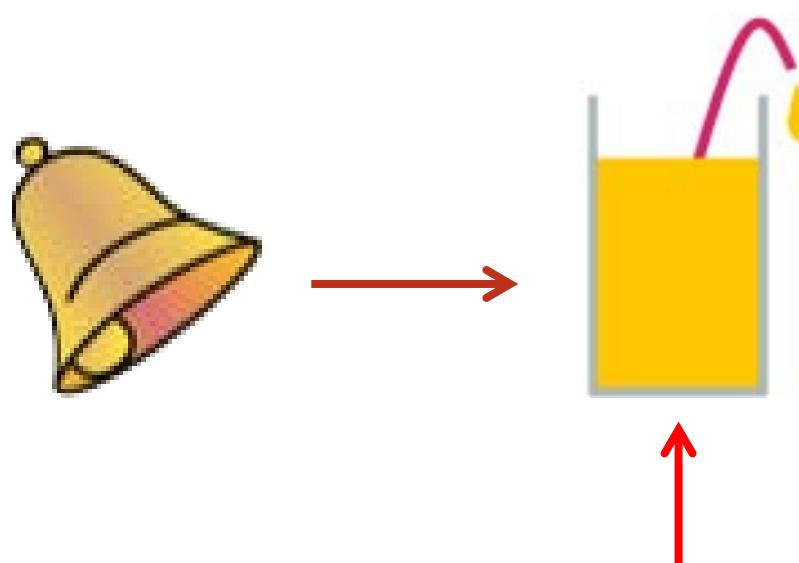


REOPENED CASE

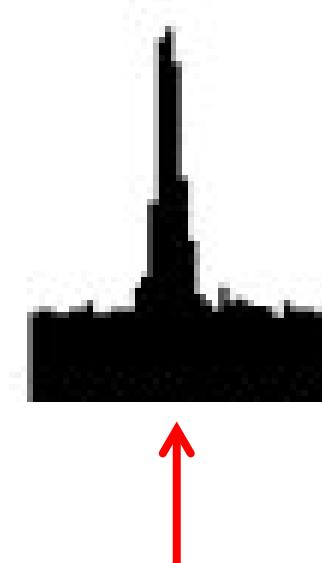


VTA DA Baseline  
firing rate of 3-5 Hz

Conditioning procedure

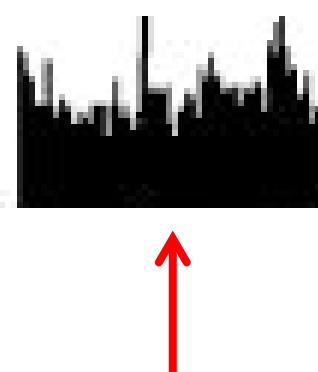


before conditioning



after conditioning

..... while the animal *STILL*  
seeks out the rewards!



Schultz ea. 1997



Having a closer look at conditioning

(a.k.a. meanwhile in computer science)



# Outline

- processing reward
- introduction to dopamine
- learning about rewards
  - error-driven learning
  - neural basis
- learning actions
  - TD learning
  - law of effect
  - second-order reinforcement
  - multiple decision systems

# Classic learning theory

Question: How do we learn to predict/seek rewards?

*Expected Value*  $EV_t$ : average amount of food following bell

On trial  $t$

- You hear the bell
- you get food, of amount  $r_t$

How to update expectation  $EV_t$ , given this experience?

Rescorla-Wagner model (delta-learning)

$$EV_{t+1} = (1 - \alpha) \cdot EV_t + \alpha \cdot r_t$$

downweight  
current belief

update with  
observation

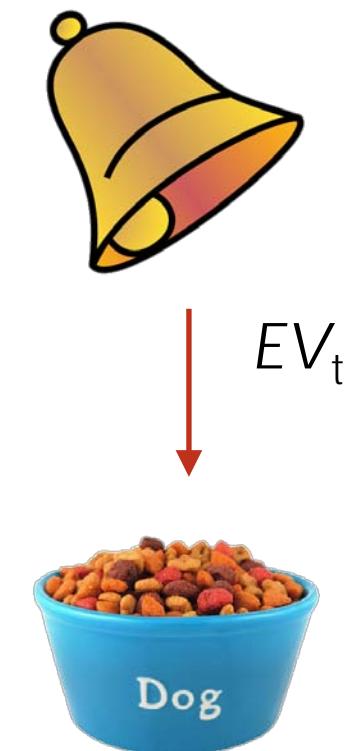
$\alpha$  is learning rate parameter

Equivalently

$$EV_{t+1} = EV_t + \alpha \cdot (r_t - EV_t)$$

add weighted prediction  
error to current belief

$r_t - EV_t$  is prediction error



# Rescorla-Wagner Rule

## I. Learning = **weighted average** of rewards

$$EV_{t+1} = (1 - \alpha) \cdot EV_t + \alpha \cdot r_t$$

recent trials are weighed more strongly

$$EV_1 = 0$$

$$EV_2 = (1 - \alpha) \cdot 0 + \alpha \cdot r_1$$

$$EV_3 = (1 - \alpha) \cdot \alpha \cdot r_1 + \alpha \cdot r_2$$

$$EV_4 = (1 - \alpha)^2 \cdot \alpha \cdot r_1 + (1 - \alpha) \cdot \alpha \cdot r_2$$

Note:

$$0 <= \alpha <= 1, \text{ so}$$

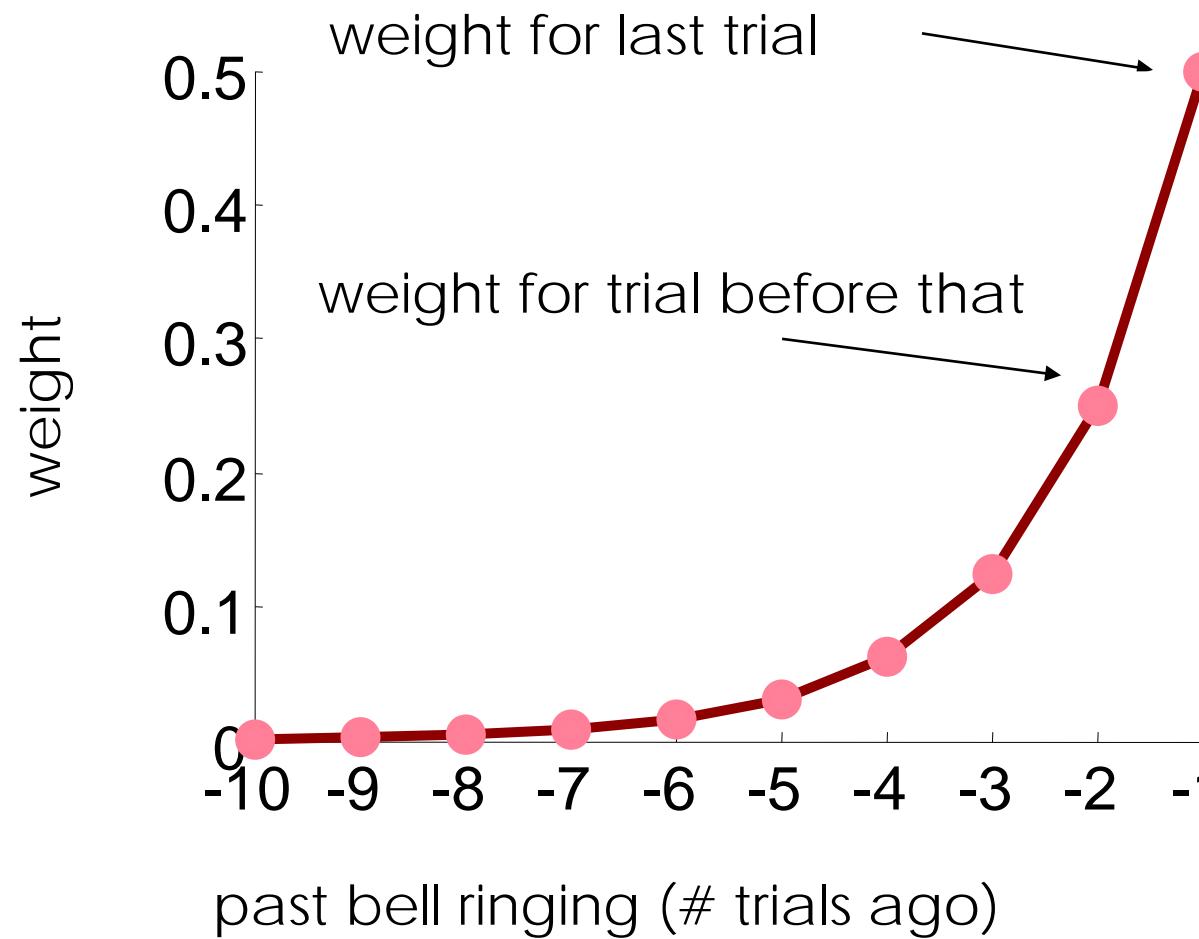
$$(1 - \alpha)^3 < (1 - \alpha)^2 < (1 - \alpha)$$

# Rescorla-Wagner Rule

## I. Learning = weighted average of rewards

$$EV_{t+1} = (1 - \alpha) \cdot EV_t + \alpha \cdot r_t$$

recent trials are weighed more strongly



## II. Learning = error-driven

$$EV_{t+1} = EV_t + \alpha \cdot (r_t - EV_t)$$

prediction error =  
experienced - expected value  
i.e. outcome – prediction

$$EV_1 = 0$$

$$EV_2 = 0 + \alpha \cdot (r_1 - 0)$$

$$EV_3 = \alpha \cdot r_1 + \alpha \cdot (r_2 - \alpha \cdot r_1)$$

$$EV_3 = \alpha \cdot r_2 + (1 - \alpha) \cdot \alpha \cdot r_1$$

rewrite as

# Rescorla-Wagner Rule



?

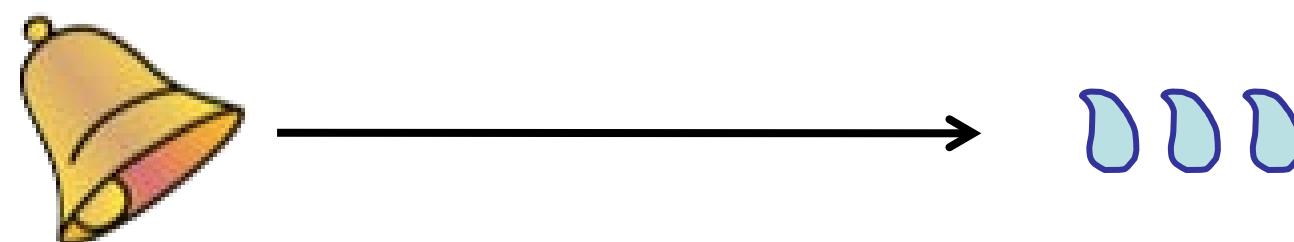
## II. Learning = error-driven

$$EV_{t+1} = EV_t + \alpha \cdot (r_t - EV_t)$$

prediction error =

experienced - expected value

i.e. outcome – prediction

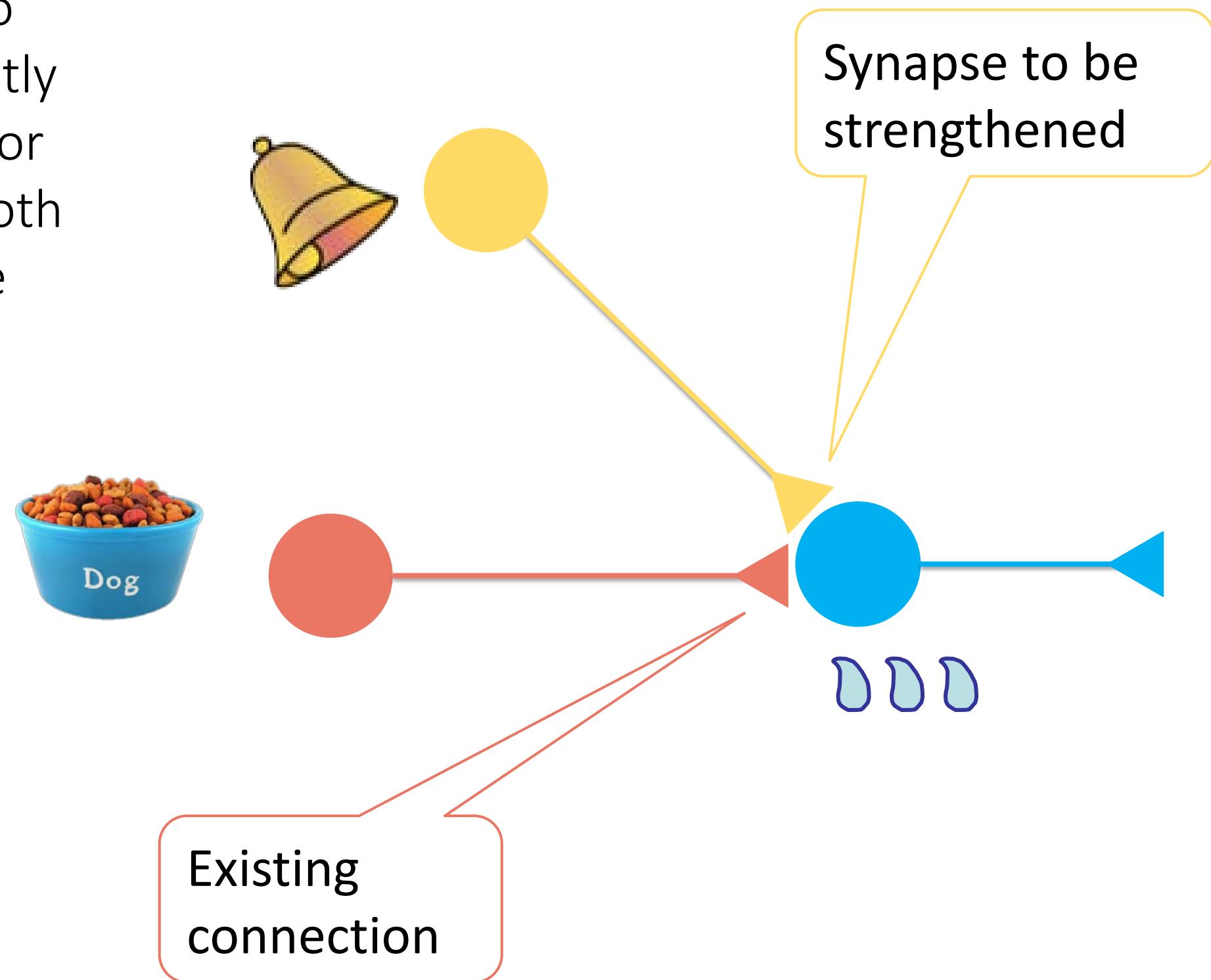


Pavlovian conditioning could be  
Hebbian / statistical learning

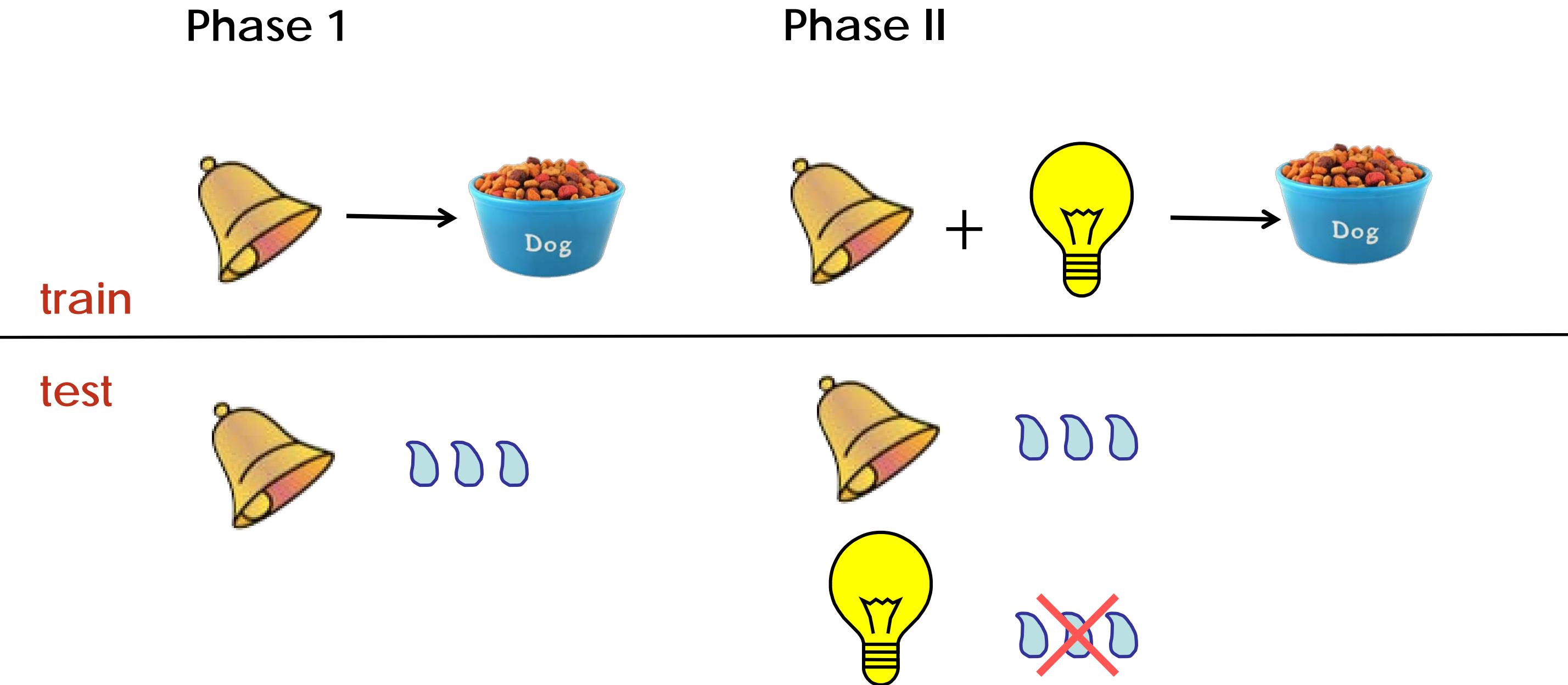
# Hebb's postulate: fire together, wire together

“When an axon of cell A is near enough to excite a cell B and repeatedly or persistently takes part in firing it. Some growth process or metabolic change takes place in one or both cells such that A’s efficiency, as one of the cells firing B, is increased”

- D.O. Hebb, 1949



# Blocking



# Blocking explained by RW rule



Rescorla & Wagner (1972) say:

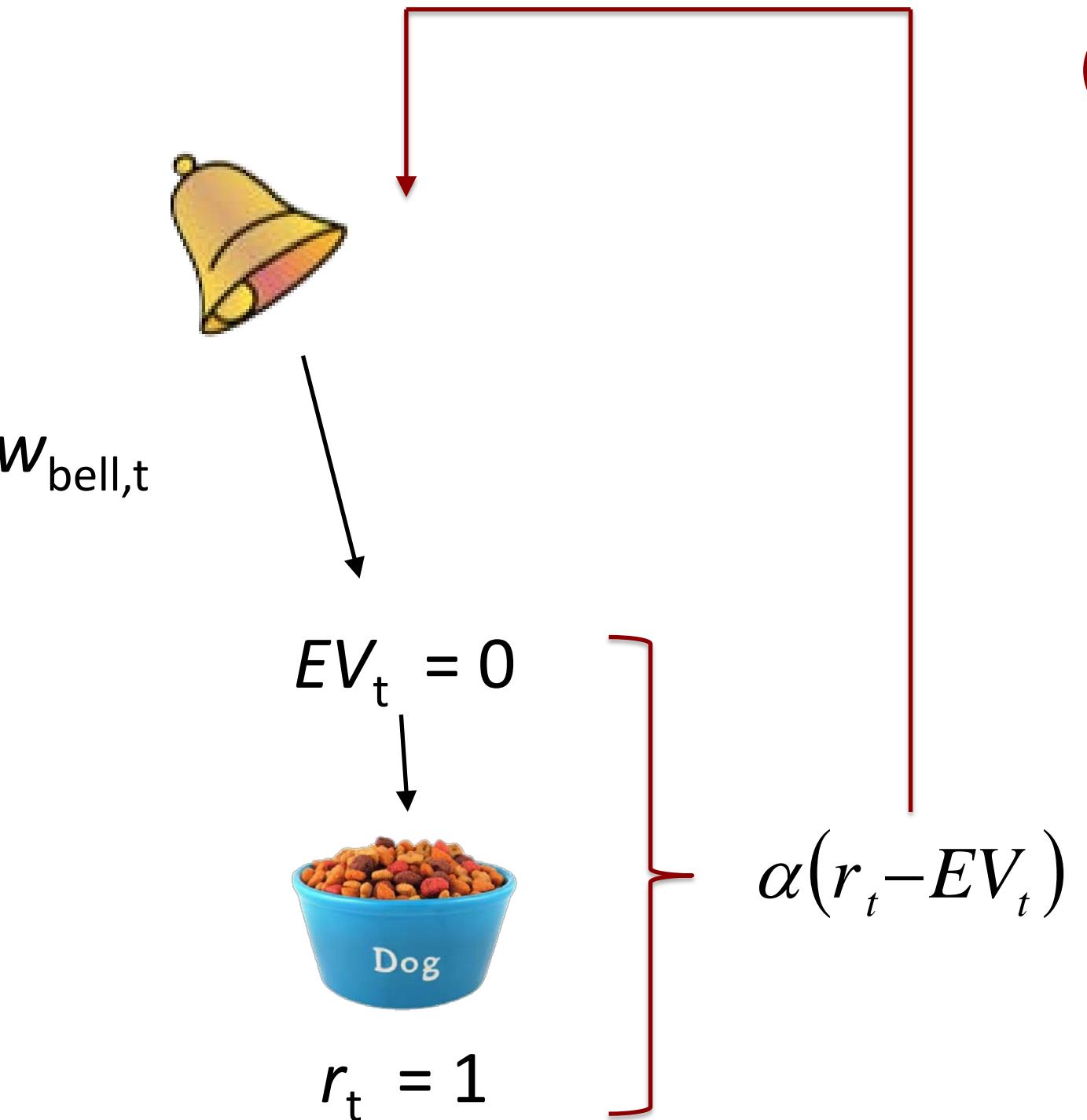
Predict total  $EV_t$ :  $EV_t = \sum_i w_{i,t}$

for each presented stimulus  $i$

Learn (based on net error):

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

for each presented stimulus  $i$





# Blocking explained by RW rule

Rescorla & Wagner (1972) say:

$$\text{Predict total } EV_t : \quad EV_t = \sum_i w_{i,t}$$

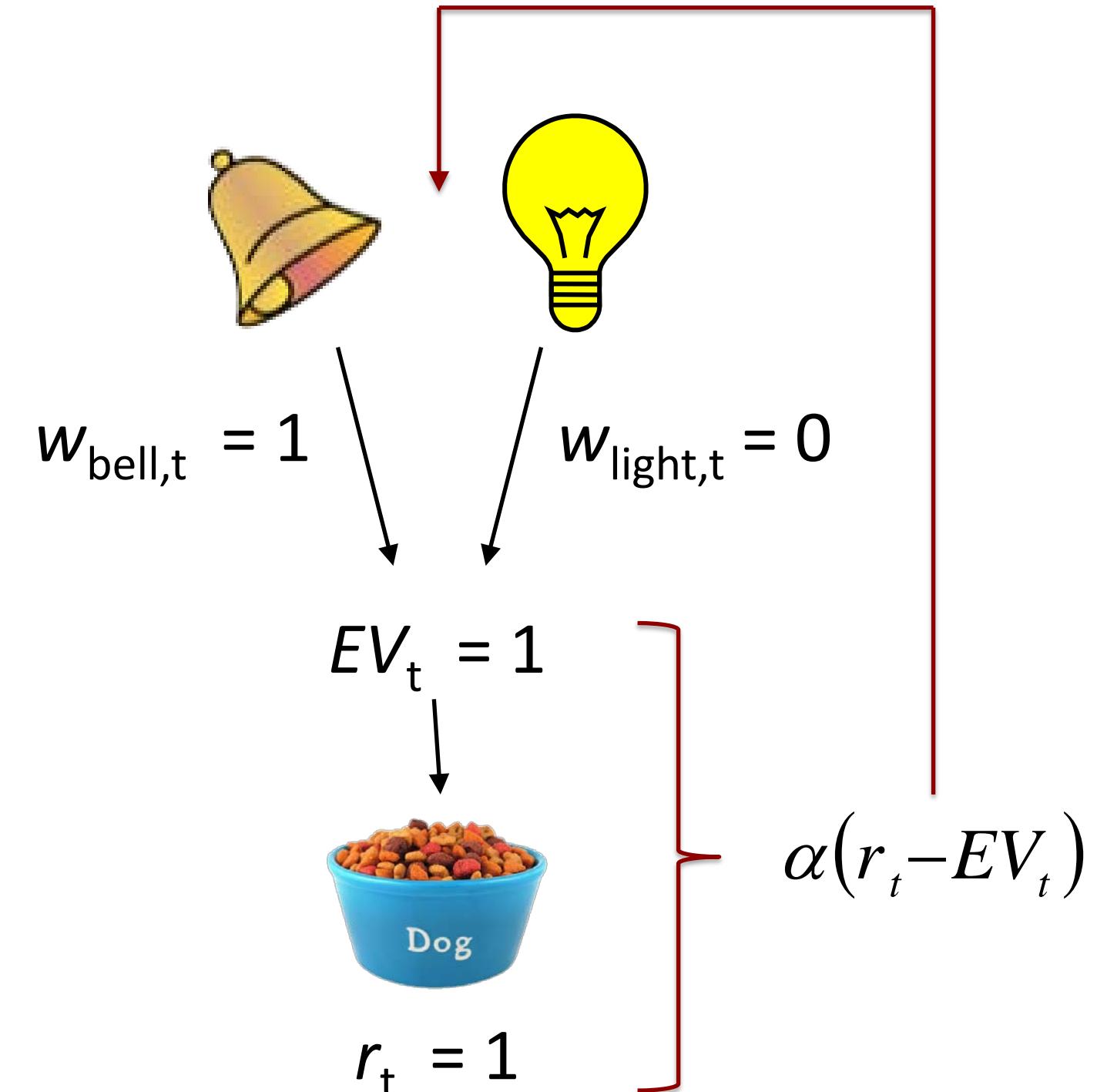
for each presented stimulus  $i$

Learn (based on net error):

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

for each presented stimulus  $i$

If  $w_{bell}$  fully explains food, then  $w_{light}$  stays 0



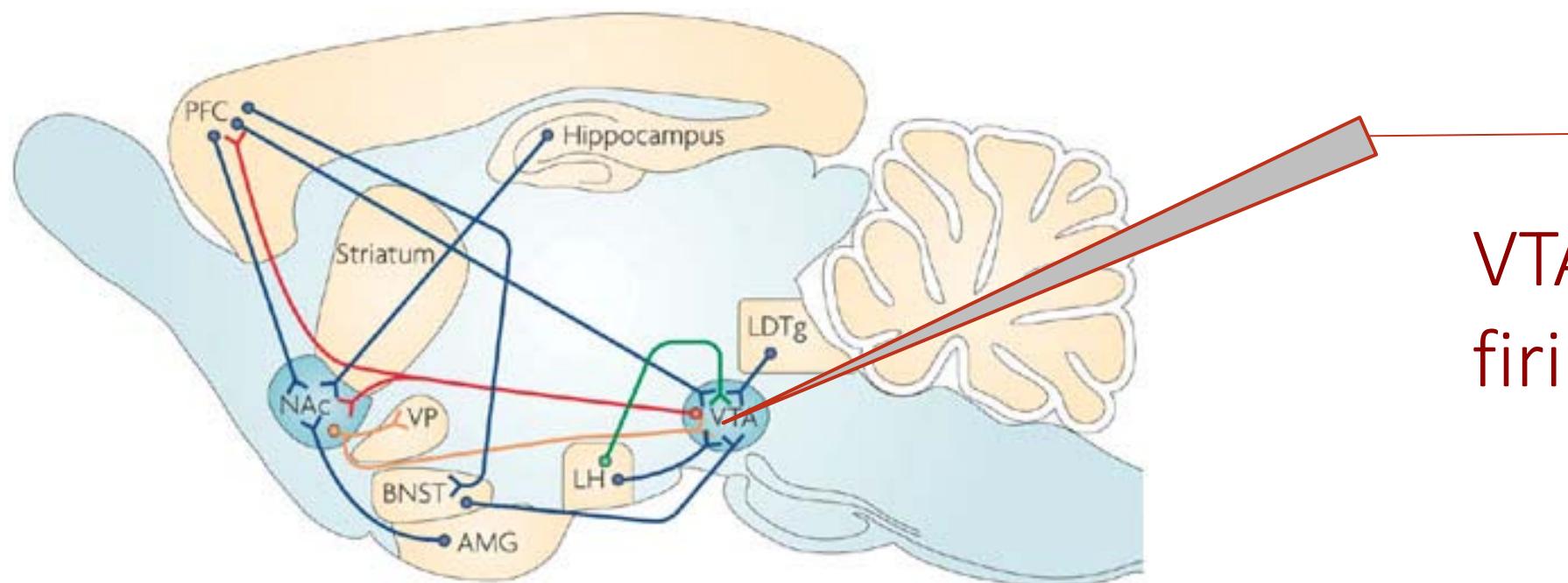
blocking phenomenon provides evidence that reinforcement learning is error-driven



Pavlovian conditioning suggests error-driven learning

.... is there a neural basis for this?

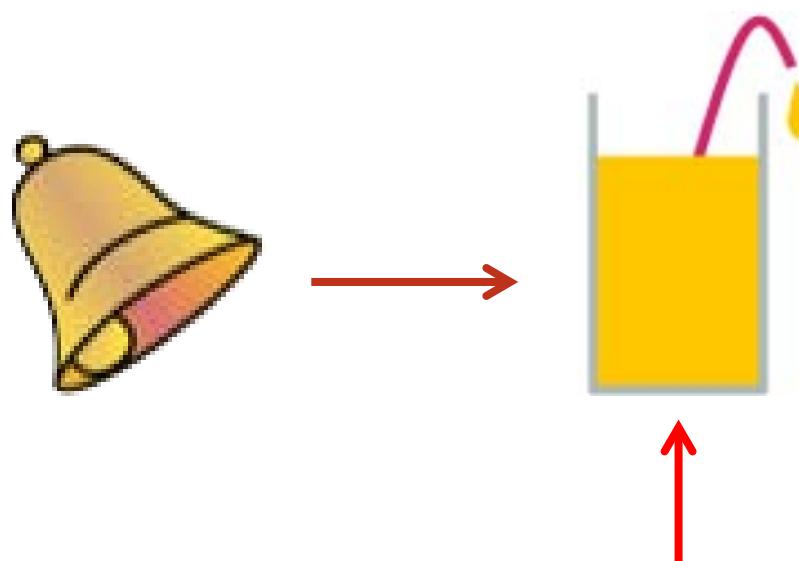
# dopamine ≠ reward?



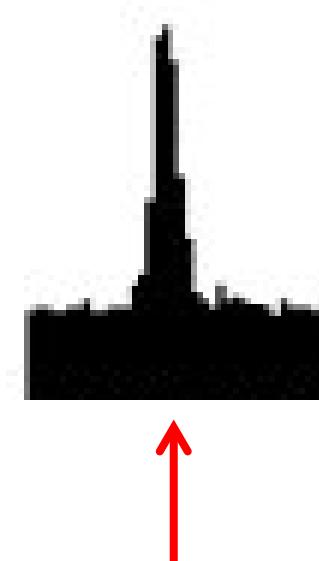
VTA DA Baseline  
firing rate of 3-5 Hz



Conditioning procedure



before conditioning



after conditioning

..... while the animal *STILL*  
seeks out the rewards!

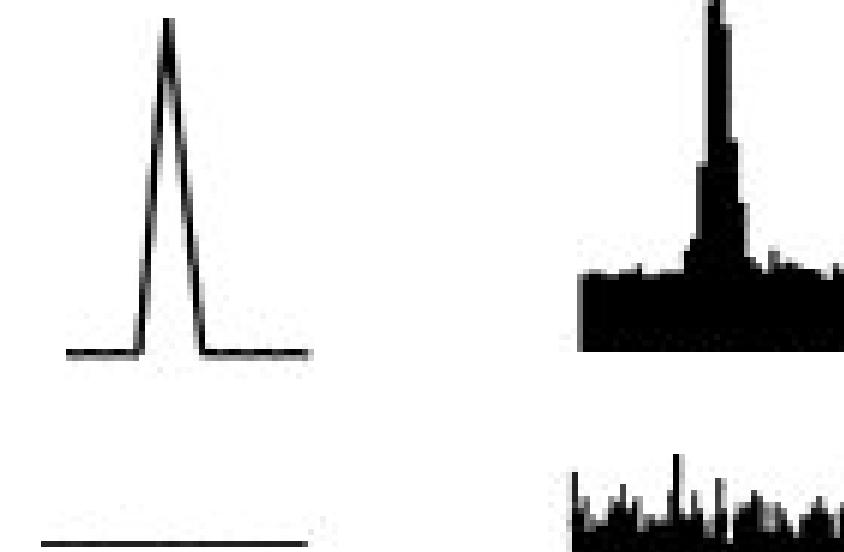


?



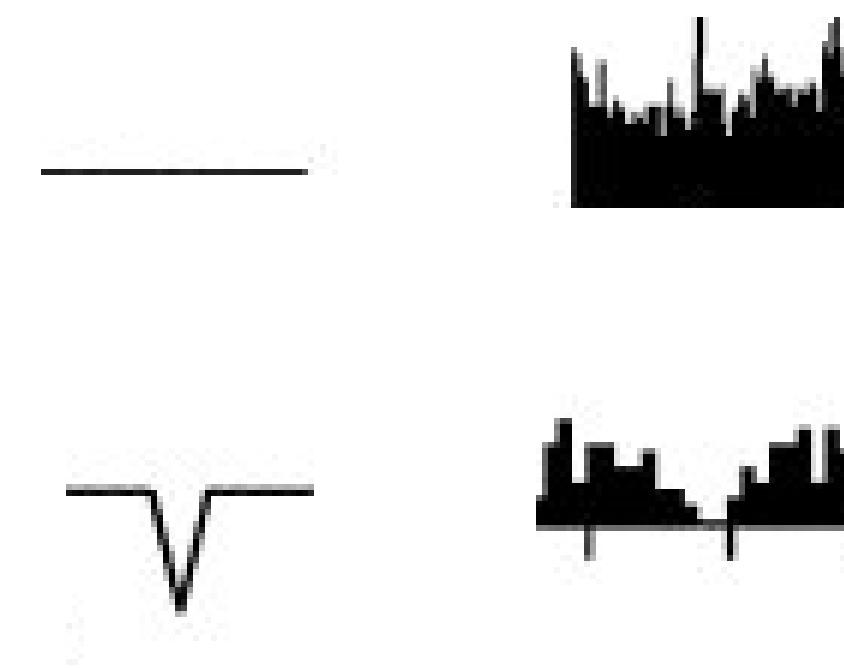
## theoretical PE DA firing

before conditioning



burst to  
unexpected reward

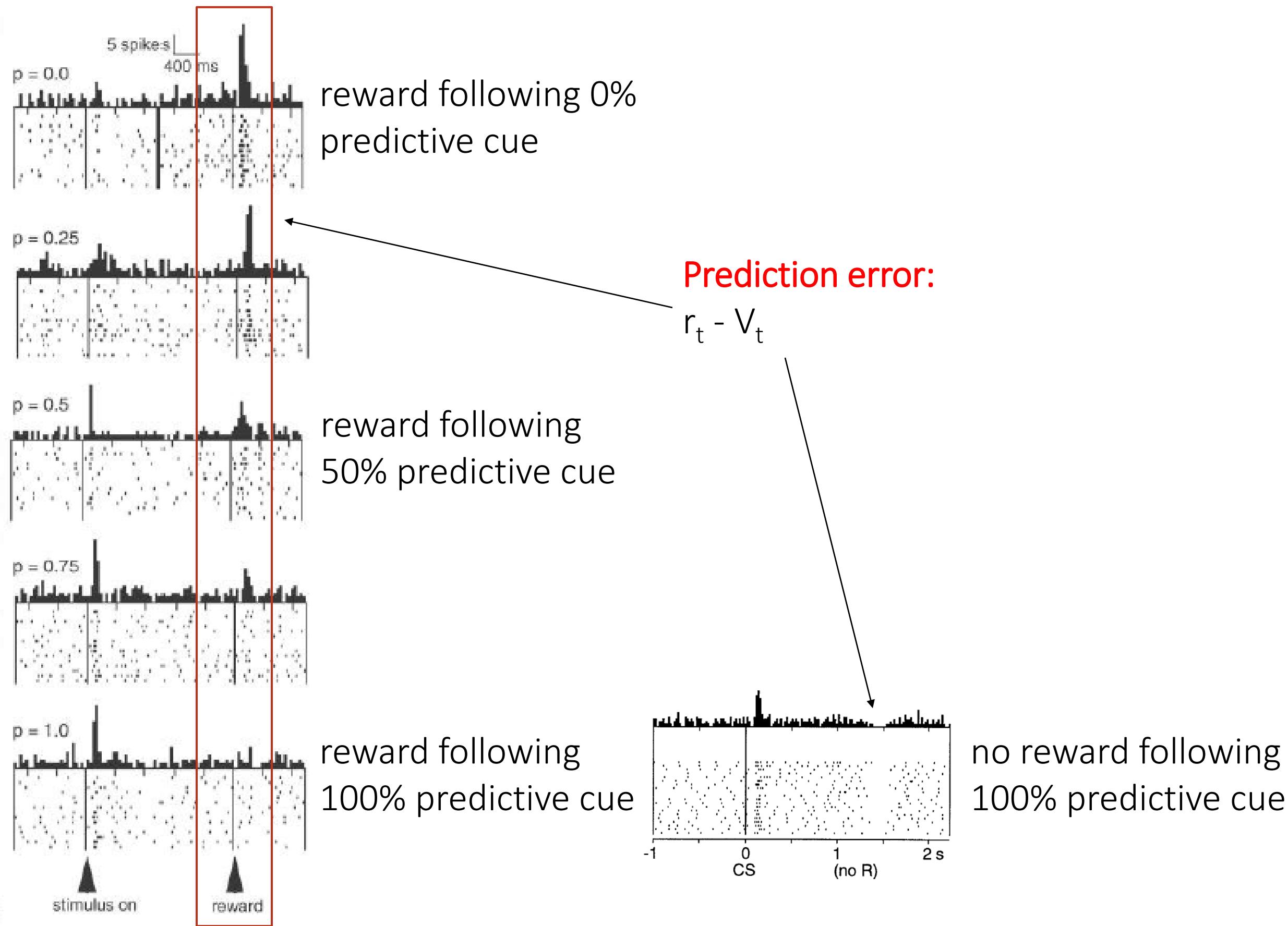
after conditioning



Dopamine firing matches  
theoretical prediction error  
signal

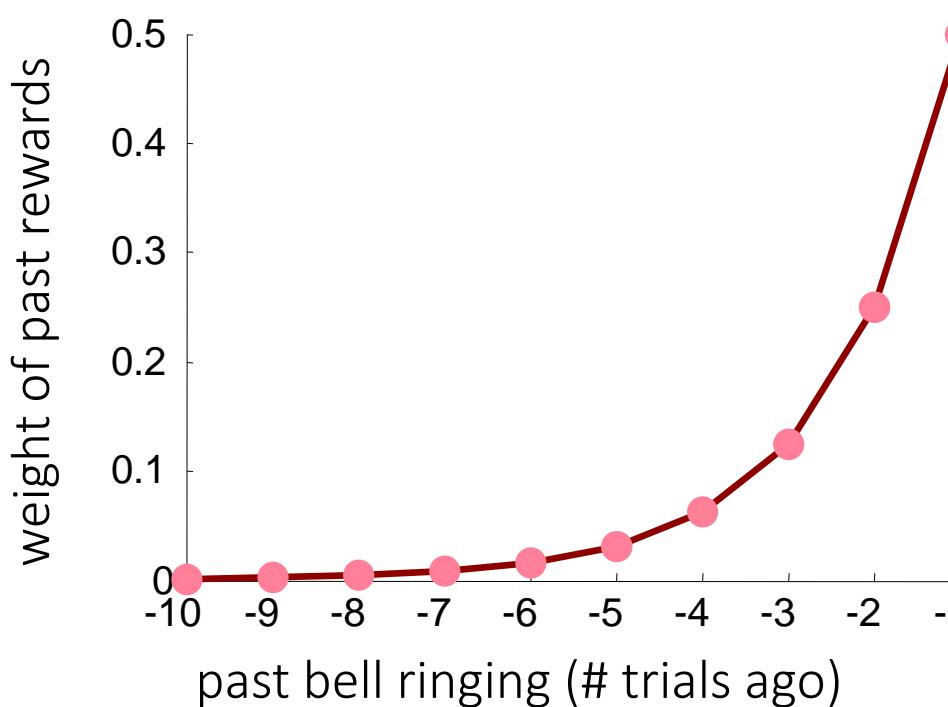
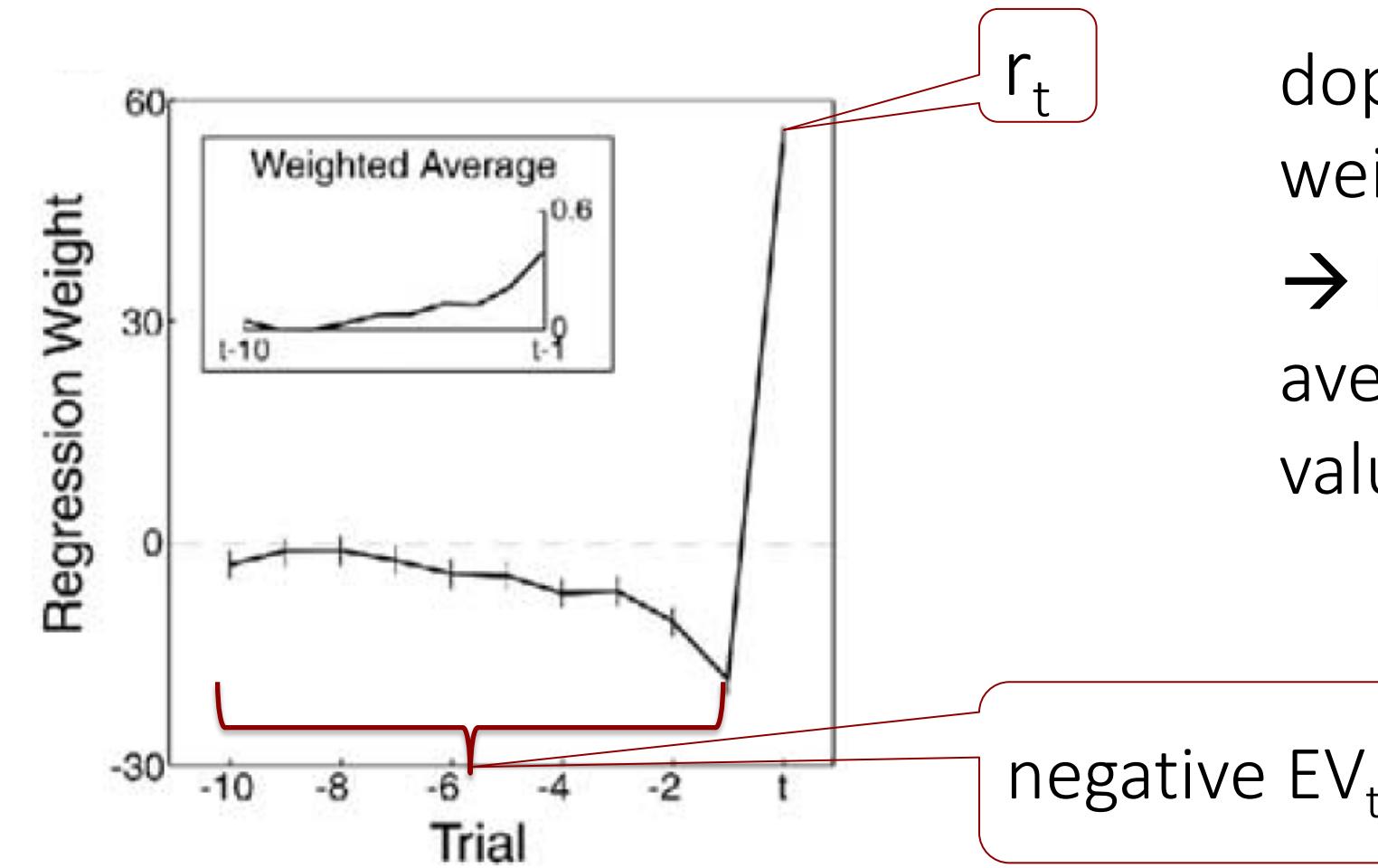
pause at time of  
omitted reward

# more dopamine responses





# prediction error



dopamine response to reward is weighted sum of current & past rewards  
 $\rightarrow$  looks like current  $r$  minus weighted average of past rewards (= estimated value): prediction error =  $r_t + (-EV_t)$

remember....

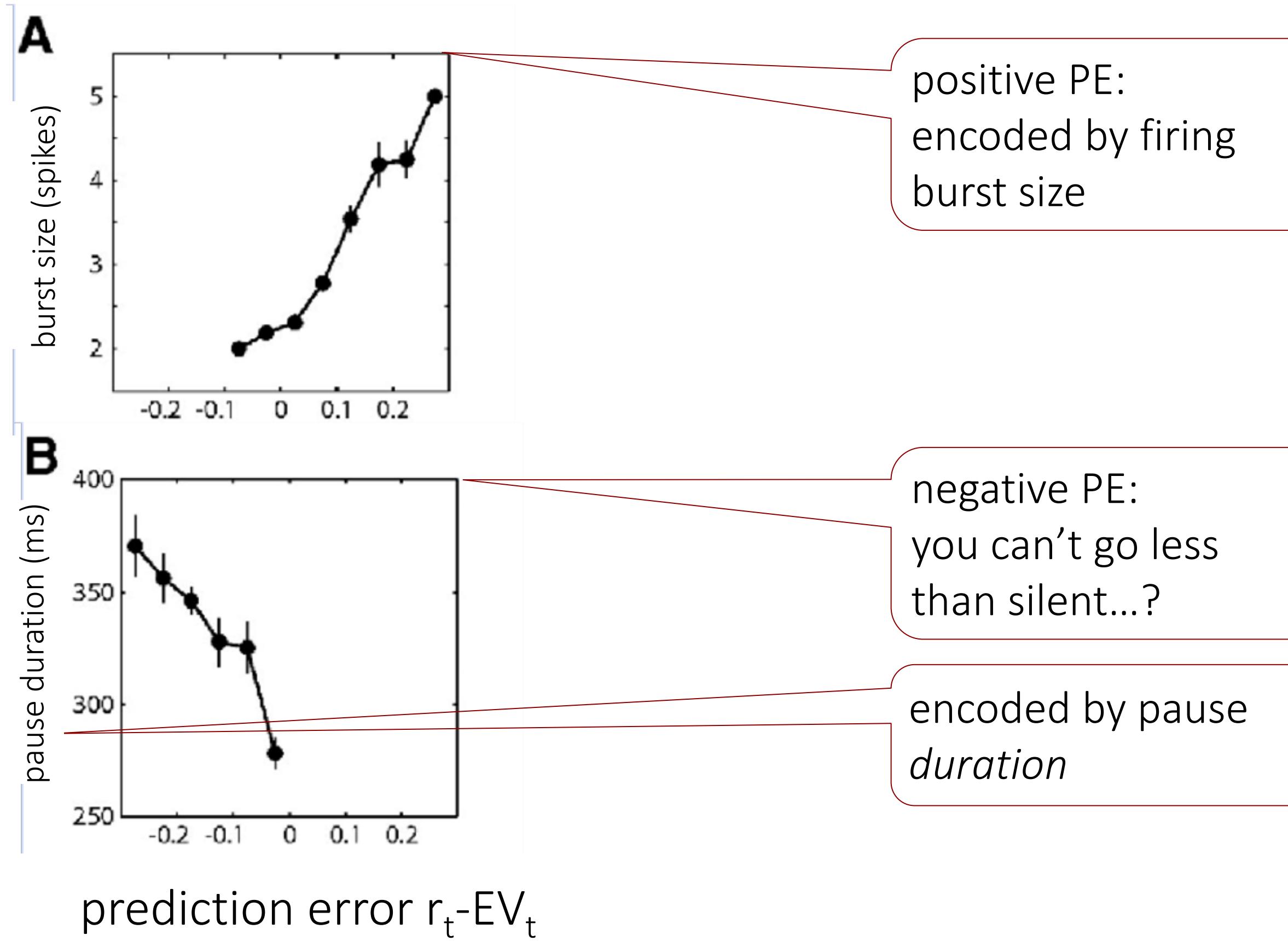
$$EV_1 = 0$$

$$EV_2 = (1-\alpha) \cdot 0 + \alpha \cdot r_1$$

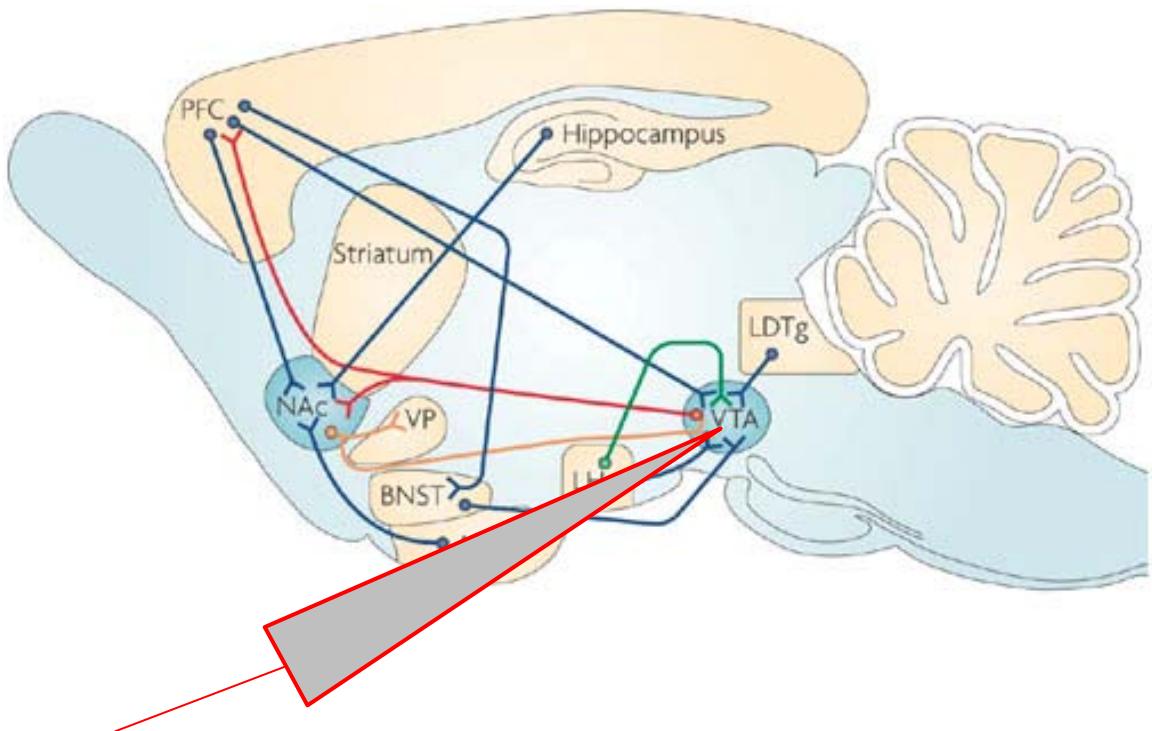
$$EV_3 = (1-\alpha) \cdot \alpha \cdot r_1 + \alpha \cdot r_2$$

$$EV_4 = (1-\alpha)^2 \cdot \alpha \cdot r_1 + (1-\alpha) \cdot \alpha \cdot r_2 + \alpha \cdot r_3$$

# Encoding positive versus negative prediction errors

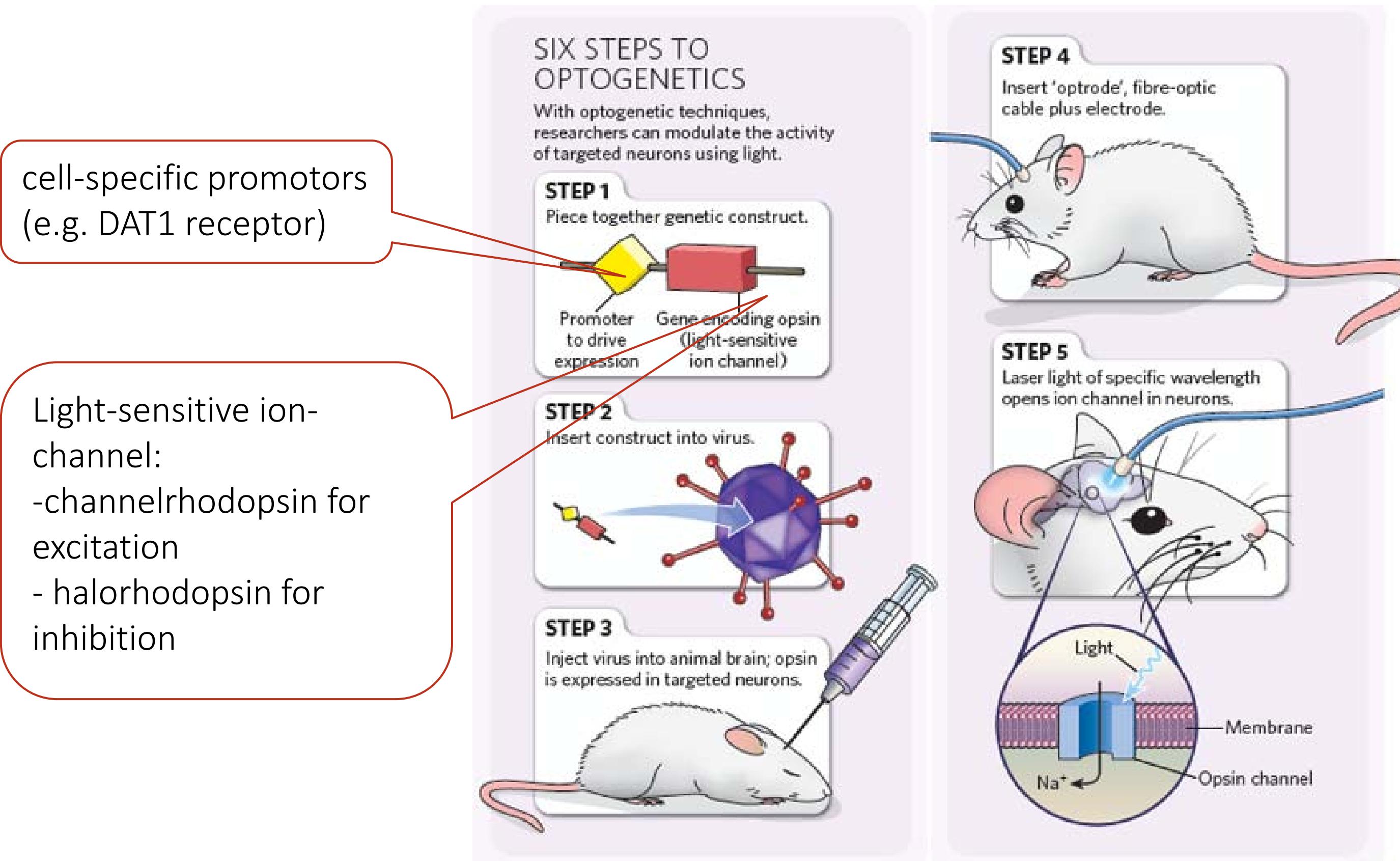


# Dopamine = Reward Prediction Error ?



midbrain dopamine firing  
... looks like a prediction error  
... but does it act like one?

# Using optogenetics to control neuronal firing

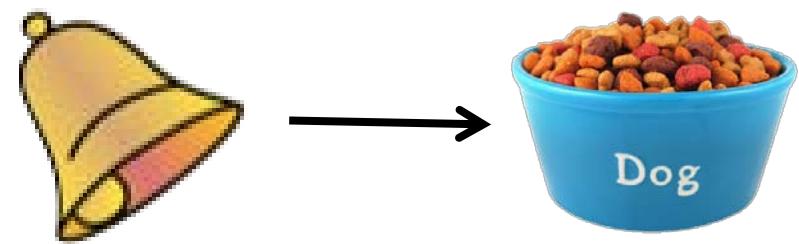




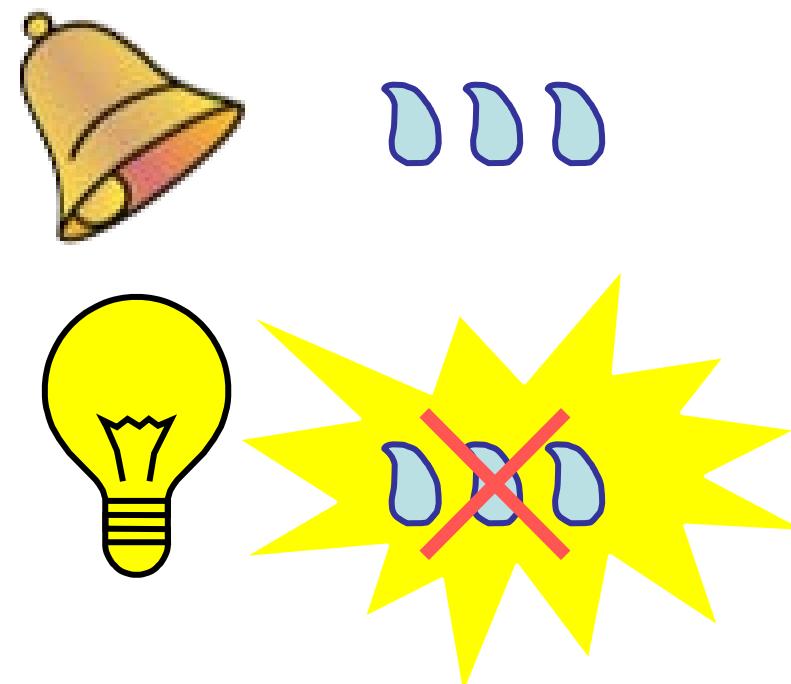
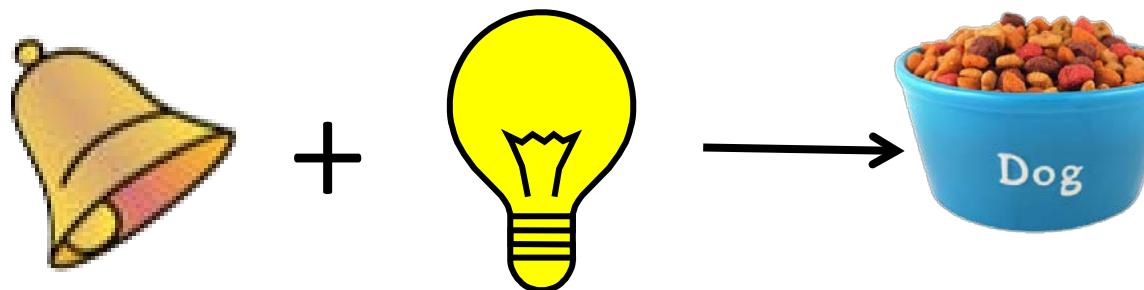
How could you prove using optogenetics, that dopamine firing acts as a reward prediction error?

# Blocking 2.0

Phase 1



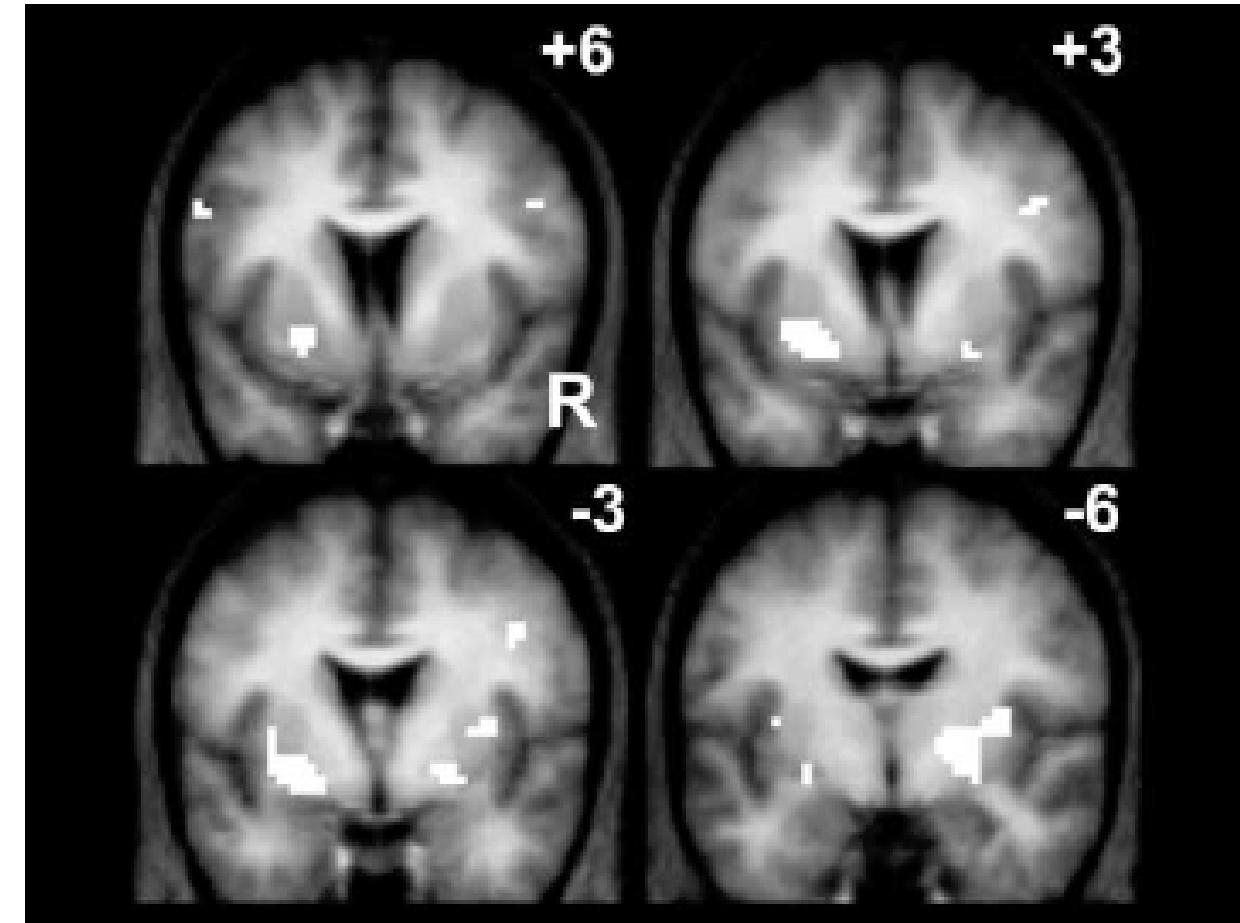
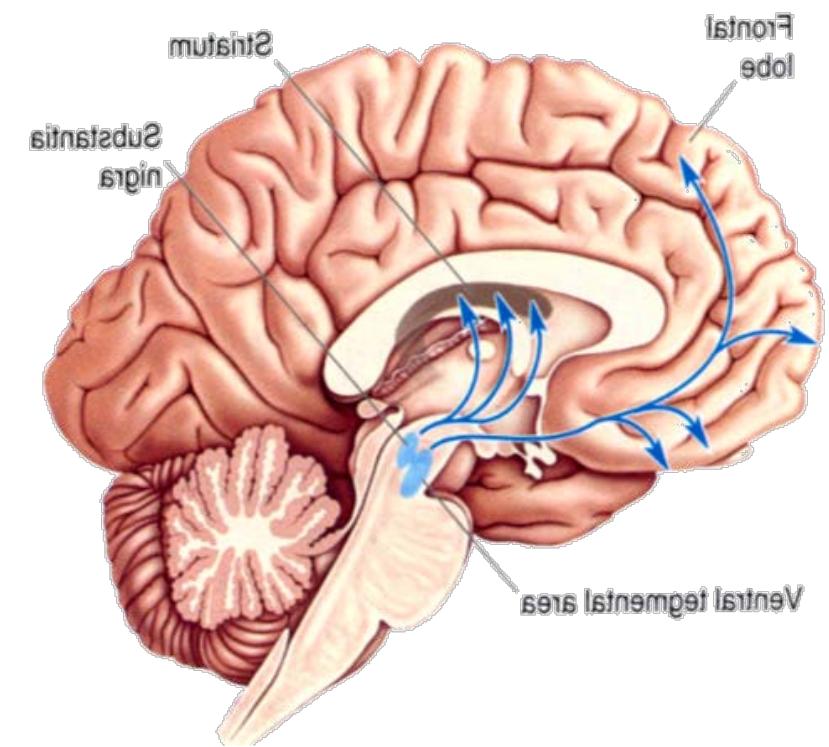
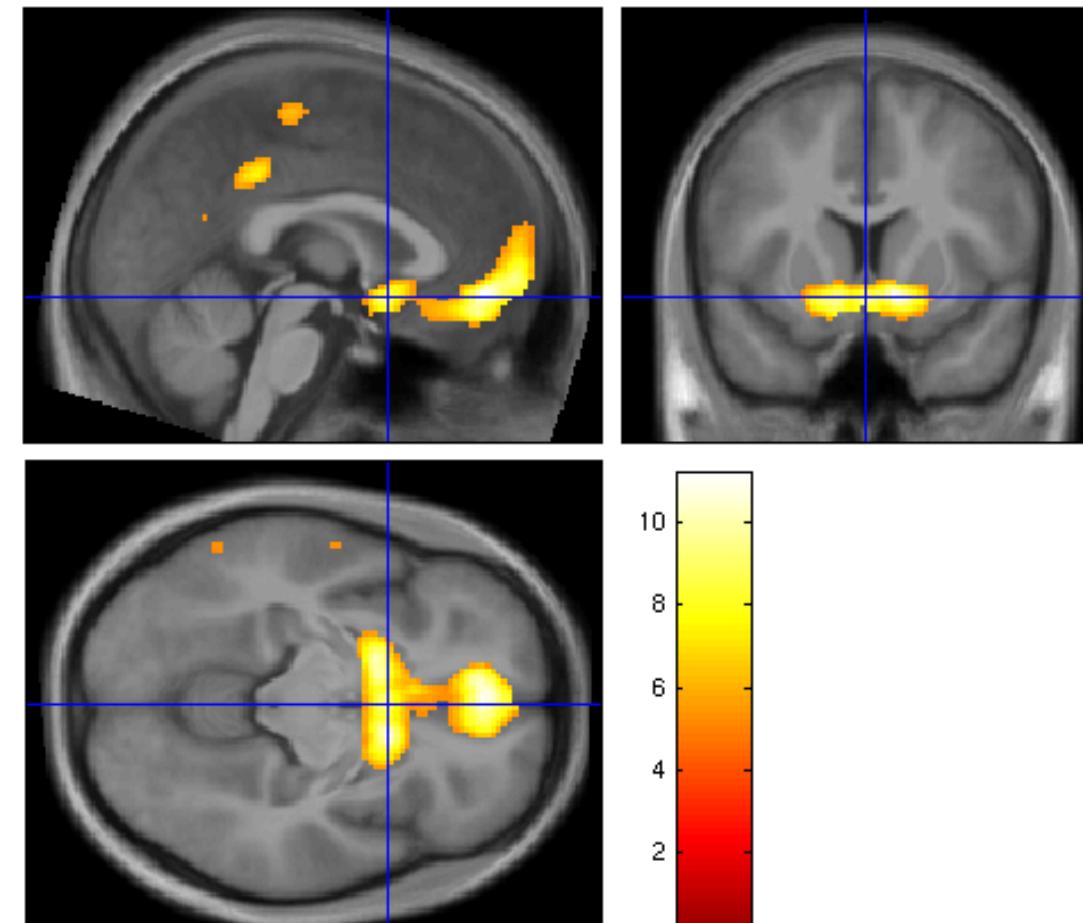
Phase II





Any evidence for dopamine prediction errors in  
humans?

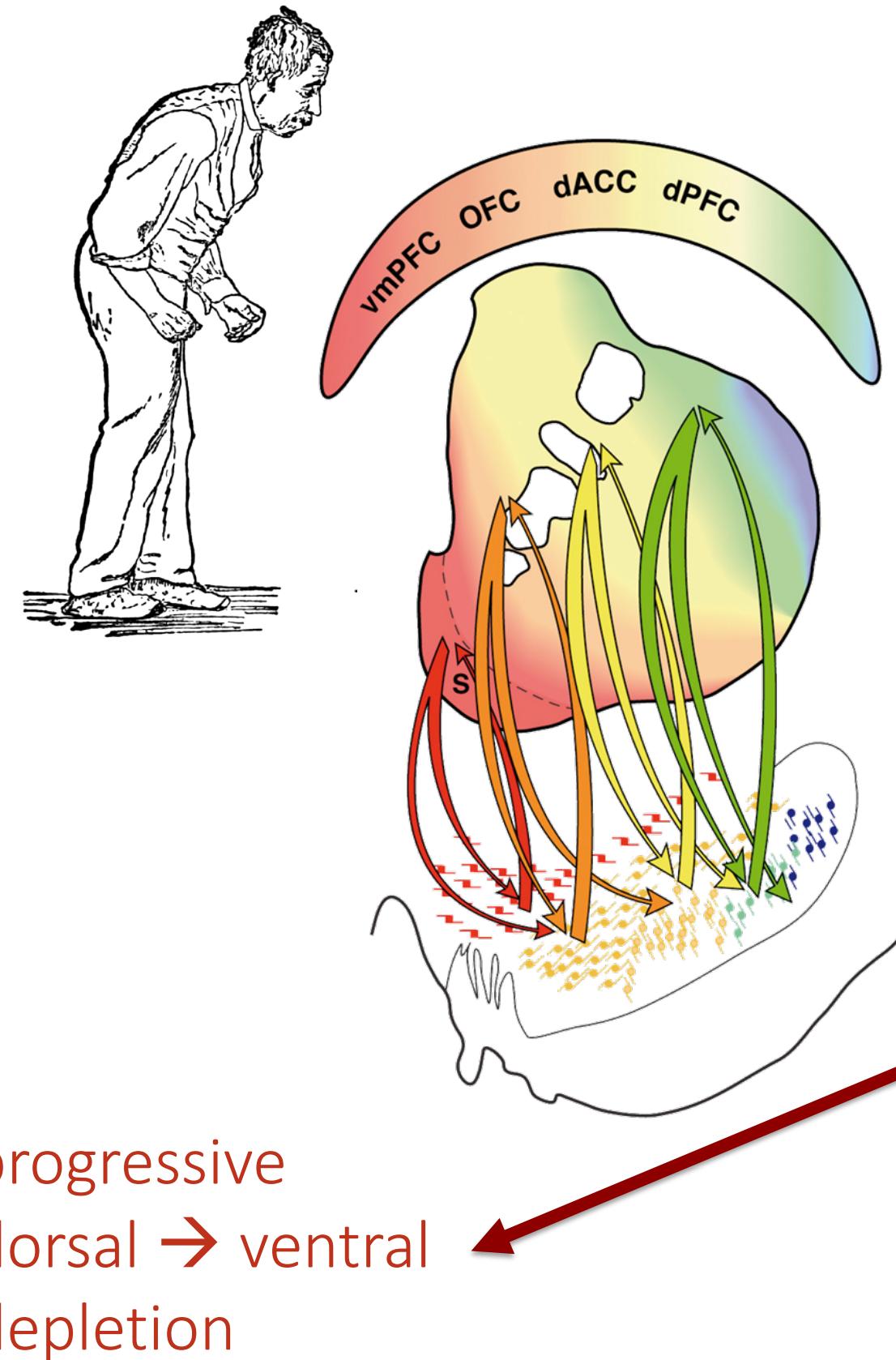
# Neural reward prediction errors in humans



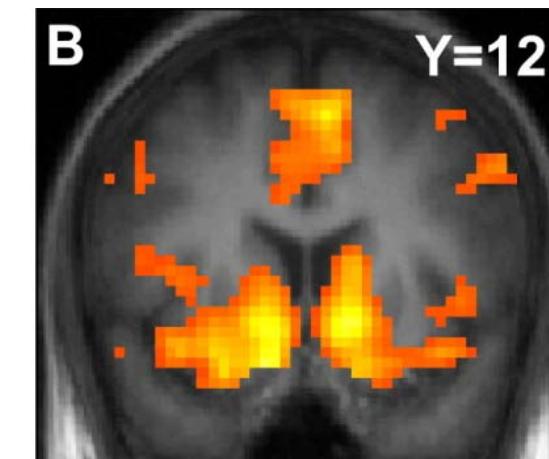
BOLD signal to reward in DA target regions is modulated by prediction error  
Juice: unexpected good – unexpected bad  
Den Ouden ea. *In prep*

BOLD signal in striatum correlates trial-by-trial with prediction error size  
O'Doherty ea. 2003

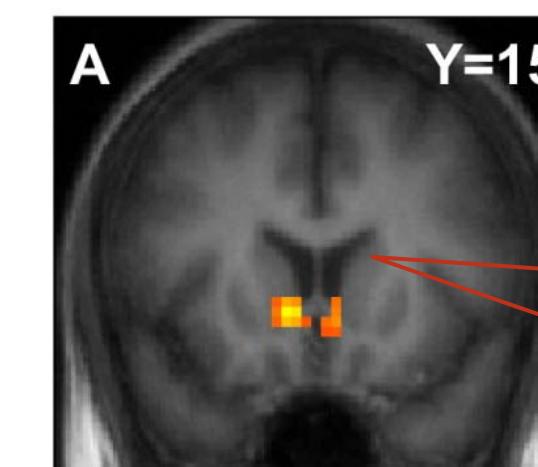
# Neural reward prediction errors in humans – dopamine?



healthy control

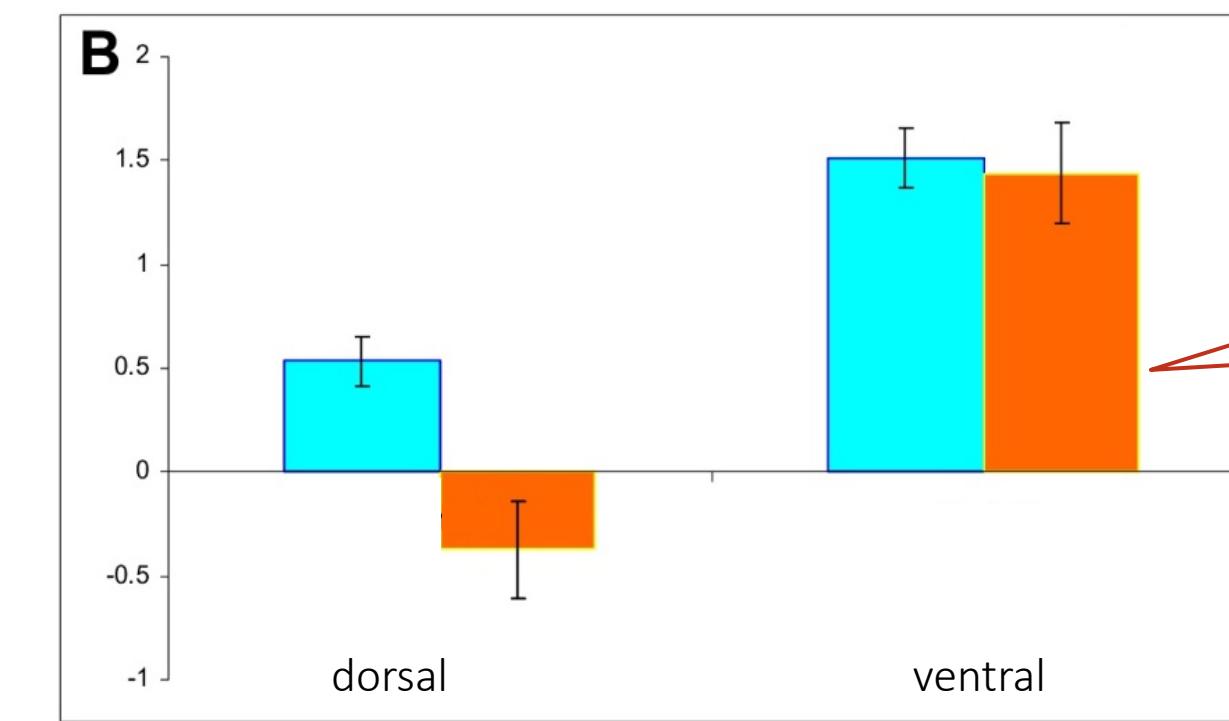


Parkinson's disease



dorsal, but not  
ventral, striatal DA  
depletion

BOLD PE effect sizes



ventral, but not  
dorsal, striatal RPE  
in PD

# How are prediction errors computed?

Remember Marr's levels of analysis?

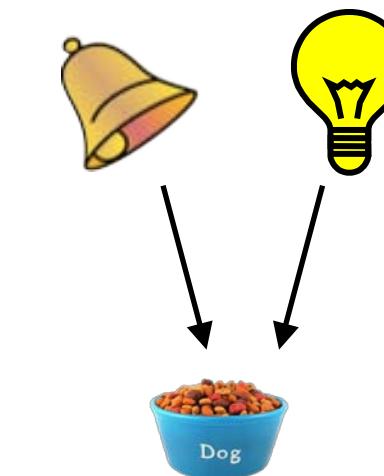
**Computational:** what problem is the brain trying to solve?

- getting rewards



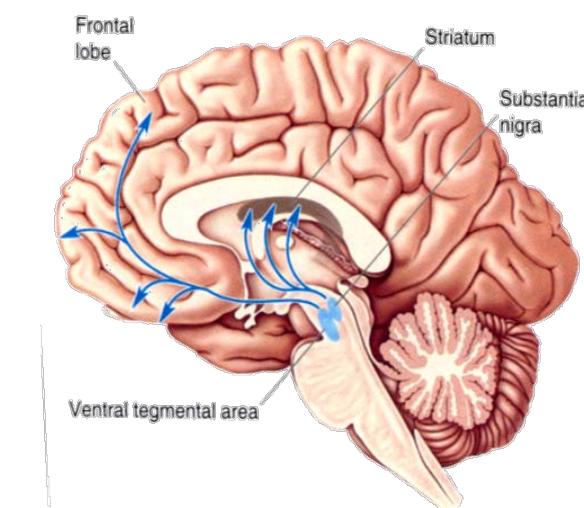
**Algorithmic:** how does the brain solve these problems:

- error-based reinforcement learning

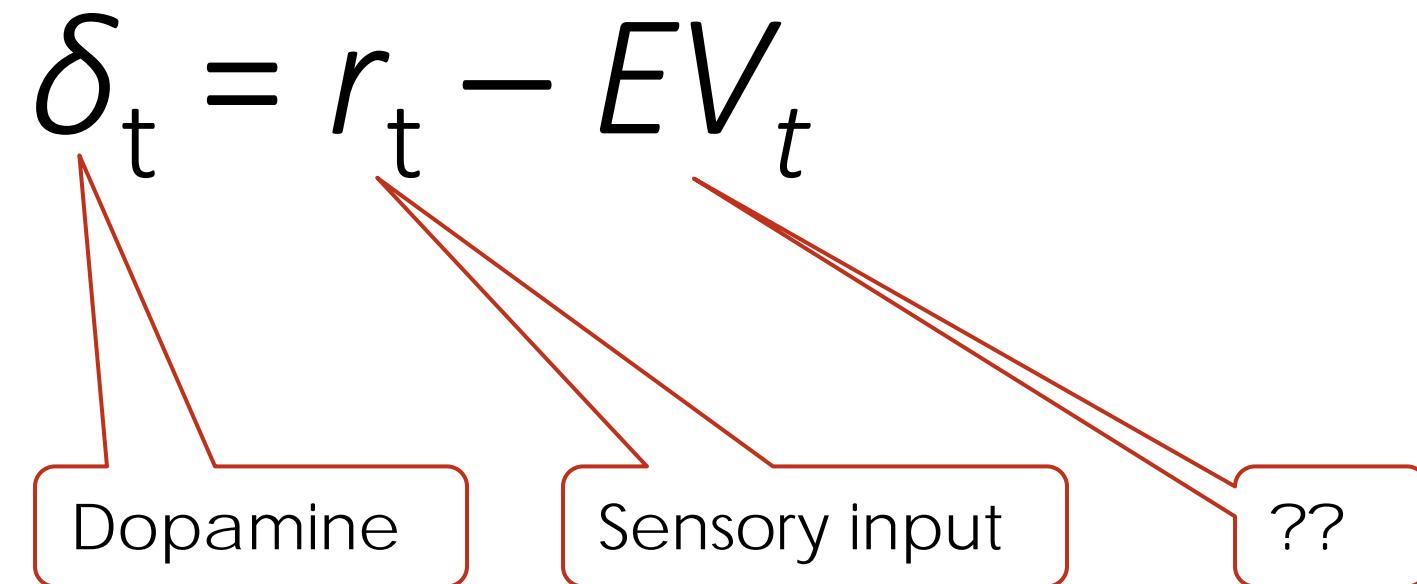


**Implementational:** how are these solutions implemented in wet-ware?

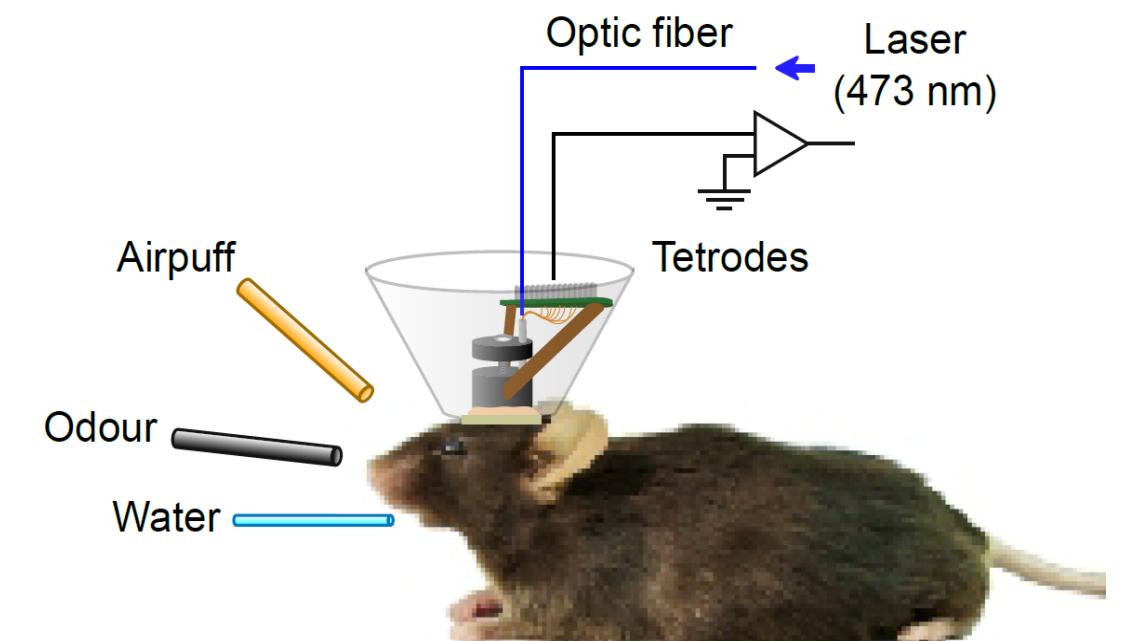
- dopamine reflects error signal  
.... but how is this arrived at?



# How are (dopamine) prediction errors computed?

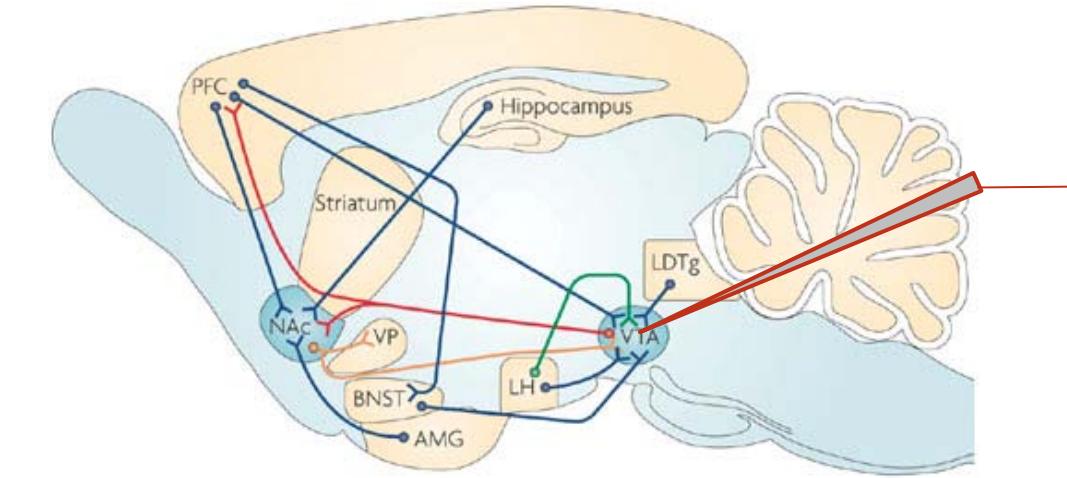
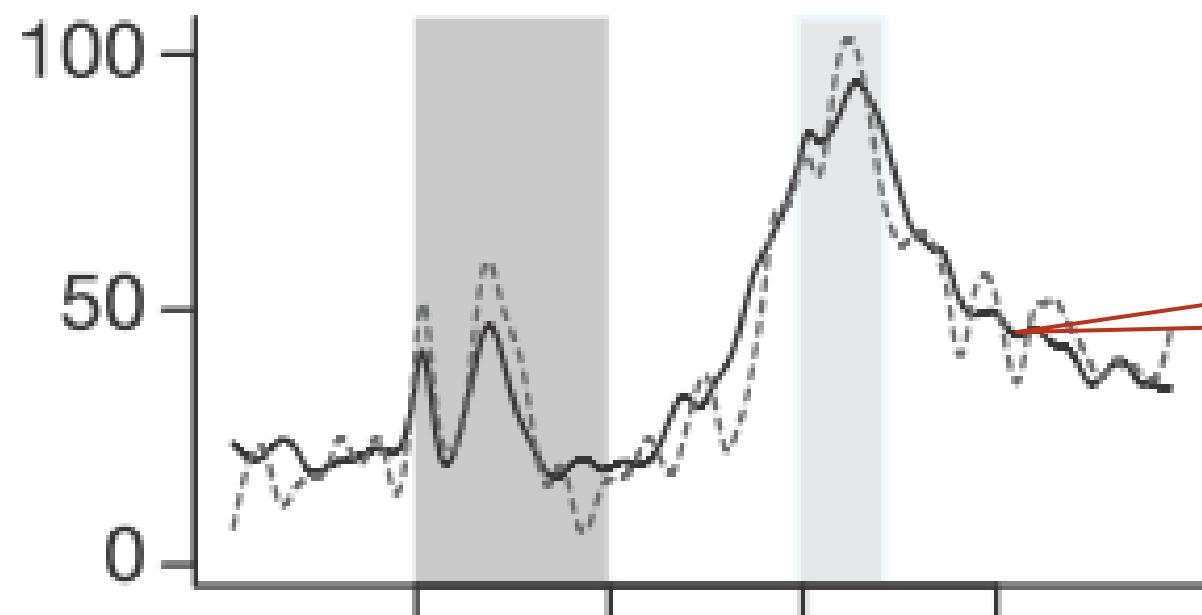


- Where does the prediction  $EV_t$  come from?
- What should that prediction signal look like?

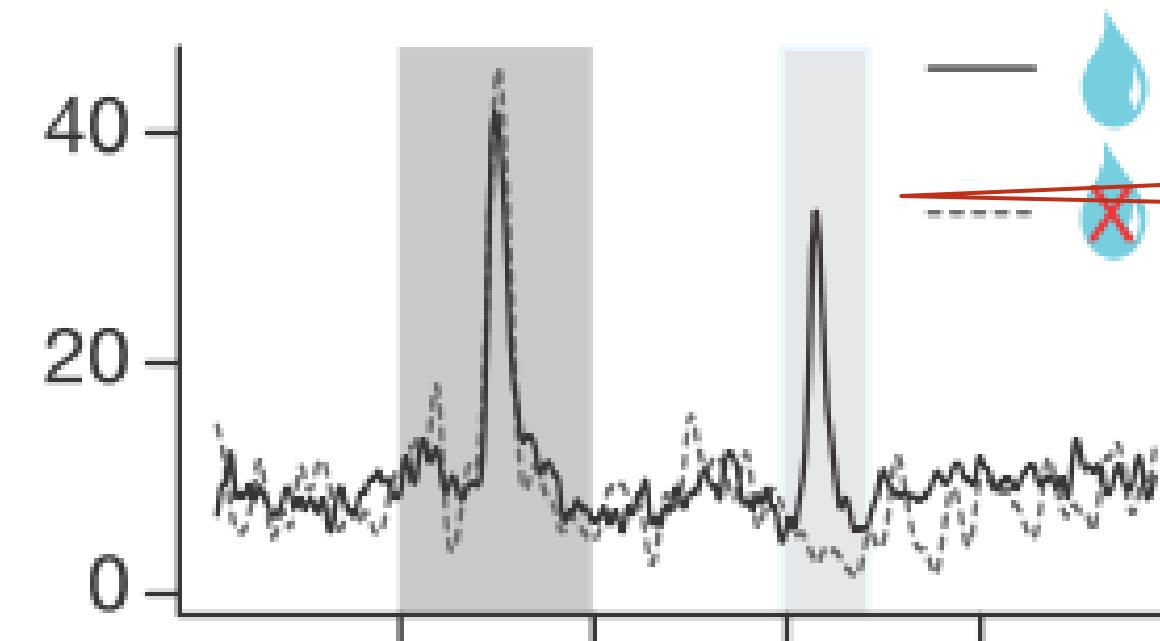


# Getting the prediction

- 2 classes of VTA neurons:



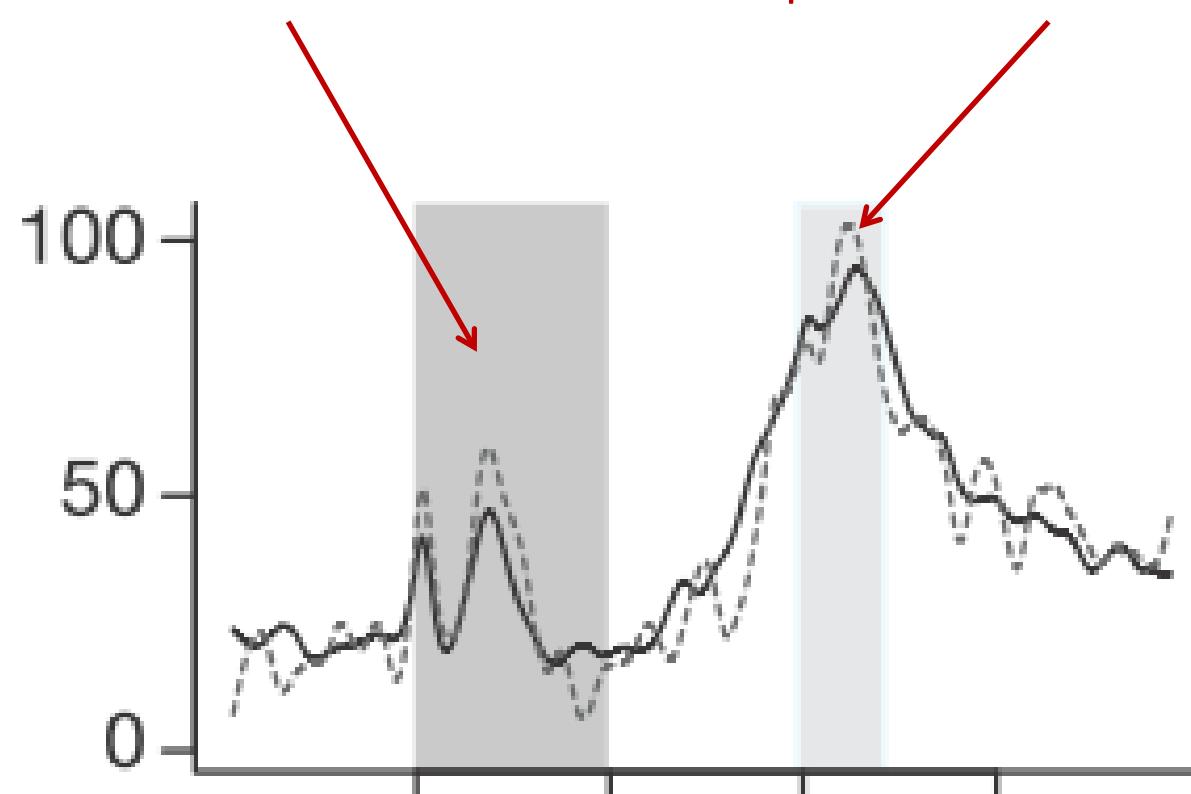
Independent of outcome →  
candidate prediction signal?  
*Which neurotransmitter?*



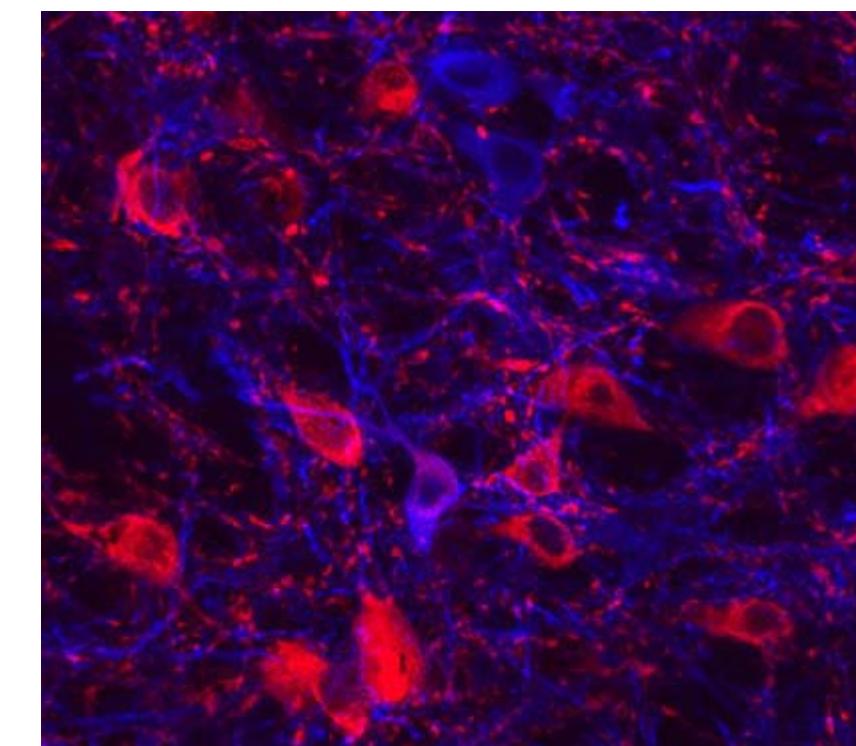
Classic dopaminergic RPE

# GABA signals prediction

cue predicting reward



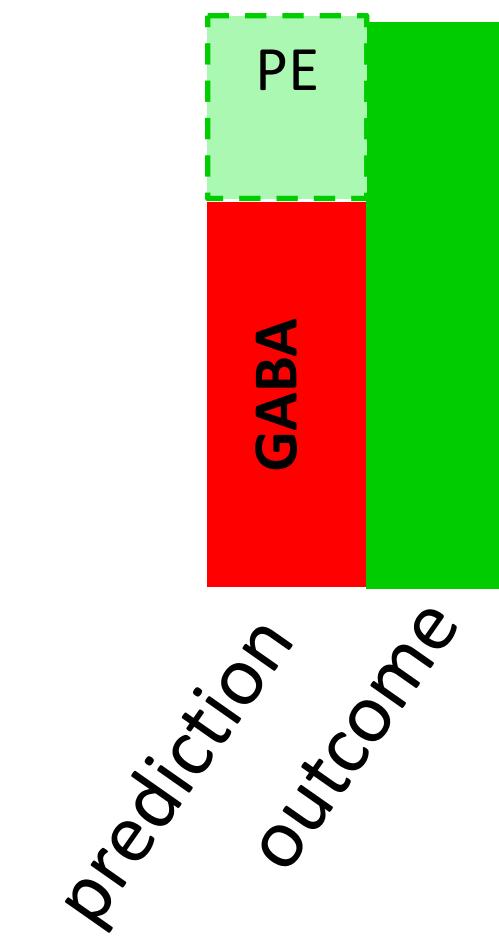
GABA peak @ time of expected delivery



Dopamine neurons (55-65 %)  
GABAergic neurons (30-40 %)

GABA sets ‘baseline inhibition’ for driving inputs on DA neurons to overcome:

Expected reward      Surprising reward



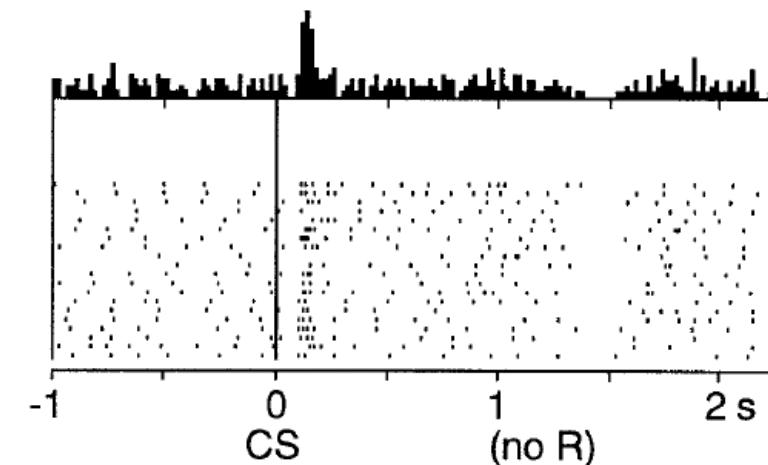


# Where are we

conditioning phenomena like blocking suggest learning to predict reward is **error-driven**



dopamine neurons (& fMRI signals at recipient structures in humans) appear to code **reward prediction errors**



next:  
what about **more complicated** tasks?





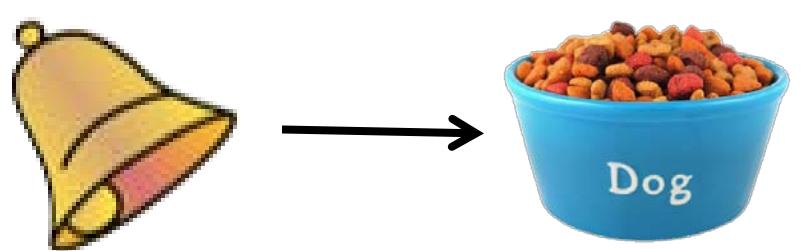
# Outline

- processing reward
- introduction to dopamine
- learning about rewards
  - error-driven learning
  - neural basis
- learning action
  - TD learning
  - law of effect
  - second-order reinforcement
  - multiple decision systems

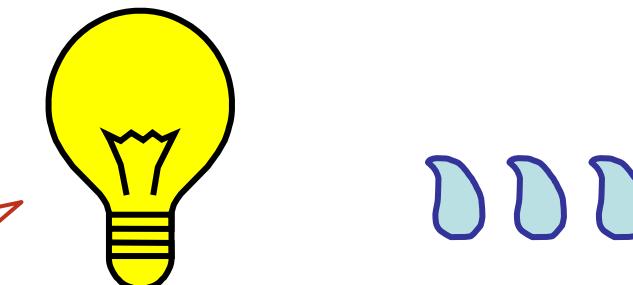
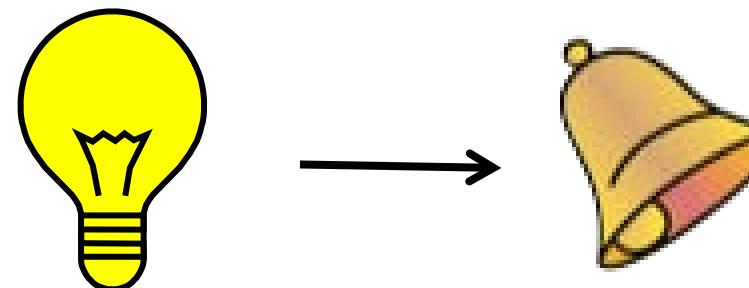
# Problems for Rescorla-Wagner: 2<sup>nd</sup> order conditioning



Phase 1



Phase II



**What does R/W predict?**  
Prediction error only at reward

**Solution: temporal difference learning**



# Temporal difference learning

## Rescorla-Wagner

- Aim:  $EV_t = r_t$
- Use prediction error:  $(r_t - EV_t)$

Predict of reward on trial  
n:  
t = discrete trials

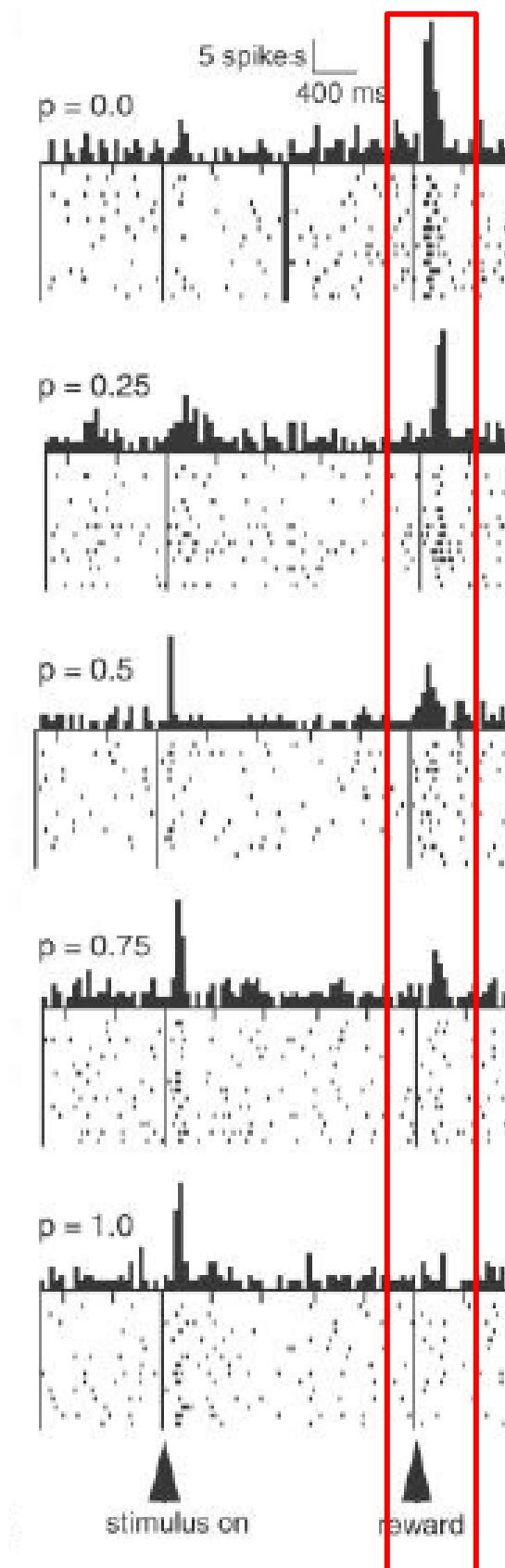
## Temporal difference learning (*Sutton & Barto*)

- Aim :  $EV_t = r_t + r_{t+1} + r_{t+2} + r_{t+3} + \dots$   
 $= r_t + EV_{t+1}$        $\leftarrow$  (clever recursive trick)
- Prediction error:  $[r_t + EV_{t+1}] - EV_t$
- learn consistent predictions based on temporal difference  $EV_{t+1} - EV_t$ 
  - if  $EV_{t+1} = EV_t$ , my predictions are consistent
  - if  $EV_{t+1} > EV_t$ , things got unexpectedly better
  - if  $EV_{t+1} < EV_t$ , things got unexpectedly worse

Predict ALL cumulative  
future reward:  
t = continuous

A prediction error is a  
change in the expectation  
of all future rewards.  
Thus, predictive cues act  
like reward to generate  
prediction errors and  
learning

# R/W rule explains DA firing @ outcome

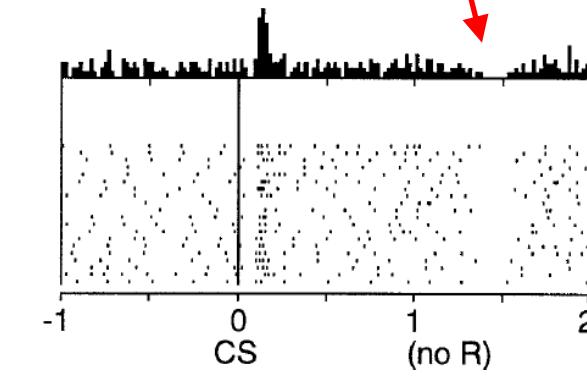


reward following  
0% predictive cue

reward following  
50% predictive cue

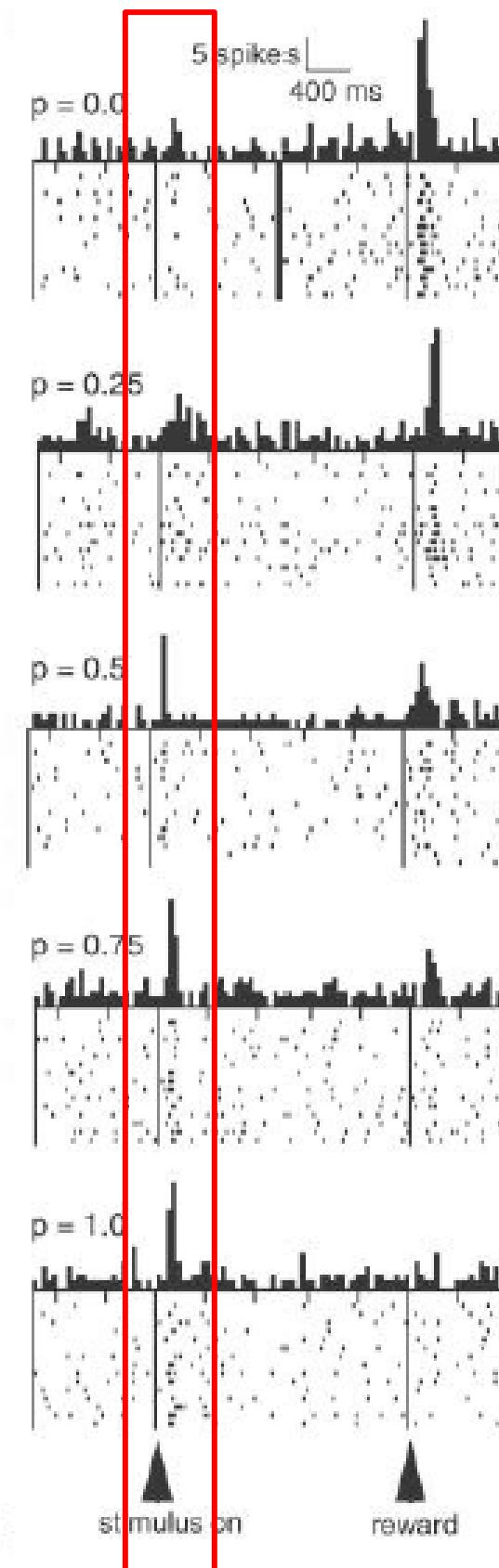
reward following  
100% predictive cue

Prediction error:  
 $r_t - EV_t$



no reward following  
100% predictive cue

# TD learning needed to explain firing @ cue

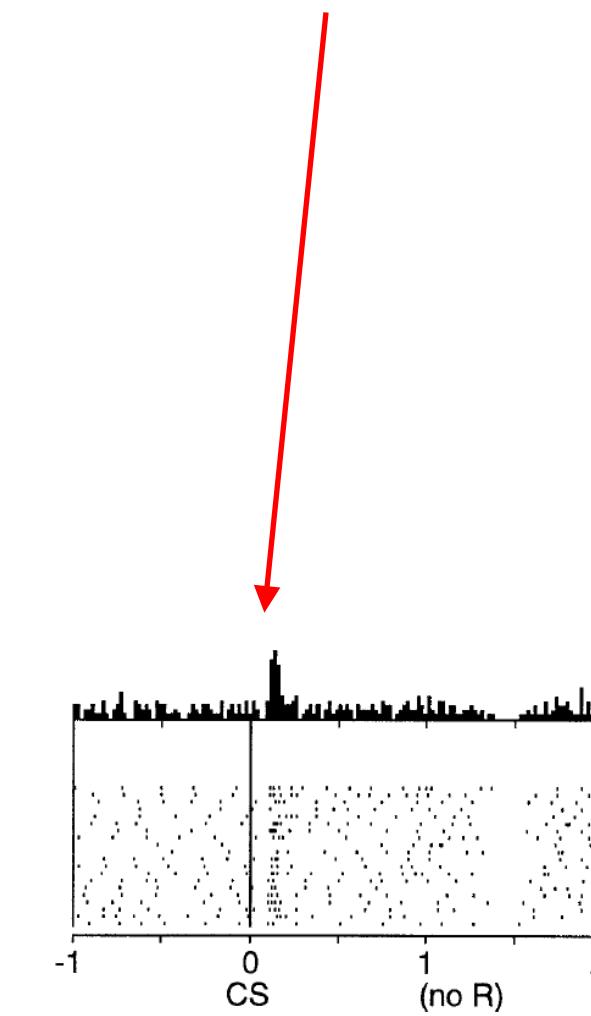


0% predictive  
cue: no chance at  
reward

50% predictive cue:  
I'll likely get a reward!

100% predictive cue: I'll  
certainly get a reward

These are prediction errors too!  
Telling you how much your future  
just got better  
 $(r_t + Ev_{t+1}) - EV_t$

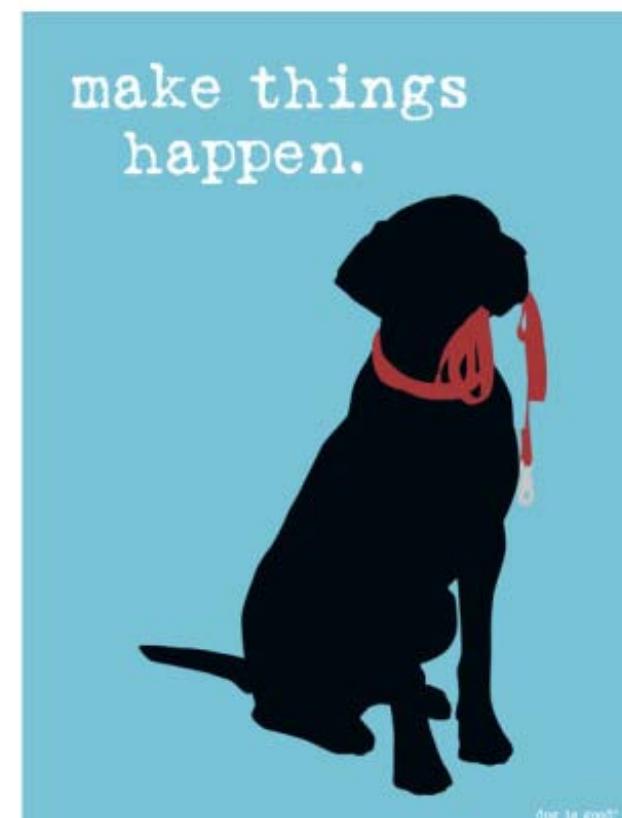


100% predictive cue:  
I'll certainly get a reward

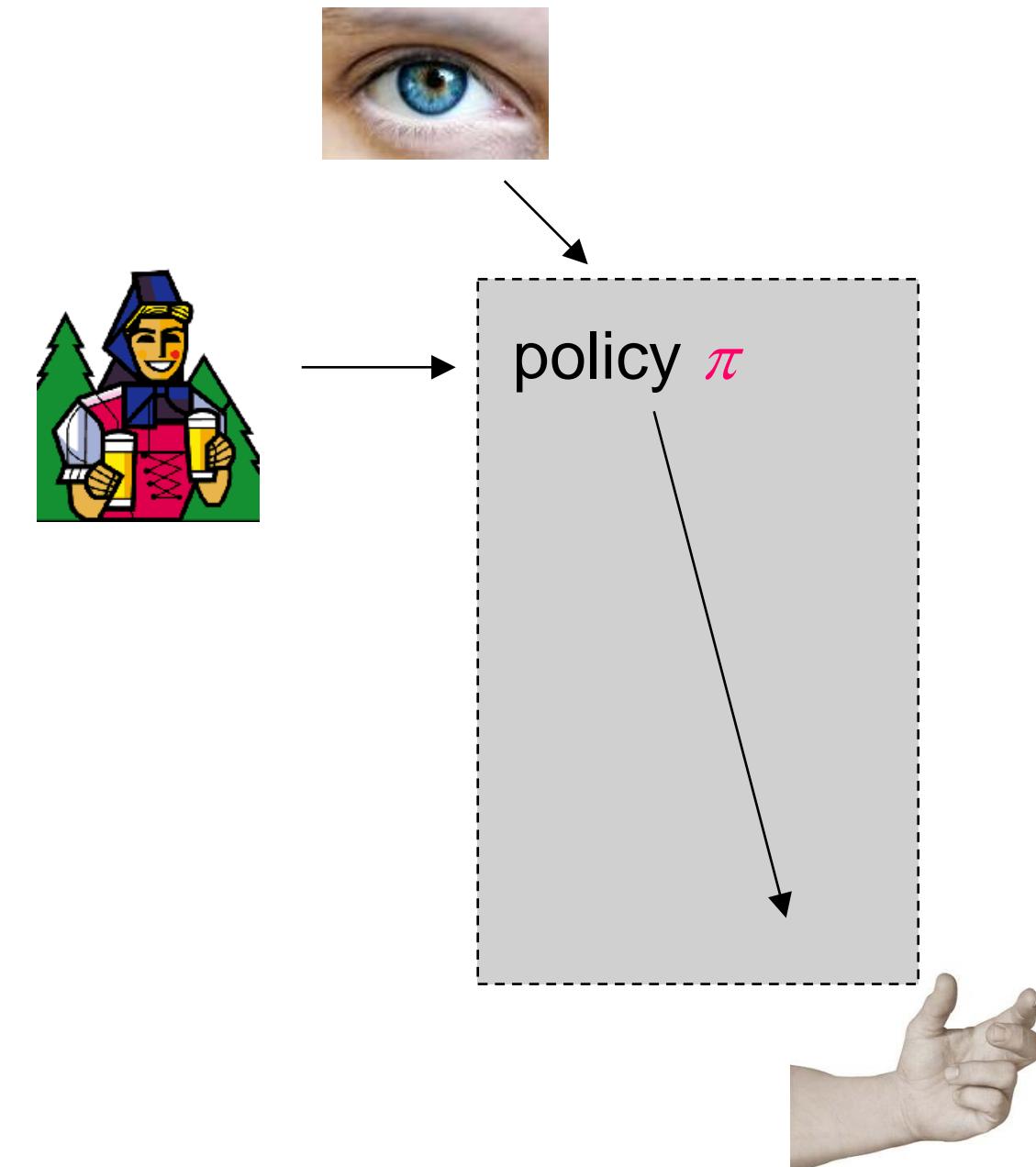
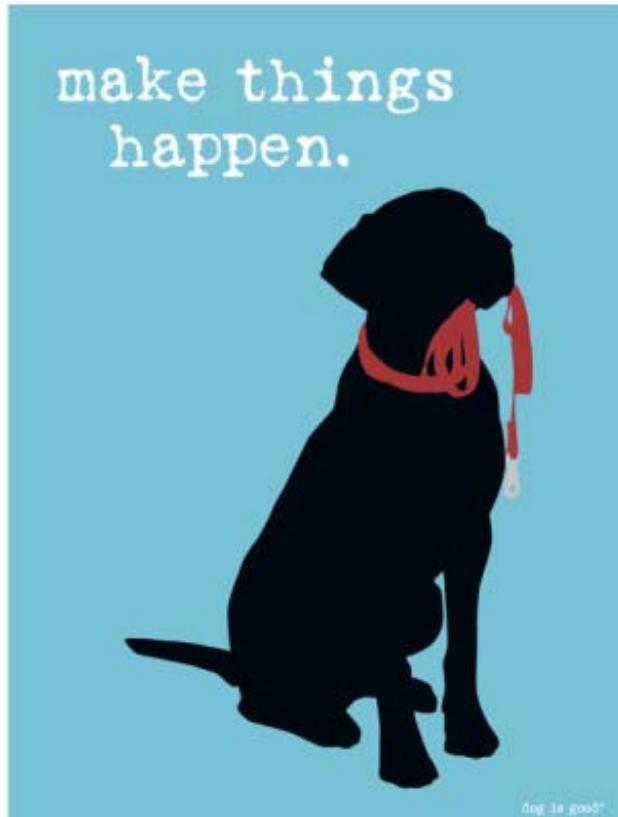


TD learning allows us to understand how we learn which stimuli predict future rewards....

But how do we learn which actions will help us to get the rewards?



# Law of effect: learning value of actions (not cues)



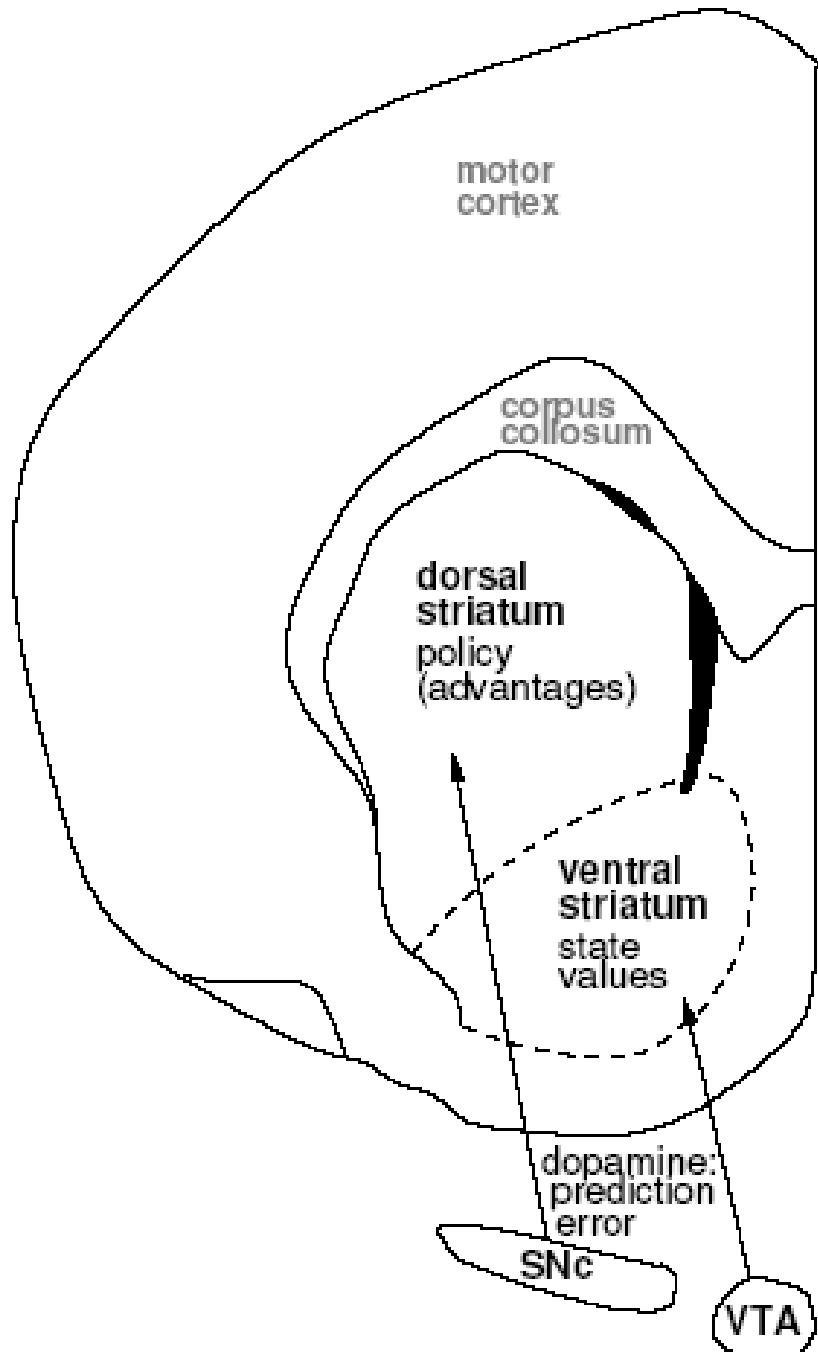
*“Of several responses made to the same situation, those which are accompanied or closely followed by satisfaction to the animal will, other things being equal, be more firmly connected with the situation, so that, when it recurs, they will be more likely to recur.”*

Thorndike (1911)



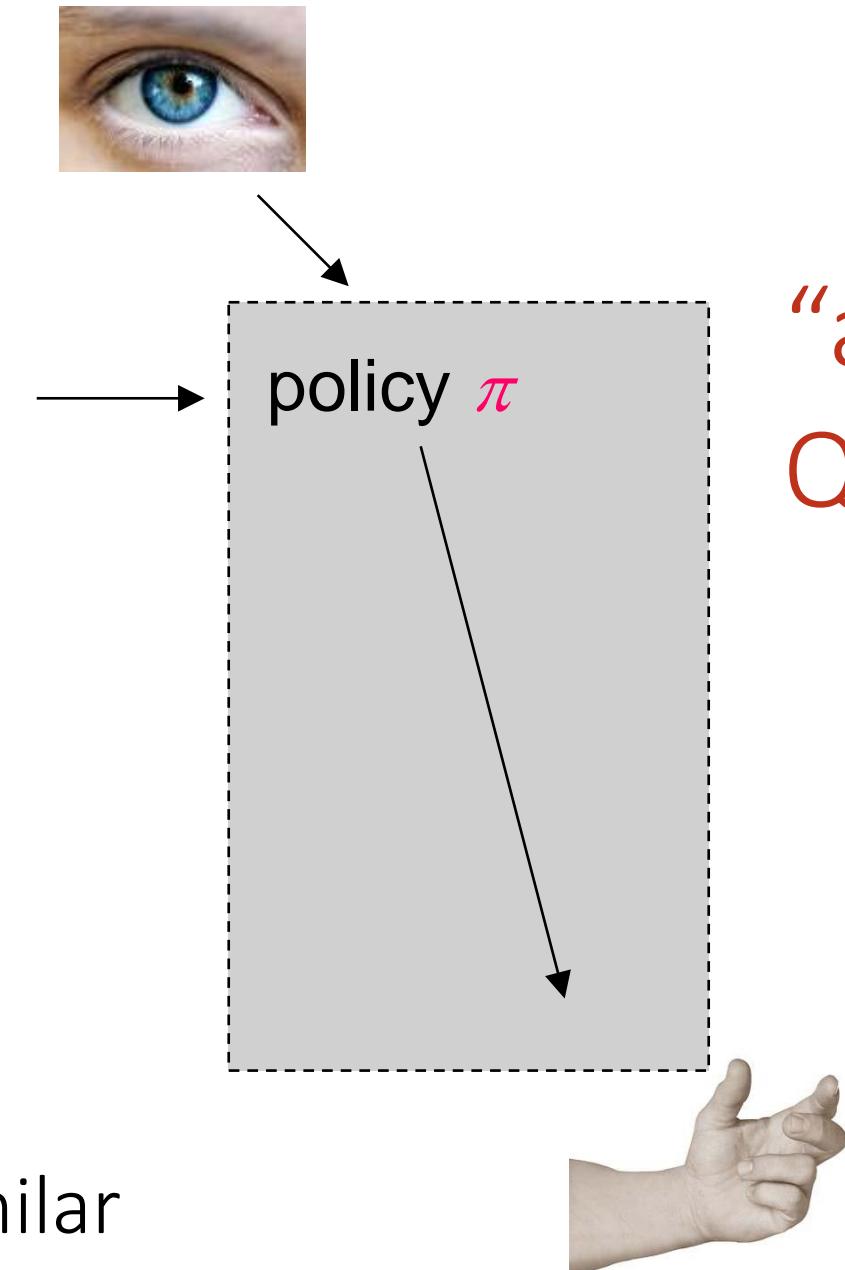
# Actor/critic

- Learning actions, learning values



**dopamine signals**  
ventral/motivational &  
dorsal/motor striatum are similar  
  
suggestion: DA prediction errors  
train both values & policies

“critic”  
 $V_{t,s}$

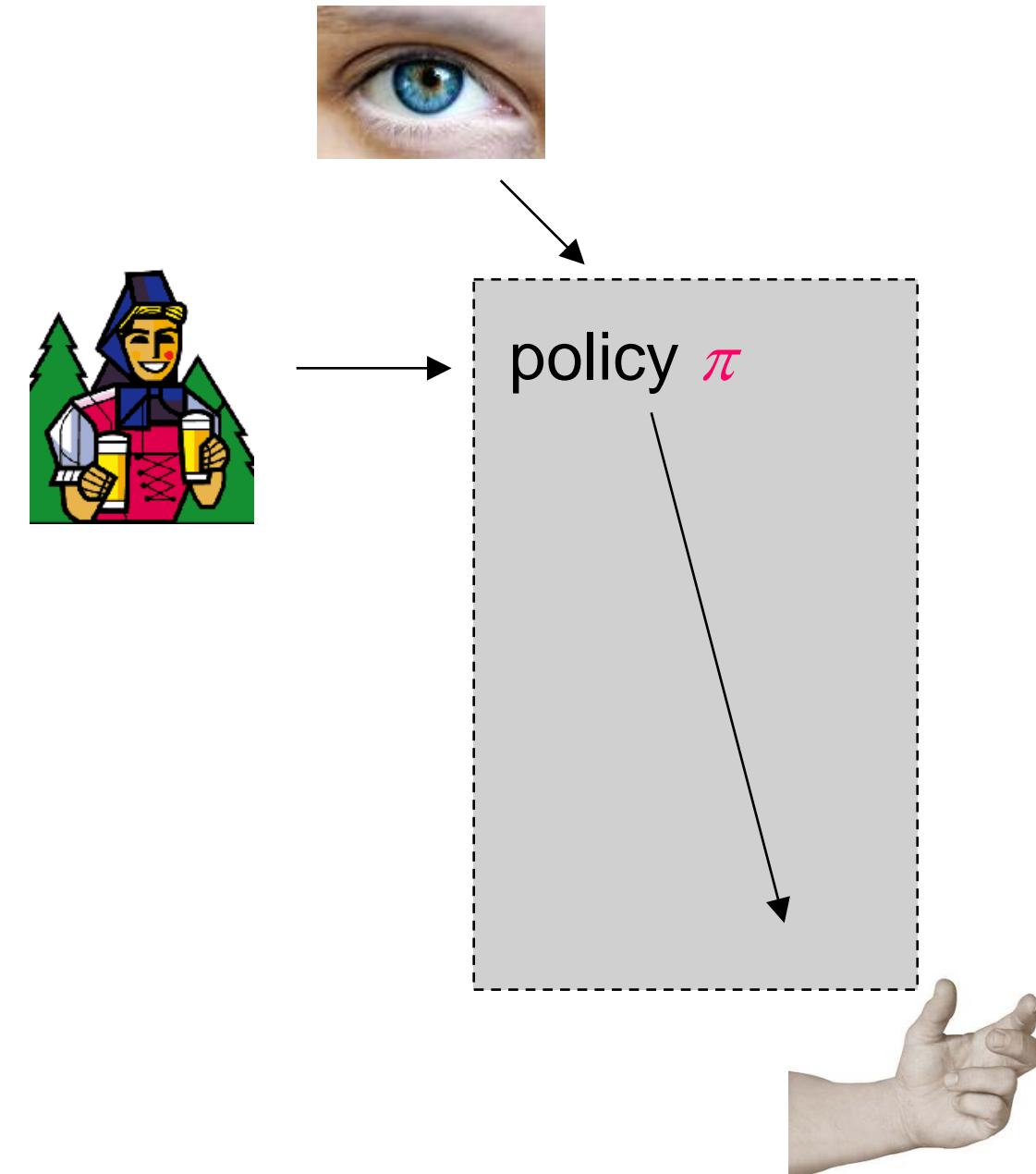
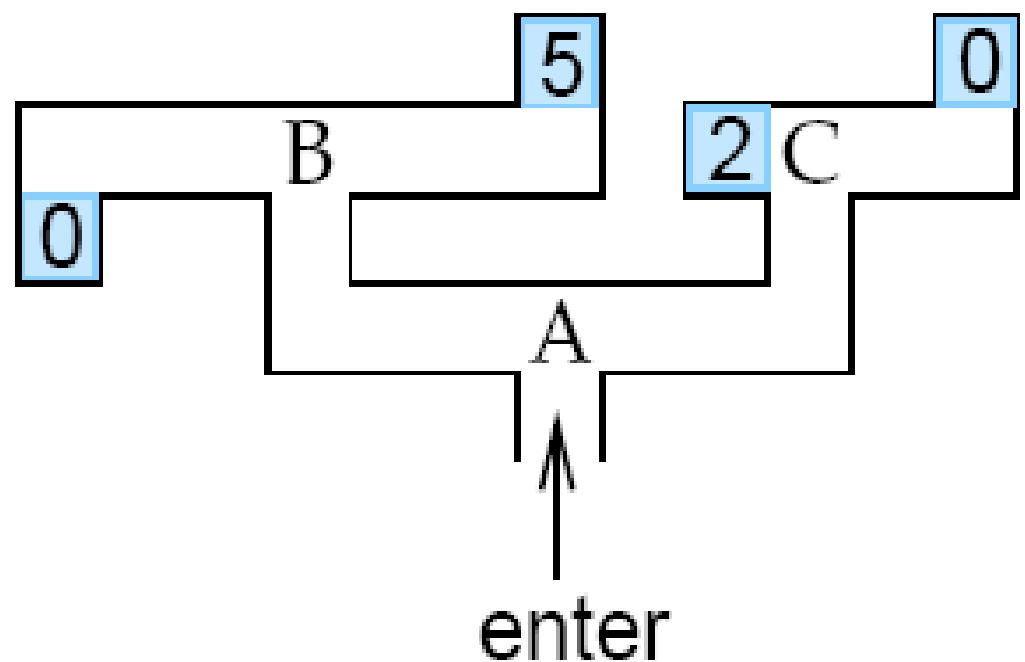


“actor”  
 $Q_{t,s|a}$

# What's the problem with law of effect?

Credit assignment:

- How do know which action led to the result?



# Again we need TD learning

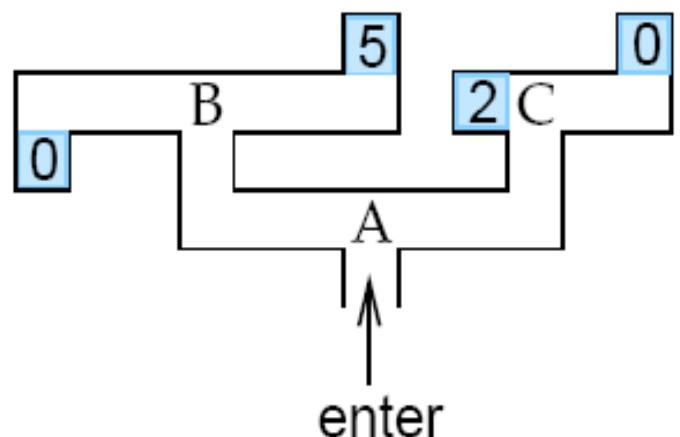
reinforce actions not by immediate reward  $r_t$

but by estimated future value

$$V_t = r_t + V_{t+1}$$

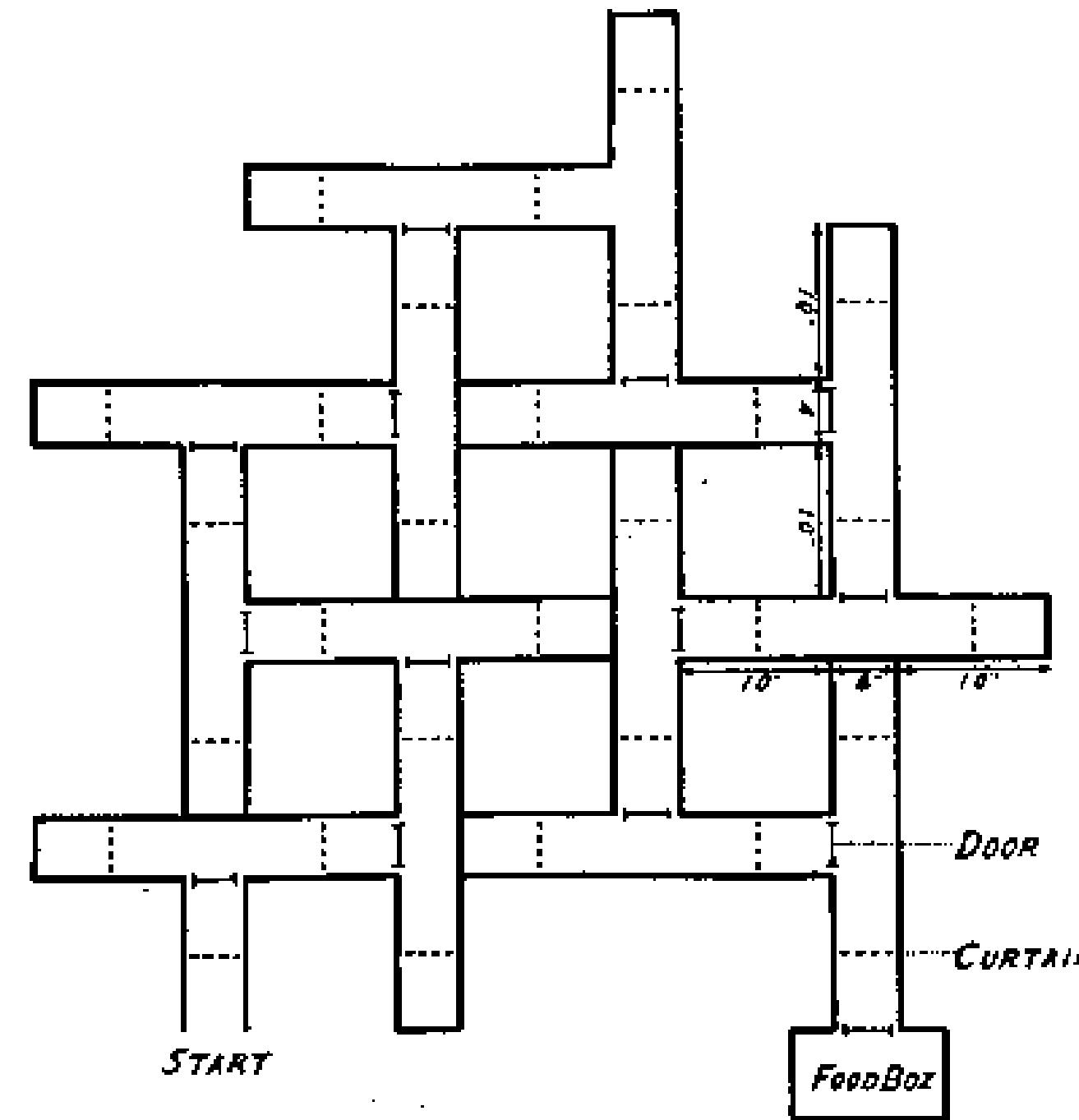
or prediction error:

$$V_t = r_t + V_{t+1} - V_t$$

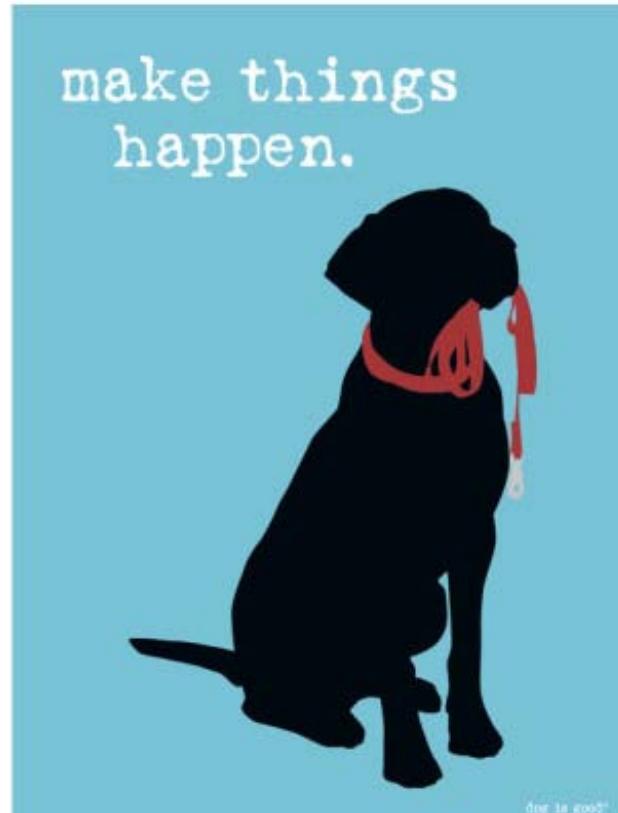


... this is how prediction helps to solve the credit assignment problem

But how to learn long sequences of actions?



# Cognitive Maps



## Early critique of S-R approach, birth of cognitive psychology

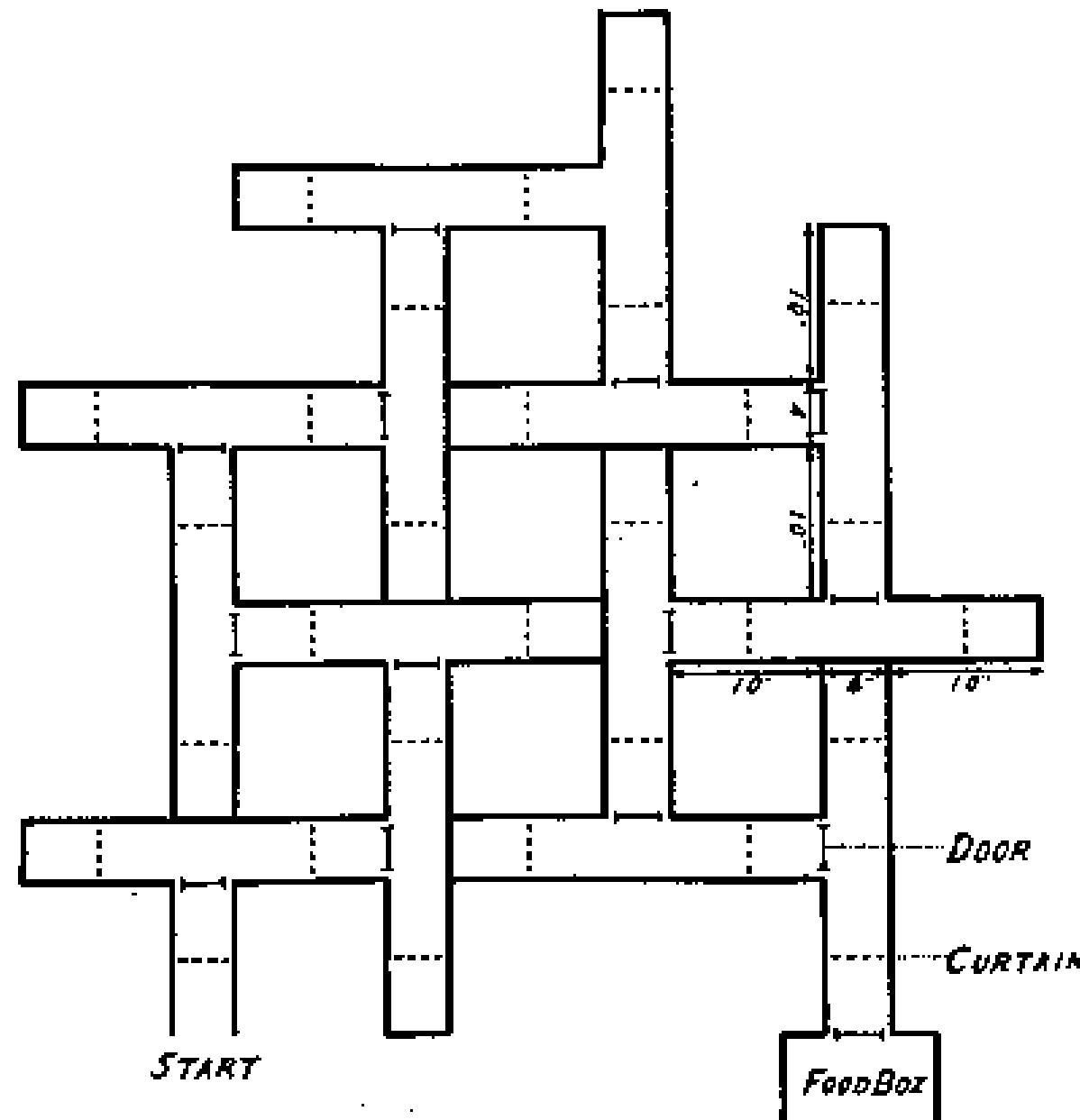
*"The stimuli are not connected by just simple one-to-one switches to the outgoing responses. Rather, the incoming impulses are usually worked over and elaborated in the central control room into a tentative, cognitive-like map of the environment. And it is this tentative map, indicating routes and paths and environmental relationships, which finally determines what responses, if any, the animal will finally release."*

Tolman (1948)





# Maze task



Evidence for cognitive maps?

Train rats to run a maze for food

compare this to a group that had been pre-exposed to maze but without food, then run for food in subsequent test

what will happen?

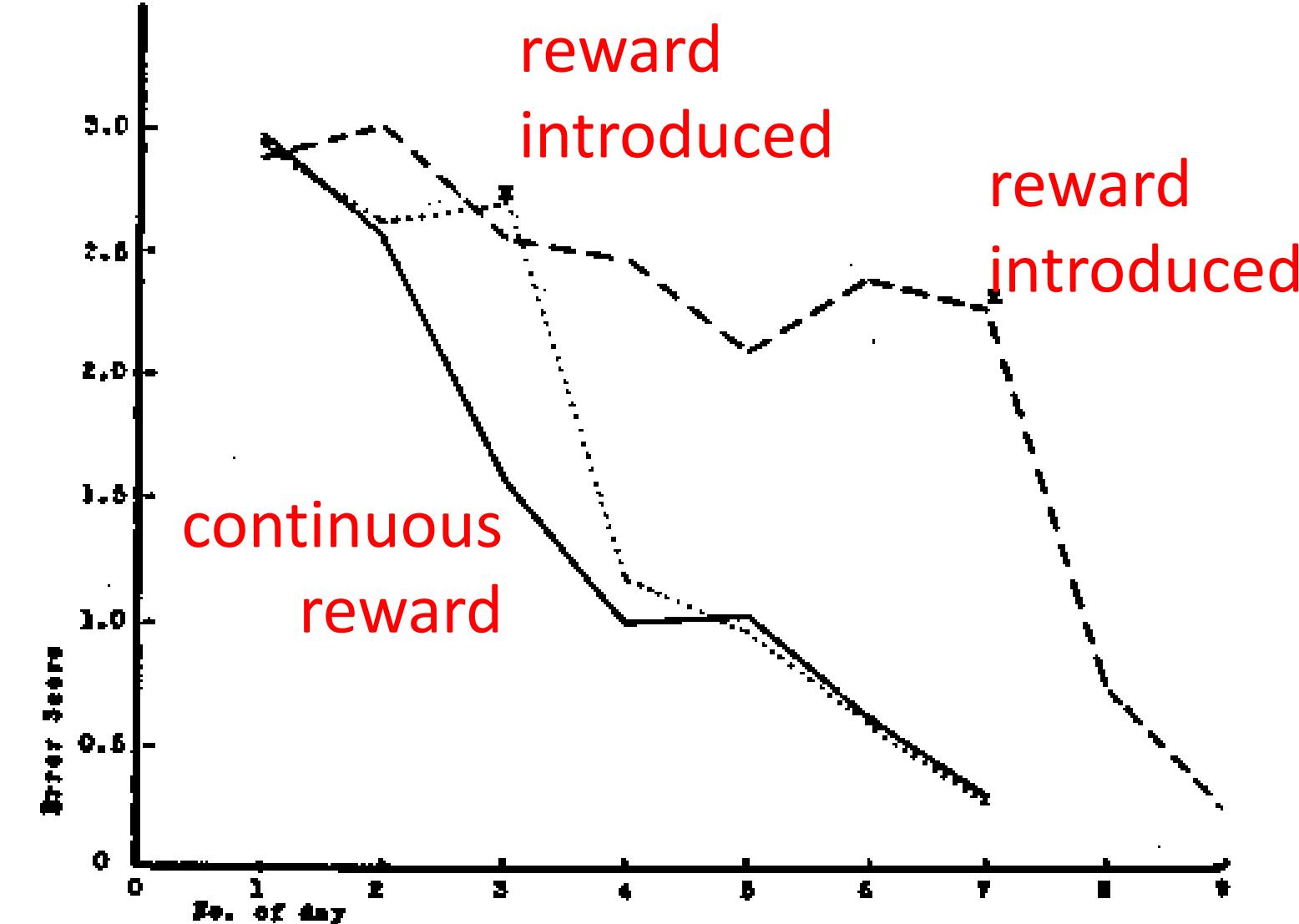
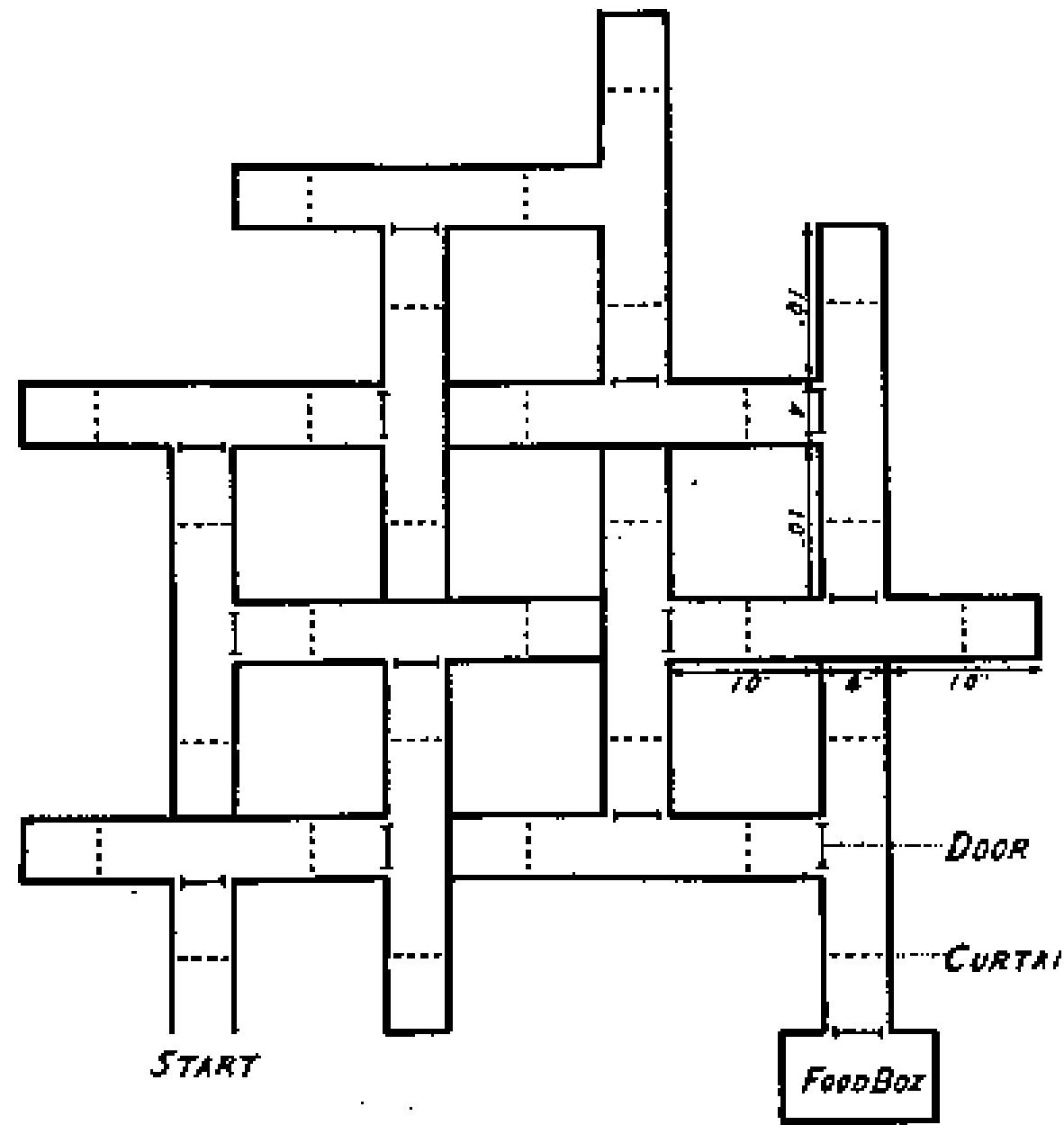
what does this demonstrate?

# Latent learning

Animals learn during unrewarded exposure

→ not just stamping in S→R

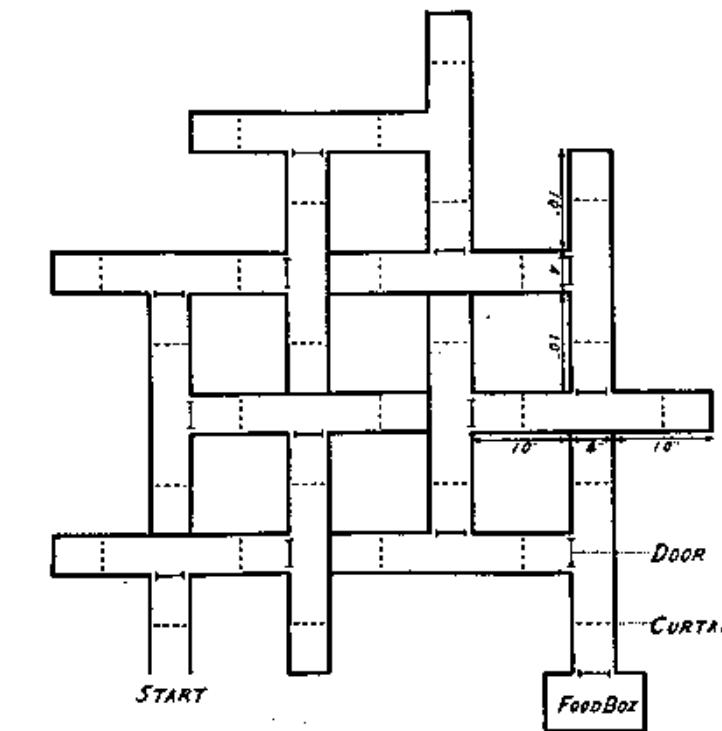
→ called latent learning



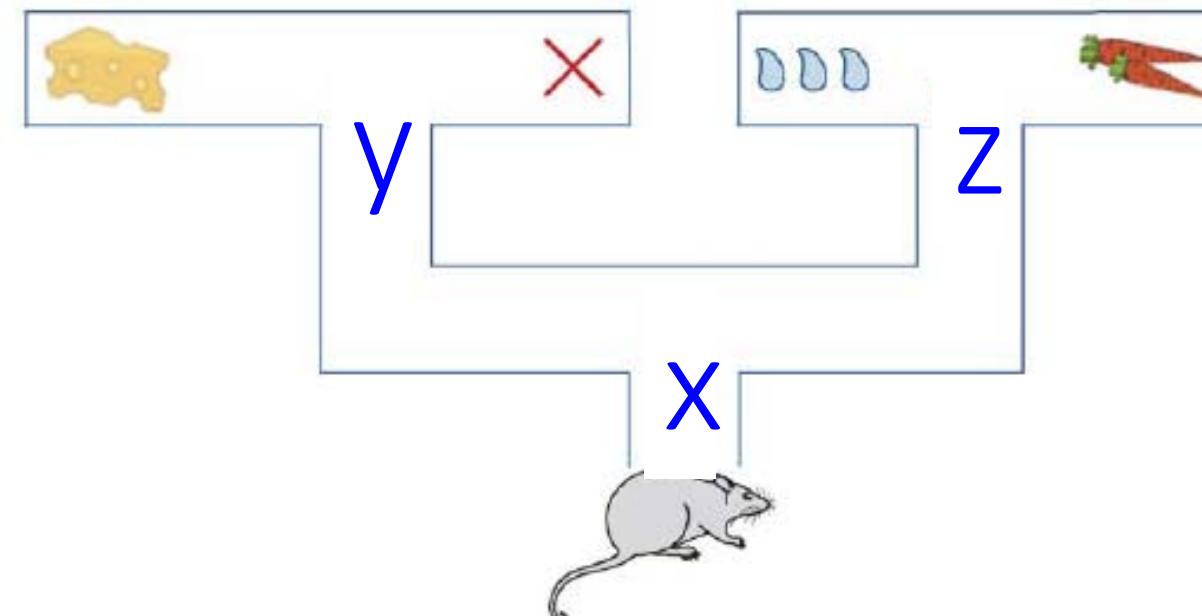


# Interim summary

- Even the humble rat can learn & internally represent spatial structure, use it to plan flexibly
  - TD learning (dopamine-style) **can't do this**
  - Note that spatial tasks are really complicated & hard to control
- Next: search for modern versions of these effects, relevance for dopamine/TD
- Key questions:
  - is S→R (TD) learning ever relevant?
  - what is there beyond it?



# Sequential decision tasks

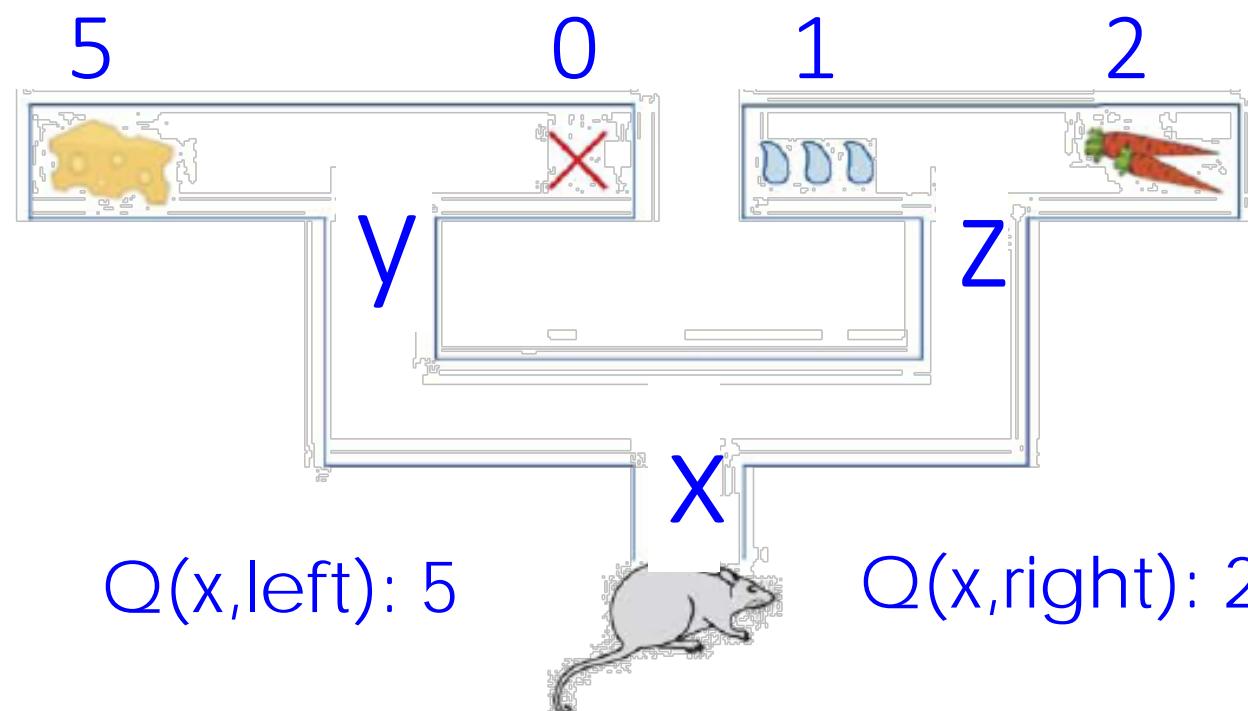


You're a rat dropped in an unknown **sequential** task

Want to learn to choose good actions

→ must predict **long term consequences** of action

# Sequential decision tasks



predict summed future rewards:

$$\begin{aligned} Q(x,\text{left}) &= r(x,\text{left}) + r(y,a_y) & [Q \text{ instead of } V \text{ when actions are involved}] \\ &= 0 & + 5 \end{aligned}$$

TD learning with actions:

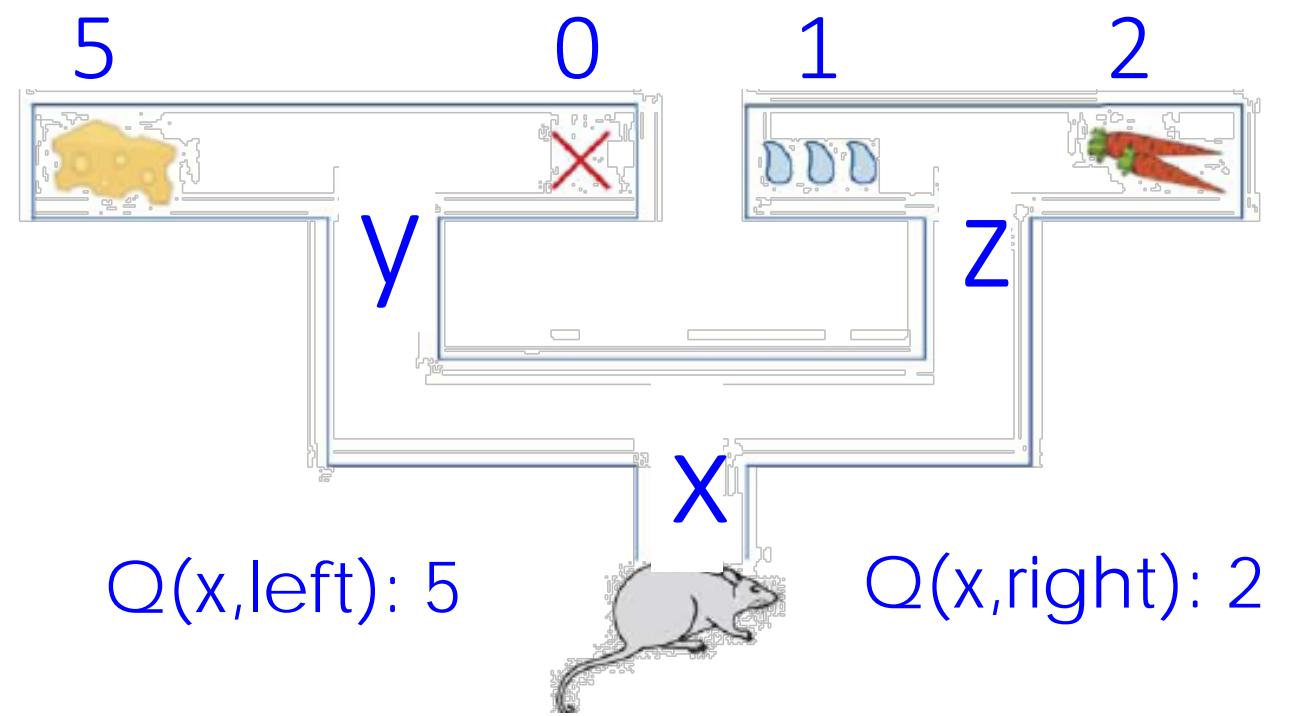
$$Q(s_t, a_t) = r(s_t, a_t) + Q(s_{t+1}, a_{t+1})$$

immediate reward

future reward(s)

Because TD learning learns from all expected future rewards, rat learns the values of turning left and turning right in state x

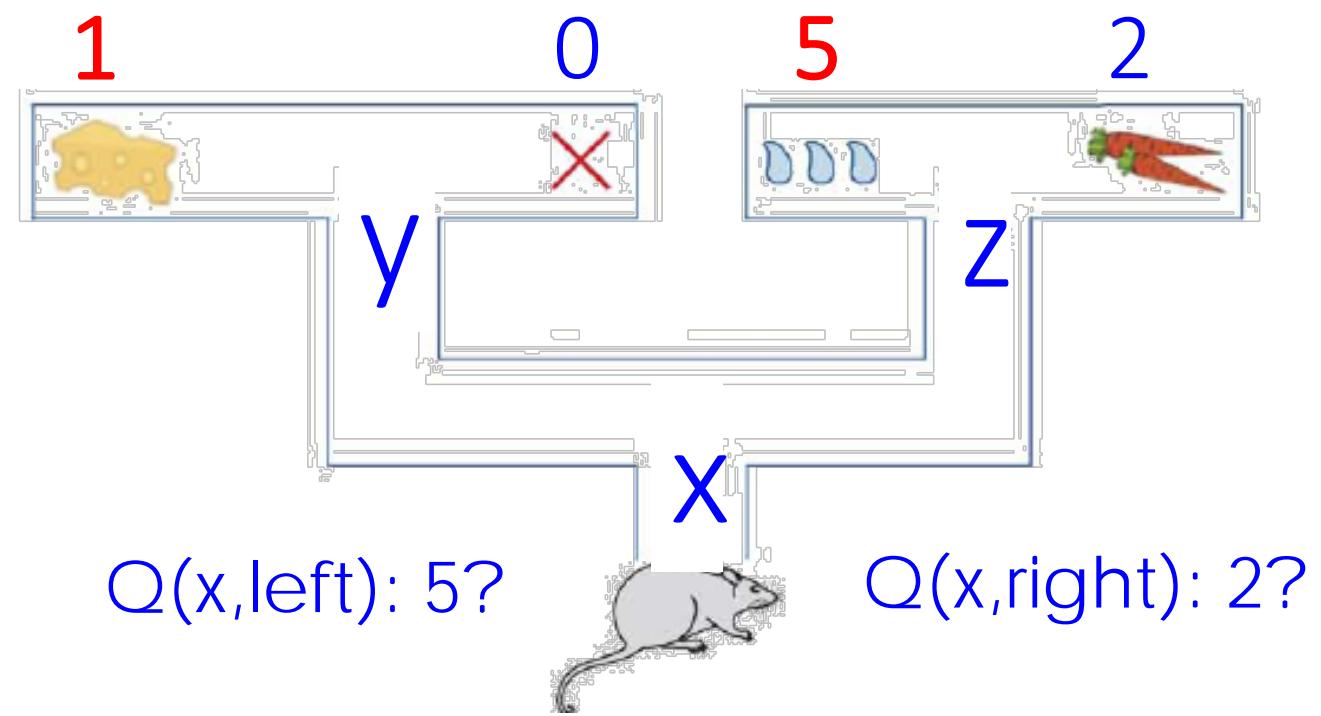
# Sequential decision tasks



What happens if the rat suddenly stops being **hungry**, and gets **thirsty**?



# Sequential decision tasks



What happens if the rat suddenly stops being **hungry**, and gets **thirsty**?

$Q(x,\text{left})$  and  $Q(x,\text{right})$  just measure how much utility I derived from the action in the past

- they do not represent the different specific outcomes (cheese, water, carrots)
- so if my motivation changes (if I become thirsty) I can't immediately update them

Agent must take the actions again and learn their new consequences. This makes a **weird** prediction

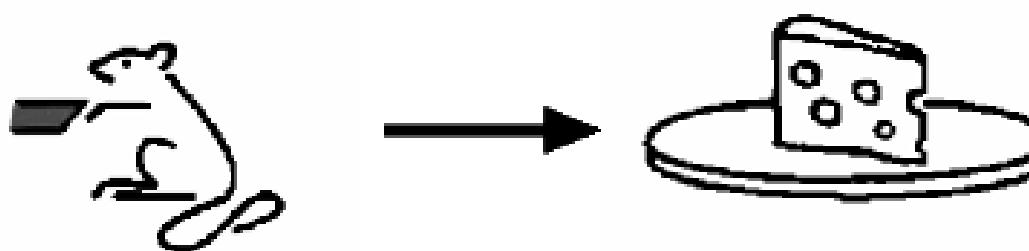
- behaviour should be blind (without retraining) to **changes in outcome value**
- .... so agents choose options they don't want?!?



# Testing this prediction

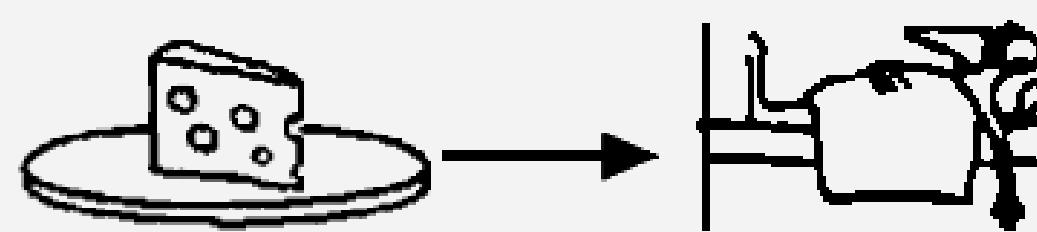
Stage

**1. training**  
(hungry)



learn to leverpress for food (choose work or not)

**2. devaluation**



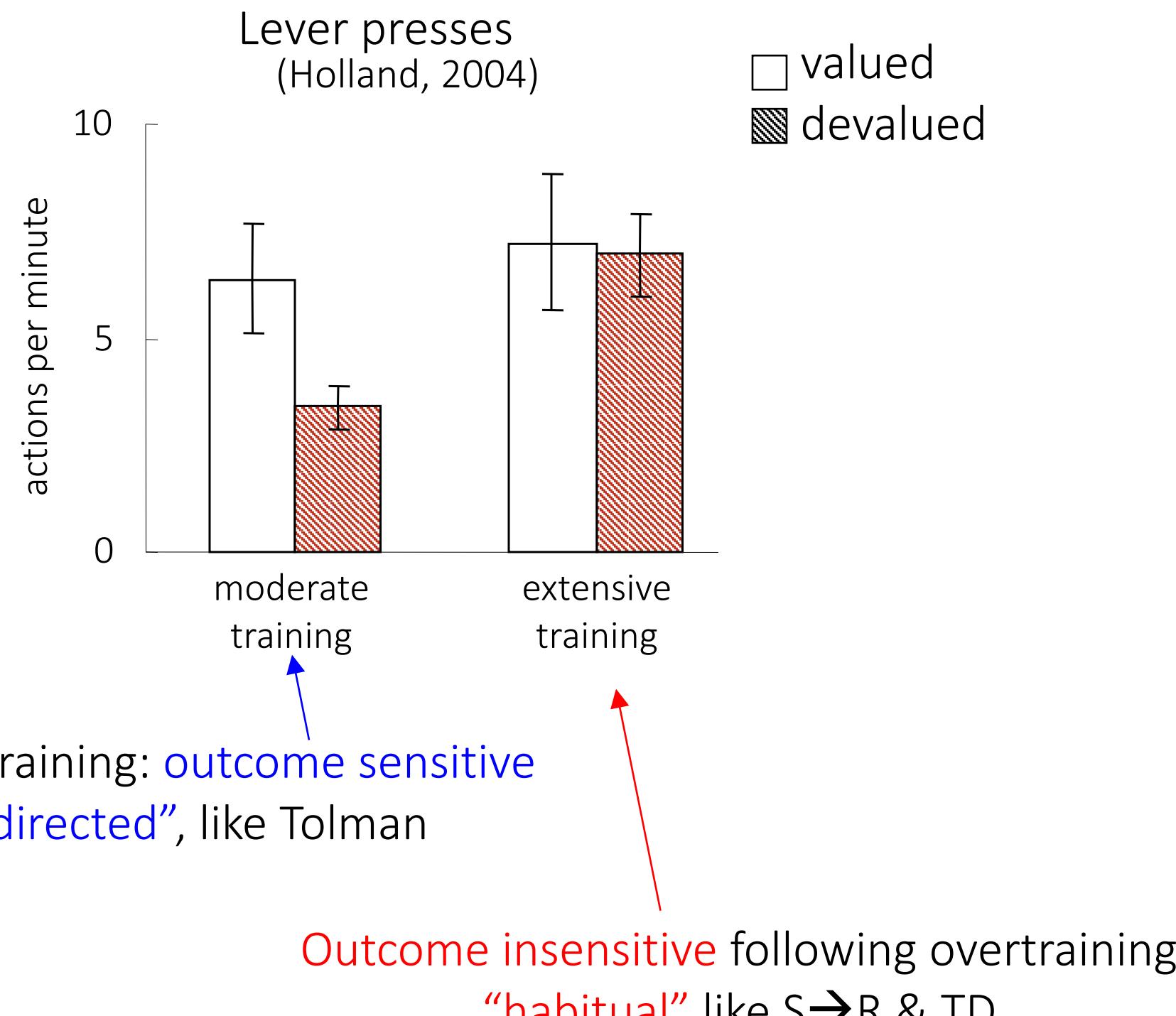
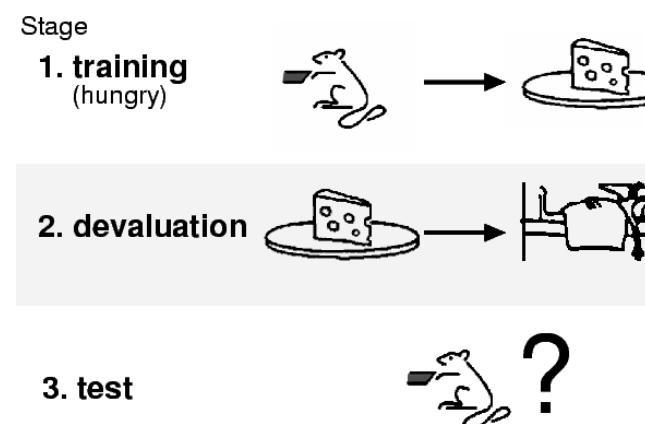
pair food with illness;  
develop aversion

**3. test**



will animals work for food **they don't want?**

# Results



Animals will work for food they don't want, **sometimes**  
 → familiar counterpart: actions become automatic with repetition

# Habitual and Goal-directed Actions

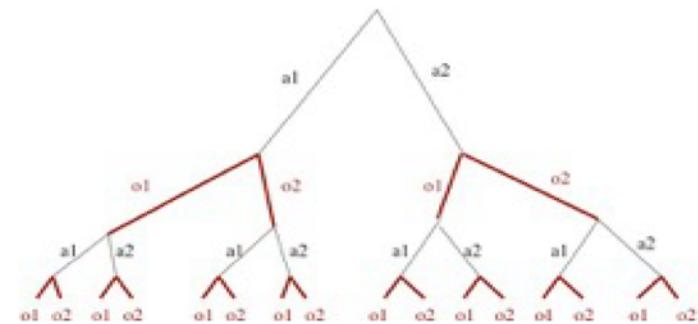
- Instrumental actions are of two kinds: **goal-directed** and **habitual** actions
- Goal-directed actions
  - Actions guided by evaluating their consequences
  - Sensitive to changes in the environment, but computationally expensive
- Habitual actions
  - Actions executed without deliberation over their future consequences
  - Less sensitive to changes in the environment, but fast and computationally efficient
- **Devaluation test** is crucial to dissociate them

# Parallel decision systems

## Computed values

### Goal-directed system

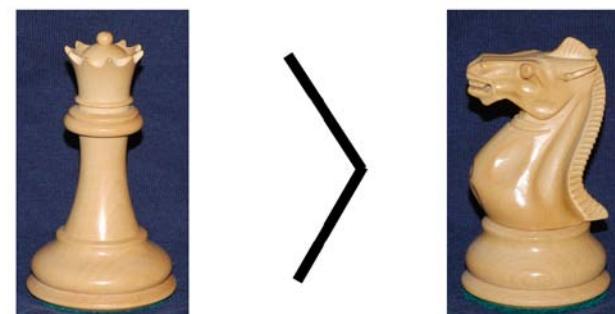
Tree search



## Cached values

### Habit system

Experience average



Exhaustive, exact, flexible  
computationally costly and slow

slow to learning, eventually correct,  
inflexible, computationally cheap



Neural basis of habits and goal-directed action

# Lesions

## Habits

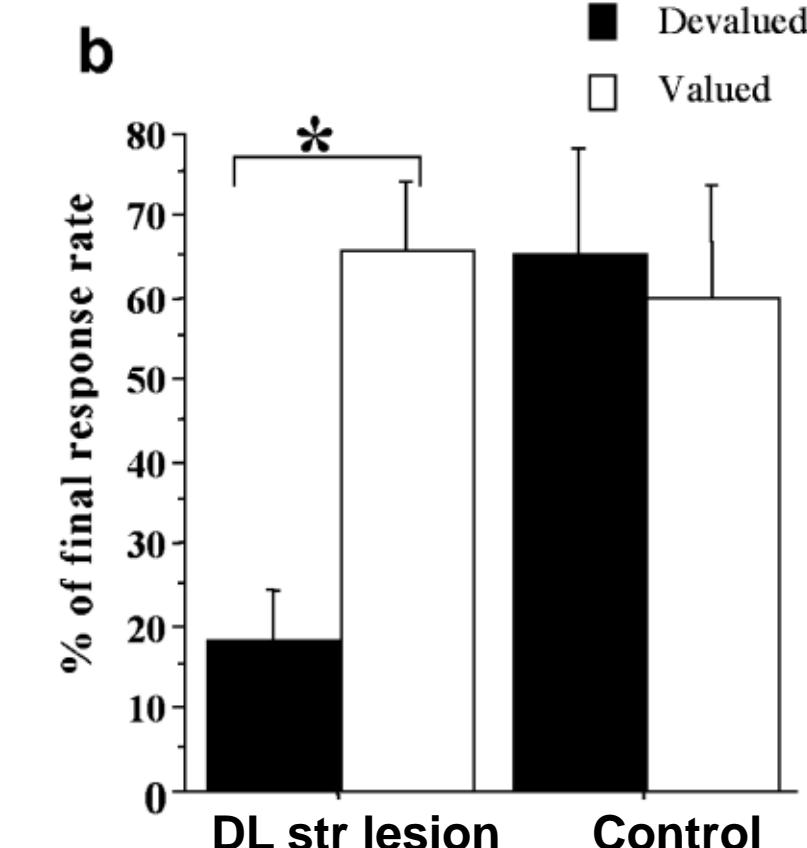
- With lesion of **dorsolateral striatum** rats acquire normally but never form habits: perpetually devaluation sensitive

## Goal-directed

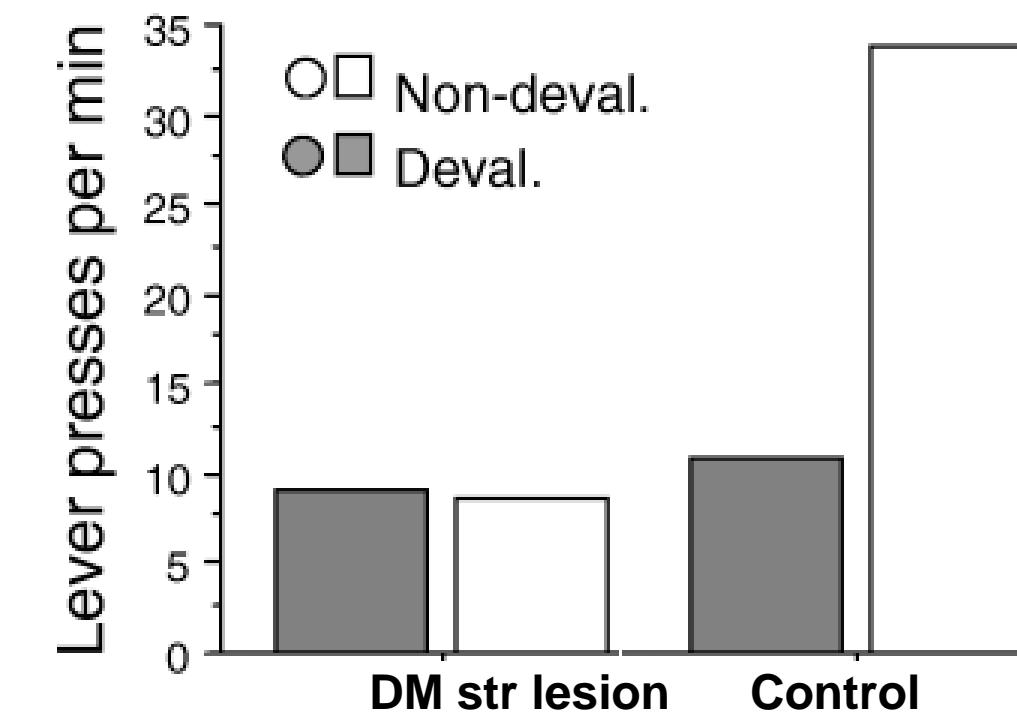
- Prefrontal areas and dorsomedial striatum** produce opposite pattern: even undertrained rats are habitual (devaluation insensitive)

→ Behavior arises from **dissociable neural systems**

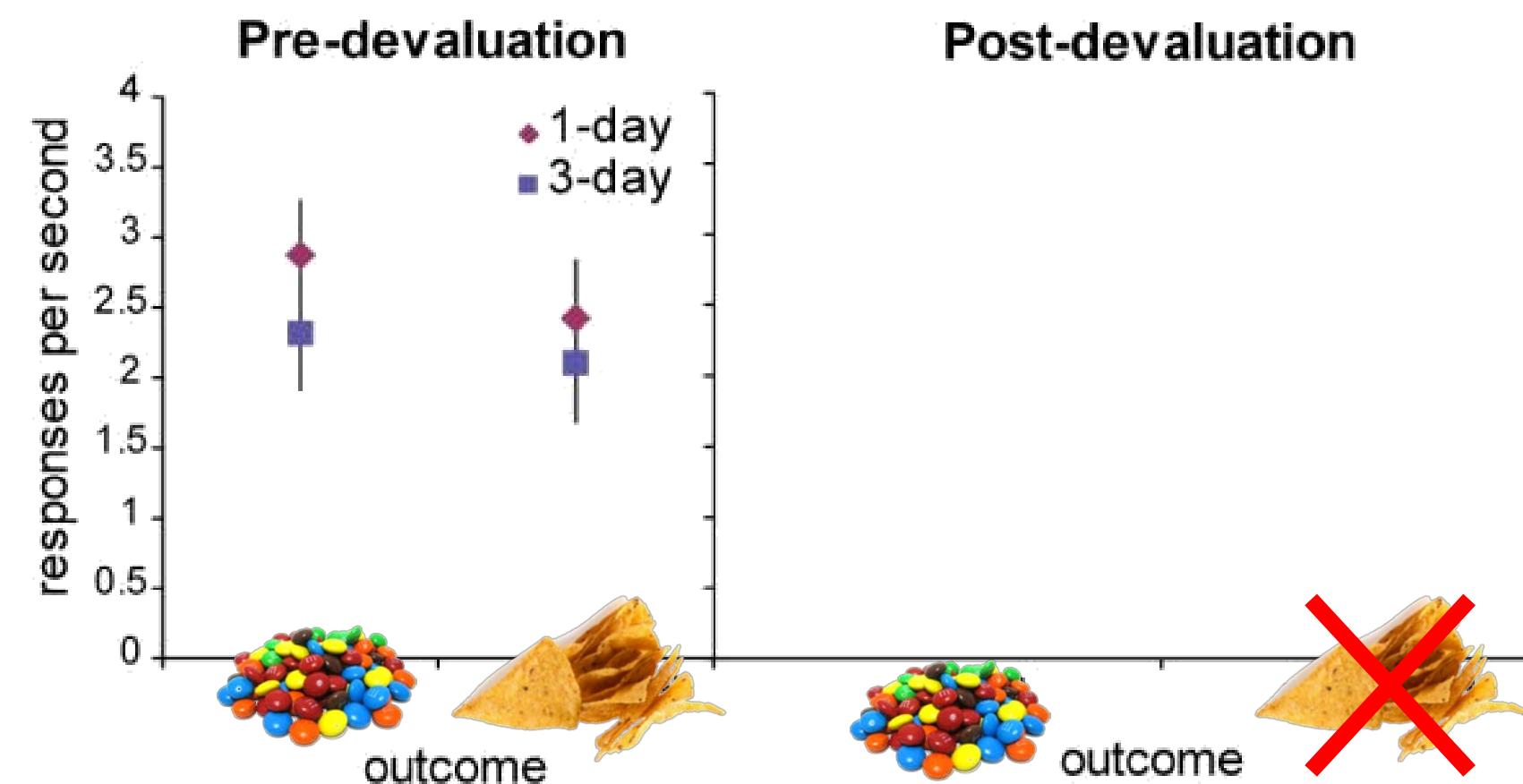
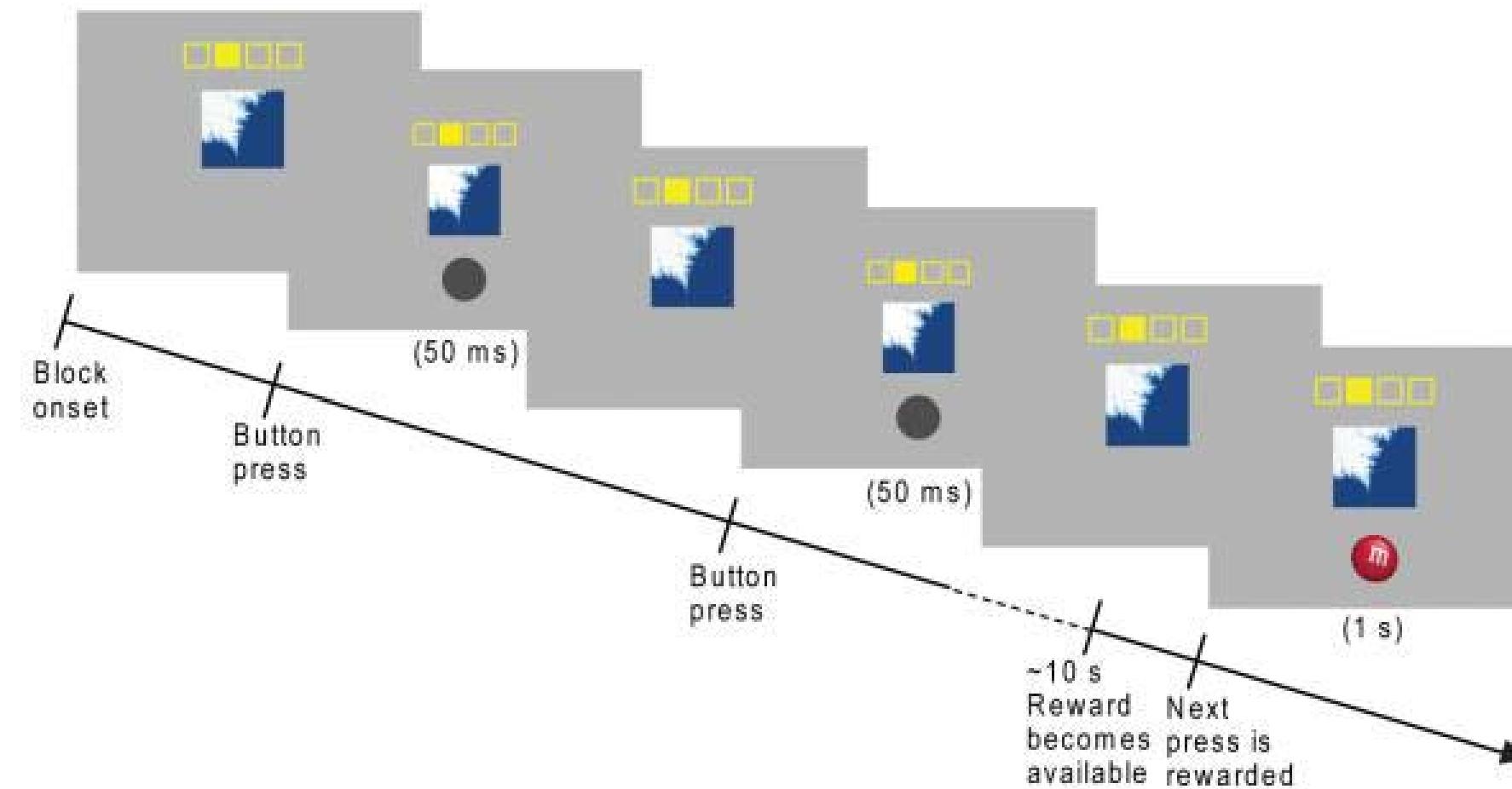
## Overtrained



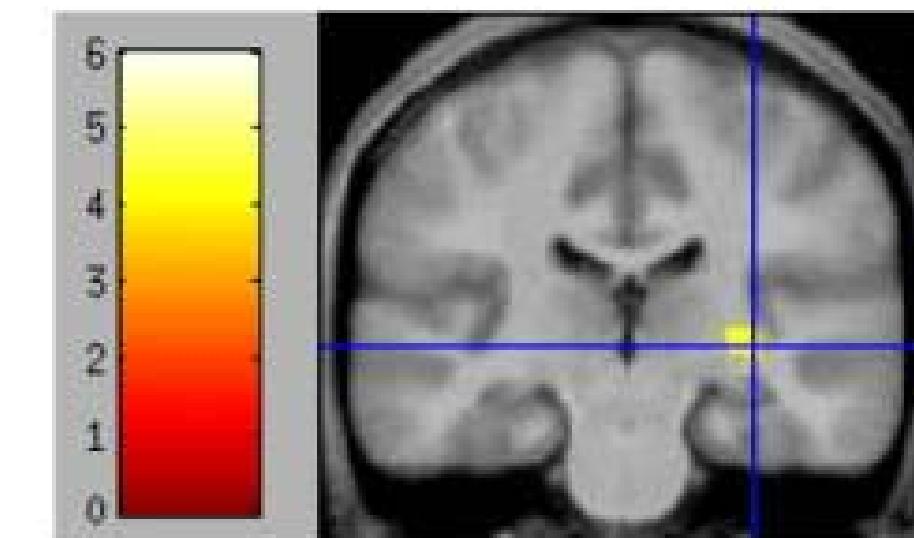
## Moderate training



# Actions and Habits in Humans



More DLS activity after overtraining



# Interim Summary: Habits vs. Goals

- same action can arise from **two behaviorally and neurally distinct systems**
  - habitual and goal-directed
  - can only distinguish with devaluation test
- **overtrained behaviour is devaluation insensitive**
  - “habitual”
  - as predicted by temporal-difference & S→R models
  - this is closely associated with what we think dopamine does
- moderately trained behaviour is devaluation sensitive
  - “goal directed”
  - demonstrates animals represent outcome, not just its cached value
  - reminiscent of Tolman cognitive map
  - probably non-dopaminergic
- possible to knock out either system with lesion; the other one takes over
  - parallel loops each involving areas of cortex and striatum
  - suggests parallel neural systems: **multiple action systems?**
  - **why is this such a crazy idea?**
  - what problems does this create?



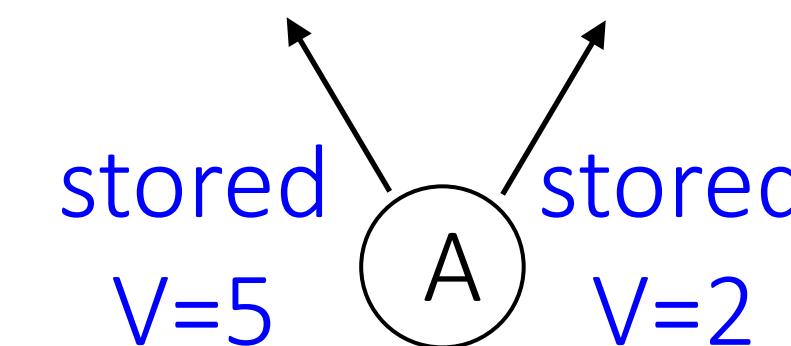
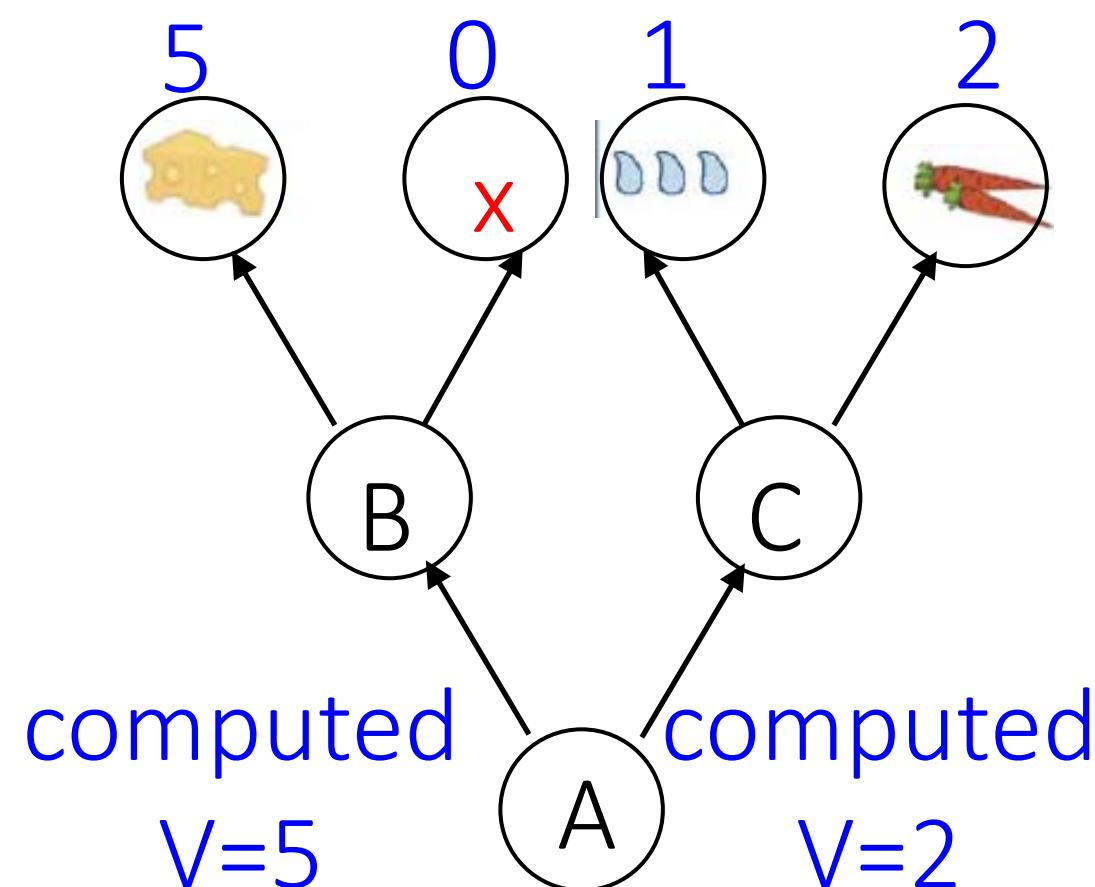
# Outcome sensitivity

## model-based

- can immediately adapt to value shifts
- like goal-directed

## model-free

- cannot immediately adapt
- like habits



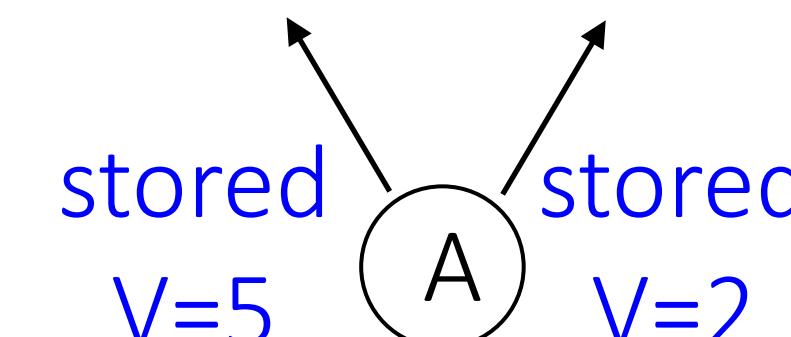
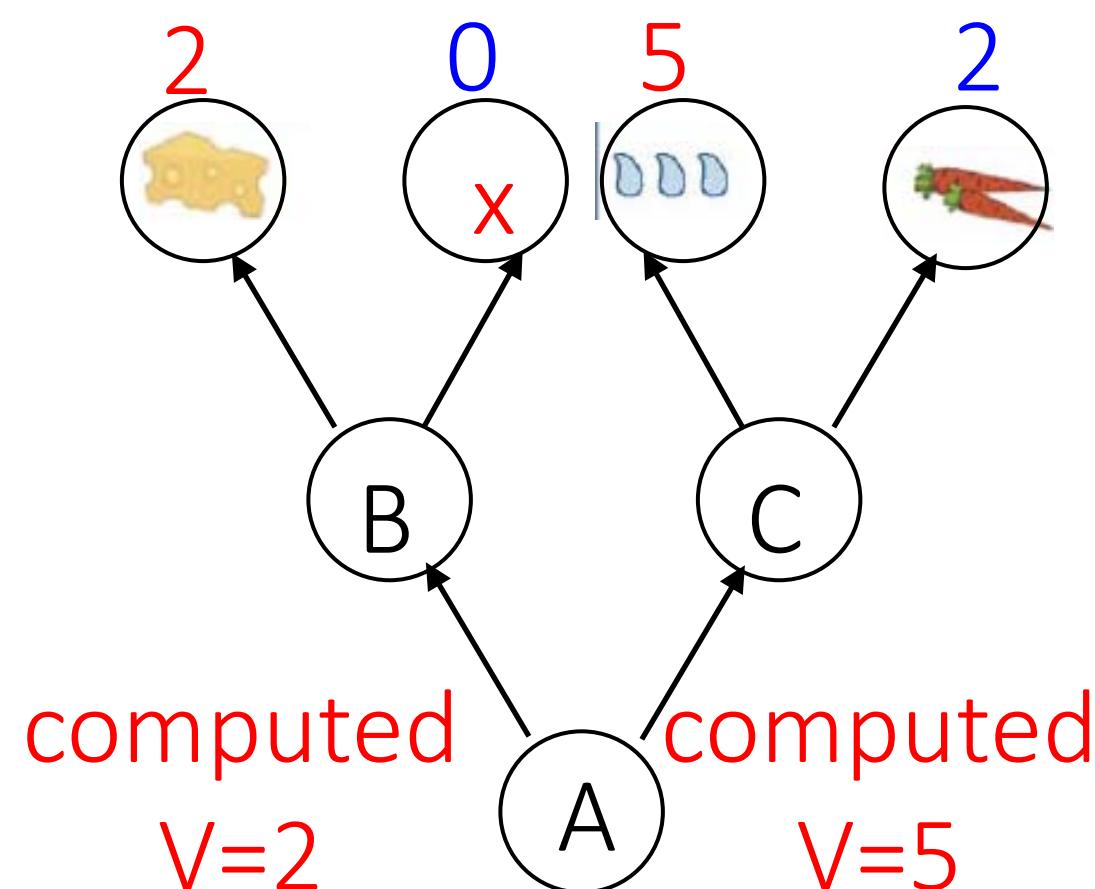
# Outcome sensitivity

## model-based

- can immediately adapt to value shifts
- like goal-directed

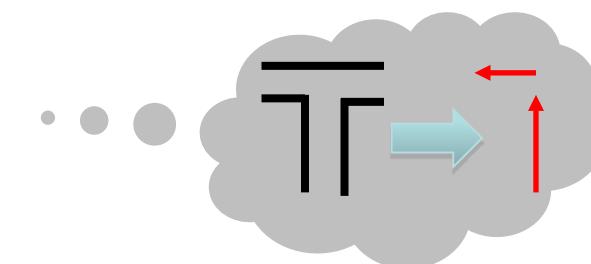
## model-free

- cannot immediately adapt
- like habits



Conflicting actions suggested by two system:  
how to arbitrate?

# Theory



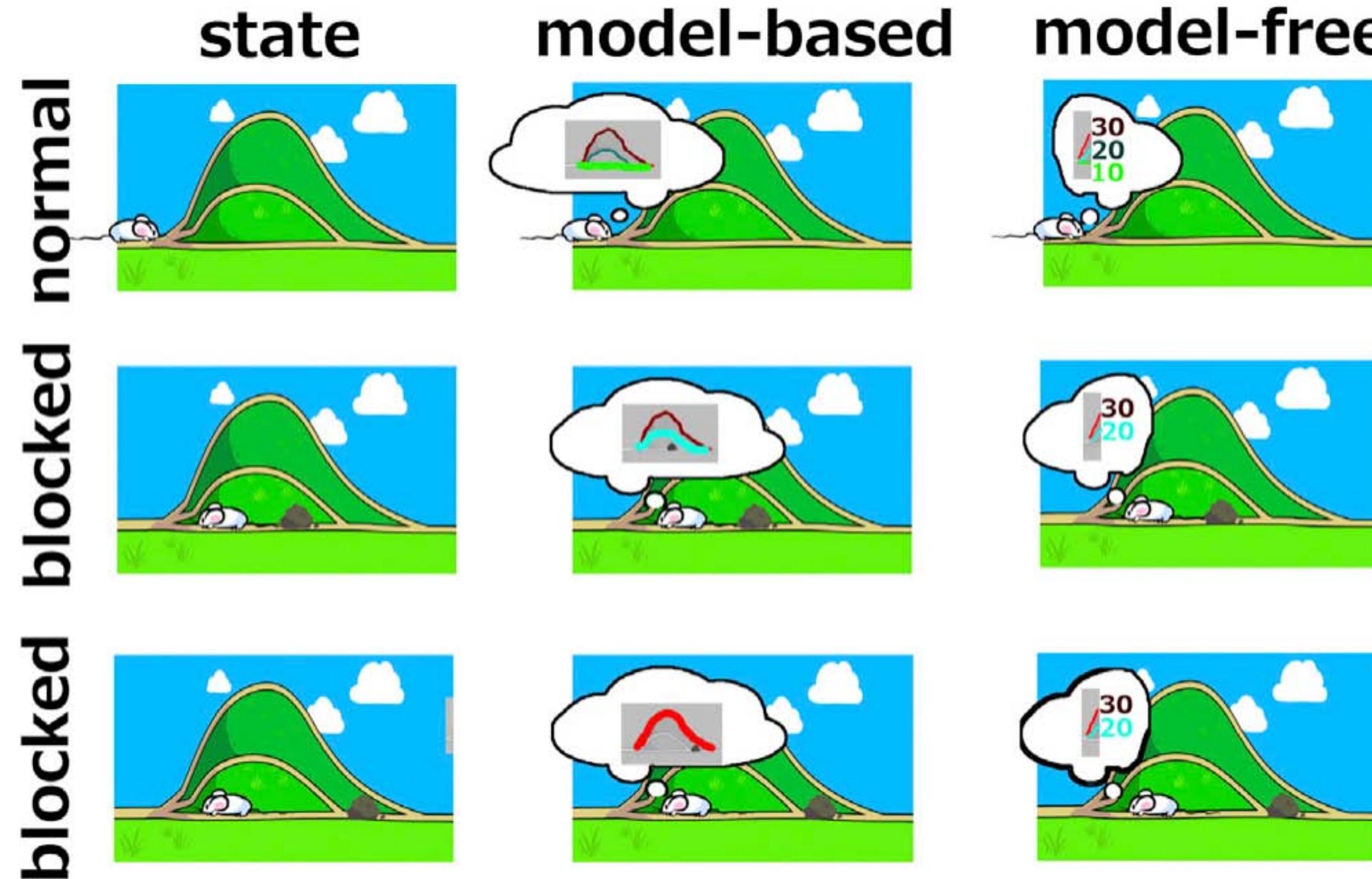
## Why have multiple systems?

- computational efficiency vs. statistical efficiency

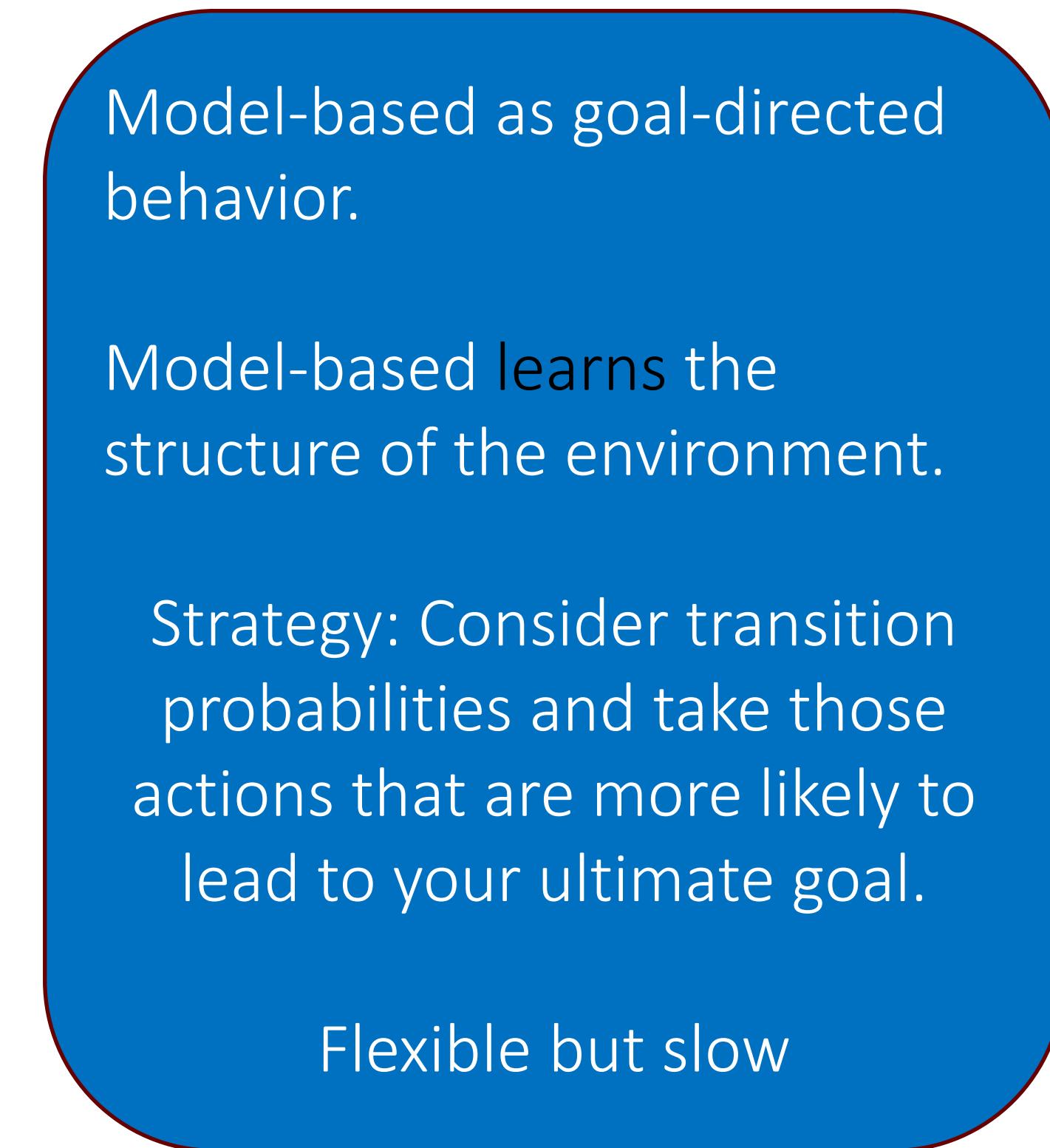
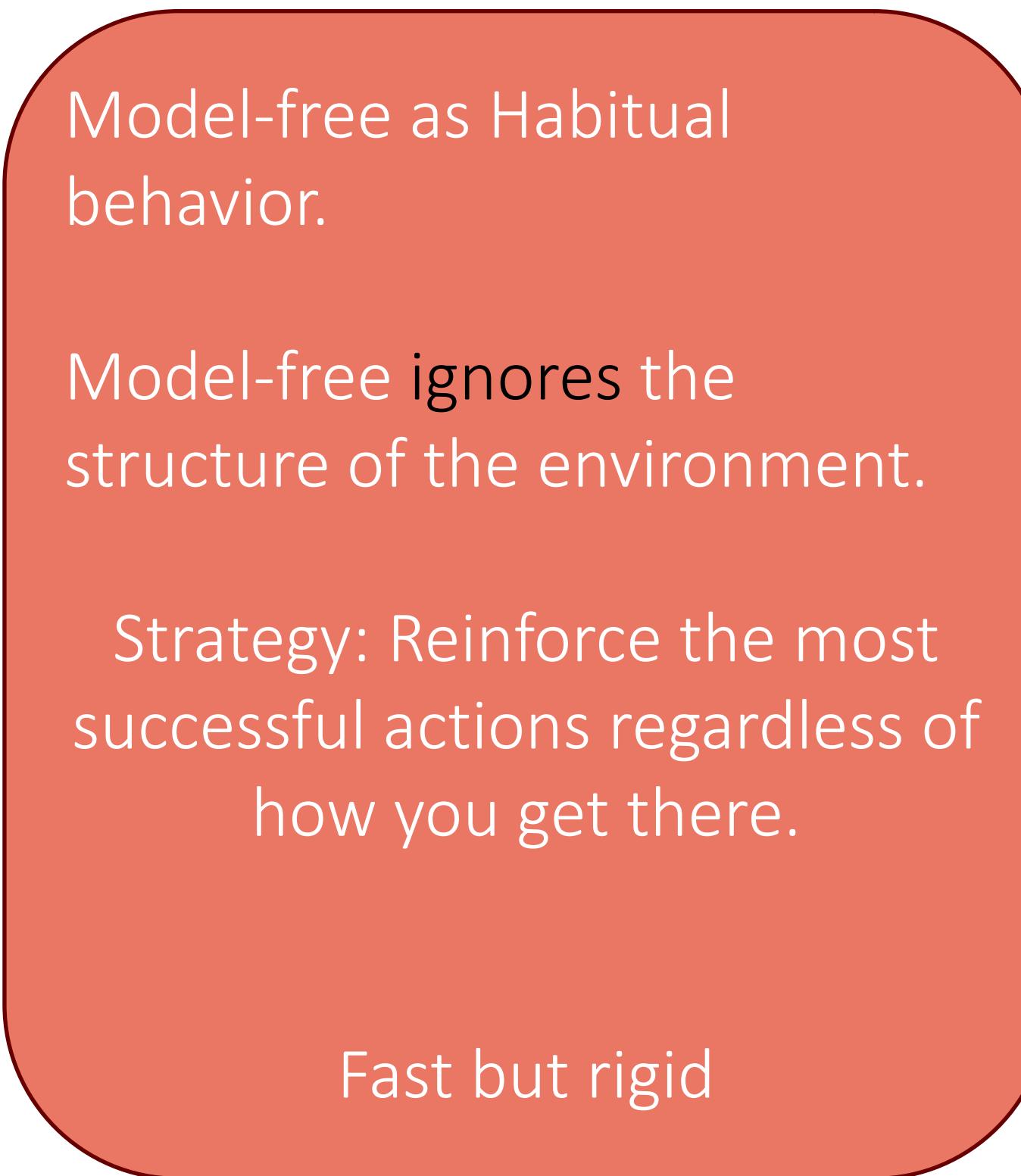
## When to favor each?

- bears on self-control, compulsion
- itself a decision-theoretic tradeoff (cf Keramati et al. 2011)
- e.g. not worth deliberating when highly practiced on stable task
- this model explains lots of data about what circumstances favor each system

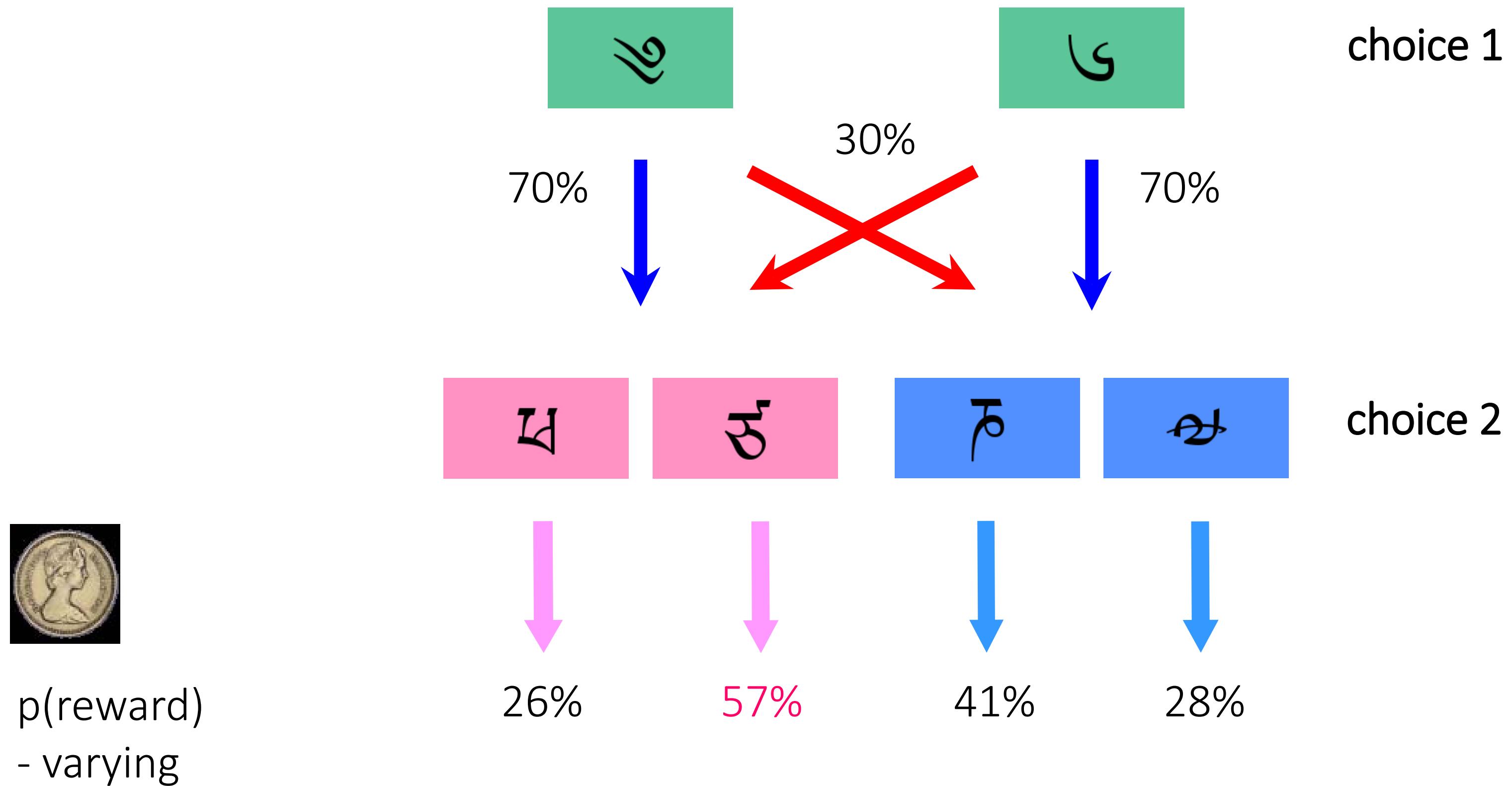
# Model-free versus Model-based



# Model-free versus Model-based



# Sequential decision task



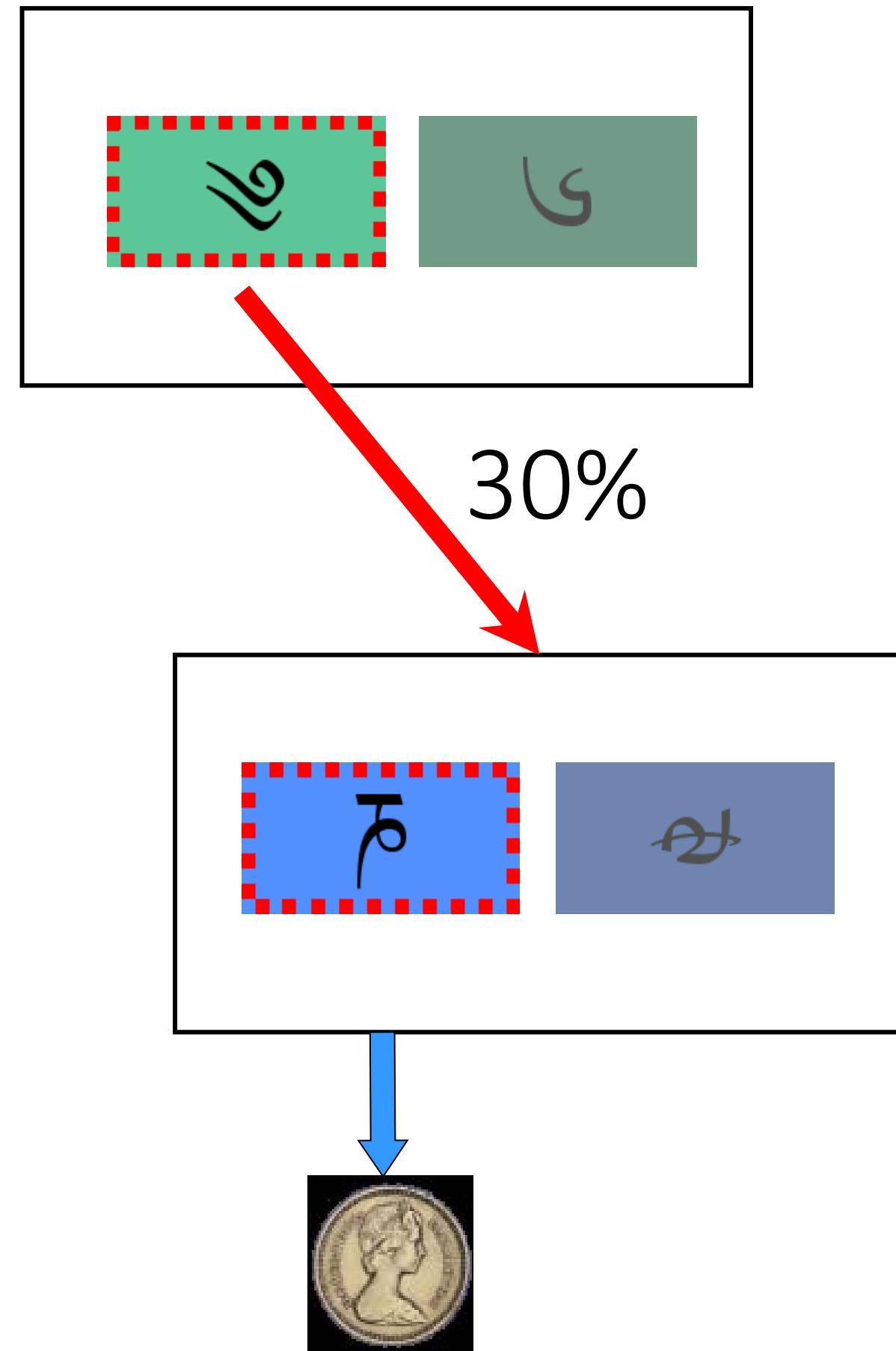


# idea

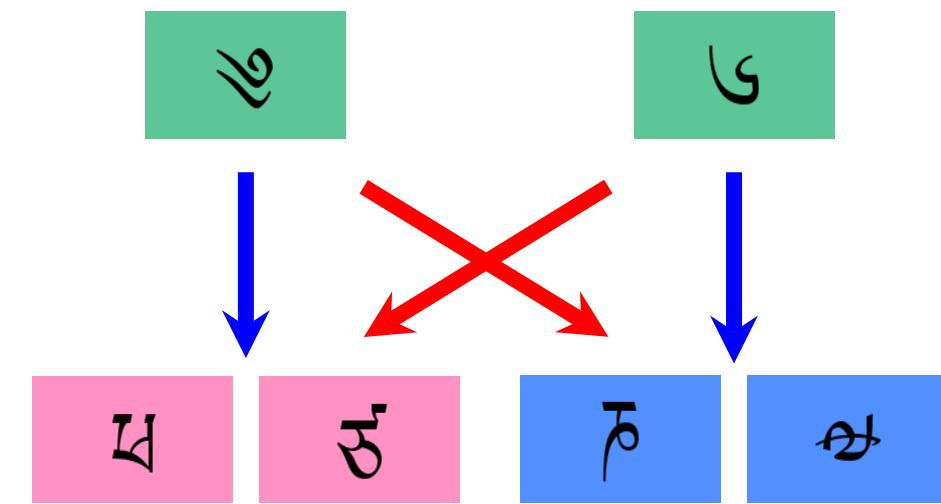
How does bottom-stage  
feedback affect top-stage  
choices?

Example: rare transition at top  
level, followed by win

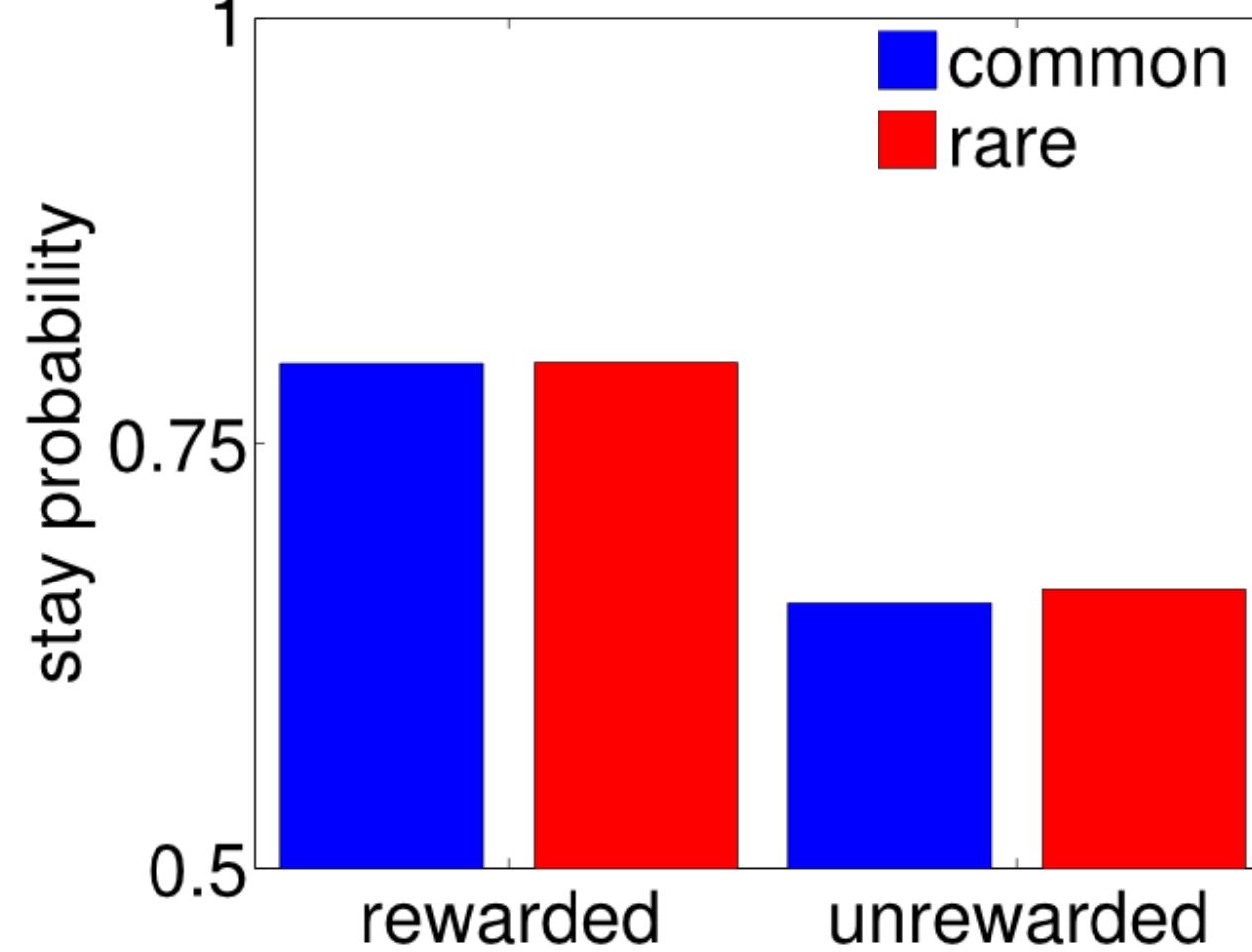
- Which top-stage action is  
now favored?



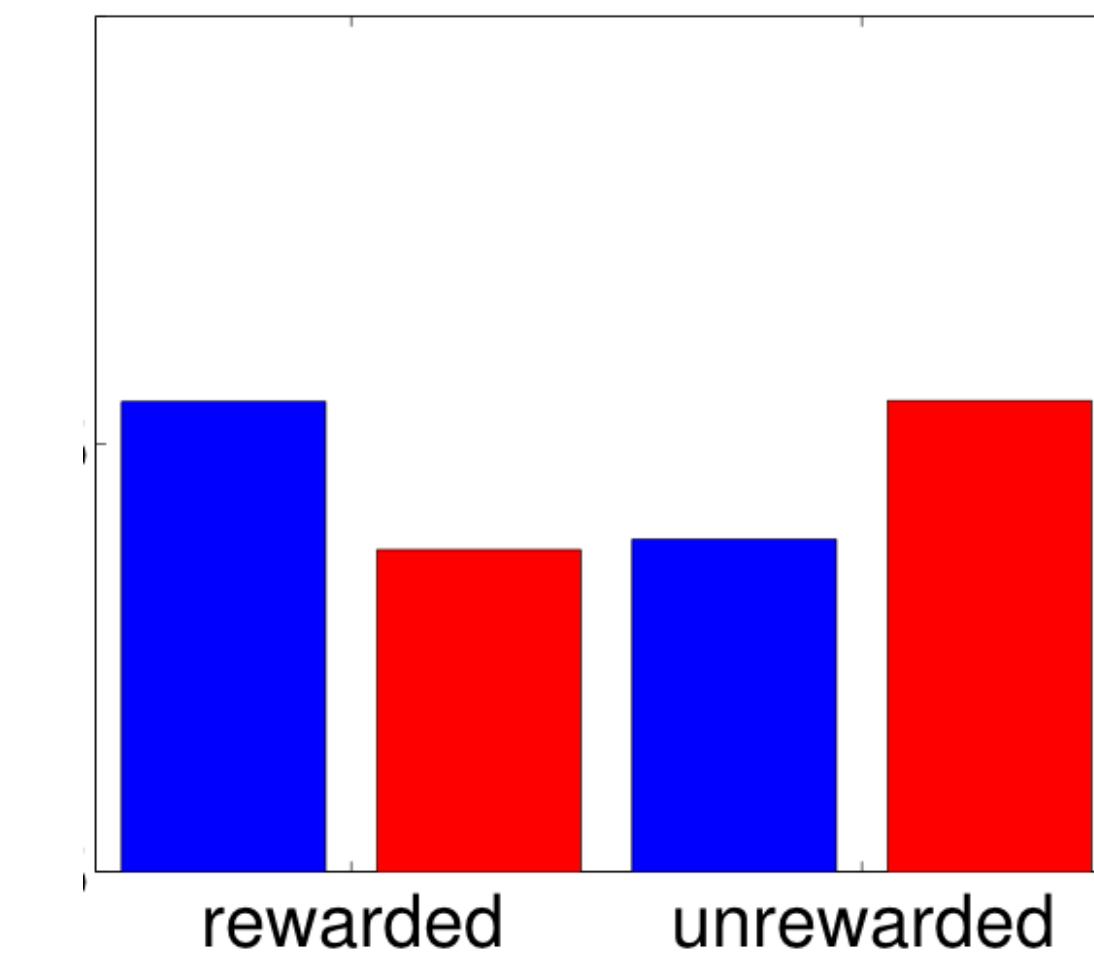
# predictions



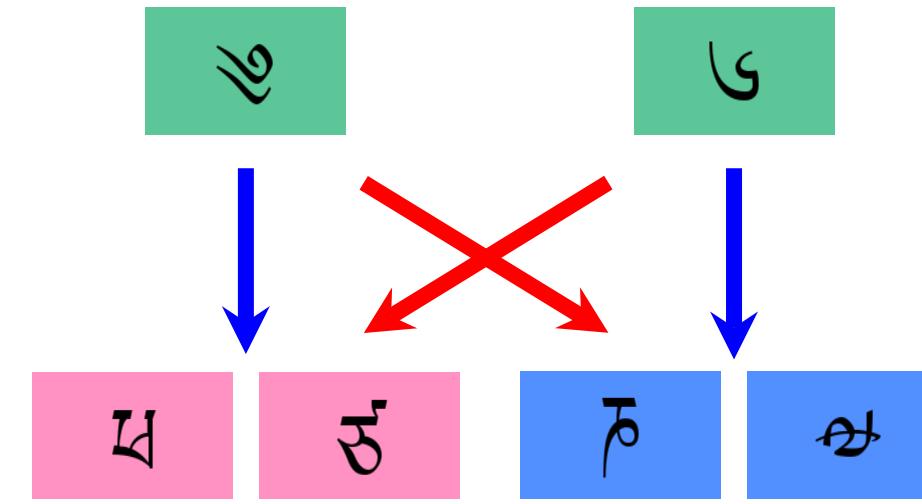
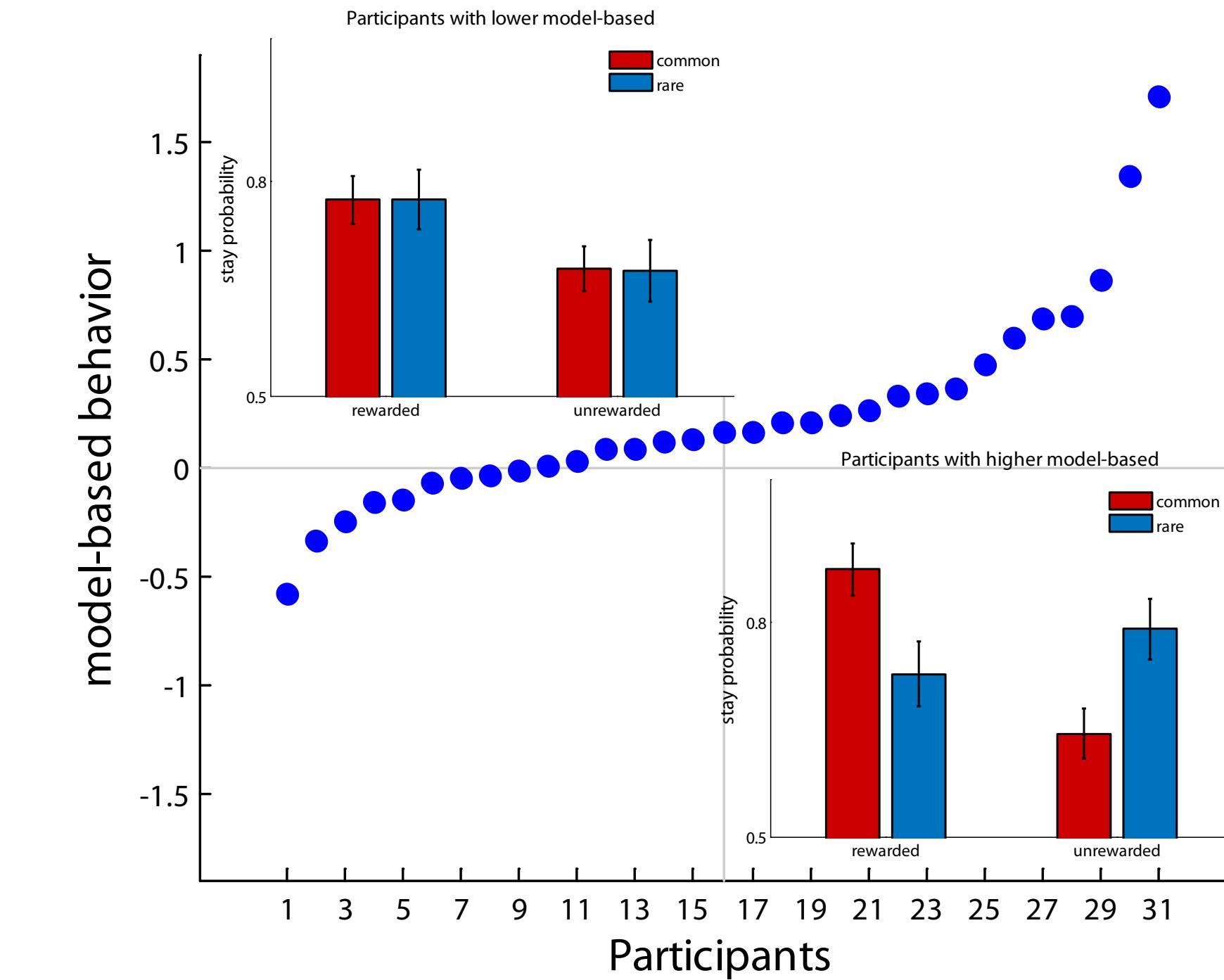
direct reinforcement  
ignores transition structure



model-based planning  
respects transition structure



# data

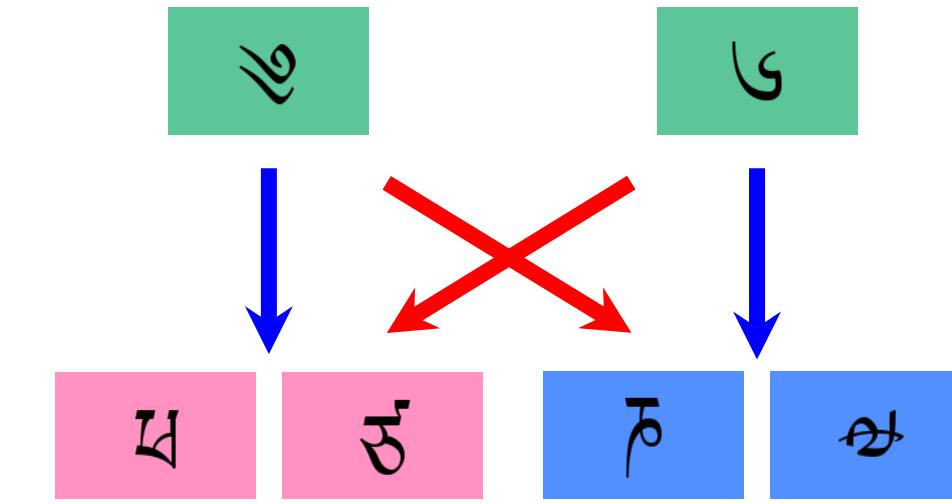
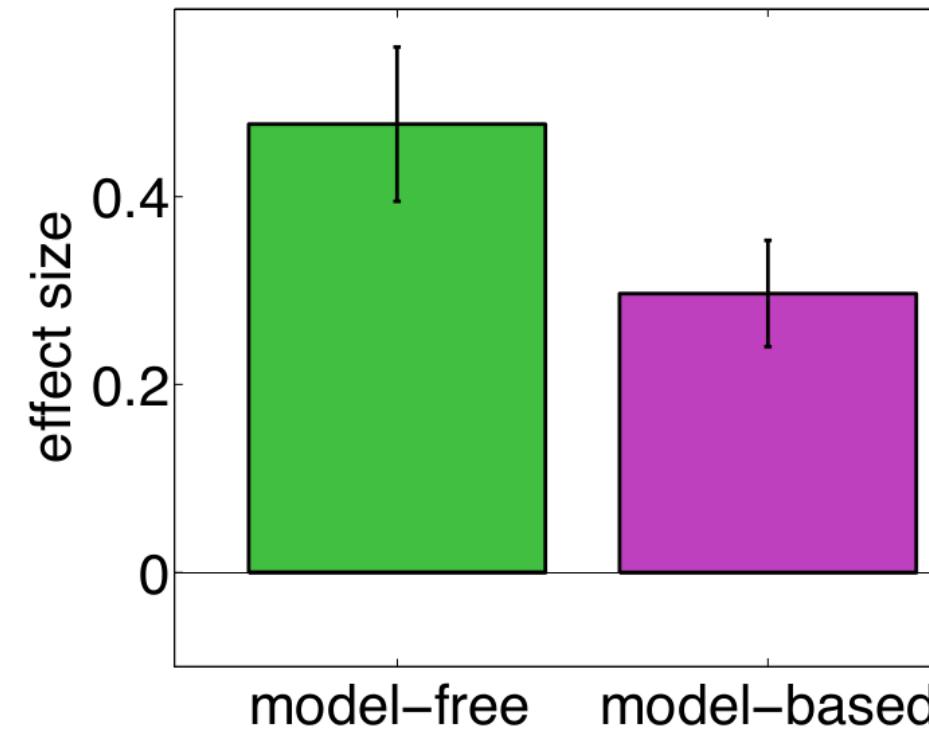
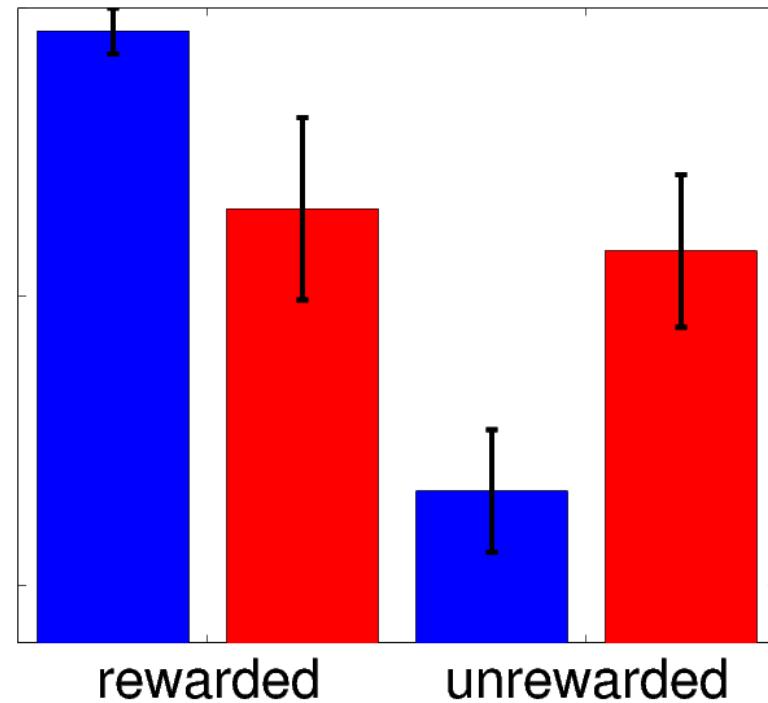


- results reject pure reinforcement models
- suggest **mixture** of planning and reinforcement processes

# data

reward:  $p < 1e-8$

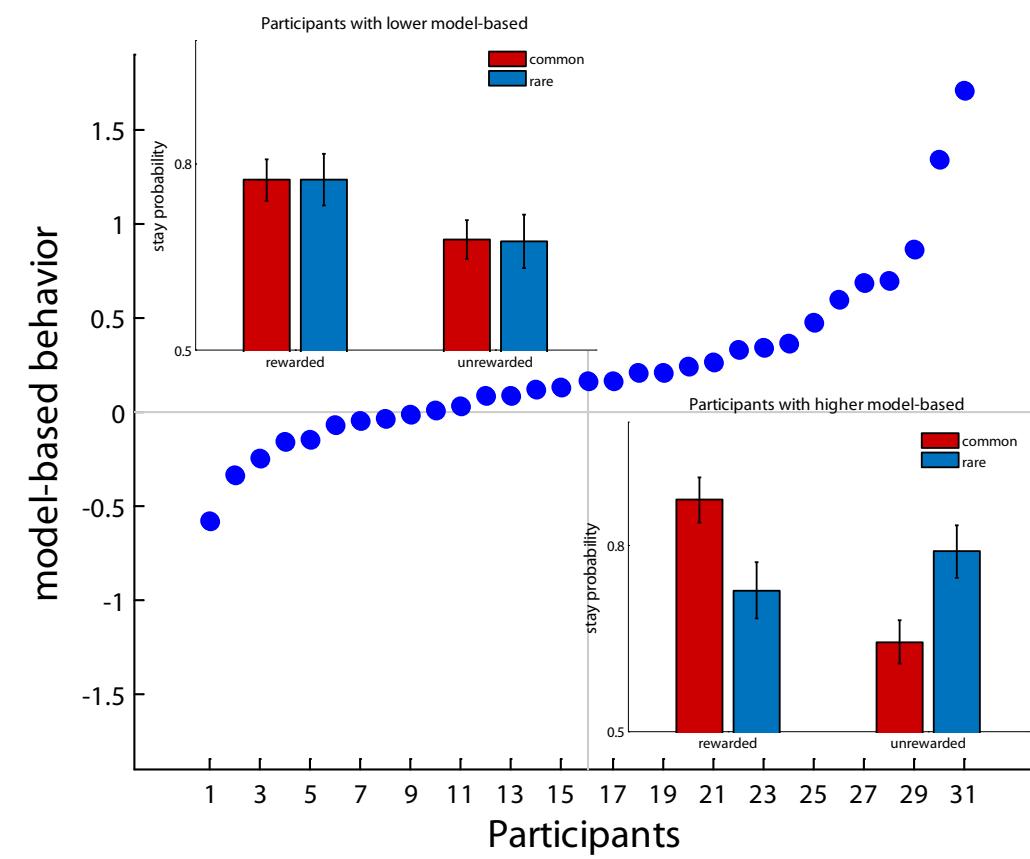
reward x rare:  $p < 5e-5$



- results reject pure reinforcement models
- suggest **mixture** of planning and reinforcement processes



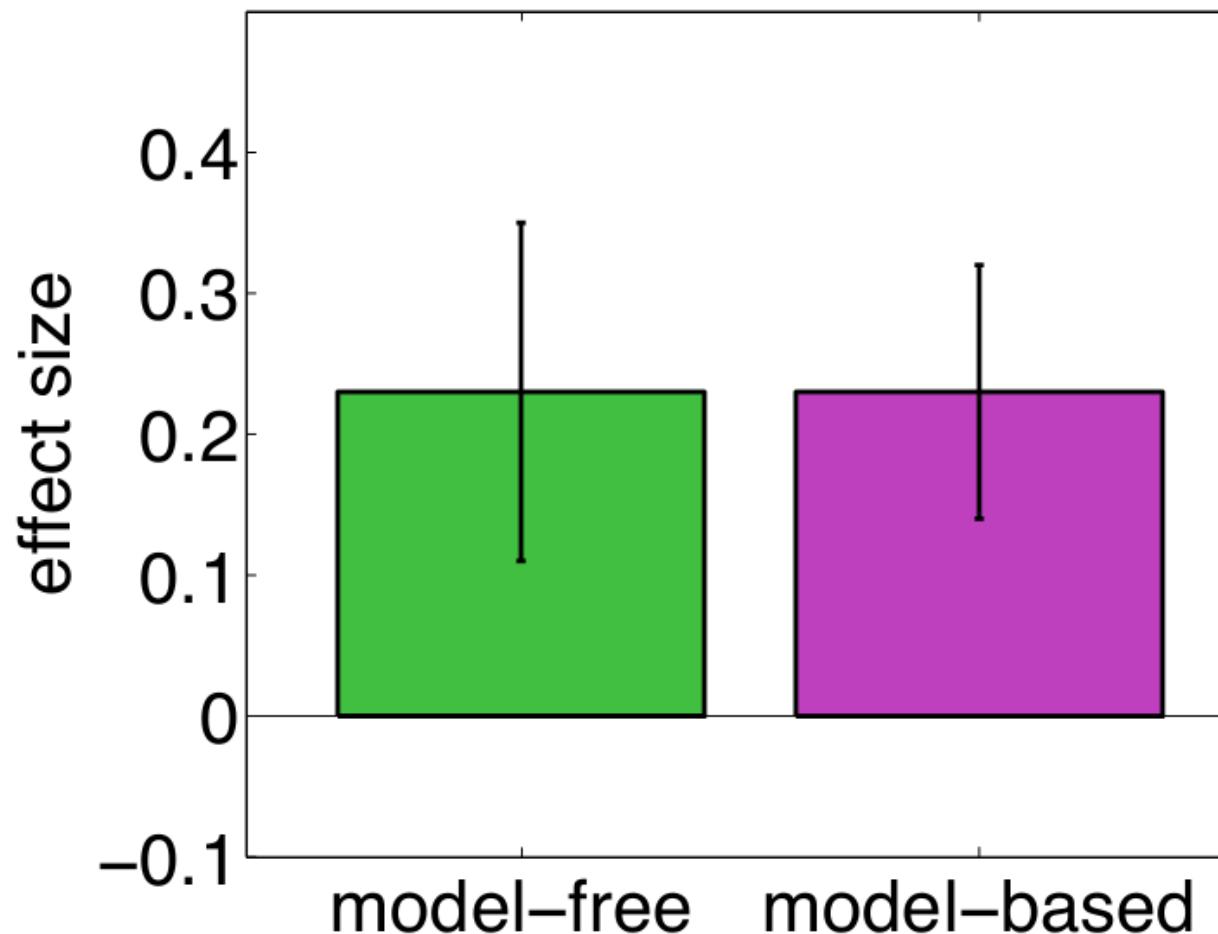
# Neural & cognitive basis of model-based versus model-free systems?



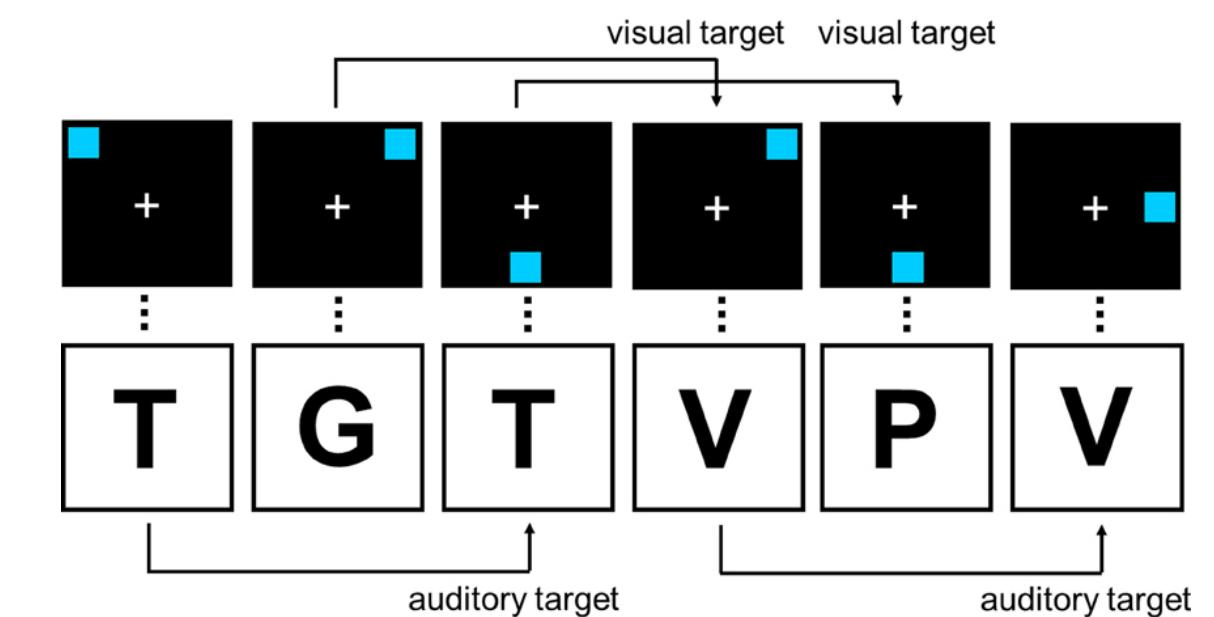
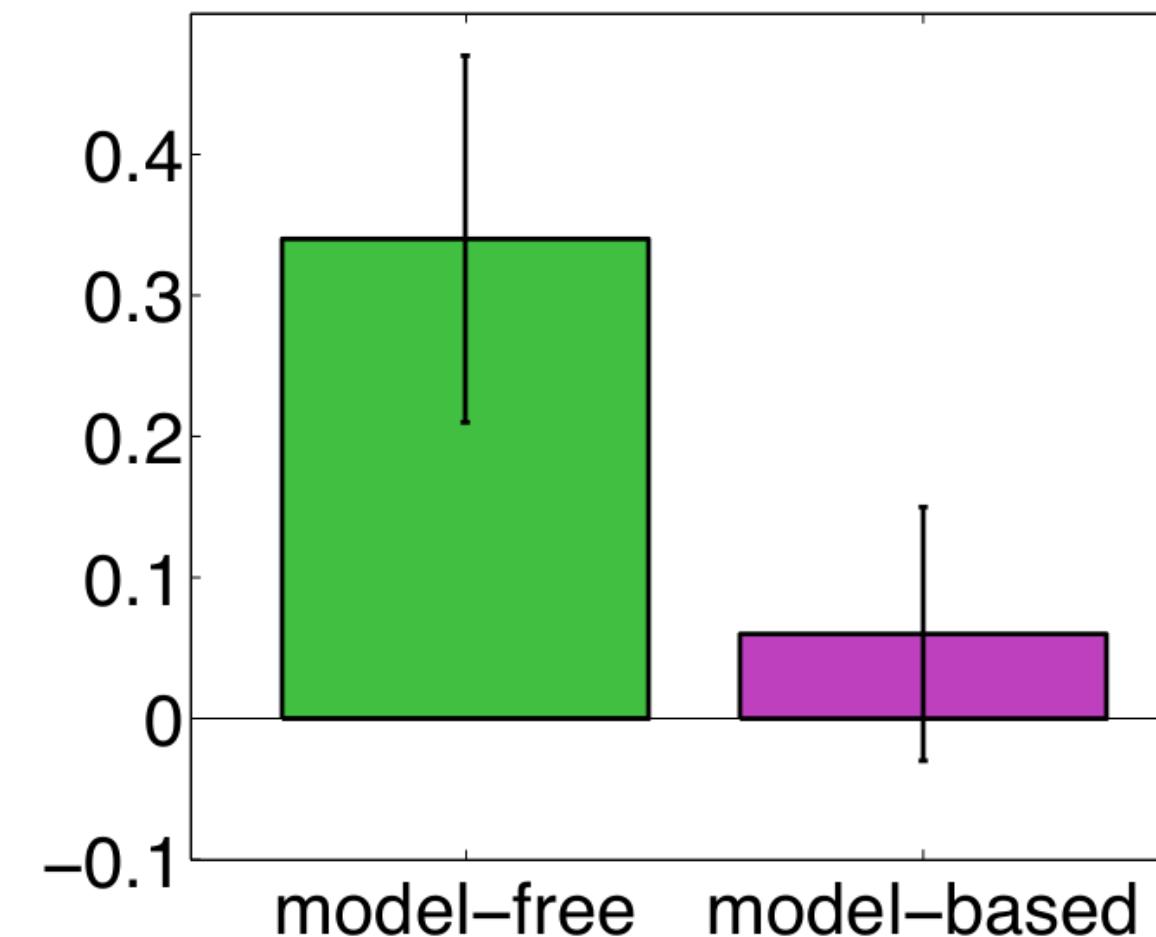
# central executive: capacity limits

- Loading the central executive: working memory task
- Prefrontal cortical basis of model-based control?

single task



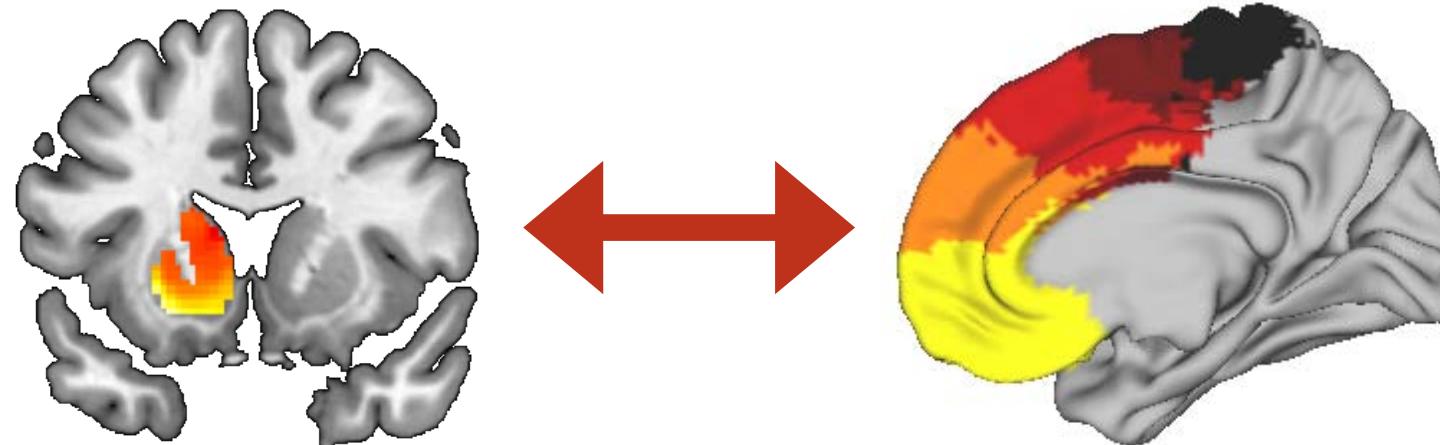
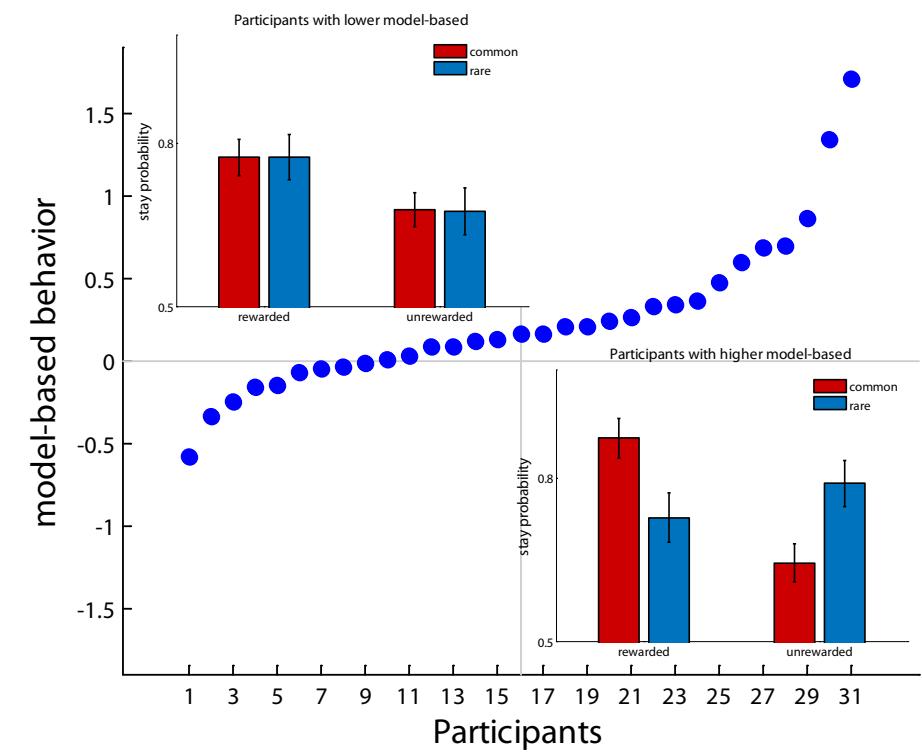
dual task



# Balancing model-based and model-free

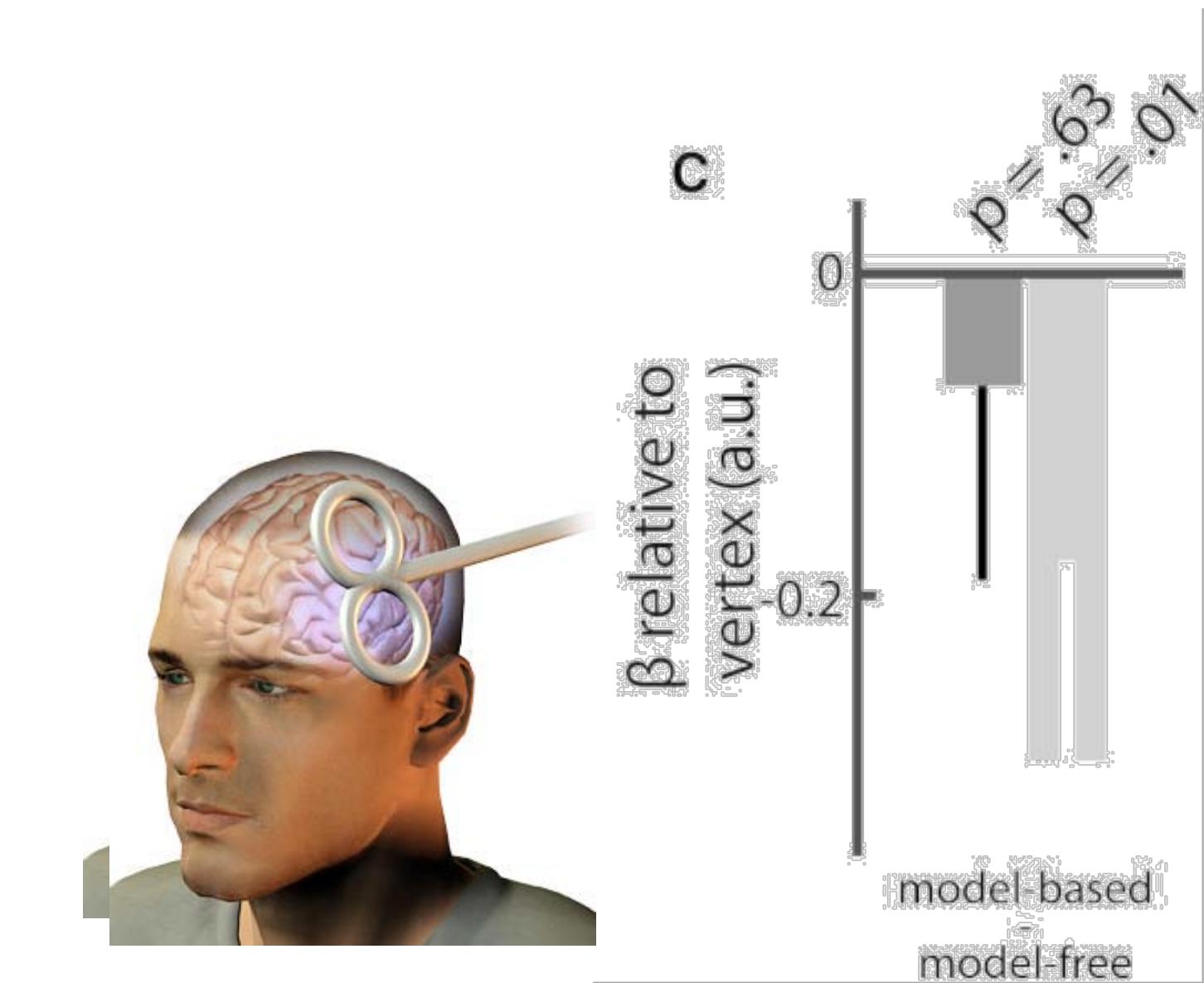
- Neural basis of individual differences: fronto-striatal connectivity (DTI)

*Piray ea. Cerebral Cortex 2016*



- Disrupting right DLPFC disrupts model-based relative to model-free control

*Smittenaar et al. Neuron 2013*



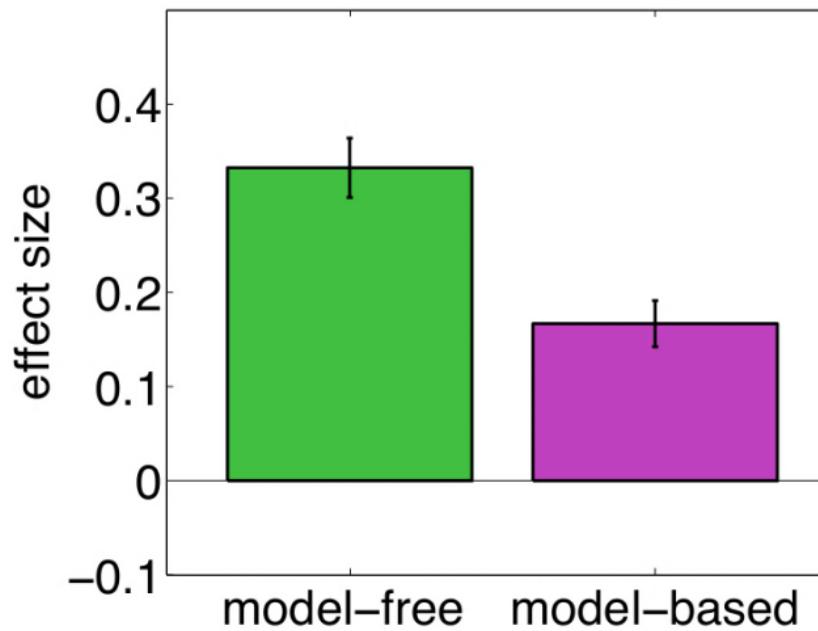


# Compulsive disorders

- “Loss of control over repetitive self-deleterious behavior seen in a range of disorders, most notably obsessive-compulsive disorder (OCD) and addiction”
- Suggested as ‘extreme habits’
  - habits tend to bring efficiency to one's life, while compulsions tend to disrupt it
- compulsion may be partially explained by an imbalance between flexible, goal-directed control and habits

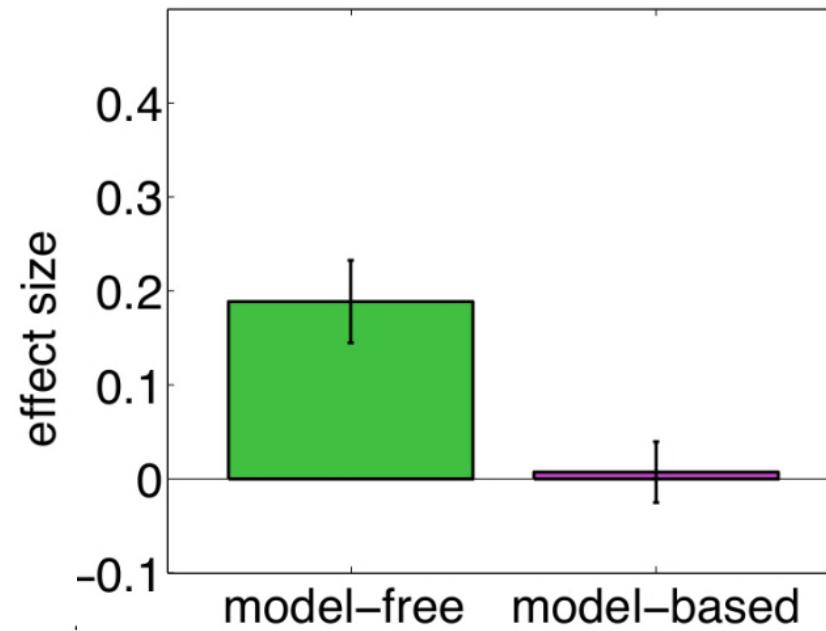


# Compulsions: pathological balance?

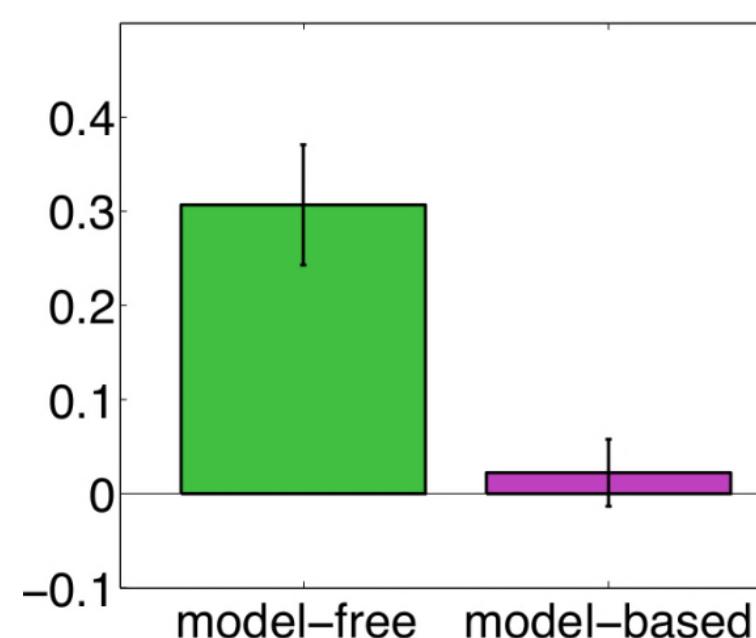


Healthy volunteers  
n=106

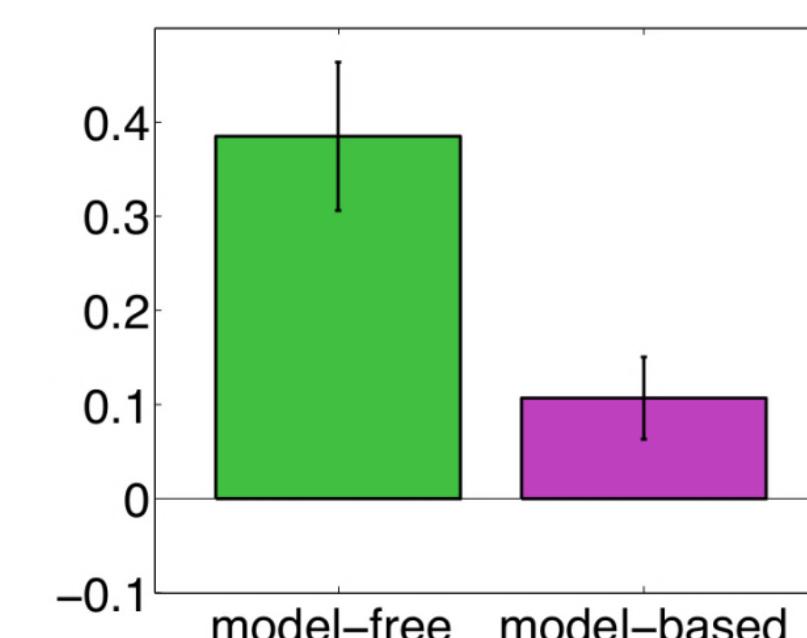
Stimulant abuse (n=36)



Binge eating (n=30)



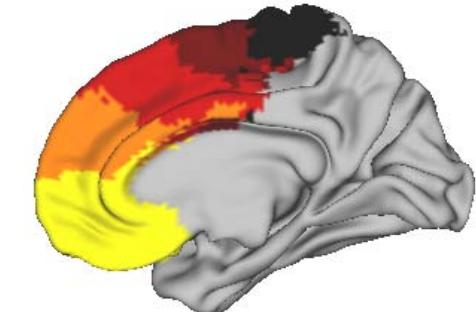
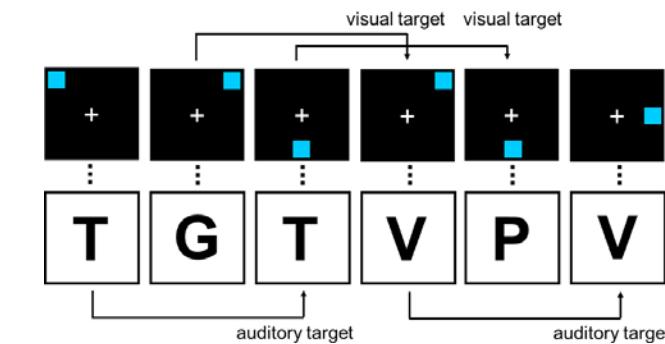
OCD (n=35)



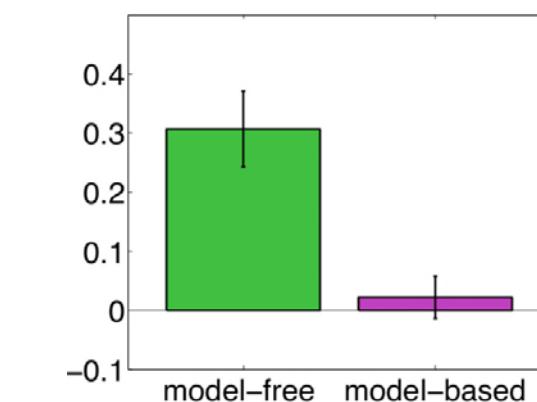


# Interim summary

- Model-based reasoning based on frontal cortex
  - 'loading' frontal processes reduces ability for model-based reasoning
  - Fronto-striatal coupling determines degree of 'model-basedness'
  - Disrupting dIPFC makes more model-free



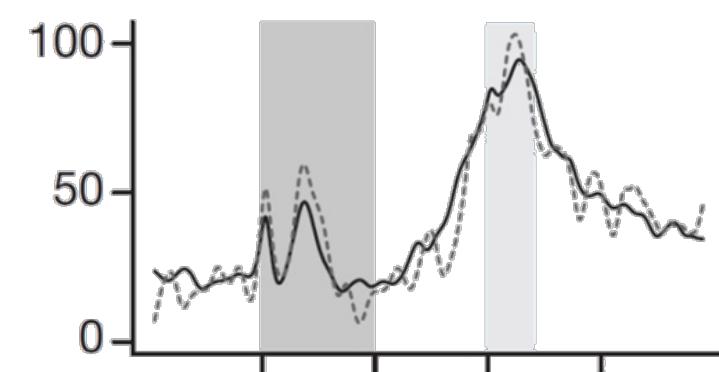
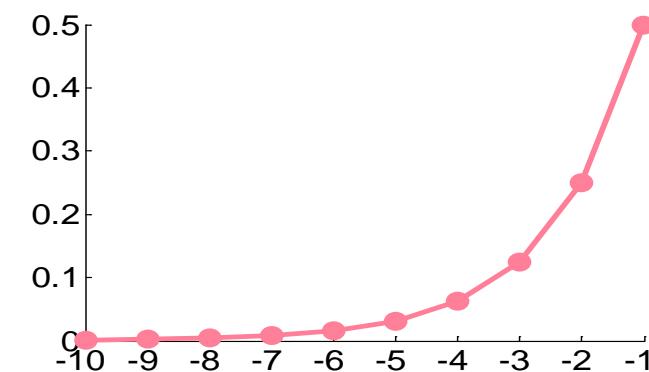
Relevance for psychiatric disorders where compulsive patients get stuck in being too model-free





# Where did we get to?

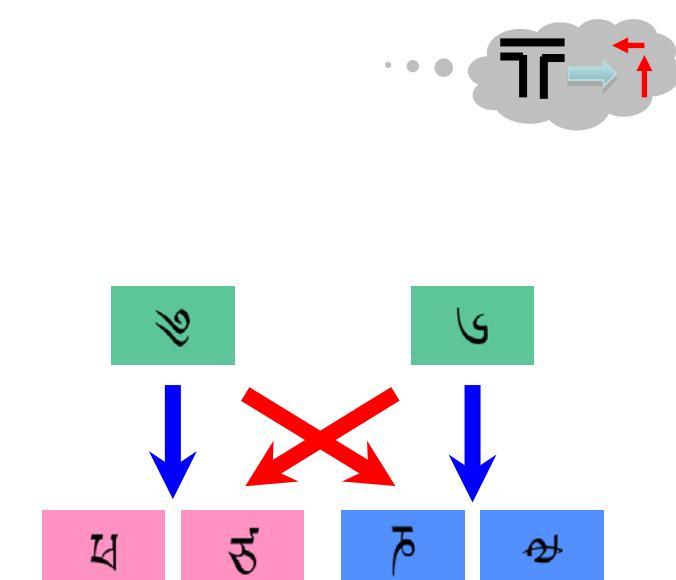
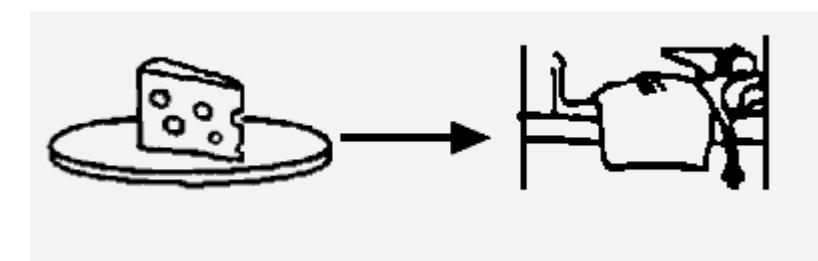
- Learning to predict reward is **error-driven**
- DA neurons code reward prediction errors
- Reinforcement learning can be modelled using a **TD** learning model
- GABAergic inhibition ‘sets the scene’, i.e. reflects predictions





# where did we get to - II

- but... error-based learning cannot explain all forms of learning, e.g. model-based learning
- Cognitive maps vs.  $S \rightarrow R$  learning
- Dopamine as  $S \rightarrow R$  like model, devaluation insensitive, “habitual”
- Shadowed by separate “goal-directed” system
  - in looking for behavioral effects of dopamine, important to consider this confound
  - Devaluation test
- Can also be understood computationally as model-free vs. model-based learning



# Reinforcement Learning

*Interested in studying how we make strategic decisions, using computational approaches?*

get in touch:



[hannekedenouden.ruhosting.nl](http://hannekedenouden.ruhosting.nl)



[h.denouden@donders.ru.nl](mailto:h.denouden@donders.ru.nl)



[@HannekedenOuden](https://twitter.com/HannekedenOuden)