The Neuroscience of Reinforcement Learning

Yael Niv
Psychology Department & Neuroscience Institute
Princeton University

yael@princeton.edu

ICML'09 Tutorial
Montreal

Goals

- Reinforcement learning has revolutionized our understanding of learning in the brain in the last 20 years
- Not many ML researchers know this!
  1. Take pride
  2. Ask: what can neuroscience do for me?
- Why are you here?
  - To learn about learning in animals and humans
  - To find out the latest about how the brain does RL
  - To find out how understanding learning in the brain can help RL research

If you are here for other reasons...

learn what is RL and how to do it

learn about the brain in general

read email

take a well-needed nap

smirk at the shoddy state of neuroscience

Outline

- The brain coarse-grain
- Learning and decision making in animals and humans: what does RL have to do with it?
- A success story: Dopamine and prediction errors
- Actor/Critic architecture in basal ganglia
- SARSA vs Q-learning: can the brain teach us about ML?
- Model free and model based RL in the brain
- Average reward RL & tonic dopamine
- Risk sensitivity and RL in the brain
- Open challenges and future directions

Why do we have a brain?

- because computers were not yet invented
- to behave
- example: sea squirt

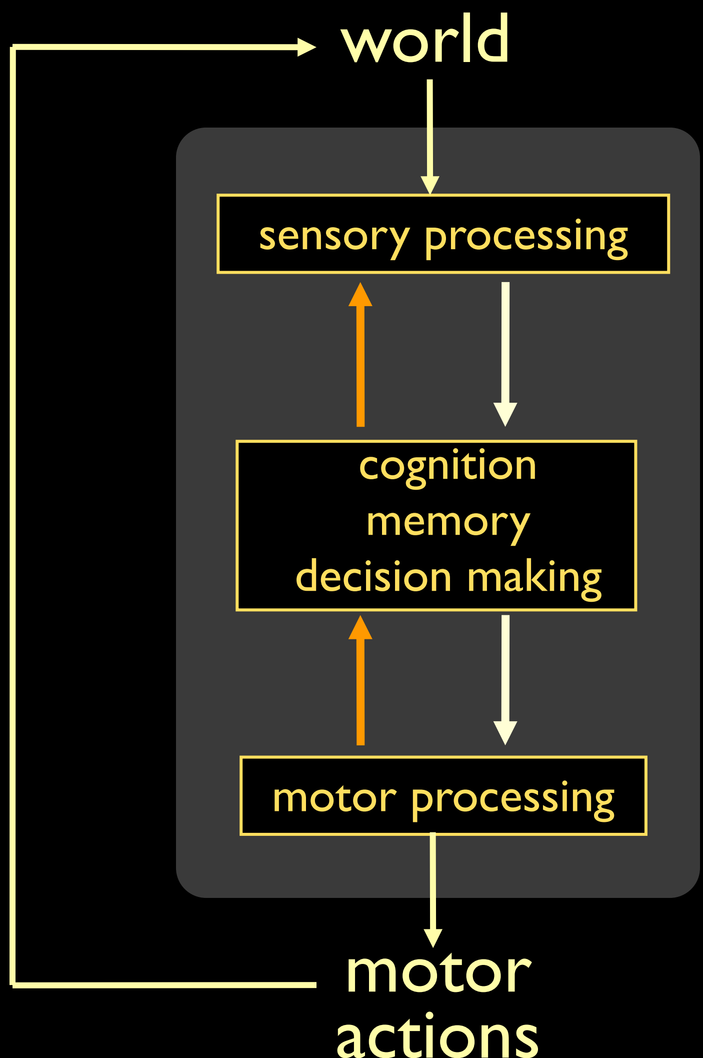
- larval stage: primitive brain & eye, swims around, attaches to a rock
- adult stage: sits. digests brain.

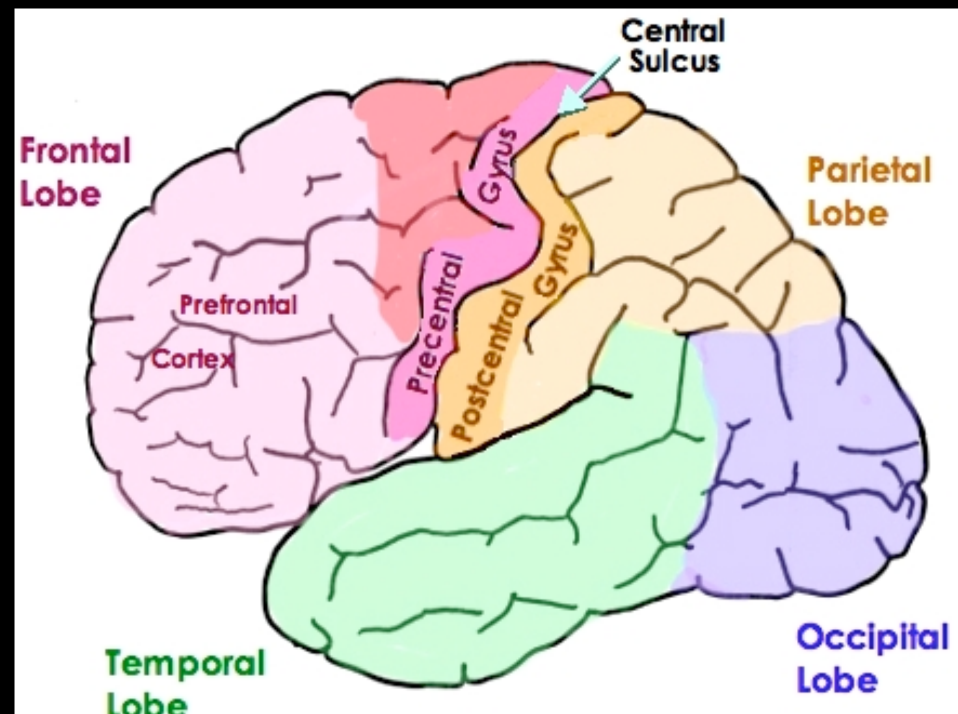
Credits: Daniel Wolpert

Why do we have a brain?

Credits: Daniel Wolpert

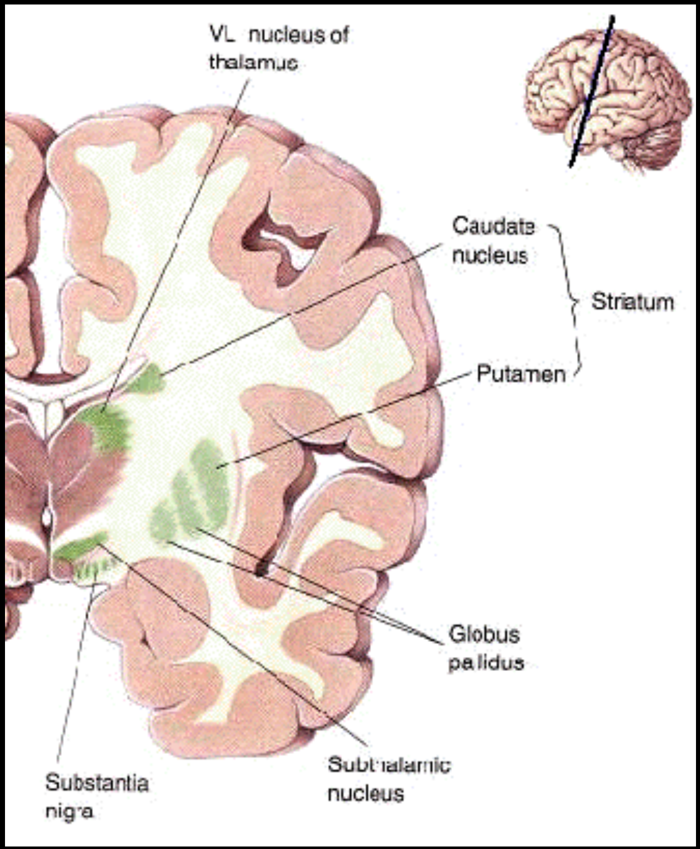
the brain in very coarse grain





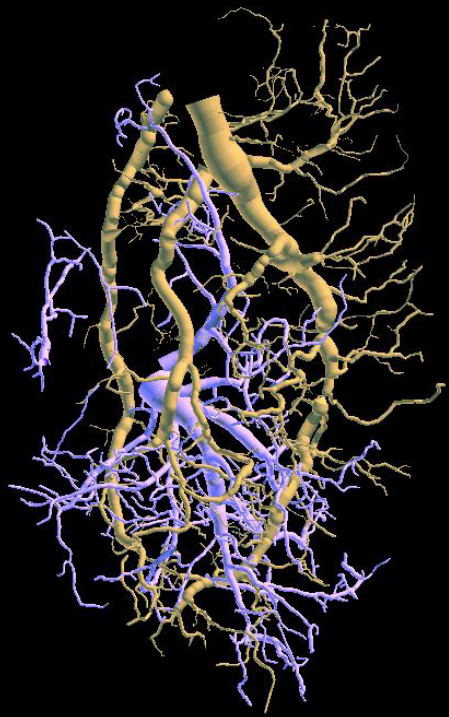
what do we know about the brain?

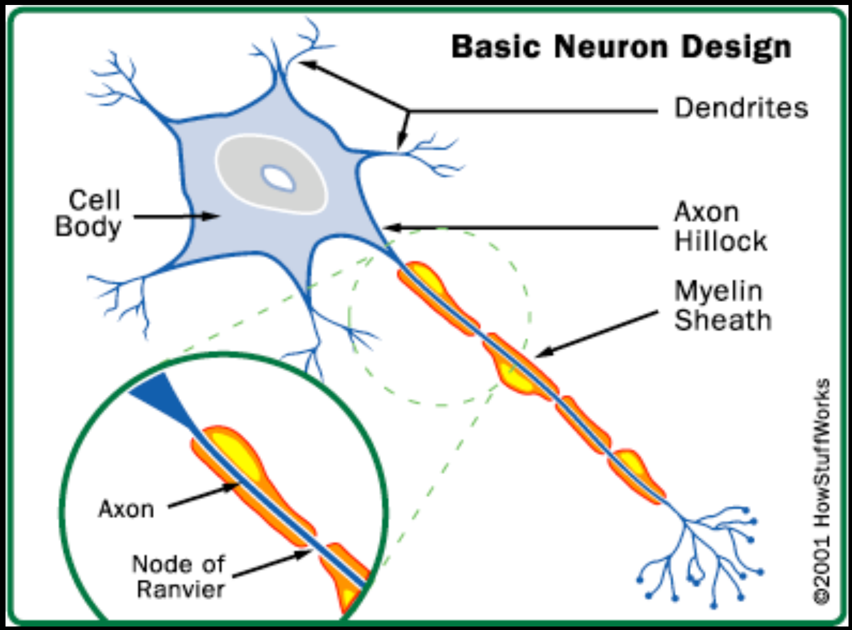
- Anatomy: we know a lot about what is where and (more or less) which area is connected to which (But unfortunately names follow structure and not function; be careful of generalizations, e.g. neurons in motor cortex can respond to color)



what do we know about the brain?

- Anatomy: we know a lot about what is where and (more or less) which area is connected to which (But unfortunately names follow structure and not function; be careful of generalizations, e.g. neurons in motor cortex can respond to color)
- Single neurons: we know quite a bit about how they work (but still don't know much about how their 3D structure affects function)





what do we know about the brain?

- Anatomy: we know a lot about what is where and (more or less) which area is connected to which (But unfortunately names follow structure and not function; be careful of generalizations, e.g. neurons in motor cortex can respond to color)
- Single neurons: we know quite a bit about how they work (but still don't know much about how their 3D structure affects function)
- Networks of neurons: we have some ideas but in general are still in the dark
- Learning: we know a lot of facts (LTP, LTD, STDP) (not clear which, if any are relevant; relationship between synaptic learning rules and computation essentially unknown)
- Function: we have pretty coarse grain knowledge of what different brain areas do (mainly sensory and motor; unclear about higher cognitive areas; much emphasis on representation rather than computation)

Summary so far...

- We have a lot of facts about the brain
- But.. we still don't understand that much about how it works
- (can ML help??)

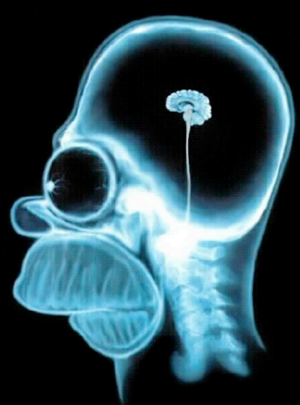
Outline

- The brain coarse-grain
- Learning and decision making in animals and humans: what does RL have to do with it?
- A success story: Dopamine and prediction errors
- Actor/Critic architecture in basal ganglia
- SARSA vs Q-learning: can the brain teach us about ML?
- Model free and model based RL in the brain
- Average reward RL & tonic dopamine
- Risk sensitivity and RL in the brain
- Open challenges and future directions

what do neuroscientists do all day?

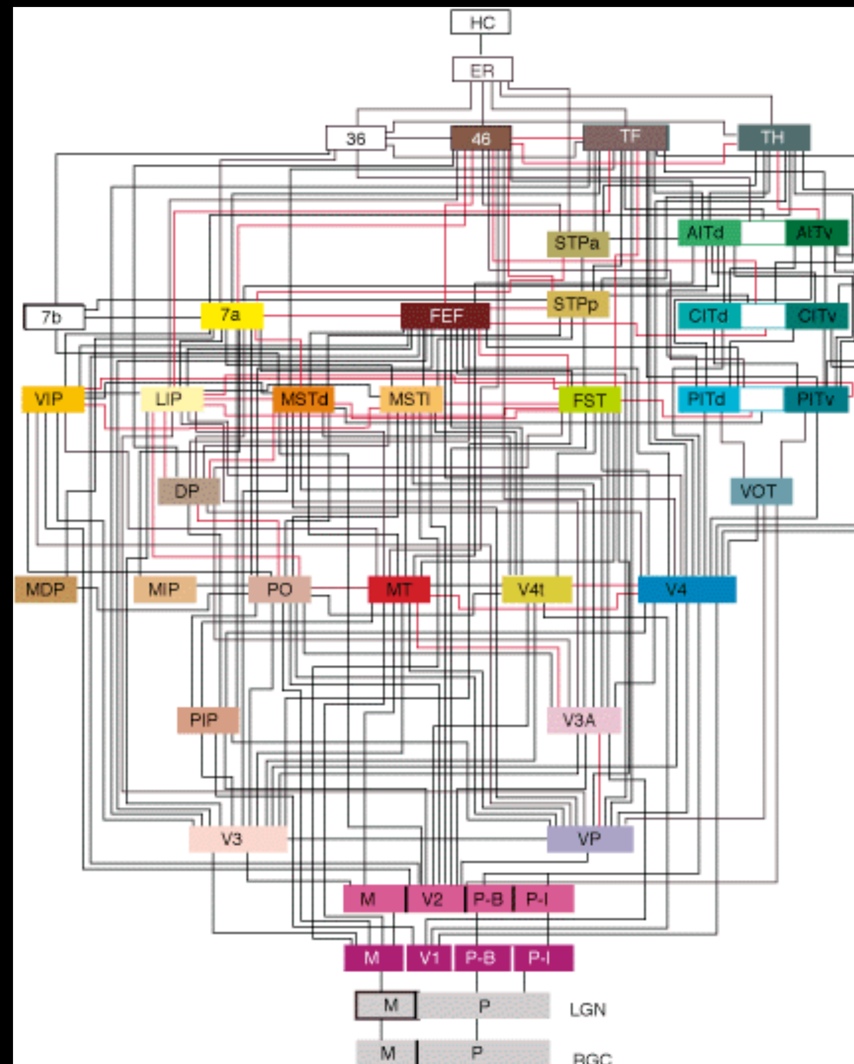
figure out how the brain generates behavior

brain ⇄ behavior



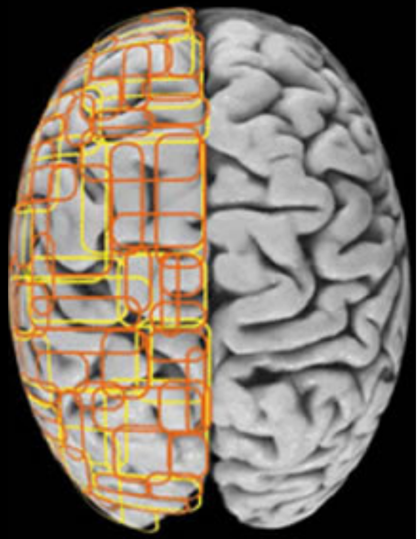
do we need so many neuroscientists for one simple question?

- Old idea:
  structure → function
- The brain is an extremely complex (and messy) dynamic biological system
- 10^{11} neurons communicating through 10^{14} synapses
- we don't stand a chance...



Credits: Peter Latham

in comes computational neuroscience



- (relatively) New Idea:
- The brain is a computing device
- Computational models can help us talk about functions of the brain in a precise way
- Abstract and formal theory can help us organize and interpret (concrete) data

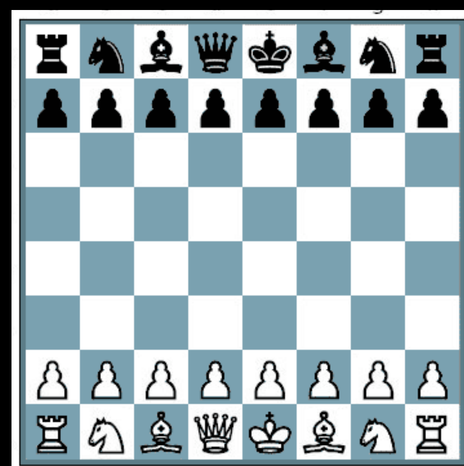
a framework for computational neuroscience

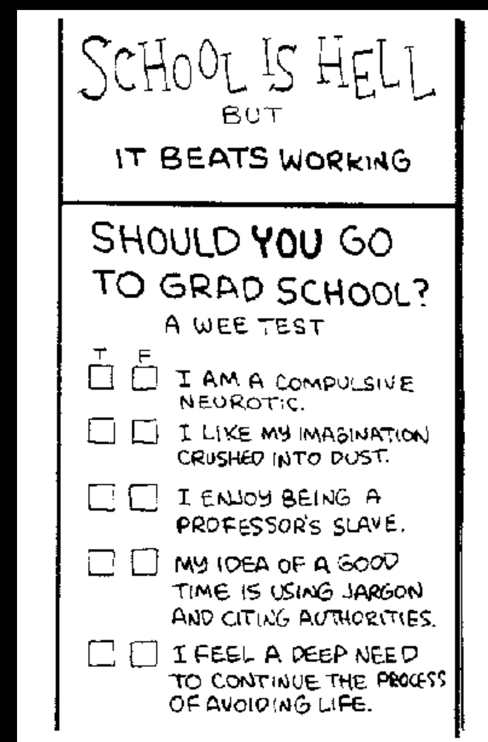
David Marr (1945-1980) proposed three levels of analysis:

1. the problem (Computational Level)
2. the strategy (Algorithmic Level)
3. how its actually done by networks of neurons (Implementational Level)

the problem we all face in our daily life

optimal decision making
(maximize reward, minimize punishment)





Why is this hard?

- Reward/punishment may be delayed
- Outcomes may depend on a series of actions

⇒ "credit assignment problem" (Sutton, 1978)

in comes reinforcement learning

- The problem: optimal decision making (maximize reward, minimize punishment)
- An algorithm: reinforcement learning
- Neural implementation: basal ganglia, dopamine

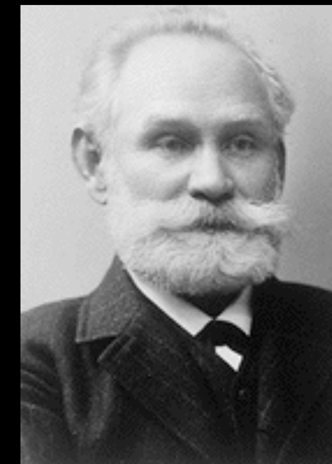
Summary so far...

- Idea: study the brain as a computing device
- Rather than look at what networks of neurons in the brain represent, look at what they compute
- What do animal's brains compute?

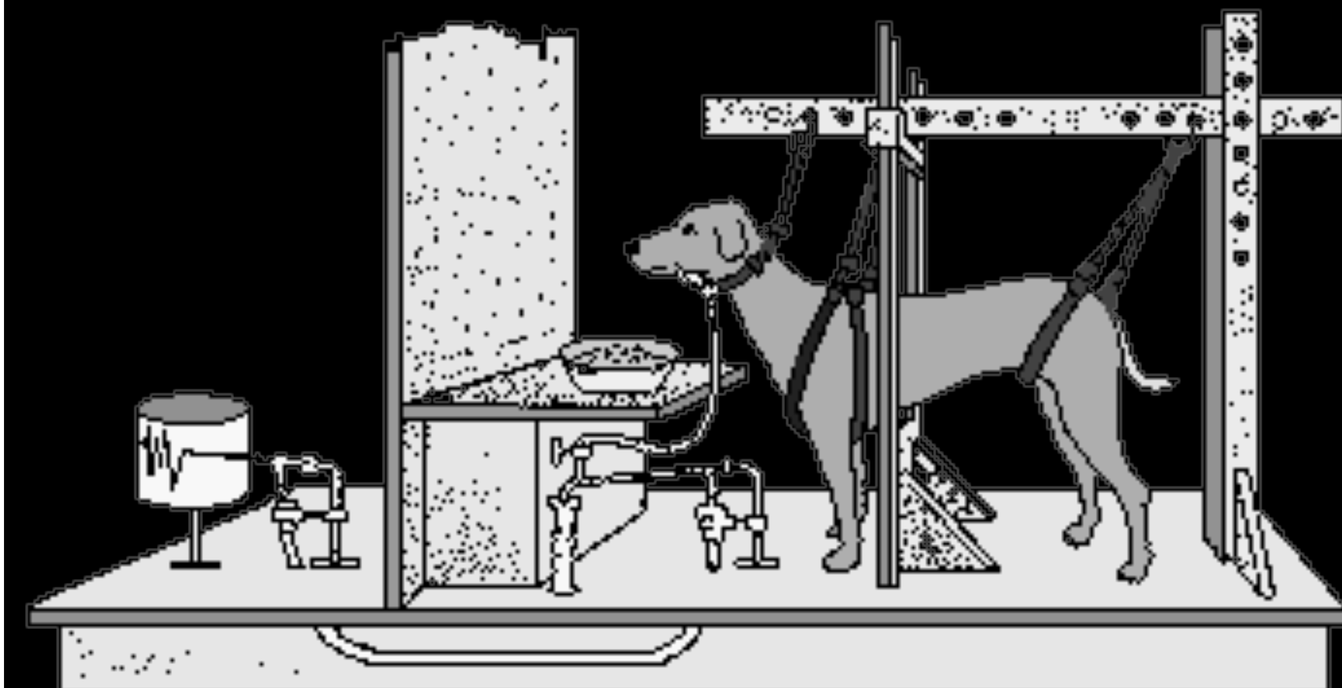
Animal Conditioning and RL

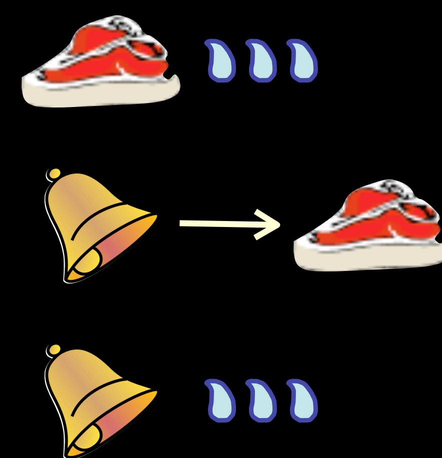
- two basic types of animal conditioning (animal learning)
- how do these relate to RL?

1. Pavlovian conditioning: animals learn predictions



Ivan Pavlov
(Nobel prize portrait)

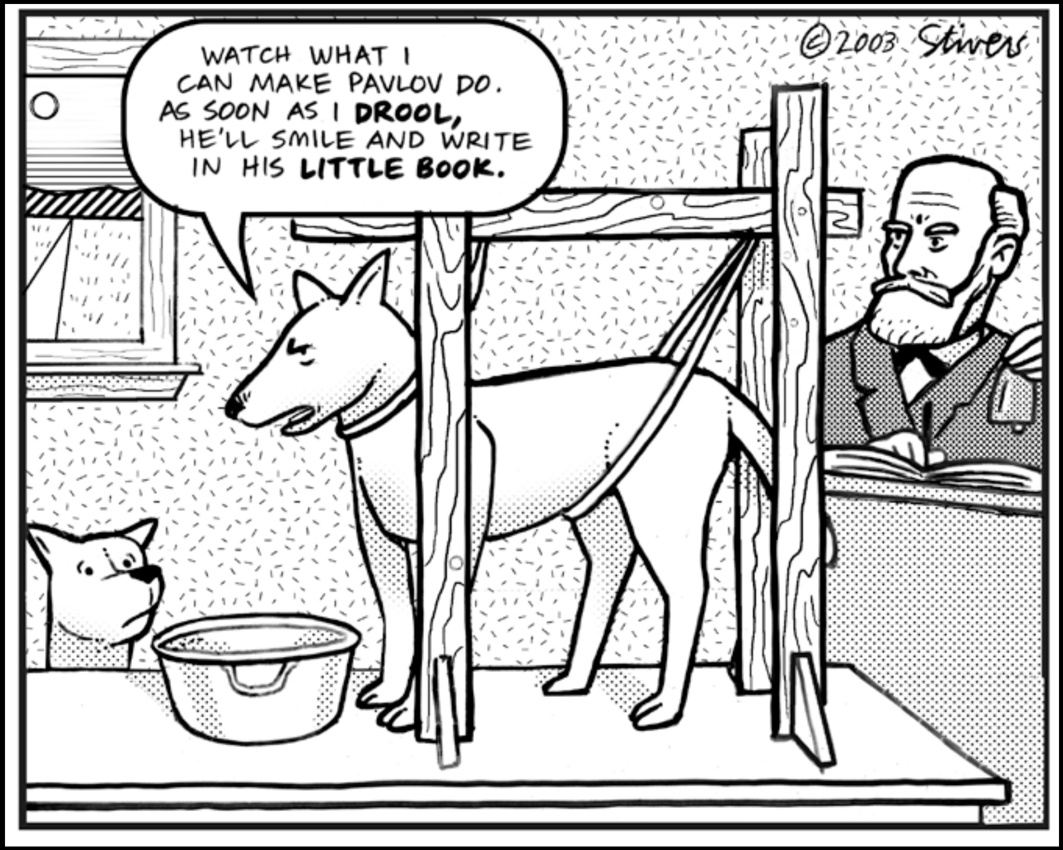




Pavlovian conditioning examples
(conditioned suppression, autoshaping)

Credits: Greg Quirk

how is this related to RL?



model-free learning of values of stimuli through experience; responding conditioned on (predictive) value of stimulus

Dayan et al. (2006) - “The Misbehavior of Value and the Discipline of the Will”

Rescorla & Wagner (1972)

The idea: error-driven learning

Change in value is proportional to the difference between actual and predicted outcome

$$\Delta V(S_i) = \eta[R - \sum_{j \in \text{trial}} V(S_j)]$$

Two assumptions/hypotheses:

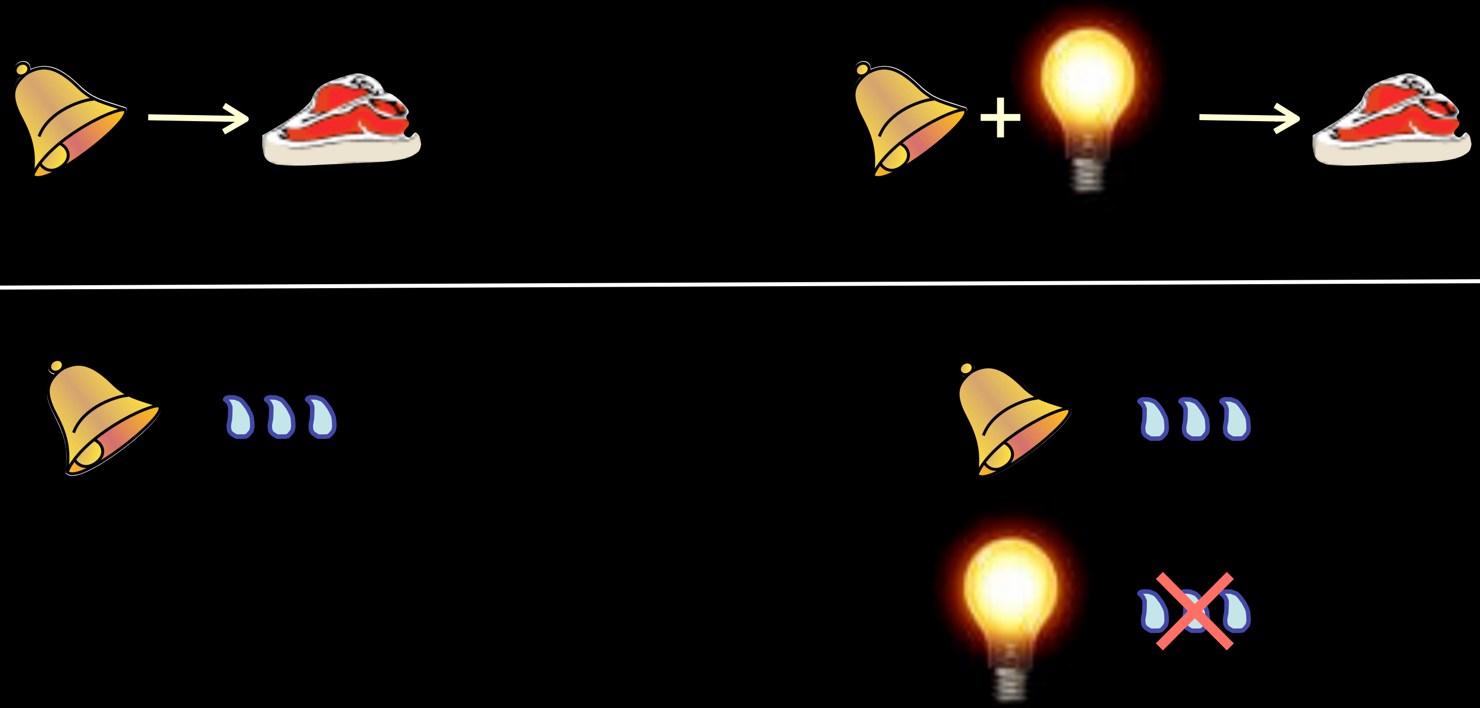
(1) learning is driven by error (formalize notion of surprise)

(2) summations of predictors is linear

How do we know that animals use an error-correcting learning rule?

Phase I

Phase II



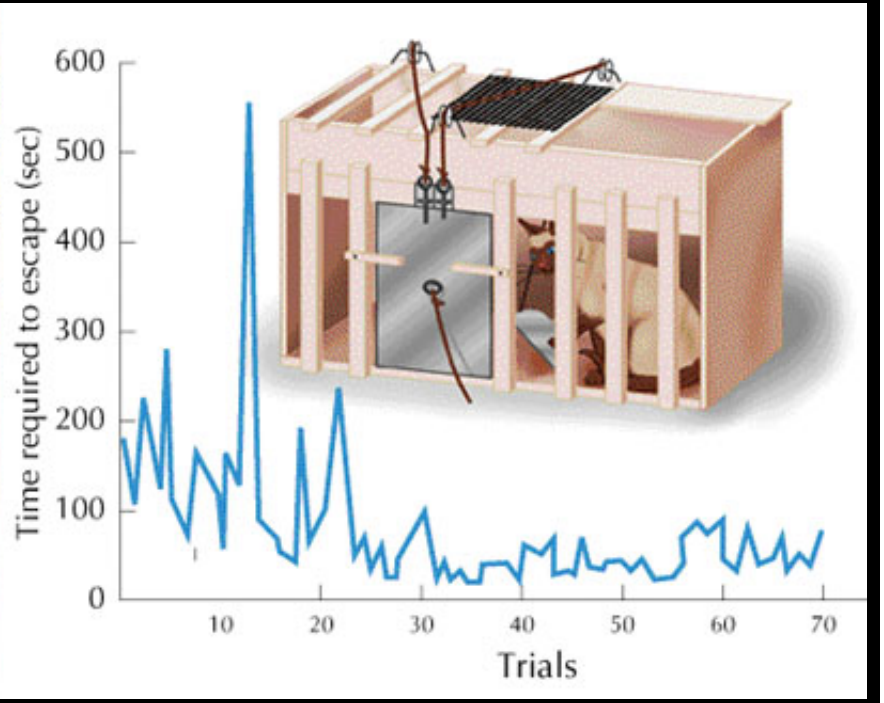
Blocking

(NB. Also in humans)

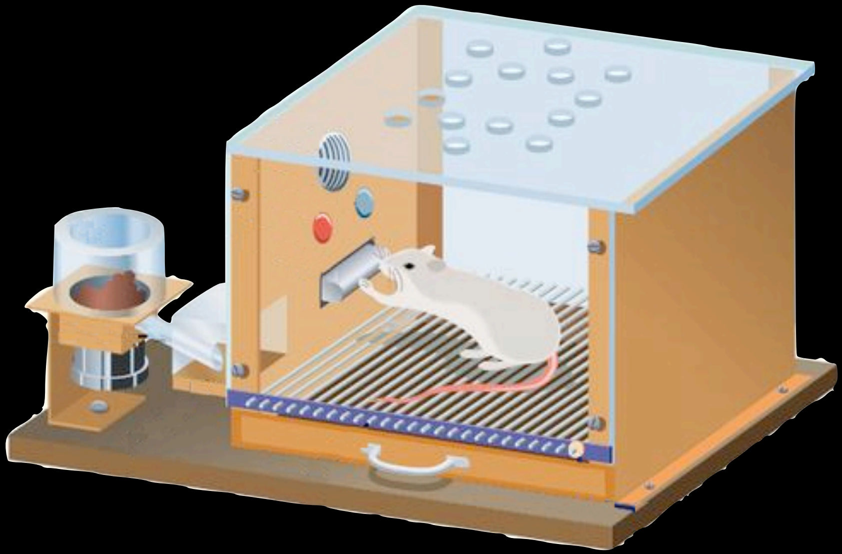
2. Instrumental conditioning: adding control

Edward Thorndike (law of effect)

- Background: Darwin, attempts to show that animals are intelligent
- Thorndike (age 23): submitted PhD thesis on "Animal intelligence: an experimental study of the associative processes in animals"
- Tested hungry cats in "puzzle boxes"
- Definition for learning: time to escape
- Gradual learning curves, did not look like 'insight' but rather trial and error



Example: Free operant conditioning (Skinner)

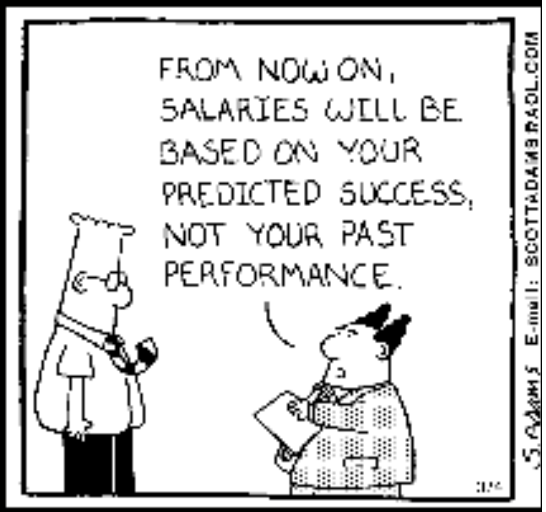


how is this related to RL?

animals can learn an arbitrary policy to obtain rewards (and avoid punishments)

Summary so far…

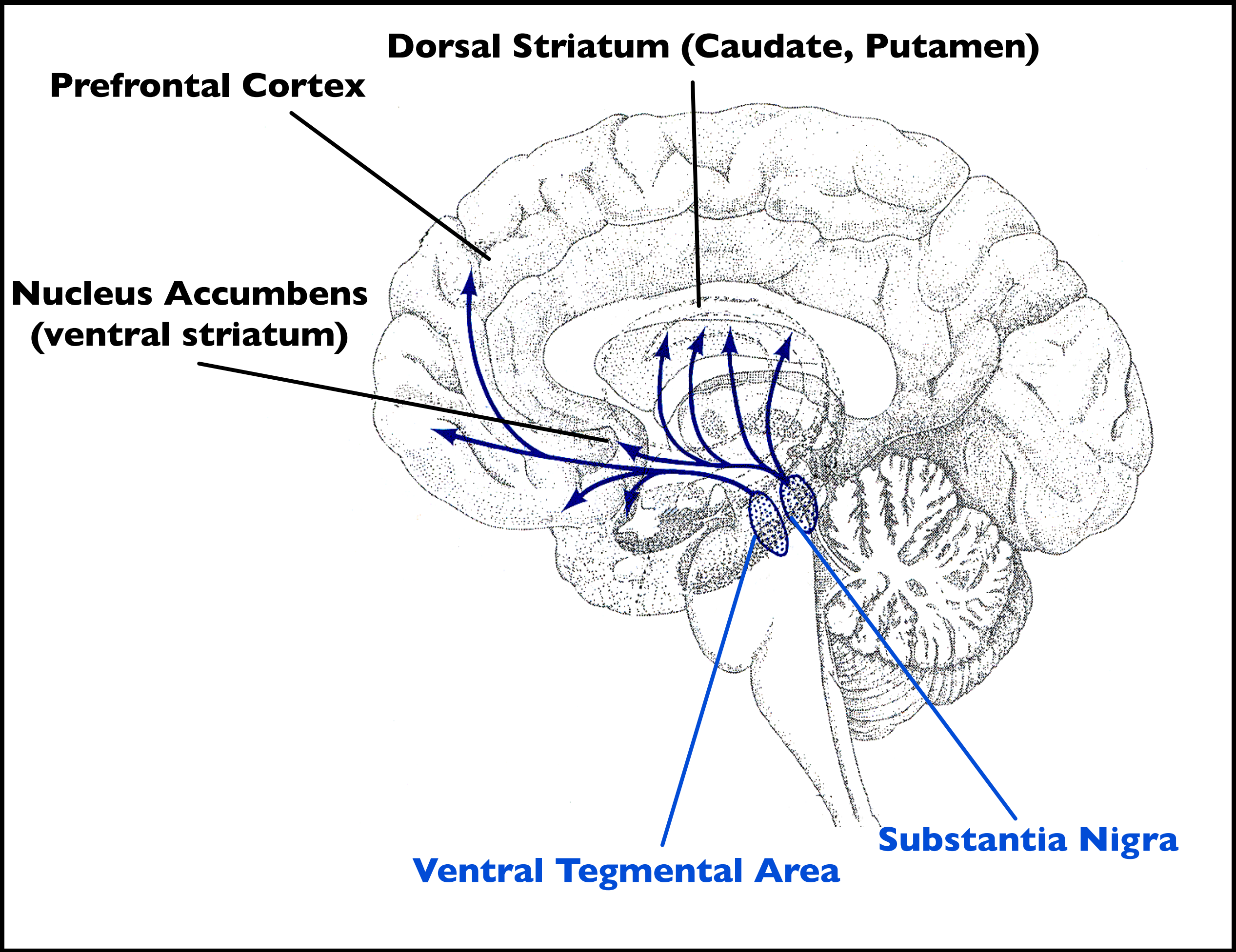
- The world presents animals/humans with a huge reinforcement learning problem (or many such small problems)
- Optimal learning and behavior depend on prediction and control
- How can the brain realize these? Can RL help us understand the brain's computations?



Outline

- The brain coarse-grain
- Learning and decision making in animals and humans: what does RL have to do with it?
- A success story: Dopamine and prediction errors
- Actor/Critic architecture in basal ganglia
- SARSA vs Q-learning: can the brain teach us about ML?
- Model free and model based RL in the brain
- Average reward RL & tonic dopamine
- Risk sensitivity and RL in the brain
- Open challenges and future directions

What is dopamine and why do we care about it?



Parkinson's Disease

→ Motor control / initiation?

Drug addiction, gambling,
Natural rewards

→ Reward pathway?

→ Learning?

Also involved in:

- Working memory
- Novel situations
- ADHD
- Schizophrenia
- …

role of dopamine: many hypotheses

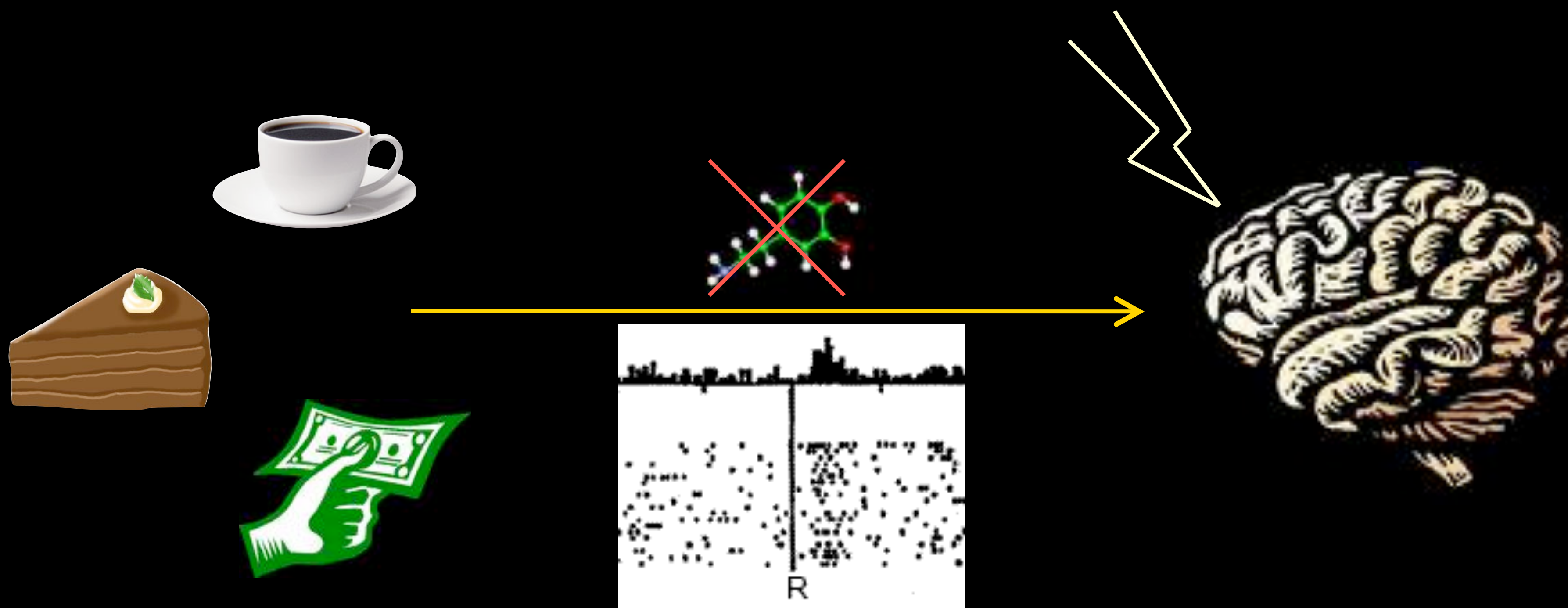
- Anhedonia hypothesis
- Prediction error hypothesis
- Salience/attention
- (Uncertainty)
- Incentive salience
- Cost/benefit computation
- Energizing/motivating behavior

the anhedonia hypothesis (Wise, '80s)

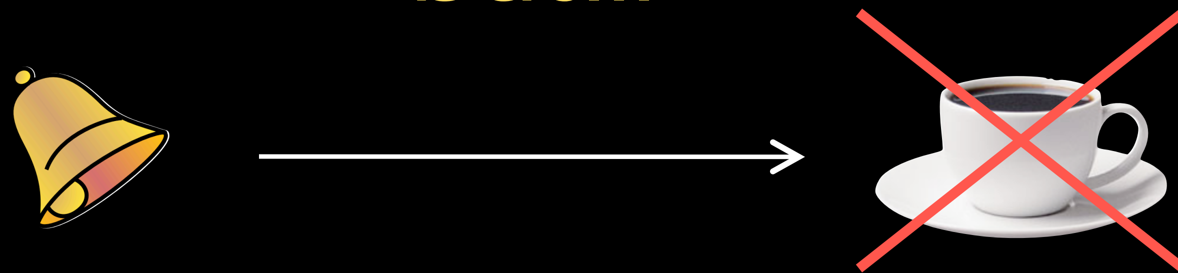
- Anhedonia = inability to experience positive emotional states derived from obtaining a desired or biologically significant stimulus
- Neuroleptics (dopamine antagonists) cause anhedonia
- Dopamine is important for reward-mediated conditioning

the anhedonia hypothesis (Wise, ’80s)

- Anhedonia = inability to experience positive emotional states derived from obtaining a desired or biologically significant stimulus
- Neuroleptics (dopamine antagonists) cause anhedonia
- Dopamine is important for reward-mediated conditioning

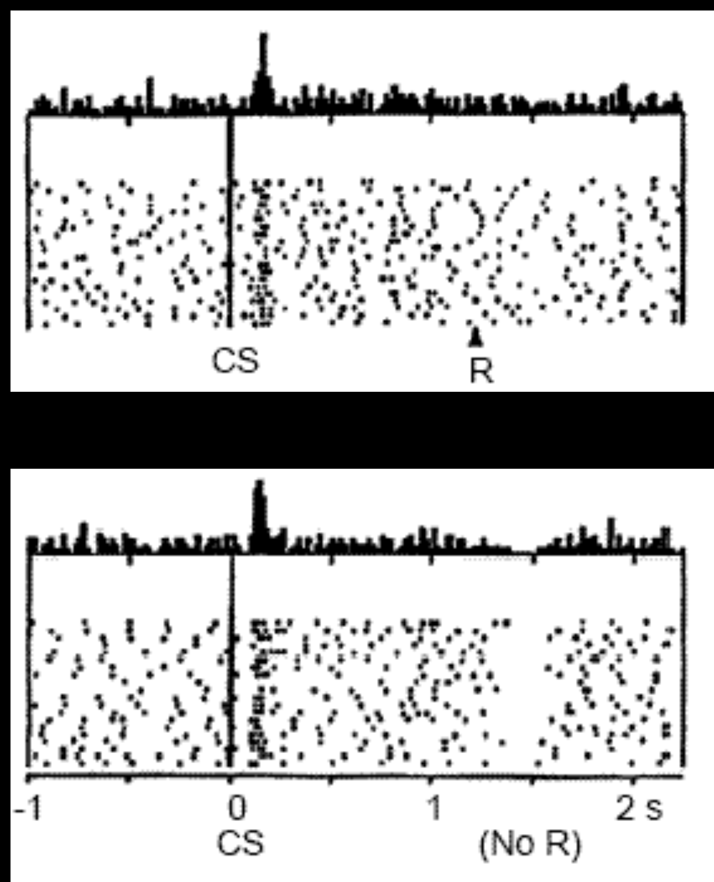


but...



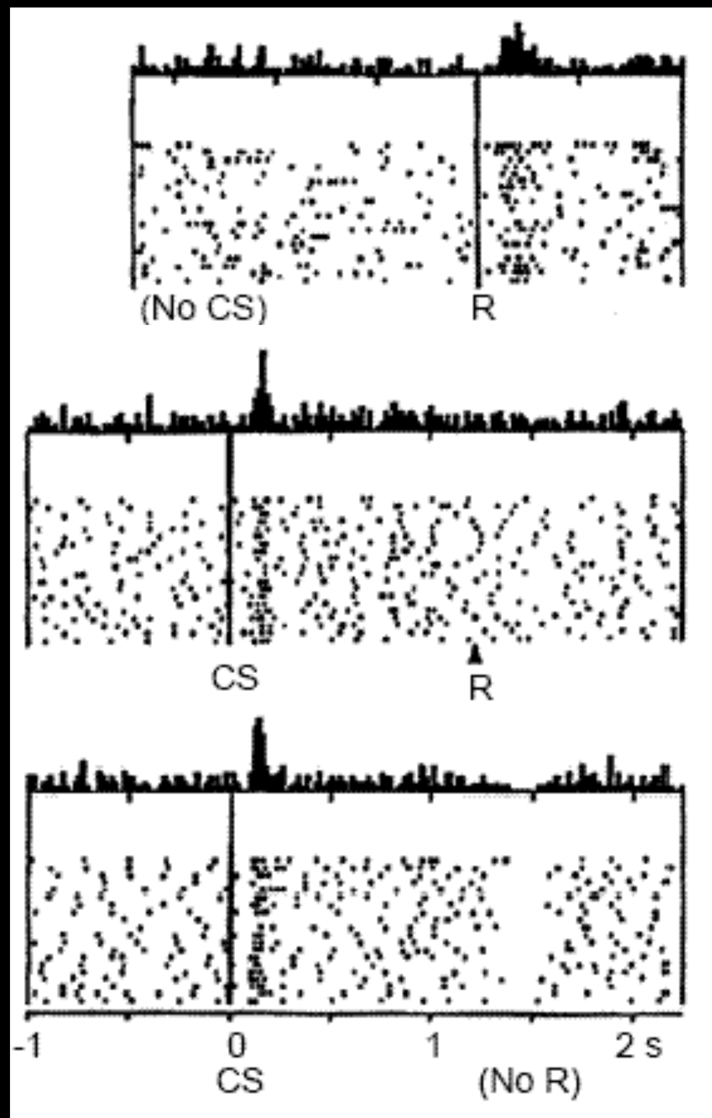
predictable
reward

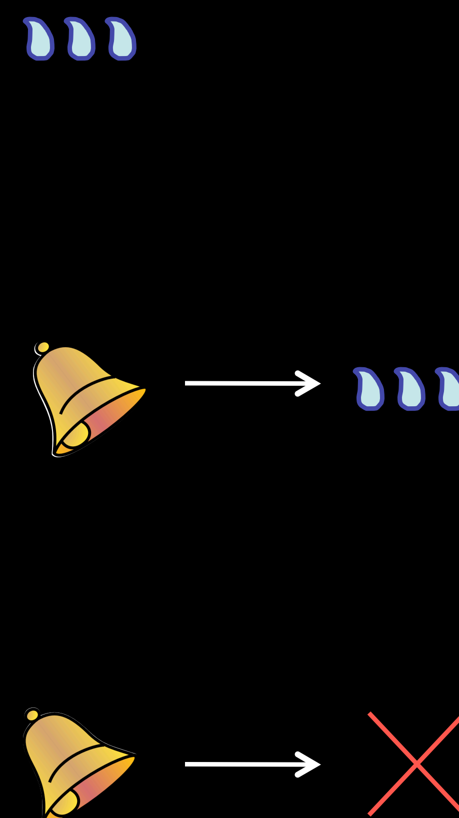
omitted
reward



(Schultz et al. '90s)

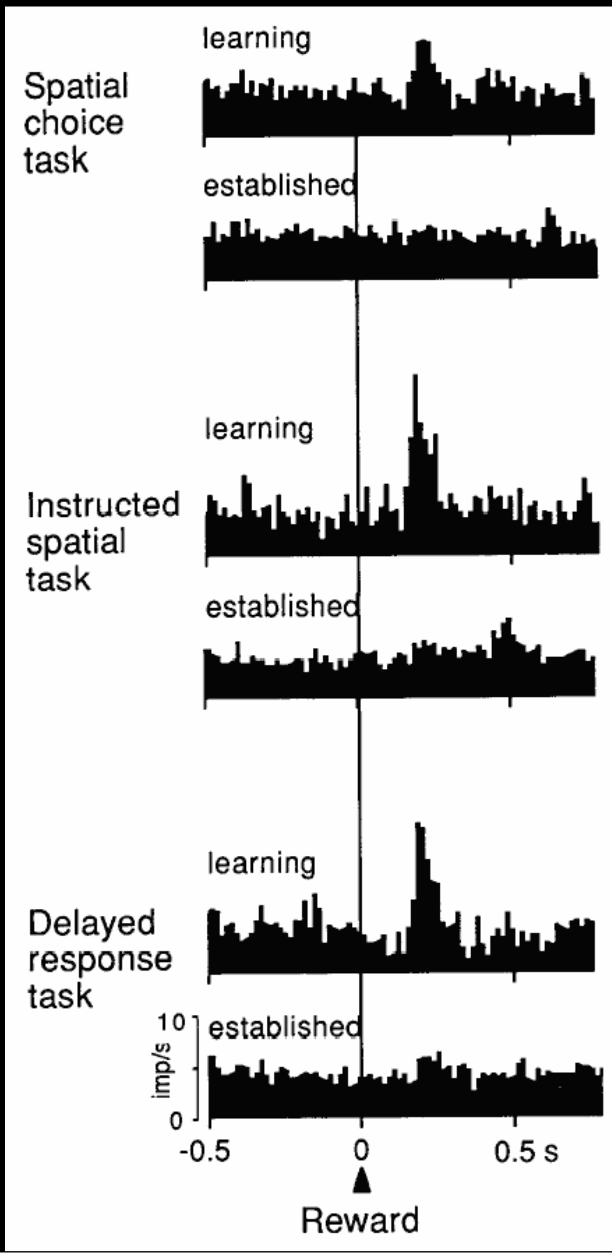
looks familiar?





Schultz, Dayan, Montague, 1997

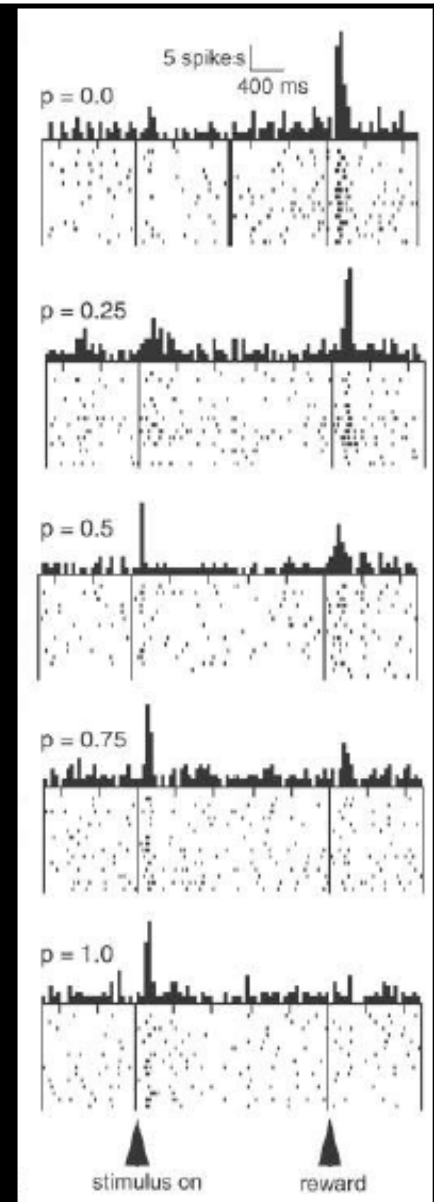
prediction error hypothesis of dopamine



Schultz et al, 1993

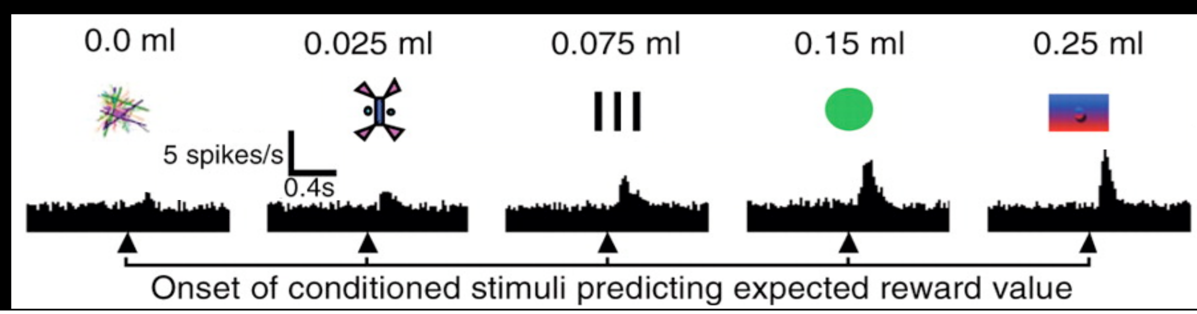
The idea: Dopamine encodes a temporal difference reward prediction error

(Montague, Dayan, Barto mid 90's)

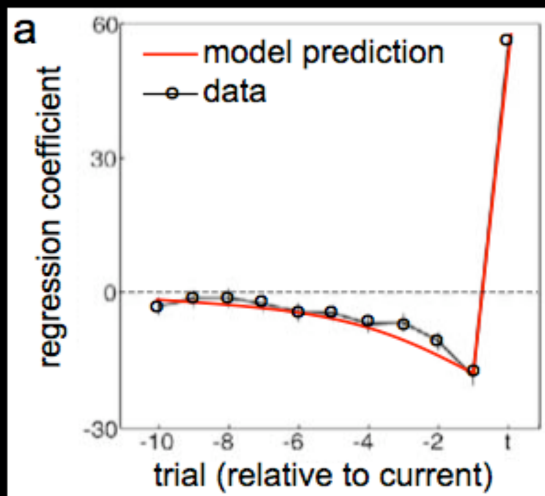


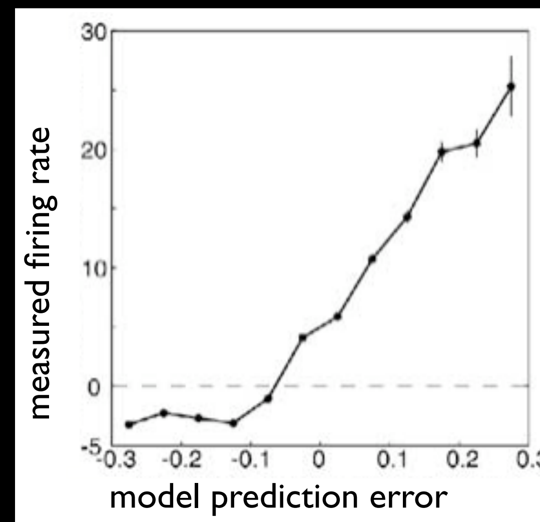
Fiorillo et al, 2003

Tobler et al, 2005

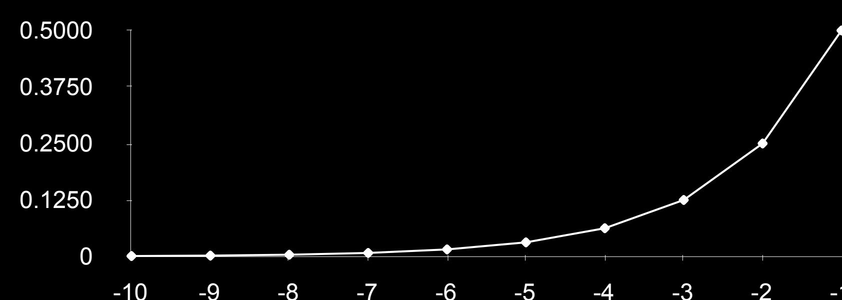


prediction error hypothesis of dopamine: stringent tests





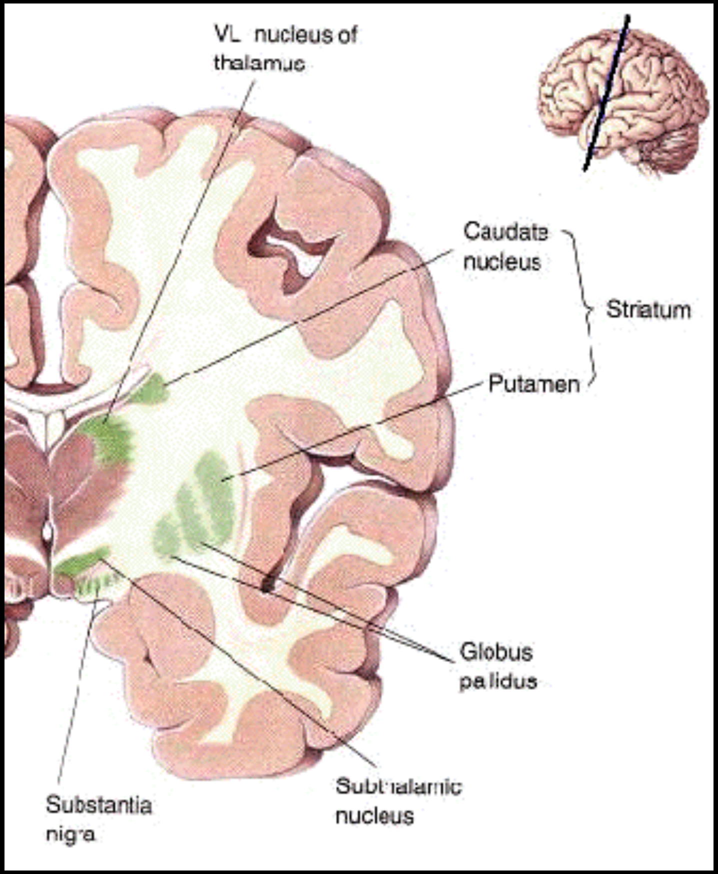
$$V_t = \eta \sum_{i=1}^{t} (1-\eta)^{t-i} r_i$$

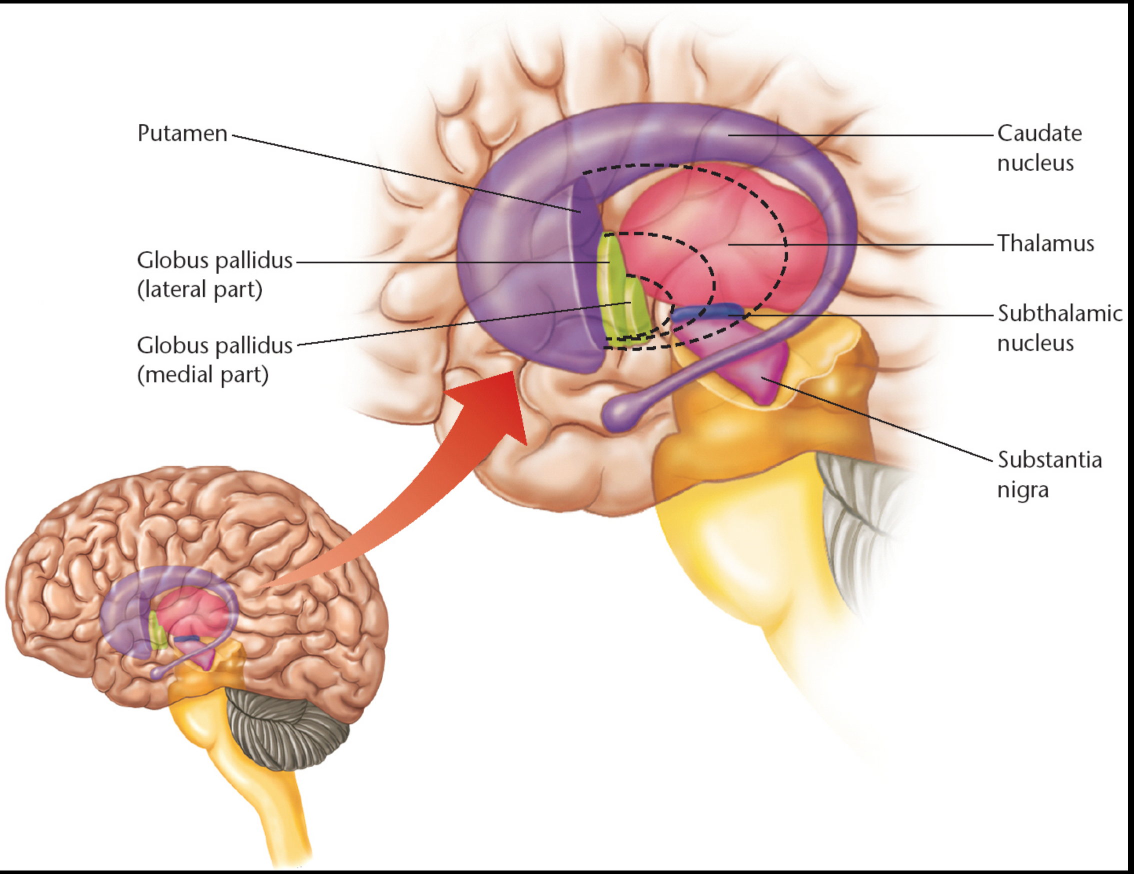


Bayer & Glimcher (2005)

where does dopamine project to?

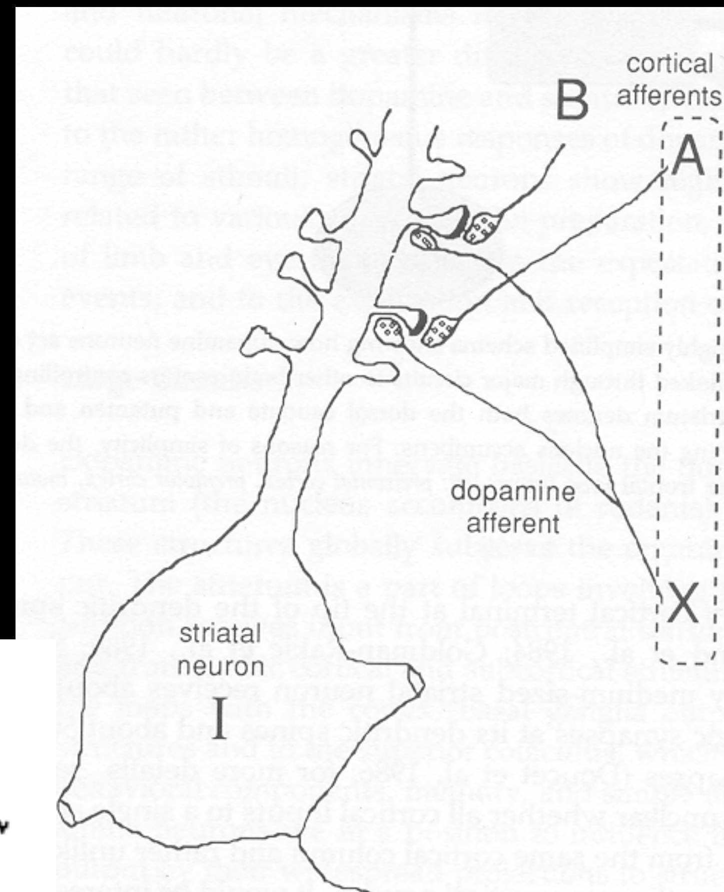
main target: basal ganglia

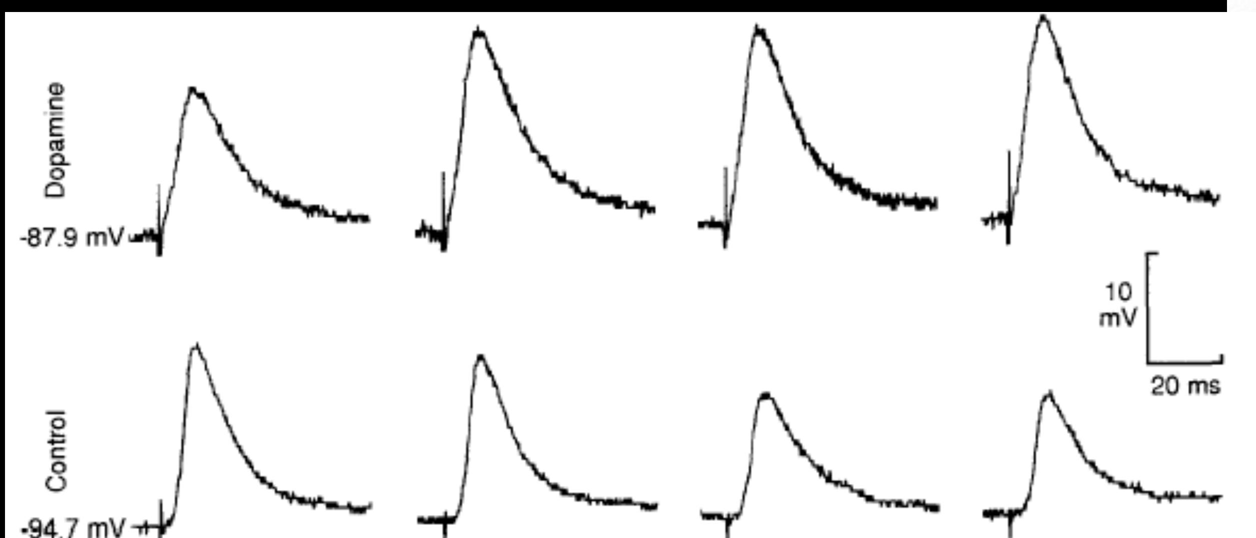




dopamine and synaptic plasticity

- prediction errors are for learning...
- cortico-striatal synapses show dopamine-dependent plasticity
- three-factor learning rule: need presynaptic+postsynaptic+dopamine





Wickens et al, 1996

Summary so far...

Conditioning can be viewed as prediction learning

- The problem: prediction of future reward
- The algorithm: temporal difference learning
- Neural implementation: dopamine dependent learning in corticostriatal synapses in the basal ganglia

$\Rightarrow$ Precise (normative!) theory for generation of dopamine firing patterns

$\Rightarrow$ A computational model of learning allows us to look in the brain for "hidden variables" postulated by the model

Outline

- The brain coarse-grain
- Learning and decision making in animals and humans: what does RL have to do with it?
- A success story: Dopamine and prediction errors
- Actor/Critic architecture in basal ganglia
- SARSA vs Q-learning: can the brain teach us about ML?
- Model free and model based RL in the brain
- Average reward RL & tonic dopamine
- Risk sensitivity and RL in the brain
- Open challenges and future directions

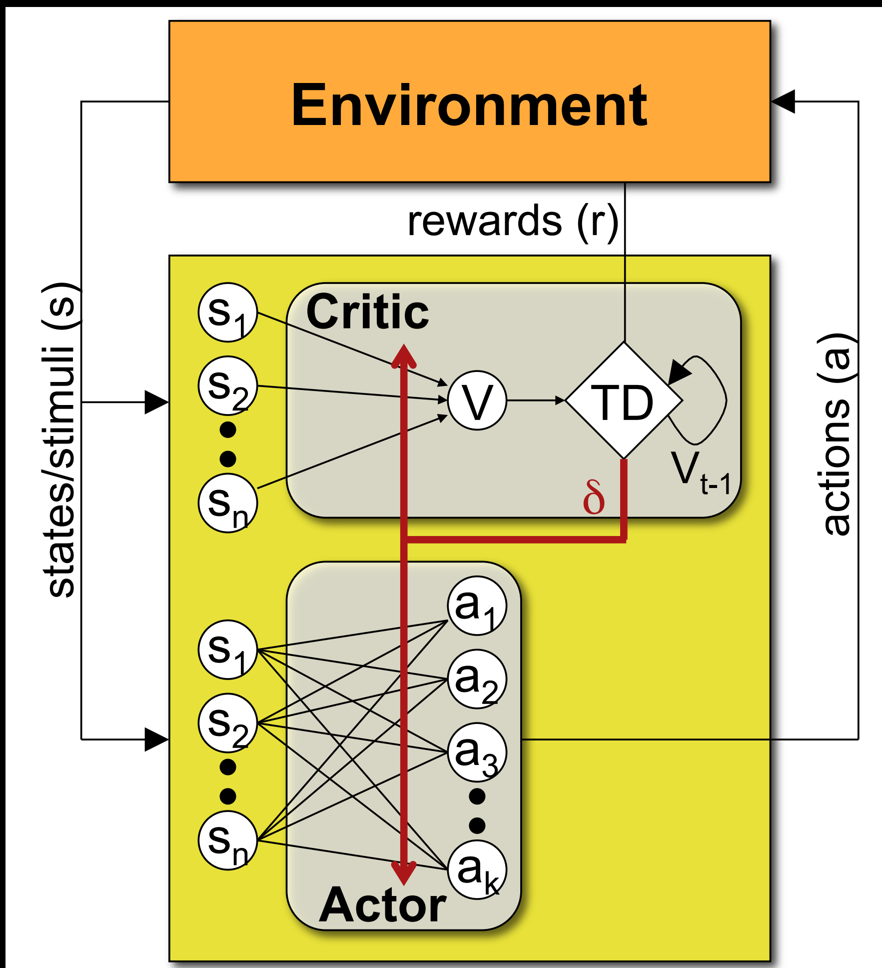
3 model-free learning algorithms

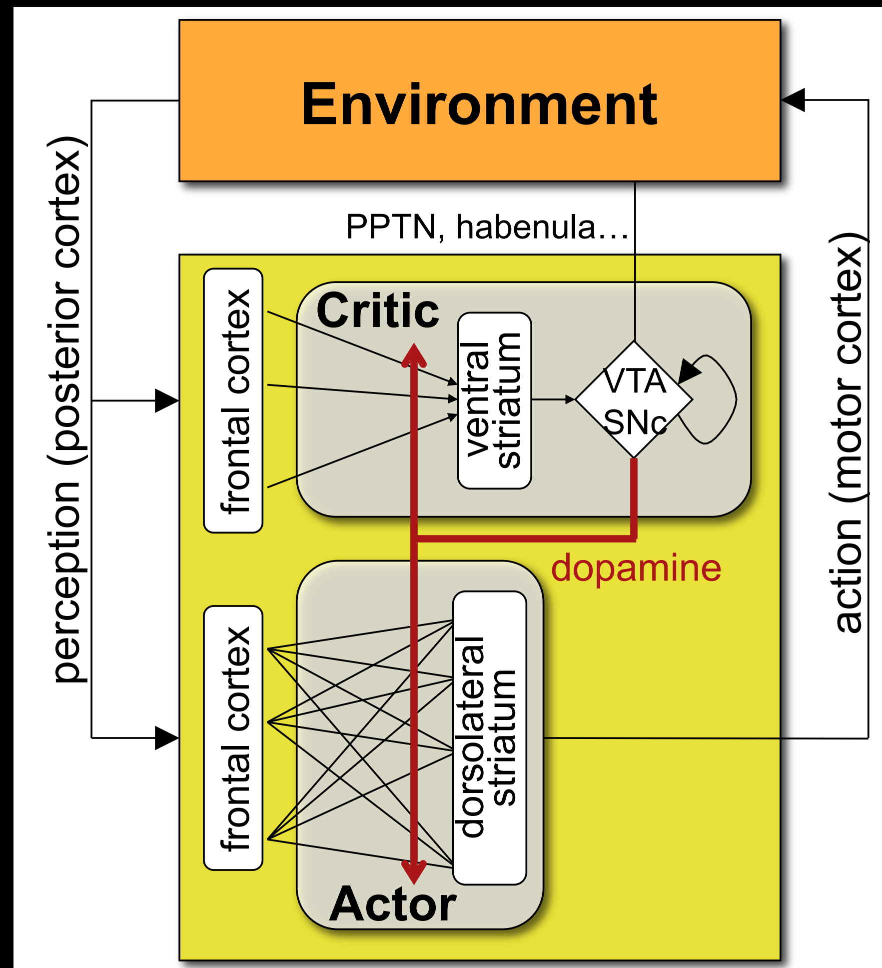
Actor/Critic

Q learning

SARSA

Actor/Critic in the brain?





evidence for this?

short aside: functional magnetic resonance imaging (fMRI)

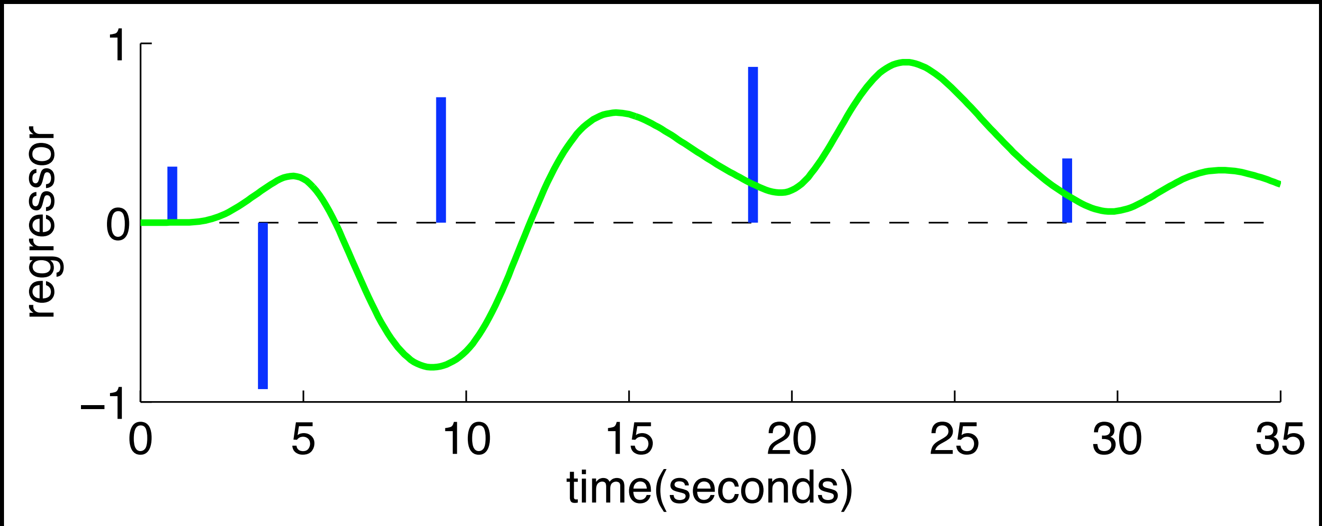
- measure BOLD ("blood oxygenation level dependent") signal
- oxygenated vs de-oxygenated hemoglobin have different magnetic properties
- detected by big superconducting magnet

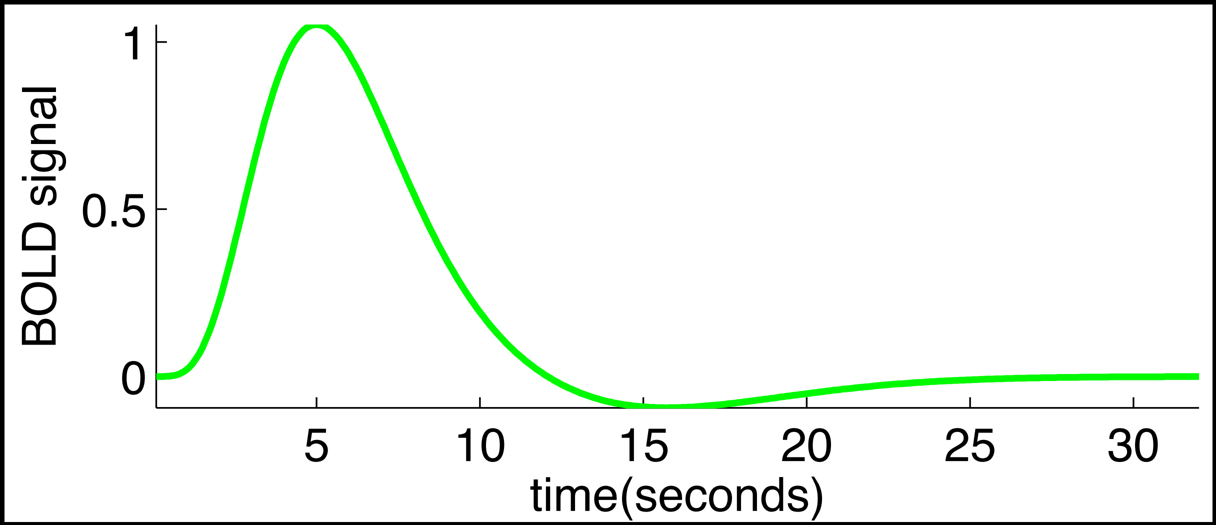
<u>Idea</u>:

- Brain is functionally modular
- Neural activity uses energy & oxygen
- Measure brain usage, not structure

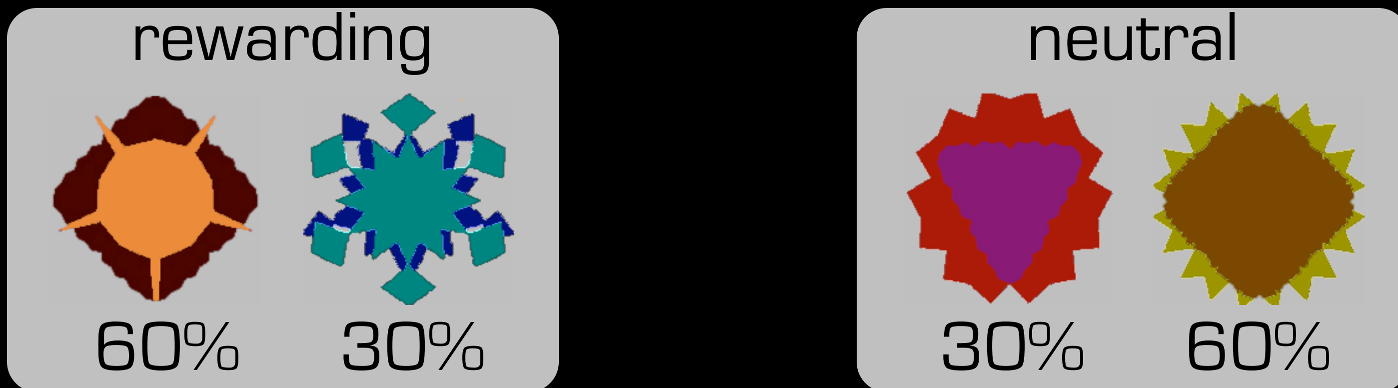
- Spatial resolution: ~3mm 3D "voxels"
- temporal resolution: 5-10 seconds

short aside: functional magnetic resonance imaging (fMRI)





Back to Actor/Critic: Evidence from fMRI

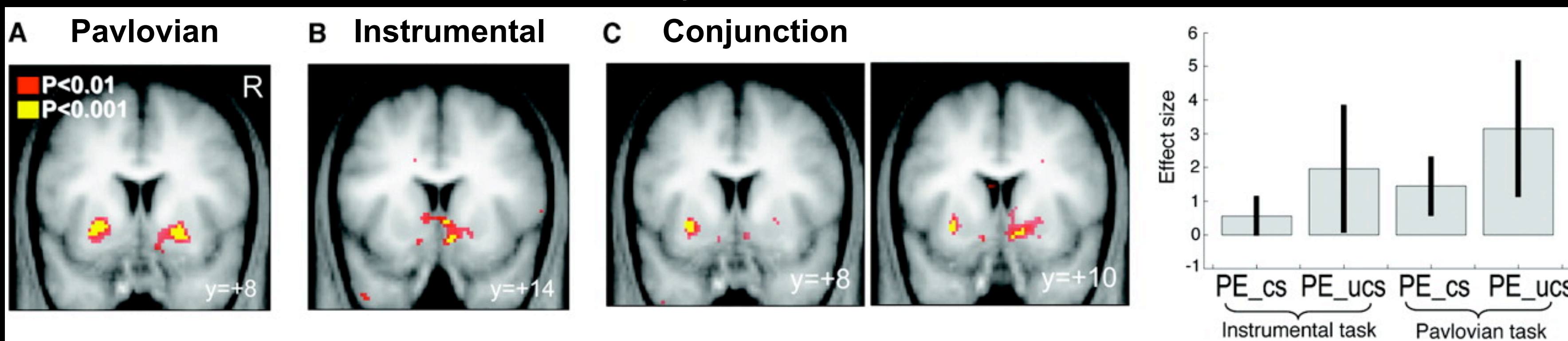


- cond 1: instrumental (choose stimuli) - show preference for high probability stimulus in rewarding but not neutral trials
- cond 2: Pavlovian - only indicate the side the 'computer' has selected (RTs as measure of learning)
- why was the experiment designed this way (hint: think of prediction errors)

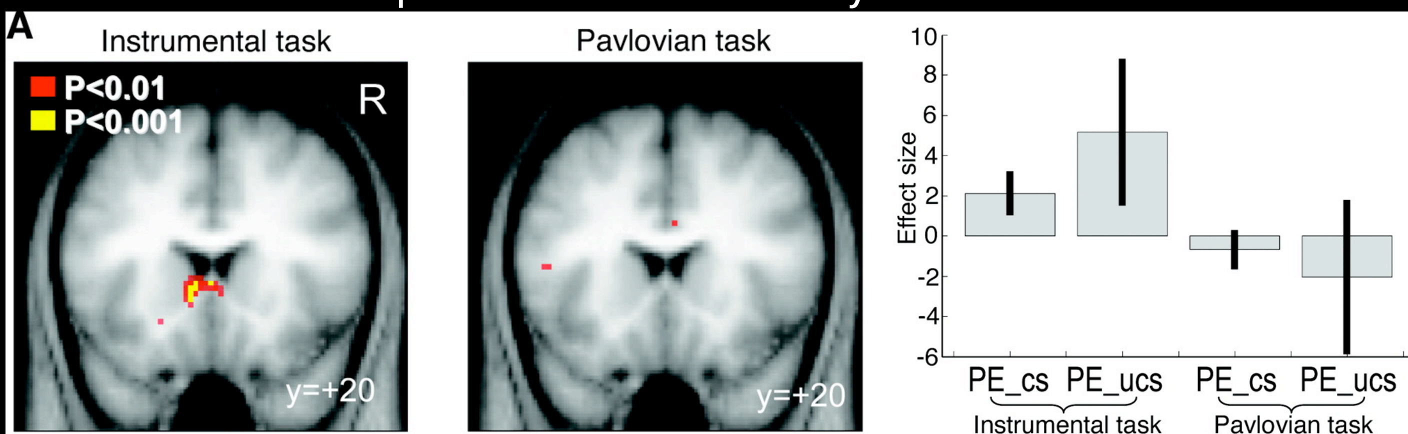
O'Doherty et al. 2004

Back to Actor/Critic: Evidence from fMRI

ventral striatum: correlated with prediction error in both conditions

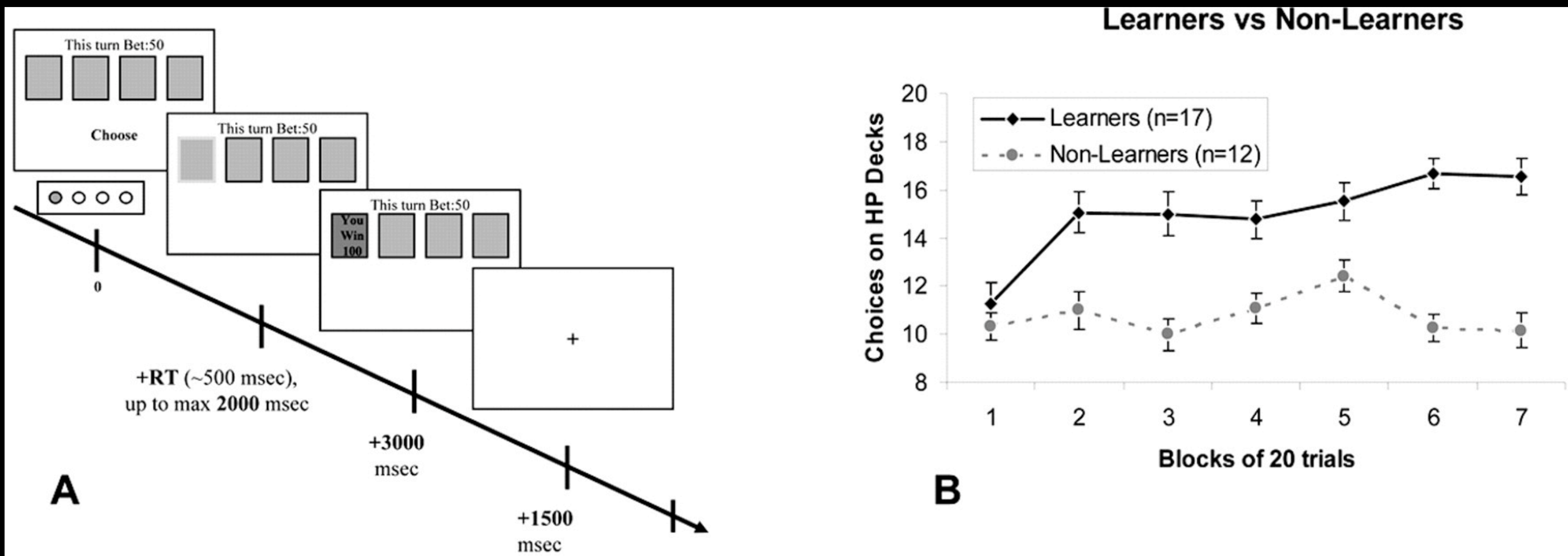


Dorsal striatum: prediction error only in instrumental task



O'Doherty et al. 2004

do prediction errors really influence learning?

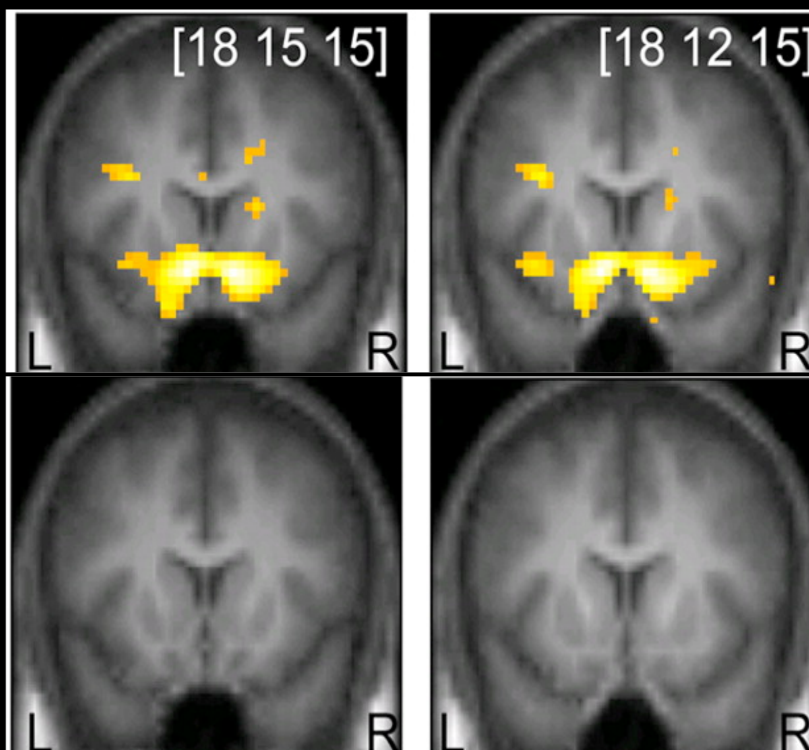


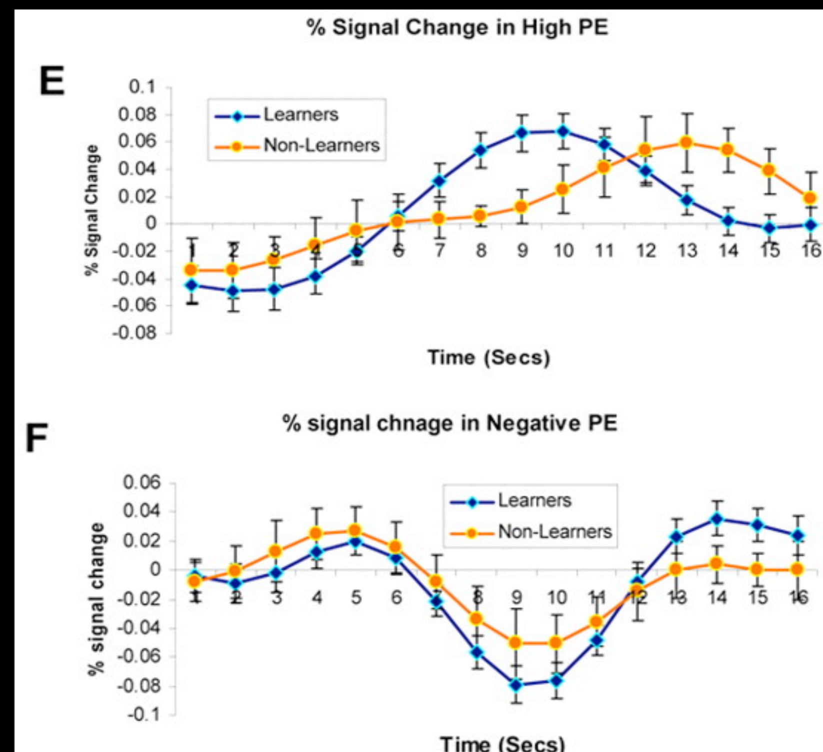
Schonberg et al. 2007

do prediction errors really influence learning?

Learners

Non-Learners





Schonberg et al. 2007

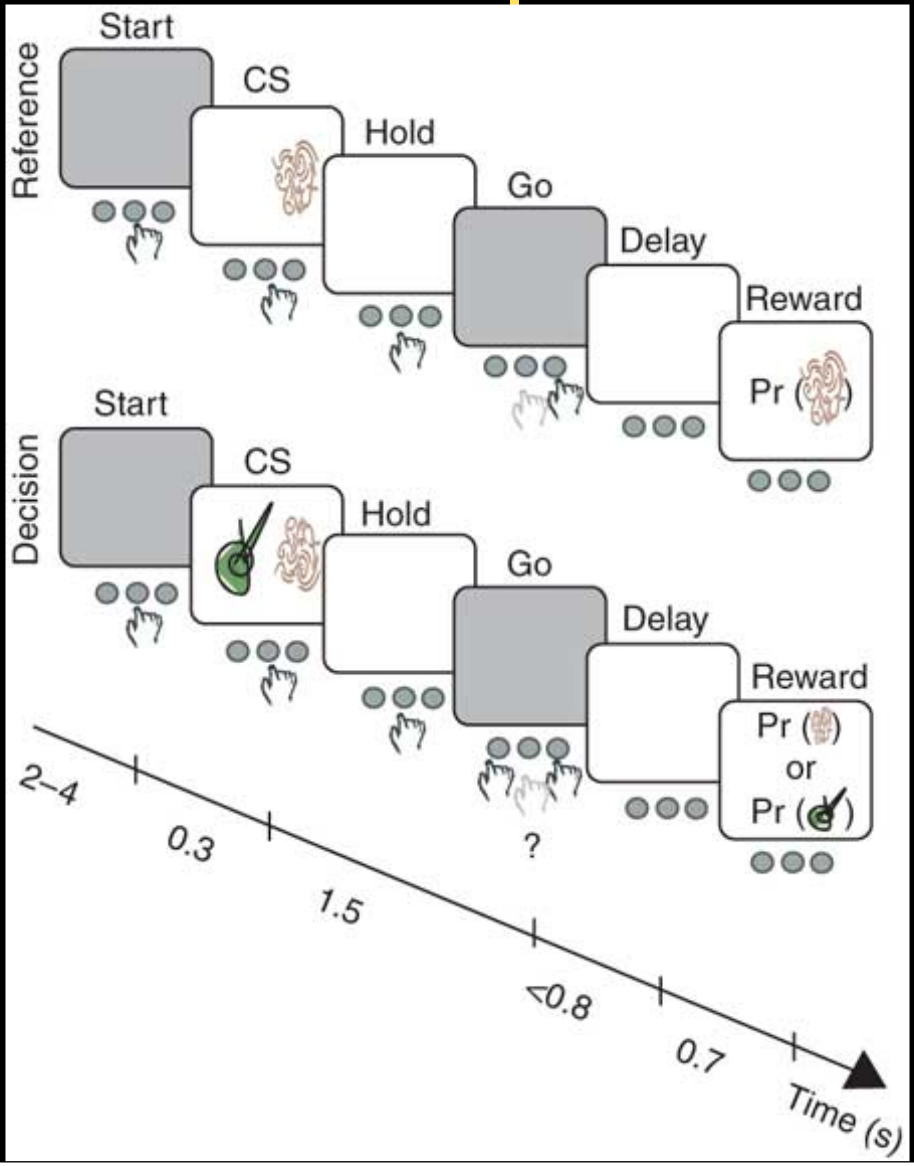
Summary so far...

- Some evidence for an Actor/Critic architecture in the brain
- Links predictions (Critic) to control (Actor) in very specific way; assumes no Q values
- (Not at all conclusive evidence)

Outline

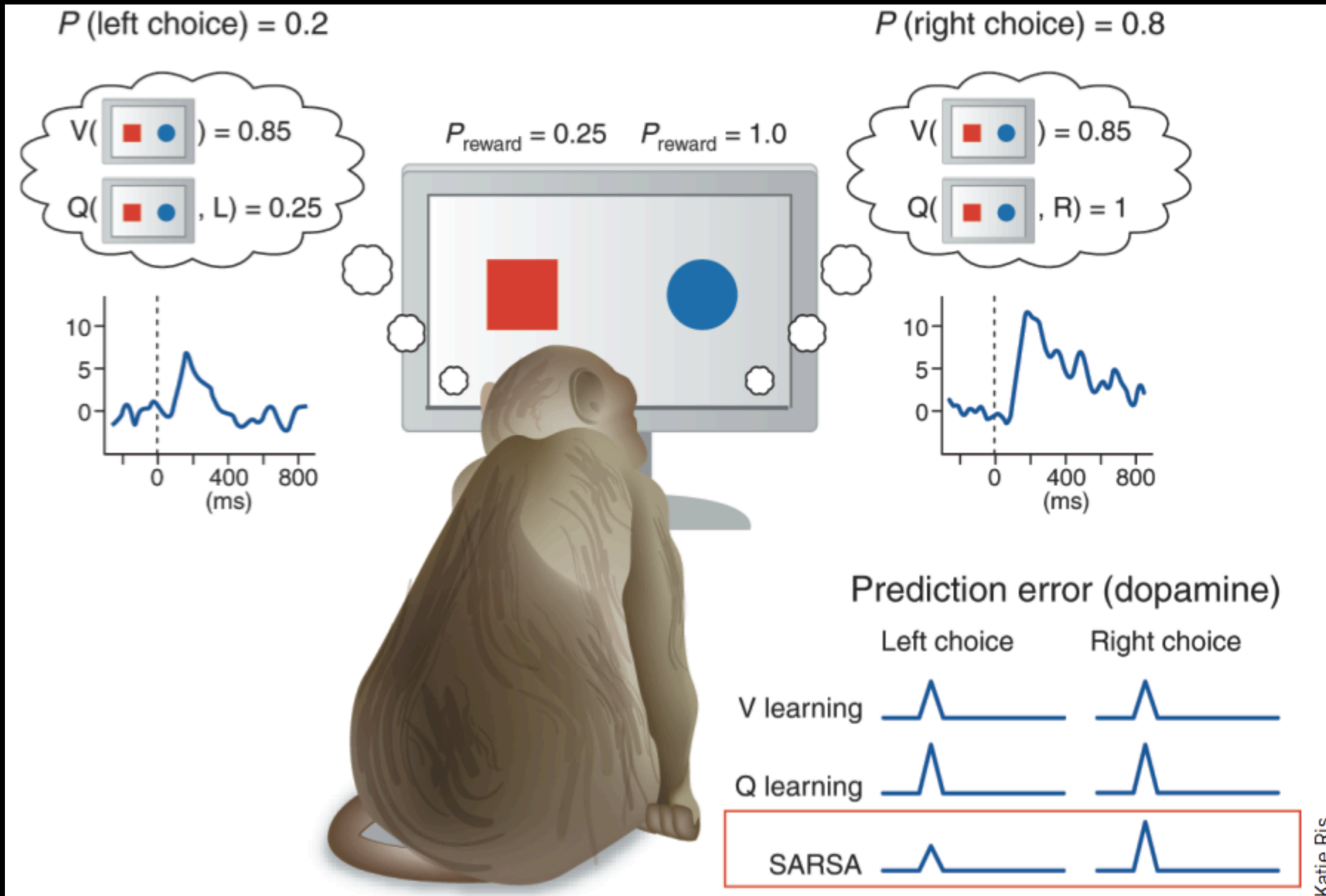
- The brain coarse-grain
- Learning and decision making in animals and humans: what does RL have to do with it?
- A success story: Dopamine and prediction errors
- Actor/Critic architecture in basal ganglia
- SARSA vs Q-learning: can the brain teach us about ML?
- Model free and model based RL in the brain
- Average reward RL & tonic dopamine
- Risk sensitivity and RL in the brain
- Open challenges and future directions

do dopamine prediction errors at trial onset represent V(S)?



Morris et al. 2005

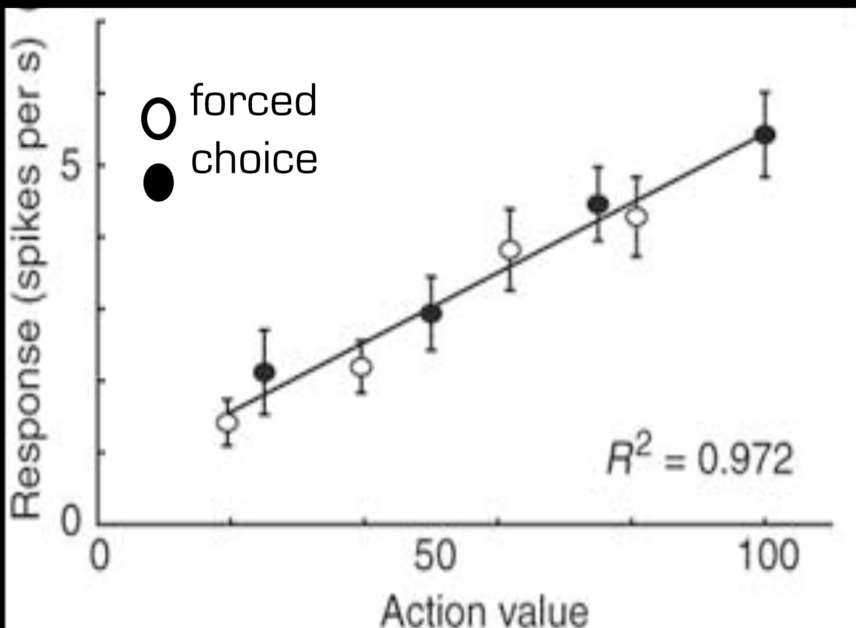
do dopamine prediction errors at trial onset represent V(S)?

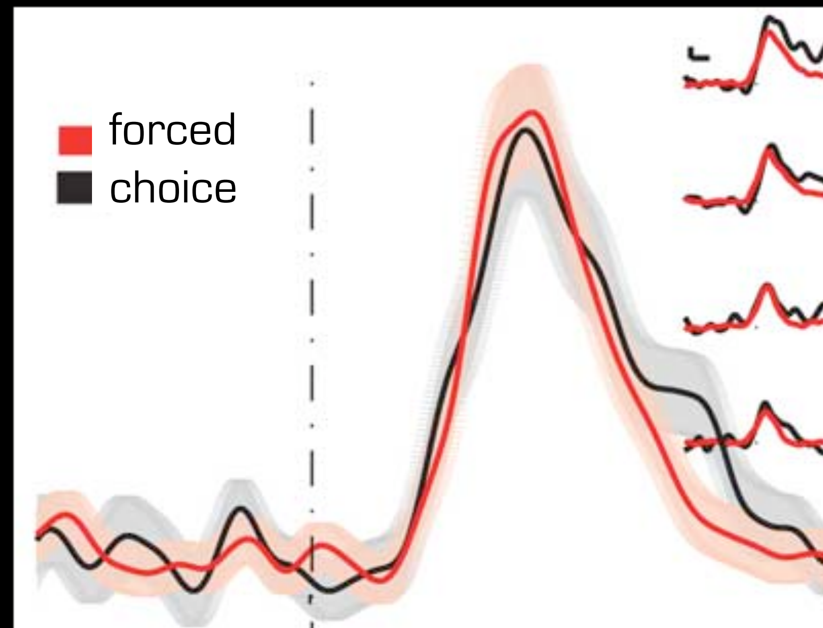


Morris et al. 2005

do dopamine prediction errors at trial onset represent V(S)?

stimulus on





Morris et al. 2005

but... another study suggests otherwise

BLOCK1 odor

immediate reward → delayed reward

BLOCK2 odor

delayed reward → immediate reward

BLOCK3 odor

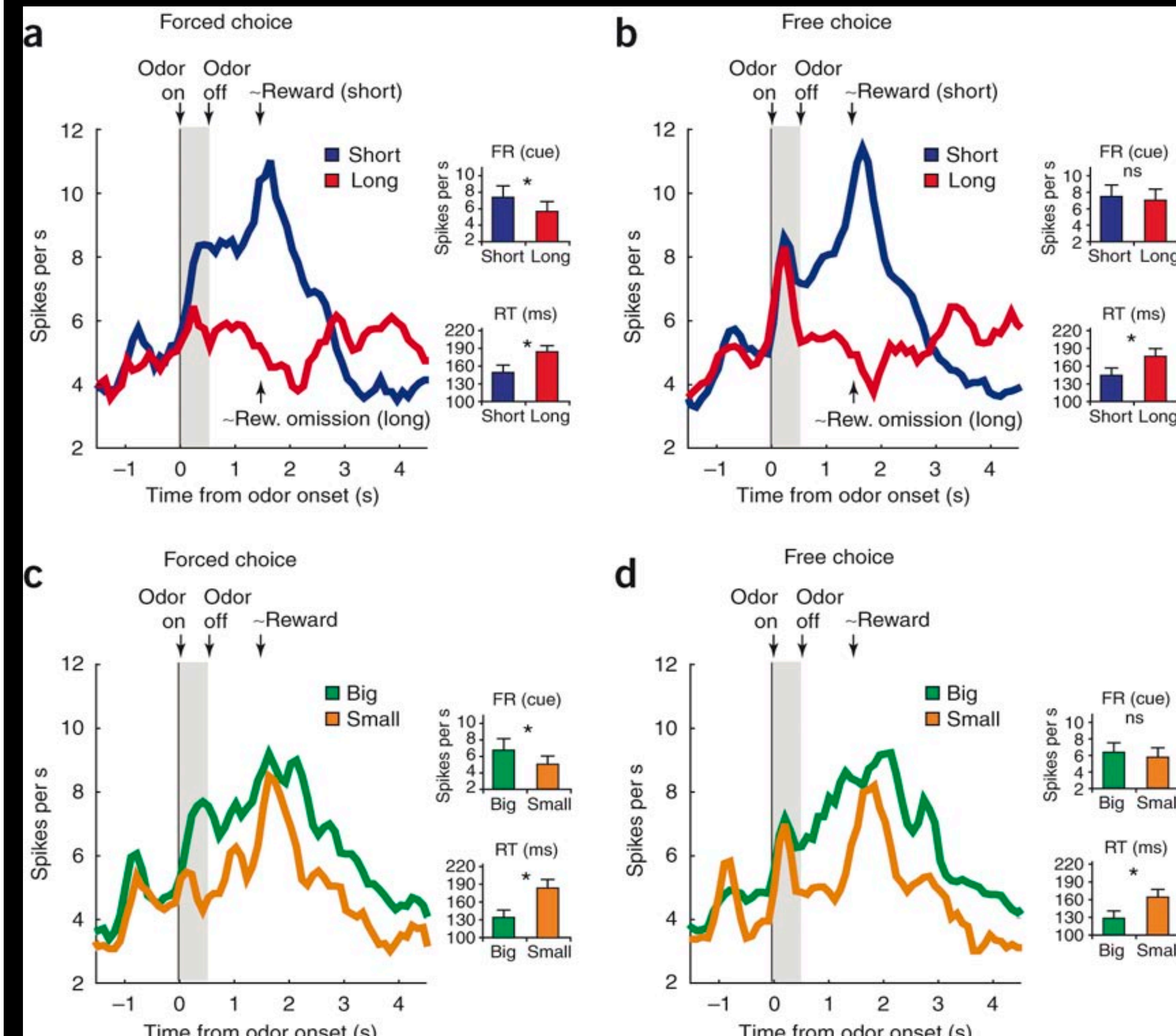
small reward → large reward

BLOCK4 odor

large reward → small reward

Roesch et al. 2007

but... another study suggests otherwise



Differences from Morris et al. (2005):

- rats not monkeys
- VTA not SNc
- amount of training
- task (representation of stimuli?)

(notice the messy signal... due to measurement or is it that way in the brain?)

Roesch et al. 2007

Summary so far...

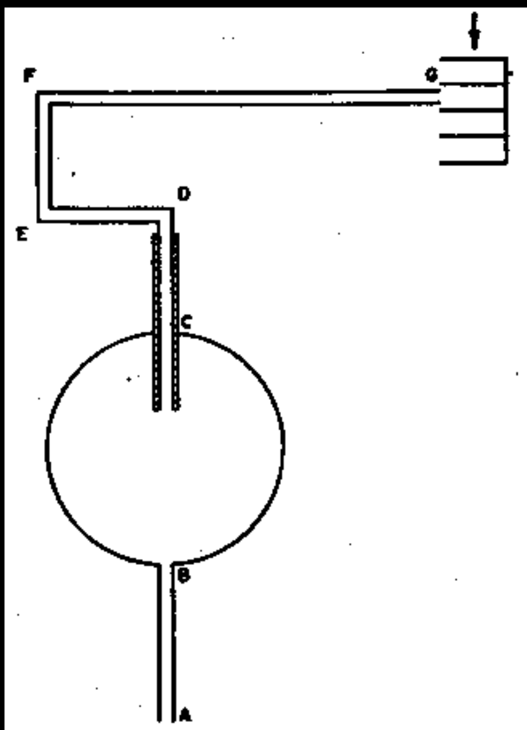
- SARSA or Q-learning? The jury is still out
- What needs to be done: more experiments recording from dopamine in telltale tasks
- The brain (dopamine) can inform RL: how does it learn in real time, with real noise, in real problems?

Outline

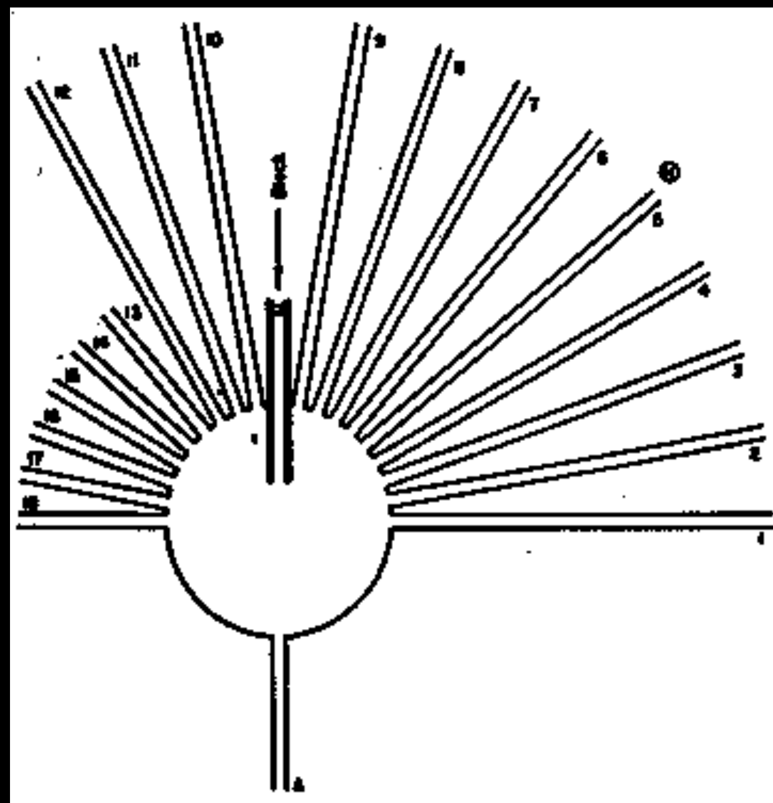
- The brain coarse-grain
- Learning and decision making in animals and humans: what does RL have to do with it?
- A success story: Dopamine and prediction errors
- Actor/Critic architecture in basal ganglia
- SARSA vs Q-learning: can the brain teach us about ML?
- Model free and model based RL in the brain
- Average reward RL & tonic dopamine
- Risk sensitivity and RL in the brain
- Open challenges and future directions

do animals only learn action policies?

training:



test:

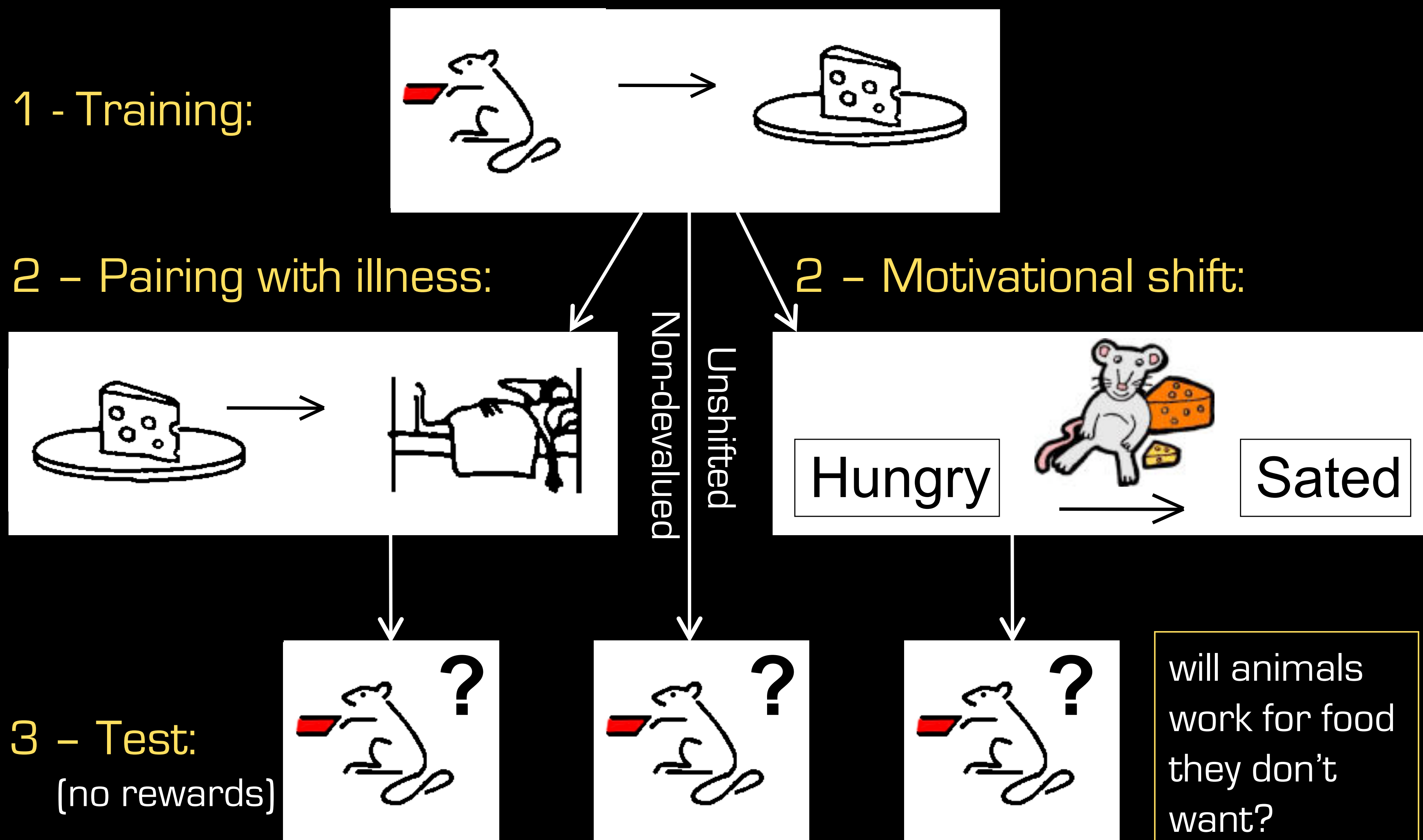


result:

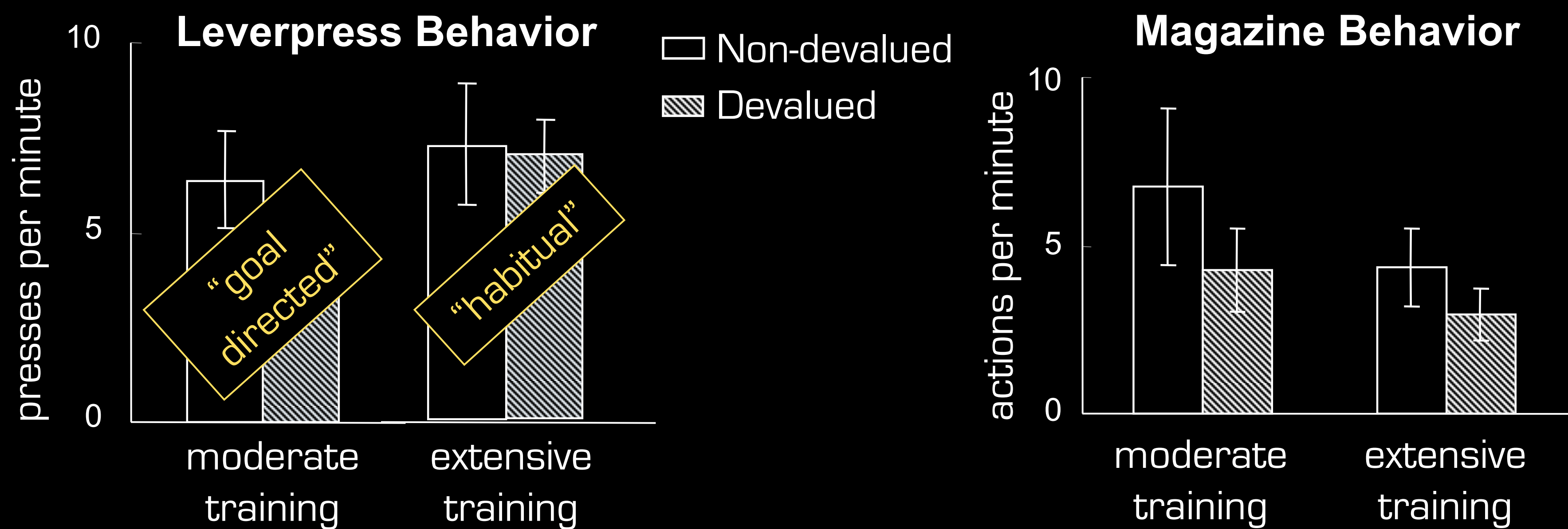
Even the humble rat can can learn spatial structure, and use it to plan flexibly

Tolman et al (1946) 60

another test: outcome devaluation



devaluation: results



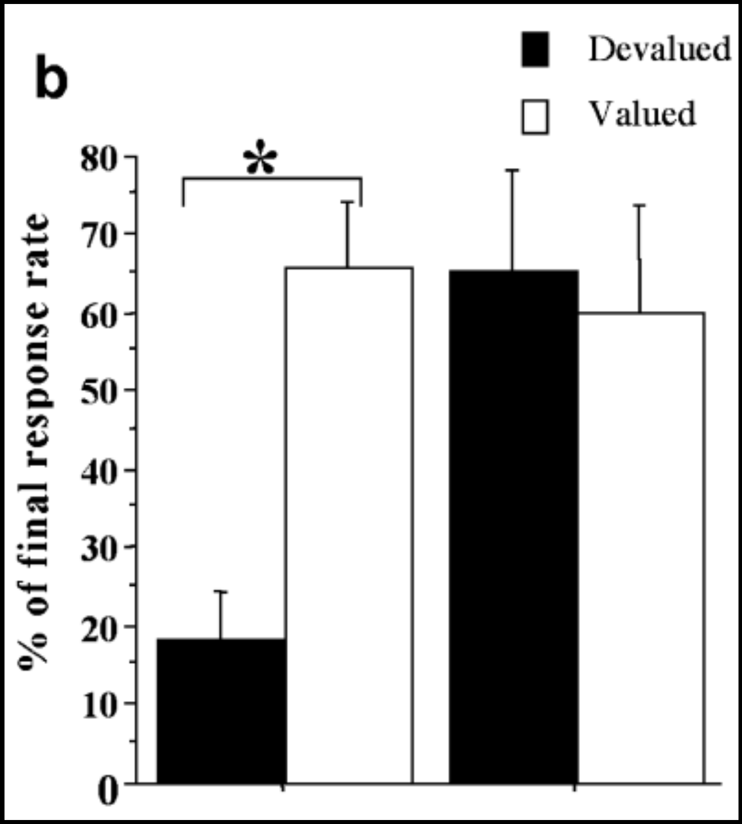
Animals will sometimes work for food they don't want!

→ in daily life: actions become automatic with repetition

Holland (2004), Kilcross & Coutureau (2003)

devaluation: results from lesions I

overtrained rats



dorsolateral striatum lesion

control (sham lesion)

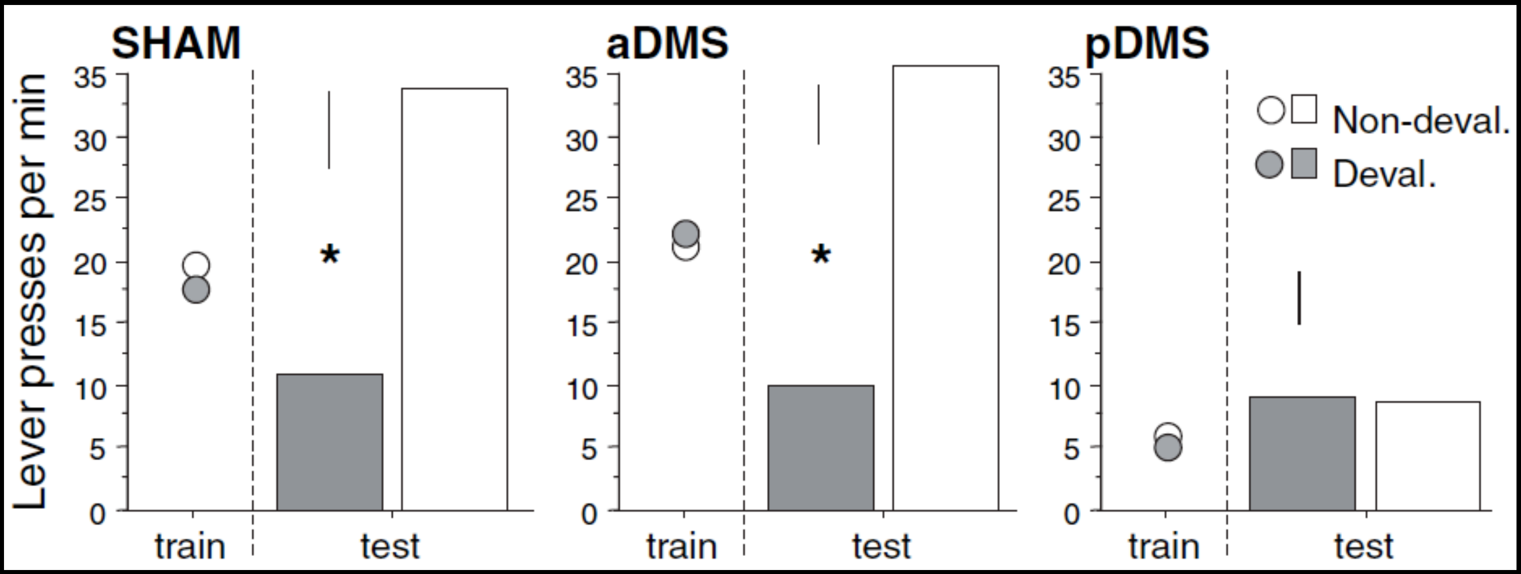
→ animals with lesions to DLS never develop habits despite extensive training

→ also treatments depleting dopamine in DLS

→ also inactivations of infra-limbic PFC after training

Yin et al (2004), Coutureau & Killcross (2003)

devaluation: results from lesions II



lesions of the pDMS cause animals to leverpress habitually even with only moderate training

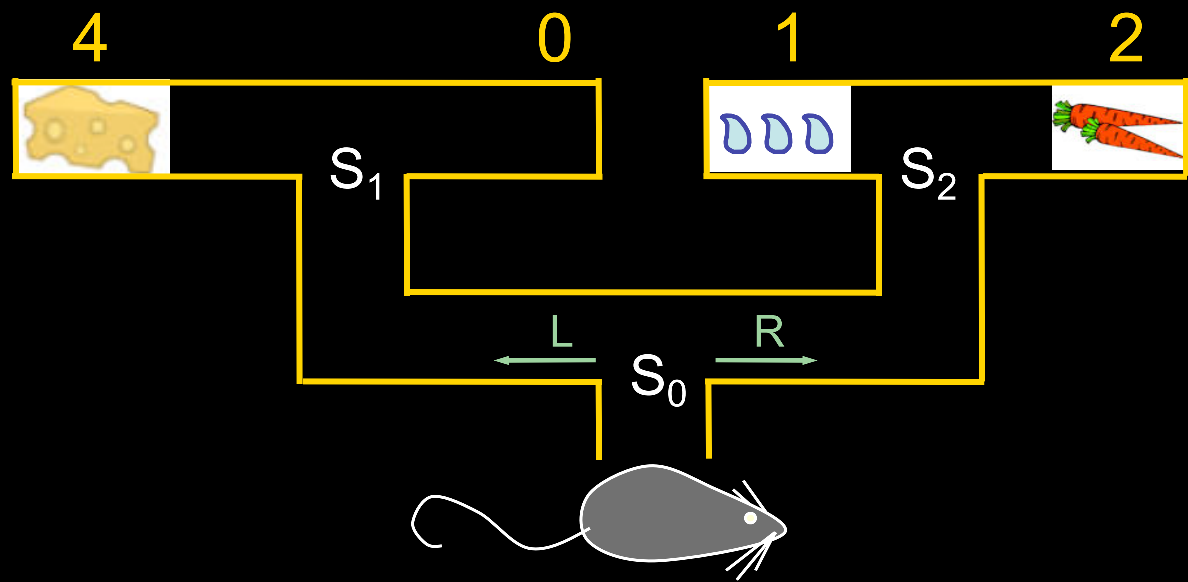
(also.. pre-limbic PFC, dorsomedial thalamus)

Yin, Ostlund, Knowlton & Balleine (2005)

what does all this mean?

- The same action (leverpressing) can arise from two psychologically dissociable pathways
    1. moderately trained behavior is "goal-directed": dependent on outcome representation
    2. overtrained behavior is "habitual": apparently not dependent on outcome representation
- Lesions suggest two parallel systems; the intact one can apparently support behavior at any stage
- Can RL help us make sense of this mess?

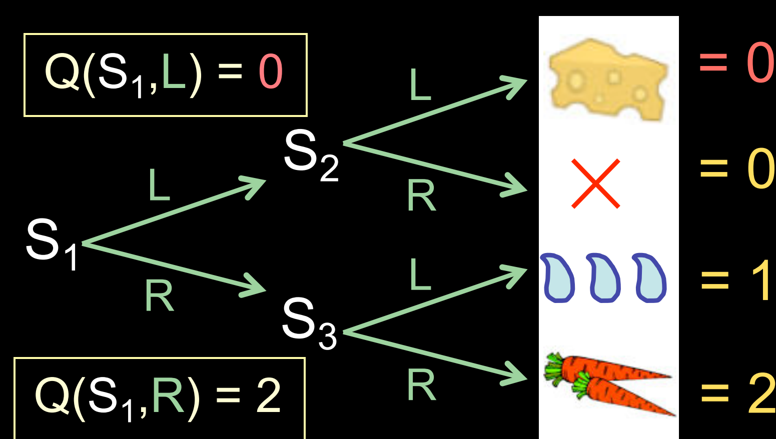
strategy 1: model based RL



learn model of task through experience

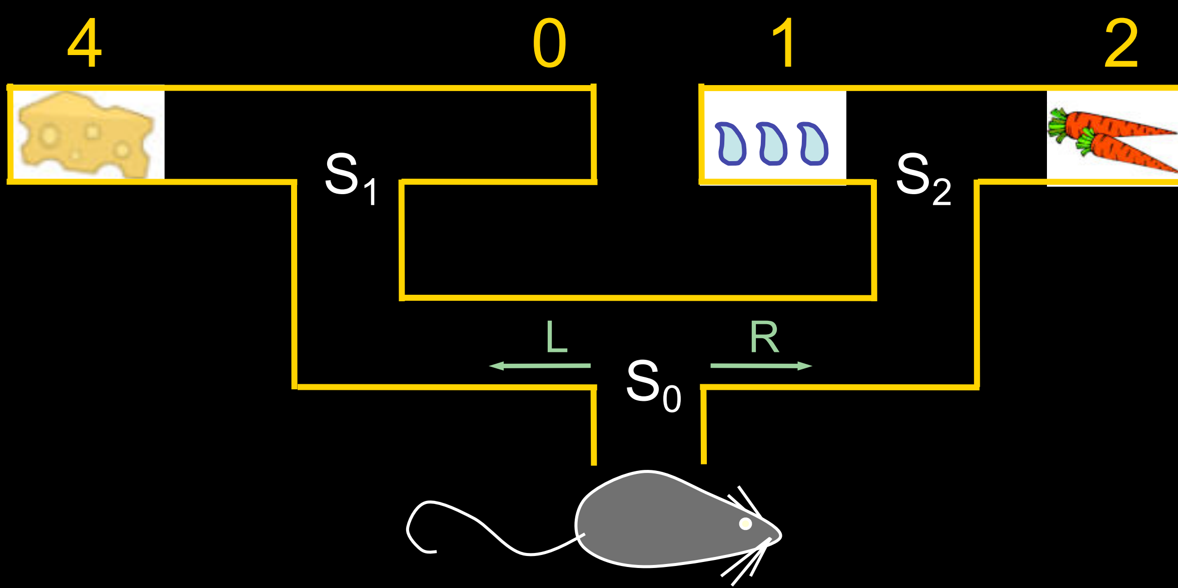
compute Q values by dynamic programming (or other method of look-ahead/planning)

computationally costly, but also flexible (immediately sensitive to change)



Daw, Niv, Dayan (2005)

strategy II: model free RL



- learn values through prediction errors
- choosing actions is easy so behavior is quick, reflexive
- but needs a lot of experience to learn
- and inflexible, need relearning to adapt to any change (habitual)

Stored:

$Q(S_0,L) = 4$
$Q(S_0,R) = 2$

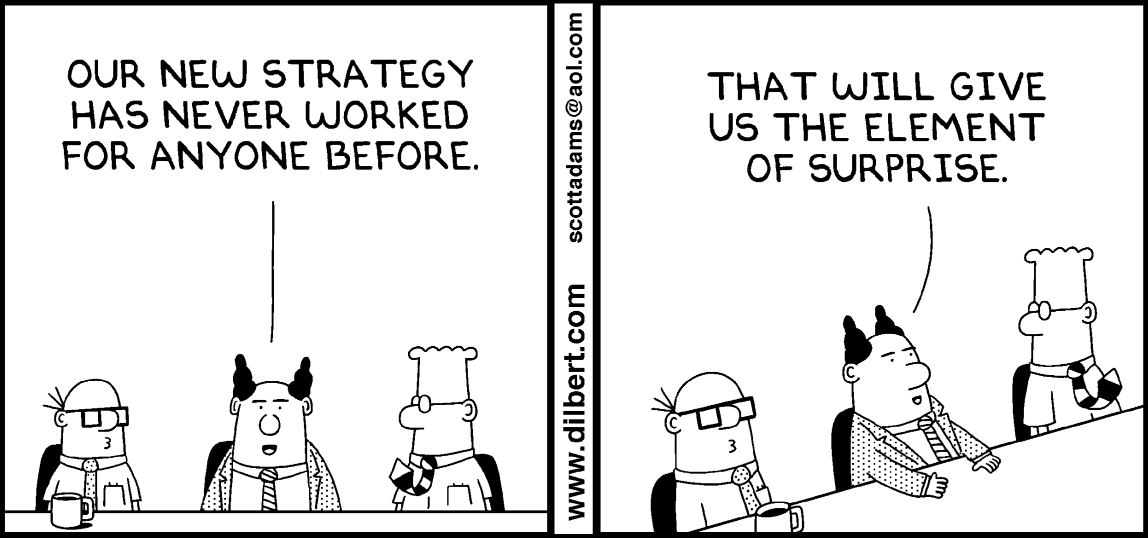
$Q(S_1,L) = 4$
$Q(S_1,R) = 0$

$Q(S_2,L) = 1$
$Q(S_2,R) = 2$

Daw, Niv, Dayan (2005)

this answer raises two questions:

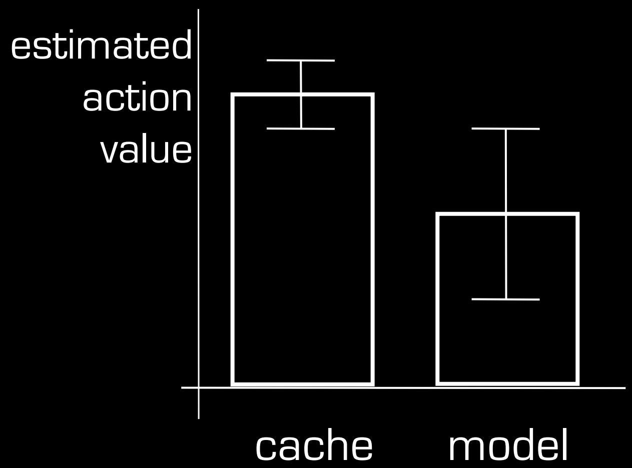
- Why should the brain use two different strategies/ controllers in parallel?
- If it uses two: how can it arbitrate between the two when they disagree (new decision making problem...)



Daw, Niv, Dayan (2005)

answers

1. each system is best in different situations (use each one when it is most suitable/most accurate)
   - goal-directed (forward search) - good with limited training, close to the reward (don't have to search ahead too far)
   - habitual (cache) - good after much experience, distance from reward not so important
2. arbitration: trust the system that is more confident in its recommendation
   - use Bayesian RL (explore/exploit in unknown MDP; POMDP)
   - different sources of uncertainty in the two systems



Daw, Niv, Dayan (2005)

Summary so far...

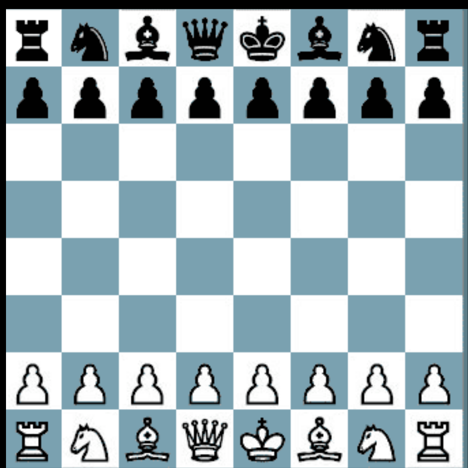
- animal conditioned behavior is not a simple unitary phenomenon: the same response can result from different neural and computational origins
- different neural mechanisms work in parallel to support behavior: cooperation and competition
- RL provides clues as to why this should be so, and what each system does

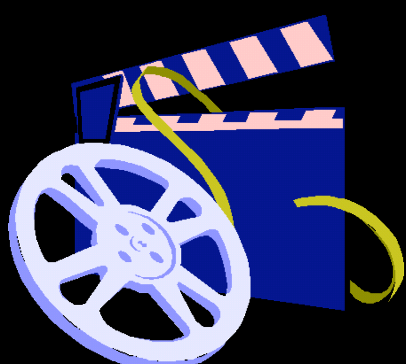
Outline

- The brain coarse-grain
- Learning and decision making in animals and humans: what does RL have to do with it?
- A success story: Dopamine and prediction errors
- Actor/Critic architecture in basal ganglia
- SARSA vs Q-learning: can the brain teach us about ML?
- Model free and model based RL in the brain
- Average reward RL & tonic dopamine
- Risk sensitivity and RL in the brain
- Open challenges and future directions

still a bunch of open questions...

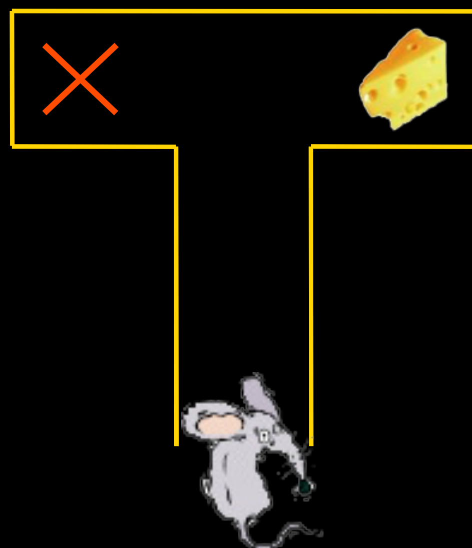
Behavior



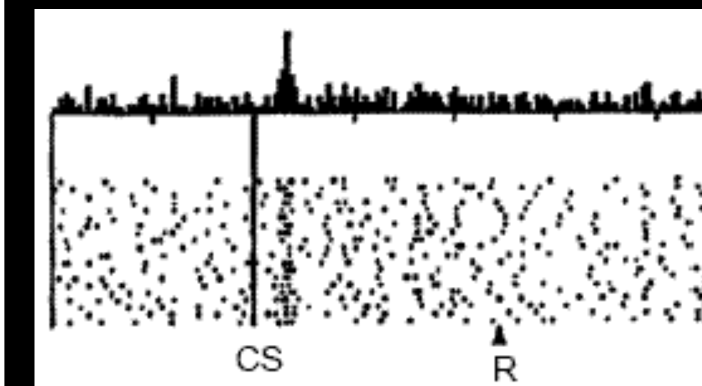


VI60 vs HR

Motivation

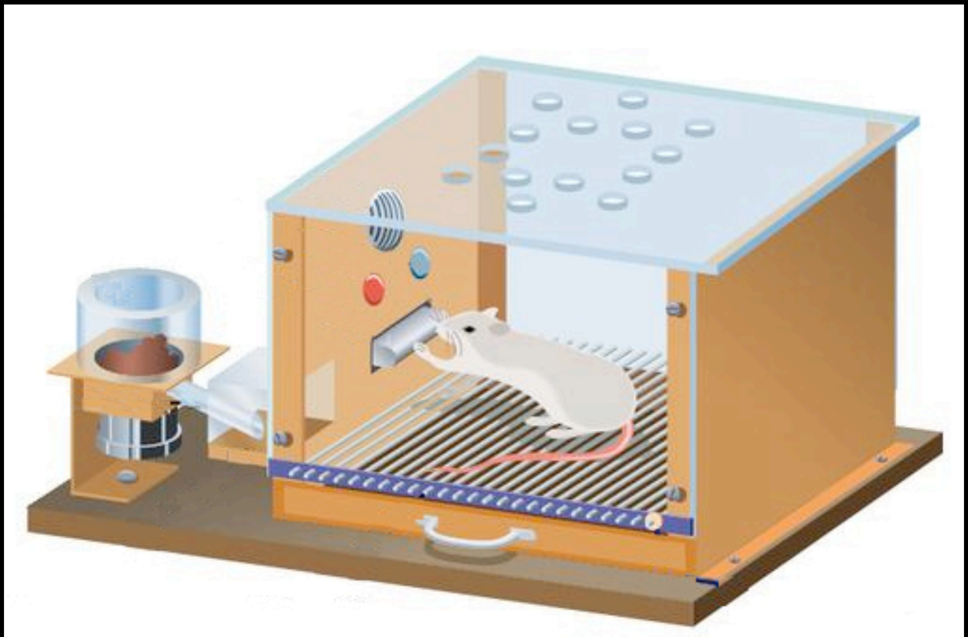


Dopamine



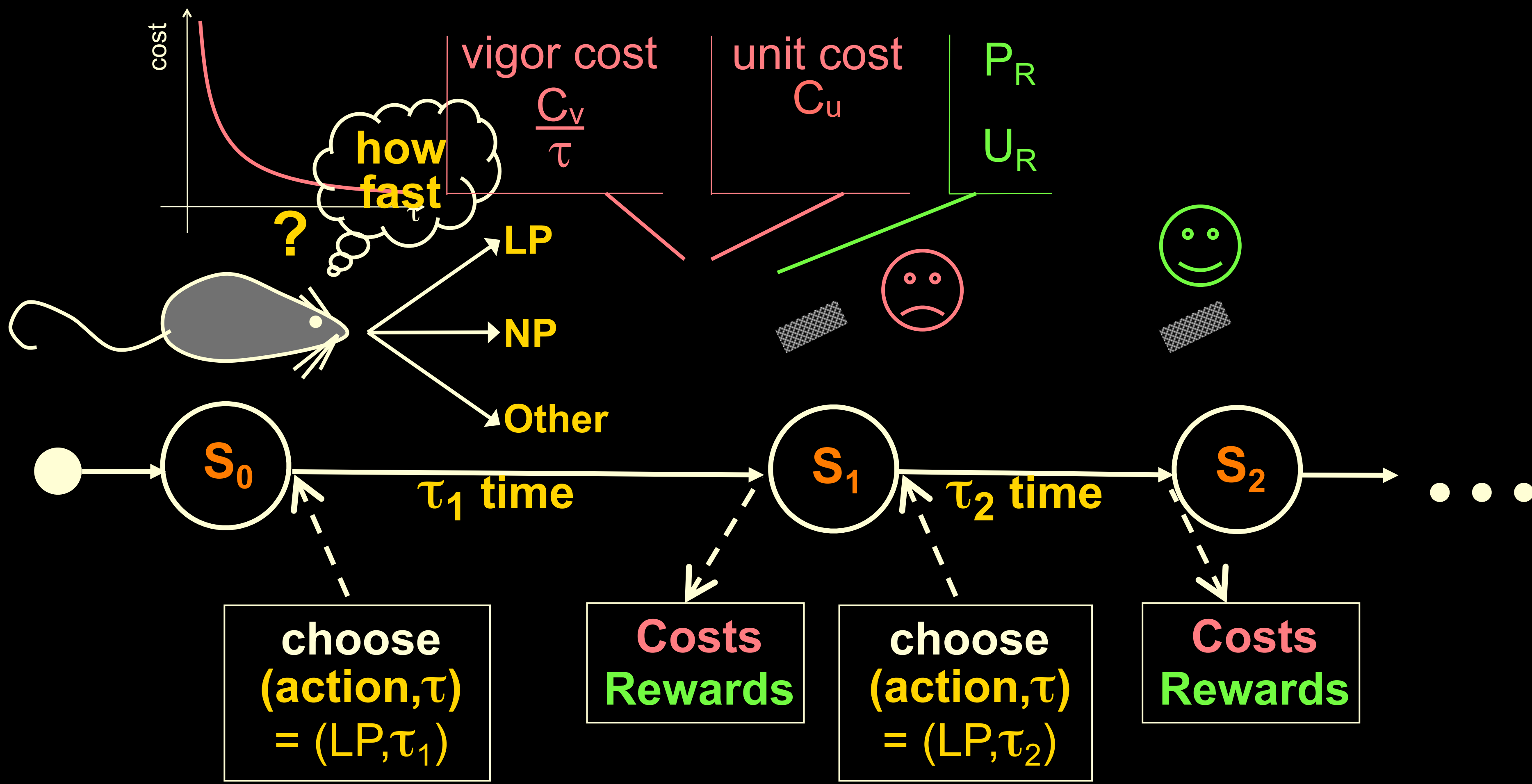
- What did you know about dopamine before today?
- What are the main effects of dopamine?

modeling response rates (vigor) using RL



Niv, Daw, Dayan (2005)

model dynamics

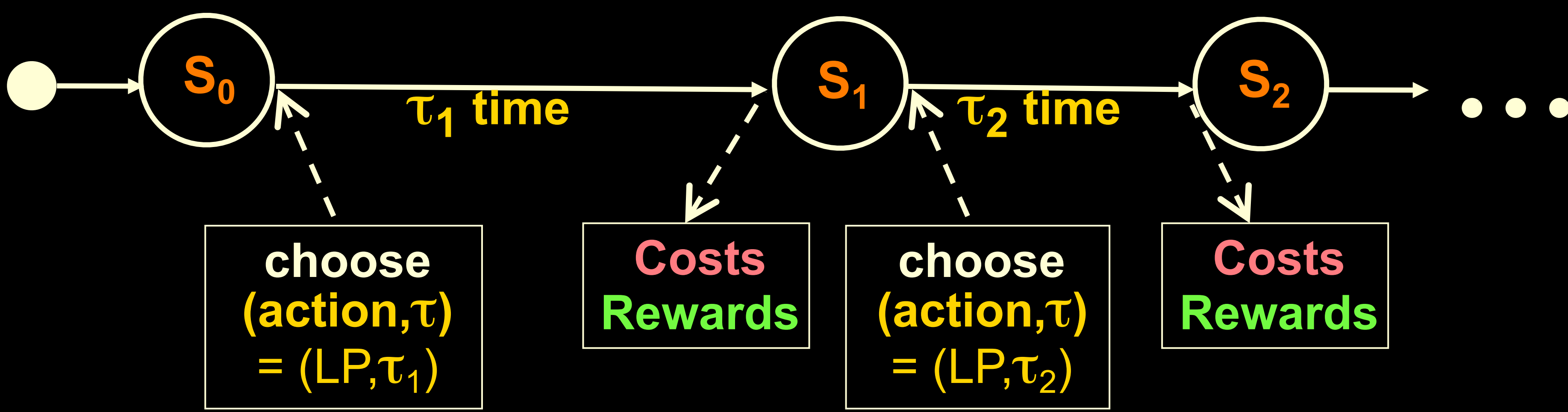


Niv, Daw, Dayan (2005)

model dynamics

<u>Goal</u>: Choose actions and latencies to maximize the average rate of return (rewards minus costs per time)

$$Q(S_t,a,\tau) = (\text{Rewards} - \text{Costs}) + V(S_{t+1}) - \tau\bar{R}$$



Niv, Daw, Dayan (2005)

cost/benefit tradeoffs

Choice of action:

- want to maximize rewards
- and minimize costs

Choice of latency:

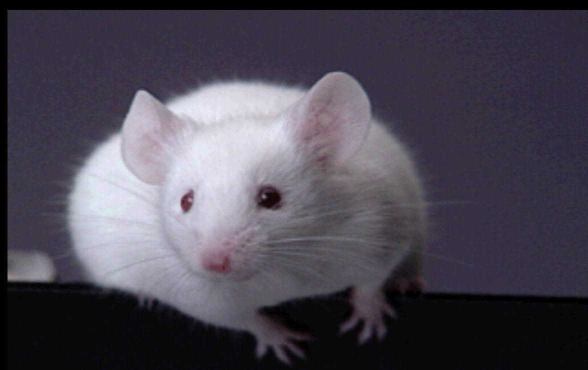
- slow → less costly (vigor cost)
- slow → delays (all) rewards (wastes time)
- what is the cost of time?

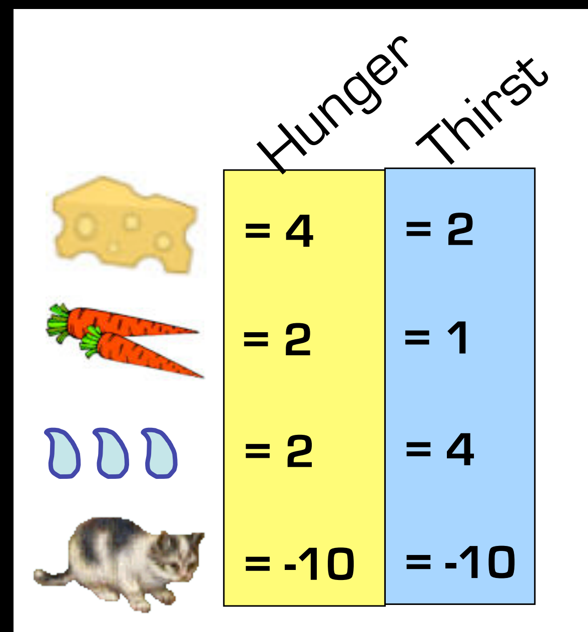
$$\boxed{Q(S_t,a,\tau) = (\text{Rewards} - \text{Costs}) + V(S_{t+1}) - \tau\overline{R}}$$

Niv, Daw, Dayan (2005)

putting motivation in the picture:

Motivation = mapping from outcomes to subjective utilities



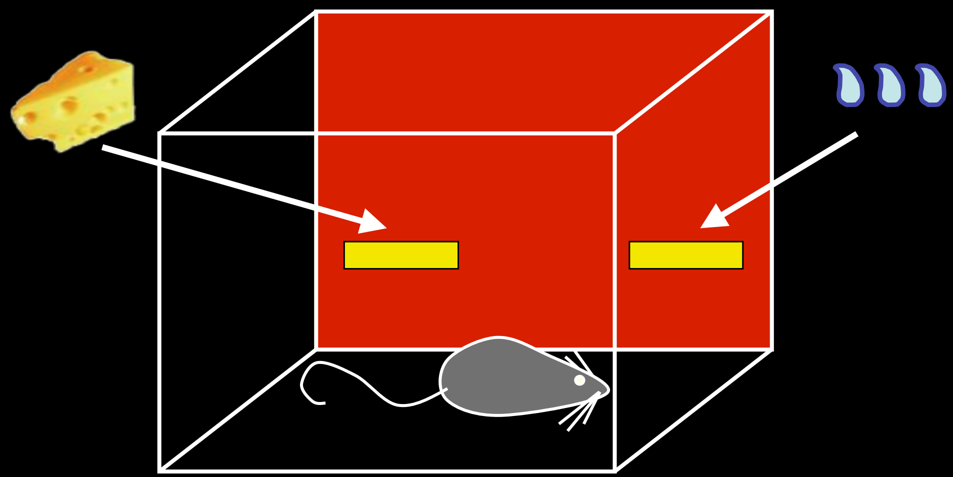


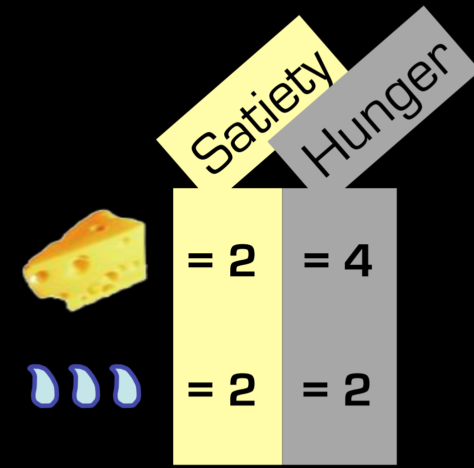
Two traditional effects of motivation in psychology:

1. Directing
2. Energizing (←this is the puzzling one; can RL explain it?)

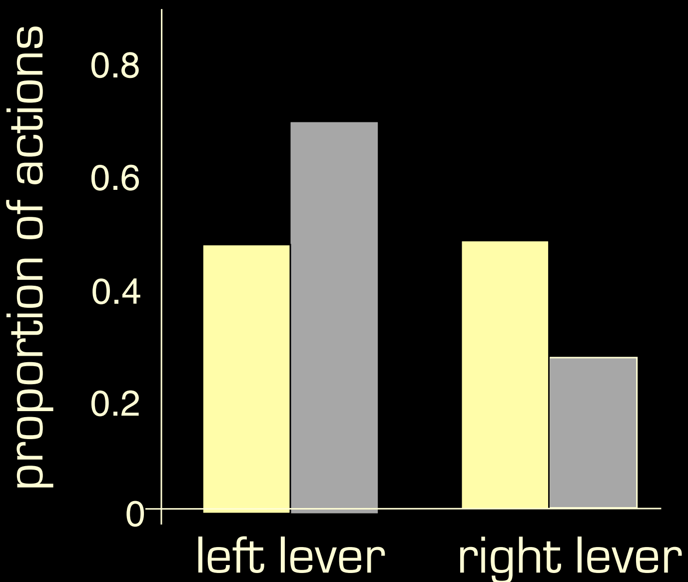
Niv, Joel, Dayan (2006)

two orthogonal effects of motivation in the model

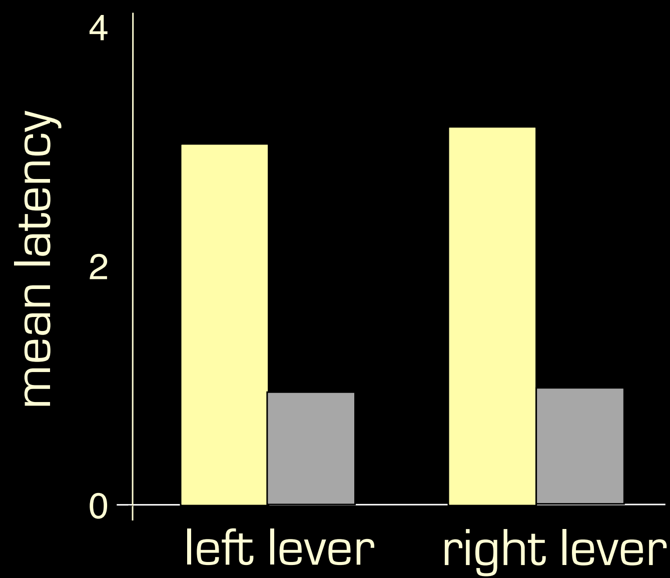




1. Outcome-specific:

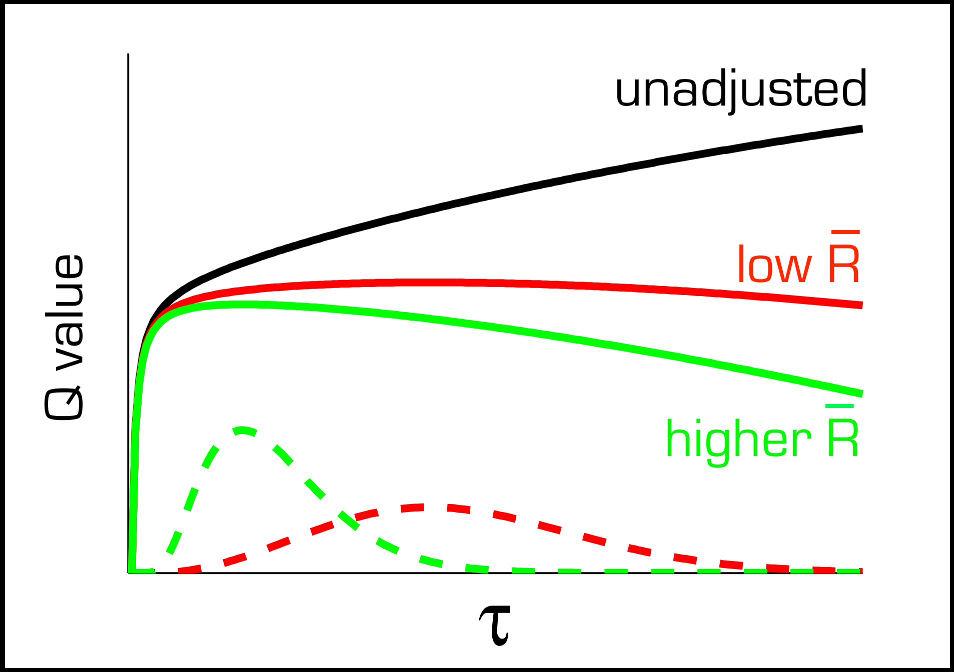


2. Outcome-general:



Niv, Joel, Dayan (2006)

behind the scenes

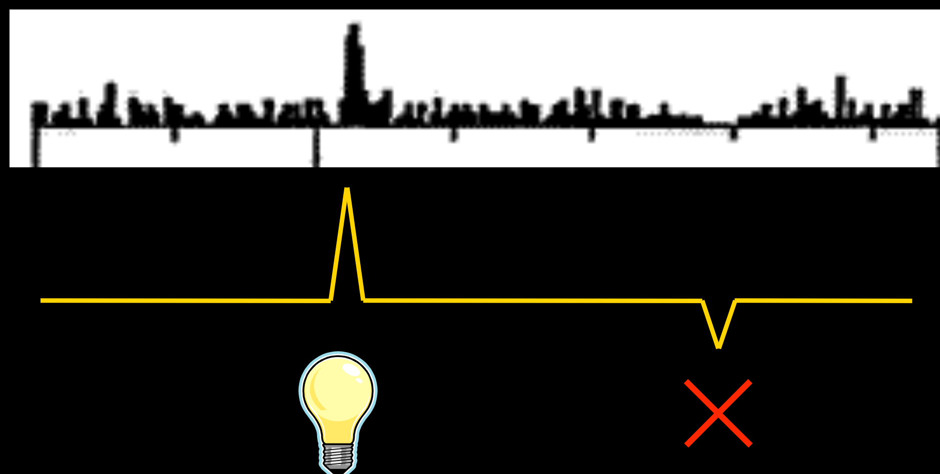


Q(a,τ,S) = Rewards – Costs + Future Returns – Opportunity Cost

- reward rate determines the “cost of sloth”
- higher rate of reward: pressure on all actions to be faster
- Energizing effect (nonspecific “drive”) is an optimal solution!

how does dopamine fit in the picture?

Phasic dopamine firing = reward prediction error



What about tonic dopamine?

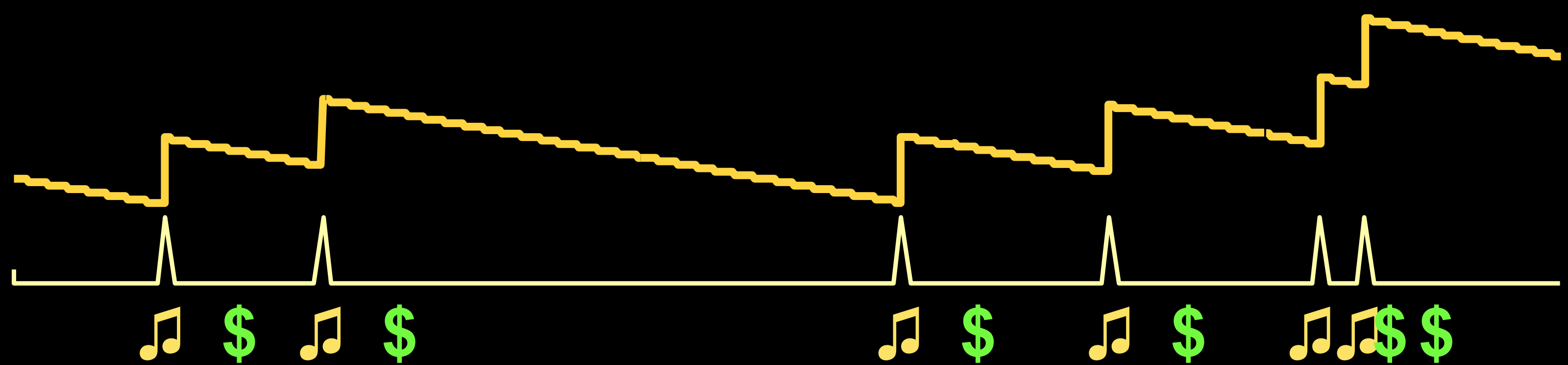
less **more**

ring any bells??

Niv, Daw, Joel, Dayan (2006)

the tonic dopamine hypothesis

tonic level of dopamine = net reward rate



NB. Phasic signal still needed as prediction error for value learning

Niv, Daw, Joel, Dayan (2006)

summary so far...

- In the real world every action we choose comes with a choice of latency
- Adding a notion of vigor or response rate to reinforcement learning models can explain much about the vigor or rate of behavior
- ... and motivation
- ... and dopamine
- suggestion: relation between dopamine and response vigor is due to optimal decision making
- some insight into disorders (Parkinson's etc.)
- insight into cost/benefit tradeoffs in model-free RL

Outline

- The brain coarse-grain
- Learning and decision making in animals and humans: what does RL have to do with it?
- A success story: Dopamine and prediction errors
- Actor/Critic architecture in basal ganglia
- SARSA vs Q-learning: can the brain teach us about ML?
- Model free and model based RL in the brain
- Average reward RL & tonic dopamine
- Risk sensitivity and RL in the brain *NEW*
- Open challenges and future directions

summary so far...

- Although we are used to thinking about expected rewards in RL...
- The brain (and human behavior) seems to fold risk (variance) into predictive values as well
- Why is this a good thing to do?
- Can this help RL applications?

Outline

- The brain coarse-grain
- Learning and decision making in animals and humans: what does RL have to do with it?
- A success story: Dopamine and prediction errors
- Actor/Critic architecture in basal ganglia
- SARSA vs Q-learning: can the brain teach us about ML?
- Model free and model based RL in the brain
- Average reward RL & tonic dopamine
- Risk sensitivity and RL in the brain
- Open challenges and future directions

Neural RL: Open challenges

- How can RL deal with noisy inputs?
- How can RL deal with an unspecified state space?
- How can RL deal with multiple goals? Transfer between tasks?
- ...

- How does the brain deal with noisy inputs? (temporal noise!)
- How does the brain deal with an unspecified state space?
- How does the brain deal with multiple goals? Transfer between tasks?
- ...

Open challenges I: structure learning

- Acquisition of hierarchical structure (parsing of tasks into their components)
- Detection of change: when to unlearn versus when to build a new model
- Learning an appropriate state space for each task

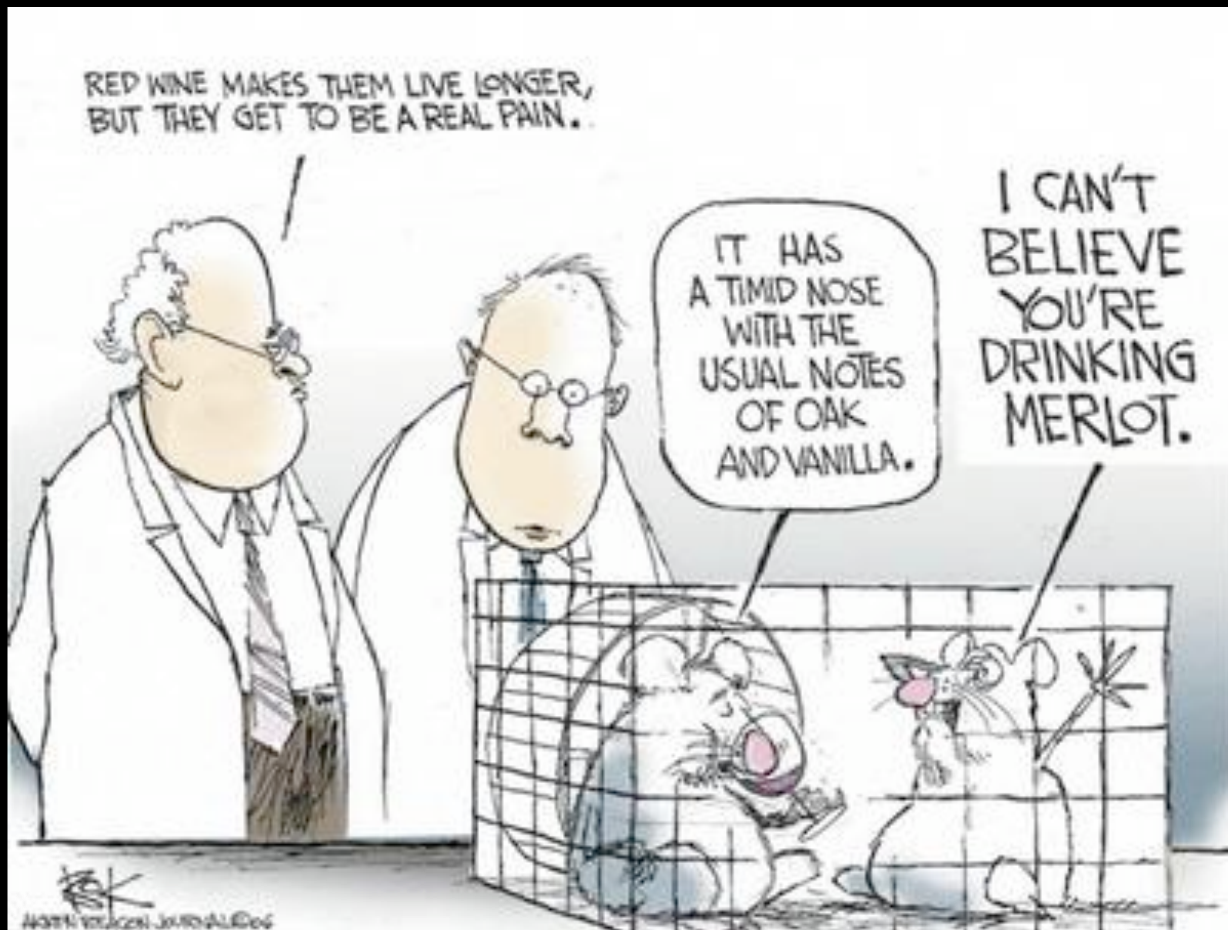
Open challenges II: model-free learning in the brain

- In some cases (eg. conditioned inhibition) dopamine prediction errors differ from simple RL → implications for RL?
- Reward versus punishment: dopamine seems to care only about the former. Why?
- Adaptive scaling of prediction errors in the brain and Kalman filtering(?)
- Diversity of prediction errors in the brain? (more experiments with complex tasks needed)
- Timing noise... (abundant in the brain; detrimental to simple TD learning)

Summary: What have we learned here?

- RL has revolutionized how we think about learning in the brain
- Theoretical, but also practical (even clinical?) implications for neuroscience
- Neuroscience continues to be a “consumer” of ML theory/algorithms
- This does not have to be a one-way street: humans solve some problems so well that it is silly not to use human learning as an inspiration for new RL methods

THANK YOU!


RED WINE MAKES THEM LIVE LONGER, BUT THEY GET TO BE A REAL PAIN.
IT HAS A TIMID NOSE WITH THE USUAL NOTES OF OAK AND VANILLA.
I CAN'T BELIEVE YOU'RE DRINKING MERLOT.

interested in reading more? some recent reviews of neural RL

- Y Niv (2009) - Reinforcement learning in the brain - The Journal of Mathematical Psychology
- P Dayan & Y Niv (2008) - Reinforcement learning and the brain: The Good, The Bad and The Ugly - Current Opinion in Neurobiology, 18(2), 185-196
- MM Botvinick, Y Niv & A Barto (2008) - Hierarchically organized behavior and its neural foundations: A reinforcement learning perspective - Cognition (online prepublication)
- K Doya (2008) - Modulators of decision making - Nature Neuroscience 11,410-416
- MFS Rushworth & TEJ Behrens (2008) - Choice, uncertainty and value in prefrontal and cingulate cortex - Nature Neuroscience 11, 389-397
- A Johnson, MA van der Meer & AD Redish (2007) - Integrating hippocampus and striatum in decision-making - Current Opinion in Neurobiology, 17, 692-697
- JP O'Doherty, A Hampton & H Kim (2007) - Model-based fMRI and its application to reward learning and decision making - Annals of the New York Academy of Science, 1104, 35-53
- ND Daw & K Doya (2006) - The computational neurobiology of learning and reward - Current Opinion in Neurobiology, 6, 199-204