reinforcement learning in humans and other animals

NIPS tutorial 2010

Nathaniel Daw

NYU

collaborators

NYU:	Aaron Bornstein
	Sara Constantino
	Nick Gustafson
	Jian Li
	Seth Madlon-Kay
	Dylan Simon
	Bijan Pesaran
Columbia:	Daphna Shohamy
	Elliott Wimmer
UCL:	Peter Dayan
	Ben Seymour
	Ray Dolan
Berkeley:	Bianca Wittmann
U Chicago:	Jeff Beeler
	Xiaoji Zhuang
Princeton:	Yael Niv
	Sam Gershman
Trinity:	John O'Doherty
Tel Aviv:	Tom Schonberg
	Daphna Joel
Montreal:	Aaron Courville
CMU:	David Touretzky

funding: NIMH, NIDA, NARSAD, McKnight Endowment, HFSP

from nips to neuroscience

reinforcement learning exemplifies two (related) ways that computer science informs behavioral neuroscience

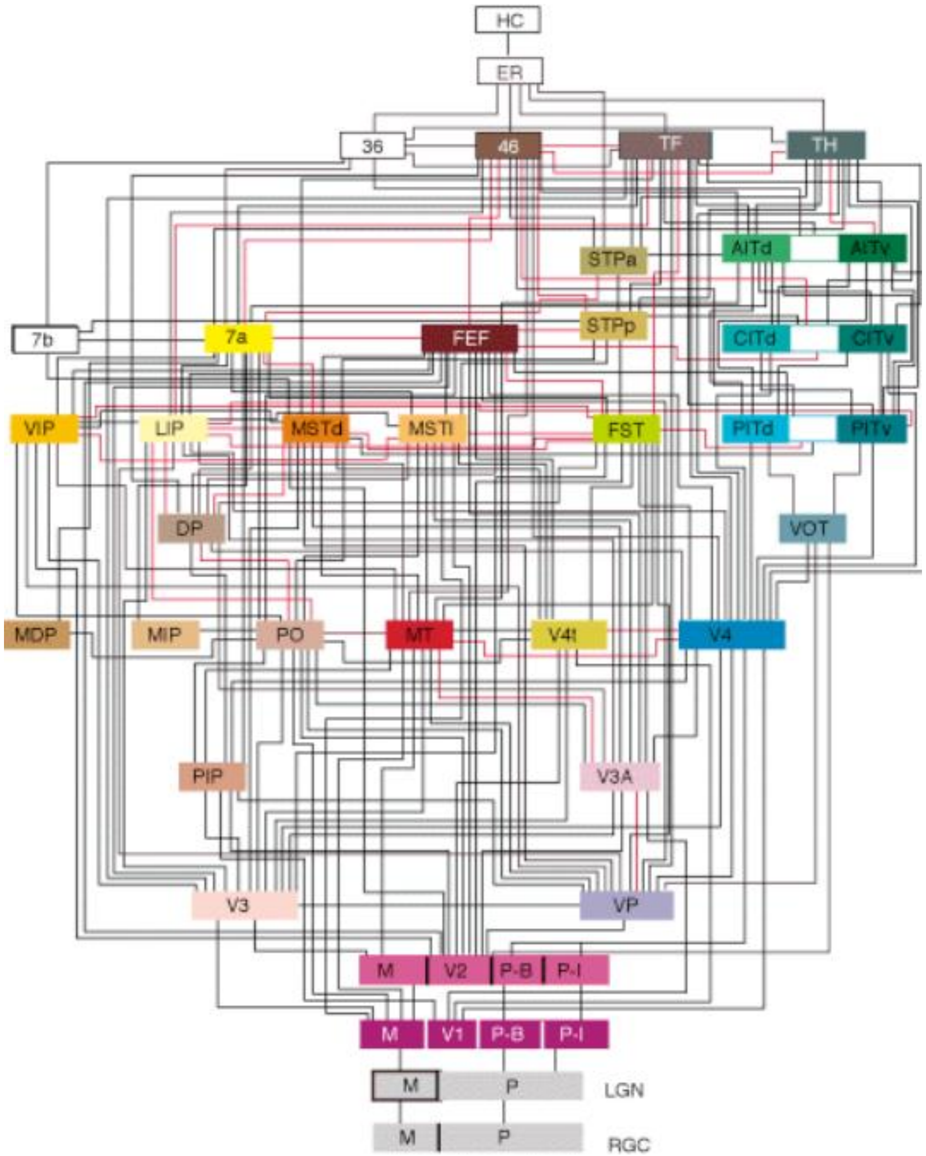
1. conceptual
   - how to characterize hard problems (formally analyzable tasks)
   - optimal (typically intractable) solution
   - approximate algorithms and their properties
   - define relevant quantities

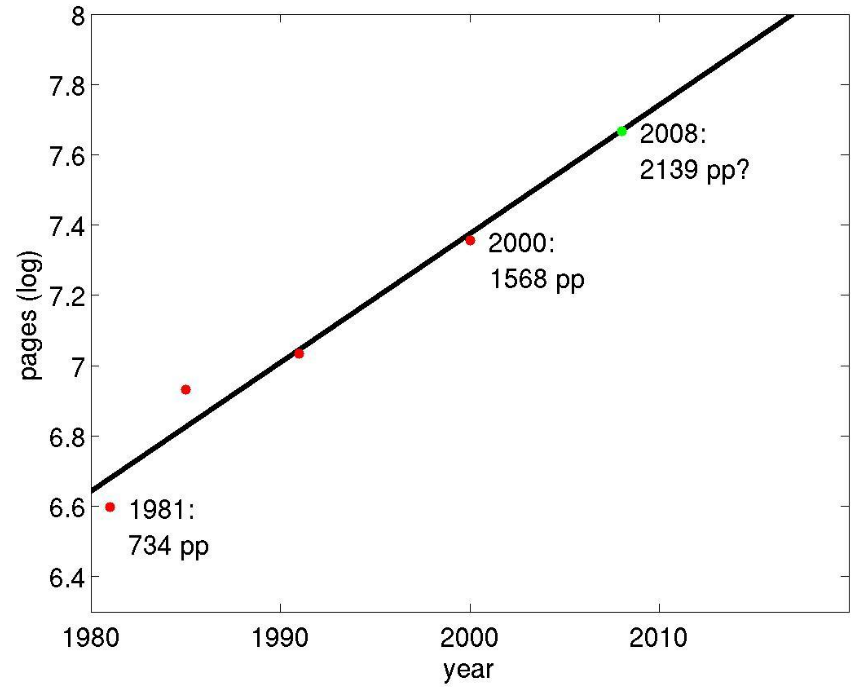
   → algorithms as hypotheses

   → common process level explanation for different kinds of data

2. analytical
   - algorithms as likelihood functions for inference from data
   - data analysis as statistical machine learning

(!= from neuroscience to nips)





plan

reinforcement learning in neuroscience (psychology, behav. economics)

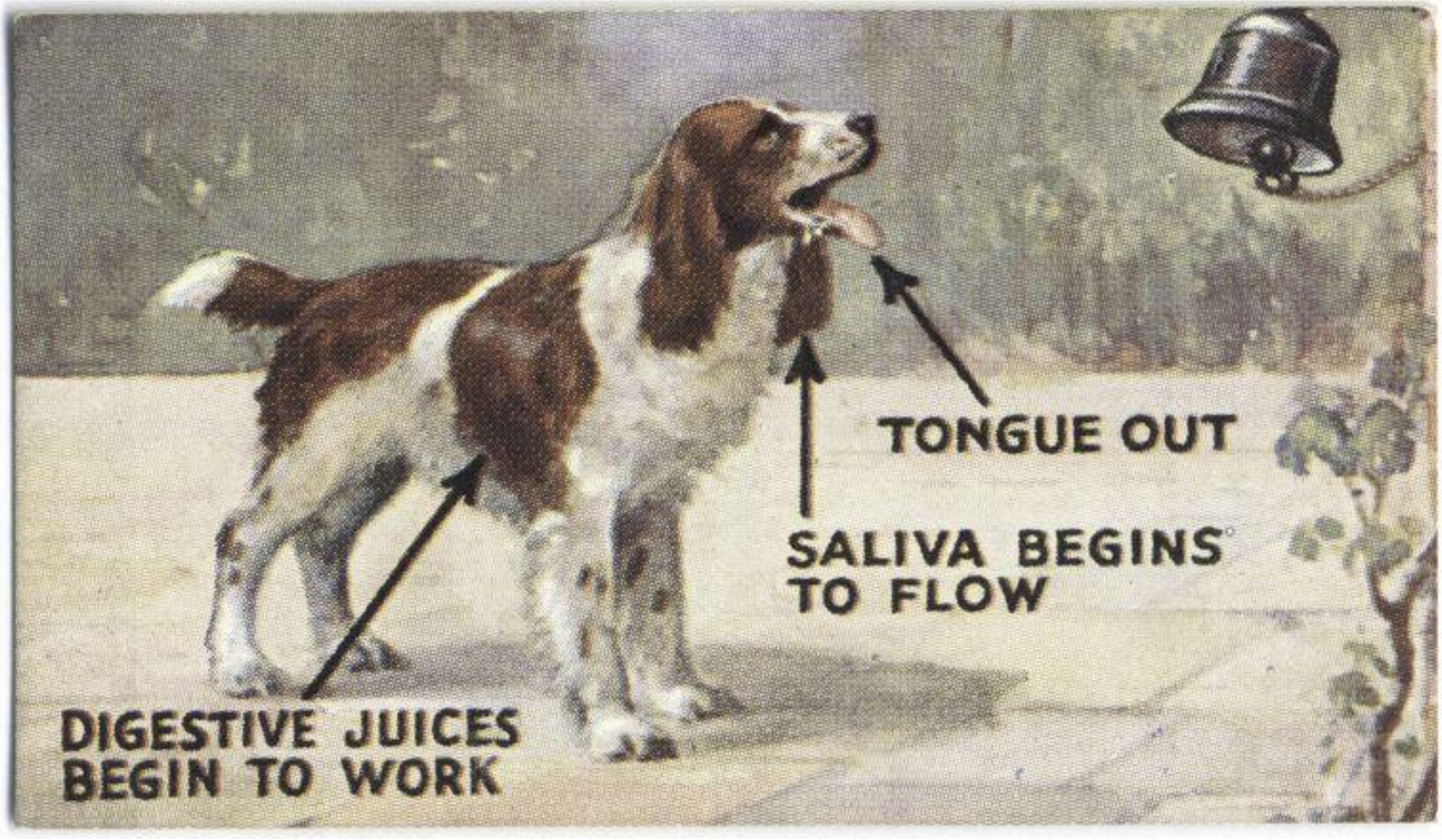
1. dopamine & the TD hypothesis
   - behavioral & analytical background
   - recordings: spiking, fMRI
   - functional neuroanatomy

2. beyond the TD hypothesis
   - states (→ POMDPs & belief states)
   - actions (→ hierarchical RL, decomposed error signals)
   - rewards (→ model-based vs model free)

basic assumption: you know some machine learning.

will try to stay at high level: sloppy notation, etc.

Pavlovian conditioning



prediction

- … revealed by behavior
- … shaped by learning

blocking

Phase 1

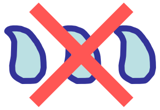
Phase II

interpretation (Rescorla & Wagner 1972):

blocking supports delta-rule ("error driven") learning, e.g.

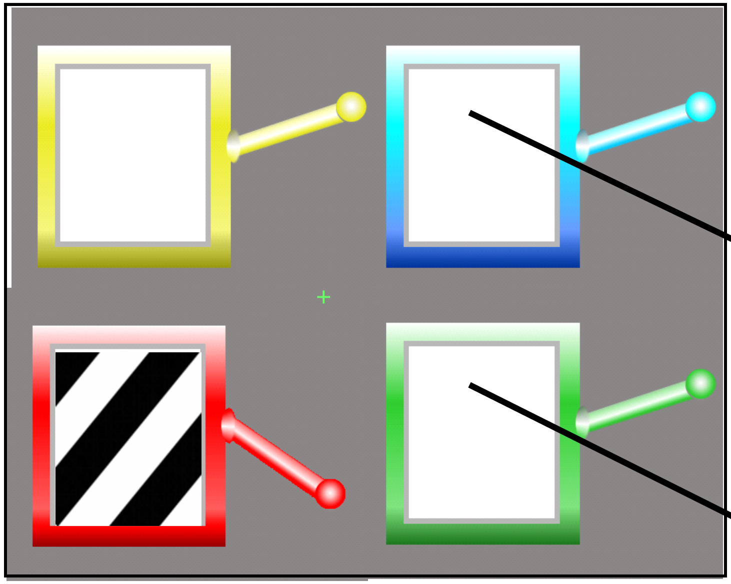
$$V = \Sigma_s w_s$$
$$\delta = r - V$$
$$w_s = w_s + \alpha\delta$$

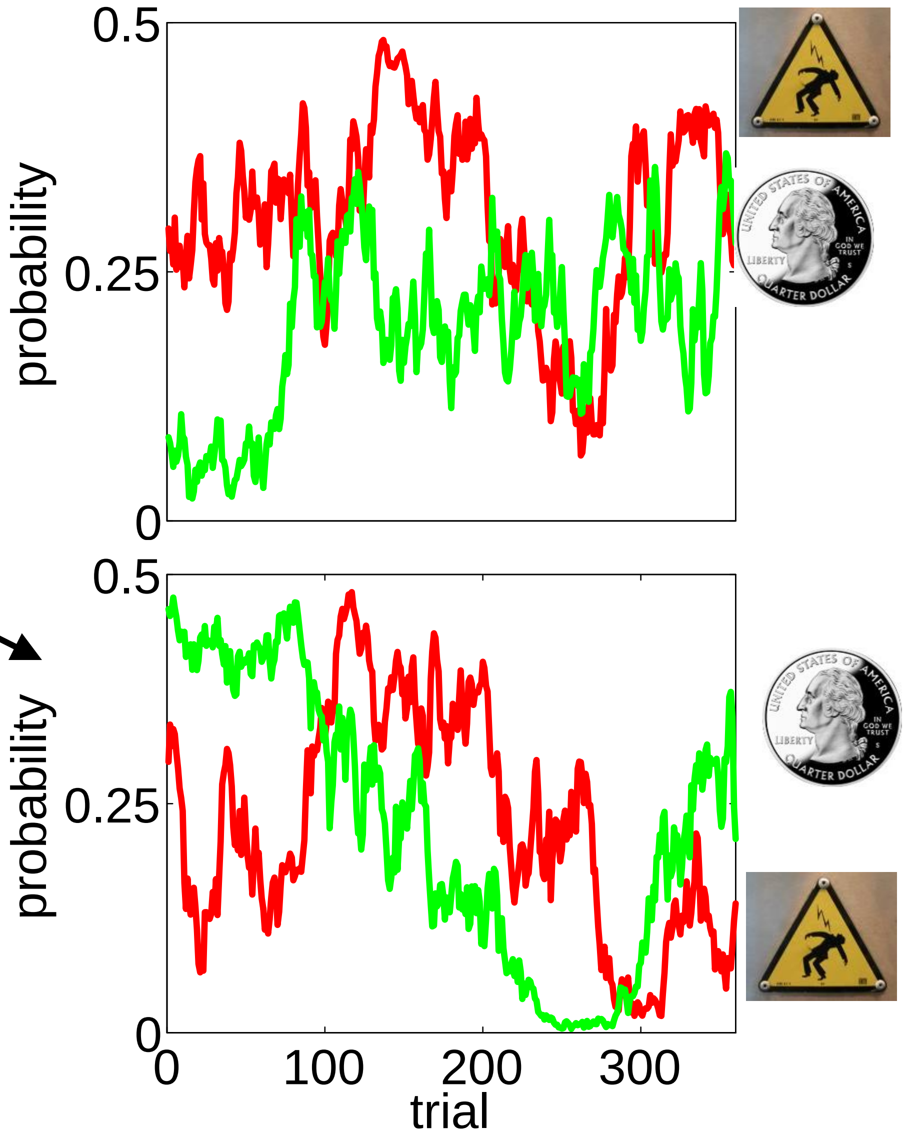
this rule can be motivated from statistical inference in appropriate model (eg Kalman filter; Kakade & Dayan 2000)



(Kamin 1968; Rescorla & Wagner 1972)

bandit tasks for primates

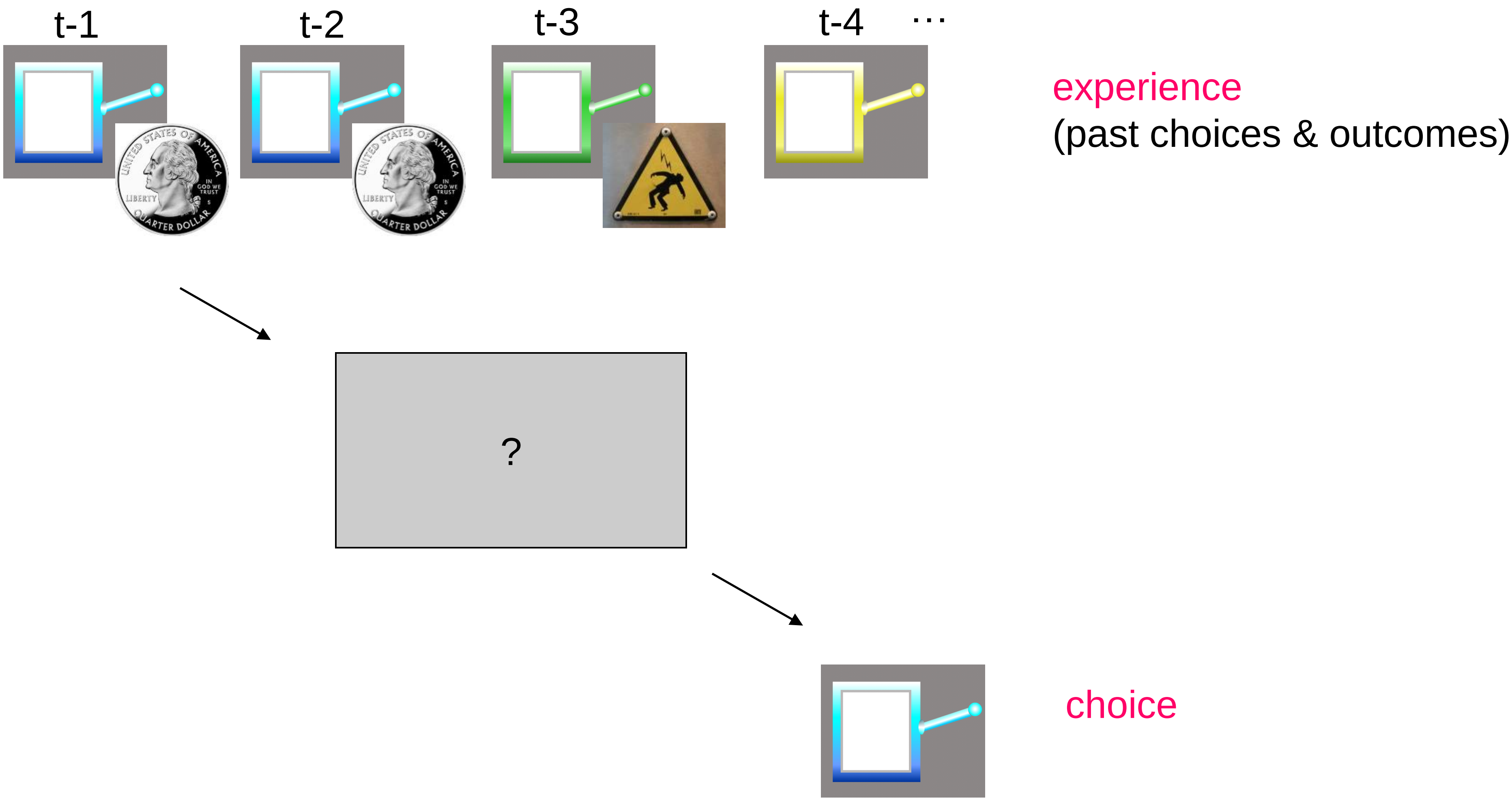




monkeys
Platt & Glimcher 1998
Sugrue et al. 2004
Samejima et al. 2005
Lau & Glimcher 2008

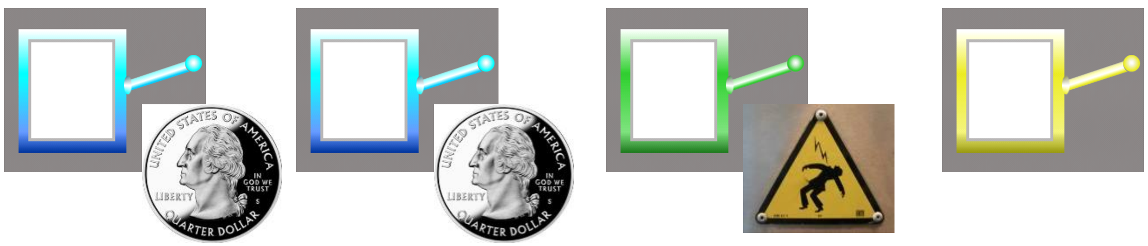
humans
Daw et al. 2006
Wittmann et al 2008
Gershman et al 2009
Schonberg et al 2007, 2010
Glascher et al. 2010

typical analysis



typical analysis

t-1 t-2 t-3 t-4 ···

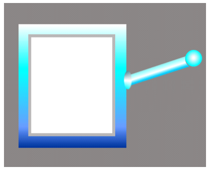


experience
(past choices & outcomes)

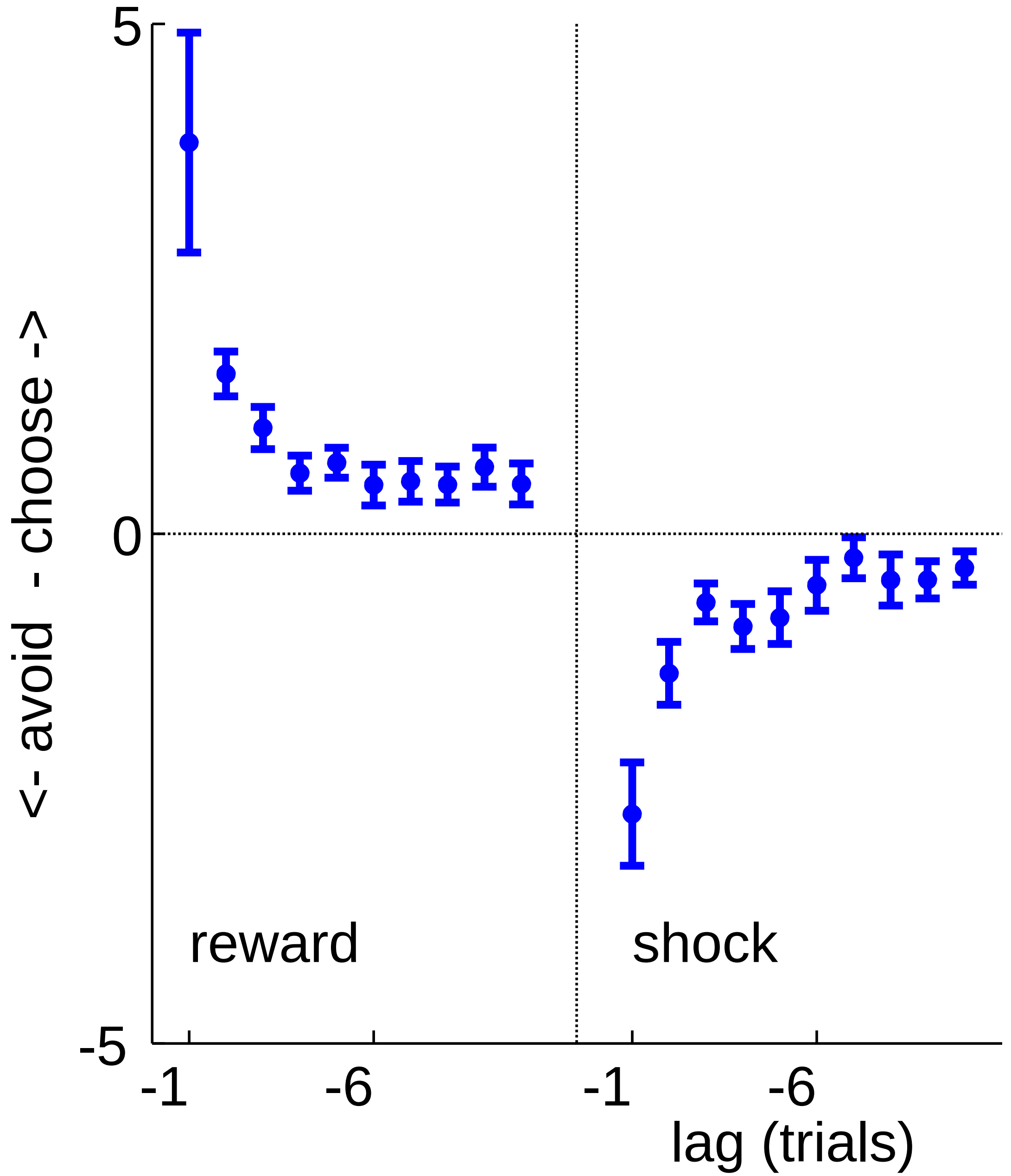
regression

(eg Sugrue et al.;
Lau & Glimcher)

model
(probabilistic algorithm:
experience → choices)



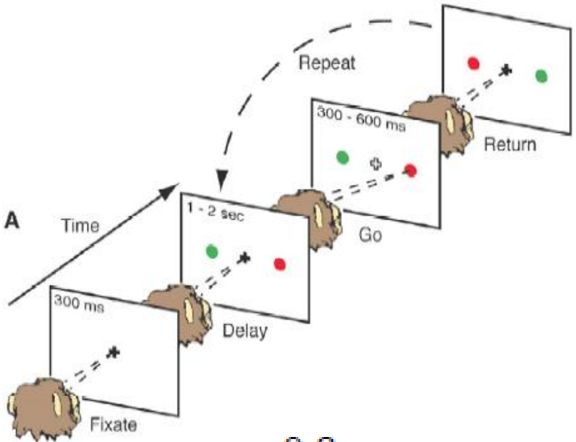
choice

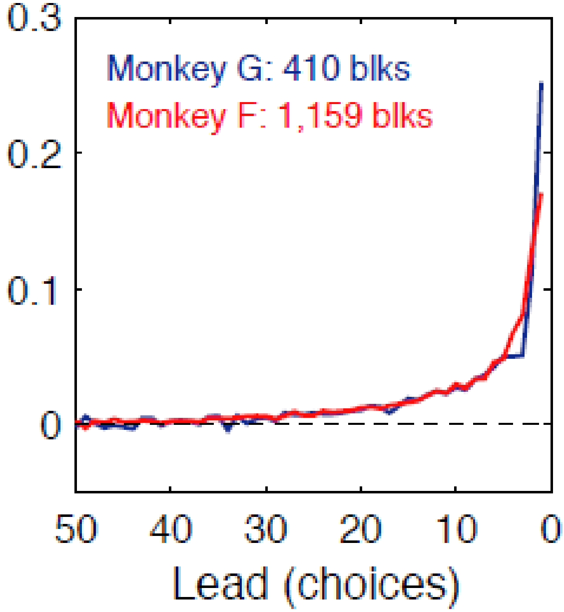


Logit regression, outcomes → choices

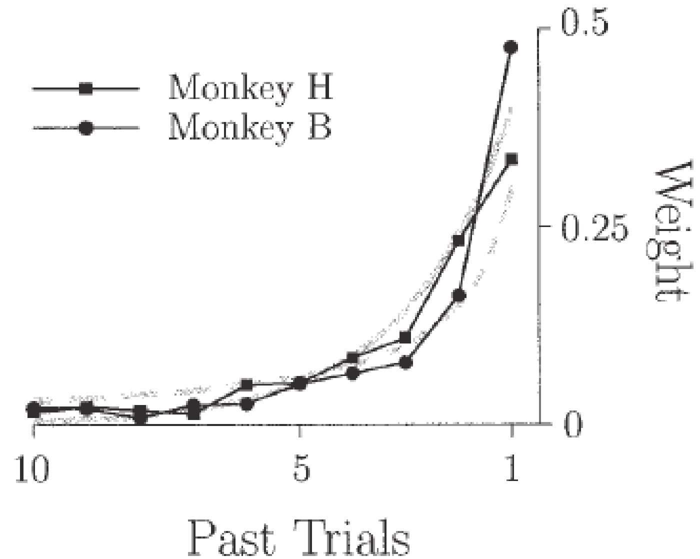
(Seymour et al., under revision)

monkeys





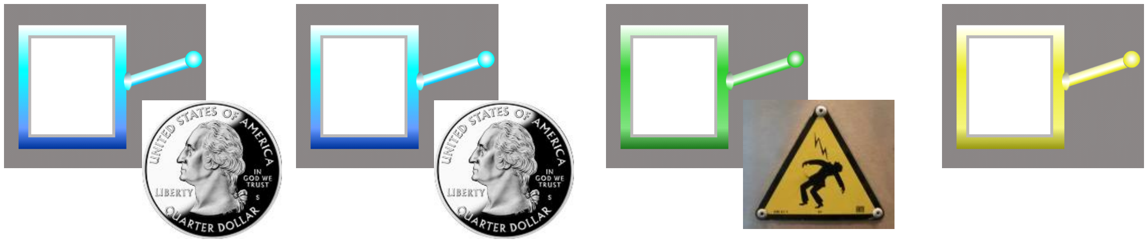
(Sugrue et al. 2004)



(Lau & Glimcher 2005)

typical analysis

t-1 t-2 t-3 t-4 …

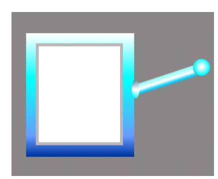


experience
(past choices & outcomes)

regression

(eg Sugrue et al.;
Lau & Glimcher)

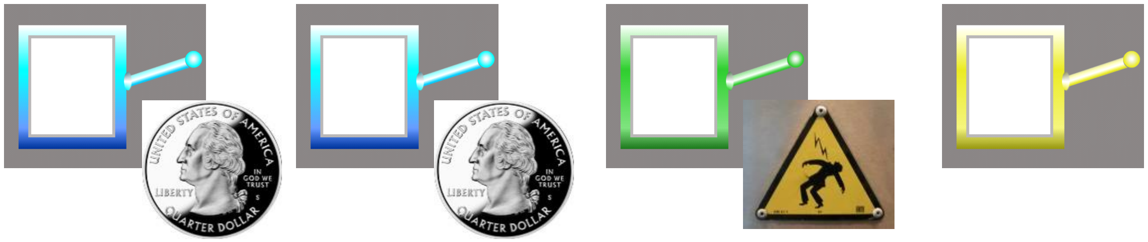
model
(probabilistic algorithm:
experience → choices)



choice

typical analysis

t-1 t-2 t-3 t-4 …



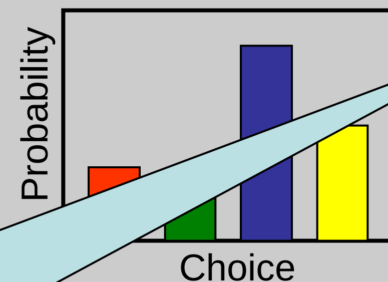
experience
(past choices & outcomes)

delta rule learning

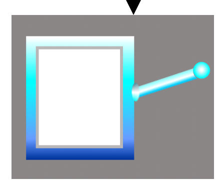
$$\delta_t = r_t - V_t(c_t)$$
$$V_{t+1}(c_t) = V_t(c_t) + \alpha\delta_t$$

logistic choice

$P(c_t{=}c) \propto exp(\beta V_t(c))$

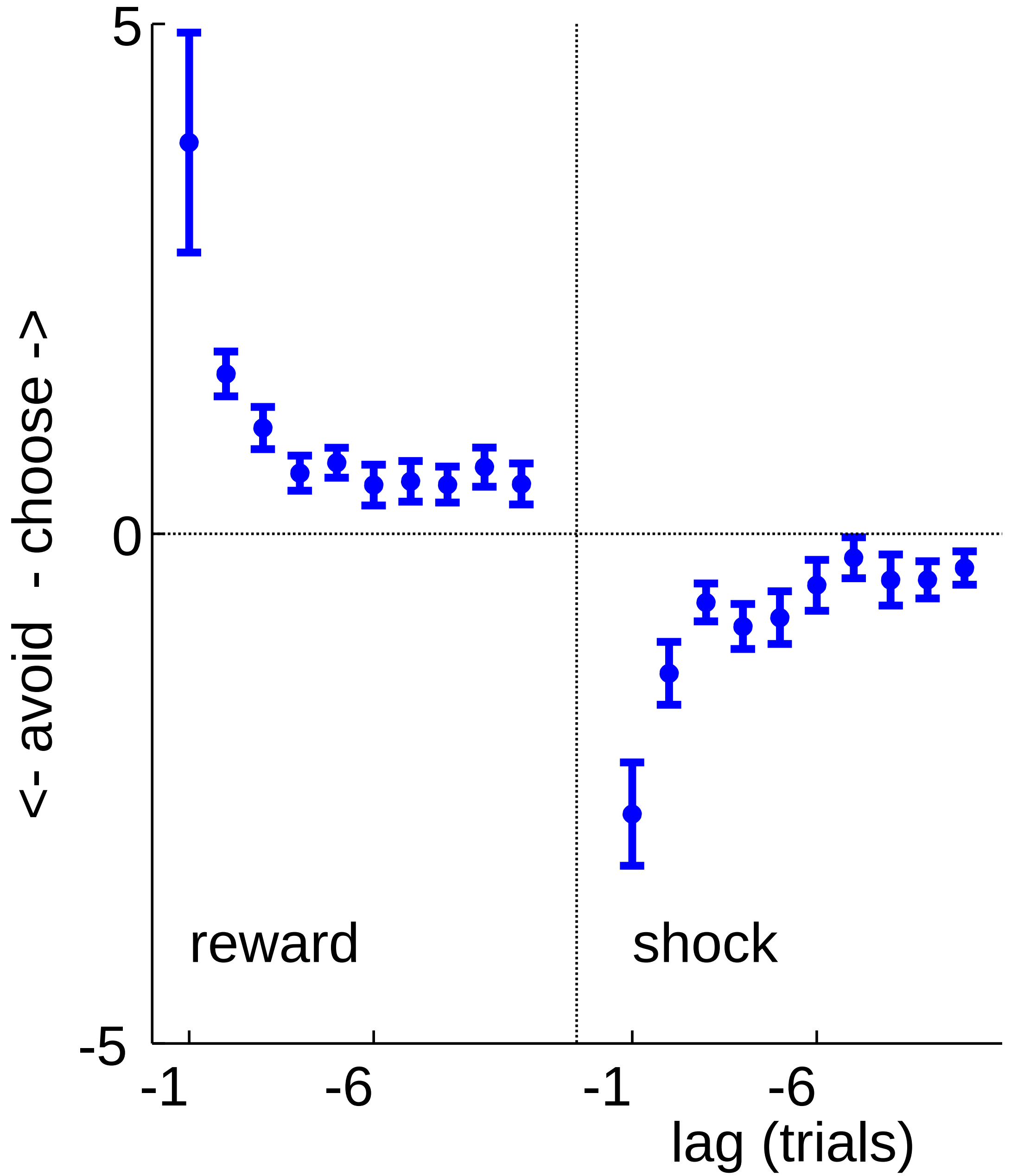


model
(probabilistic algorithm:
experience → choices)



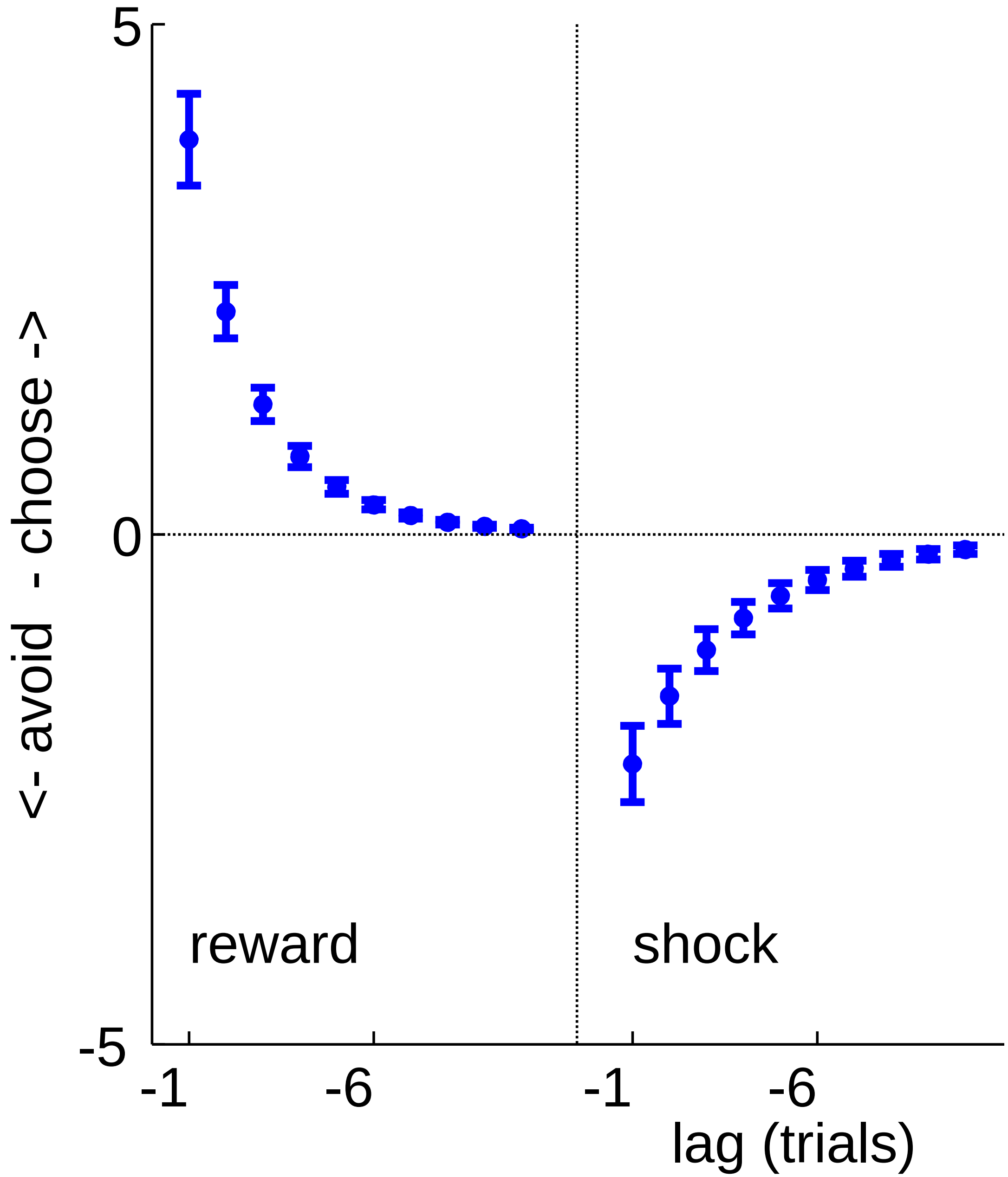
choice

behavior:
Bayesian (or ML) inference: which model & parameters make observed choices most likely?
P(α,β|D,M) ∝ P(D|α,β,M) P(α,β|M)



Logit regression, outcomes → choices

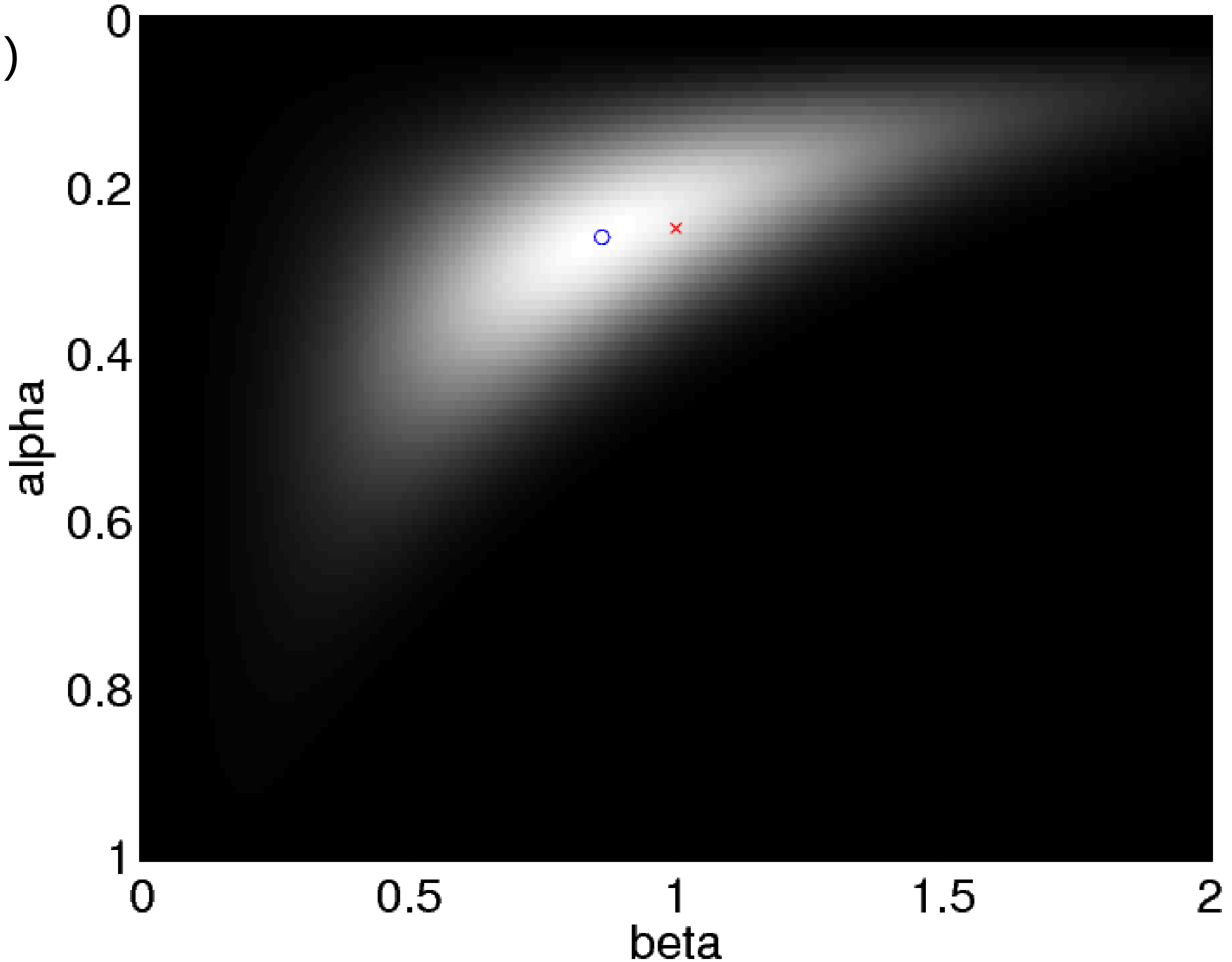
(Seymour et al., under revision)



Fit of delta-rule model, visualized same way
(5x fewer parameters, almost as good fit to data) (Seymour et al. submitted)

model as likelihood

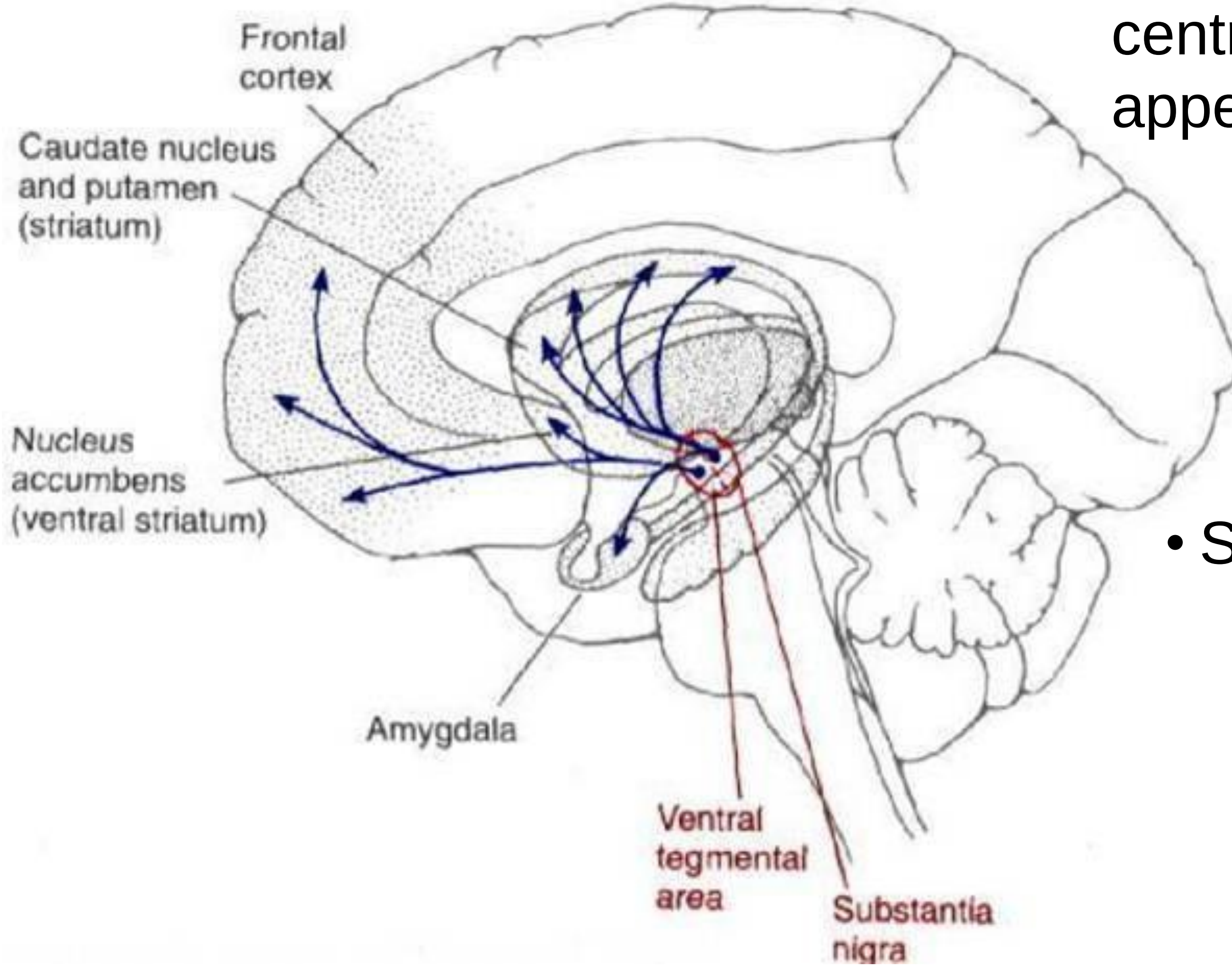
P(D|α,β,M)



nb: RL algorithms as likelihood functions tend to be poorly behaved
(e.g., correlations between parameters a posteriori)

(Daw 2010)

Dopamine

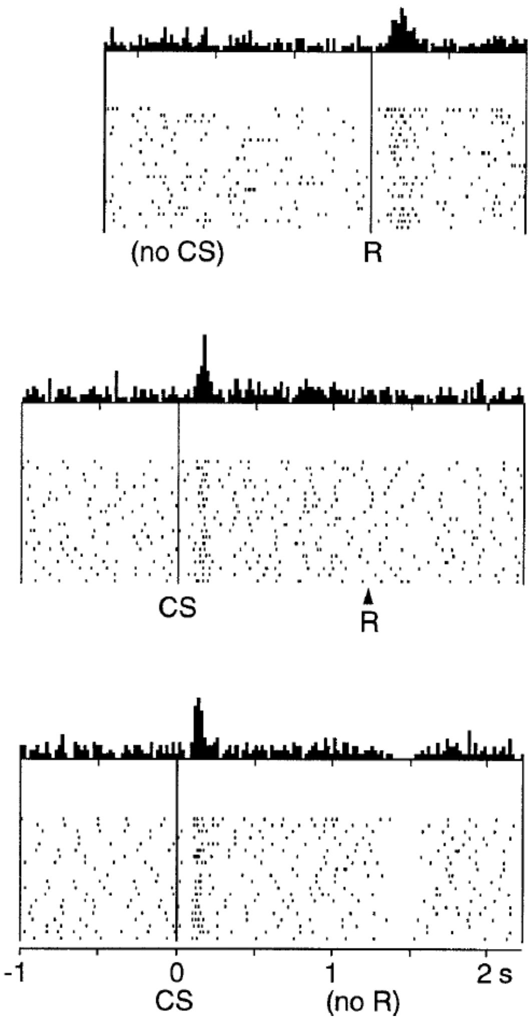


central tension:
appetitive vs motor

- Movement
- Reward
- Addiction
- Self-stimulation
- Synaptic plasticity

(Kandel and Schwartz)

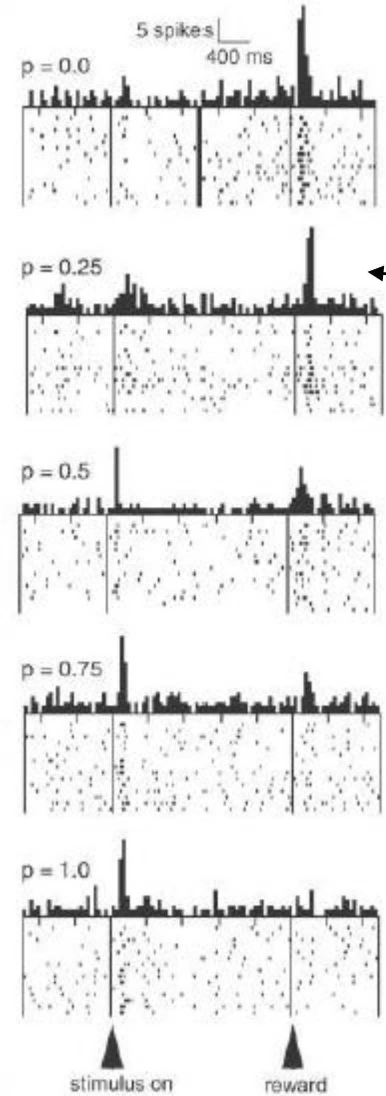
Dopamine responses



Schultz et al. 1997

- Burst to unexpected reward
- Response transfers to reward predictors
- Pause at time of omitted reward

Dopamine responses



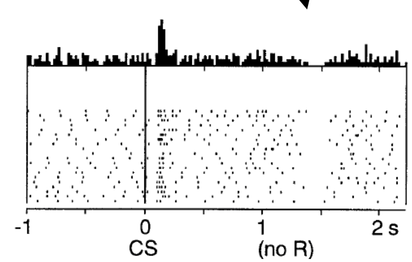
reward following
0% predictive cue

reward following
50% predictive cue

reward following
100% predictive cue

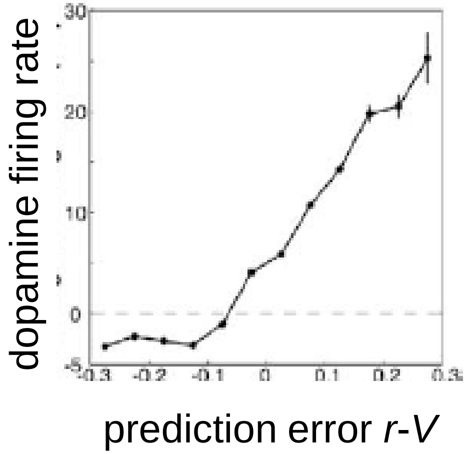
(Fiorillo et al 2003)

Prediction error:
$\delta_t = r_t - V_t$
(Houk et al. 1995;
Montague et al. 1996)

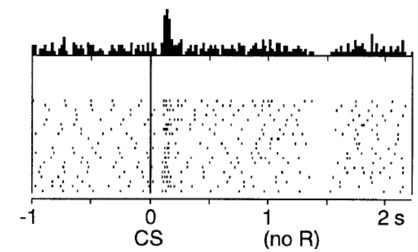


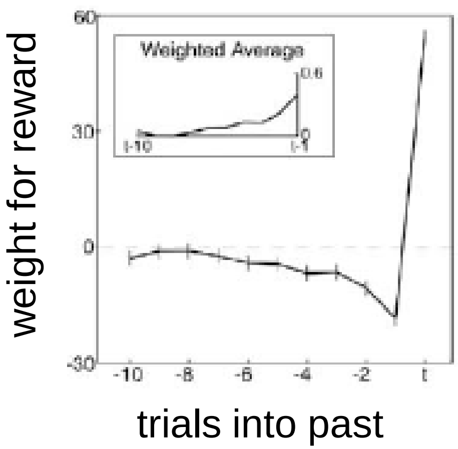
no reward following
100% predictive cue

Prediction error



coding: dopamine response to reward as function of prediction error *r* – (estimated) *V*
→ quite linear; negative error cut off due to low baseline response

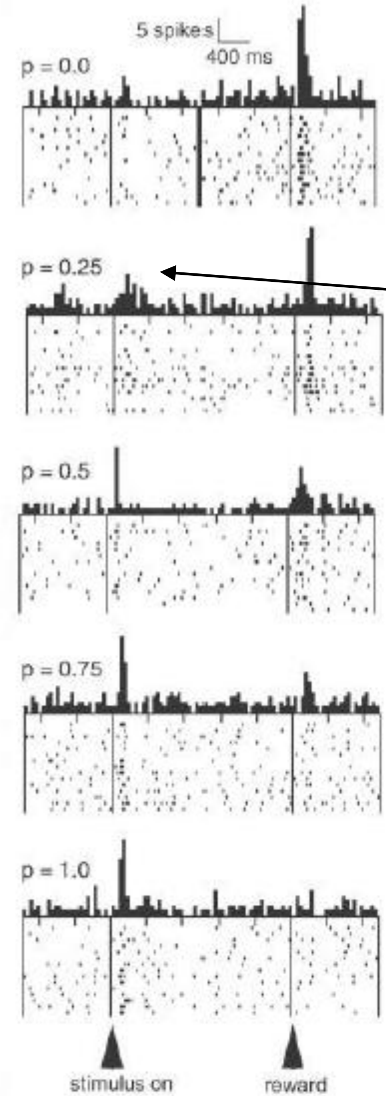




learning: express dopamine response to reward as weighted sum of current & past rewards
→ looks like current *r* minus weighted average of past *r*s (*r* – *V*)

(Bayer & Glimcher 2004)

More dopamine responses

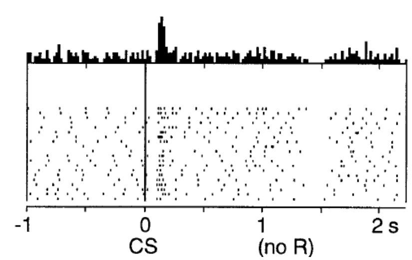


reward following
0% predictive cue

reward following
50% predictive cue

reward following
100% predictive cue

what about these?



no reward following
100% predictive cue

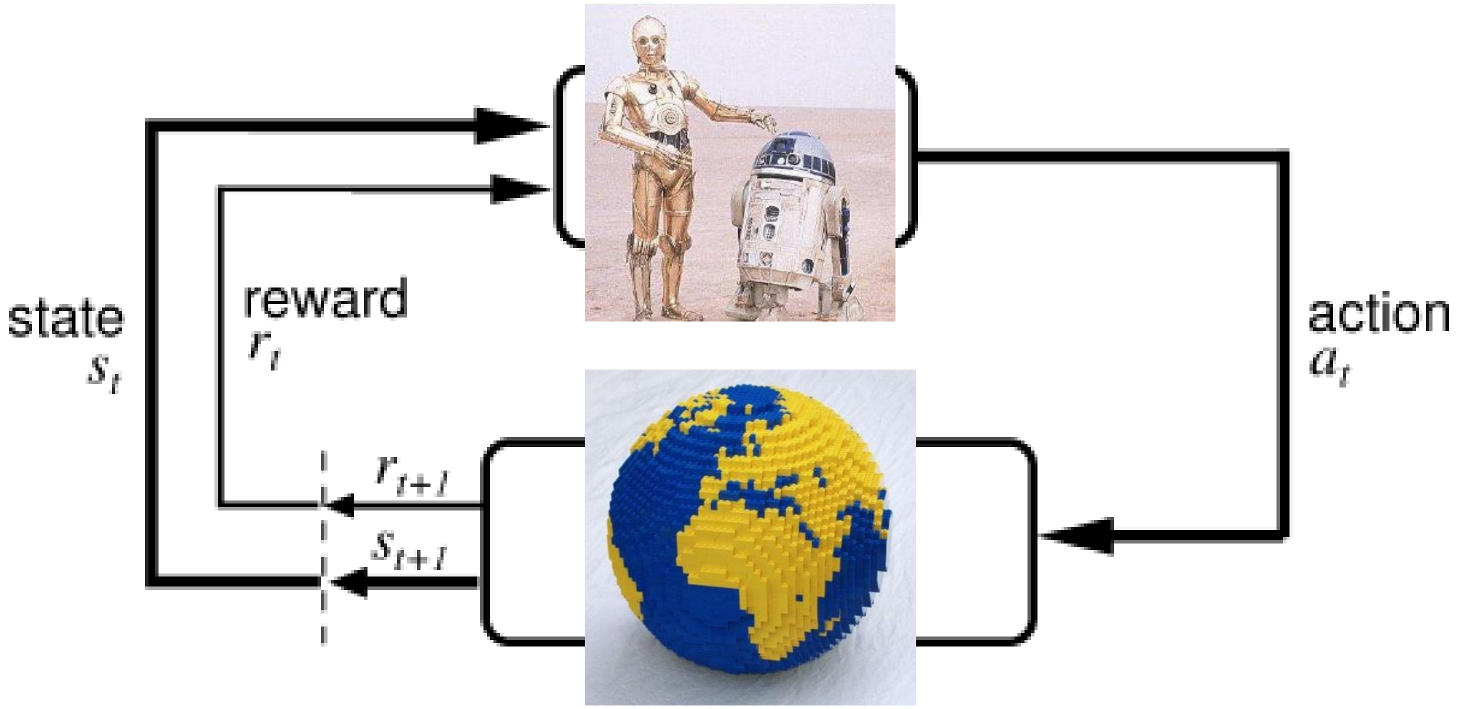
(Fiorillo et al 2003)

Markov Decision Process

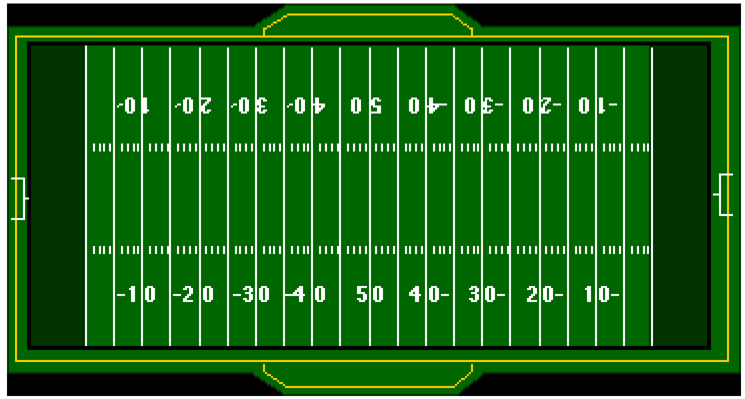
class of stylized tasks with

states, actions & rewards

– at each timestep *t* the world takes on state s_t and delivers reward r_t, and the agent chooses an action a_t



sequential decision problem



total score is not just immediate points scored on play

V(state) = E[immediate reward + next reward + next next reward + …]

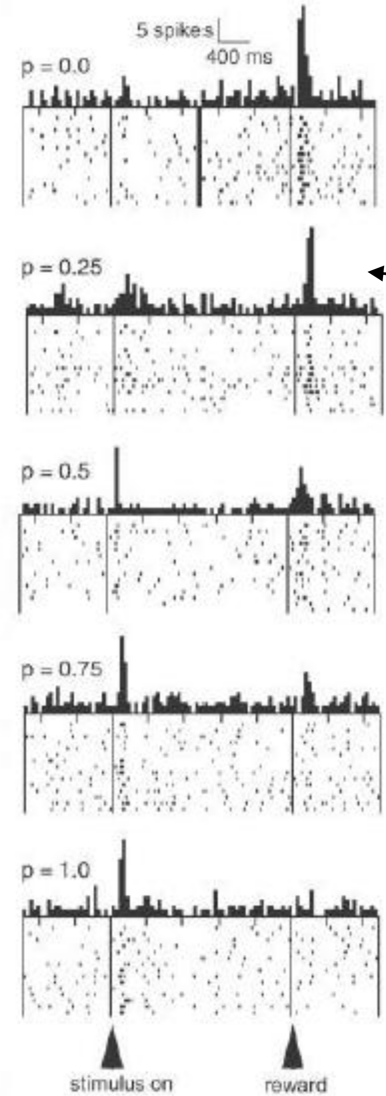
V(next state) = E[next reward + next next reward + …]

V(state) = E[immediate reward + V(next state)] *(Bellman equation)*

→ temporal difference methods (Sutton 1992) based on sampling Bellman residual:

δ(state) = [immediate reward + V(next state)] - V(state)

More dopamine responses



reward following
0% predictive cue

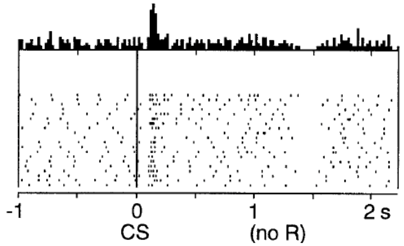
reward following
50% predictive cue

reward following
100% predictive cue

Prediction error:

$V_{t+1} = 0$

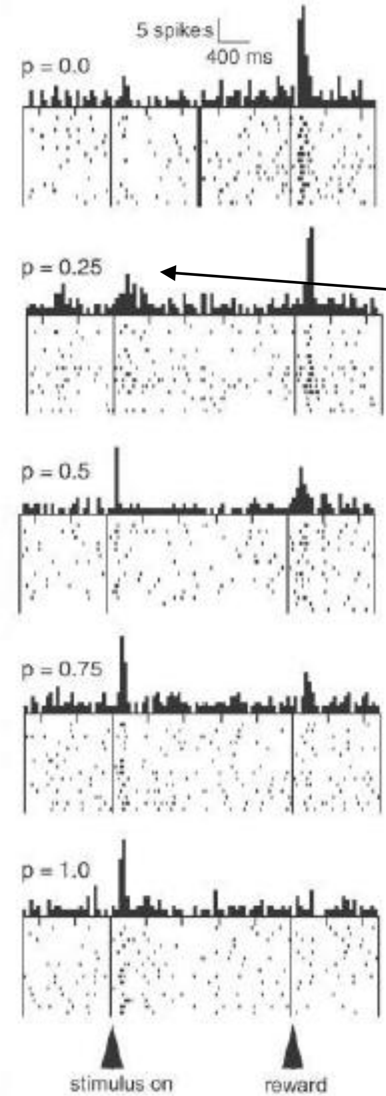
$\delta_t = r_t - V_t + V_{t+1}$



no reward following
100% predictive cue

(Fiorillo et al 2003)

More dopamine responses



reward following
0% predictive cue

reward following
50% predictive cue

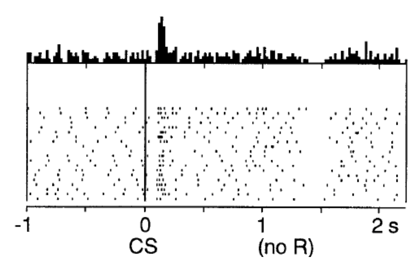
reward following
100% predictive cue

(Fiorillo et al 2003)

Same story here

$V_t = 0; r_t = 0$

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

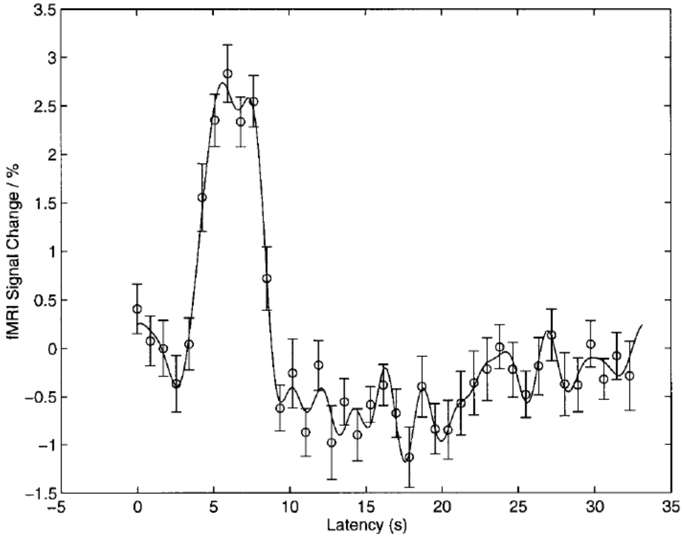


no reward following
100% predictive cue

aside: fMRI

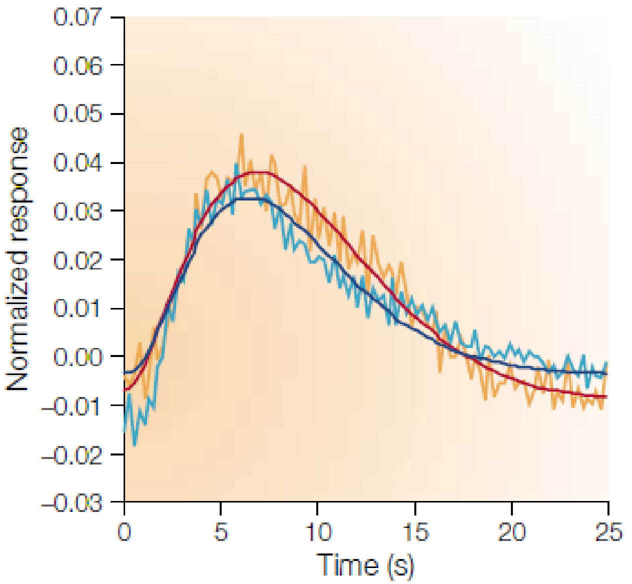
- Functional magnetic resonance imaging
  - "functional": measuring brain usage, not structure
  - useful technology for studying neural function in humans
- Concept: measure BOLD ("blood oxygenation level dependent") signal
  - oxygenated vs de-oxygenated hemoglobin have different magnetic properties
  - detected by big superconducting magnet
- Brain is functionally modular
- Synaptic activity uses energy
  - & oxygen
  - (activity apparently reflects input more than local firing?)
- Spatial resolution: ~3mm "voxels"
- temporal resolution: maybe 5-10 secs

hemodynamic impulse response



(Josephs et al 1997)

single words,
auditory cortex



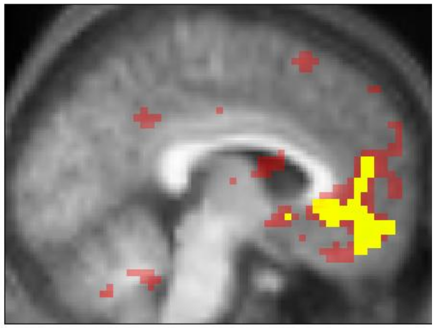
(Logothetis et al 2001)

checkerboard,
visual cortex

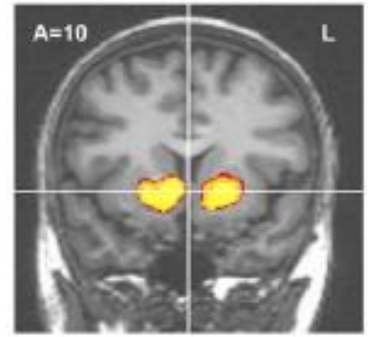
- Slow
- Localized
- Event-related
- Negative & positive portions

Broad findings

Reward or reward anticipation activates ventromedial prefrontal cortex & orbitofrontal cortex, striatum (sometimes midbrain)

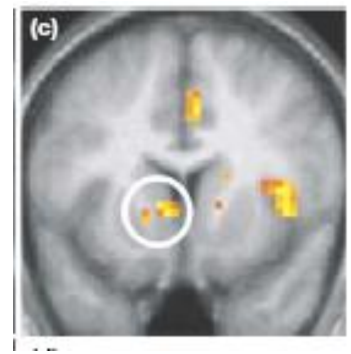


money
value predicted
(Daw et al 2006)



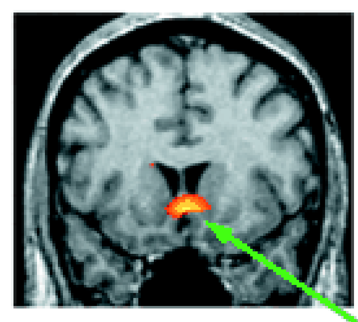
money
gain vs loss
(Kuhnen & Knutson 2005)

faces
attractiveness
(O'Doherty et al 2003)



food odors
valued vs devalued
(Gottfreid et al 2003)

Coke or Pepsi
degree favored
(McClure et al. 2004)

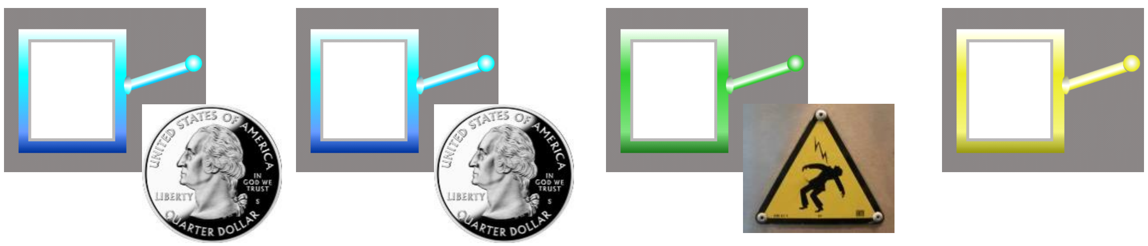


juice
unpredictable vs predictable
(Berns et al 2001)

→ commonality of responding across reinforcers suggests generalized appetitive function

typical analysis

t-1 t-2 t-3 t-4 …



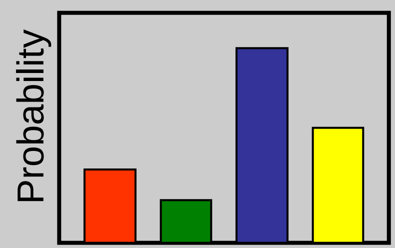
experience
(past choices & outcomes)

delta rule learning

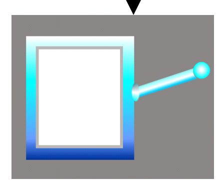
$$\delta_t = r_t - V_t(c_t)$$
$$V_{t+1}(c_t) = V_t(c_t) + \alpha\delta_t$$

logistic choice

$P(c_t{=}c) \propto exp(\beta V_t(c))$



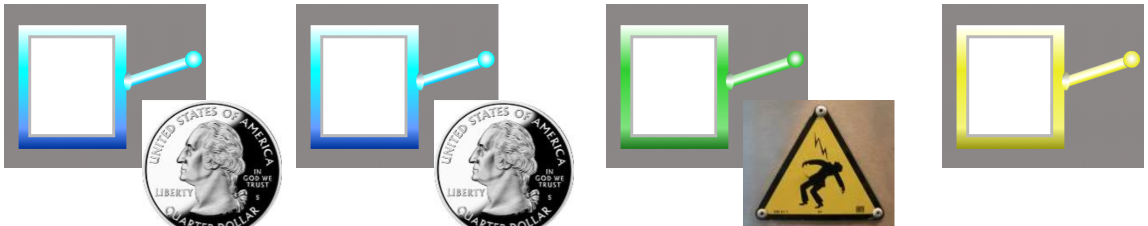
model
(probabilistic algorithm:
experience → choices)



choice

typical analysis

t-1 t-2 t-3 t-4 …

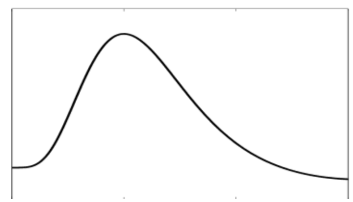


experience
(past choices & outcomes)

delta-rule learning

$$\delta_t = r_t - V_t(c_t)$$
$$V_{t+1}(c_t) = V_t(c_t) + \alpha\delta_t$$

*

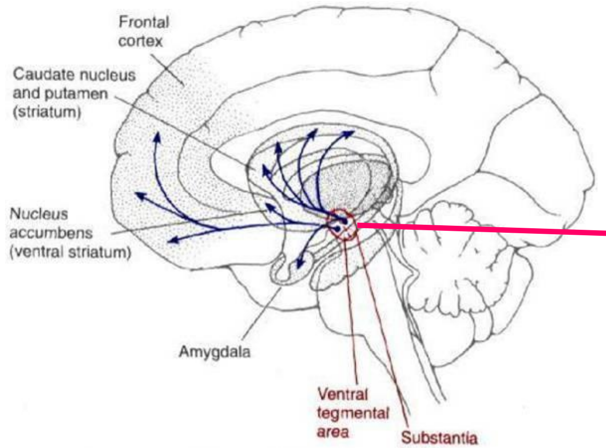


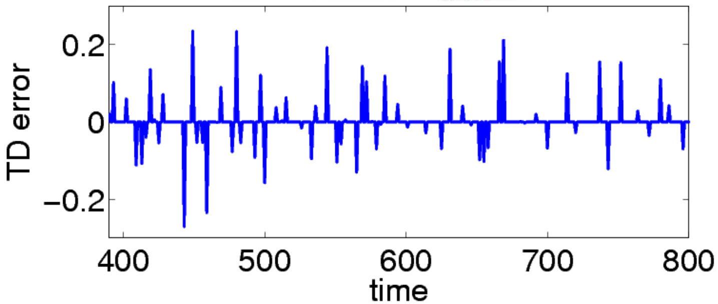
model
(probabilistic algorithm:
experience → BOLD)

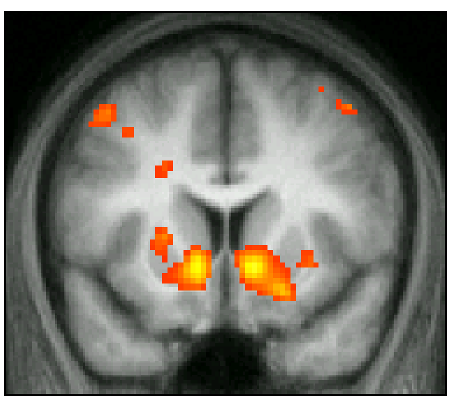
fMRI:
search for neural correlates;
how do they behave?

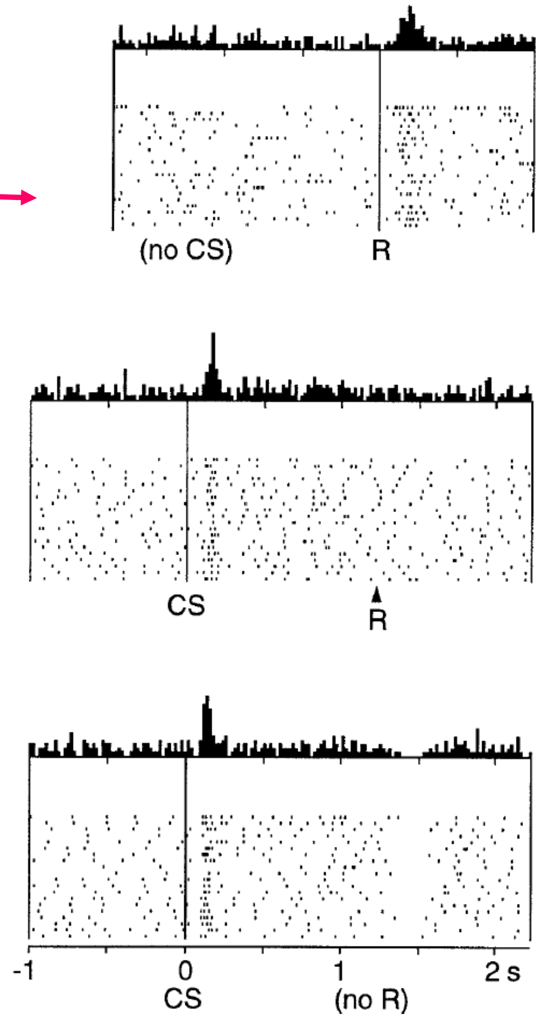
neural response

dopamine & RL



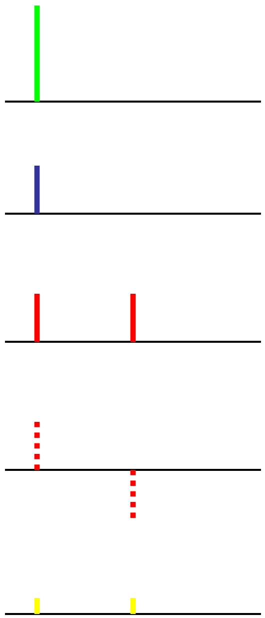




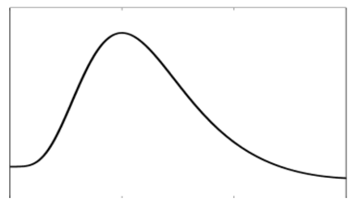


(Schultz et al. 1997)

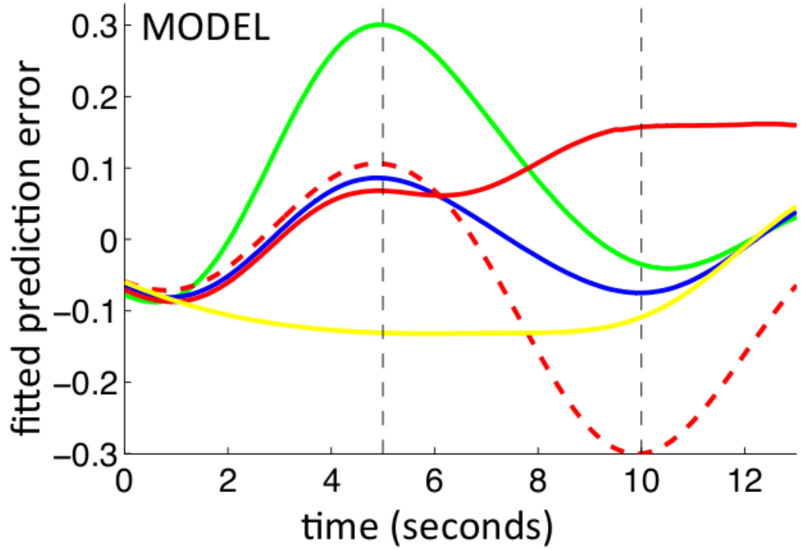
striatal BOLD and PE

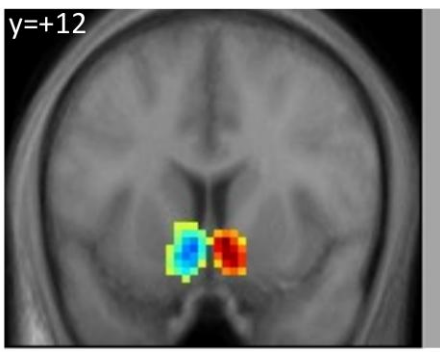


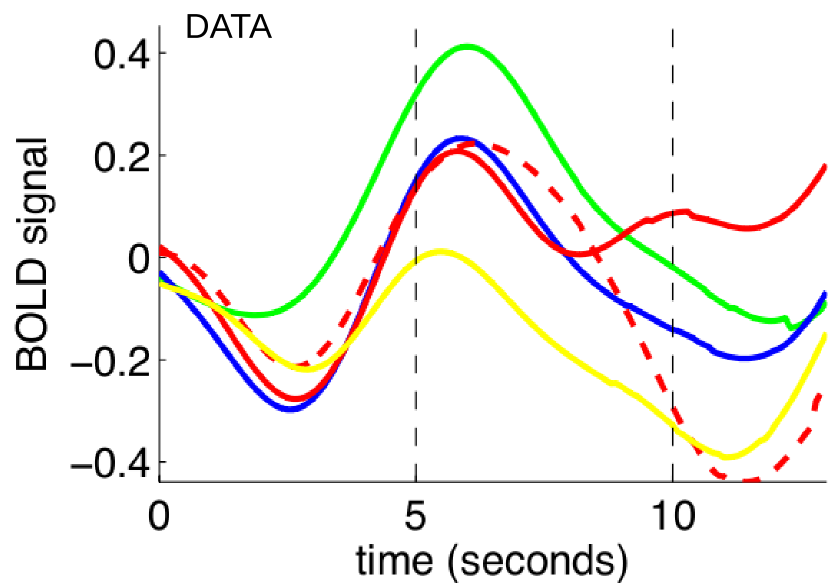
*



=



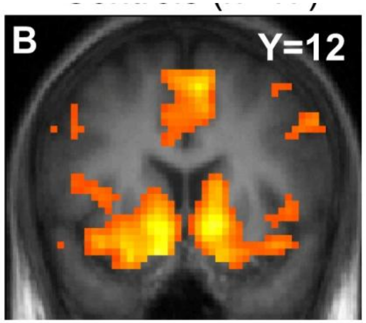




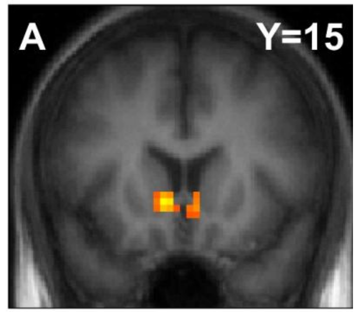
(Niv et al. under review)

Striatal BOLD, DA, and PE

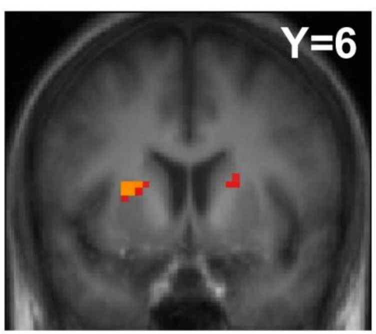
healthy control



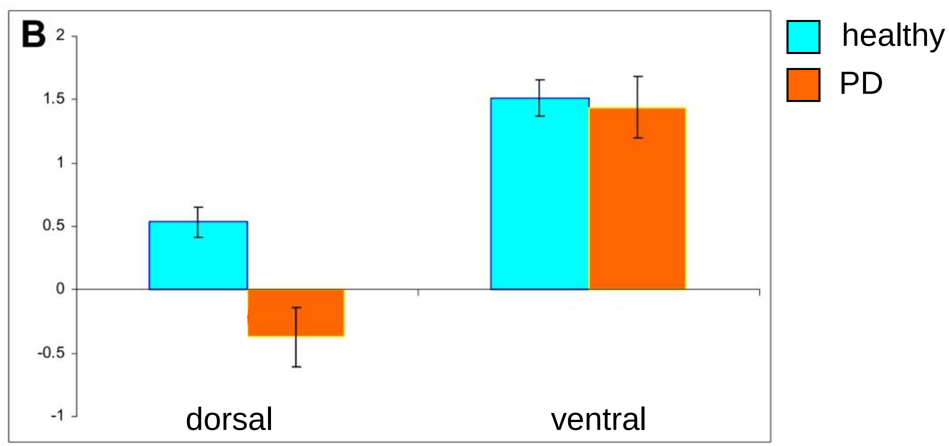
Parkinson’s disease



difference



BOLD PE effect sizes



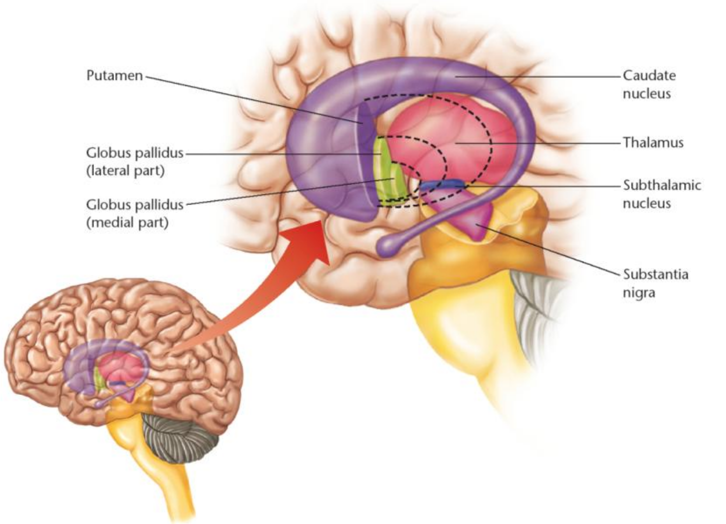
(Schonberg et al 2010)

where are we

- behavioral suggestions of delta-rule learning
- phasic dopamine response well characterized by TD prediction error signal
  - animals, human fMRI
  - nb: some anomalous responses
  - suggests very specific mechanism for learning/prediction in sequential tasks
  - (but this is a causal/functional claim and not uncontroversial)

Basal ganglia

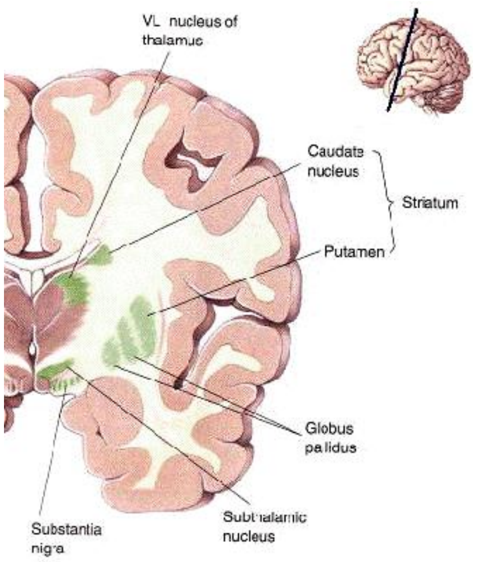
Several large subcortical nuclei, unfortunate nomenclature



essential puzzle (as with dopamine): motor control *plus* (drugs, reward, motivation)

various specific ideas

- limbic-motor gateway
- action selection, facilation/suppression
- behavioral sequencing
- behavioral monitoring
- …



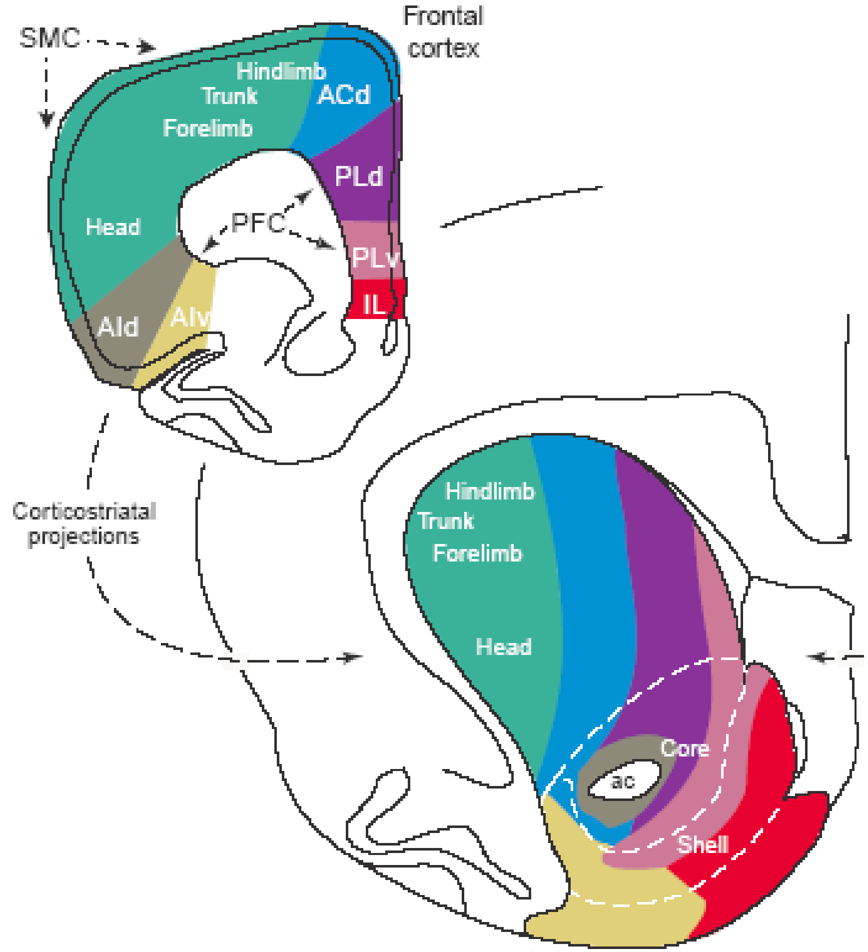
prediction error

- what should prediction error do?
  - drive learning
  - …about expected rewards (eg state values)
  - …to guide choice (eg policies, action values)

→ this fits well with the multifarious roles of dopamine & its targets

striatum: basal ganglia input

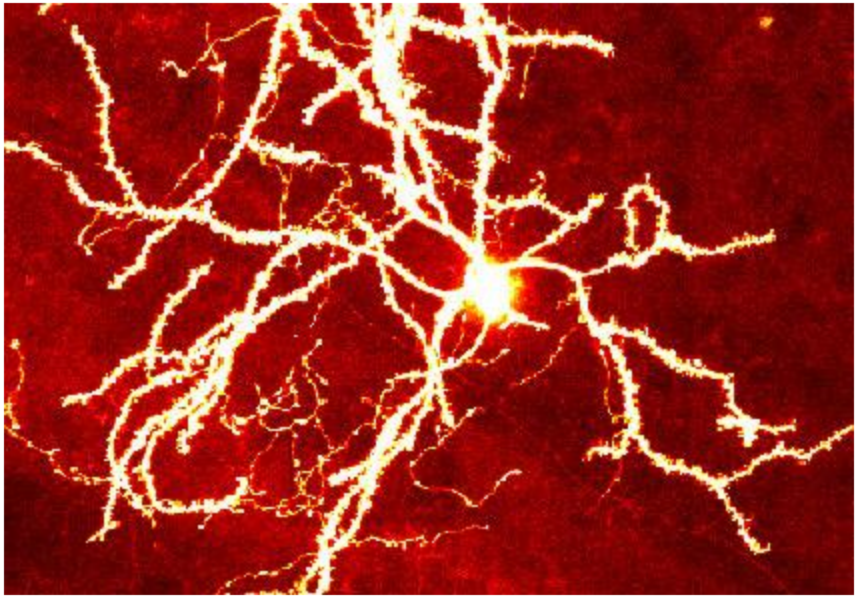
- Projection from entire cortex (including sensory, motor, associative areas) to striatum
- Topographic



Voorn et al 2004

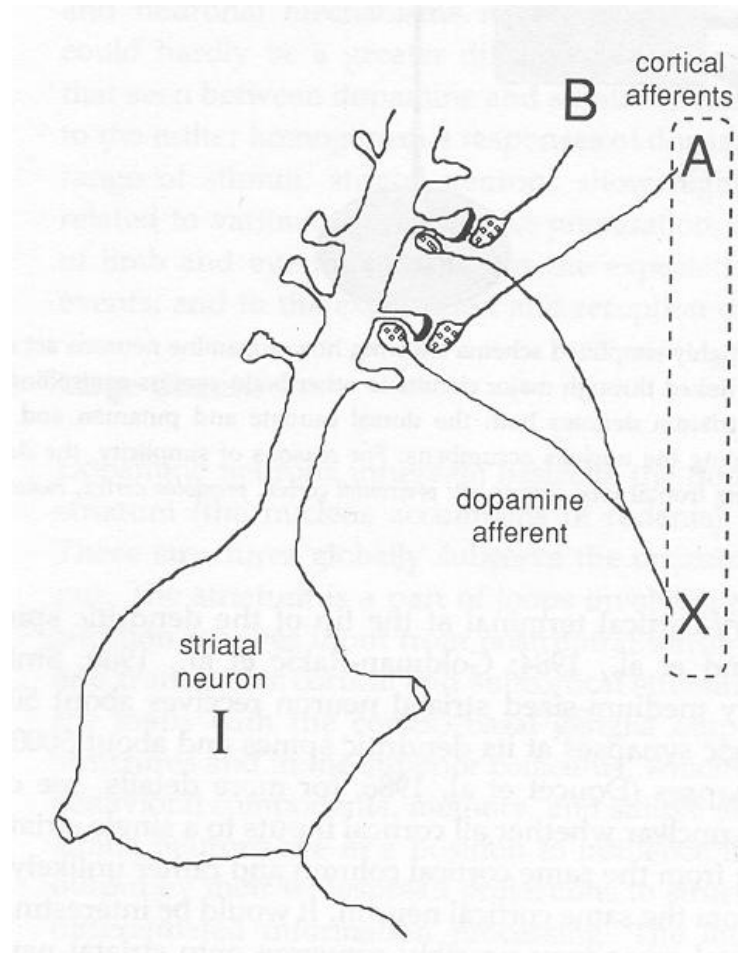
Medium spiny neurons

- Principal neuron type in striatum
- Recipient of corticostriatal inputs
- Extensive dendrites – each receives input from 10,000 fibers
- Unusual: GABAergic (inhibitory) projections
  - Also collaterals (competitive network)

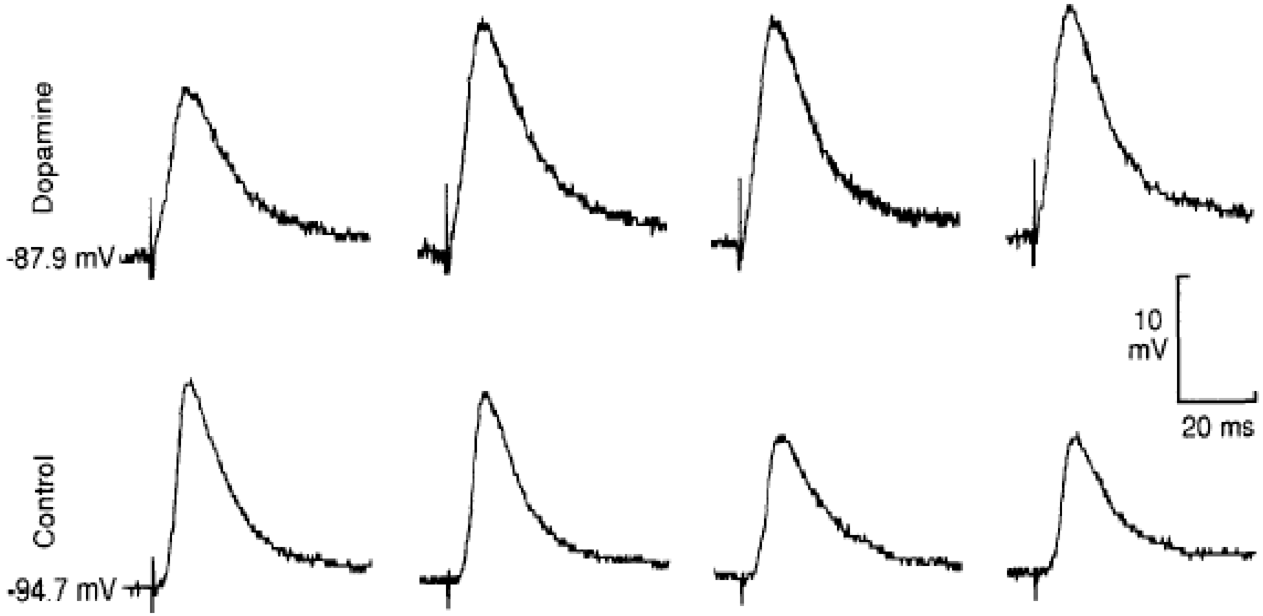


Dopamine and plasticity

- If dopamine carries a prediction error, where does learning happen?
- Potentially, the cortico-striatal synapse



DA and corticostriatal plasticity



Wickens et al. 1996

Three-factor learning rule? (pre/post/dopamine)

Actor / critic

Critic (values)

$$V(s_t) \leftarrow V(s_t) + \eta \cdot \delta_t$$

Actor (policy)

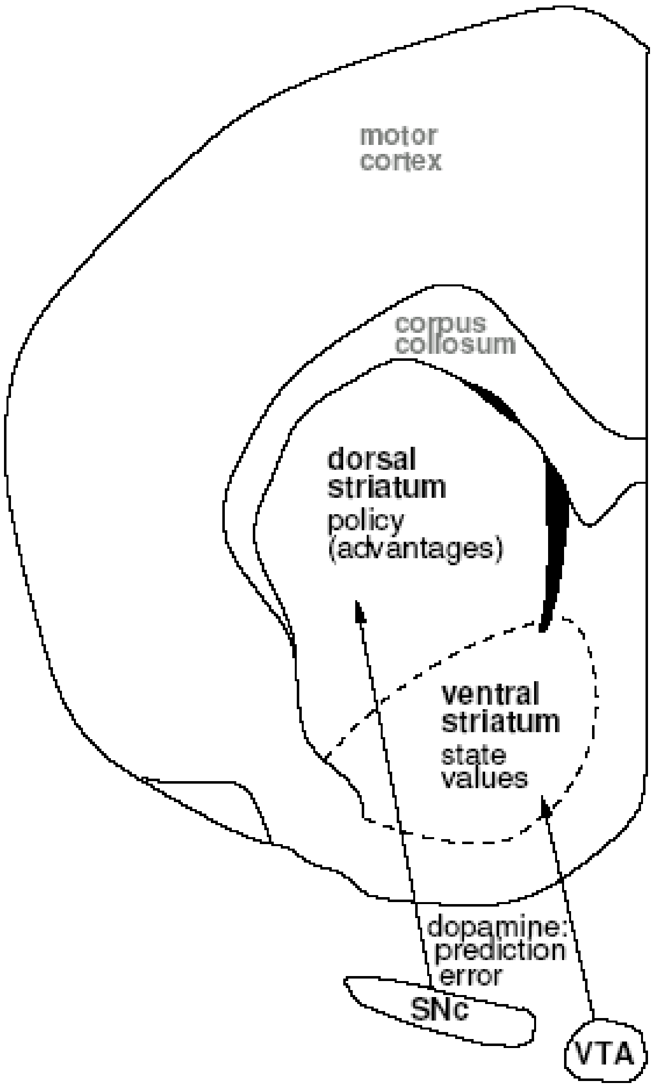
$$\pi(s_t, a_t) \leftarrow \pi(s_t, a_t) + \varepsilon \cdot \delta_t$$

$$\delta_t = r_t + V(s_{t+1}) - V(s_t)$$

(choose according to π)

- Same error signal for values and policies
- Theory of interaction of Pavlovian (prediction) and instrumental (action choice) conditioning
- gradient ascent on V wrt π

actor/critic

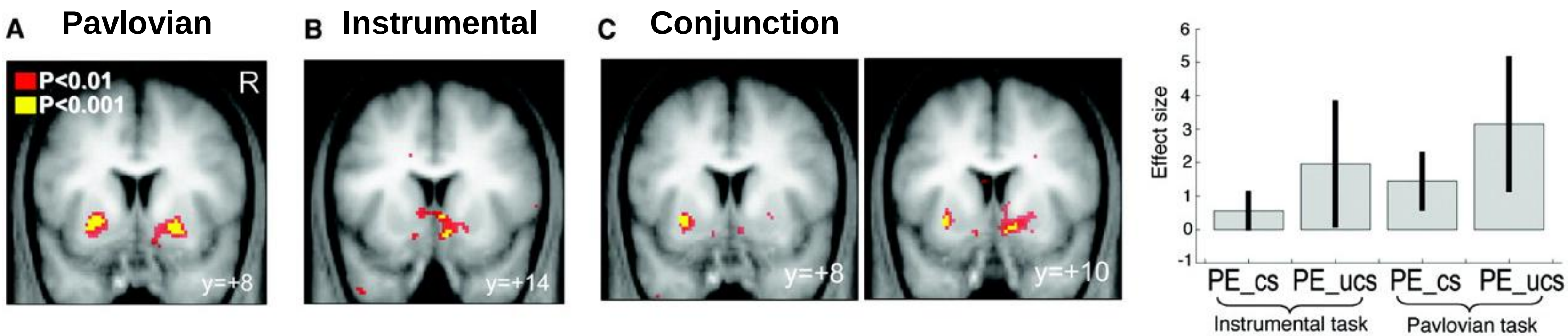


dopamine signals to both motivational & motor striatum appear, surprisingly the same

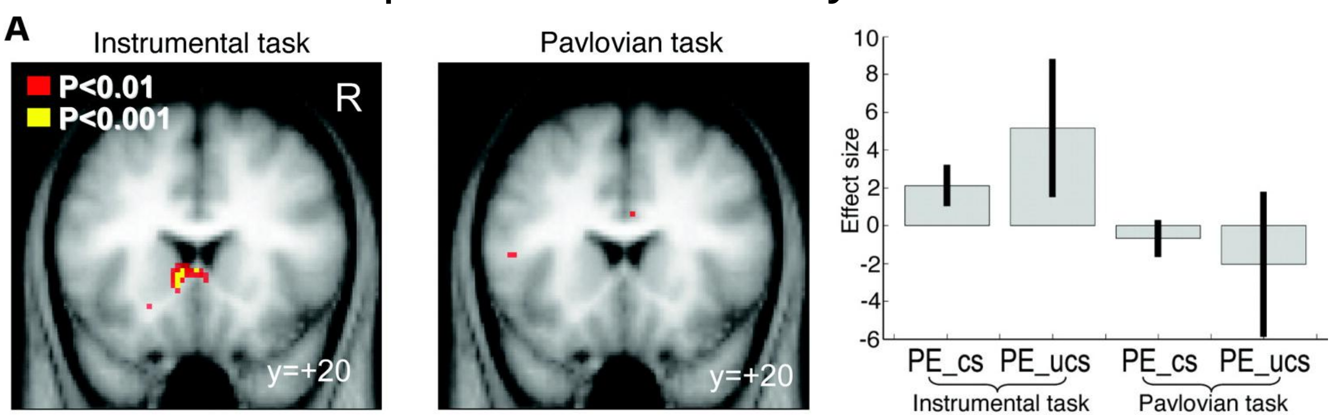
suggestion: training both values & policies

actor/critic in fMRI?

ventral striatum: correlated with prediction error in both conditions

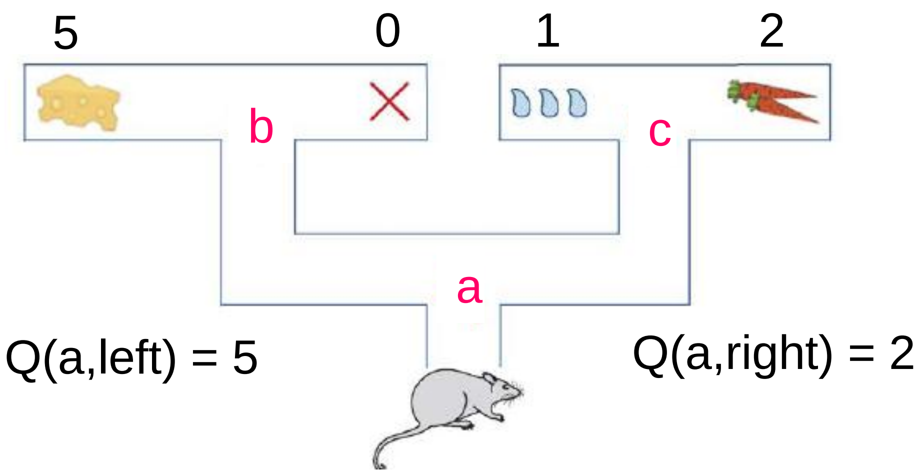


Dorsal striatum: prediction error only in instrumental task



(O'Doherty et al. 2004)

Q learning



another version learns state-action values (but doesn’t distinguish actor from critic)

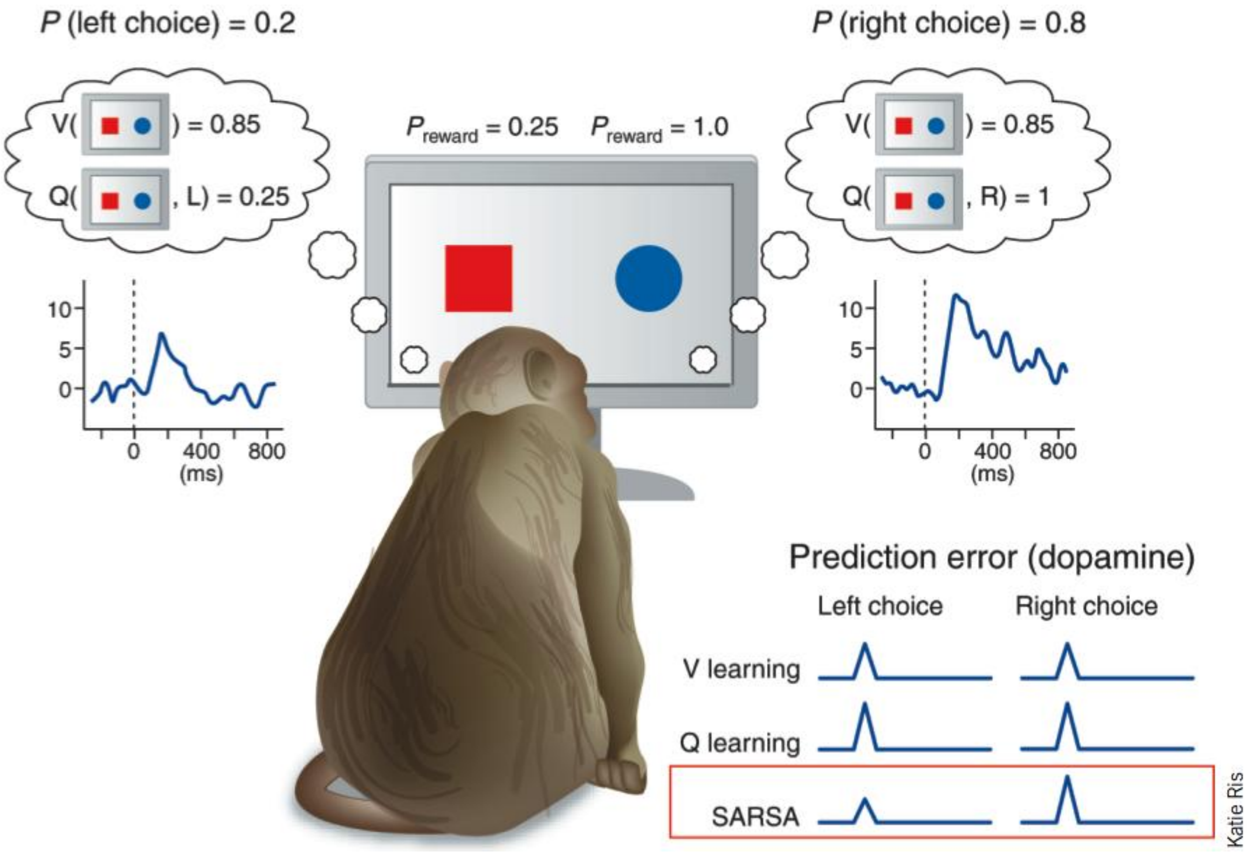
$\delta_t = r_t + Q(s_{t+1},a_{t+1}) - Q(s_t,a_t)$ (SARSA)

or

$\delta_t = r_t + \mathrm{argmax}_a[Q(s_{t+1},a)] - Q(s_t,a_t)$ (Q-learning)

$Q(s_t,a_t) = Q(s_t,a_t) + \eta\delta_t$

SARSA?



NIv et al 2006 after Morris et al. 2006

where are we

- dopamine responses (+ various aspects of their functional neuroanatomy) seem well accounted for by TD learning
  - though not without questions!
- how can this possibly scale up to real-world, e.g. embodied, behavior?
  - especially given the constraint that at the heart there is apparently a simple TD system?

plan

reinforcement learning in neuroscience (psychology, behav. economics)

1. dopamine & the TD hypothesis
   - behavioral & analytical background
   - recordings: spiking, fMRI
   - functional neuroanatomy

2. beyond the TD hypothesis
   - states (→ POMDPs & belief states)
   - actions (→ hierarchical RL, decomposed error signals)
   - rewards (→ model-based vs model free)

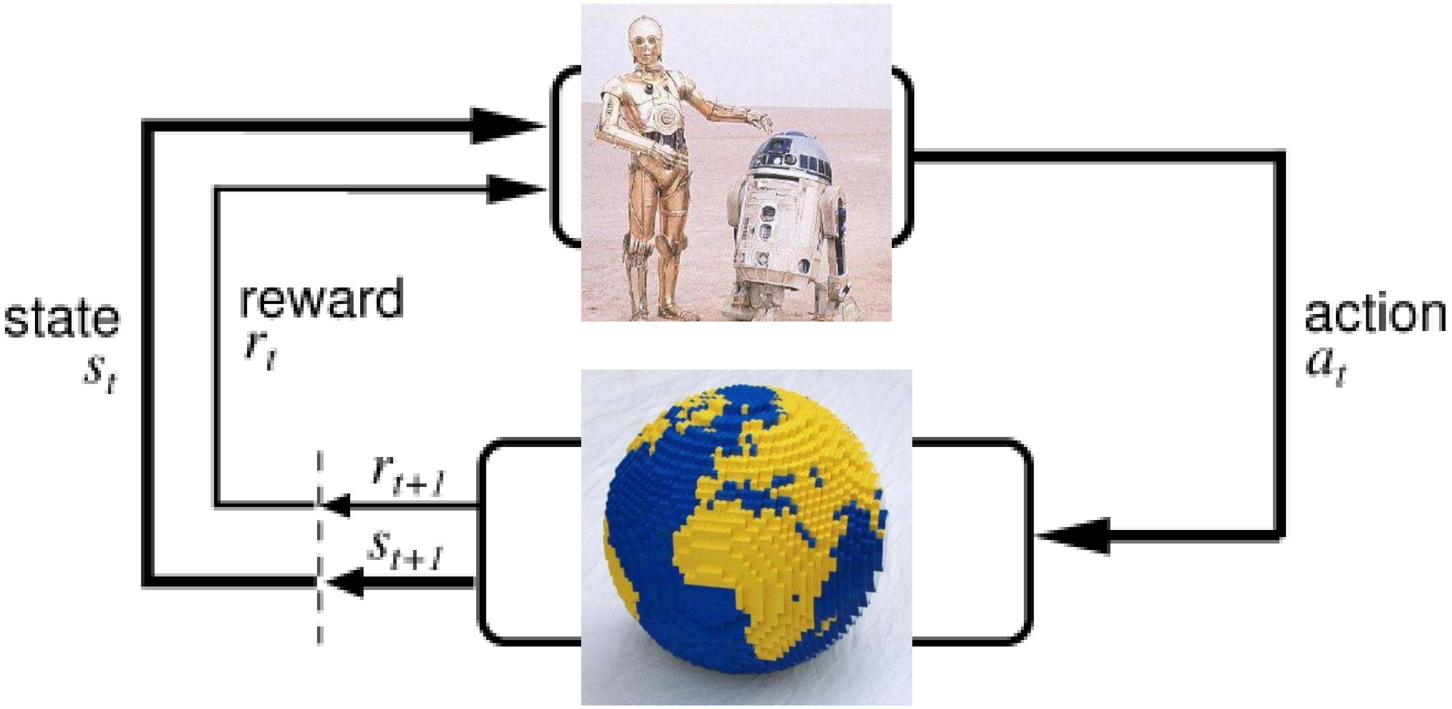
.

Markov Decision Process

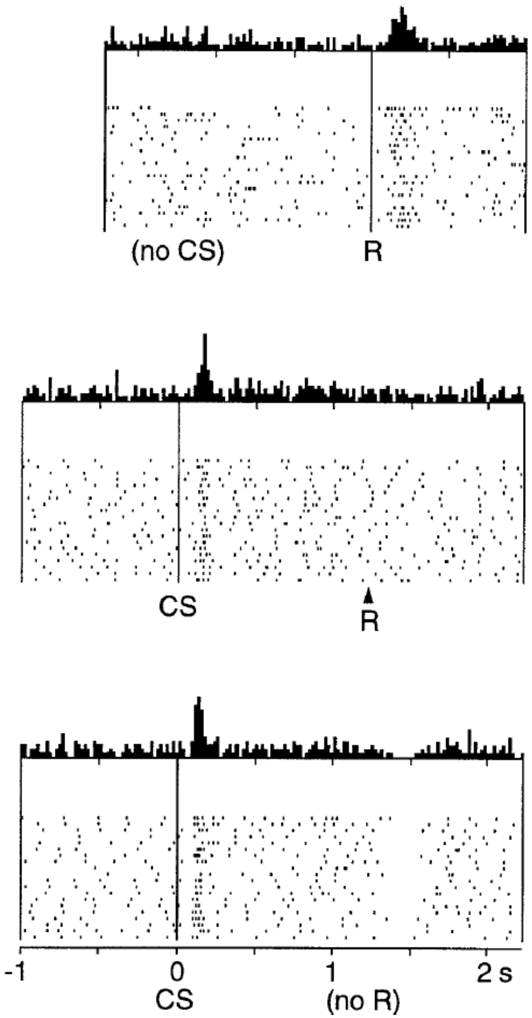
class of stylized tasks with

states, actions & rewards

→ what do these correspond to in biology?



state & history



(Schultz et al. 1997)

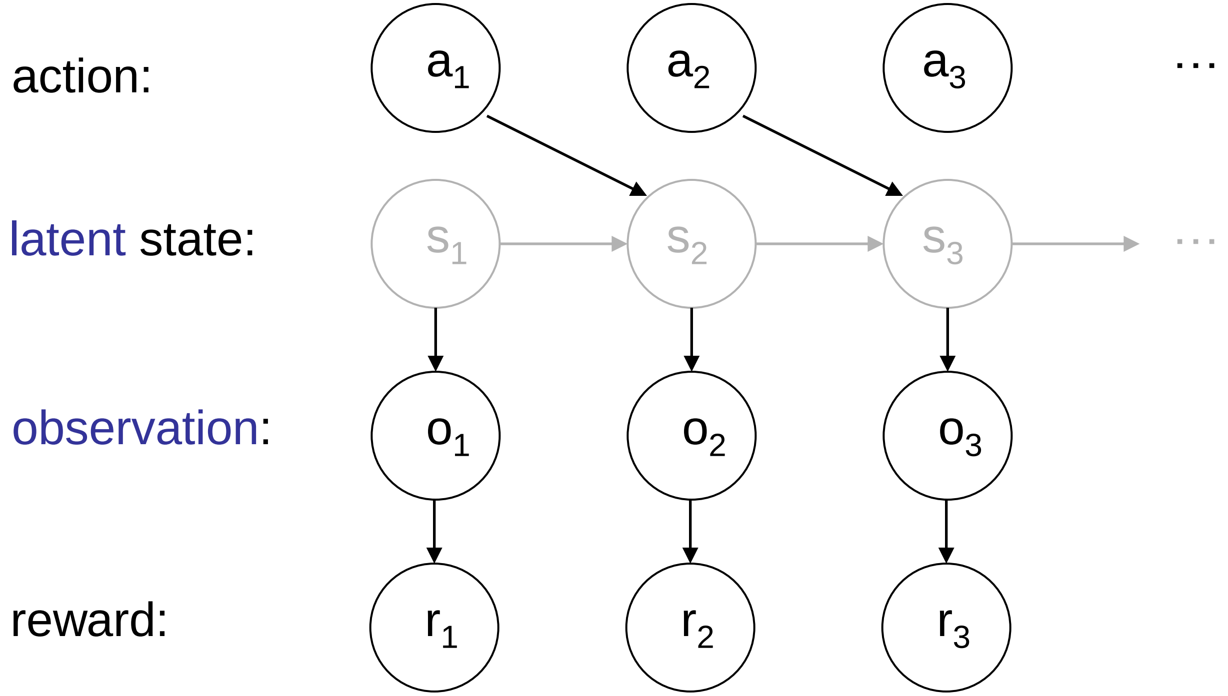
- What are the sensory events in this task?
- What is the state for this task?
- What tells the neuron when to pause for omitted reward?

→ raw sensory events are clearly non-Markovian

various approaches: history, POMDP

Partially observable MDP

- MDP but state is unobserved



Transition function: $P(s_{t+1} \mid s_t, a_t)$

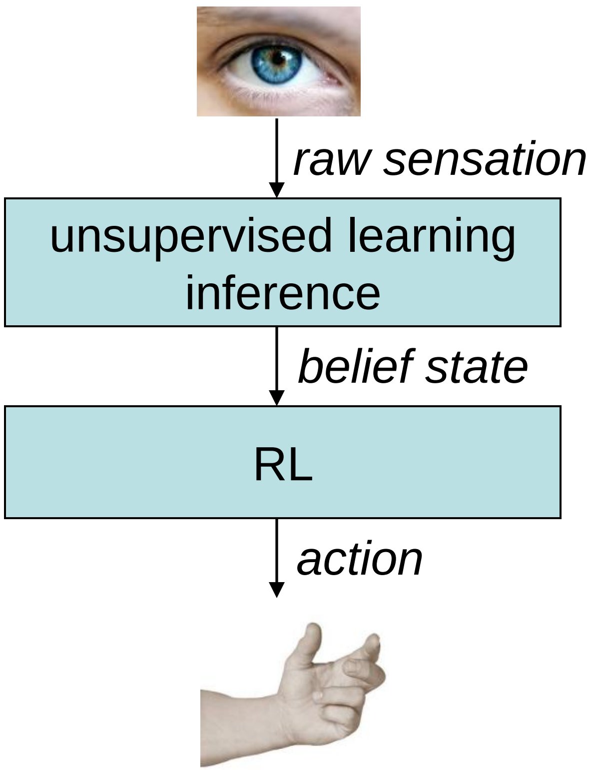
Observation function: $P(o_t \mid s_t)$

Reward function: $P(r_t \mid s_t)$

Belief state MDP

- belief states (ie inferred state distributions) in a POMDP themselves form the states of an MDP (Kaelbling et al 1995)

- Thus in principle we can use "standard" RL in the space of belief states
  - Framework: sensory analysis infers belief state; this becomes state for RL (BG etc)
  - This fits well with Bayesian models of sensation & sensory cortex
  - severe practical issues related to dimensionality/continuity of belief state



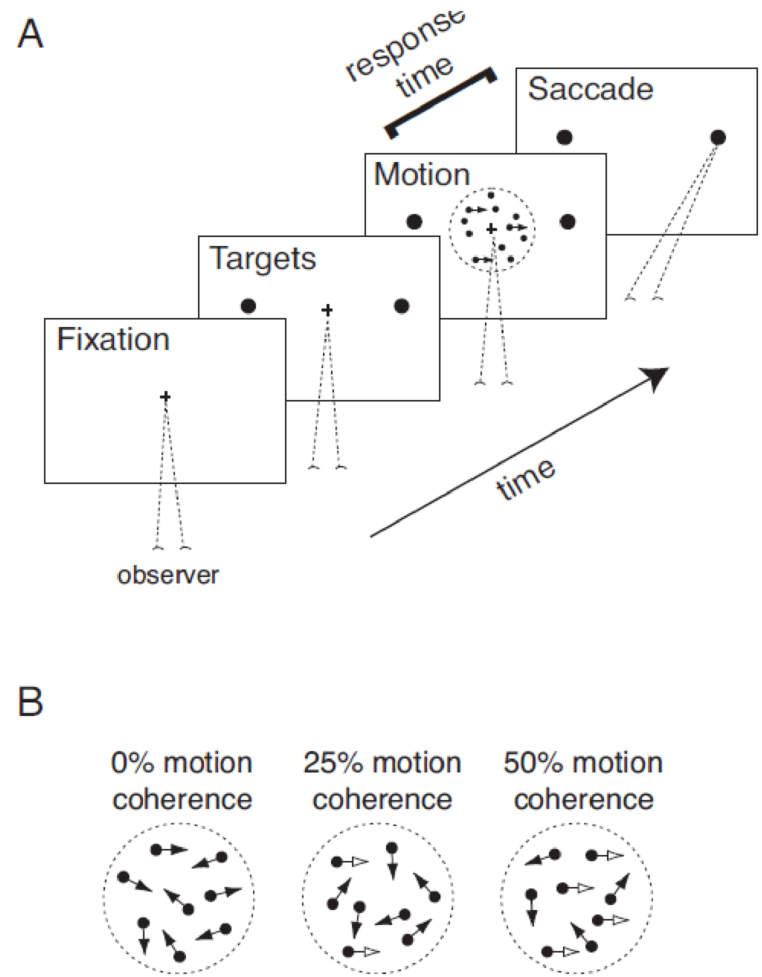
Daw et al. 2006
Dayan & Daw 2008
Gershman & Niv 2009
Rao 2010

example

Shadlen, Newsome, Movshon, etc

"sensory decision" task: are the snowy dots moving left or right?

- coherence varied (hard or easy)
- watch till you're ready to answer ("reaction time" task)
- signal answer with left or right saccade
- no real learning



Palmer et al 2005

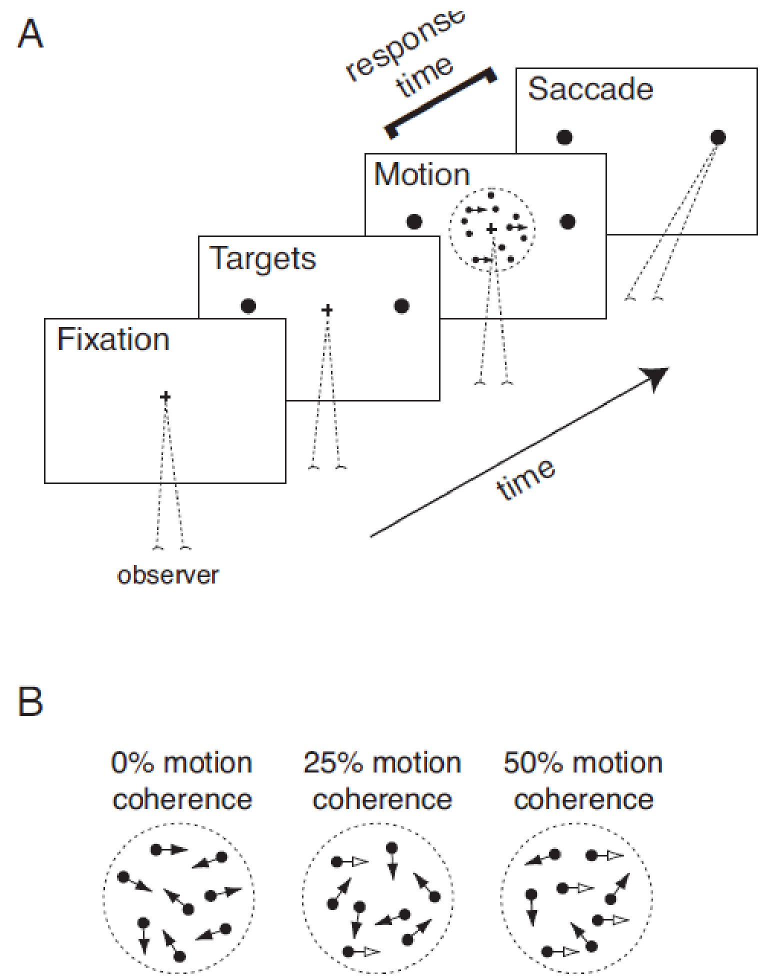
task

idealized task:

- you don’t know if dots are moving left or right
- at each step you may respond “left” or “right” or watch another burst of noise

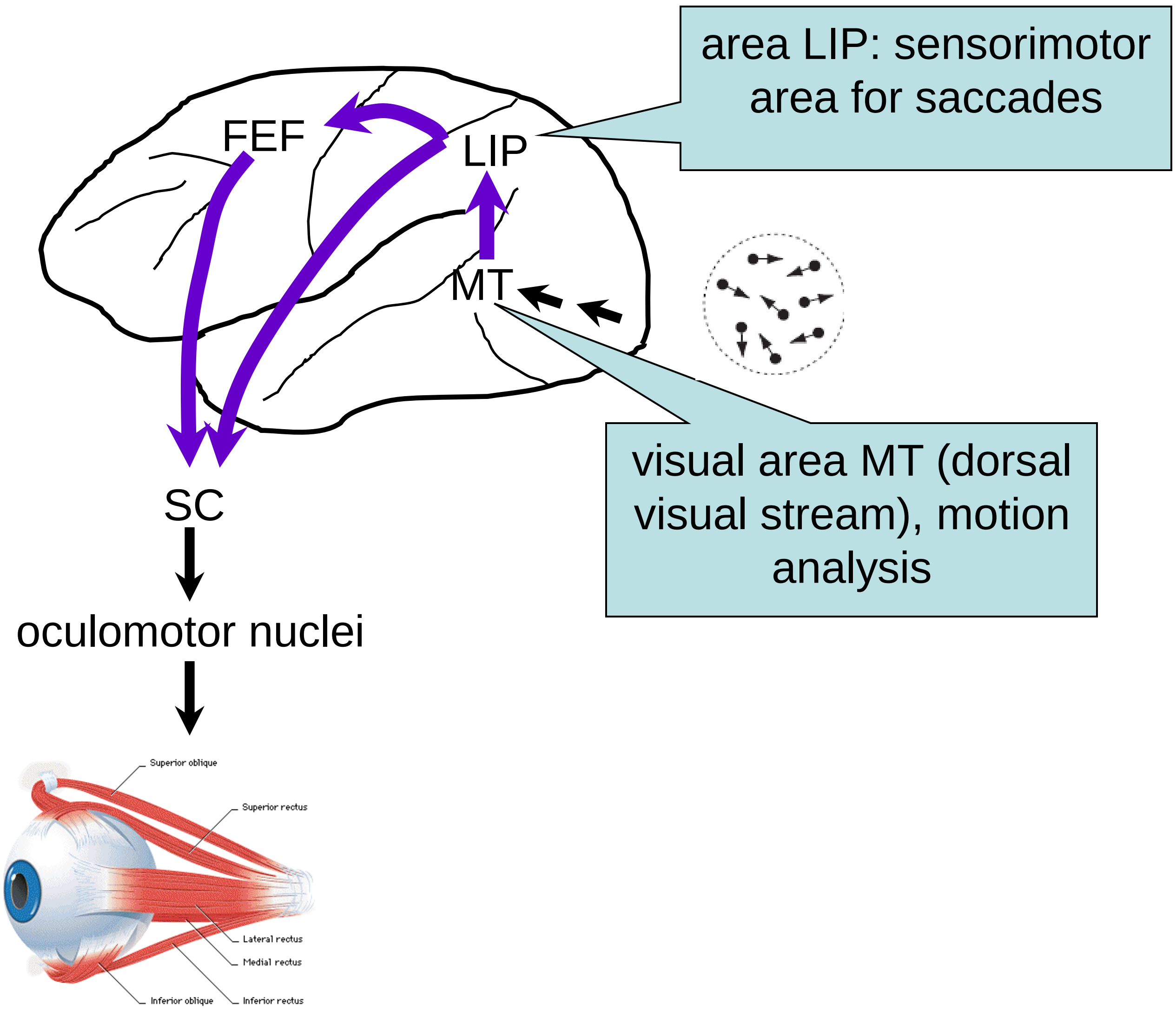
→isomorophic to POMDP tiger problem

→solution tracks posterior prob of underlying state (right or left) given data; responds on threshold (SLRT; Gold & Shadlen 2002)



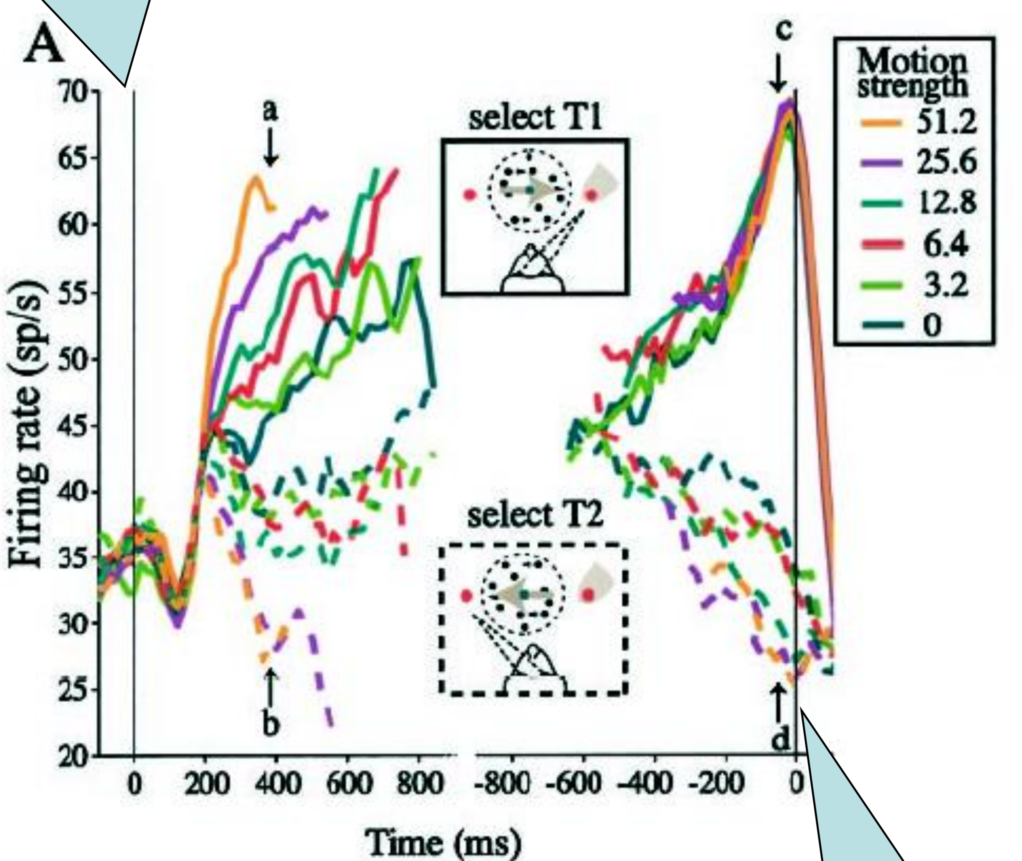
Palmer et al 2005

Key neural players



belief state?

align on motion onset



align on saccade

neurons in area LIP

ramps prior to saccade
faster for larger coherence

saccade occurs when
response hits threshhold

Roitman & Shadlen

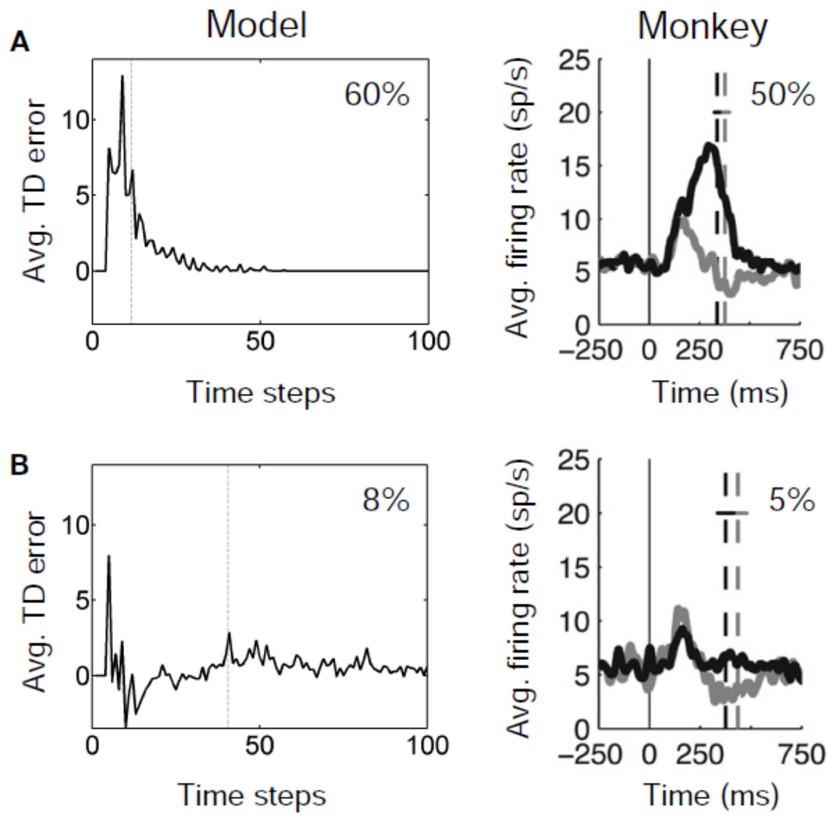
belief state as state

Nomoto et al 2010:
Dopamine neurons in this task show 2-stage responding

1) quick response related to dots onset

2) slower response related to trial difficulty

→ Rao 2010: latter tracks change in value due to evolving internal belief state (over additional latent variable of trial difficulty)



model: Rao 2010
data: Nomoto et al 2010

where are we

- huge issue: where does state come from

- “state” driving DA response is internal
  - evolves with passage of time
  - evolves with non-Markovian input

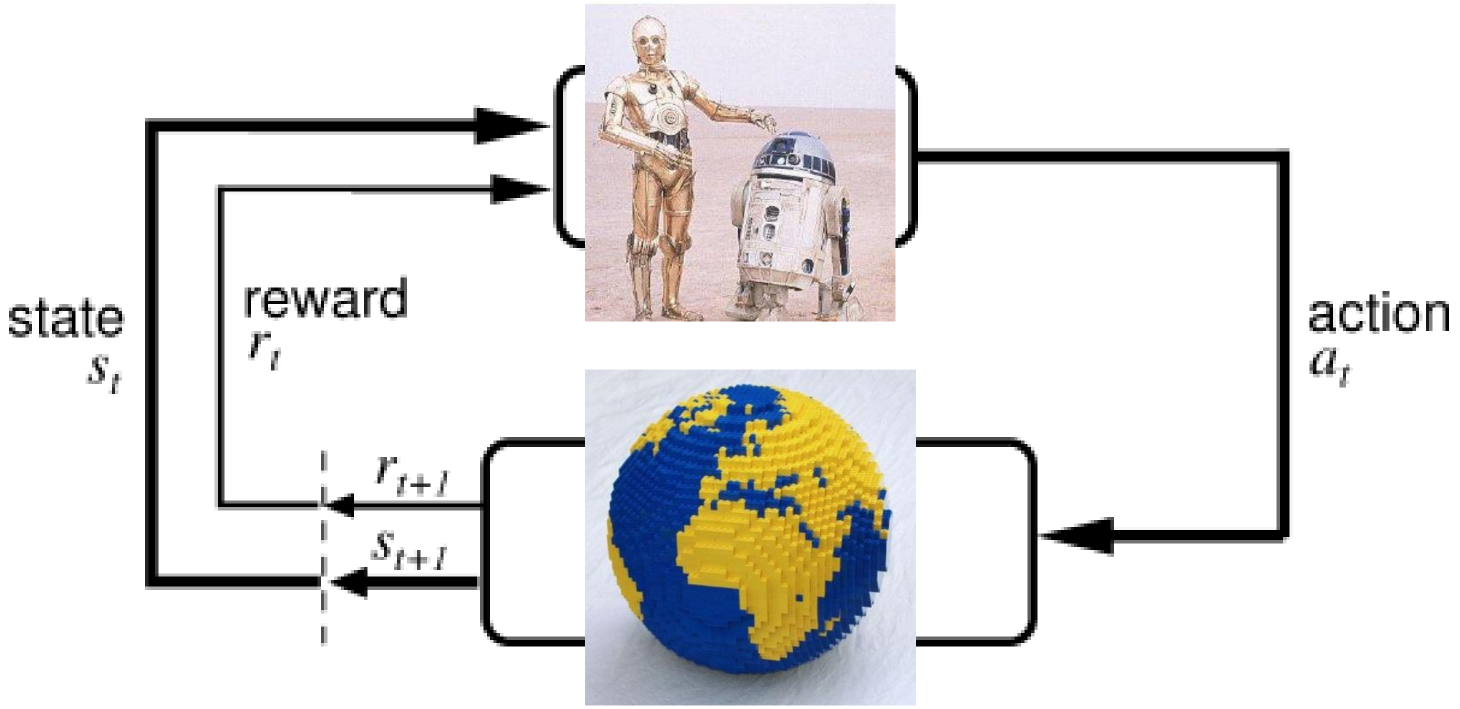
- one way to conceptualize this is POMDP belief state
  - leads to many more questions
  - but encapsulates simple RL mechanism behind fancy sensory inference

Markov Decision Process

class of stylized tasks with

states, actions & rewards

→ what do these correspond to in biology?

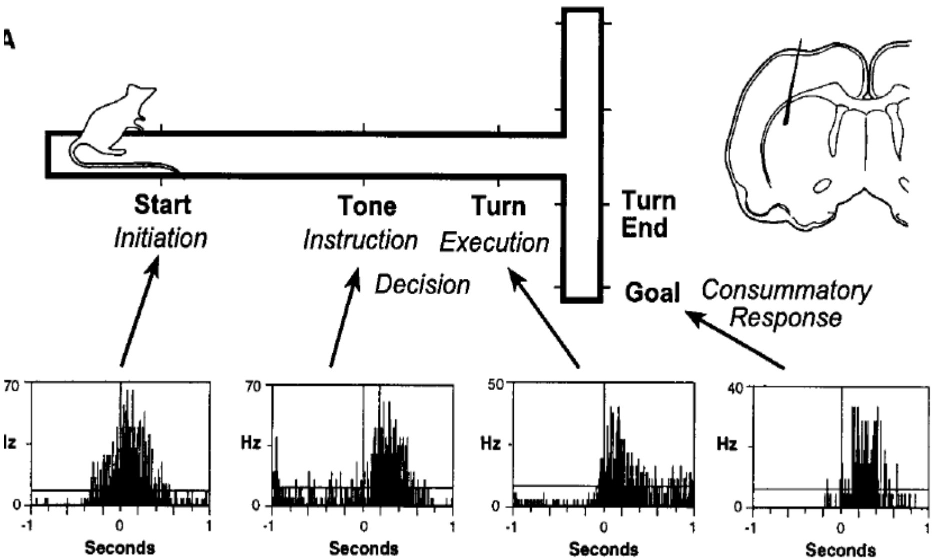


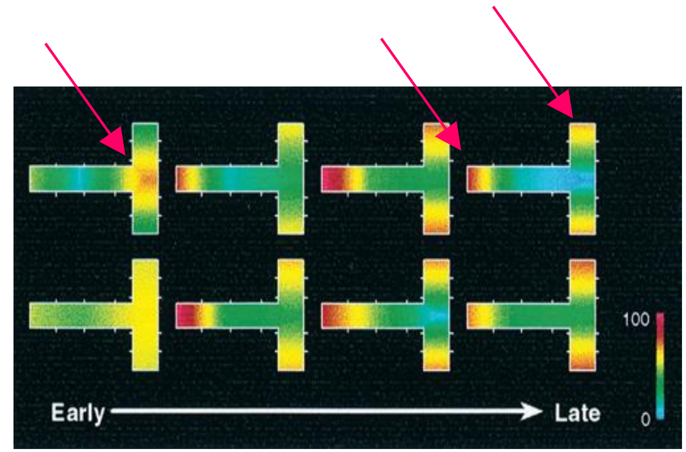
action

- the simple notion of action in a bandit task is also not good enough
  - Ballard example of encapsulation

- three examples involving curse of dimensionality in action space
  - sequential action & hierarchical structure
  - multieffector action
  - vigor

action "chunking"

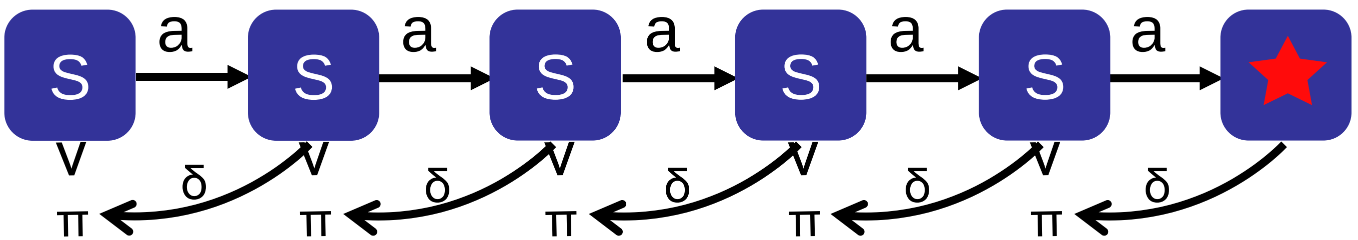




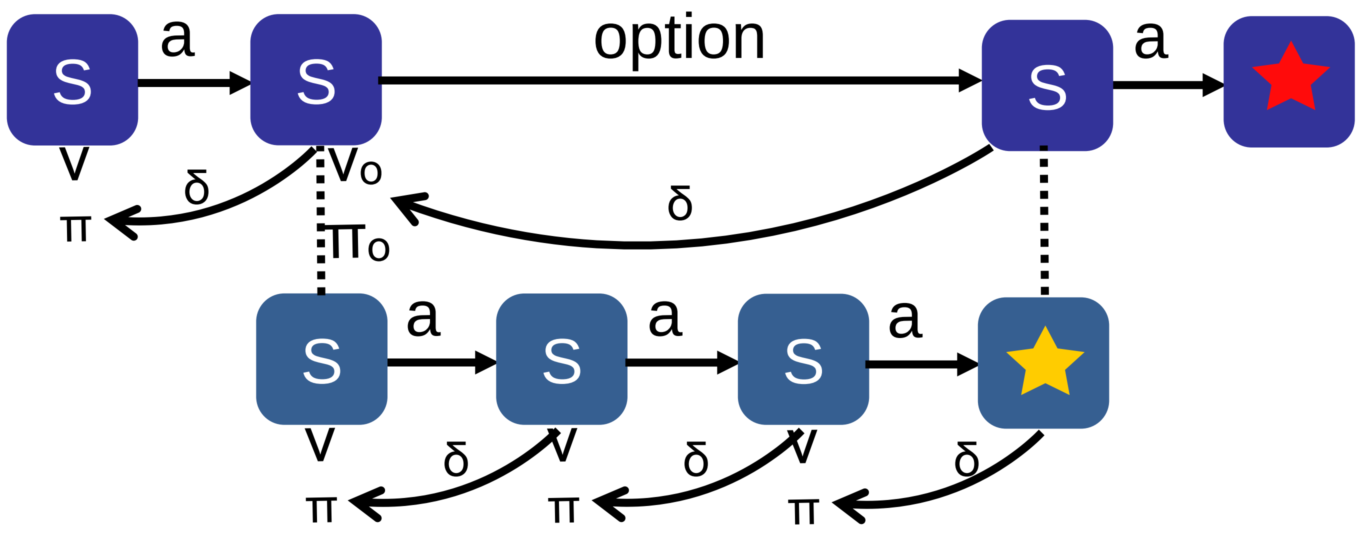
(Jog et al 1999)

activity patterns in rat DL striatum change with overtraining
→ responses move to beginning and end of action sequence
→ reminiscent of hierarchical RL, eg options (Precup et al. 1998; Botvinik et al 2009)

standard RL

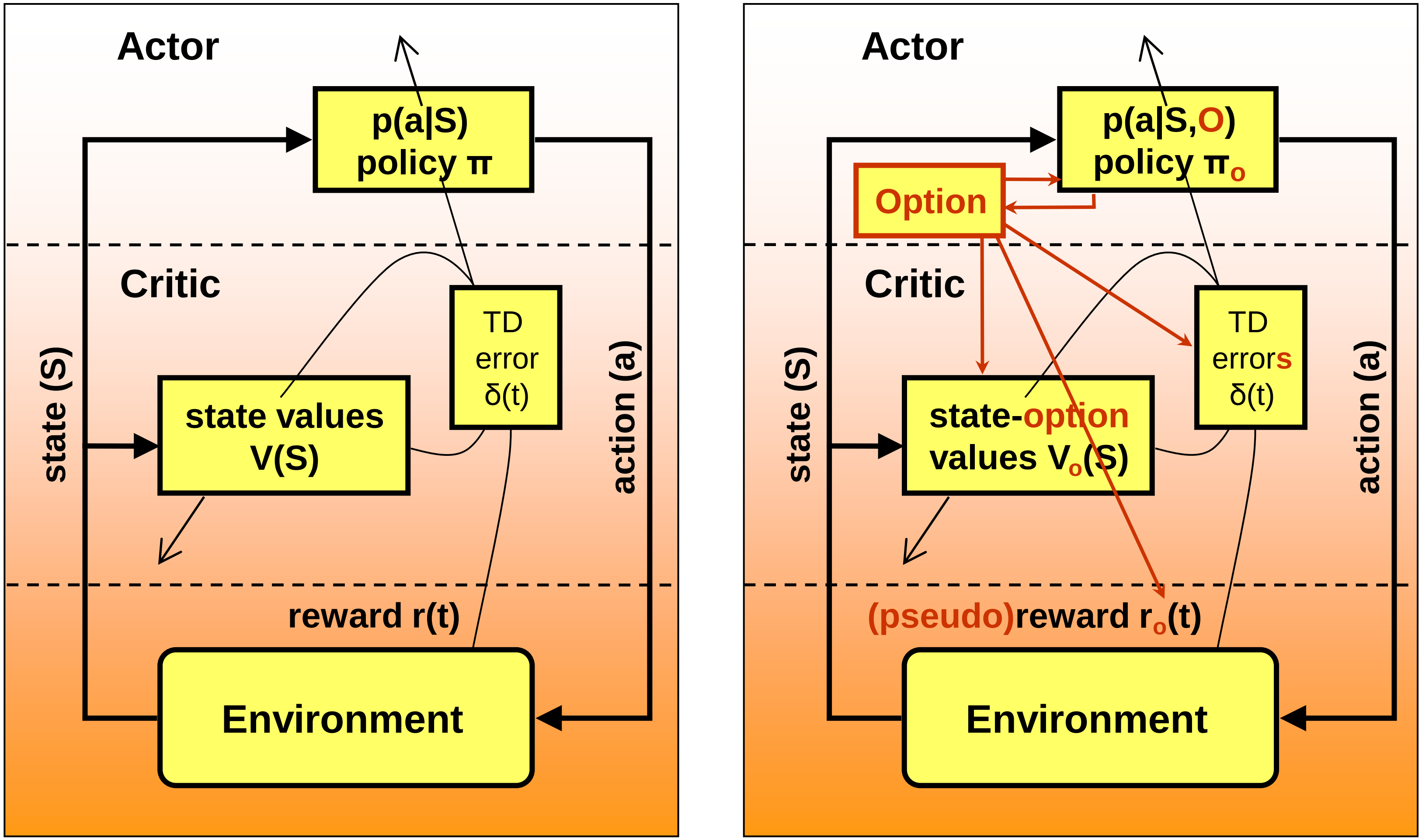


hierarchical RL



options (Precup et al.): macro-actions learned by TD methods
→ but with multiple error signals (within and outside option, pseudorewards)

implications



Botvinik et al. 2009 review changes to standard actor/critic story this necessitates
see also Badre et al 2010; Reynolds & O'Reilly
- experiments underway

multieffector learning

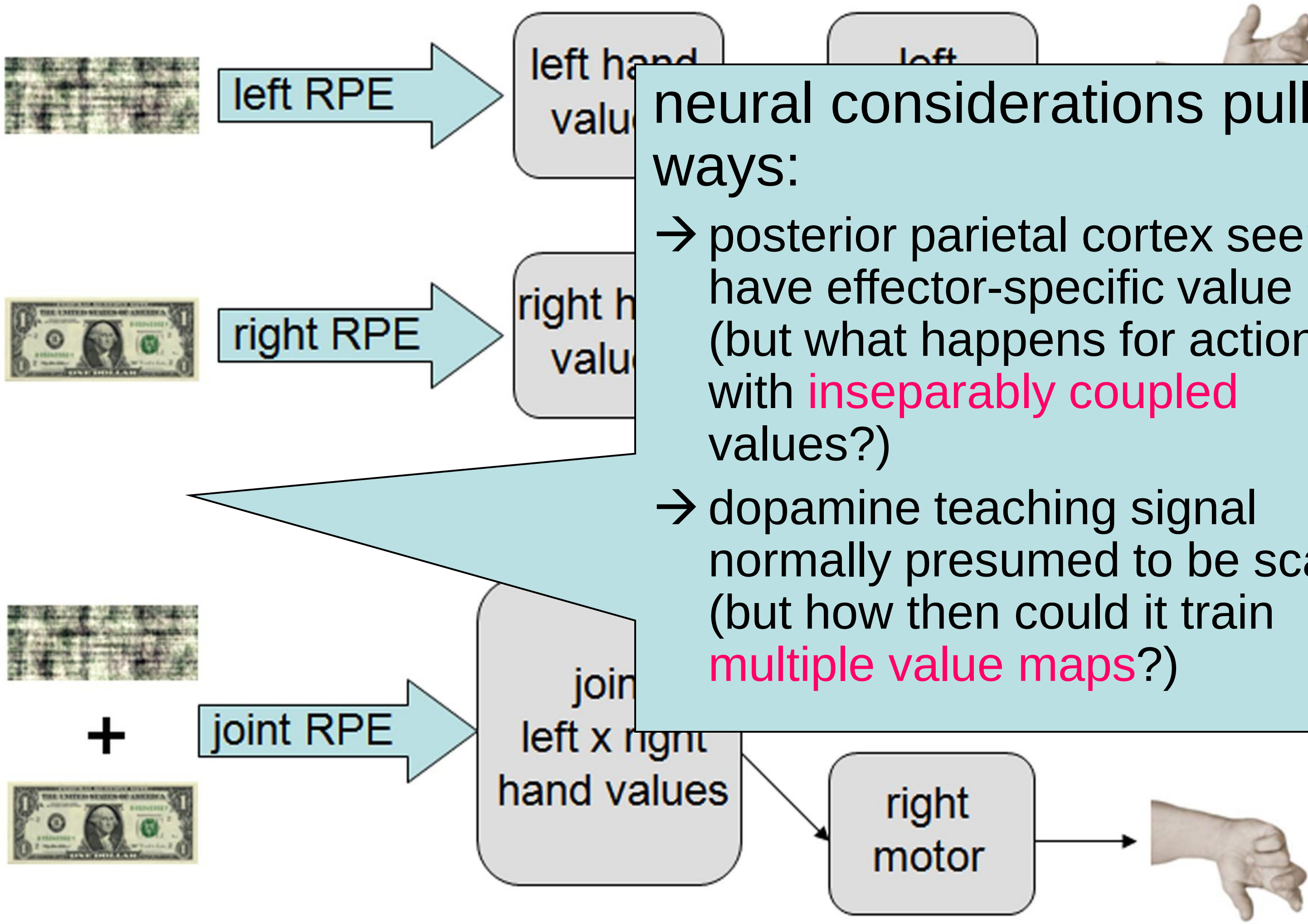
example 2:

even at a single step, due to the multieffector body, there is still a curse of dimensionality

general computational approach (eg Russell and Zimdars, 2003; Chang, Ho and Kaelbling, 2003; Rothkopf & Ballard): divide & conquer, exploiting structure of problem

- insofar as possible, learn separately at each effector
- this may involve a credit assignment problem (though not in the experiment to follow)

divide & conquer

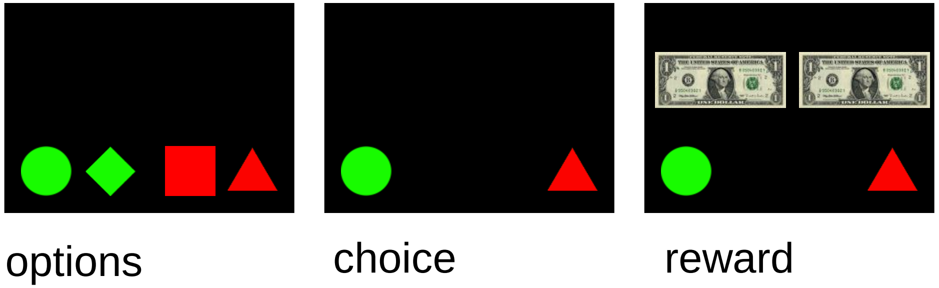


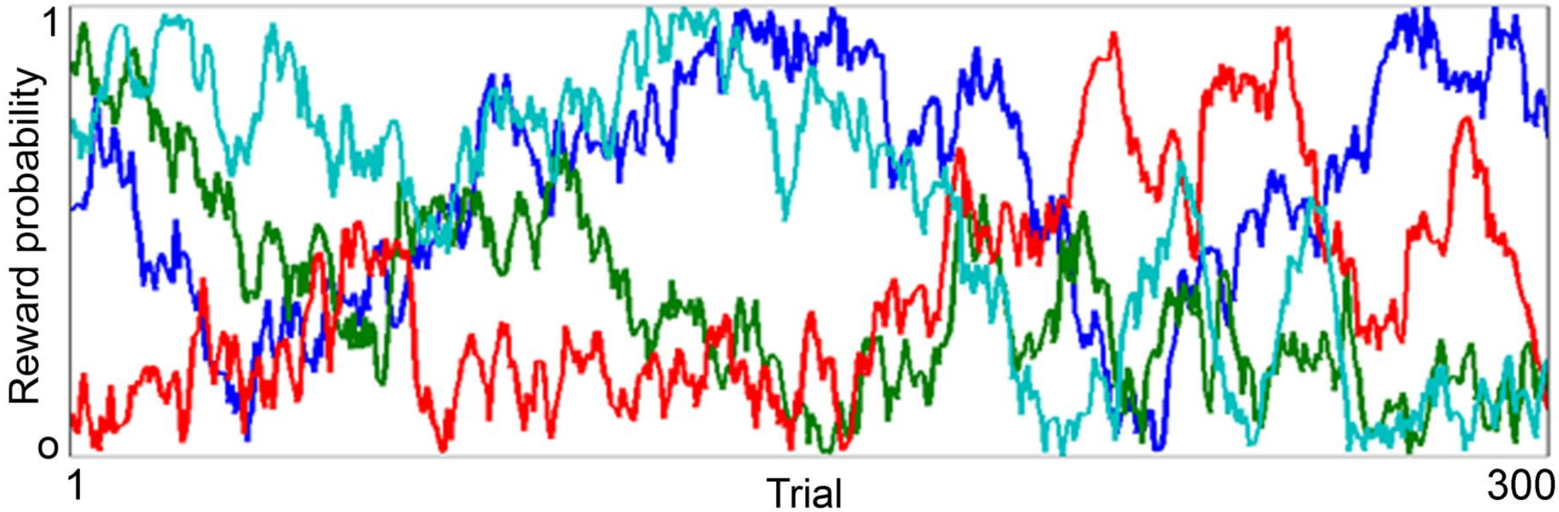
neural considerations pull two ways:

- → posterior parietal cortex seems to have effector-specific value maps (but what happens for actions with inseparably coupled values?)
- → dopamine teaching signal normally presumed to be scalar (but how then could it train multiple value maps?)

(Gershman et al. 2009)

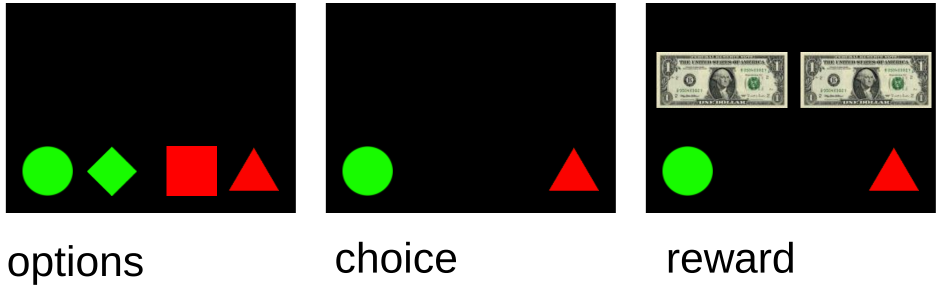
task





(Gershman et al. 2009)

models



Joint model: learns values for pairs of choices

$$Q(\bullet\blacktriangle) \leftarrow Q(\bullet\blacktriangle) + \alpha\delta(\bullet\blacktriangle)$$
$$\delta(\bullet\blacktriangle) = R(\bullet\blacktriangle) - Q(\bullet\blacktriangle)$$

Decomposed model*:* learns values for each hand separately

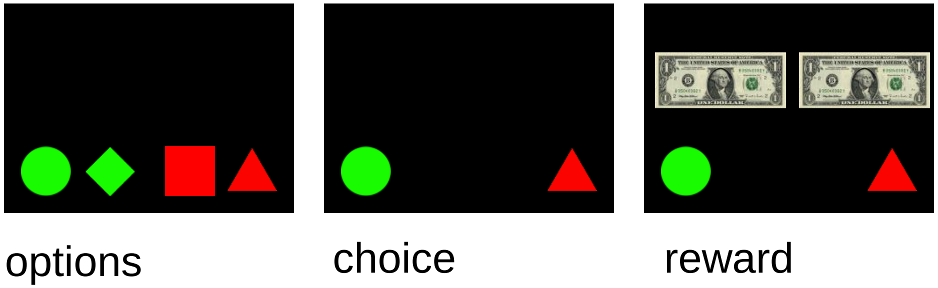
$$Q(\bullet) \leftarrow Q(\bullet) + \alpha\delta(\bullet)$$
$$\delta(\bullet) = R(\bullet) - Q(\bullet)$$

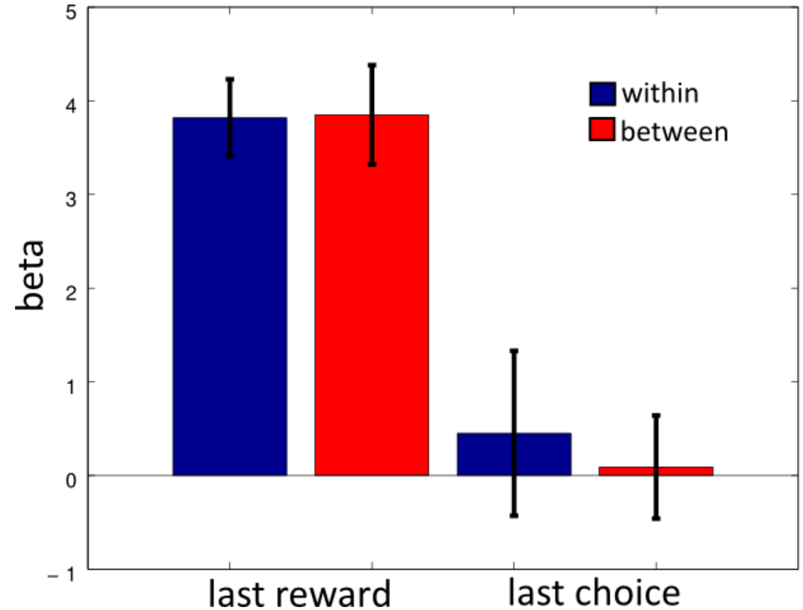
$$Q(\blacktriangle) \leftarrow Q(\blacktriangle) + \alpha\delta(\blacktriangle)$$
$$\delta(\blacktriangle) = R(\blacktriangle) - Q(\blacktriangle)$$

Does behavior reflect decomposed values?

Do neural signals?

behavior

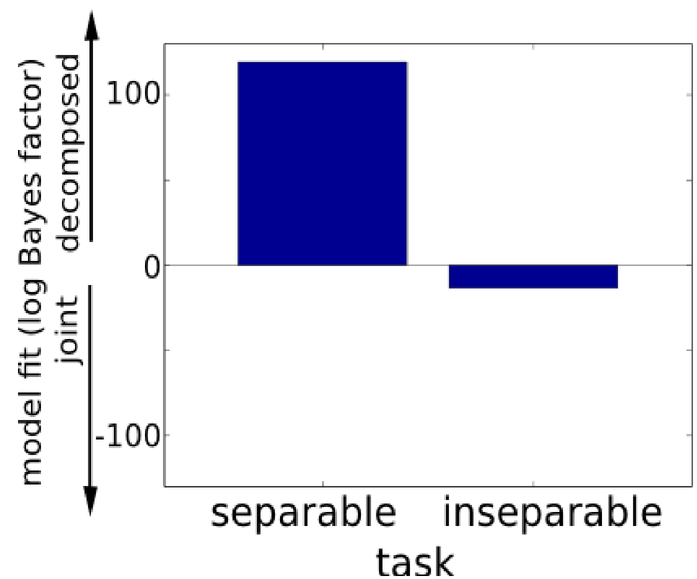




1. regression consistent with separable solution (effect of outcomes bleeds between joint actions; P<.001)

2. comparison of full RL models favors separable one (15/16 subjects)

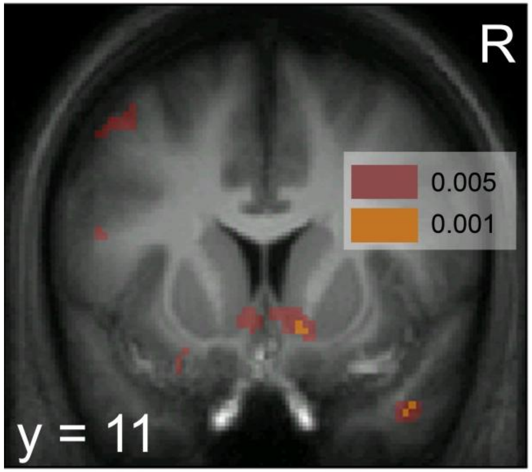
3. in separate behavioral experiment with inseparable rewards, inseparable model is favored



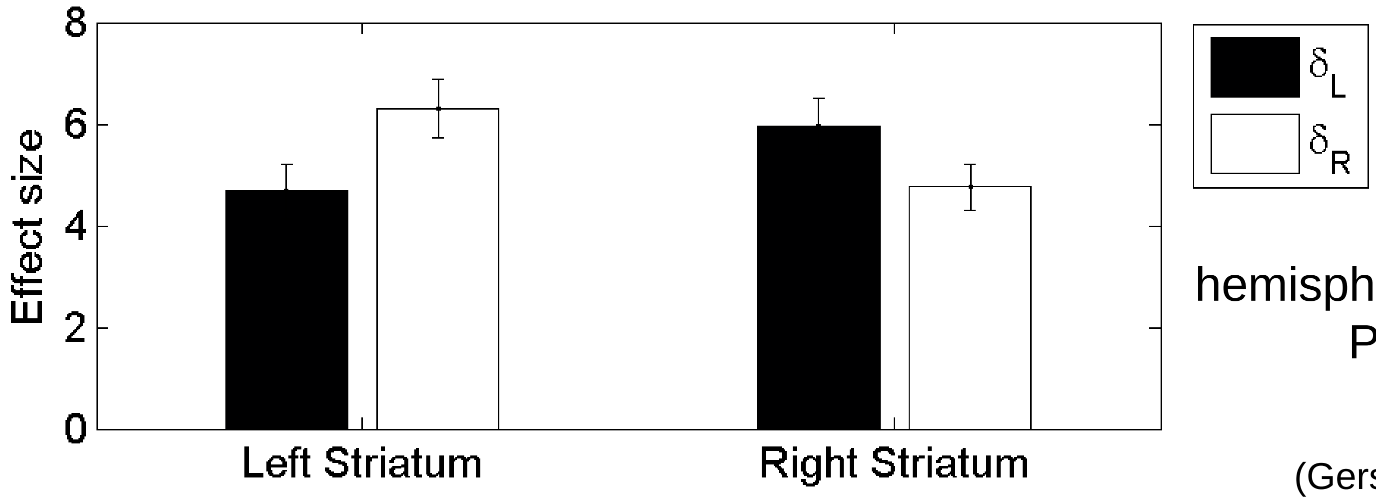
(Gershman et al. 2009)

striatal prediction error

separable task: do PEs decompose?



net PE
$\delta_L + \delta_R$



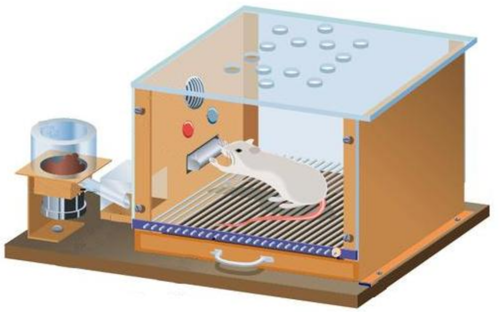
hemisphere x effector
$P<.005$

(Gershman et al. 2009)

summary

- both hierarchy example and multieffector example suggest decomposed, vector-valued error signal
  - this, apparently, can be observed
  - nb also useful in POMDPs
  - cf Frank et al. “Making working memory work”

vigor



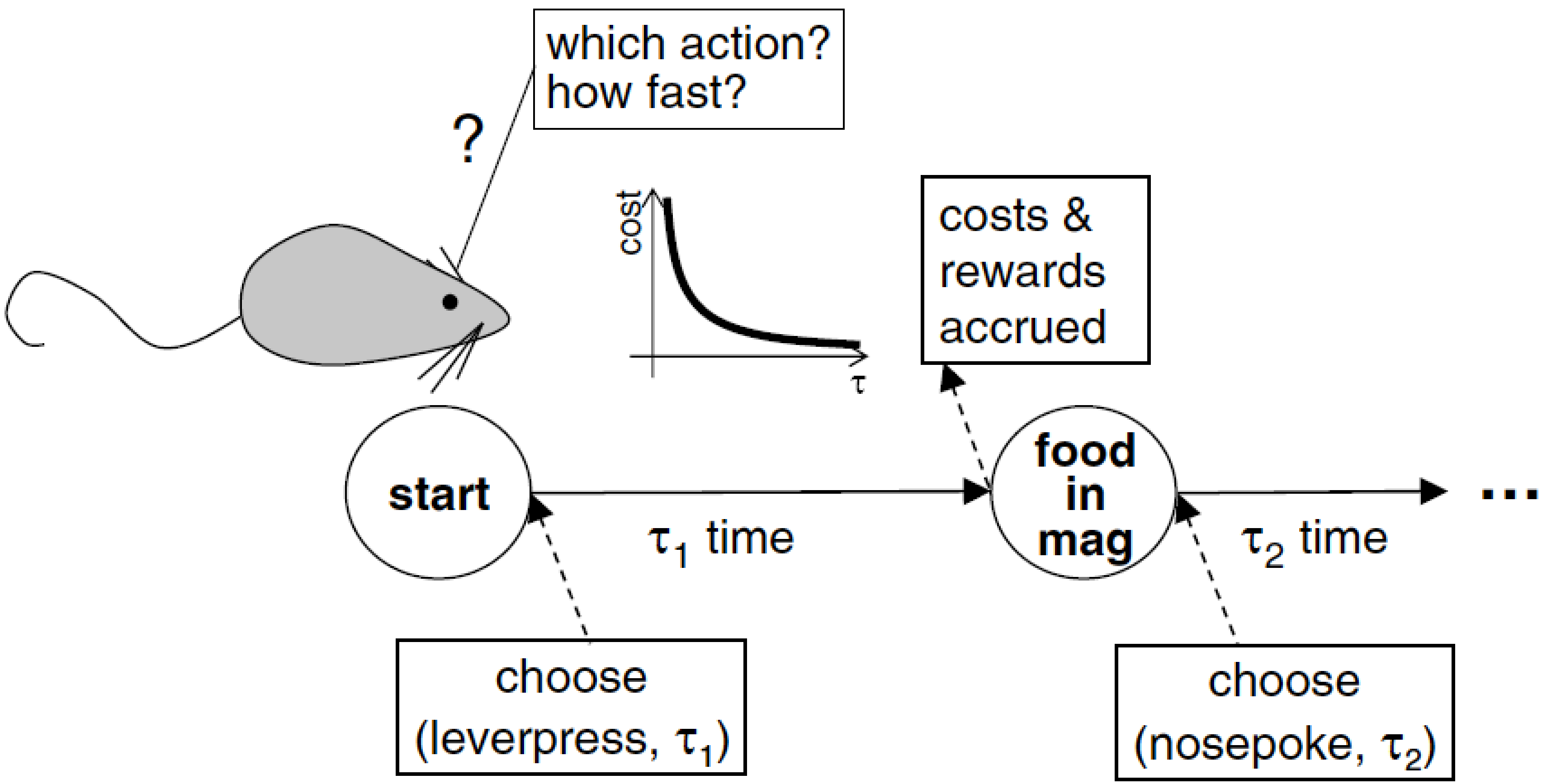
- in many conditioning tasks, dependent variable is vigor (eg rate of leverpressing, speed of running), not a discrete choice

- causal manipulations of dopamine (eg drugs, Parkinson’s disease) have most obvious effects on behavioral activation, hard to attribute to learning
  - suggestion (e.g. Berridge 2008) that dopamine is involved in performance, not learning

- Niv et al (2005, 2007): formulate RL-like optimization problem with choice of action + vigor

Formalism

$Q(S,a,\tau) =$ immediate reward or cost - vigor cost - time cost + expected future rewards minus costs



Formalism

$Q(S,a,\tau) =$ immediate reward or cost - vigor cost - time cost + expected future rewards minus costs

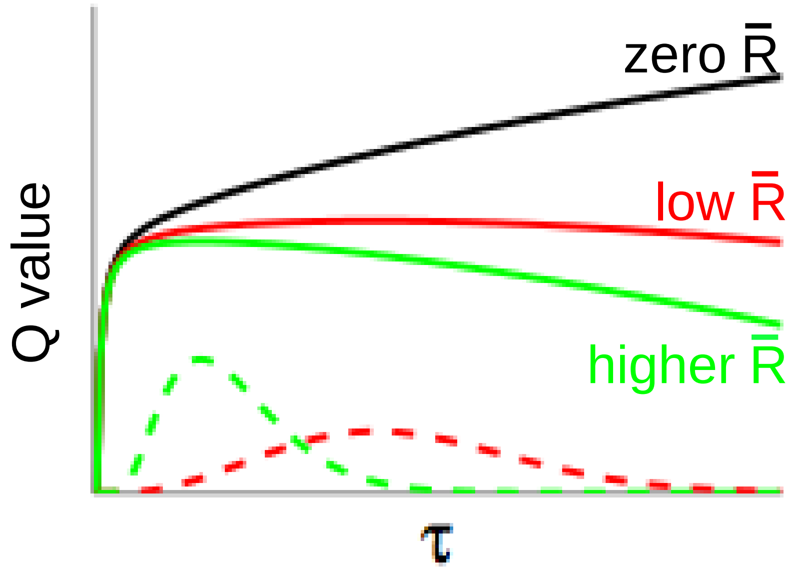
choice of latency:
slow → less vigor cost
but more time cost

from average reward RL time (opportunity) cost is

$$\tau \overline{R}$$

for average reward $\overline{R}$

opportunity cost and vigor

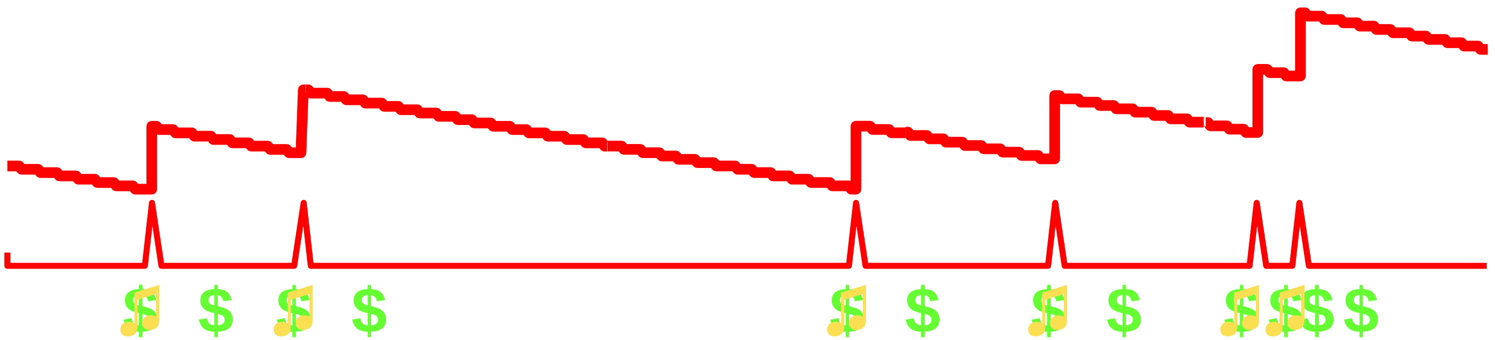


- reward rate determines the "cost of sloth"
- higher rate of reward: pressure on all actions to be faster
- suggests causal control mechanism: track reward rate, energize behavior

the tonic dopamine hypothesis

average prediction error = net reward rate

thus, automagically: dopamine viewed more slowly (tonic DA) could carry $\overline{R}$



→ suggests reconciliation of phasic (reward, teaching) and tonic (activational) functions of DA

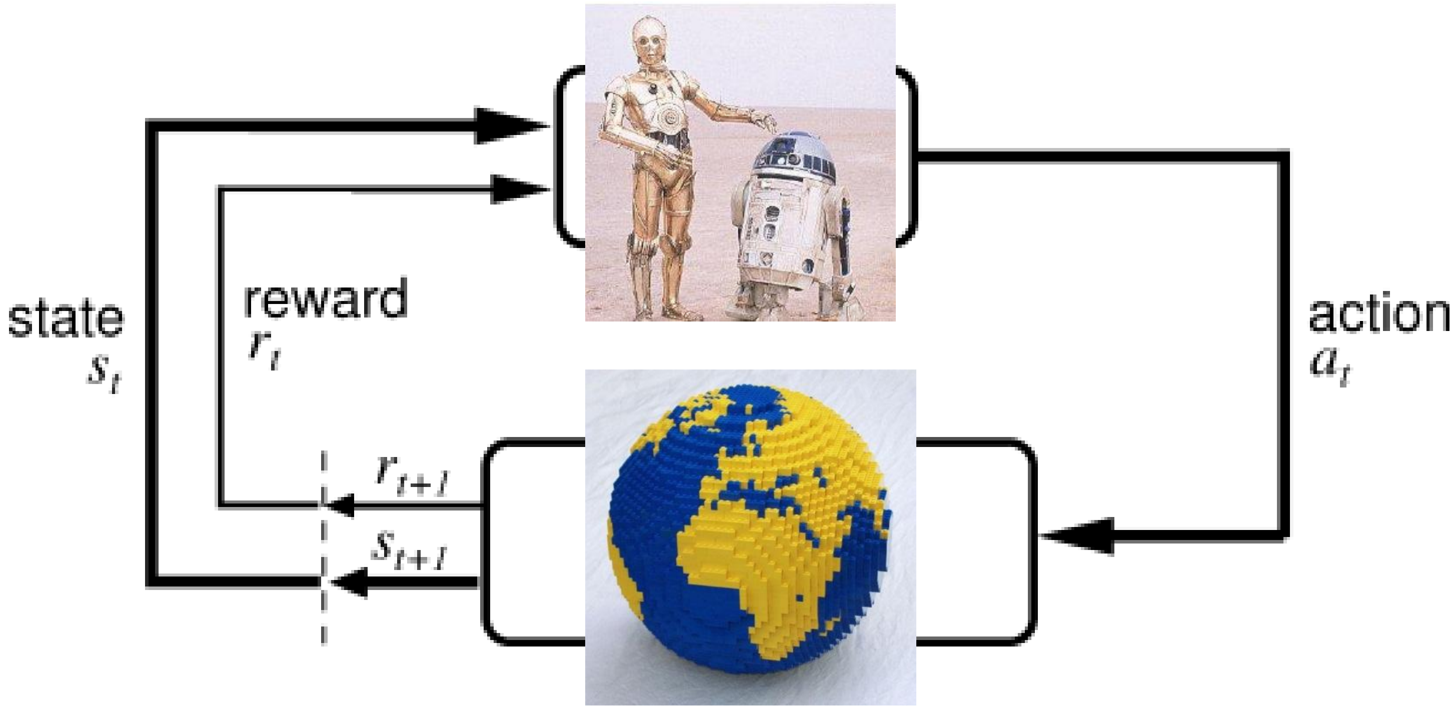
Niv et al 2007

Markov Decision Process

class of stylized tasks with

states, actions & rewards

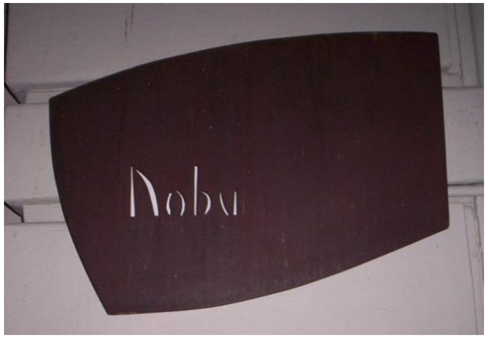
→ what do these correspond to in biology?



rewards

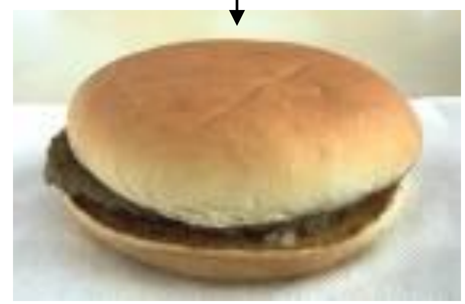
- in RL, rewards are scalars
- in psychology and biology, rewards may serve a number of roles
  – reinforcement (~ model-free RL)
  – goal/incentive (~ model-based RL)

- this relates to a classic disagreement in psychology as to what is learned from reward



Â

Â

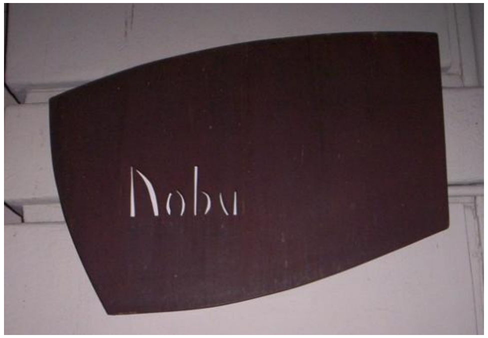


Á

The New York Times

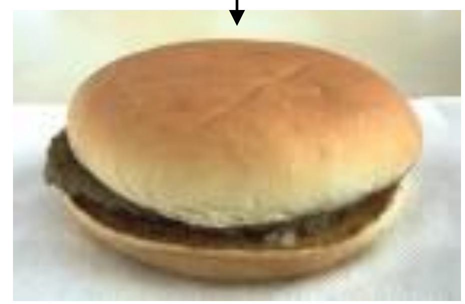
Tainted Fish

Tuna sushi purchased from 20 restaurants and stores in Manhattan
The New York Times in October was tested for mercury. Analysts
examined at least two pieces of sushi from each place and calculat
the level of methylmercury, a form linked to health problems, in part
per million. They then determined how many pieces it would take to
reach what the Environmental Protection Agency calls a weekly
reference dose (RfD), what it considers an acceptable level to be
regularly consumed. (Pieces varied in size.) Figures below are for tl
piece of sushi with the highest level of mercury at each place.



?

Á



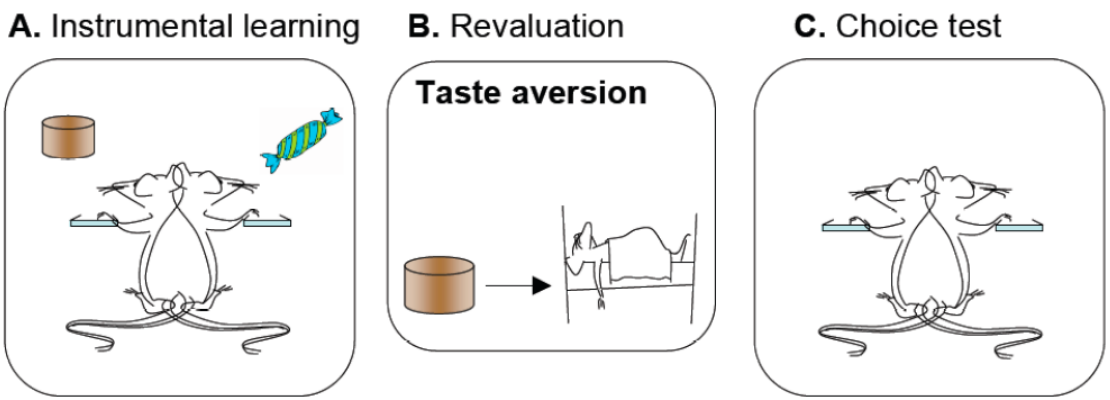
The New York Times

Tainted Fish

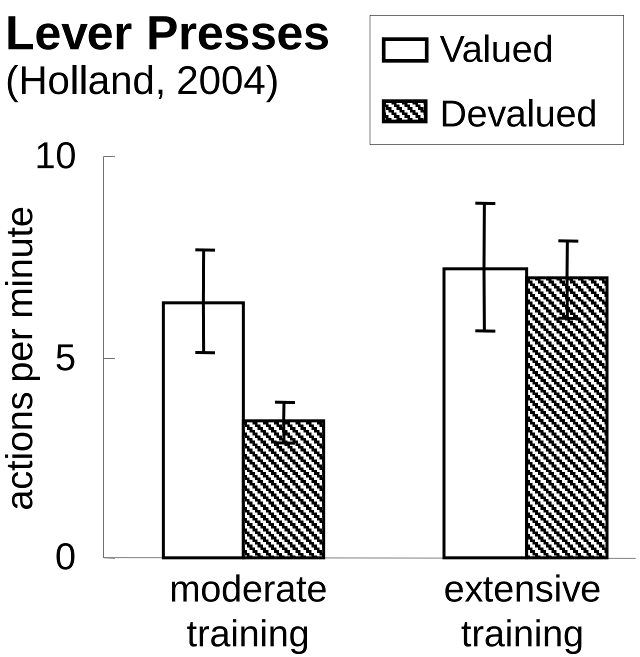
Tuna sushi purchased from 20 restaurants and stores in Manhattan
The New York Times in October was tested for mercury. Analysts
examined at least two pieces of sushi from each place and calculat
the level of methylmercury, a form linked to health problems, in part
per million. They then determined how many pieces it would take to
reach what the Environmental Protection Agency calls a weekly
reference dose (RfD), what it considers an acceptable level to be
regularly consumed. (Pieces varied in size.) Figures below are for th
piece of sushi with the highest level of mercury at each place.

$$E[V(a)] = \Sigma_o P(o|a)\ V(o)$$

rat version



(Balleine, Daw & O'Doherty, 2009)



two behavioral modes:
devaluation-sensitive
("goal directed")

devaluation-insensitive
("habitual", like TD)

→neurally dissociable with lesions
(Dickinson, Balleine, Killcross)

Lesions

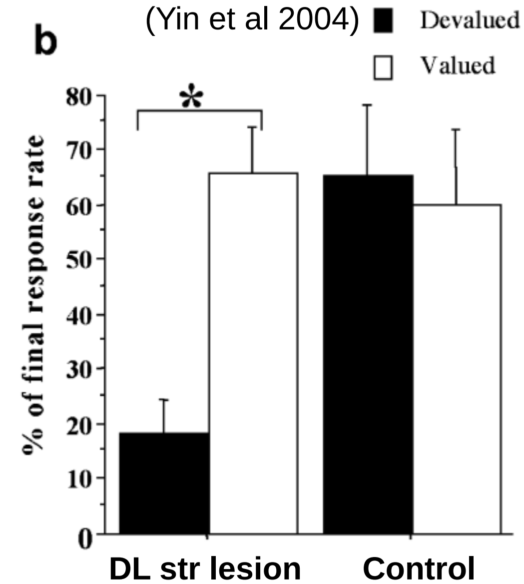
- With lesion of dorsolateral striatum (also its DA input) rats acquire normally but never form habits: perpetually devaluation sensitive

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

→ Behavior arises from dissociable neural systems

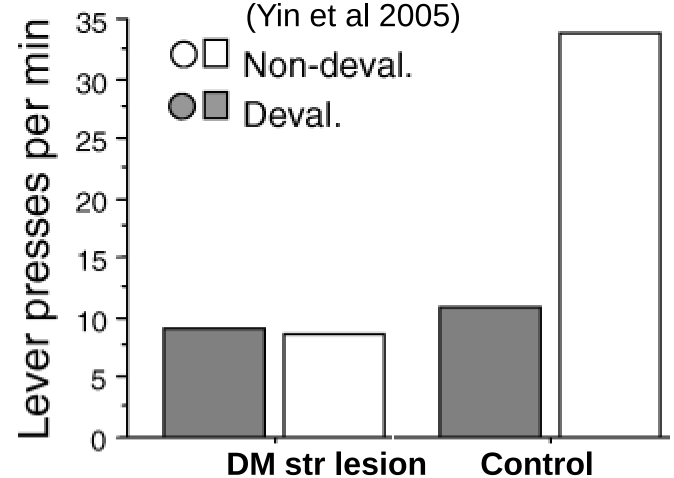
Overtrained

(Yin et al 2004)



Moderate training

(Yin et al 2005)



interim summary

same action (leverpressing) can arise from two behaviorally and neurally distinct systems; can only distinguish with devaluation test

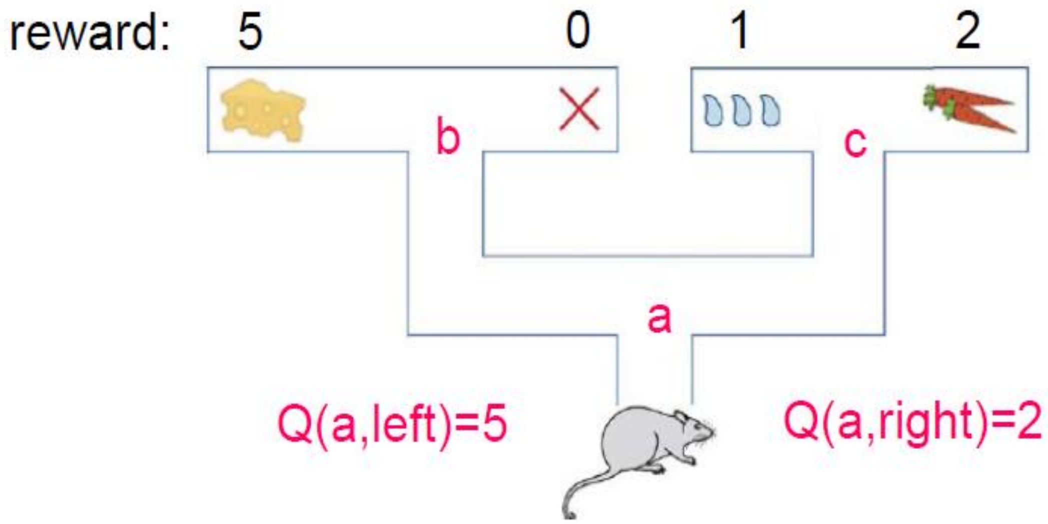
- overtrained leverpressing is devaluation insensitive
  - “habitual”
  - as predicted by temporal-difference & S→R models
  - this is closely associated with what we think dopamine does

- moderately trained leverpressing is devaluation sensitive
  - “goal directed”
  - demonstrates animals represent outcome internally
  - this is probably nondopaminergic (?)

→ possible to knock out either system with lesion; the other one takes over
  - parallel loops each involving areas of cortex and striatum
  - suggests really parallel neural systems: multiple action systems?
  - why is this such a crazy idea?
  - what problems does this create?

- how do we think about all this in terms of RL?
- is there an RL account for goal directed behavior?

Reinforcement learning

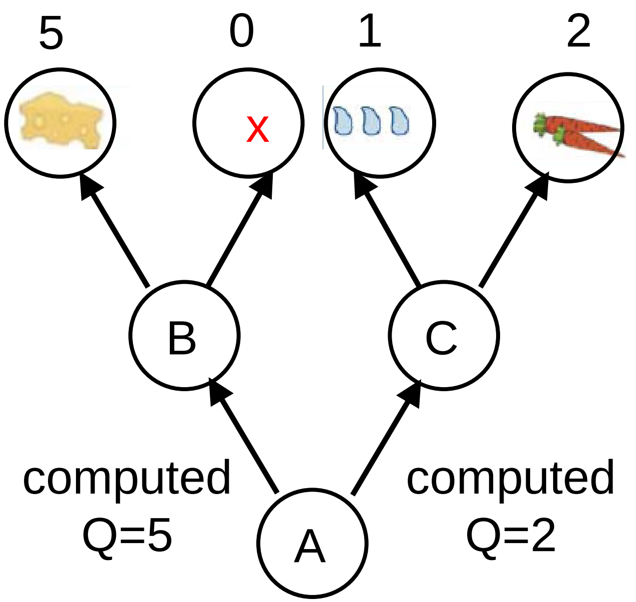


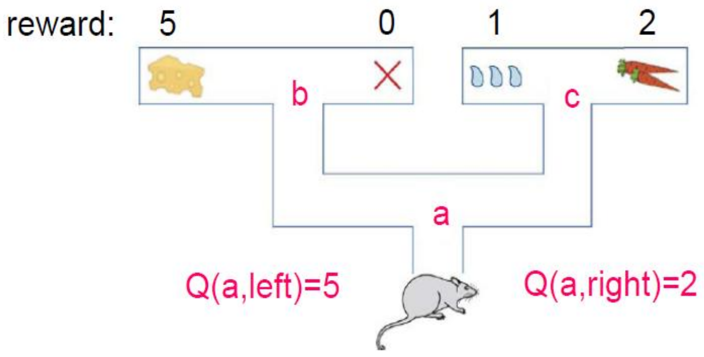
- must learn about long term consequences of actions to choose the best

→how to build these up from past morsels of short term experience? there are different strategies

Approach 1: Model-based RL

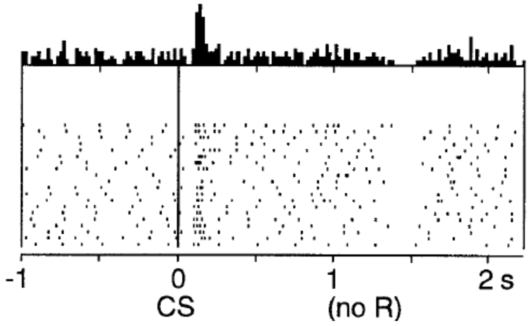
- Learn a “model”of problem…
  - state-action-state transitions
  - state-reward mappings
  - like a “cognitive map”of task (need not be spatial)
- …and you can iteratively search it to forecast long-term value of an action
  - dynamic programming
- obviously hard online, in large problmes

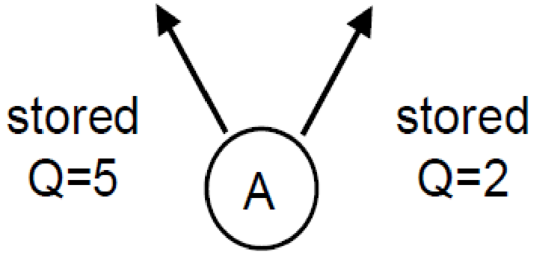




Approach 2: Model-free RL (TD)

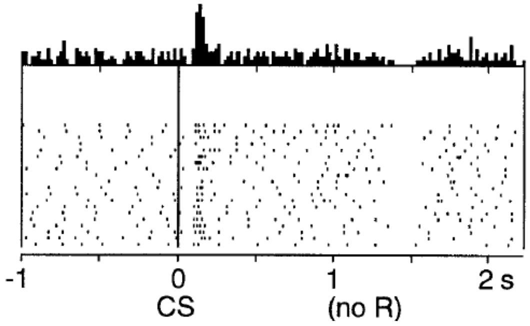
- Shortcut: store long-term values
  - then simply retrieve them to choose the best
- you can learn these directly from experience
  - without building or searching a model
  - by incremental "sampling"

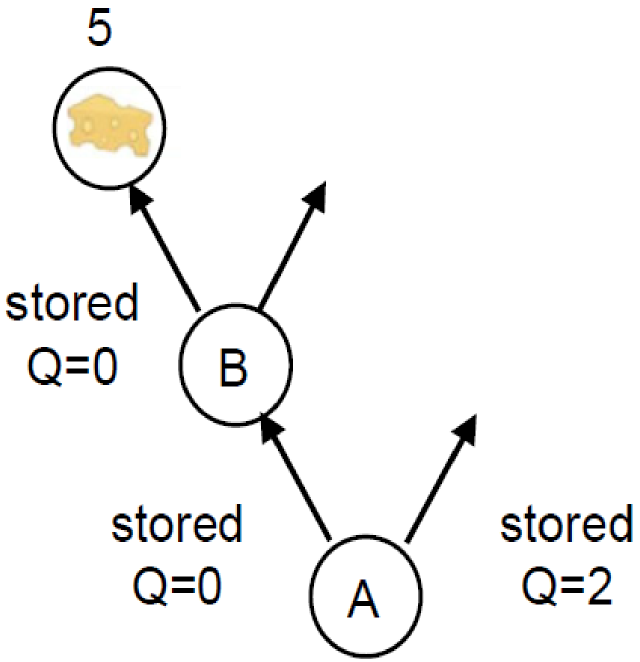




Approach 2: Model-free RL (TD)

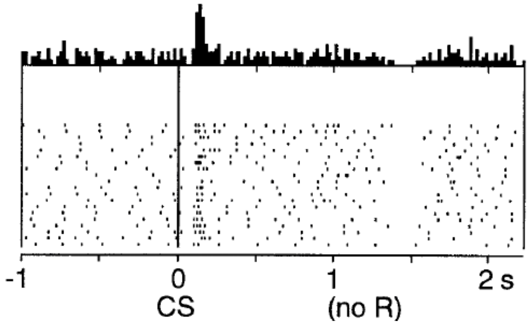
- Shortcut: store long-term values
  - then simply retrieve them to choose the best
- you can learn these directly from experience
  - without building or searching a model
  - by incremental “sampling”

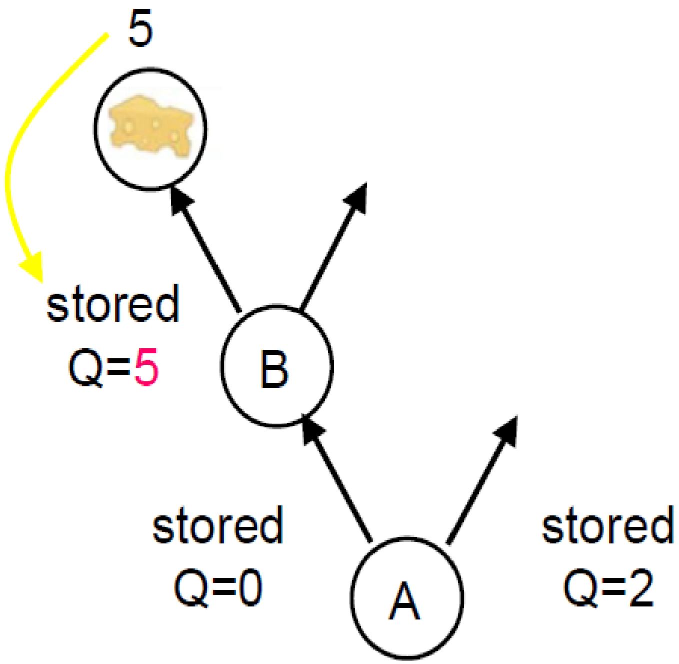




Approach 2: Model-free RL (TD)

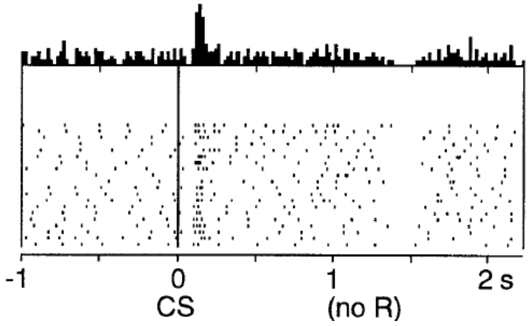
- Shortcut: store long-term values
  - then simply retrieve them to choose the best
- you can learn these directly from experience
  - without building or searching a model
  - by incremental “sampling”

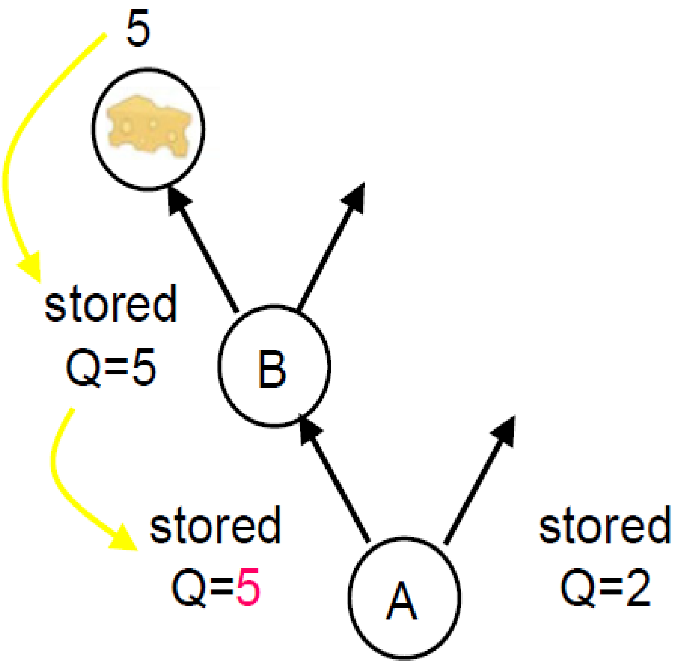




Approach 2: Model-free RL (TD)

- Shortcut: store long-term values
  - then simply retrieve them to choose the best
- you can learn these directly from experience
  - without building or searching a model
  - by incremental "sampling"

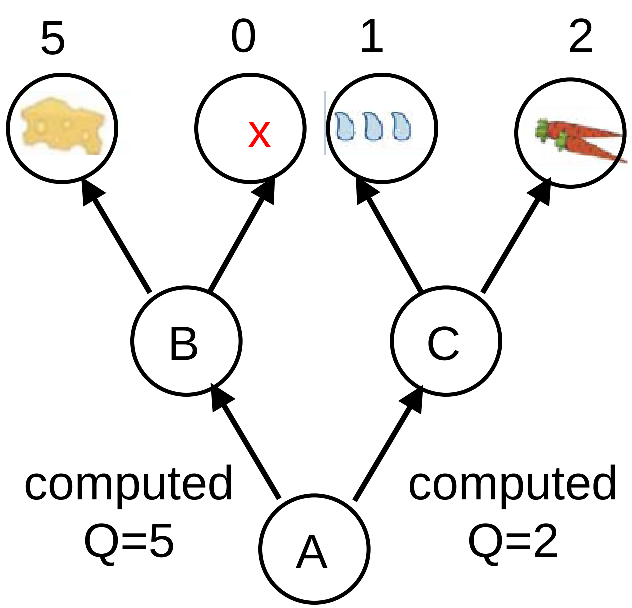


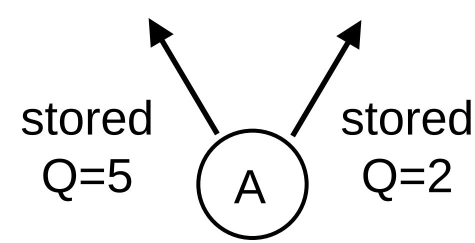


outcome sensitivity

model-based:
can immediately adapt to value shifts
like goal-directed

model-free:
cannot immediately adapt
like habits

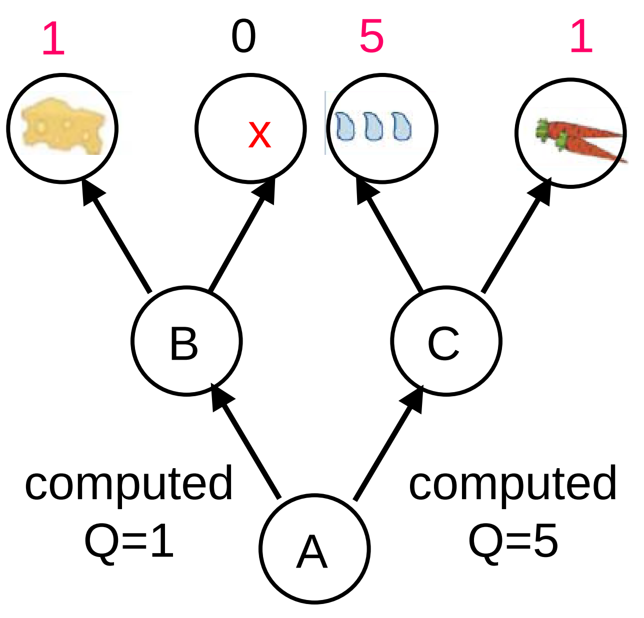


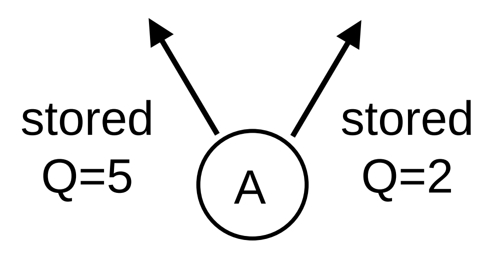


outcome sensitivity

model-based:
can immediately adapt to value shifts
like goal-directed

model-free:
cannot immediately adapt
like habits





additionally

- how to trade off these approaches online (meta-control)?
- Daw et al. 2005: basic tradeoff between cost vs accuracy of model-based search explains a lot of data (like overtraining effect); formalized with uncertainty

- parallels in behavioral economics (Ho & Camerer; Hampton et al.): do you approach multiplayer interactions by learning model of opponents + best responding?

- a lot of ongoing work trying to separate model-based vs model-free signaling in the brain
  - emerging finding: more integrated than expected

from nips to neuroscience

reinforcement learning exemplifies two (related) ways that computer science informs behavioral neuroscience

1. conceptual
   - how to characterize hard problems (formally analyzable tasks)
   - optimal (typically intractable) solution
   - approximate algorithms and their properties
   - define relevant quantities
   - → algorithms as hypotheses
   - → common process level explanation for different kinds of data

2. analytical
   - algorithms as likelihood functions for inference from data
   - data analysis as statistical machine learning

for further information

me: daw@cns.nyu.edu

don't miss Rangel talk at main meeting

Reviews of RL & the brain

- Niv (2009), Reinforcement learning in the brain, The Journal of Mathematical Psychology
- Maia (2009), Reinforcement learning, conditioning, and the brain: successes & challenges, Cognitive Affective and Behavioral Neuroscience
- Balleine, Daw, & O'Doherty (2008), Multiple forms of value learning and the function of dopamine, in Neuroeconomics
- Dayan & Niv (2008), Reinforcement learning and the brain: The Good, The Bad and The Ugly, Current Opinion in Neurobiology
- Doya (2008), Modulators of decision making, Nature Neuroscience

RL for data analysis

- Daw (2010), Trial by trial data analysis using computational models, in Attention and Performance 23
- 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