Reinforcement learning in the brain

Nathaniel Daw

ndaw@princeton.edu

outline

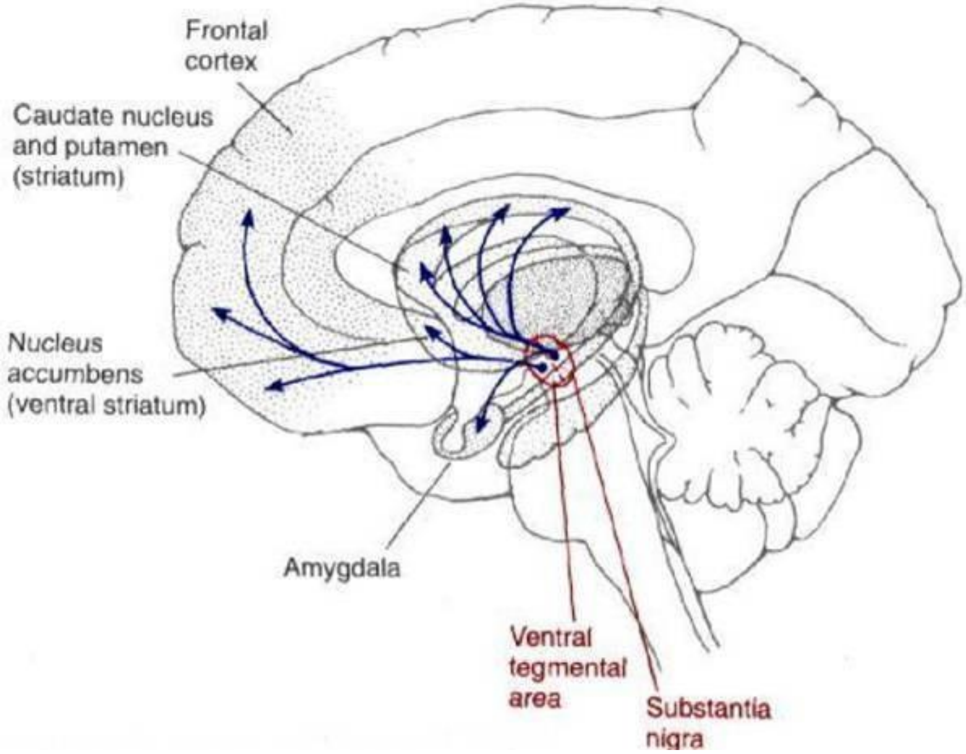
Estimating action values: model-based vs. model-free learning

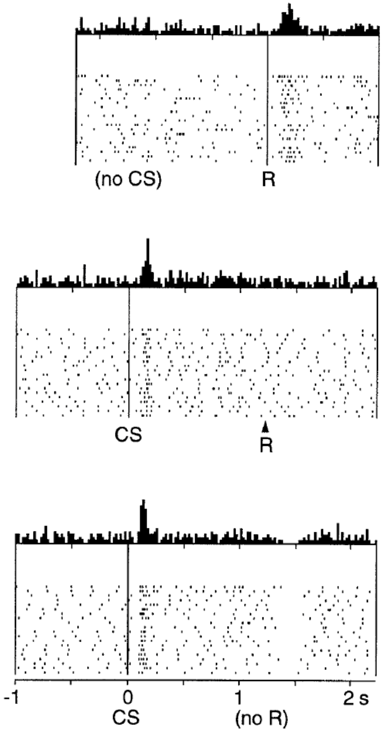
1. Intro: dopamine and credit assignment
2. Examples
   - habits and instrumental reward devaluation
   - rodent spatial navigation
   - RL in humans; compulsion
3. Hippocampal replay and planning

outline

Estimating action values: model-based vs. model-free learning

1. Intro: dopamine and credit assignment
2. Examples
    - habits and instrumental reward devaluation
    - rodent spatial navigation
    - RL in humans; compulsion
3. Hippocampal replay and planning


Frontal cortex
Caudate nucleus and putamen (striatum)
Nucleus accumbens (ventral striatum)
Amygdala
Ventral tegmental area
Substantia nigra



(Schultz et al 1997)

Error driven learning: $V_t \leftarrow V_t + \alpha\,(r_t\ (+\ V_{t+1}\ ) - V_t)$

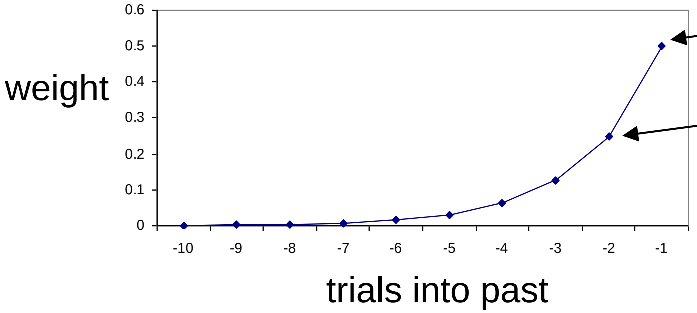
Equivalently: $= \alpha\, r_t + (1 - \alpha)\ V_t$

$= \alpha\, r_t + \alpha\,(1 - \alpha)\ r_{t-1} + \alpha\,(1 - \alpha)^2\ r_{t-2} + \ldots$

the delta rule estimates its expected reward using a weighted running average of rewards received during stimuli

recent trials are weighted more strongly (steepness determined by 1-α)

- why does this make sense?



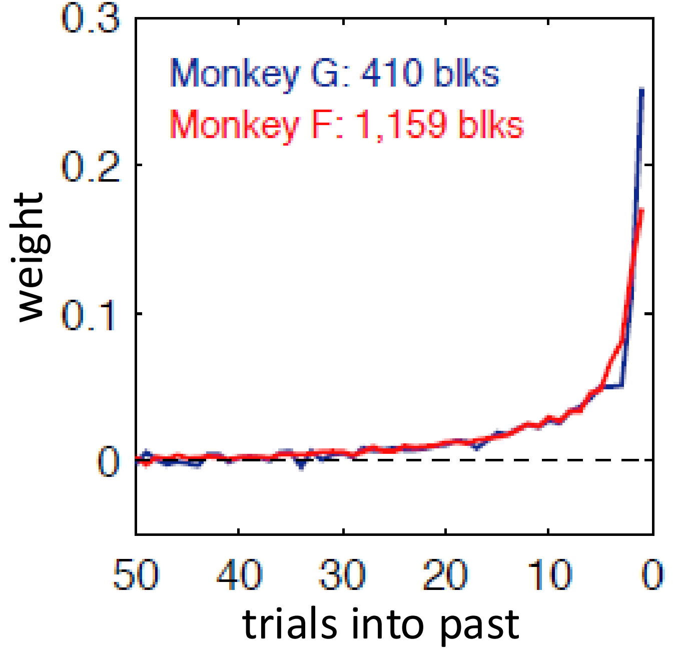
weight for last trial

weight for trial before that

etc

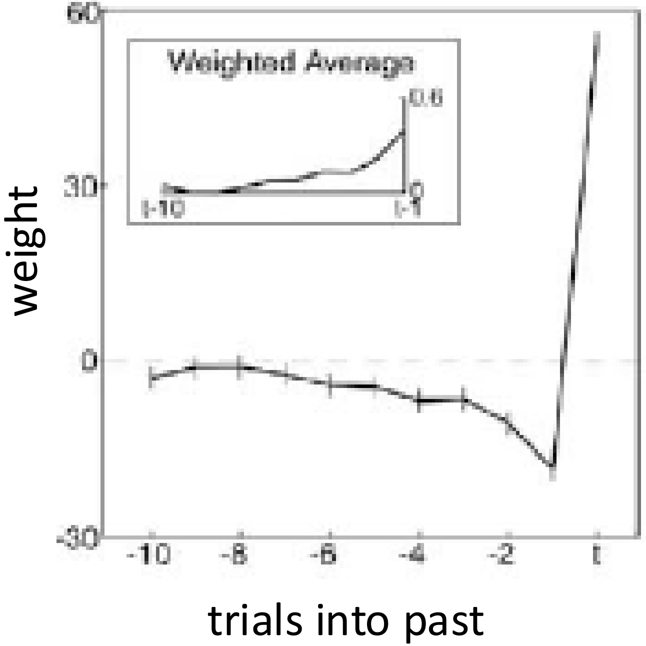
error-driven estimation

monkeys' choices



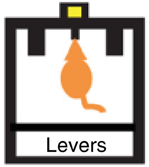
(Sugrue et al. 2004)

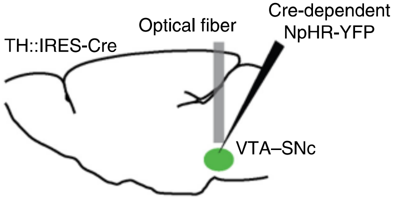
dopamine neurons



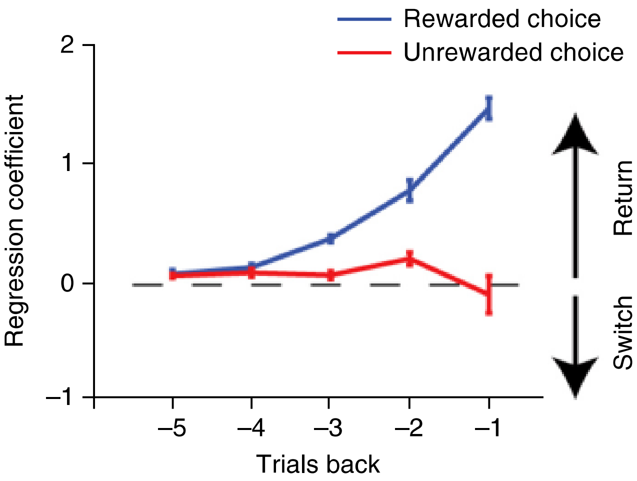
(Bayer and Glimcher 2005)

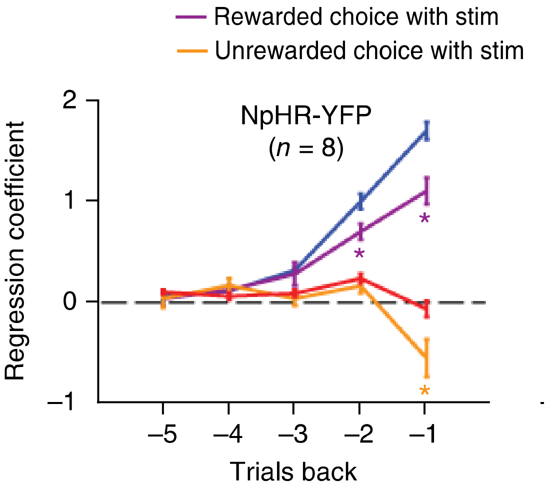
Causality: DAergic reinforcement in mice





timed suppression of dopamine neurons on 10% of trials





(Parker et al., Nature Neuroscience 2016)

temporal difference learning

Temporal-difference learning (Sutton & Barto):

Want $V(s_t) = r(s_t) + r(s_{t+1}) + r(s_{t+2}) + \ldots$

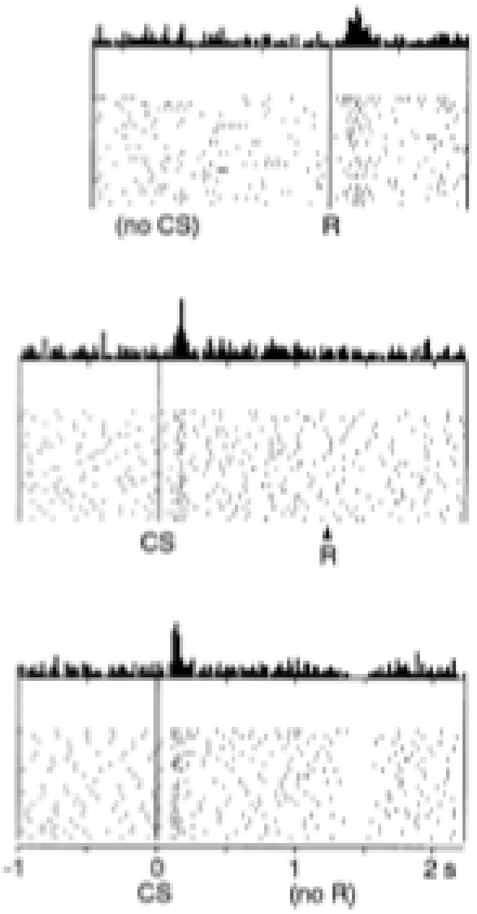
$= r(s_t) + V(s_{t+1})$

Use prediction error $\delta_t = [r(s_t) + V(s_{t+1})] - V(s_t)$

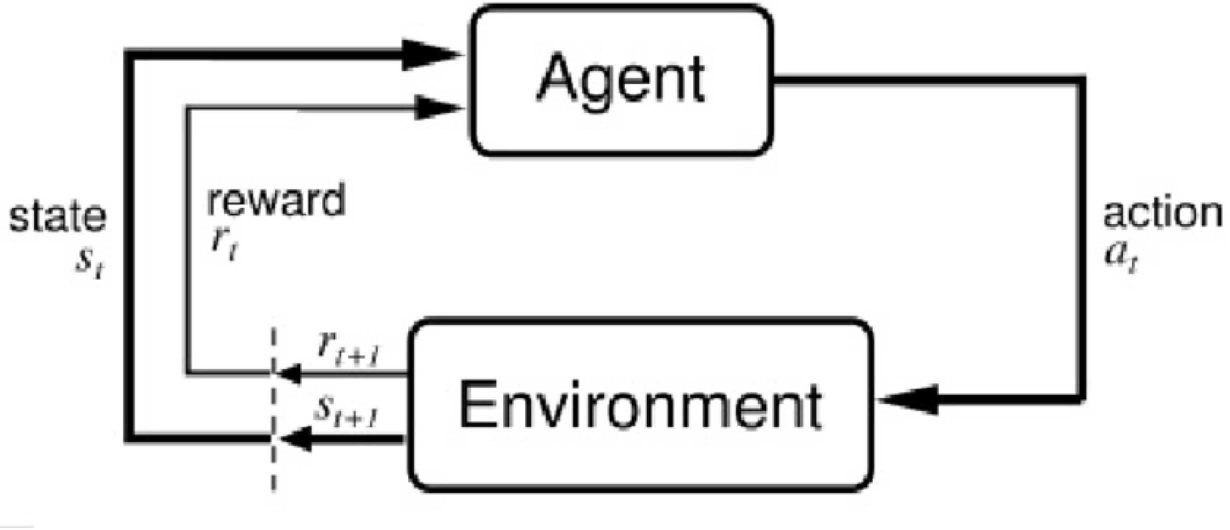
- learn to predict cumulative future rewards $r(s_t) + r(s_{t+1}) + r(s_{t+2}) + \ldots$
- learn using what I predict at time t+1 ($V(s_{t+1})$) as stand in for all future rewards
    - so I don't have to wait forever to learn
    - at t+1 I learn what is s_{t+1} (remember, this can be unexpected)

- learn consistent predictions based on temporal difference $V(s_{t+1}) - V(s_t)$
    - if $V(s_{t+1}) = V(s_t)$, my predictions are consistent
    - if $V(s_{t+1}) > V(s_t)$, things got unexpectedly better
    - if $V(s_{t+1}) < V(s_t)$, things got unexpectedly worse

→ and these act like reward to generate prediction error and learning



The setting



Agent

state s_t

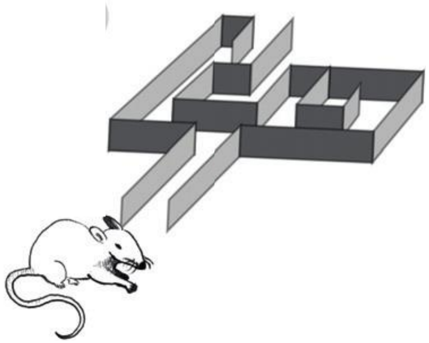
reward r_t

action a_t

r_{t+1}

s_{t+1}

Environment

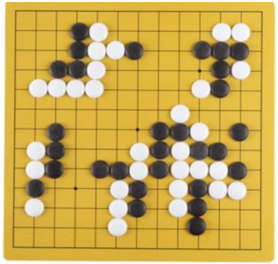


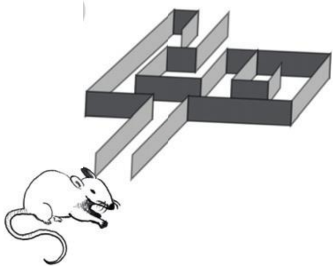
Trial and error learning in sequential tasks, where choices lead to more choices

- Maximize long-term objective (expected total points; chance of final win)
- ”Value function”: expected cumulative, discounted reward

What makes this difficult?

$$Q(s_t, a_t) = r(s_t) + \sum_{s_{t+1}} P(s_{t+1}|s_t, a_t) \left[r(s_{t+1}) + \sum_{s_{t+2}} P(s_{t+2}|s_{t+1}, a_{t+1}) [r(s_{t+2}) + \cdots] \right]$$





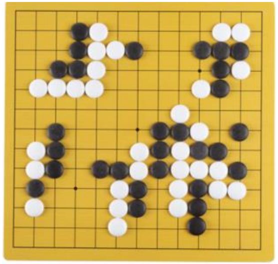
Outcomes of actions are delayed & contingent

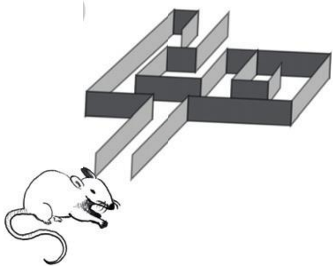
- choice requires connecting actions to consequences nonlocally over space and time
  → “planning,” “mental simulation”
  → “credit assignment”

- hard to learn by trial and error
- hard even to compute given full knowledge

The objective function

$$Q(s_t, a_t) = r(s_t) + \sum_{s_{t+1}} P(s_{t+1}|s_t, a_t)\left[r(s_{t+1}) + \sum_{s_{t+2}} P(s_{t+2}|s_{t+1}, a_{t+1})[r(s_{t+2}) + \cdots]\right]$$





expected cumulative (discounted) future reward
→ … over "tree" of future states (nested sums)
→ This is hard to compute, even if you know the one-step contingencies
→ Knowing it reduces choice to comparison

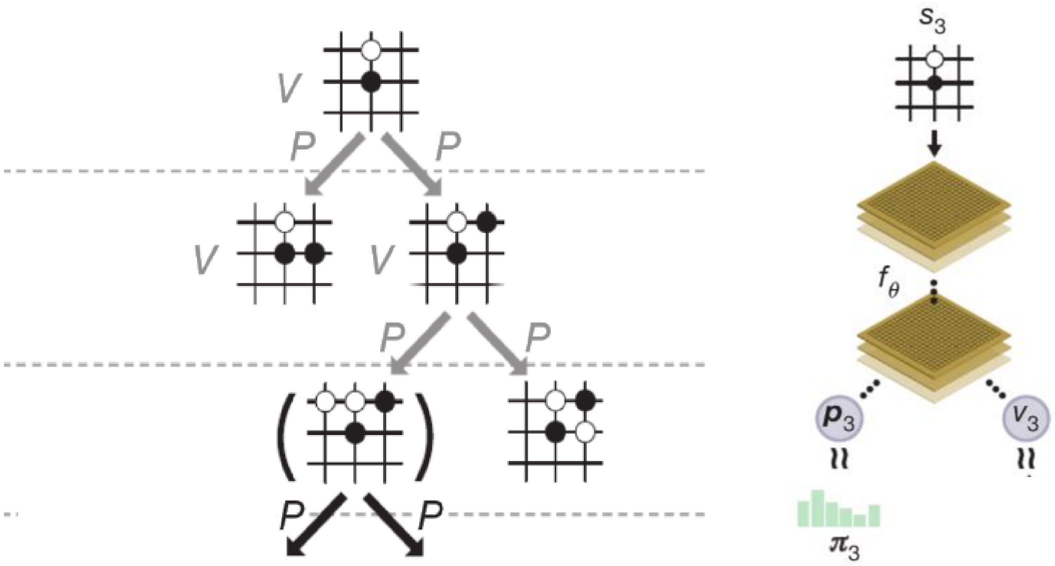
How do we estimate this (particularly in trial-and-error learning)?
→ two predominant approaches in AI

“model-based” learning

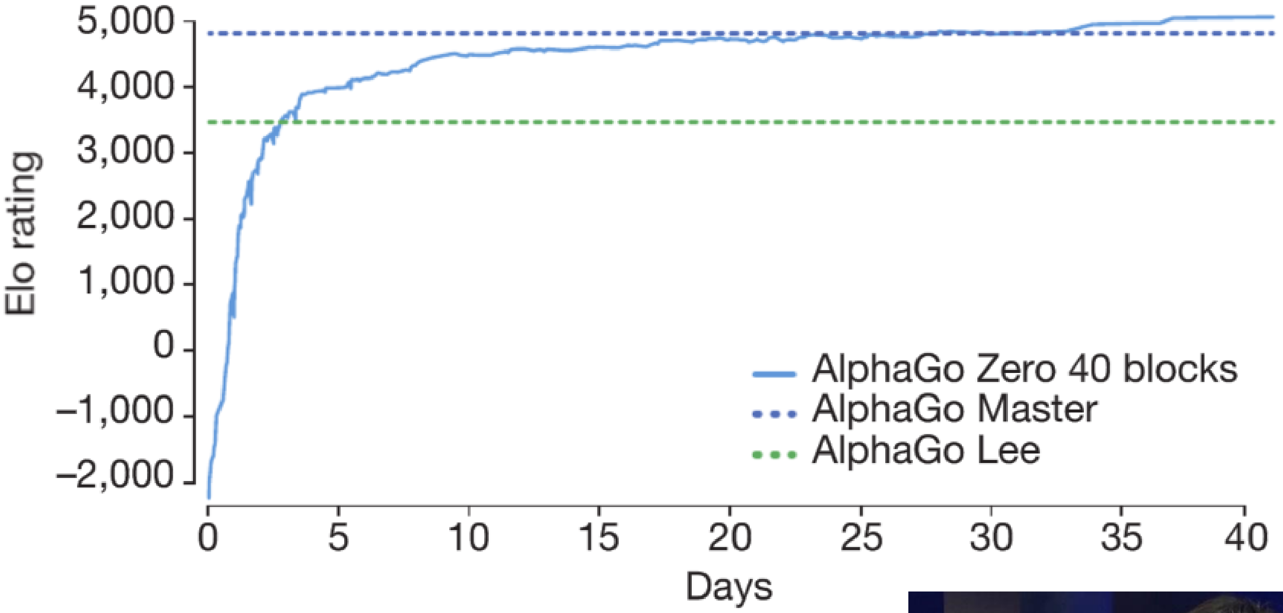
$$Q(s_t, a_t) = r(s_t) + \sum_{s_{t+1}} P(s_{t+1}|s_t, a_t) \left[r(s_{t+1}) + \sum_{s_{t+2}} \dots \right]$$

- Easy part: learn one-step reward $P(r_t|s_t)$ & transition “map” $P(s_{t+1}|s_t, a_t)$
  - (why is this easy?)

- Hard part: iterative, tree-structured computation at choice time;
  - … like mental simulation
  - example: AlphaGo



(Silver et al, 2017)

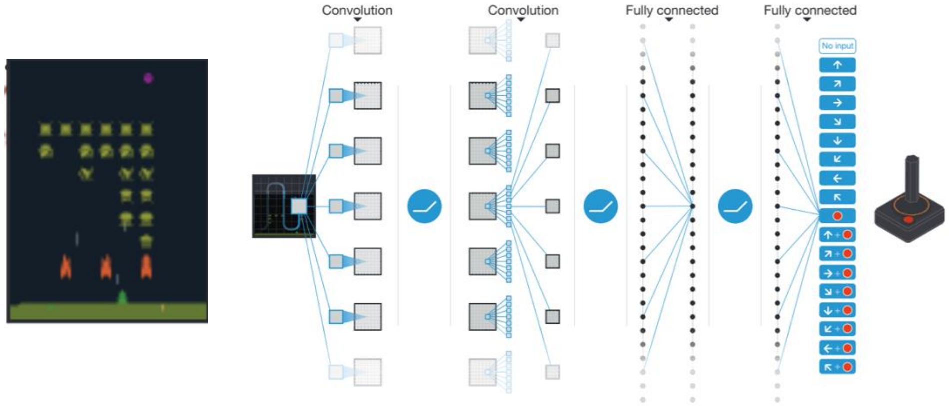


Lee Sedol (human master): ~3500

(Silver et al, 2017)

“model-free” learning

$$Q(s_t, a_t) = r(s_t) + \sum_{s_{t+1}} P(s_{t+1}|s_t, a_t)\left[r(s_{t+1}) + \sum_{s_{t+2}} \ldots\right]$$



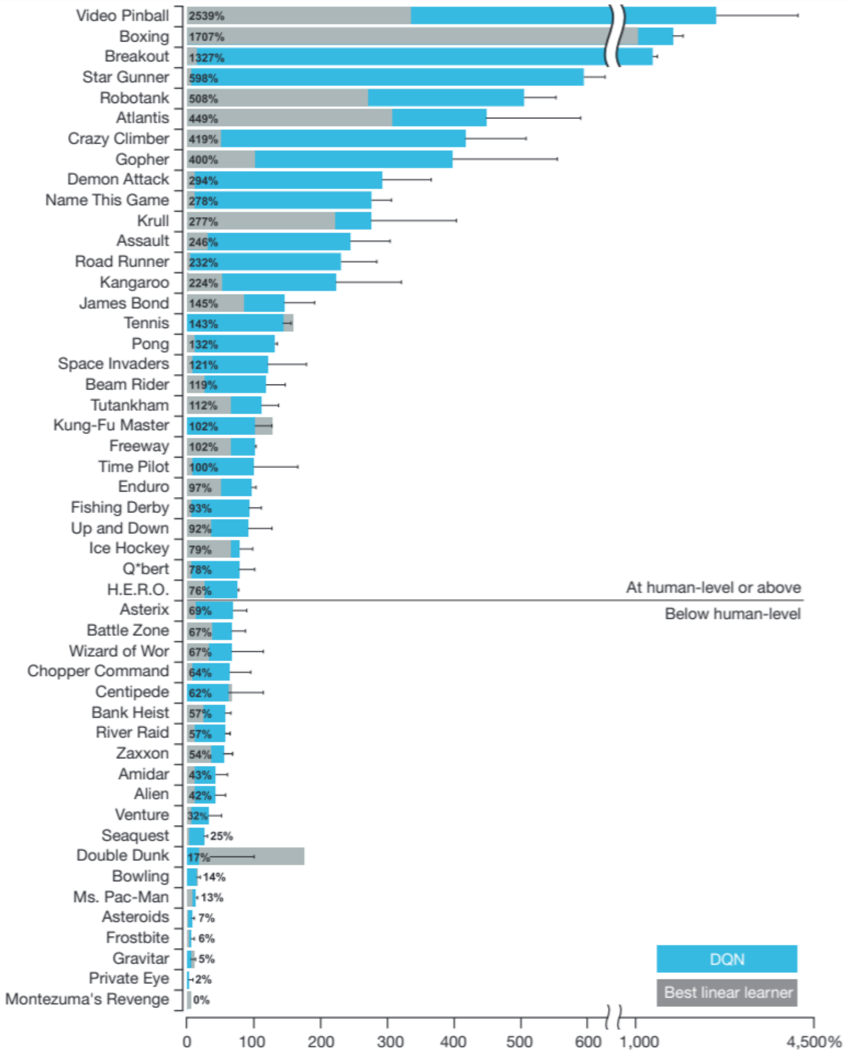
shortcut: store endpoints of computation (long-run action values)

- these can be learned directly from experience, ”model free” (TD learning)

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

- simplifies choice-time computation (just retrieve)
- example: DeepMind Atari “Deep Q Network”

(Mnih et al., 2015)

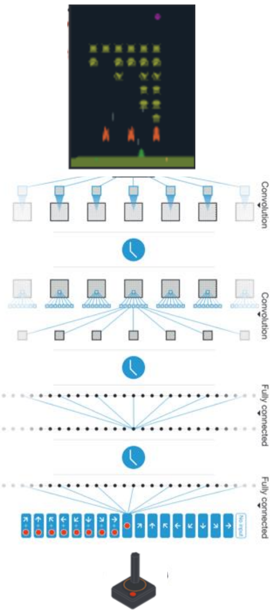


Mnih et al, 2015

Model-based and model-free learning

$$Q(s_t, a_t) = r(s_t) + \sum_{s_{t+1}} P(s_{t+1} | s_t, a_t) \left[r(s_{t+1}) + \sum_{s_{t+2}} \dots \right]$$

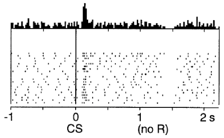
DQNs



(Mnih et al 2015)

← "Model-free" learning

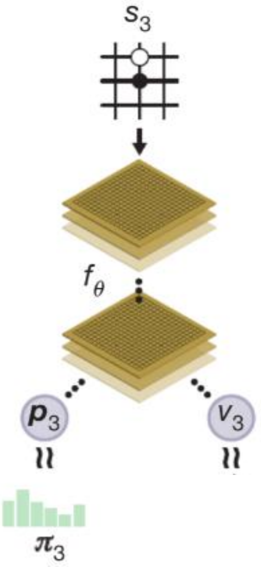
- brain: dopamine, prediction errors
- behavior: habits, slips of action



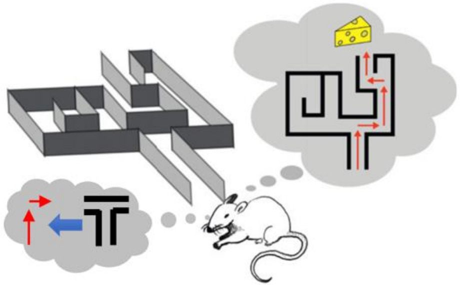
"Model-based" learning →

- brain: anticipatory activity e.g. spatial paths in hippocampus
- behavior: flexible planning

AlphaGo



(Silver et al 2017)



Idea (Daw ea 2005): the brain implements both approaches in parallel

outline

Estimating action values: model-based vs. model-free learning

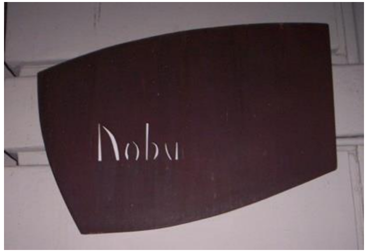
1. Intro: dopamine and credit assignment
2. Examples
    - habits and instrumental reward devaluation
    - rodent spatial navigation
    - RL in humans; compulsion
3. Hippocampal replay and planning

MF learning

Idea: brain learns long-run action values Q experientially chooses by comparing them

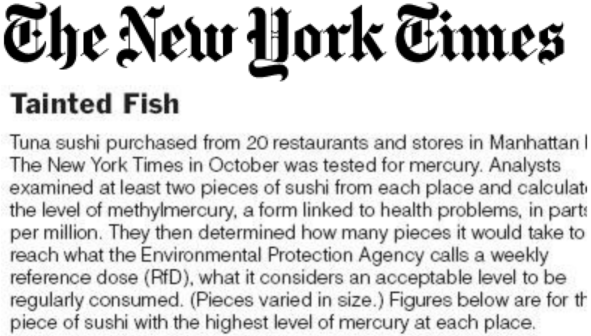
- Behavioral idea: goes back to Thorndike “law of effect”
- Neural idea: dopamine, prediction errors, temporal difference learning (Schultz, Dayan, Montague)

Weird prediction: if decision variable is scalar summary of previous experiences, animals should be blind to certain changes in task contingencies (until they relearn action values from experience)



?

<



The New York Times

Tainted Fish

Tuna sushi purchased from 20 restaurants and stores in Manhattan l 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.

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

"model-free"

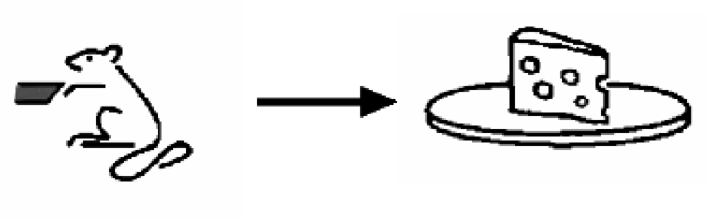
"model-based"

(Daw et al. 2005)

Classic test for MB vs MF

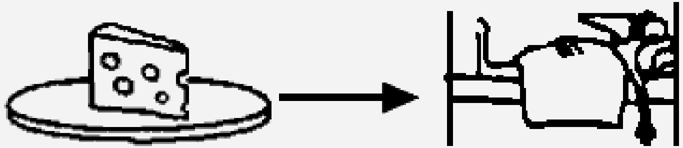
Stage

1. training
(hungry)



learn to leverpress
for food

2. devaluation

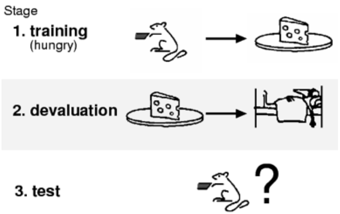


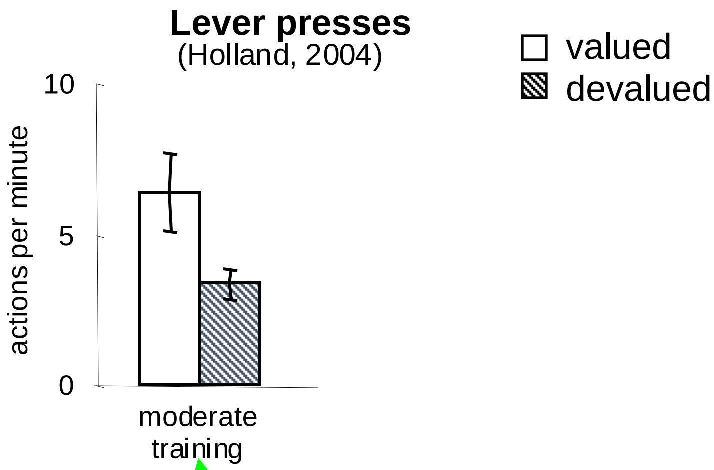
pair food with illness;
develop aversion
control: no pairing

3. test

will animals work for
food they don't want?
(compared to animals
who skipped stage 2)

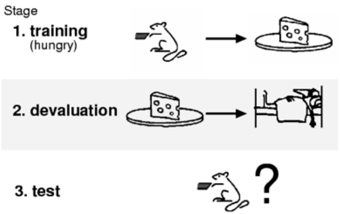
results

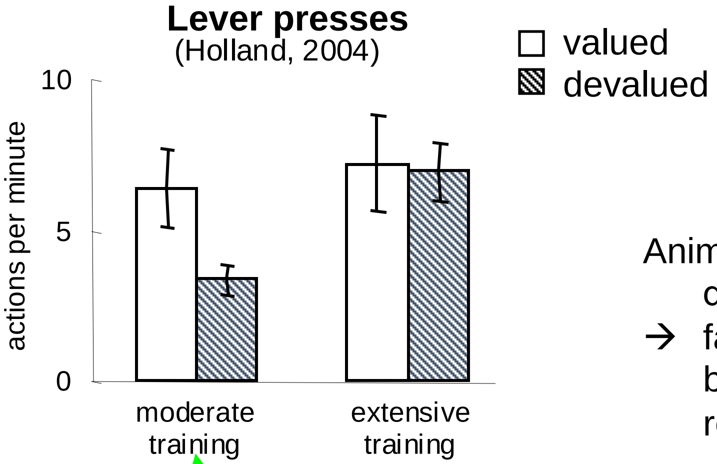




Moderate training: outcome sensitive
“goal directed”, like MB

results





Animals will work for food they don’t want, sometimes

→ familiar counterpart: actions become automatic with repetition

Moderate training: outcome sensitive
“goal directed”, like MB

Outcome insensitive following overtraining
“habitual” like MF

Lesions

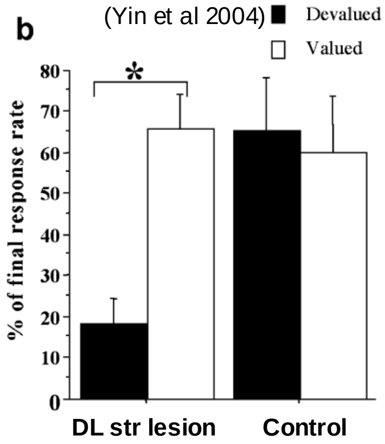
Lesions to different networks appear to differentially disable these modes of behavior

- Dorsolateral striatum loop: perpetually devaluation sensitive (never form habits)

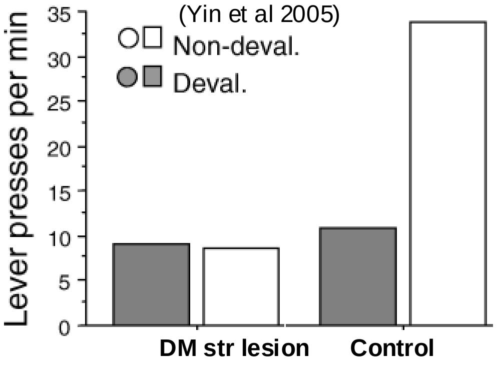
- PFC-dorsomedial striatum loop: animals: always devaluation insensitive (no MB stage)

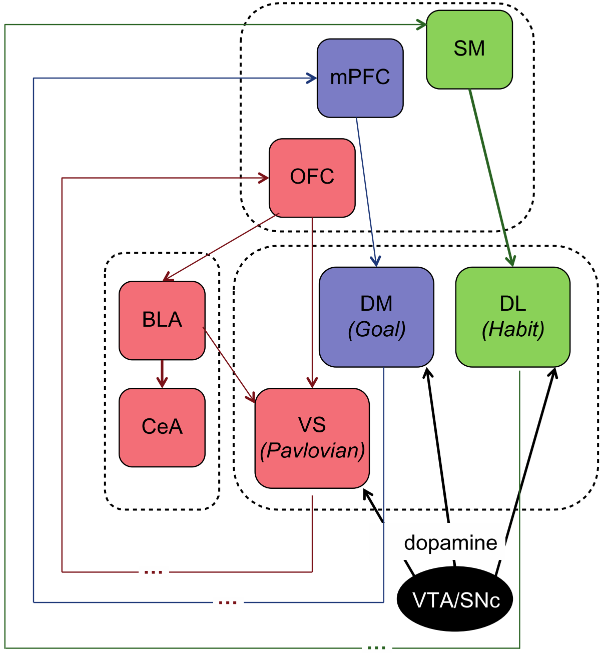
→Behavior arises from dissociable neural systems

Overtrained



Moderate training



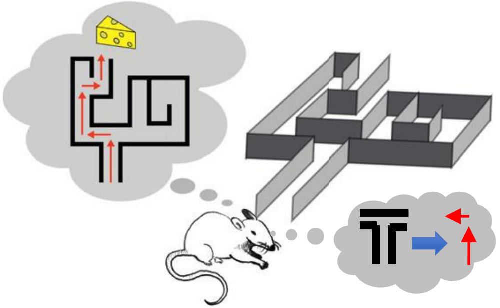


(Daw & O'Doherty 2013)

rational dual-system arbitration

Interest in dual-system architectures for healthy & disordered behavior

- Healthy: automaticity, habits, slips of action, self-control, willpower
- Dysfunction: compulsion, drugs of abuse (eg Everitt & Robbins, 2005)
    - hope to ground symptoms of mental illness in basic mechanisms



implied question: arbitration / control

idea: cost-benefit think vs. act tradeoff

- deliberation costly (delay); when is it likely to benefit: improve choice, earn more reward?
- e.g.: not usually worthwhile for highly practiced actions in stable environment
- For math see Keramati et al (PLoS CB 2011)
- cost-benefit arbitration captures many factors affecting habits in rodents

(Daw, Niv & Dayan, *Nature Neuroscience* 2005)

outline

Estimating action values: model-based vs. model-free learning

1. Intro
2. Examples
    - habits and instrumental reward devaluation
    - rodent spatial navigation
    - RL in humans; compulsion
3. Hippocampal replay and planning

nonlocal credit assignment by rodents

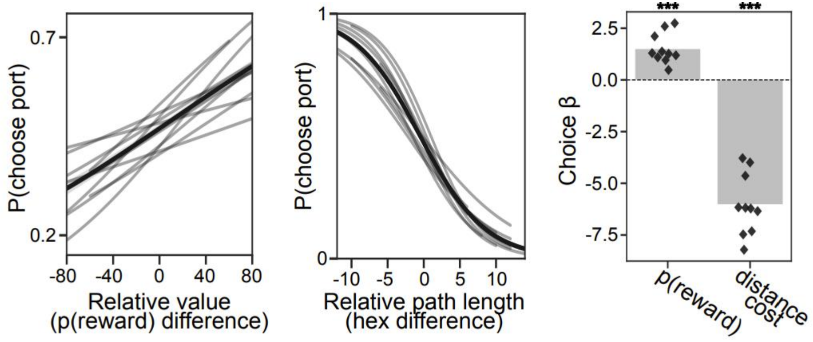
Rats receive stochastic rewards at corners

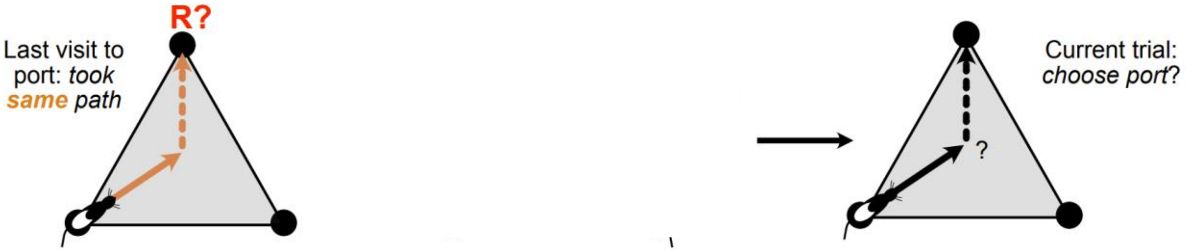
- repeatedly choose next corner balancing reward probability and distance
- continually learn facing periodic changes to barriers or outcome probabilities

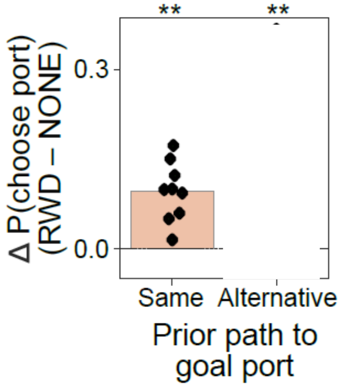
(Krausz, Comrie, Kahn, Frank, Daw & Berke, *Neuron* 2023)

choice in the hex maze

- Choices balance reward probability & distance
- this must be learned.



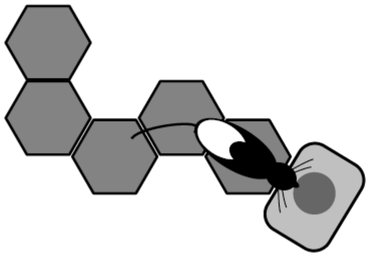

R?
Last visit to port: took same path
Current trial: choose port?
?

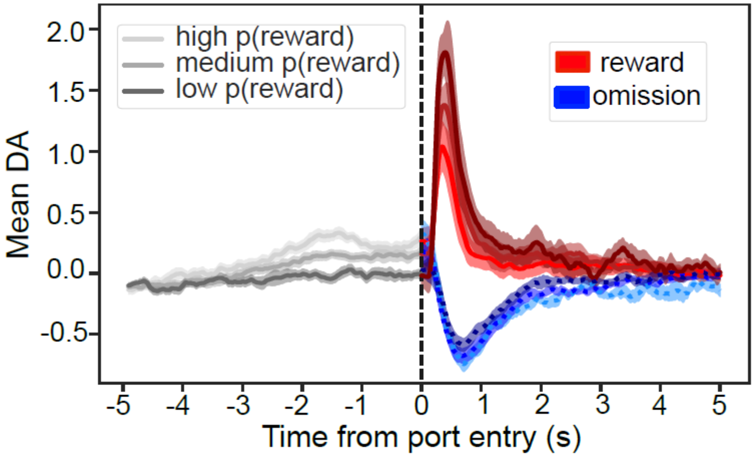


- Outcome (R=0/1) at A affects animals' next A vs. B choice (long-range credit assignment)
- … and next C vs A choice also affected (off-trajectory credit assignment)

(Krausz, Comrie, Kahn, Frank, Daw & Berke, *Neuron* 2023)

history: TD, dopamine



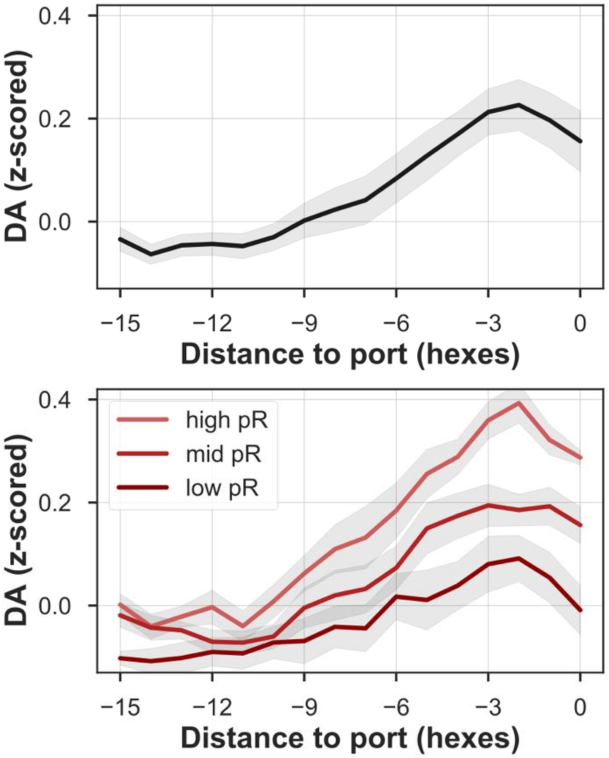


(Krausz, Comrie, Frank, Daw & Berke, bioRxiv 2023)

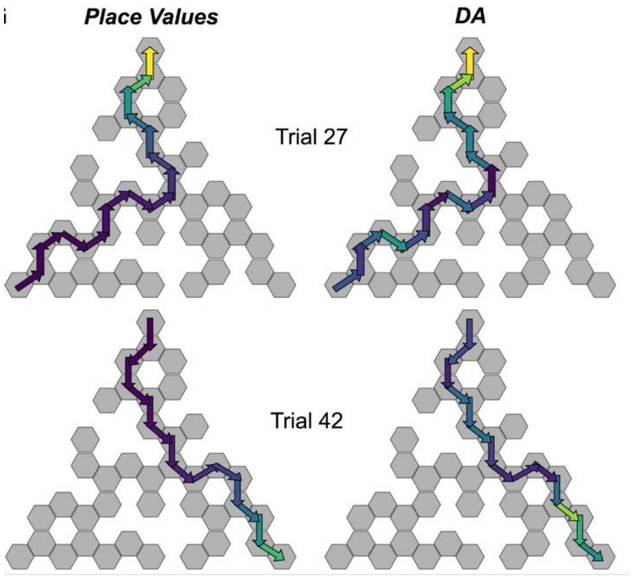
classic work (Montague, Dayan, Schultz): phasic DA responses carry reward prediction error signal

- including to reward predictors, theoretically linked to chaining value backward along repeatedly experienced paths
- but does value really spread this way? *unclear!*
- & is such experiential, model-free learning enough to explain behavior? *no!*

measuring the value function neurally

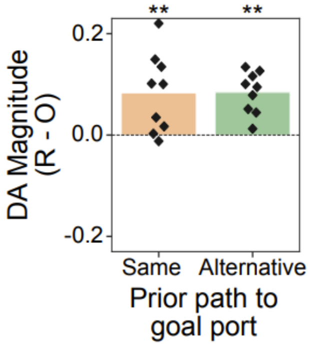


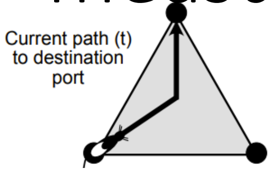
- Between phasic events, DA ramps up
- appears to track instantaneous value function, even on single trials

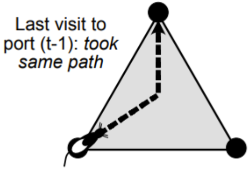


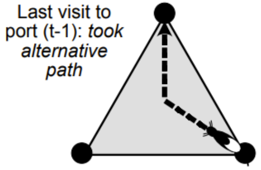
(Krausz, Comrie, Kahn, Frank, Daw & Berke, *Neuron* 2023)

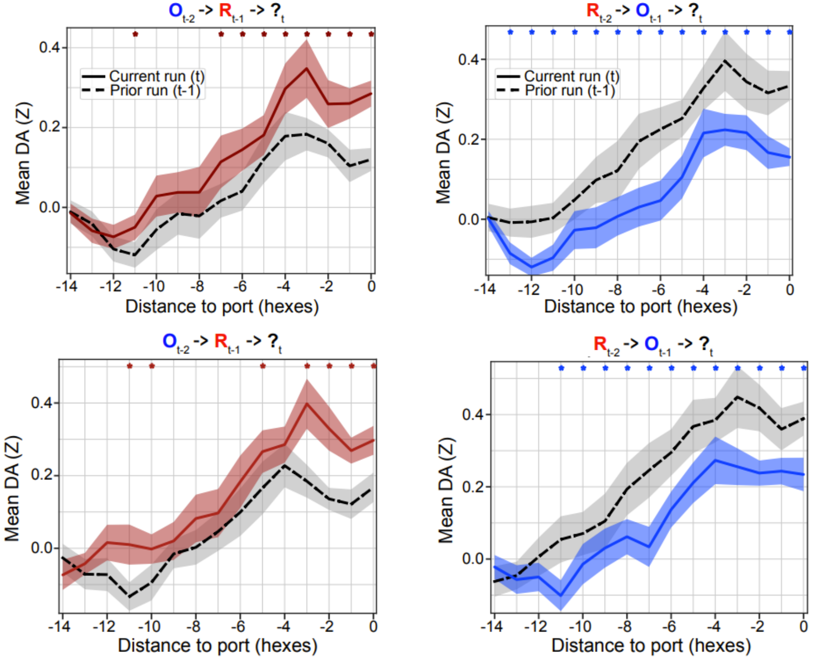
measuring value update neurally




Current path (t) to destination port


Last visit to port (t-1): *took same path*


Last visit to port (t-1): *took alternative path*



Reward propagates long-distance over experienced trajectory

And also does so similarly over nonlocal trajectory (model-based?)

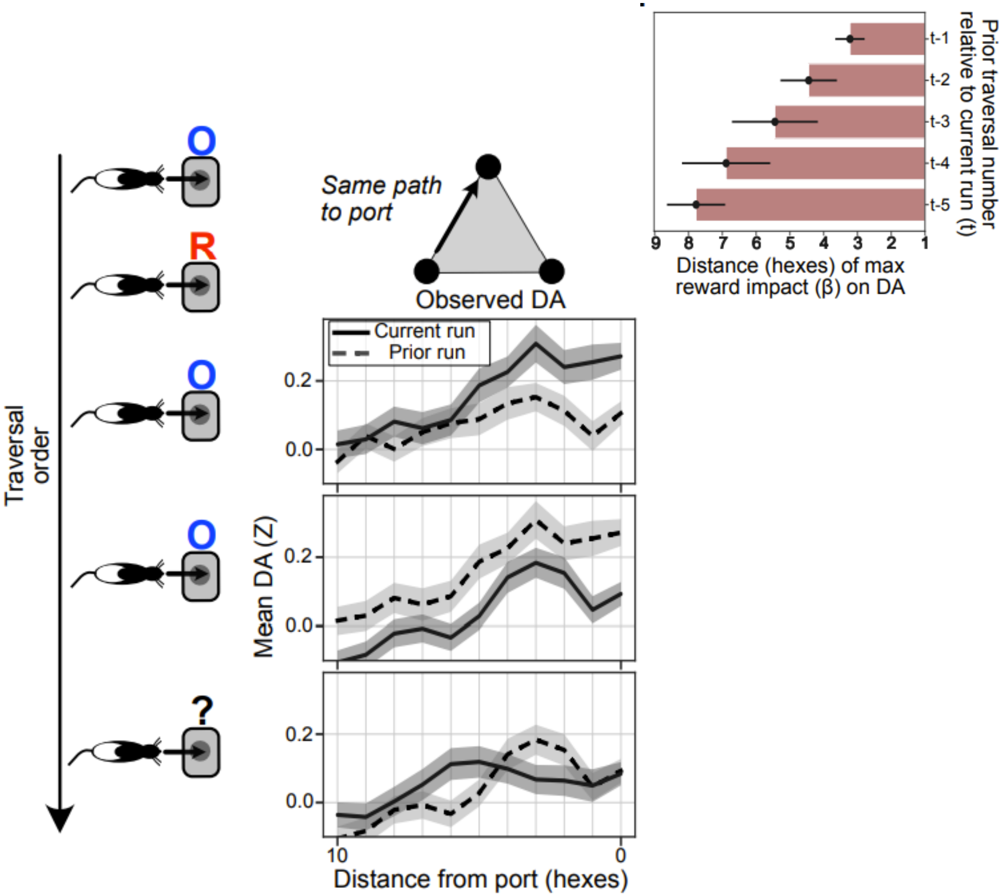
(Krausz, Comrie, Kahn, Frank, Daw & Berke, *Neuron* 2023)

TD-0 like effects also

- Can see “bumps” from individual rewards propagating backward along paths

- (clearly not doing the work for the large scale behavior!)

- we think this is distinct update mechanism from the long-term ramps (not just TD-lambda)



(Krausz, Comrie, Kahn, Frank, Daw & Berke, *Neuron* 2023)

summary

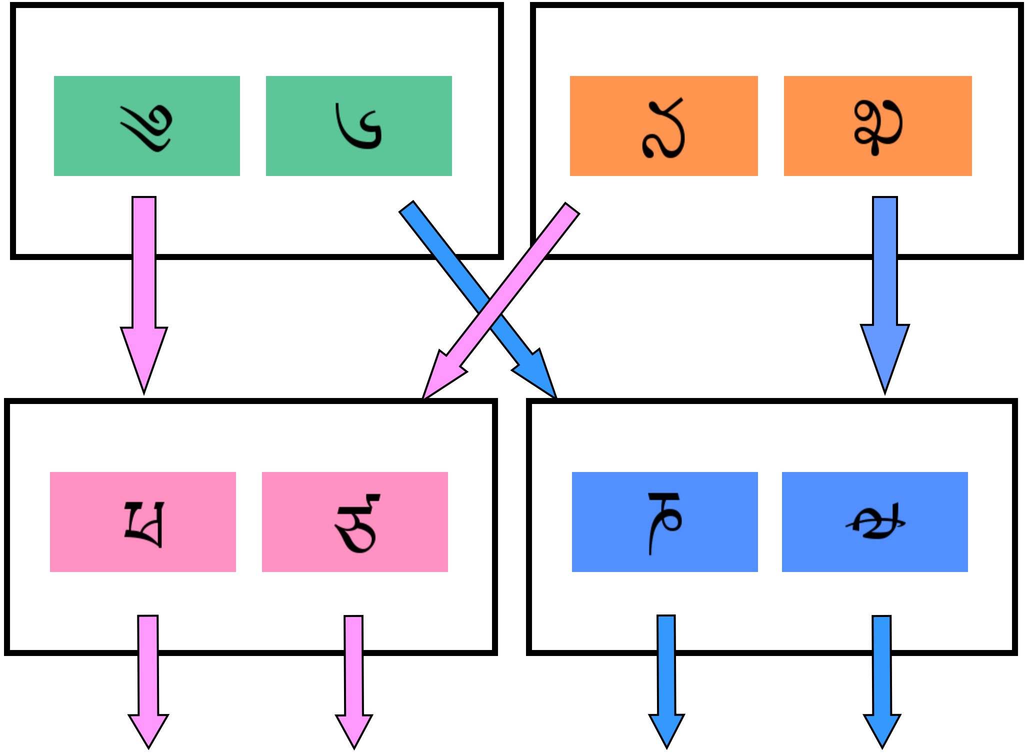
- using dopamine, can directly visualize credit assignment over space
- can see TD(0) chaining but in addition to that
    - value (and choice) affected on next trial at long distance
    - not just over experienced paths (model-based?)
    - these are reflected in choice behavior

outline

Estimating action values: model-based vs. model-free learning

1. Intro
2. Examples
   - habits and instrumental reward devaluation
   - rodent spatial navigation
   - RL in humans; compulsion
3. Hippocampal replay and planning

sequential decision task



with prob: 26% 57% 41% 28%

(*all slowly changing*)

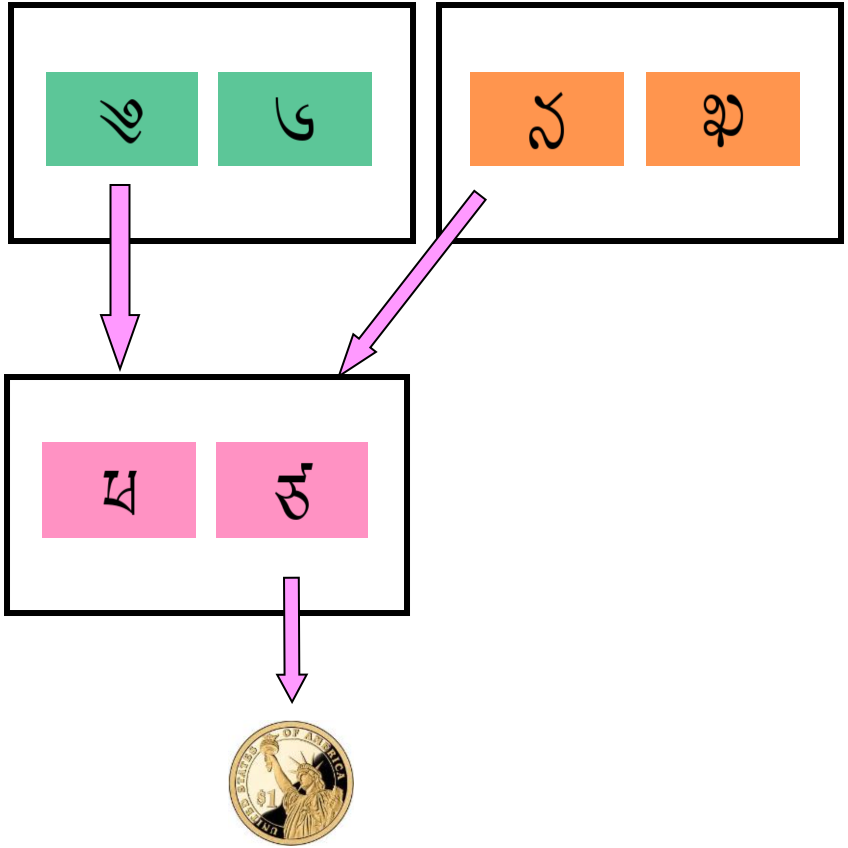
(Doll, Duncan, Simon, Shohamy & Daw *Nature Neuroscience* 2015)

idea

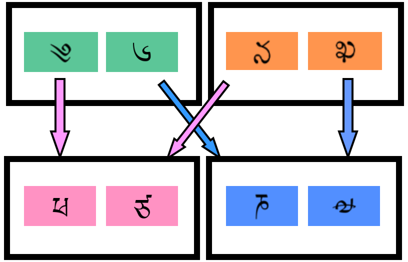
How does bottom-stage feedback affect top-stage choices?

Model-based: actions considered in terms of second-stage state
→Feedback generalizes between equivalents

Model-free: actions reinforced by consequences
→ Feedback does not generalize

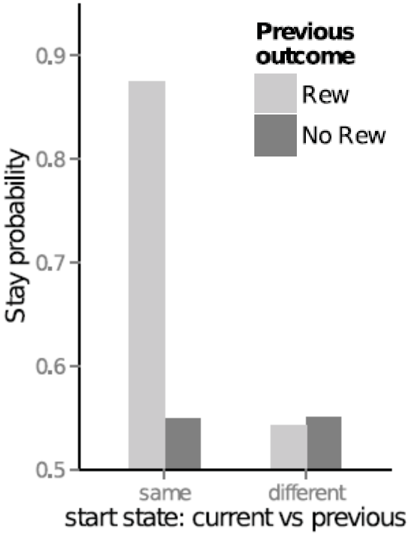


predictions



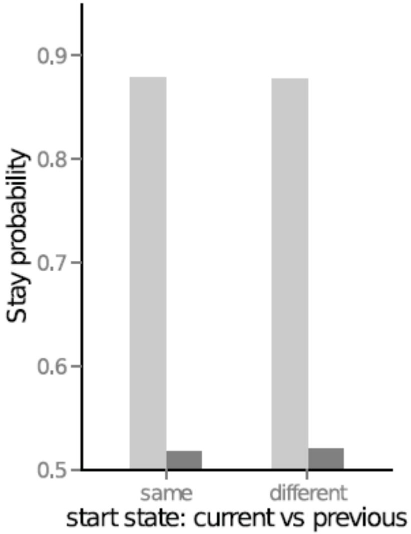
model-free

no generalization



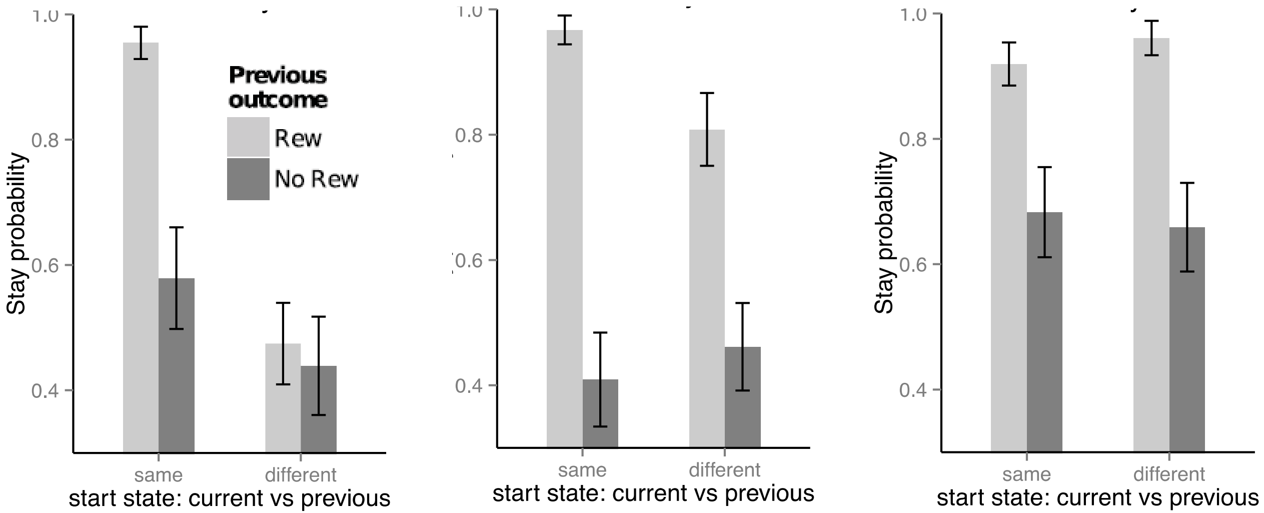
model-based

generalization

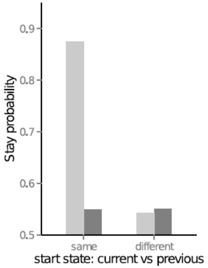


(Doll, Duncan, Simon, Shohamy & Daw *Nature Neuroscience* 2015)

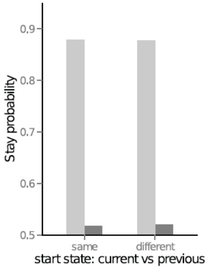
data



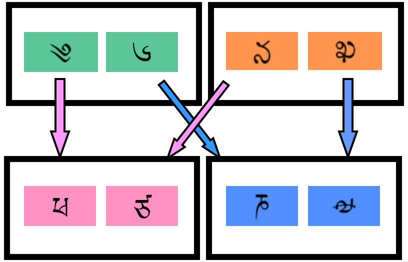
model-free



model-based

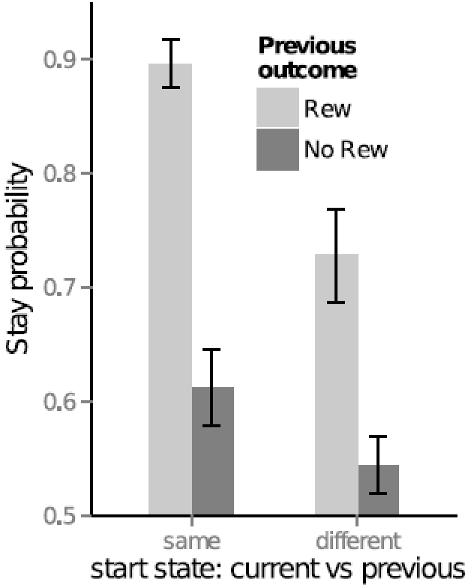


(Doll, Duncan, Simon, Shohamy & Daw *Nature Neuroscience* 2015)



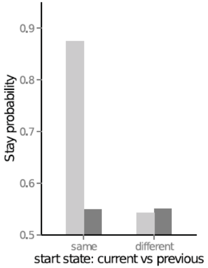
data

20 subs x 272 trials each

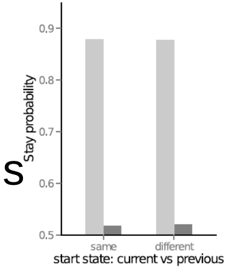


reward (MB): p<.0001
reward x same (MF) p<.005
(mixed effects logit)

model-free



model-based

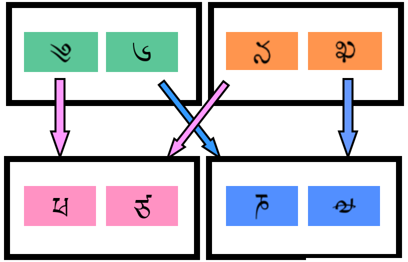


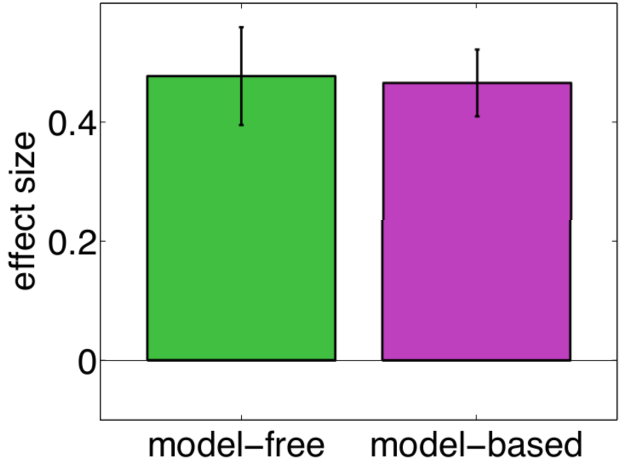
results reject pure reinforcement models
→ suggest mixture of planning and reinforcement processes

(Doll, Duncan, Simon, Shohamy & Daw *Nature Neuroscience* 2015)

data

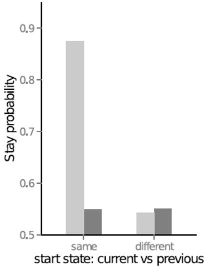
20 subs x 272 trials each



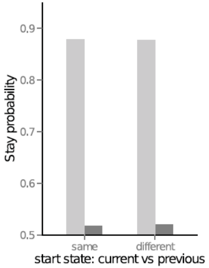


reward (MB): $p<.0001$
reward x same (MF) $p<.005$
(mixed effects logit)

model-free

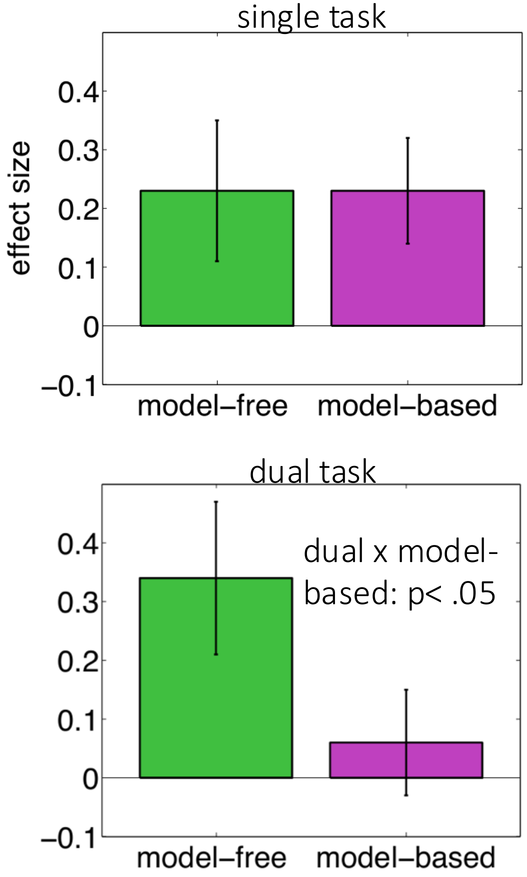


model-based



(Doll, Duncan, Simon, Shohamy & Daw *Nature Neuroscience* 2015)

interference



(Otto et al Psych Science, 2013)

Also:

Individual differences

•Development (Decker ea, 2016)

•Aging (Eppinger ea 2013)

•IQ (Schad ea 2014; Gillan ea 2016)

•cognitive control (Otto ea 2015)

•stress (Otto ea 2015**)**

•Psychopathology (Gillan ea 2016)

PFC (& dopamine there)

•PFC TMS (Smittenaar ea 2013)

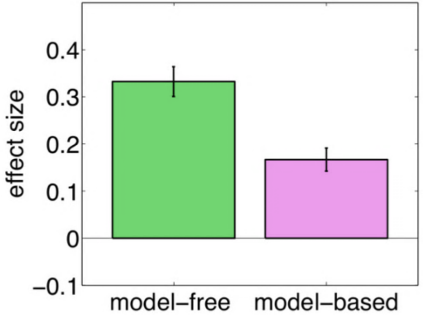
•COMT (PFC DA) genotype (Doll ea 2016))

•PFC dopamine PET (Desserno ea 2015)

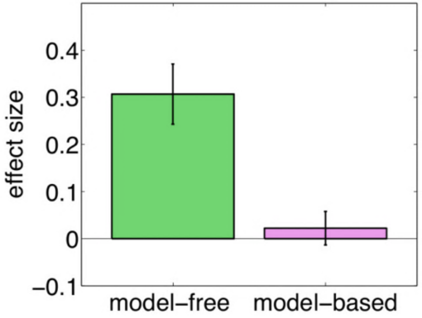
Hippocampus

- Rodents (Miller et al., 2017)
- Humans (Vikhbladh et al., 2019)

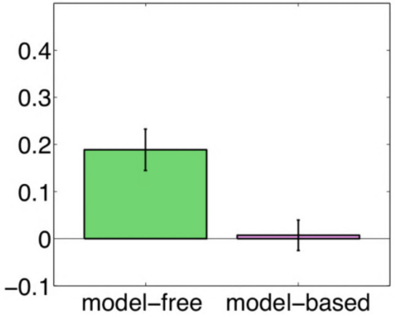
Healthy volunteers, n=106



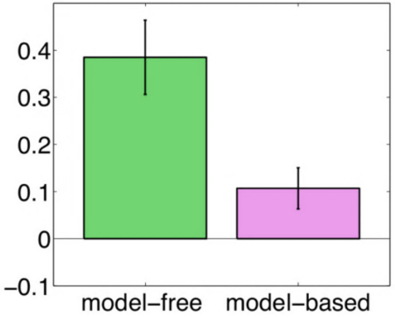
Binge eating disorder, n=30



Stimulant abusers, n=36



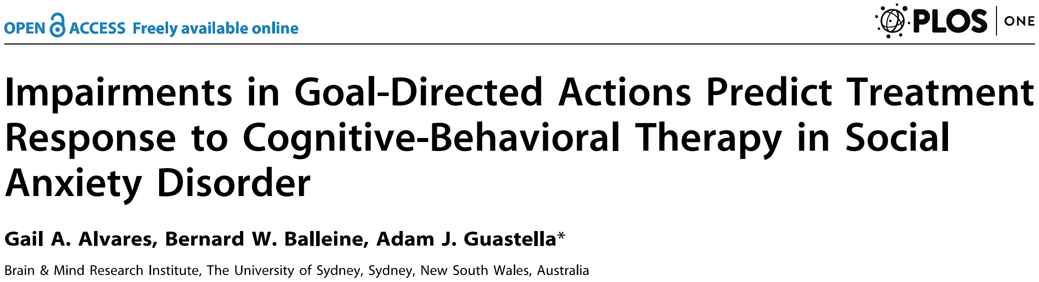
OCD, n=35



Methamphetamine/cocaine
Abstinent at least 1 wk

(Voon et al., Biological Psychiatry, 2014)

however…



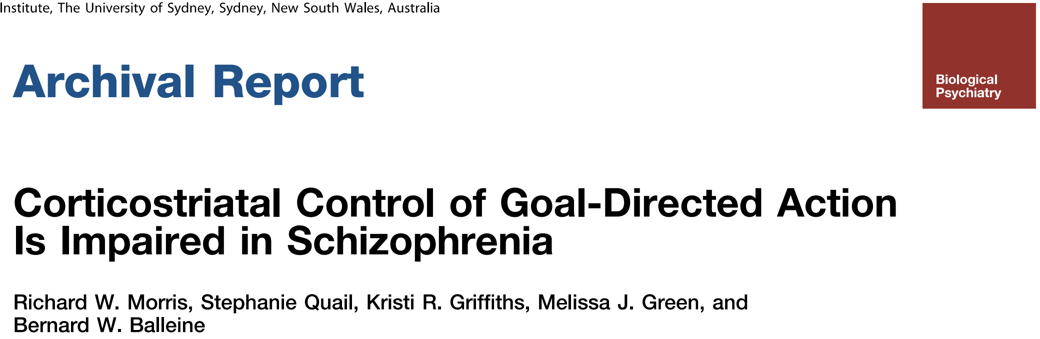
OPEN ACCESS Freely available online

PLOS | ONE

Impairments in Goal-Directed Actions Predict Treatment Response to Cognitive-Behavioral Therapy in Social Anxiety Disorder

Gail A. Alvares, Bernard W. Balleine, Adam J. Guastella*

Brain & Mind Research Institute, The University of Sydney, Sydney, New South Wales, Australia

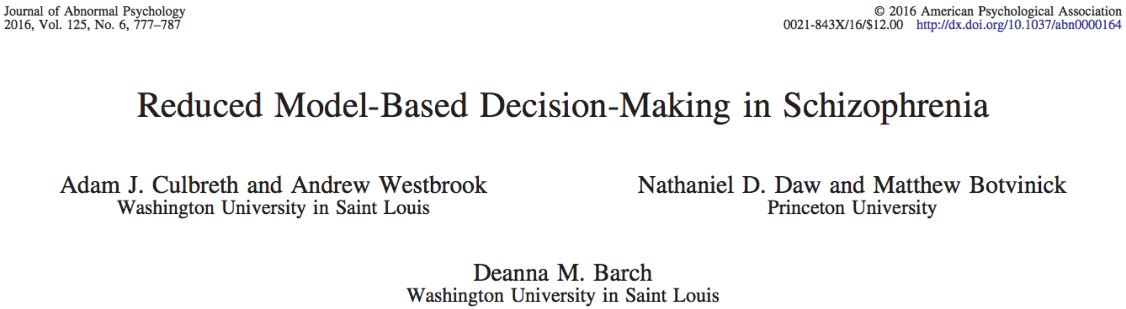


Archival Report

Biological Psychiatry

Corticostriatal Control of Goal-Directed Action Is Impaired in Schizophrenia

Richard W. Morris, Stephanie Quail, Kristi R. Griffiths, Melissa J. Green, and Bernard W. Balleine



Journal of Abnormal Psychology
2016, Vol. 125, No. 6, 777–787

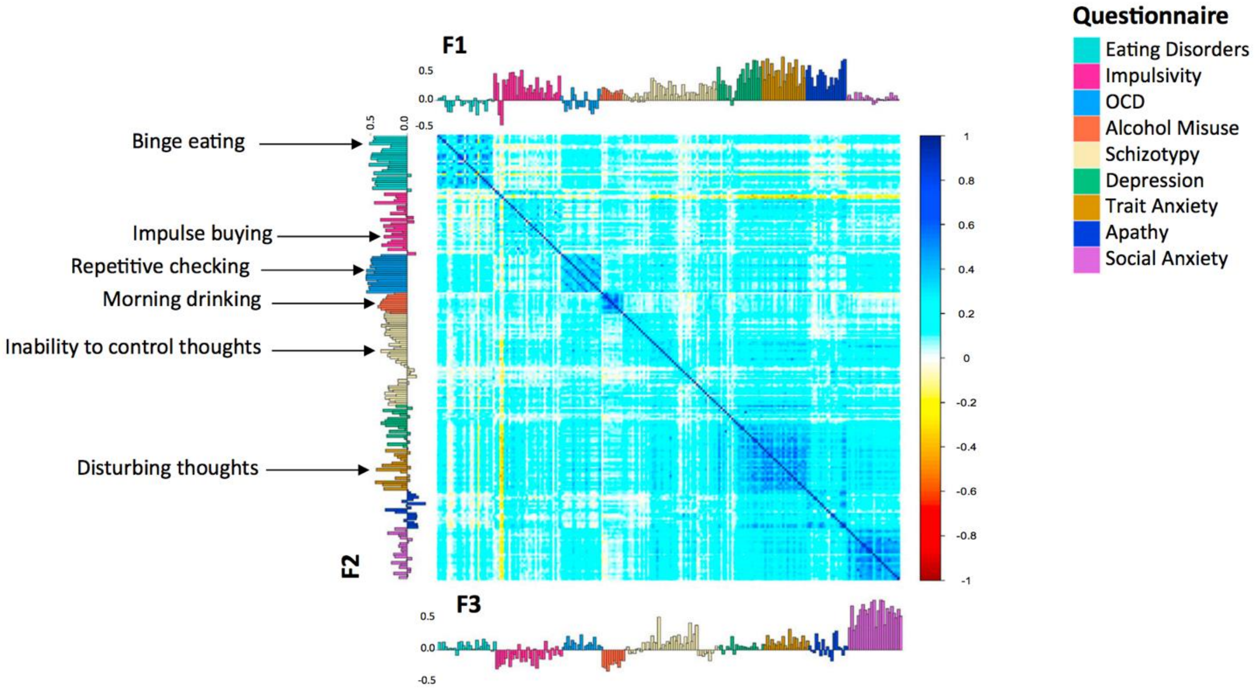
© 2016 American Psychological Association
0021-843X/16/$12.00 http://dx.doi.org/10.1037/abn0000164

Reduced Model-Based Decision-Making in Schizophrenia

Adam J. Culbreth and Andrew Westbrook
Washington University in Saint Louis

Nathaniel D. Daw and Matthew Botvinick
Princeton University

Deanna M. Barch
Washington University in Saint Louis



Gillan, Kosinski, Whelan, Phelps & Daw, *eLife* 2016

recap

- RL: connecting actions to outcomes over space and time
- Exact MB planning flexible but intractable
  - Speed up with prioritization
  - … or MF learning (caching long run values or policy)
  - … or in between like SR/DR (caching long run trajectories)
  - Connections with psychiatry

outline

Estimating action values: model-based vs. model-free learning

1. Intro: dopamine and credit assignment
2. Examples
   - habits and instrumental reward devaluation
   - rodent spatial navigation
   - RL in humans; compulsion
3. Hippocampal replay and planning

ideas: the value function, value updating

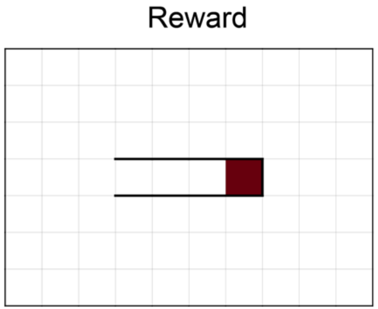
value function:

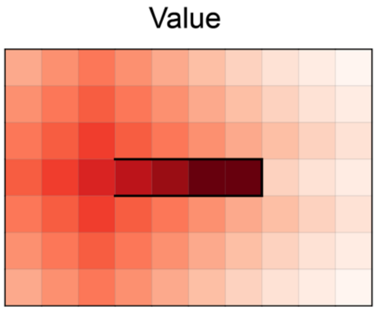
- measures proximity to reward
- makes sequential choice local

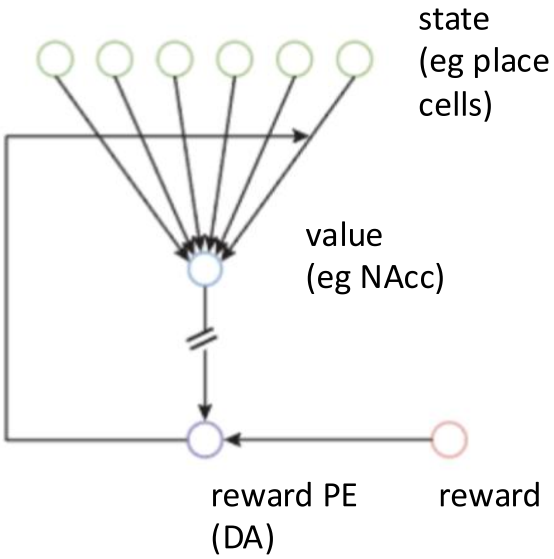
learning a value function:

- experiential updating: prediction errors, phasic dopamine (MF)
- inferential updating by driving same circuit: (MB "planning is learning from simulated experience")

→ suggests more granular control (over updates rather than choice)

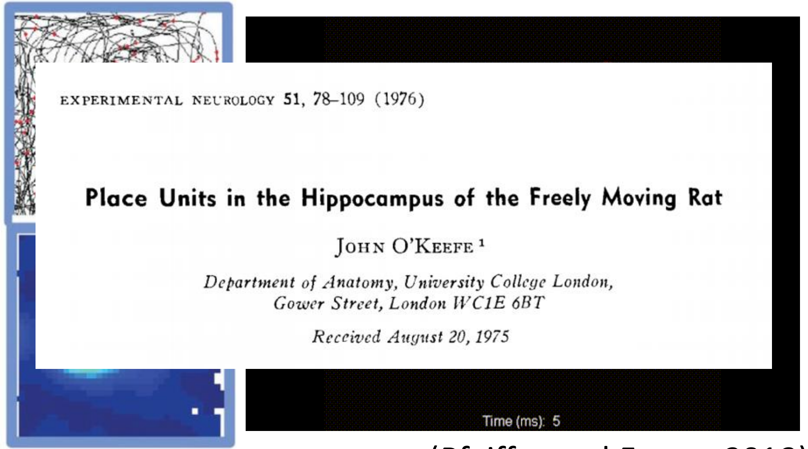






(after Montague et al. 1996)

potential mechanism: nonlocal “replay”



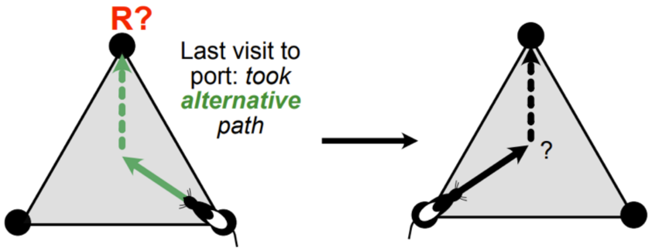
EXPERIMENTAL NEUROLOGY 51, 78–109 (1976)

Place Units in the Hippocampus of the Freely Moving Rat

JOHN O'KEEFE [1]

Department of Anatomy, University College London,
Gower Street, London WC1E 6BT

Received August 20, 1975



(Pfeiffer and Foster, 2013)

representation of location in hippocampus can run far ahead of animals

potential substrate for on-line mental simulation with world model

→ could access evaluation/ choice by driving same learning mechanisms as experience

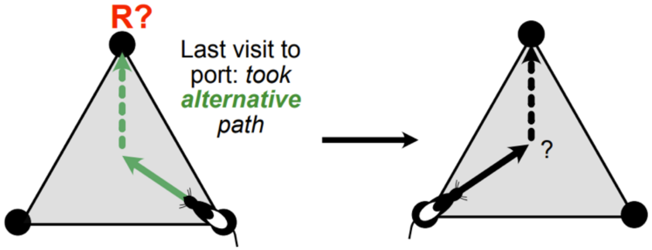
→ if so, it could give us a window into microstructure of planning

→ what can we learn from this?

planning by replay

what can we learn about planning from hippocampal SWR replay patterns?

1. replay happens one path at a time (*search is serial, must be prioritized*)
2. … only while the animal is stopped (*opportunity cost*)
3. … not only ahead but also backward, nonlocal (both *planning and credit assignment?*)



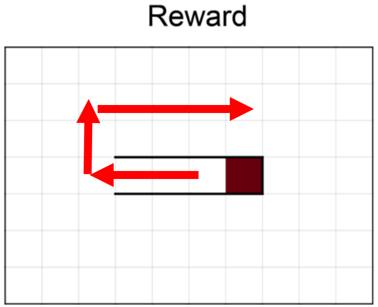
→ highlights selection: what to think about & when?

→can this explain why these patterns occur in different circumstances?

(Mattar & Daw *Nature Neuroscience* 2018; Agrawal, Mattar, Cohen & Daw *Psych Review* 2021)

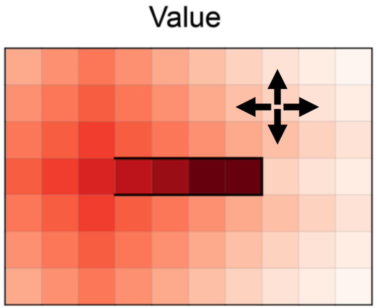
computational ideas

how do we connect actions to outcomes distant in space and time?

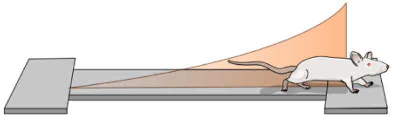


value function:

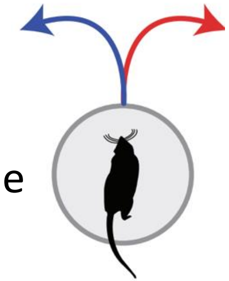
- measures proximity to reward
- makes sequential choice more local



two nonlocal operations:

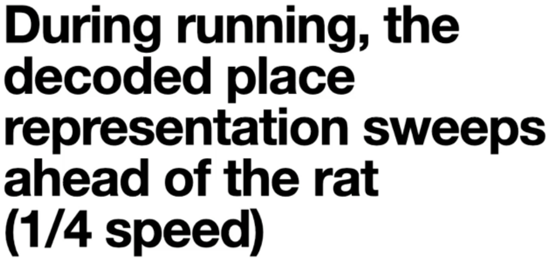


- **update**
  - build value function, e.g. propagate received reward to distal locations
  - long distance activations



- **retrieve**
  - figure out where to go by querying nearby value
  - short distance retrieval during behavior

suggestive data



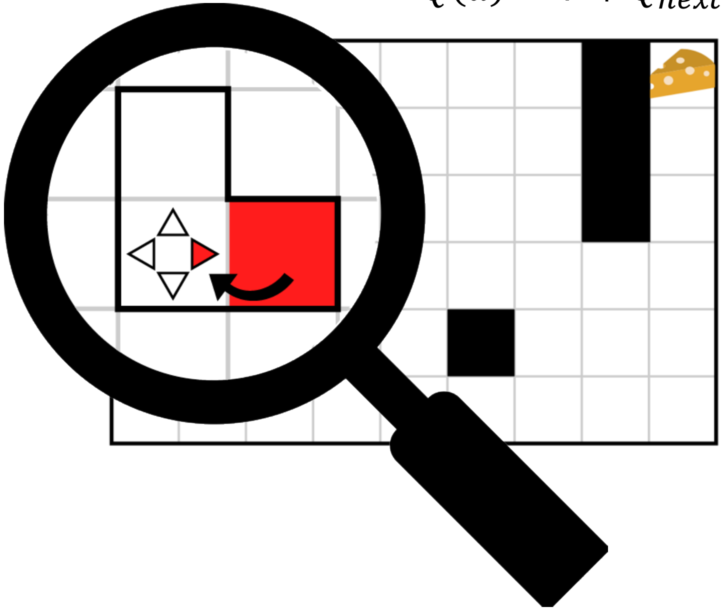
During running, the decoded place representation sweeps ahead of the rat (1/4 speed)

Krausz, Berke et al unpub

new model: prioritized backups

basic operation: Bellman backup (Dyna; Sutton 1991)

$$Q(a) \leftarrow r + Q_{next}$$



Fundamental building block

- Pushes value between adjacent states
- Over actual trajectories (experiential TD learning)
- Over **stored/simulated experiences** (local or remote planning)
- Both can use same DA circuit

question: at which locations to perform backups, in what order?

proposal: at each step, prioritize by utility ("expected value of backup")

→ why does planning carry utility??

(Mattar & Daw *Nature Neuroscience* 2018)

expected value of backup

$EVB(s, a)$: how much (cumulative future discounted) reward do I expect to earn following a backup at that location, compared to before?

$$EVB(s, a) = Need(s) \cdot Gain(s, a)$$

how likely am I to visit s soon?
→ drives activity forward

expected, discounted future occupancy
$\sum_{\tau=t}^{\infty} \gamma^{\tau-t} \delta_{s_\tau, s}$

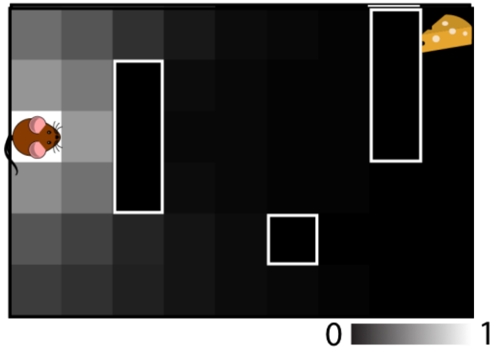
if I get there, how much more will I earn?
→ drives activity backward

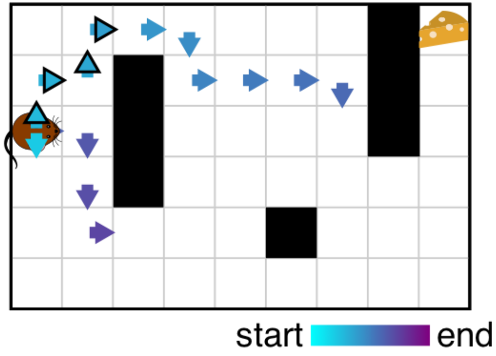
value change under updated policy
$\sum_a \left(\pi_{new}(a|s_k) - \pi_{old}(a|s_k)\right) Q_{\pi_{new}}(s_k, a)$

→ idea: prioritize retrieval according to EVB, balancing need and gain

(Mattar & Daw *Nature Neuroscience* 2018)

need

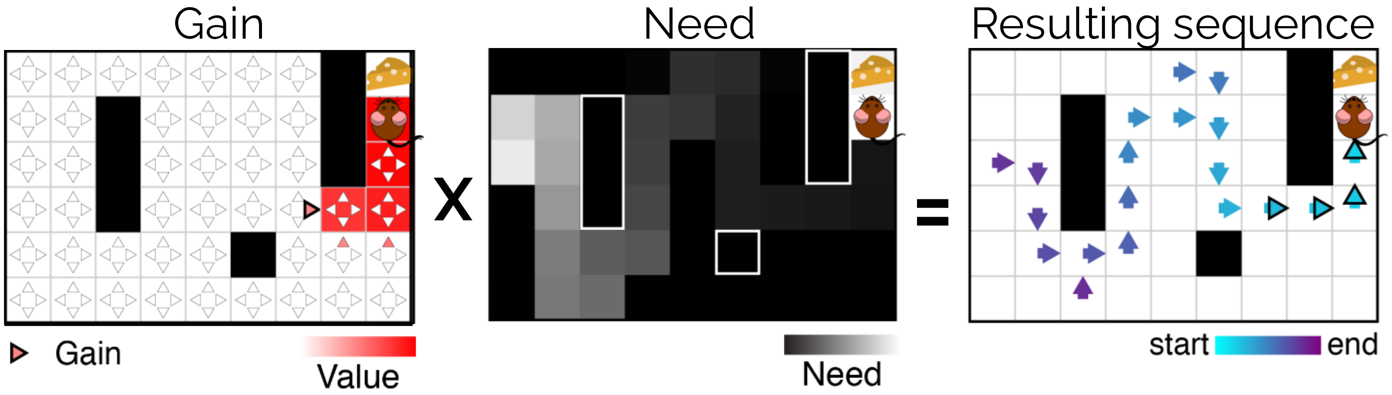




- higher for locations likely to be visited soon
- favors forward replay of imminent trajectories

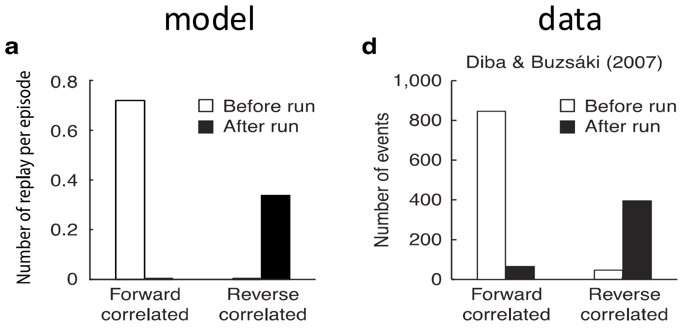
(Mattar & Daw *Nature Neuroscience* 2018)

gain



- how much can I learn at a particular spot?
- drives reverse replay upon learning new information

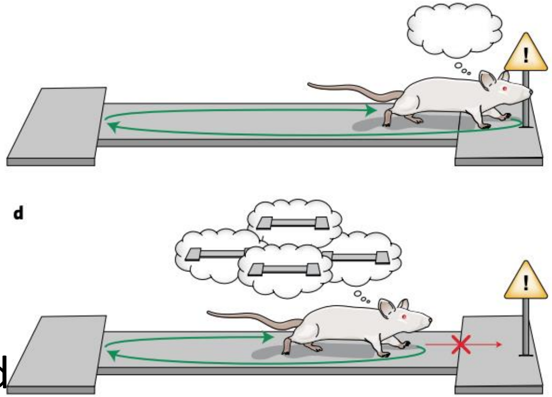
theory predicts place cell replay



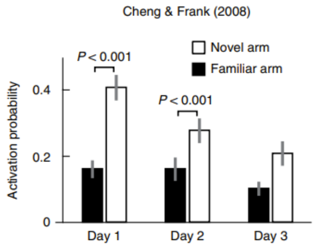
need vs gain promote forward vs backward replay in different circumstances

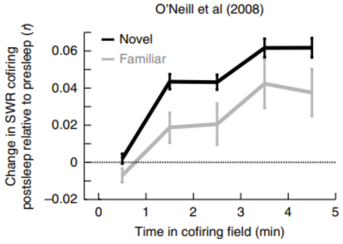
gain explains why some surprises trigger replay while others don’t

& why some trajectories are replayed but avoided



(Widloski & Foster 2018)





seemingly contradictory changes in replay with experience explained by evolution of need vs gain

(Mattar & Daw *Nature Neuroscience* 2018)

recap, thoughts

1. behavior, neural value correlates suggest the brain does nonlocal ("model-based") credit assignment, but not exhaustively (MF, habits)
2. hippocampal replay as a window into this process
   - key role for selection: which locations to consider when

→more granular view on metacontrol

- real issue is not so much whether to think, but what to think about, when

experimental tests

- examine (& intervene upon) predicted relationships in animals doing RL tasks

  experience → replay → value (or model) update → ramps, choices)

psychiatry

- generalizes habit models beyond neglect, to highlight importance of precomputation & selection
  - worry, rumination, craving, obsession, re-experiencing trauma

Other topics

- DAergic heterogeneity (Engelhard et al 2018; Lee et al 2024)
- fitting RL models to choice and neural data (Daw, 2010)
- States and generalization (deep RL; but also latent state inference, Gershman et al. 2010)
- Hierarchical, continuous, or high dimensional actions (connections with motor control; Botvinick et al, 2009; Shadmehr tomorrow)
- Exploration (Agrawal et al., 2021)
- Punishment and avoidance (Uchida; Palminteri et al, 2015)
- Uncertainty, volatility, and learning rate control (Behrens et al 2007; Piray & Daw 2021)
- Connections with sensory uncertainty, perceptual decision making (Lak et al, 2017)

ndaw@princeton.edu