Uncertainty-based arbitration & fMRI studies of reinforcement learning

Nathaniel Daw
New York University
IPAM summer school

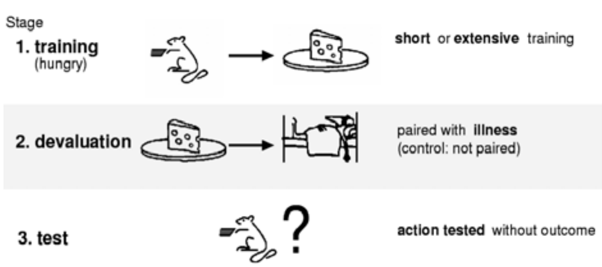
Yael Niv, Peter Dayan, John O'Doherty, Ben Seymour, Ray Dolan, Tom Schonberg, Daphna Joel, Bianca Wittman

Overview

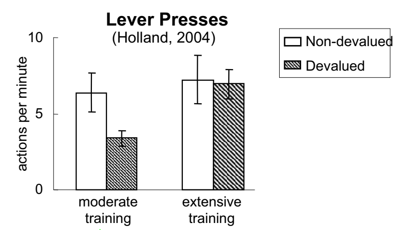
- Approximations for credit assignment: an RL view on goal-directed vs habitual systems
  - Model-based vs model-free RL
  - Uncertainty; arbitration
- fMRI studies of reinforcement learning
  - vmPFC
  - striatum
  - digging deeper: approximations for exploration

Behavioural experiment

- Because TD learners represent only value function, they should be systematically blind to inferences requiring transition/reward model (= contingency, outcome)
- Outcome devaluation used to probe this (Balleine, Dickinson, Killcross)



Behavioural results



Action initially outcome-sensitive
"goal-directed"

Outcome-insensitive following overtraining
"habitual"

→ Animals do behave like TD learners, sometimes
Lesion double dissociations: neurally dissociable systems
Many additional factors impact trade-off (eg preclude habitisation)

Questions

Data suggest behaviorally/neurally distinct systems

1. How to understand goal-directed behavior in RL terms?
2. Why have multiple systems?
3. When to use each?
   – lots of data on when animals actually do

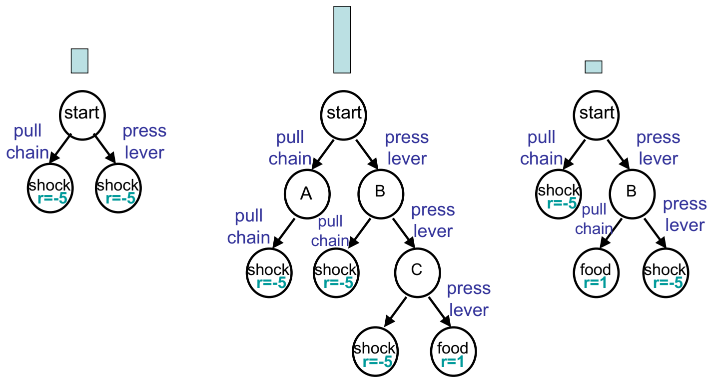
'Model-based' RL

What would Bayes do?

1) Figure out which MDP obtains ('world model')
   - ie, being Bayesian, identify distribution over MDPs
   - $P(state_{t+1}|state_t,action_t)$; $P(r_t|state_t)$
   - Easy! (just counting: Beta & Dirichlet distributions)

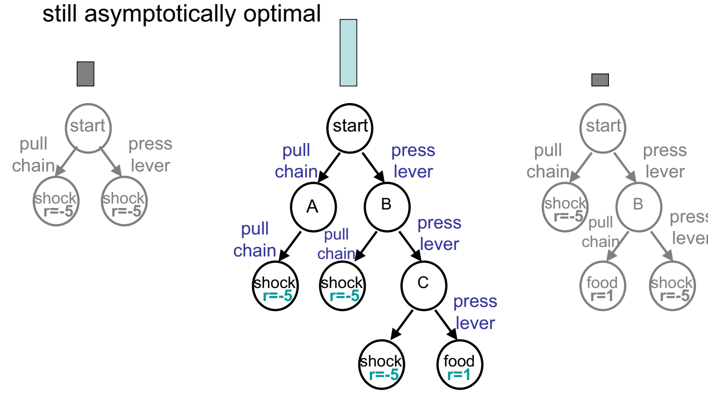
2) Solve it
   - ie compute Q(s,a): expected reward for actions in state
   - with respect to uncertainty in transitions, rewards, MDP
   - "dynamic programming" – explicit search through trajectories of states (think of chess)
   - Hard!

Shortcuts



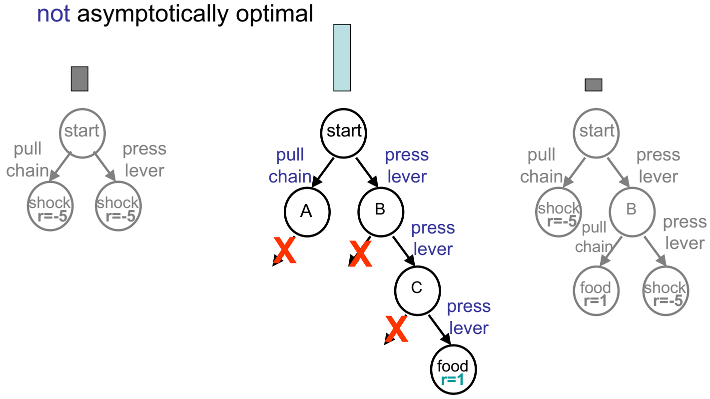
Shortcuts

simplification #1: certainty equivalent
still asymptotically optimal

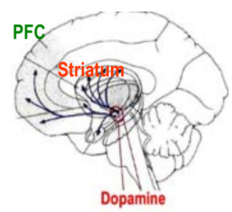


Shortcuts

simplification #2: pruning
not asymptotically optimal



Model-based RL

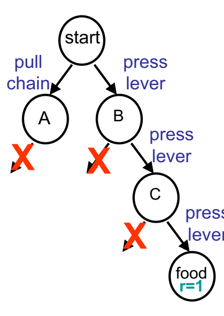


- **Psychology**:
  - cognitive model
  - “goal-directed” behaviour
- **Neuroscience**:
  - prefrontal cortex & planning
  - lesions implicate broader network (BLA, OFC?, etc)

Advantage:
Statistically efficient (inference is Bayes optimal)

Disadvantage:
Computationally prohibitive
In practice, pruning introduces error
This error persists even given infinite data

Model-based RL

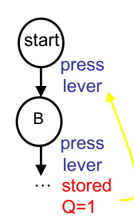


Advantage:
Statistically optimal use of experience (in principle)

Disadvantage:
Computing values is computationally prohibitive
In practice, pruning introduces error
This error persists even given infinite data

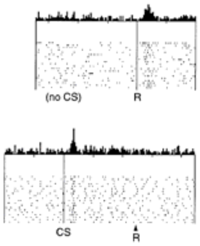
Model-free RL

- Temporal difference learning: Sample intermediate state value (‘bootstrapping’)



$Q(s_t,a_t) \leftarrow r_t + Q(s_{t+1},a_{t+1})$

Model-free RL



- **Psychology**: Habitual behaviour
- **Neuroscience**: Dopamine / TD, basal ganglia, addiction

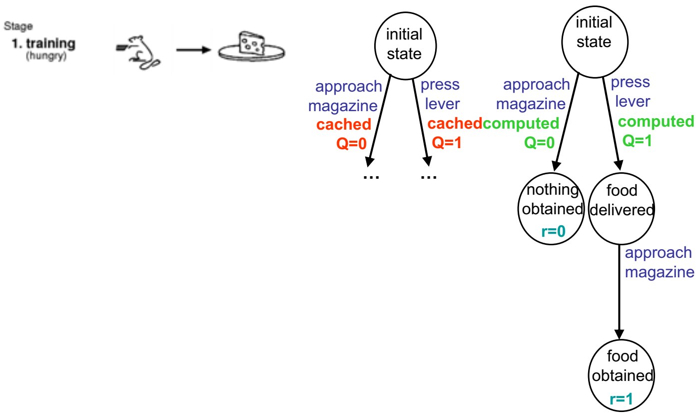
Advantage: Computationally simple Asymptotically optimal

Disadvantage: Sampling & bootstrapping are statistically inefficient when data are scarce

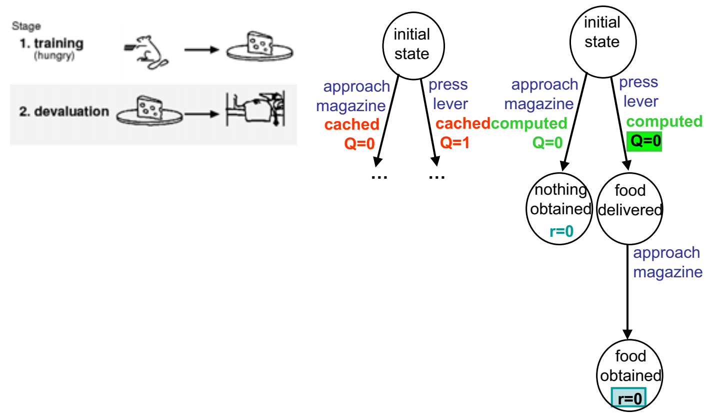
Model-free vs model-based

- Two different shortcuts for obtaining the same quantities
  - Cached values sampled model-free from experience
  - Computed values from search through transition & reward model
- Differentially accurate in different circumstances
  - Model learning more accurate initially (data efficiency)
  - Sampling more accurate asymptotically (computational efficiency)
- Explains why have multiple systems, when to favor each

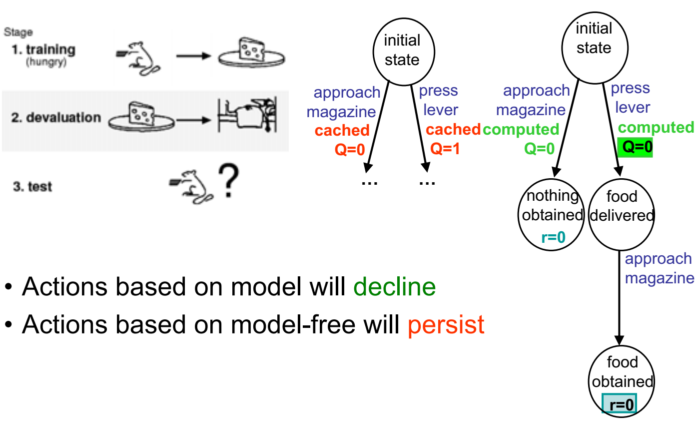
Behavioural experiment



Behavioural experiment



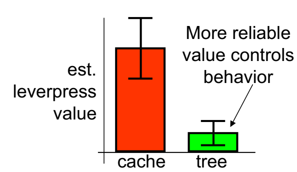
Behavioural experiment



- Actions based on model will decline
- Actions based on model-free will persist

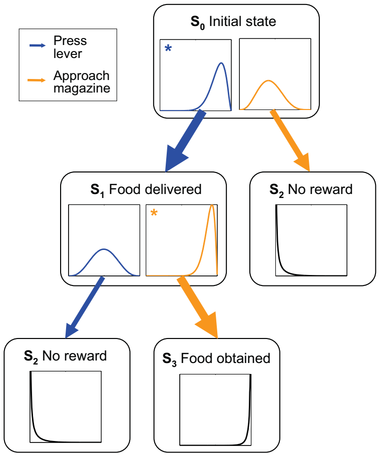
Suggested model

- Parallel controllers:
  - TD/caching (habits, dopamine/striatum)
  - Tree search (goal-directed, PFC)
- **Use each system when it is most accurate**: Assess accuracy with uncertainty
  - Quantifies ignorance about true value (not risk)
  - Treat as evidence reconciliation problem
  - Can also treat decision theoretically (costs vs benefits of expanding tree)

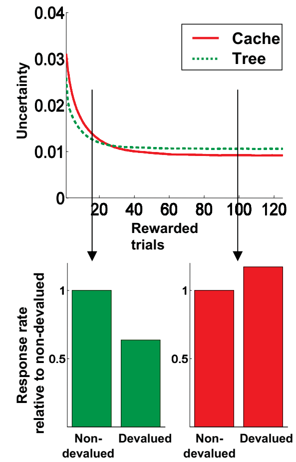


Uncertainty

- Approximate values with distributional value iteration (e.g. Mannor et al. 2004)
- Values accumulate uncertainty through search from uncertainty about MDP (~ error due to certainty equivalence)
- Pruning error modeled with fixed uncertainty per step
- Similar methods used for TD (Dearden et al. 1998)



Simulations

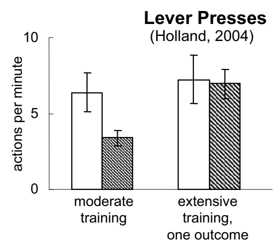


Additionally

- Model-based RL more useful near horizon
- Statistical inefficiency of model-free RL more difficult to overcome in more complex tasks

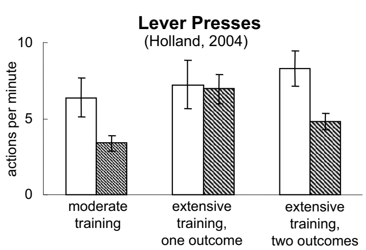
→ Both factors should oppose habitization

Behavioural results



Habitisation with overtraining

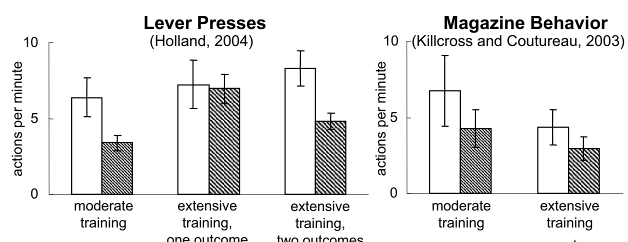
Behavioural results



Habitisation with overtraining

… but not in tasks with multiple outcomes

Behavioural results

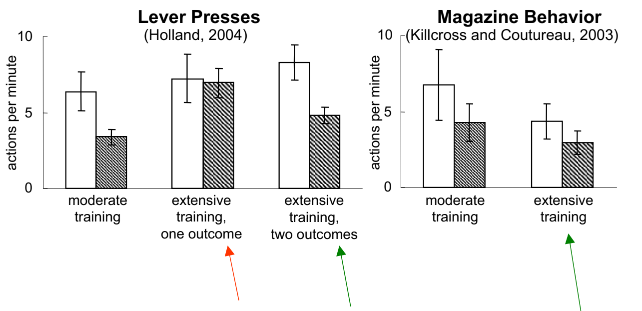


Habitisation with overtraining

… but not in tasks with multiple outcomes

… and not for actions proximal to reward

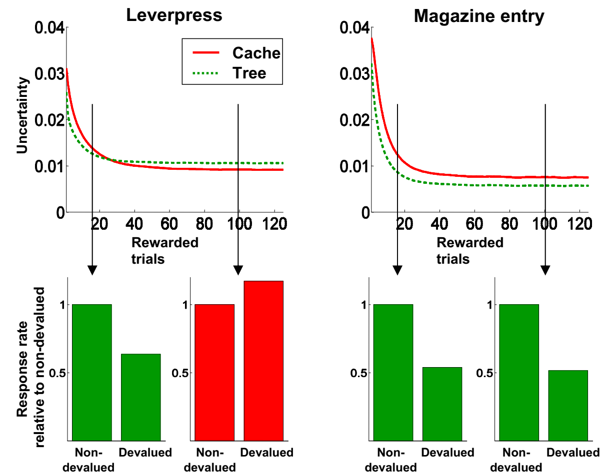
Behavioural results



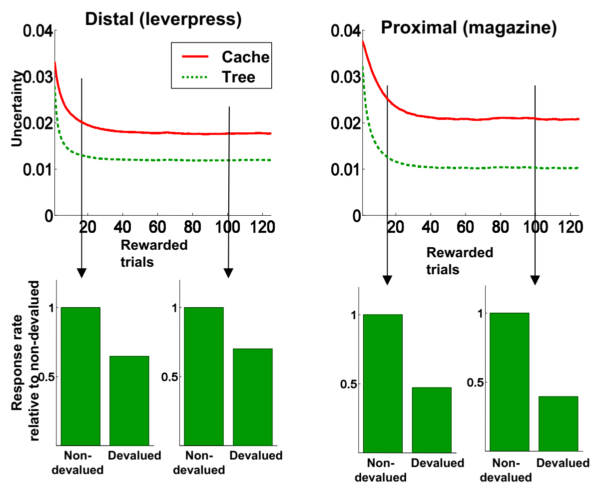
Data efficiency: overtraining and task complexity

Computational efficiency: search depth

Simulations



Two actions/two outcomes



Summary

- Model-based RL as model of “cognitive” action control
- Why have two systems? Different approximations are appropriate to different circumstances
- When do animals use each system? Under those circumstances to which it is most appropriate.
- How could they determine this? Uncertainty.

Qs: Neural substrates for uncertainty (Ach? ACC?), arbitration (ACC?), dynamic programming (attractors?)

Overview

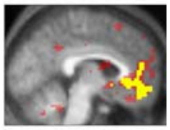
- Approximations for credit assignment: an RL view on goal-directed vs habitual systems
  – Model-based vs model-free RL
  – Uncertainty; arbitration
- fMRI studies of reinforcement learning
  – vmPFC
  – striatum
  – digging deeper: approximations for exploration

fMRI

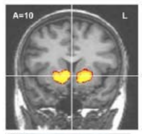
- Measure blood oxygenation level dependent (BOLD) signal. Difficult to pin down neural source.
- Good spatial resolution (eg 3mm^3). Poor temporal resolution (impulse response peaks about 5 secs late)
- Univariate tests at each voxel regressing hypothesized to observed signals.
  – Random effects over population.
  – Correct for multiple comparisons.
- Trend: fit computational models to behavior to estimate subjective trial-trial signals (like Russell’s IRL)
  – Value expectation, prediction error, uncertainty
  – Use the estimates to study neural representations (eg generate regressors, look for correlations)
  – Compare neural and behavioral fits, individual differences

General findings

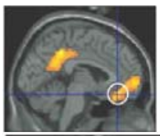
Variety of rewards or reward anticipation activates vmPFC/OFC, striatum (sometimes midbrain)



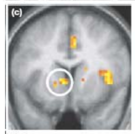
money
value predicted
(Daw et al 2006)



money
gain vs loss
(Kuhnen & Knutson 2005)

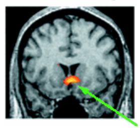


faces
attractiveness
(O'Doherty et al 2003)



food odors
valued vs devalued
(Gottfreid et al 2003)

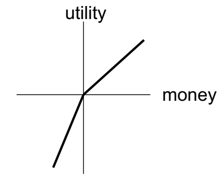
Coke or Pepsi
degree favored
(McClure et al. 2004)

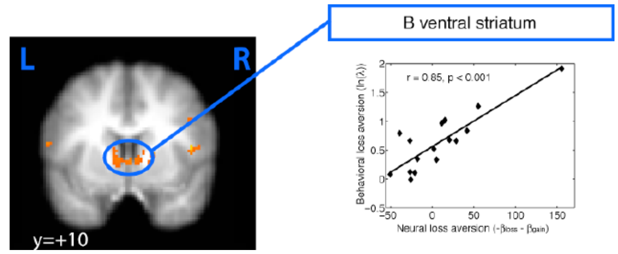


juice
unpredictable vs predictable
(Berns et al 2001)

Behavioral validity

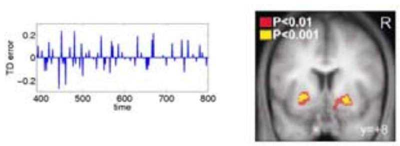
Tom et al (2007): compare loss aversion estimated from neural value signals to behavioral loss aversion from choices



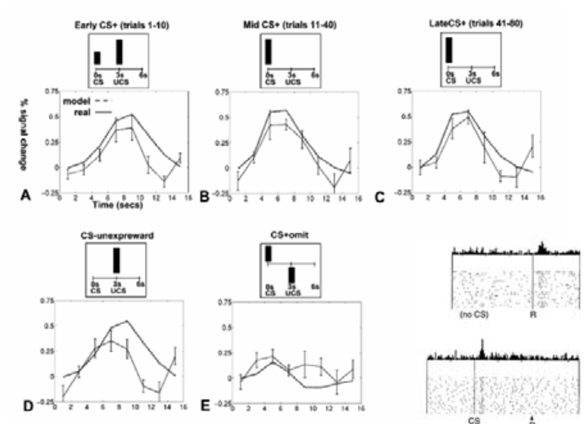


What's really going on in striatum?

TD error (O'Doherty et al 2004; cf 2003 and lots of other papers)



Striatal timecourses

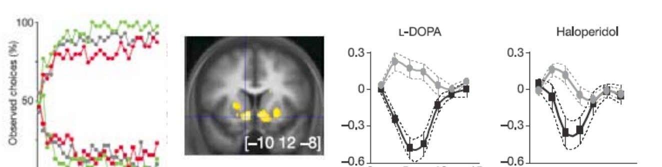


(O'Doherty et al 2003)

(Schultz et al. 1997)

Striatal BOLD, learning, dopamine

Linked to learning; may reflect dopaminergic input



(Pessiglione et al 2006)

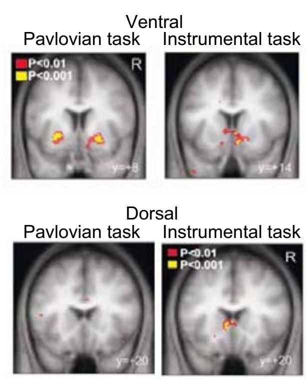
Learners

Nonlearners

hi – I had to remove data from my friends and collaborators that hasn't been published yet. please contact me personally if you would like to see it.

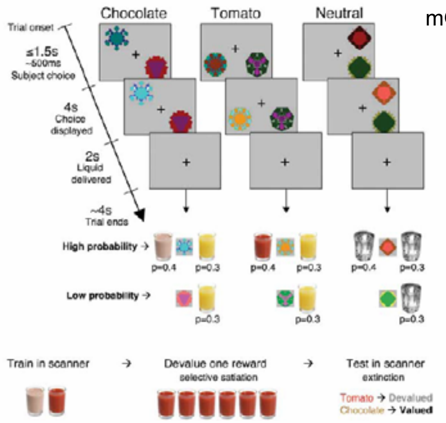
(Schonberg et al under review)

Dorsal / ventral in FMRI

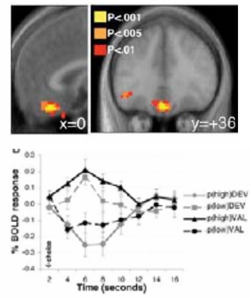


(O'Doherty et al. 2004; cf Delgado et al)

Goal-directed prediction



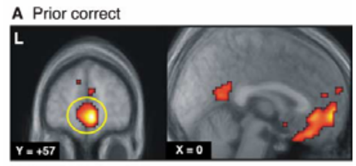
mOFC respects devaluation

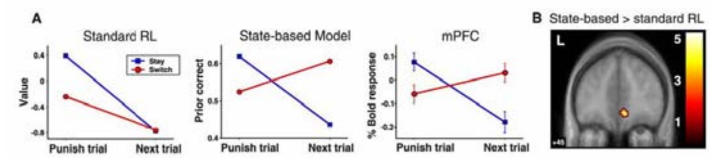


(Valentin et al 2007)

Model-based knowledge

Another example where vmPFC knows more than simple TD: respects higher-order structure in serial reversal task





(Hampton et al 2006)

Summary

- Network activated in appetitive tasks
- vmPFC/OFC: prediction (also outcomes)
  – seems to have model-based knowledge
  – no luck so far determining whether striatum does
- ventral striatum: prediction errors
  – linked to behavior, dopamine
- Can interrogate these responses further to understand neural substrates

Conclusions

- Model-free RL
  – dopamine, striatum: imaging, ephys, lesions
  – very well understood, eg exploration heuristics
  – no model: systematic ignorance
- Model-based RL
  – PFC, other parts of striatum
  – less well understood but many hints
  – RL view on richer cognitive representation
- arbitration: meta-rational analysis of approximation