RL part 2

Levels of analysis

Marr's (1982) hierarchy:

Computation

interpretation: why?

Algorithm

Implementation

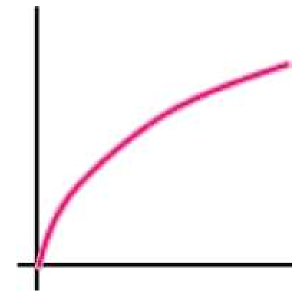
simulation: how?

Levels of analysis

Marr's (1982) hierarchy:

Computation
interpretation: why?

eg expected utility theory



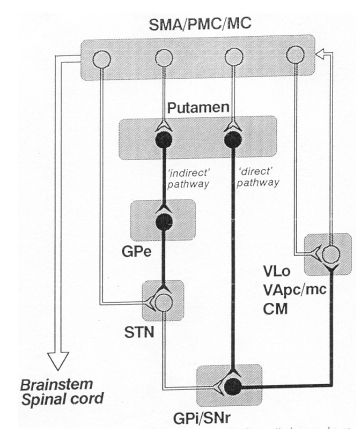
Algorithm

eg R/W learning

$\delta_t = r_t - V_t$

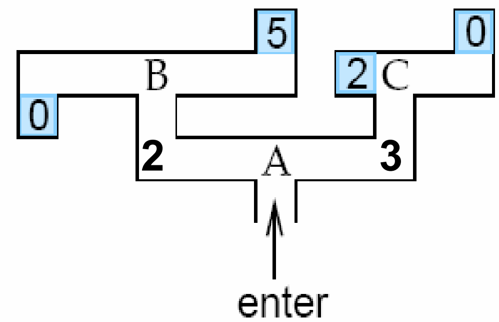
Implementation
simulation: how?

eg dopamine, BG loops



Markov Decision Processes (MDPs)

- Sequential decision tasks
  - Like a maze
  - [state,action]$\rightarrow$[reward,new state]
  - Can be stochastic

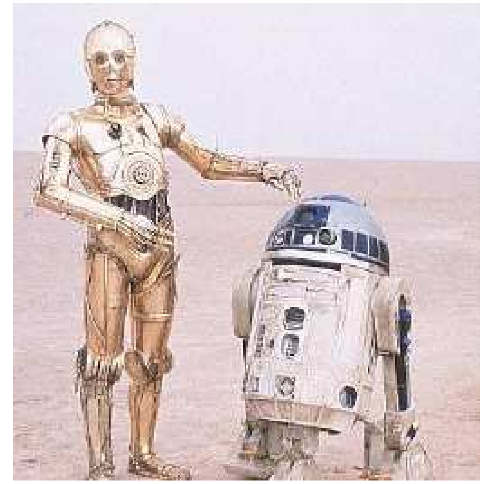


- Want to choose actions to optimize

$$E\left[\sum_{\tau=t}^{end} r_\tau\right] \quad \text{or} \quad E\left[\sum_{\tau=t}^{\infty} \gamma^{\tau-t} r_\tau\right]$$

where the expectation is over stochasticity in transitions & reward deliveries

Online policy learning



The task:

World: You are in state 34.
Your immediate reward is 3. You have 3 actions.

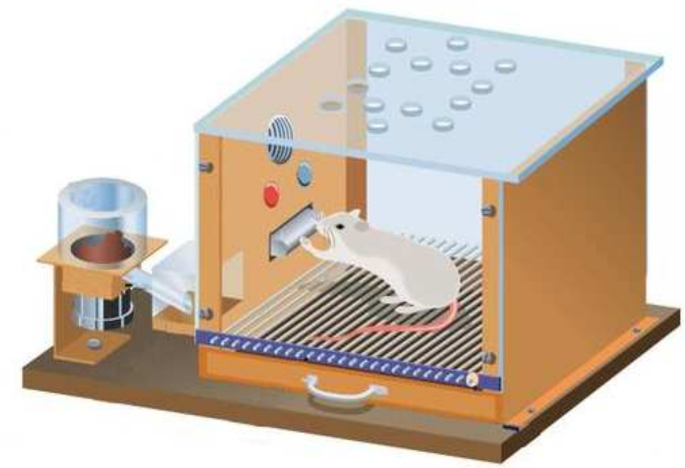
Robot: I'll take action 2.

World: You are in state 77.
Your immediate reward is -7. You have 2 actions.

Robot: I'll take action 1.

World: You're in state 34 (again).
Your immediate reward is 3. You have 3 actions.

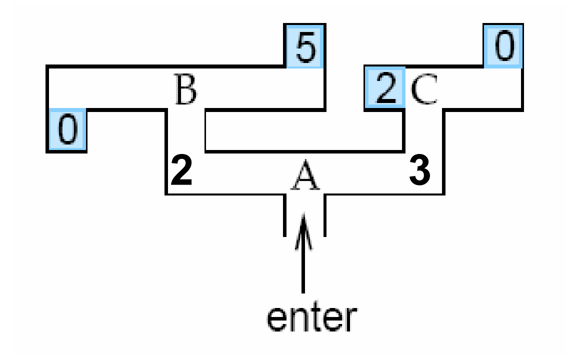
Choice in unknown MDPs



- General facts:
  - Algorithms exist that can asymptotically choose optimally
  - Very few guarantees during learning (explore/exploit, eg Kearns & Singh, 1998)
  - Only one special case really nailed (the Gittins index for n-armed bandit)

Markov Decision Processes

Sequential decision tasks



- Difficulty is optimizing long-term quantity
- 'Credit assignment problem'
- Use prediction to simplify

As before:

1. Predict long-term value of action in state: 'Q(s,a)'
2. Choose based on this

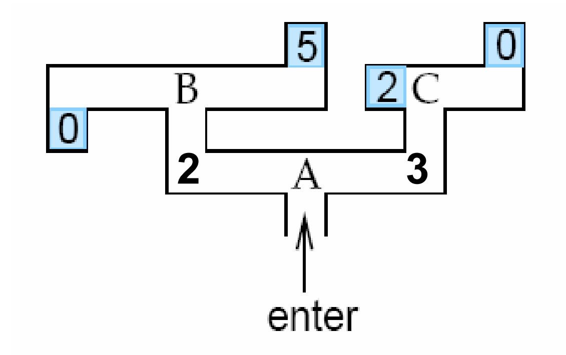
TD learning

What to do at A?

Define:

$$Q(s_t, a\ ) = E\left[r_t + \gamma r_{t+1} + \gamma^2 r_{t+2} + \ldots\right]$$

$$= E\left[\, r_t + \gamma Q(s_{t+1}, a_{t+1}) \right]$$



So:

$$\delta_t = r_t + \gamma Q(s_{t+1}, a_{t+1}) - Q(s_t, a_t) \text{ should equal } 0$$

2, if we went left

Q(B,right or left) eg 5

Use in R/W update rule as before:

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

Behavior

TD caches values V or Q

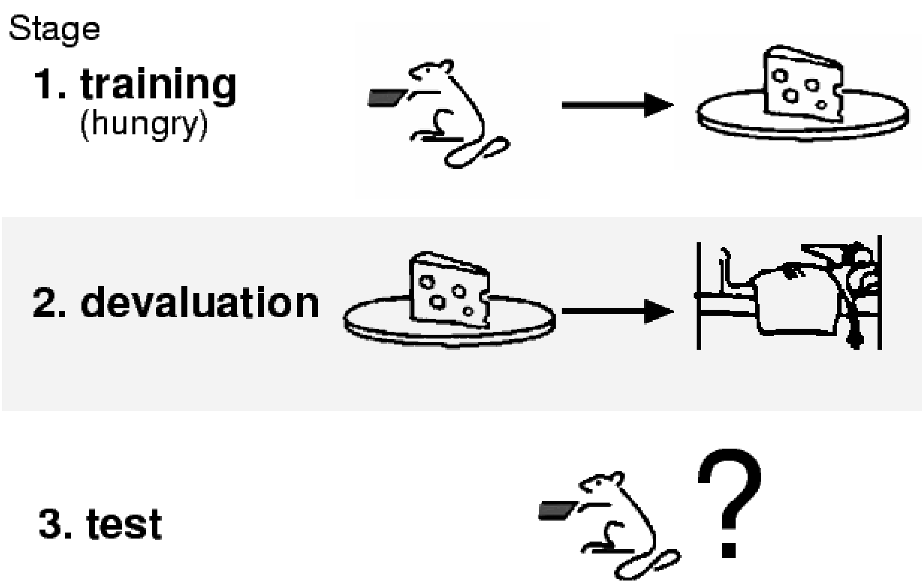
Divorced from representation of specific outcome (like food)

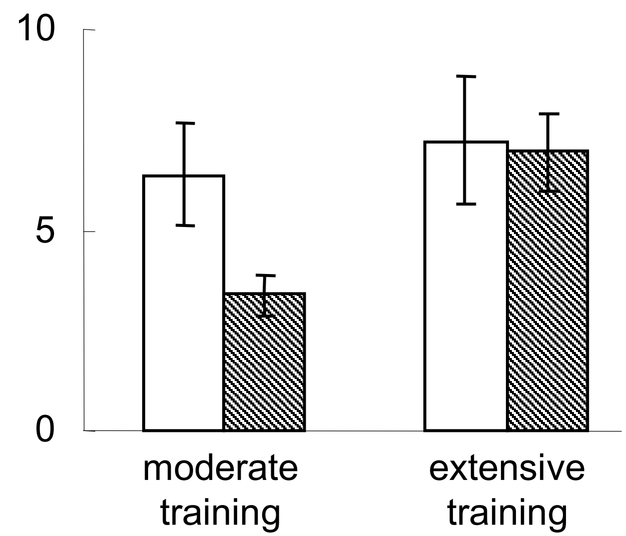
- This is a computationally simple approximation to explicit planning (about which, more later)

This approximation has weird consequences

- e.g. should be blind (without retraining) to changes in outcome value
- Satiety, illness etc.

Test





(Holland, 2004)

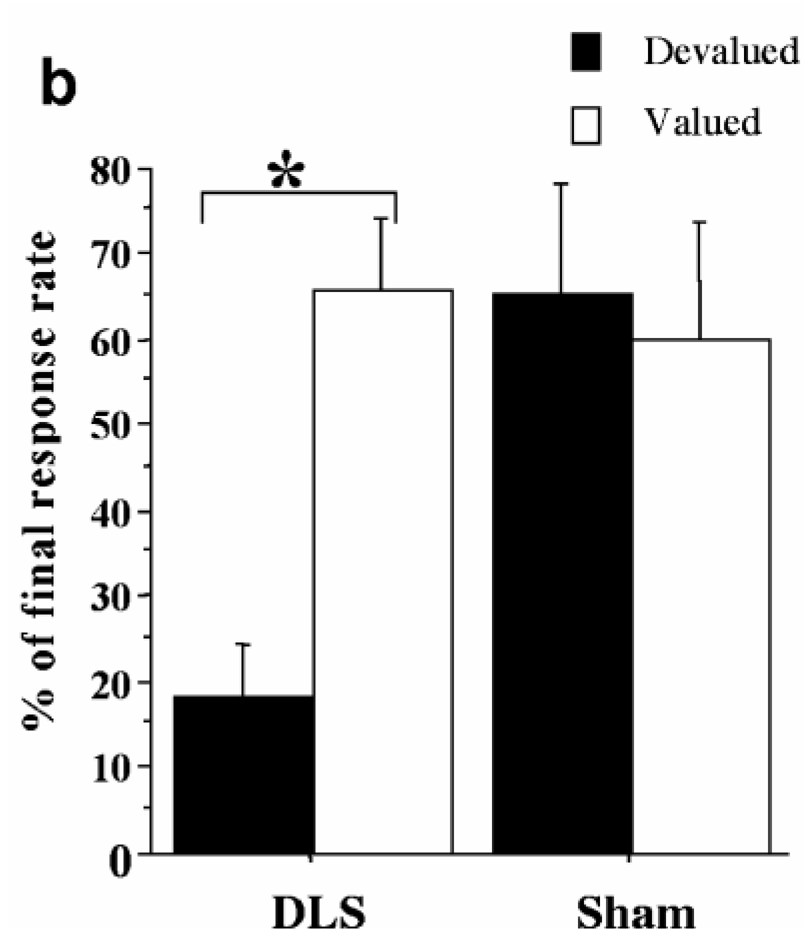
Animals behave in accord with TD, sometimes

- Experiments, lesions suggest two parallel decision paths
- Broadly, striatum associated with TD and PFC with planning
- Lots more behavioral data on when the systems trade off

Lesions

- With lesion of dorsolateral striatum (also its DA input) rats acquire normally but never habitize

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



Yin et al 2004

Some questions

(Daw, Niv, Dayan 2005)

- What is this second decision system?
- Why would there be two?
- How would you choose between them?

‘Model based’ RL

What would Bayes do?

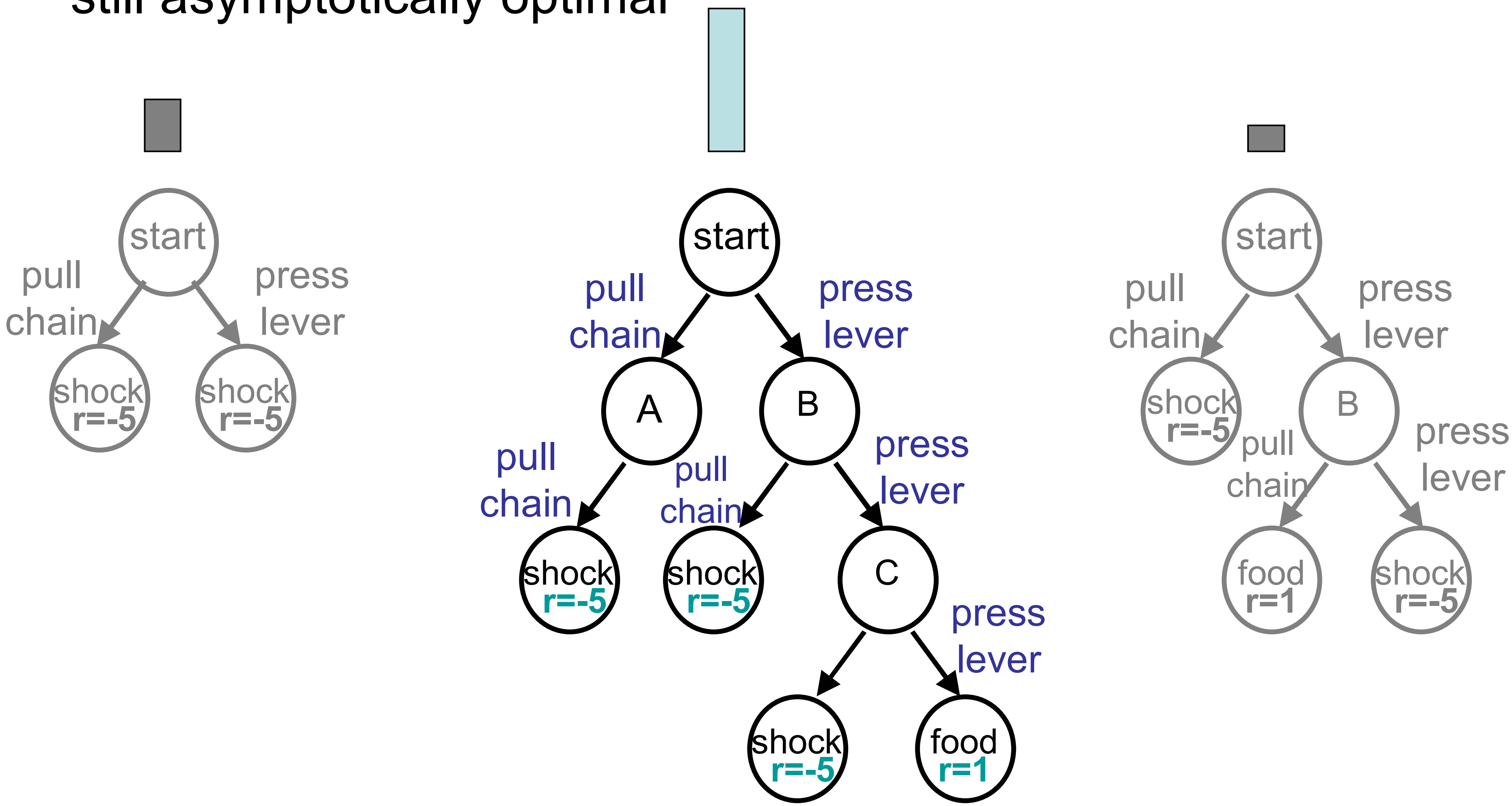
1) Figure out which MDP obtains (‘world model’)
   - ie, being Bayesian, identify distribution over MDPs
   - $P(\text{state}_{t+1}|\text{state}_t,\text{action}_t)$; $P(r_t|\text{state}_t)$
   - Easy! (just counting: Beta & Dirichlet distributions)

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 (cf Colin’s games, think of chess)
   - Hard!

Shortcuts

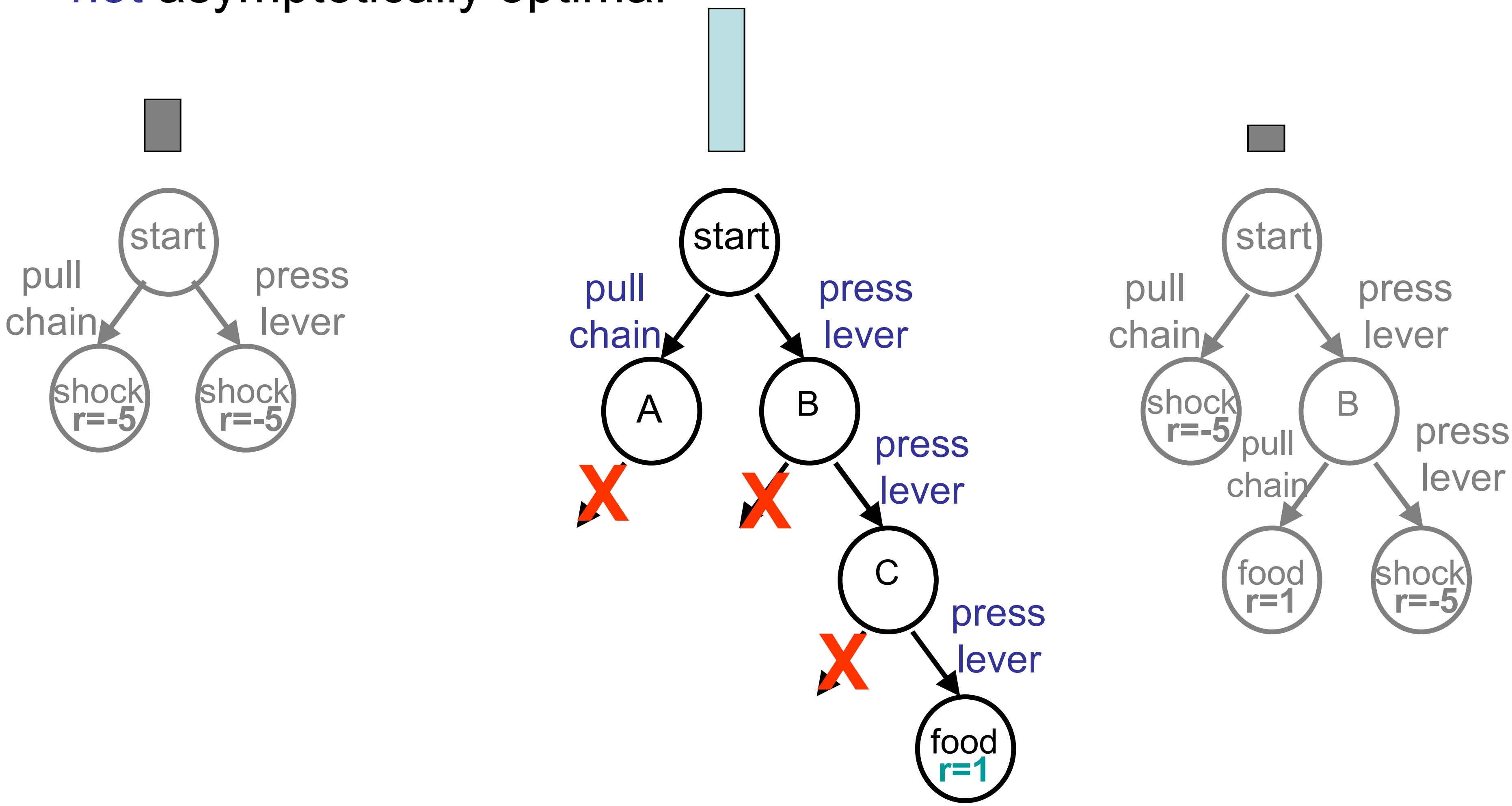
simplification #1: certainty equivalent

still asymptotically optimal

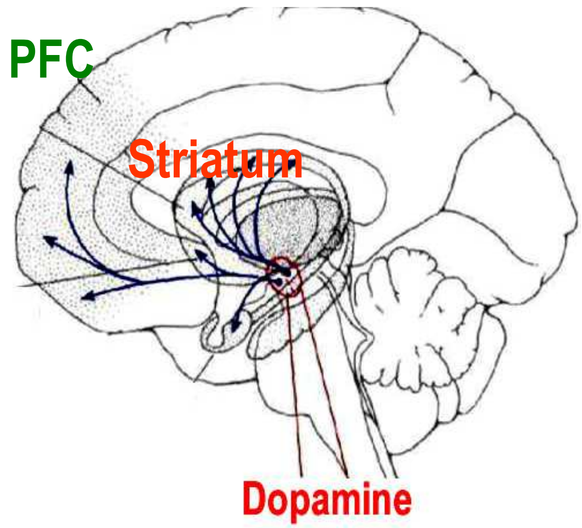


Shortcuts

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 optimal use of experience (in principle)

Disadvantage:

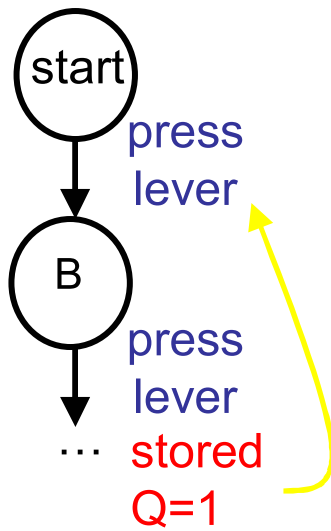
Computationally prohibitive

In practice, pruning introduces error

This error persists even given infinite data

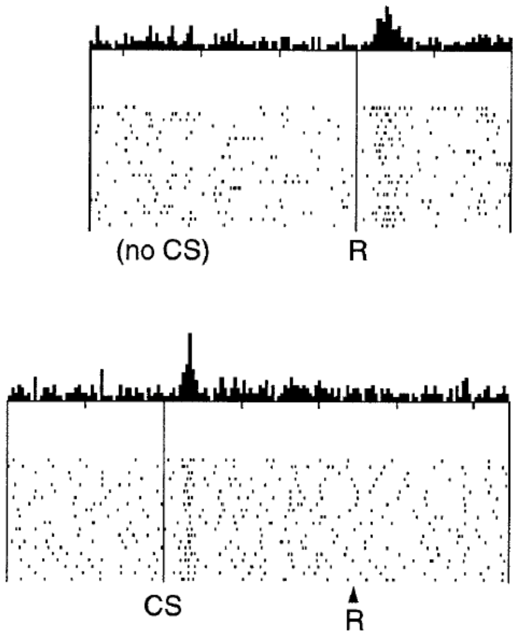
approach 2: Model-free RL

- we've already seen: 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

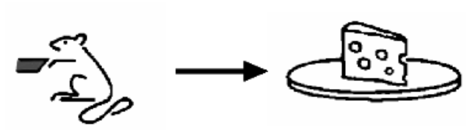
Model-free vs model-based

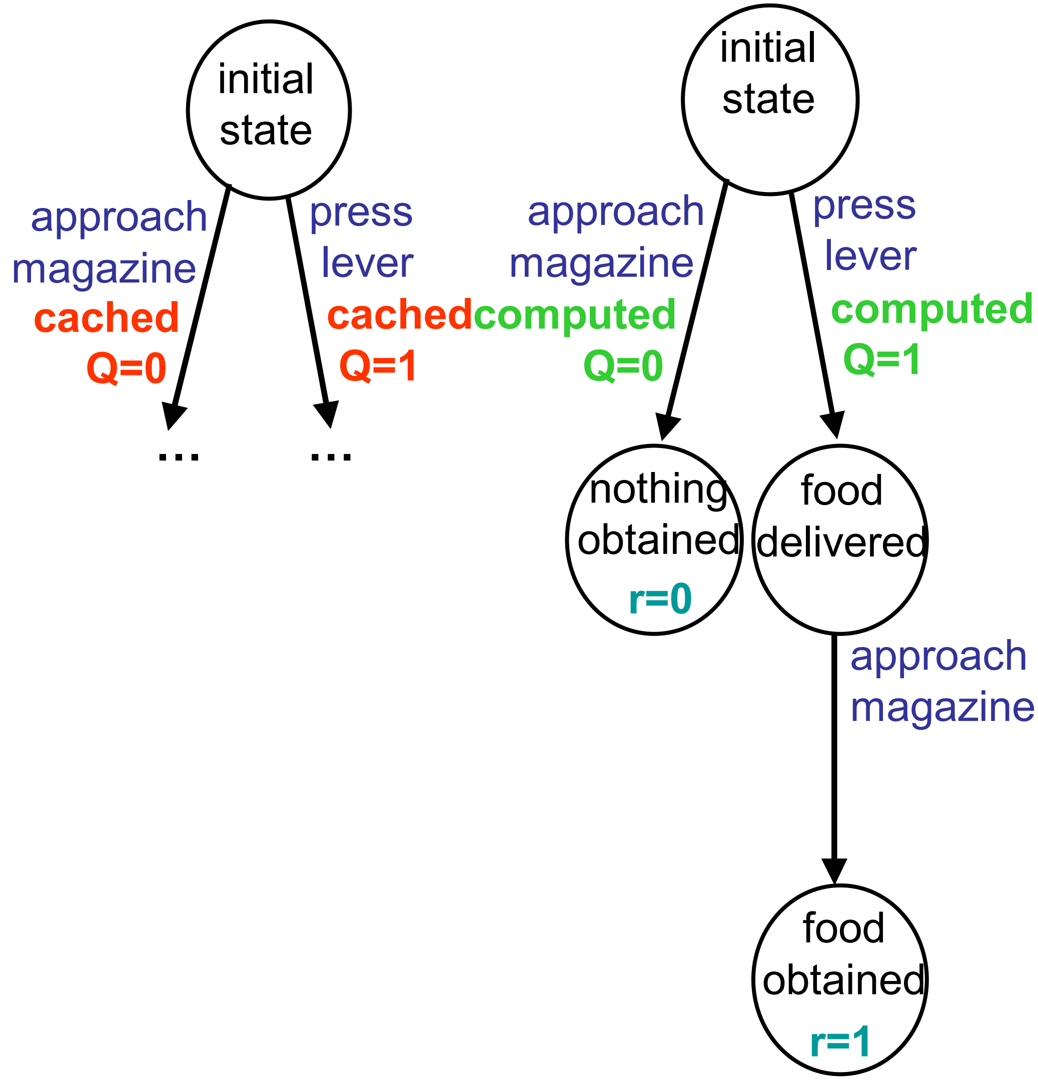
- Two different shortcuts for obtaining the same quantities
  - Cached values sampled model-free from experience
  - Computed values from search through transition & reward model
- Differentially accurate in different circumstances
  - Model learning more accurate initially (data efficiency)
  - Sampling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

Behavioural experiment

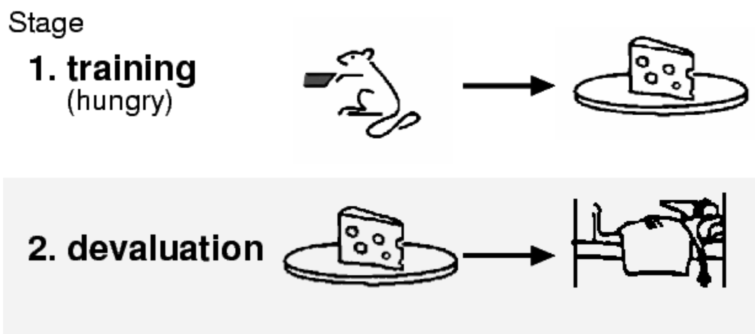
Stage

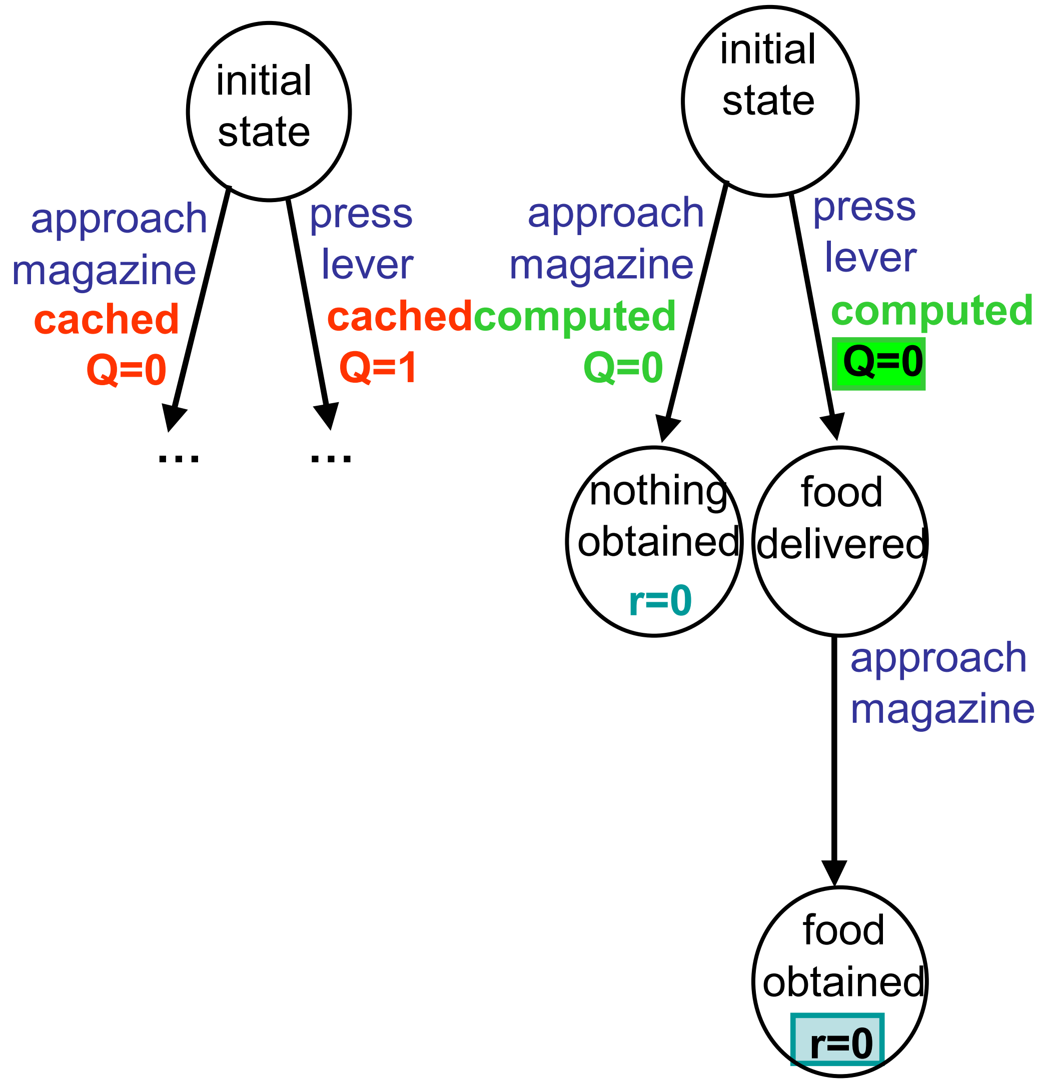
1. training
(hungry)



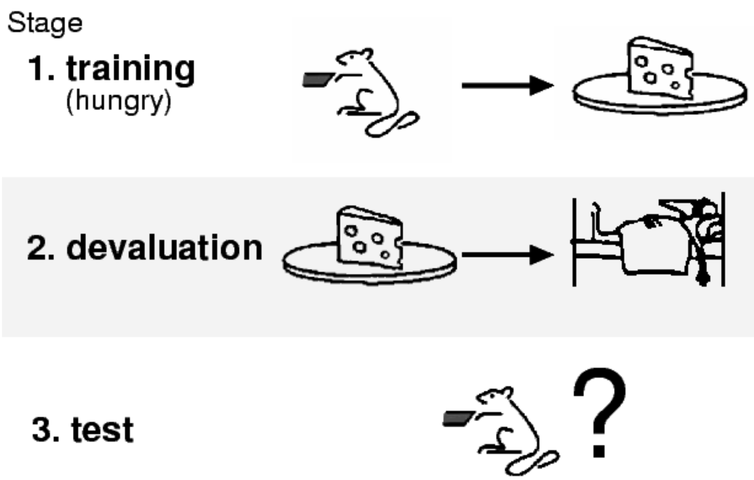


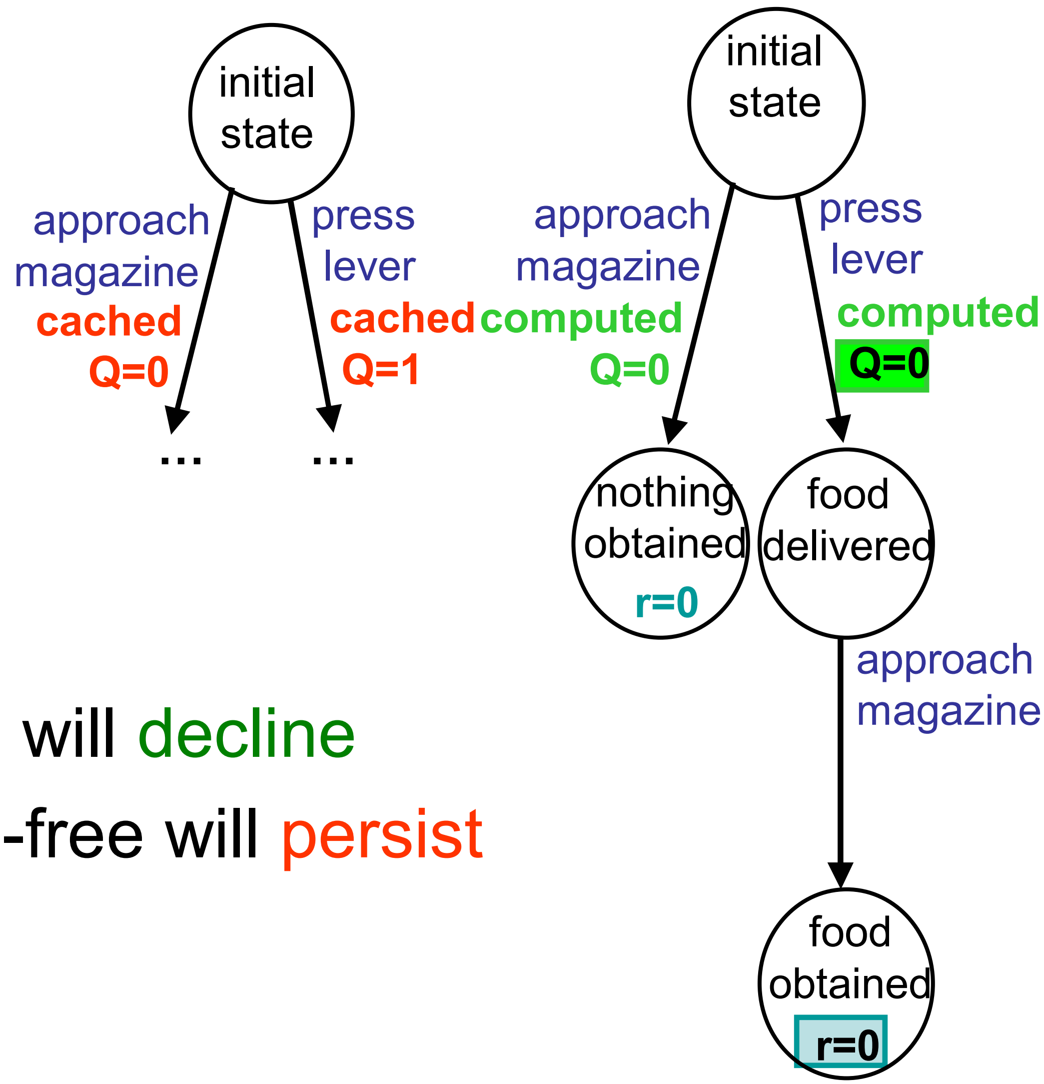
Behavioural experiment





Behavioural experiment

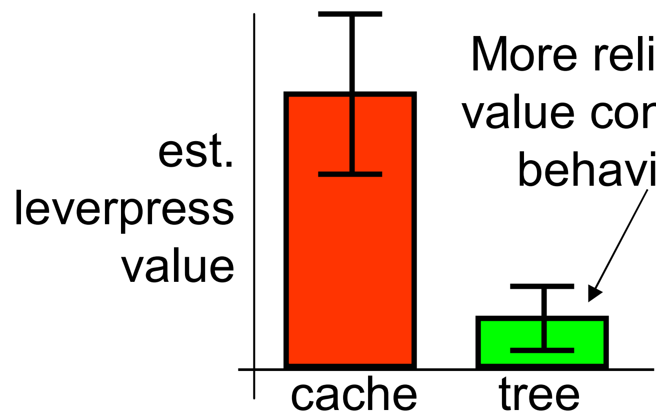




- Actions based on model will decline
- Actions based on model-free will persist

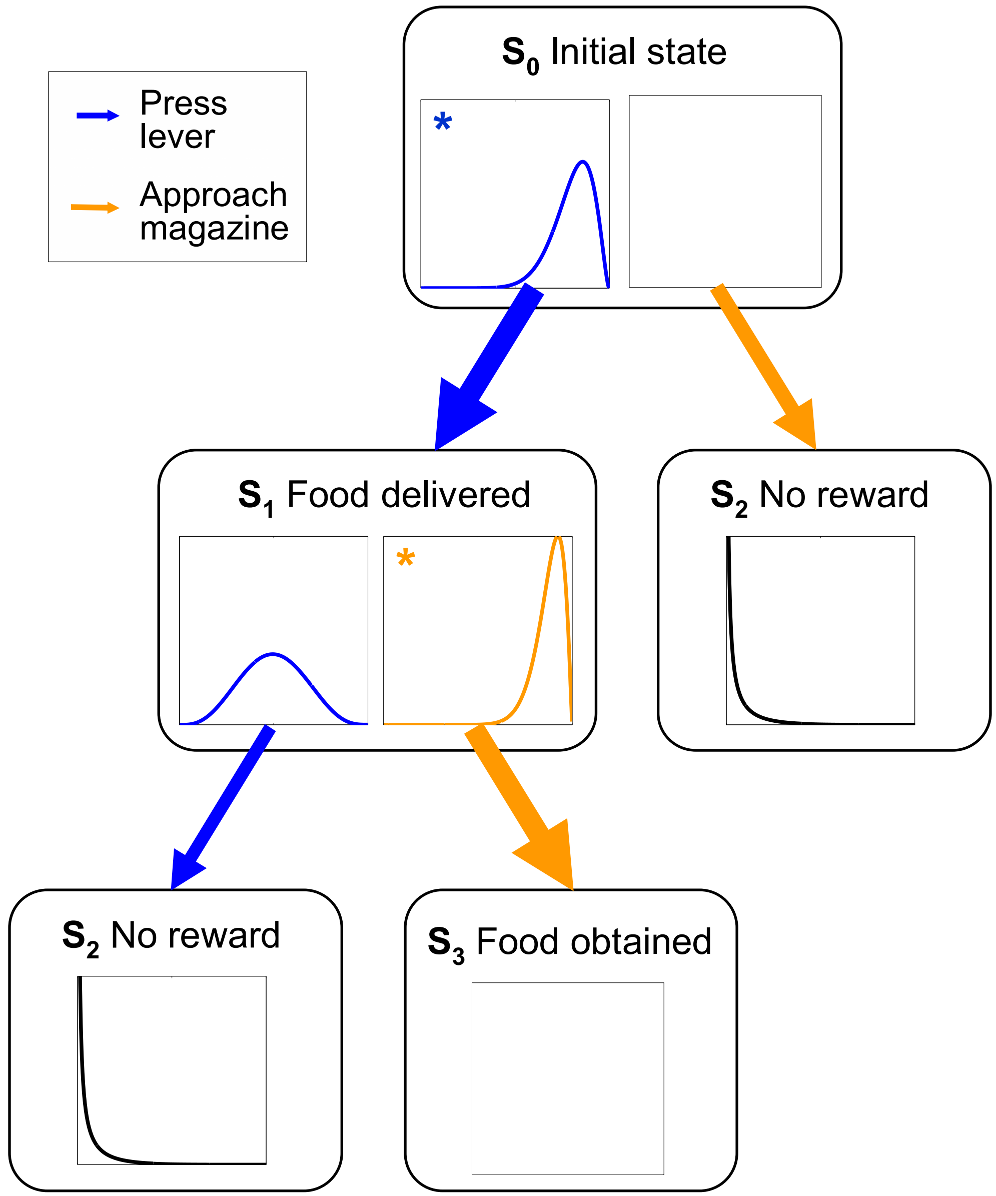
Suggested model

- 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)

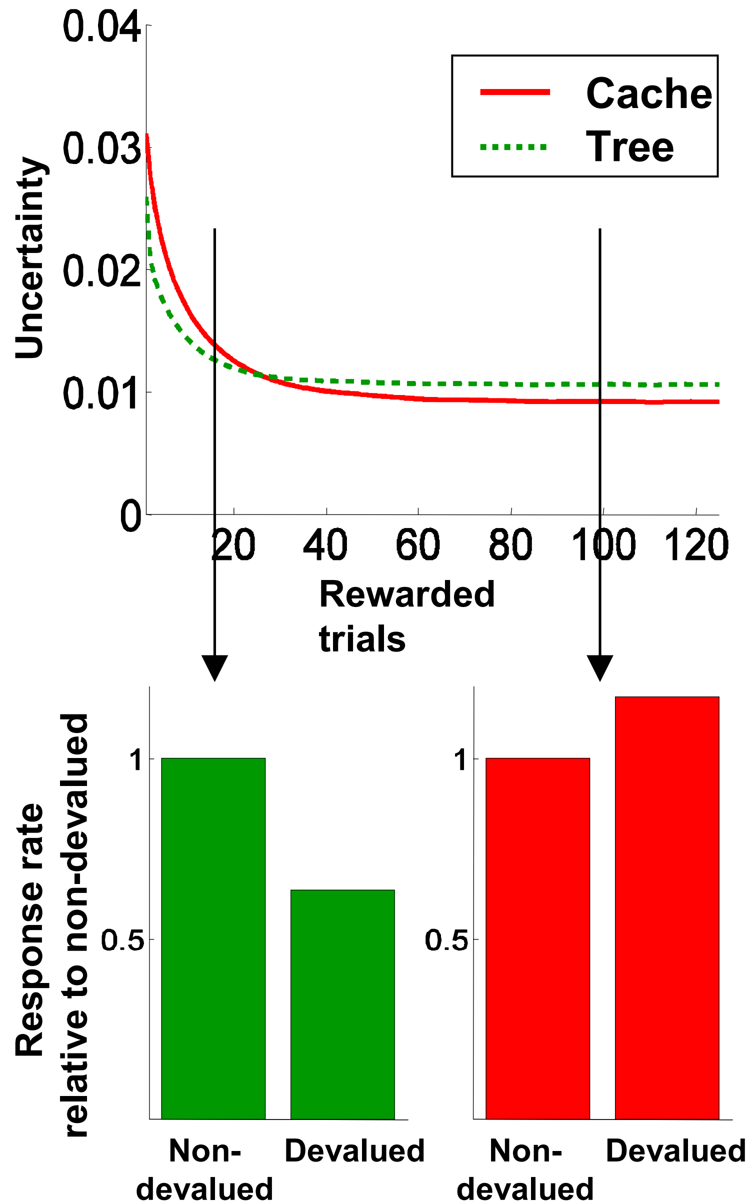


Uncertainty

- 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



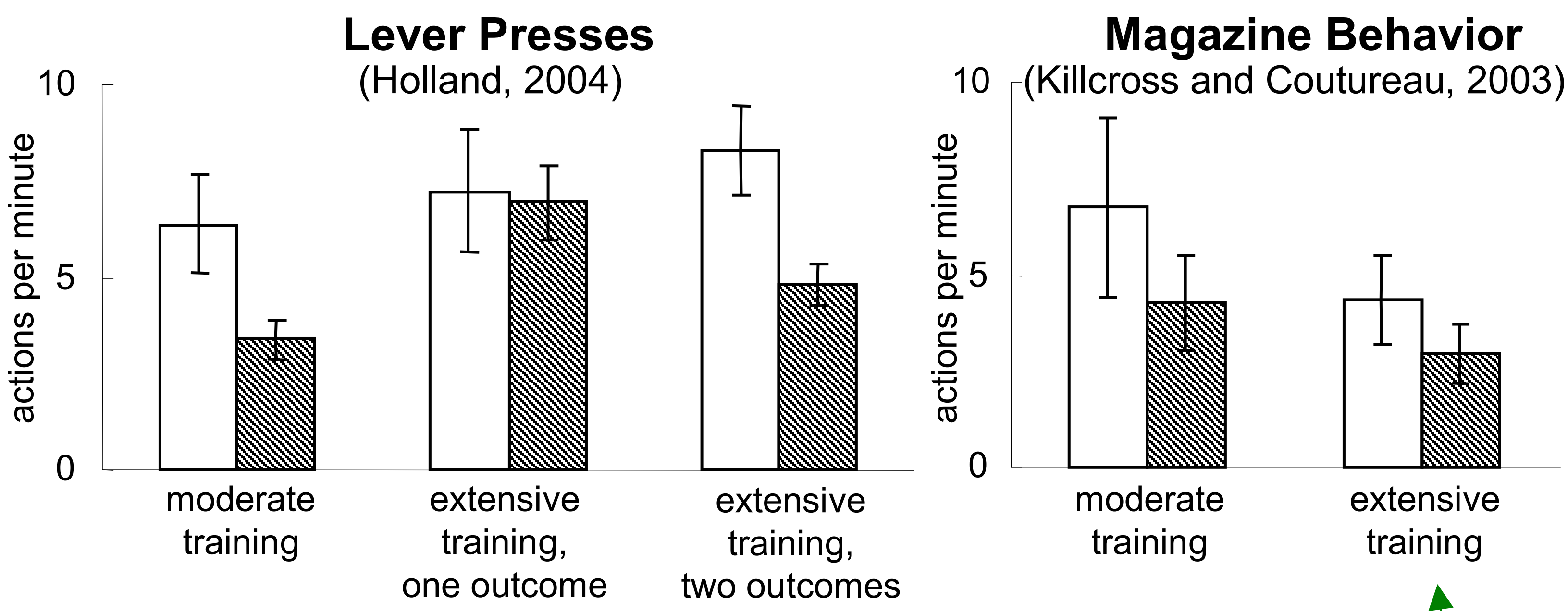
Additionally

- Model-based RL more useful near horizon

- Statistical inefficiency of model-free RL more difficult to overcome in more complex tasks

$\rightarrow$ Both factors should oppose habitization

Behavioural results

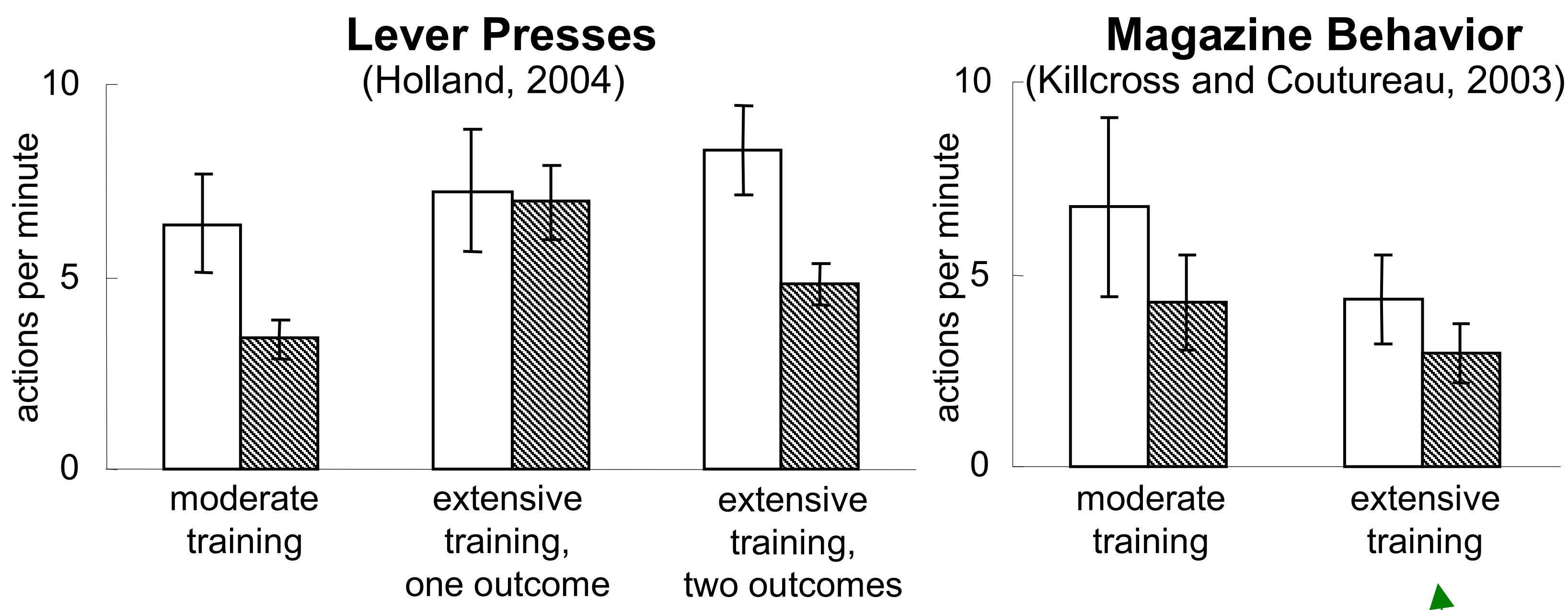


Habitisation with overtraining

… but not in tasks with multiple outcomes

… and not for actions proximal to reward

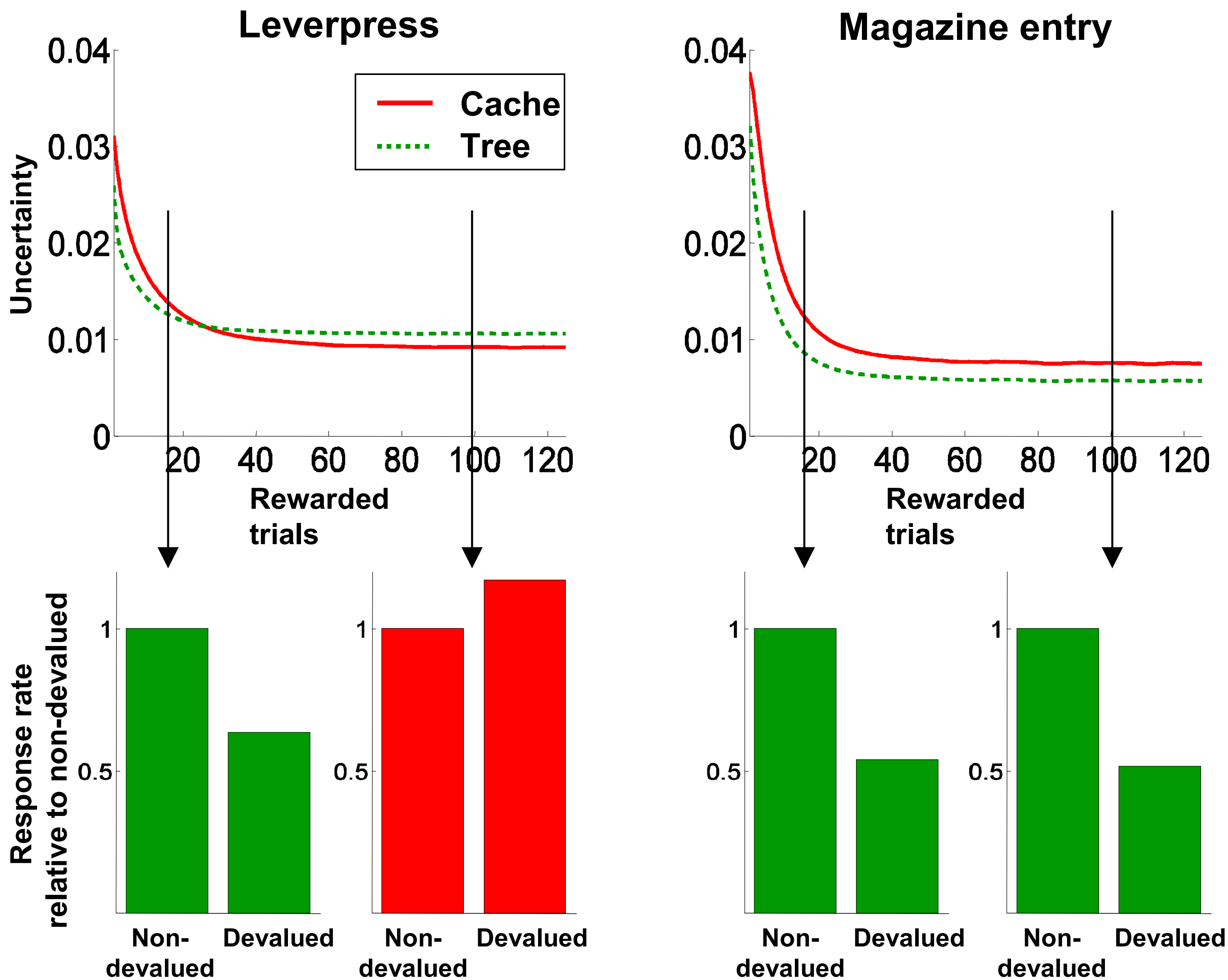
Behavioural results



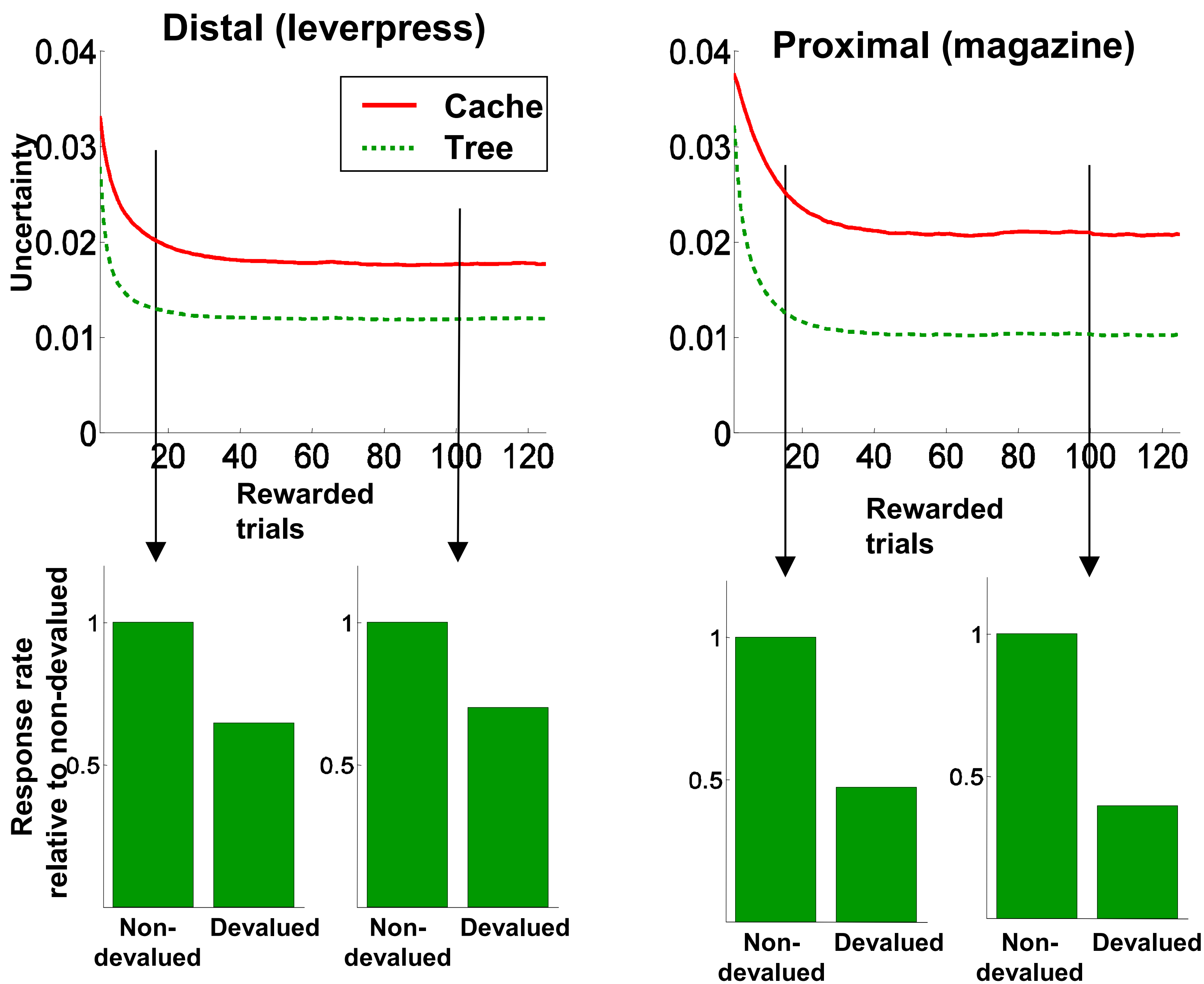
Data efficiency: overtraining and task complexity

Computational efficiency: search depth

Simulations



Two actions/two outcomes



Summary

- Dopaminergic learning for sequential choice
- 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?)