model-based planning I: motivations and methods

ccnss

2018.07.06

slides and references available at

http://aaron.bornstein.org/ccnss/

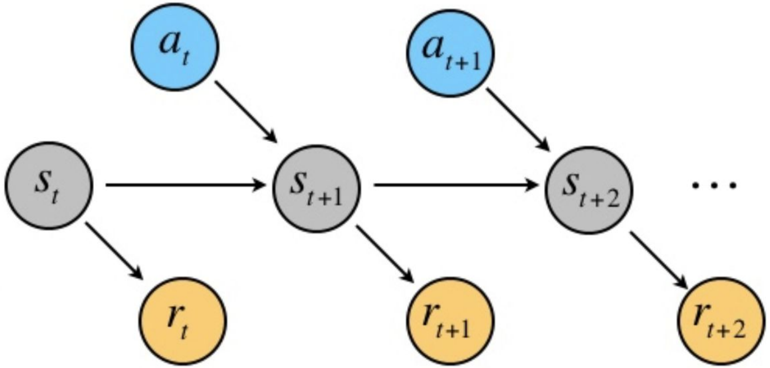
(p)review: mdps + tdrl

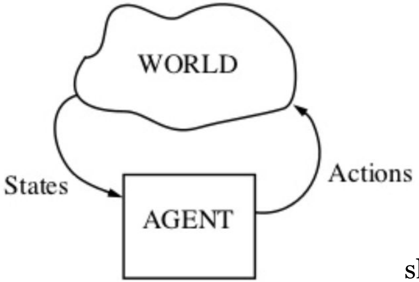
Known

- S - Set of States
- A - Set of Actions
- $\Pr(s' \mid a, s)$ - Transitions
- α - Starting State Distribution
- γ - Discount Factor

?

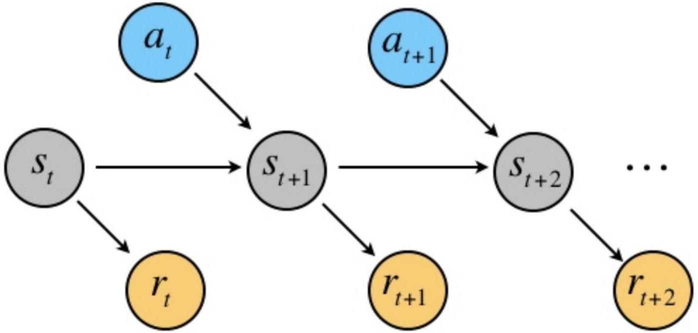
- $r(s)$ - Reward [or $r(s,a)$]





slide from kevin regan, U of T

(p)review: mdps + tdrl



$$Q(S, A) = Q(S, A) + \alpha[R - Q(S, A)]$$

(p)review: multi-armed bandit, action selection

(p)review: selection policy - decide how to decide

when Q(Left) > Q(Right):

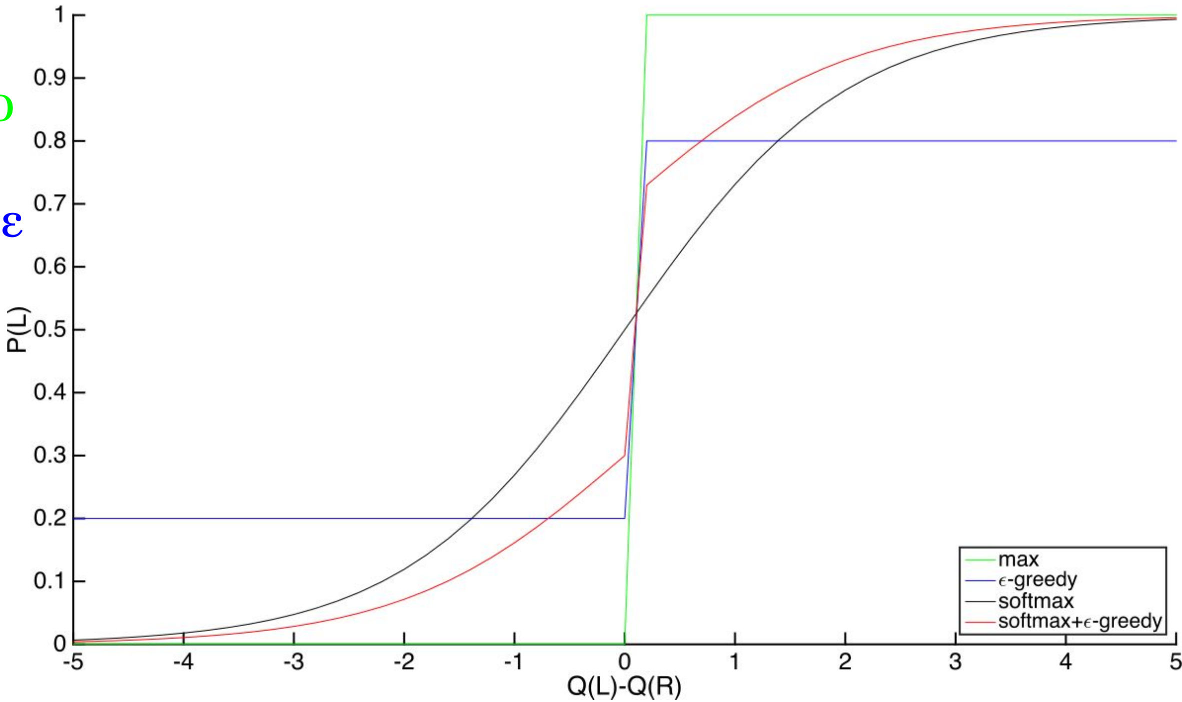
max: P(Left) = 1.0, P(Right) = 0.0

ε-greedy: P(Left)=1-ε, P(Right)=ε

softmax: P(Left) > P(Right)

ε-softmax: P(Left)-ε > P(Right)

$$P(L) = \frac{e^{\beta Q(L)}}{e^{\beta Q(L)} + e^{\beta Q(R)}}$$



ϵ-softmax: shteingart et al 2013

(p)review: q-learning with stochastic action selection

Q-learning (off-policy TD control) for estimating $\pi \approx \pi_*$

Algorithm parameters: step size $\alpha \in (0, 1]$, small $\varepsilon > 0$
Initialize $Q(s, a)$, for all $s \in \mathcal{S}^+, a \in \mathcal{A}(s)$, arbitrarily except that $Q(terminal, \cdot) = 0$

Loop for each episode:
    Initialize S
    Loop for each step of episode:
        Choose A from S using policy derived from Q (e.g., ε-greedy)
        Take action A, observe R, S'
        $Q(S, A) \leftarrow Q(S, A) + \alpha\left[R + \gamma \max_a Q(S', a) - Q(S, A)\right]$
        $S \leftarrow S'$
    until S is terminal

outline

I. motivations

II. behavioral signatures

III. neural substrates

IV. if time: open questions

outline

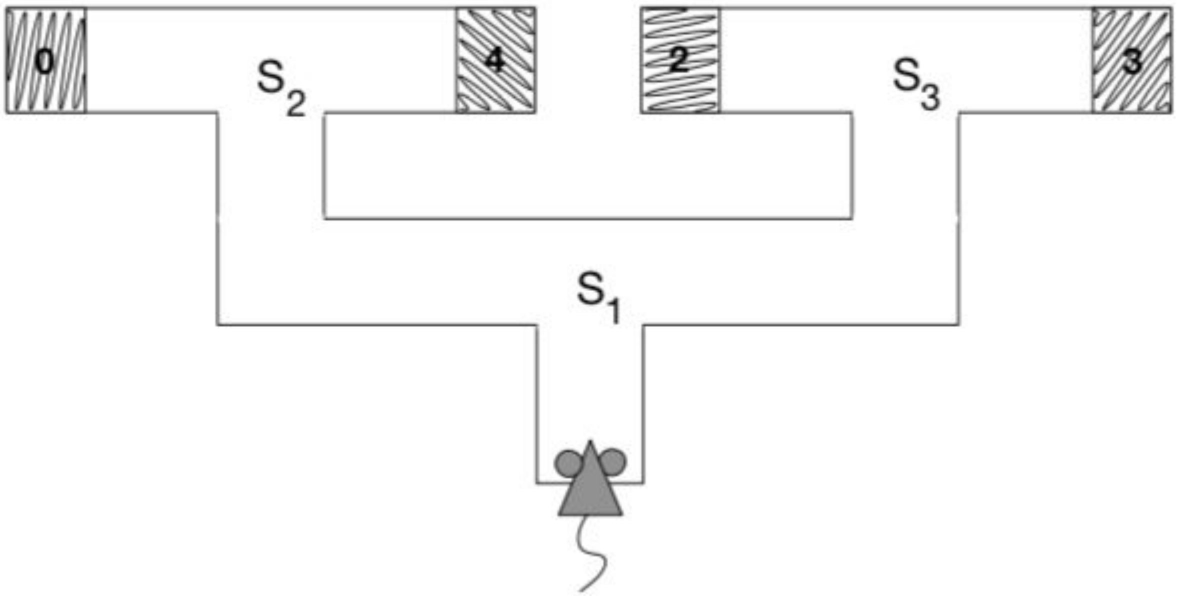
I. motivations

II. behavioral signatures

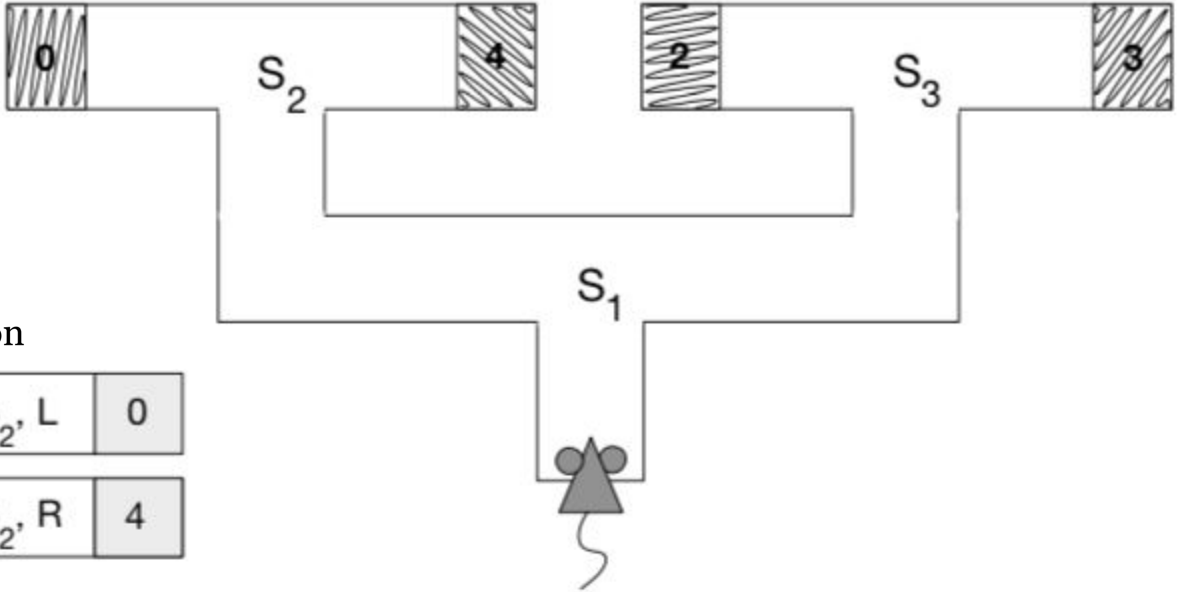
III. neural substrates

IV. if time: open questions

multi-step decisions



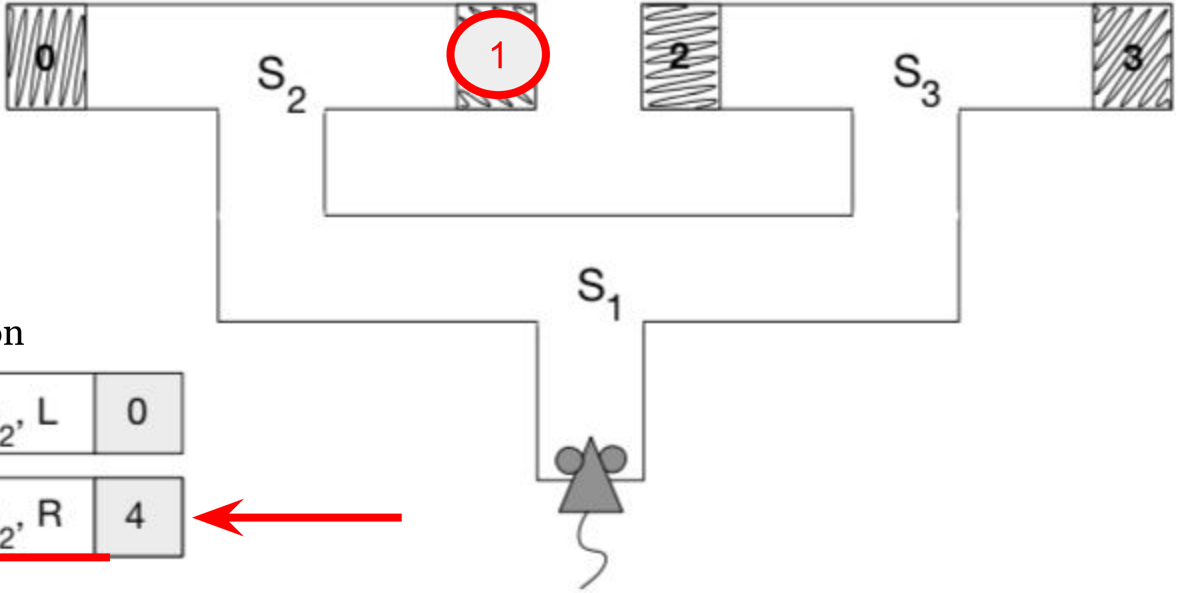
multi-step decisions



value function

S_1, L	4
S_1, R	3
S_2, L	0
S_2, R	4
S_3, L	2
S_3, R	3

outcome devaluation

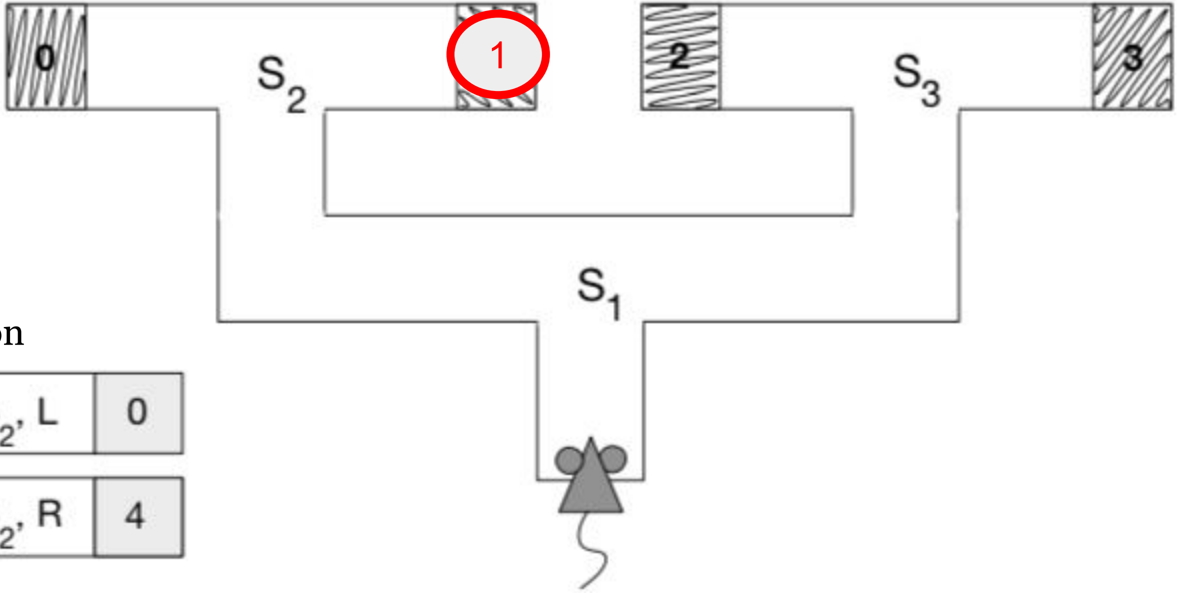


value function

S_2, L	0
S_2, R	4
S_1, L	4
S_1, R	3
S_3, L	2
S_3, R	3

(adams & dickinson 1981)

tdrl is "devaluation insensitive"

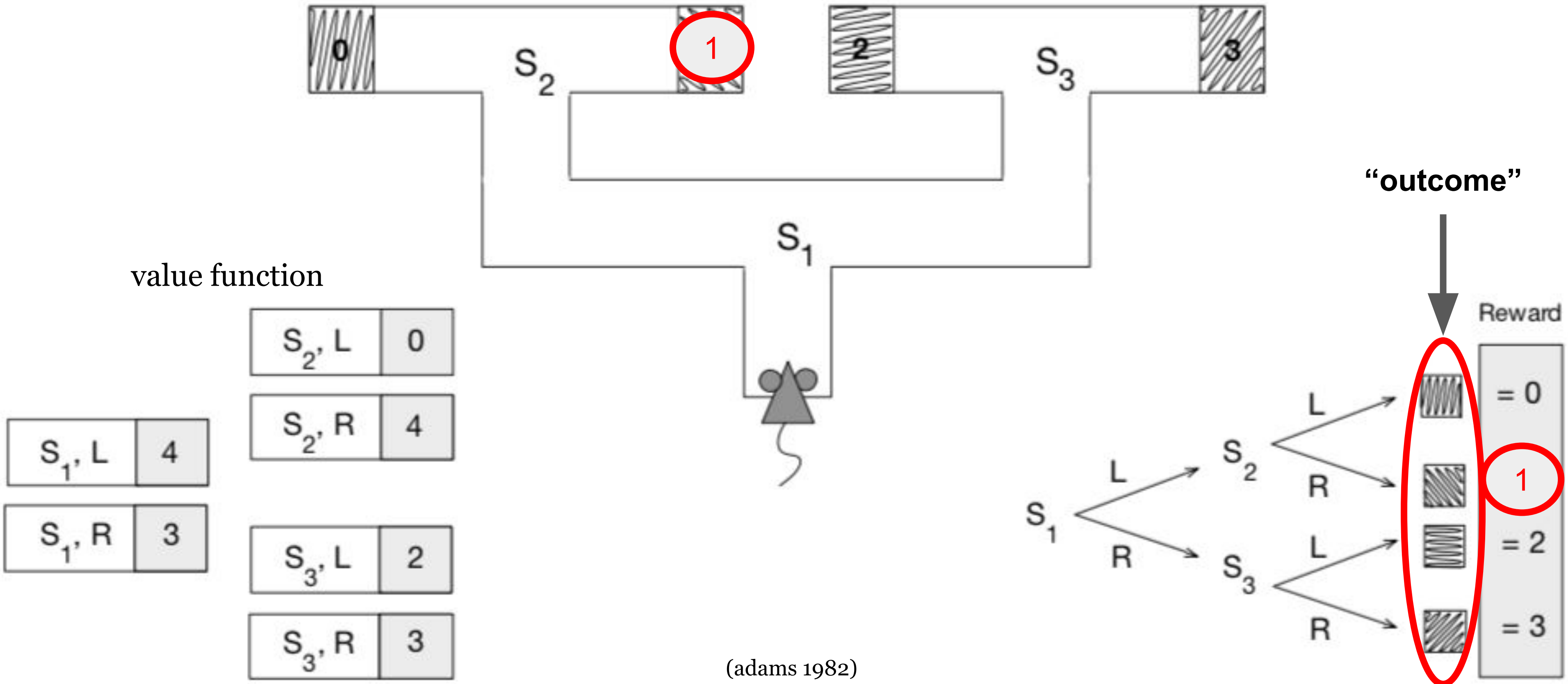


value function

S_1, L	4
S_1, R	3
S_2, L	0
S_2, R	4
S_3, L	2
S_3, R	3

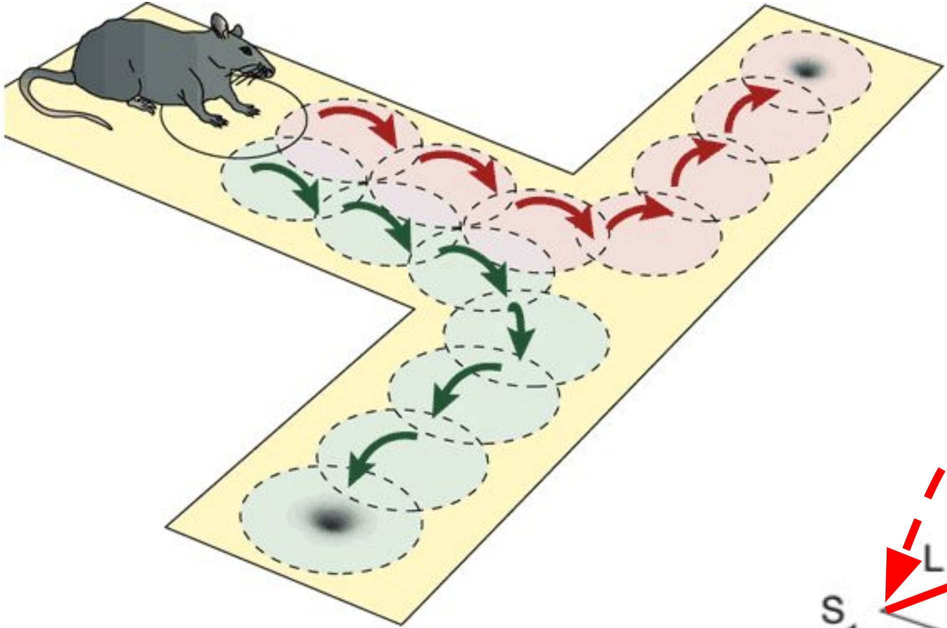
(adams & dickinson 1981)

outcome-sensitive



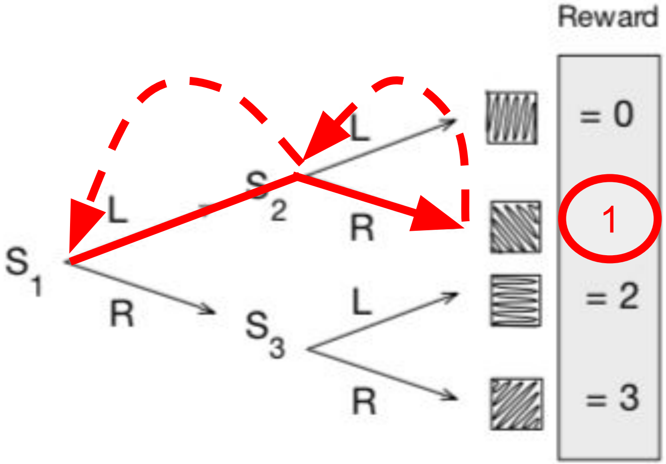
(adams 1982)

“online planning” with simulated experience

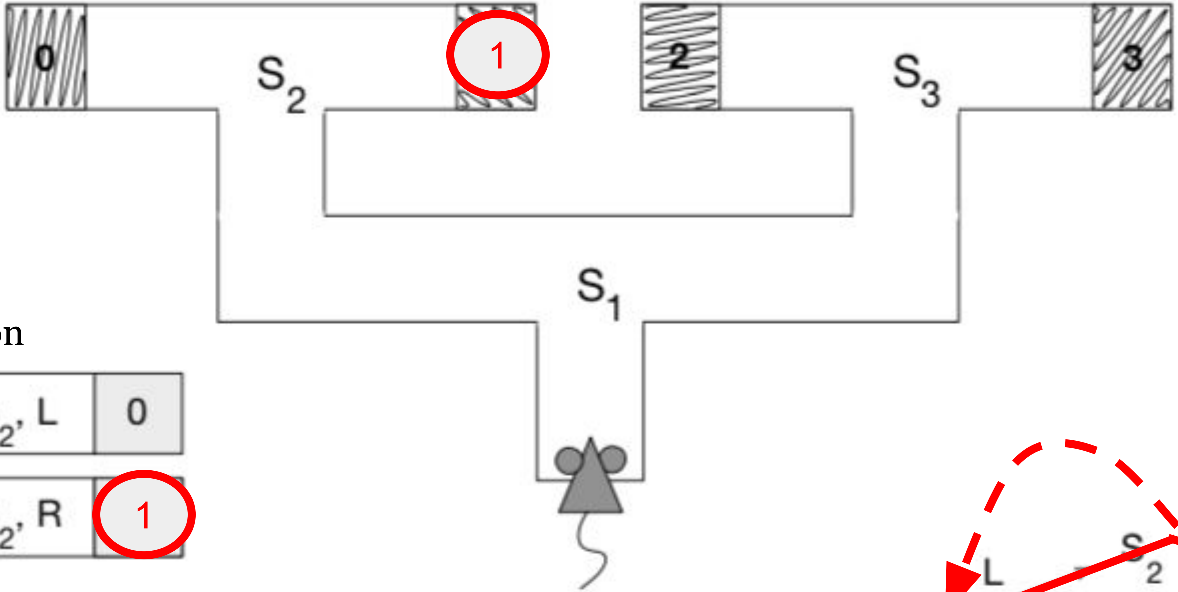


value function

S_1, L	1
S_1, R	3

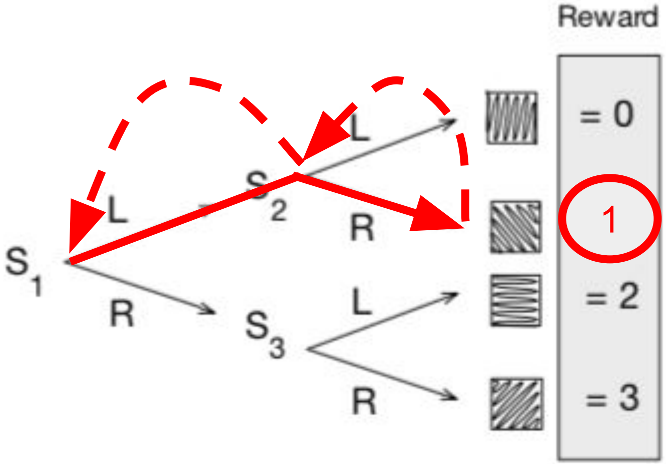


“offline planning” update via simulated outcomes



value function

S2, L	0
S2, R	1
S1, L	1
S1, R	3
S3, L	2
S3, R	3



dyna-q: "offline" updates using previous experience

Tabular Dyna-Q

Initialize $Q(s,a)$ and $Model(s,a)$ for all $s \in \mathcal{S}$ and $a \in \mathcal{A}(s)$

Loop forever:

(a) $S \leftarrow$ current (nonterminal) state

(b) $A \leftarrow \varepsilon\text{-greedy}(S, Q)$

(c) Take action A; observe resultant reward, R, and state, S'

(d) $Q(S,A) \leftarrow Q(S,A) + \alpha[R + \gamma \max_a Q(S',a) - Q(S,A)]$

(e) $Model(S,A) \leftarrow R, S'$ (assuming deterministic environment)

(f) Loop repeat n times:

  $S \leftarrow$ random previously observed state

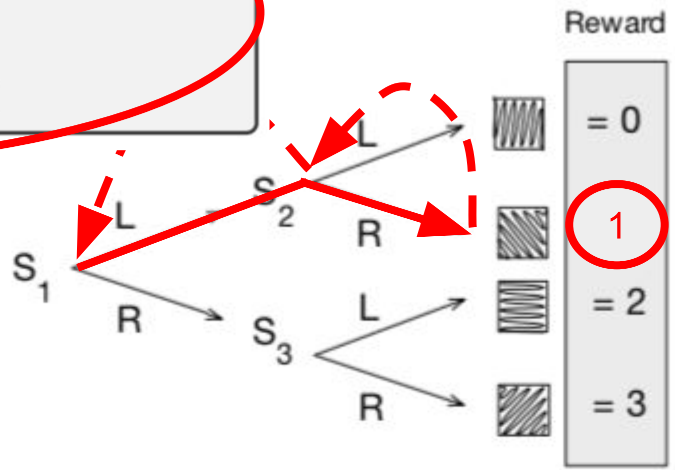
  $A \leftarrow$ random action previously taken in S

  $R, S' \leftarrow Model(S,A)$

  $Q(S,A) \leftarrow Q(S,A) + \alpha[R + \gamma \max_a Q(S',a) - Q(S,A)]$

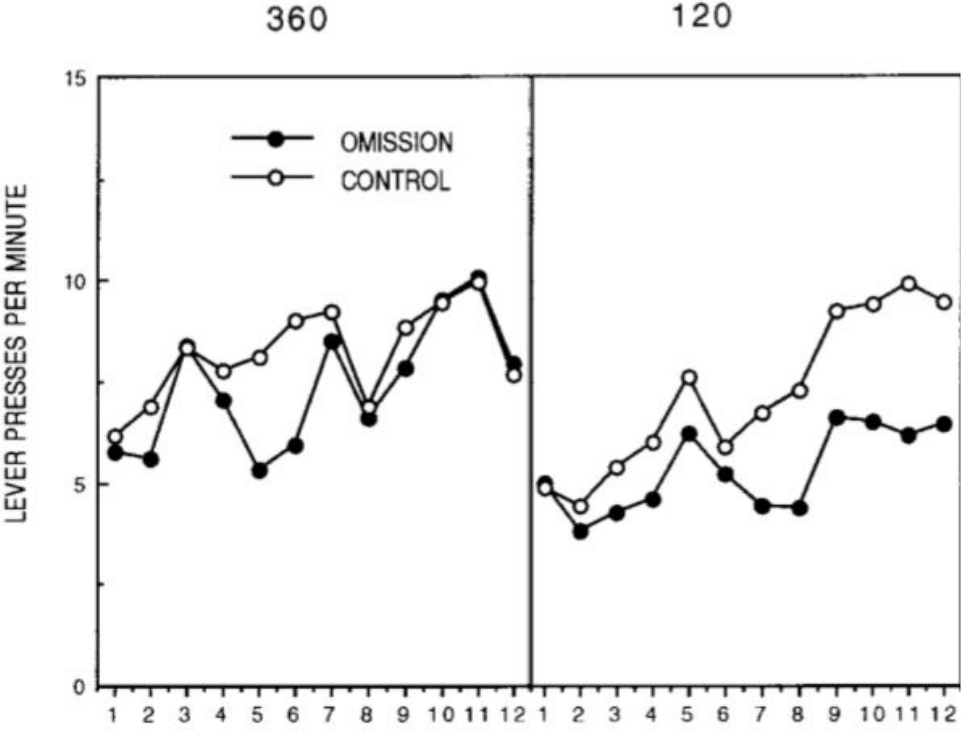
value function

S_1, L	1
S_1, R	3
S_2,	
S_2, R	1
S_3, L	2
S_3, R	3



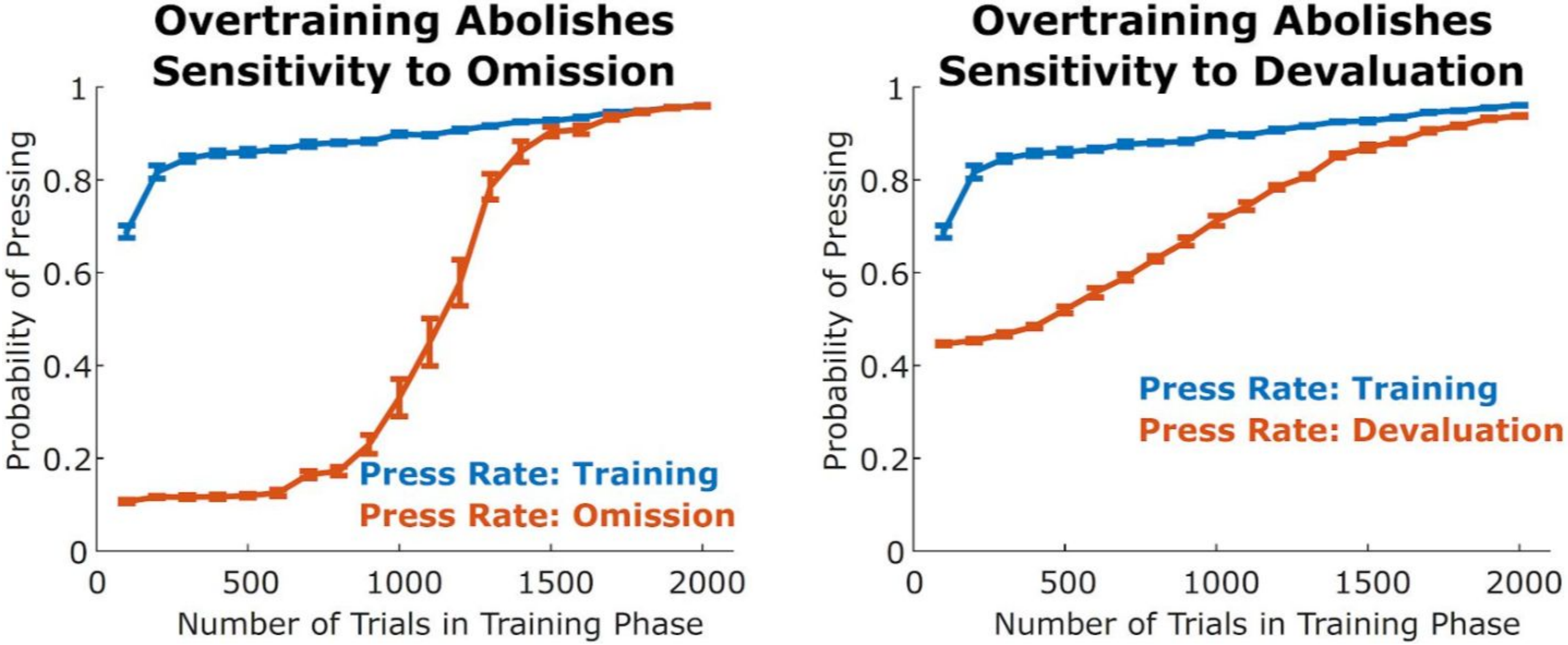
sutton 1992

“overtraining”



(adams 1982; dickinson et al 1998)

“overtraining”



(miller et al 2018)

interim summary

internal model ⇨ *simulated* experience

decision-time ("online") planning
background ("offline") planning

- allows sensitivity to changes in outcome value ("devaluation-sensitive")
    - *even with no direct experience!*
    - animals are, mostly, devaluation-sensitive
        - inference: they are using a "flexible" "action-*outcome*" (A-O) representation
    - … *unless* they are "overtrained"
        - inference: some other "stimulus-response" (S-R) representation takes over

outline

I. motivations

II. behavioral signatures

III. neural substrates

IV. if time: open questions

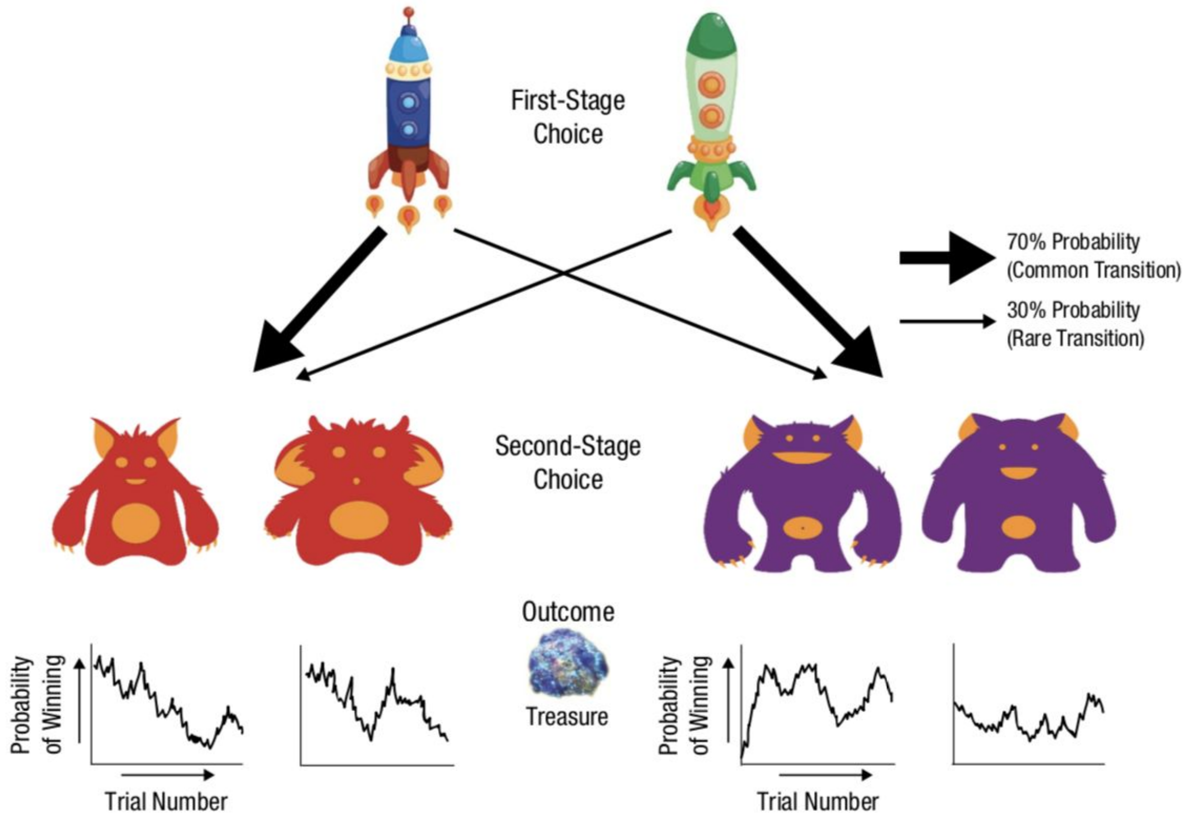
signatures of model-based planning

- sensitivity to outcome devaluation is one signature of model-based planning

- but not the most useful, in practice:
    - difficult to elicit overtraining / devaluation insensitivity in healthy humans
    - blocked tasks with coarse behavioral transition between “overtrained” and non-
    - would like a task that can elicit model-based and/or model-free behaviors, repeatedly

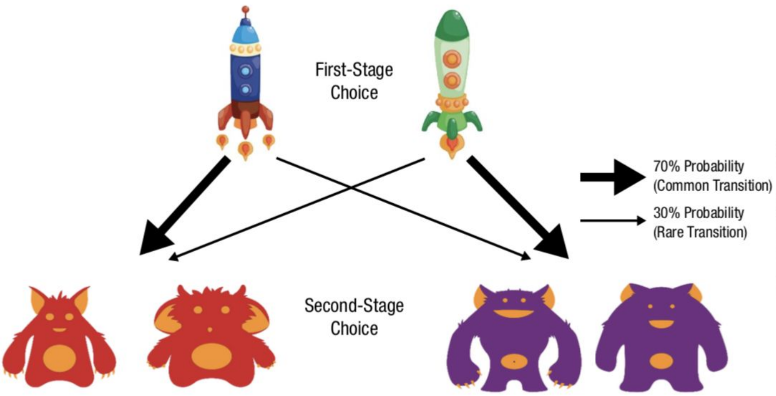
- another idea: test the model *update*

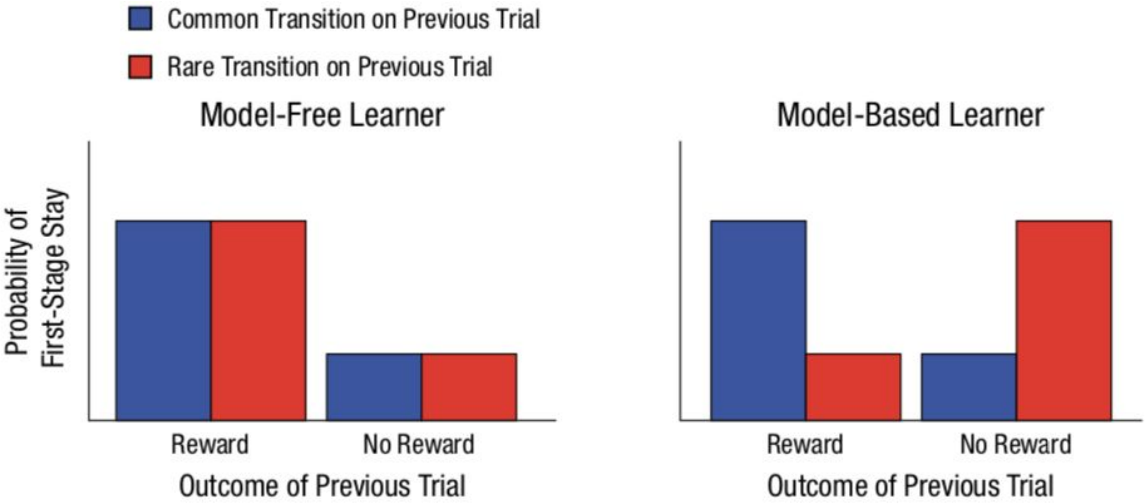
the “two-step task”



(daw et al 2011; decker et al 2016)

the “two-step task”

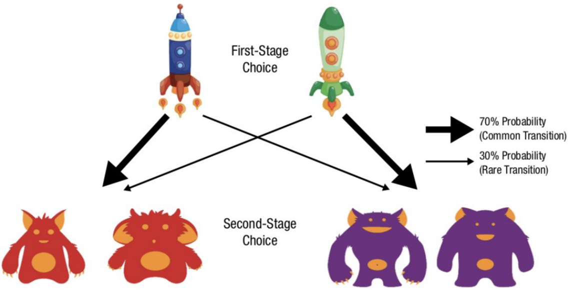


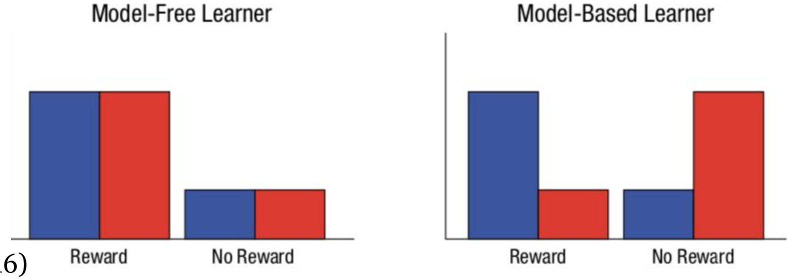


$$ModelFreeIndex = P(stay|RC) + P(stay|RU) - P(stay|OC) - P(stay|OU)$$
$$ModelBasedIndex = P(stay|RC) - P(stay|RU) - P(stay|OC) + P(stay|OU)$$

(daw et al 2011; decker et al 2016)

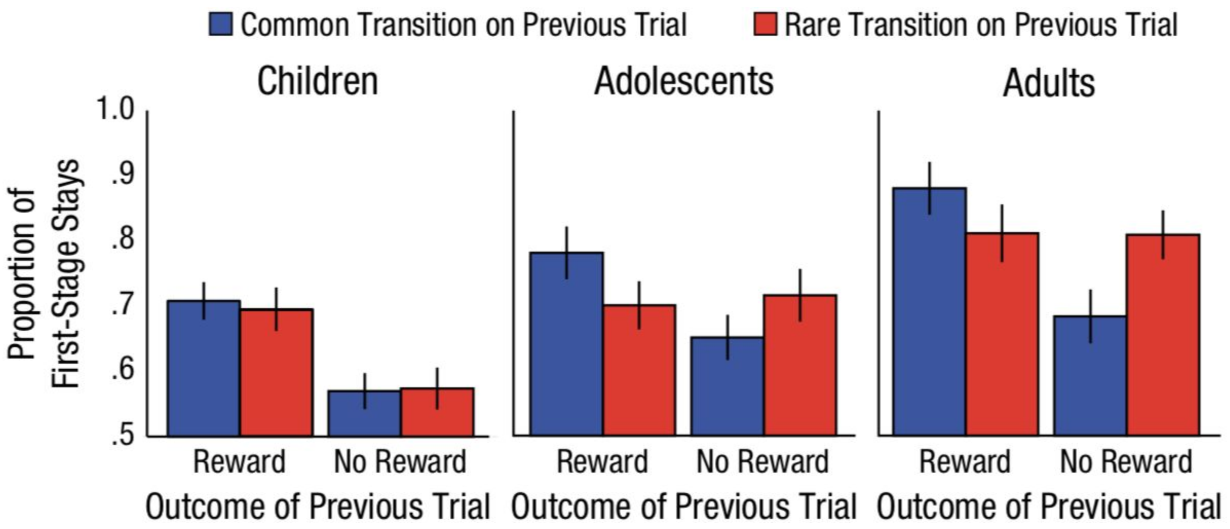
the “two-step task”

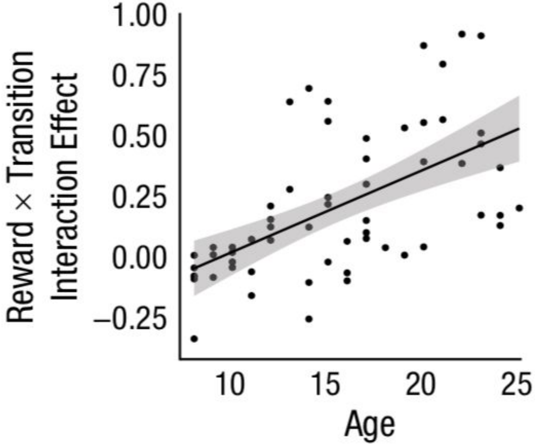


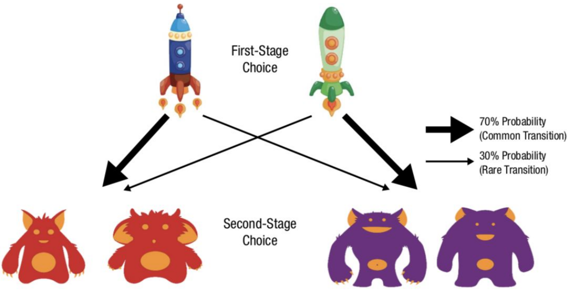


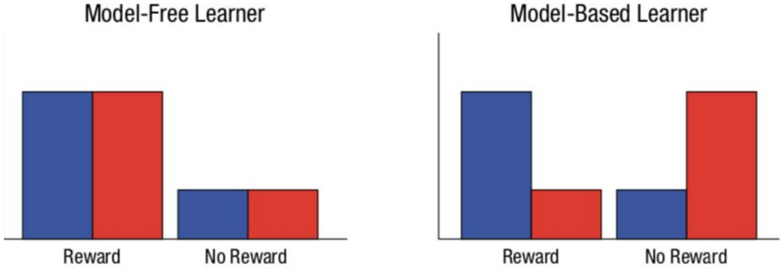
(daw et al 2011; decker et al 2016)

models in development



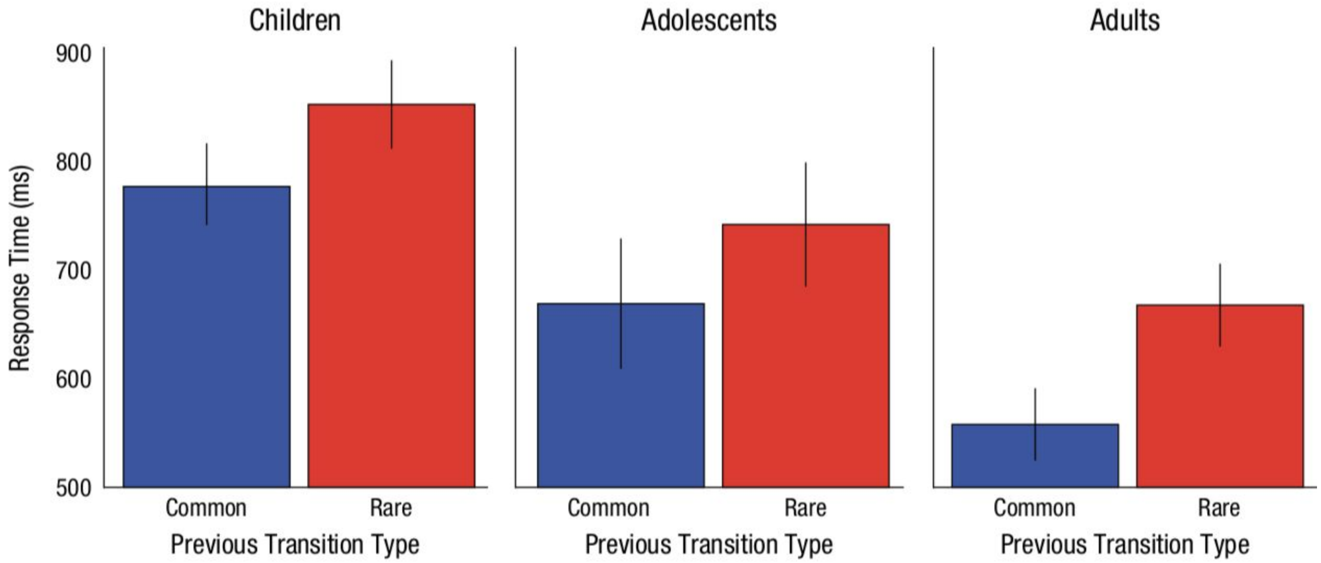


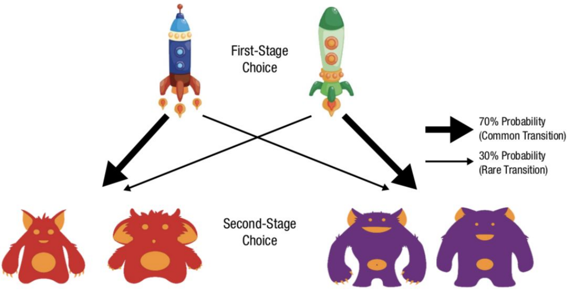




(decker et al 2016)

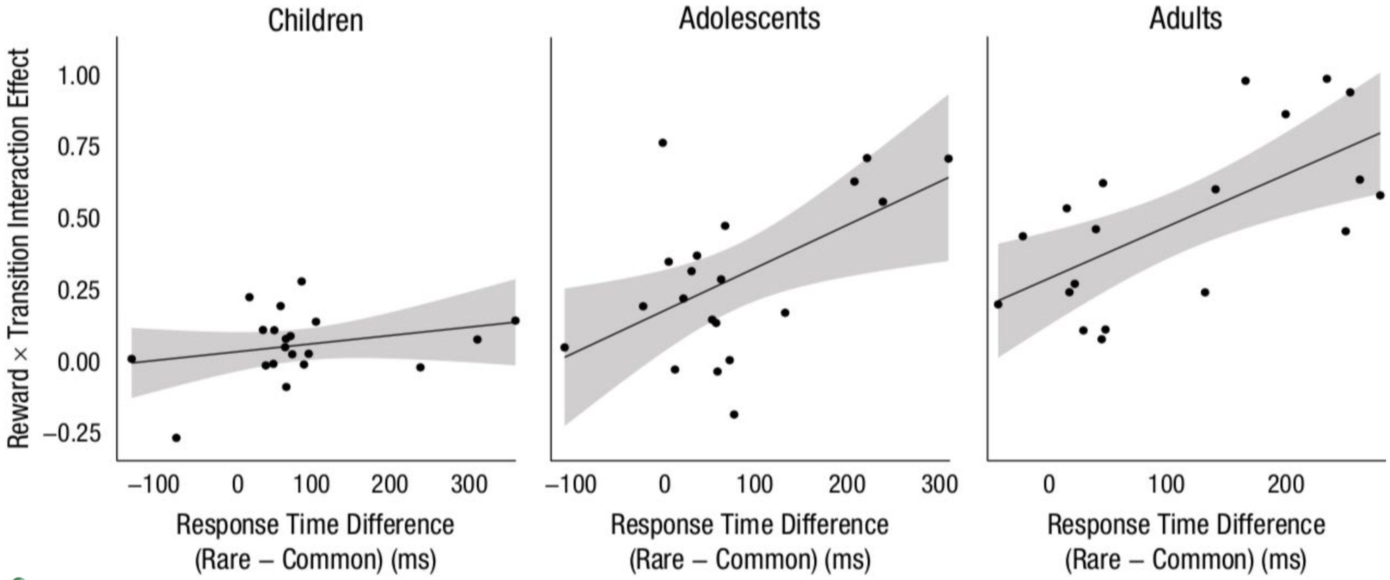
“implicit” model-based

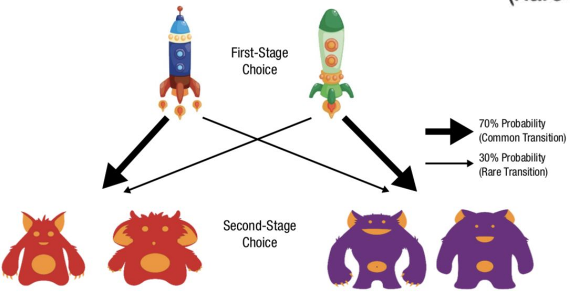




(decker et al 2016)

“implicit” model-based

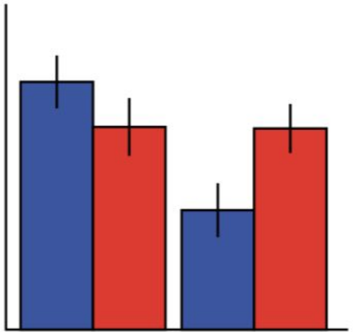




(decker et al 2016)

uncertainty-based arbitration

- if these are to be combined, how might they be combined?
- idea: "uncertainty-based arbitration" (daw et al 2005)
  - at state *S*, each controller (mb, mf) produces a candidate action *A*
  - these are **Bayesian**, not point estimates - they carry distributions over Q(s,a)
  - thus they code for the **uncertainty** of each controller
- the source of the uncertainty depends on the controller
  - model-free uncertainty arises from little experience
    - width of the posterior of Q(s,a)
  - model-based uncertainty arises from
    - estimation variance, e.g. width of the posterior of the transition function, due to computational "noise" — presumed heuristics (such as tree search strategies) of online planning
- explains transition from flexible to inflexible behavior



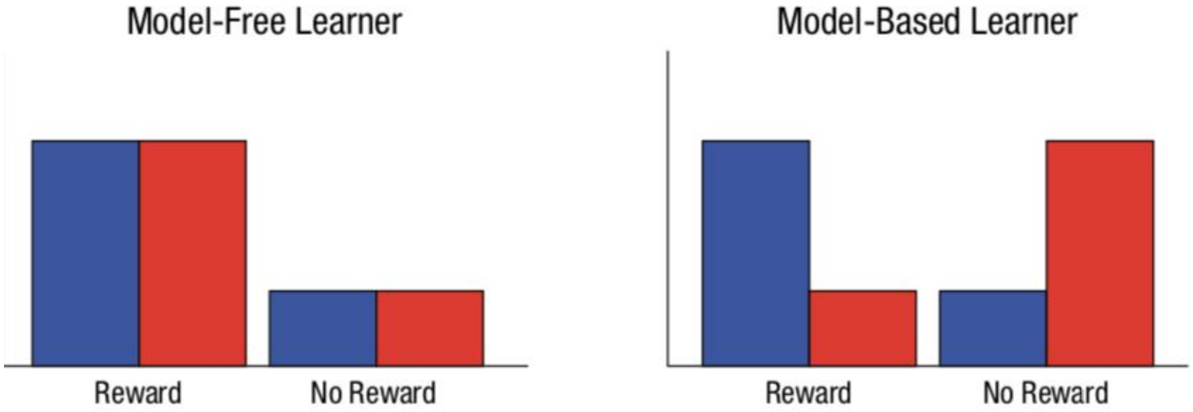
(daw et al 2005)

“model-basedness” as a personality trait

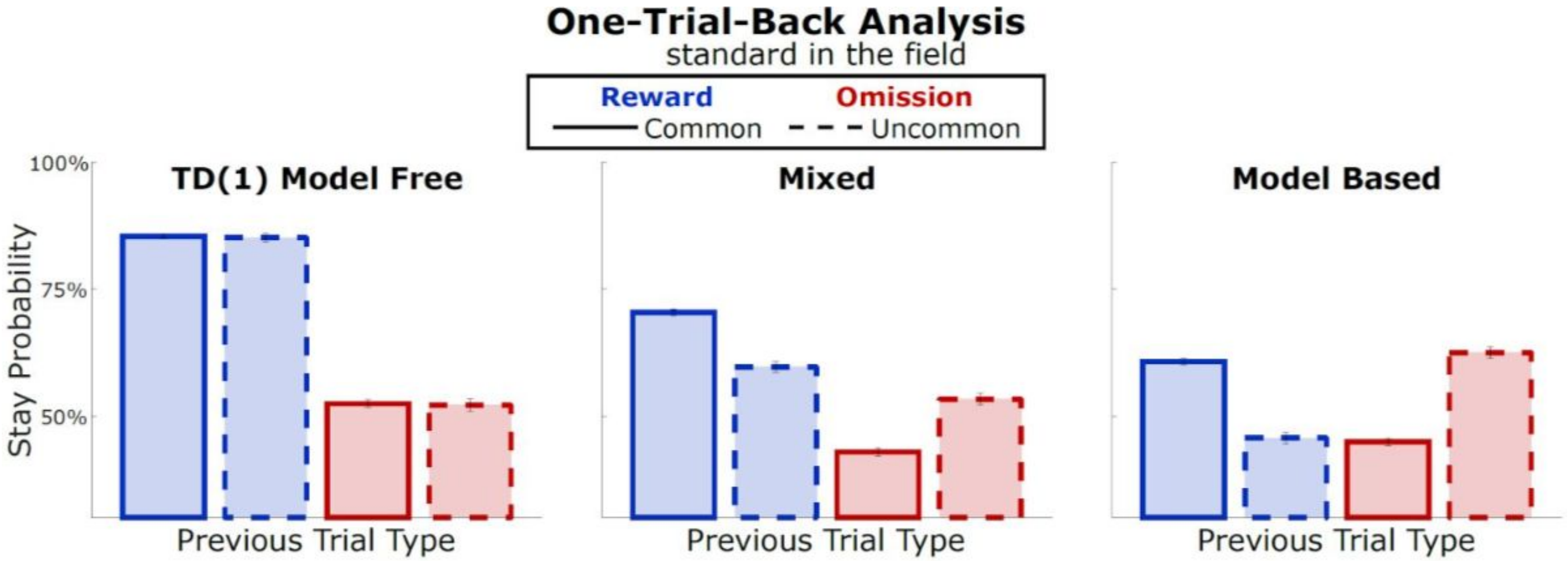
- “model-based” index correlates with a variety of stable or semi-stable personality traits
    - working memory span (otto et al 2014)
    - moral judgements (crockett 2016)
    - negatively with compulsion disorders (gillan et al 2015, 2016; voon et al 2015)
    - negatively with schizophrenia symptoms (culbreth et al 2016)
    - patience in *deliberative* (not reflexive) intertemporal choice (shenhav et al 2016; hunter, bornstein, hartley in prep; cf solway et al 2017)

one back stay/switch

$$ModelFreeIndex = P(stay|RC) + P(stay|RU) - P(stay|OC) - P(stay|OU)$$
$$ModelBasedIndex = P(stay|RC) - P(stay|RU) - P(stay|OC) + P(stay|OU)$$

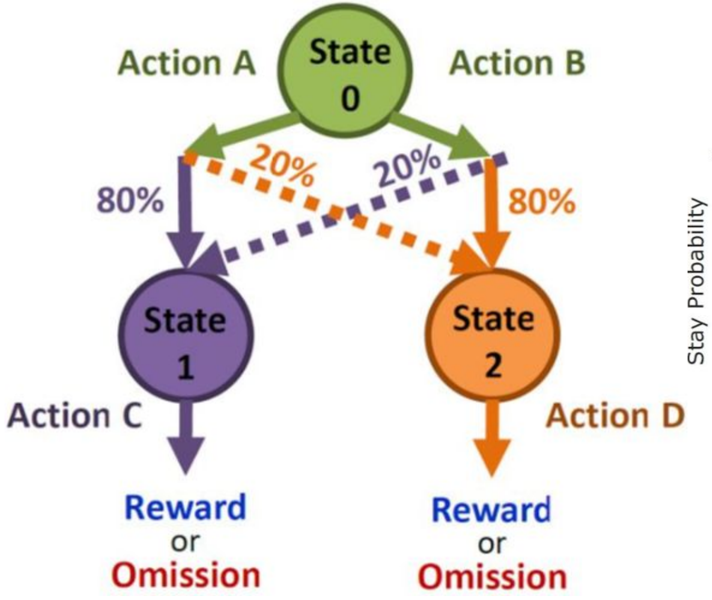


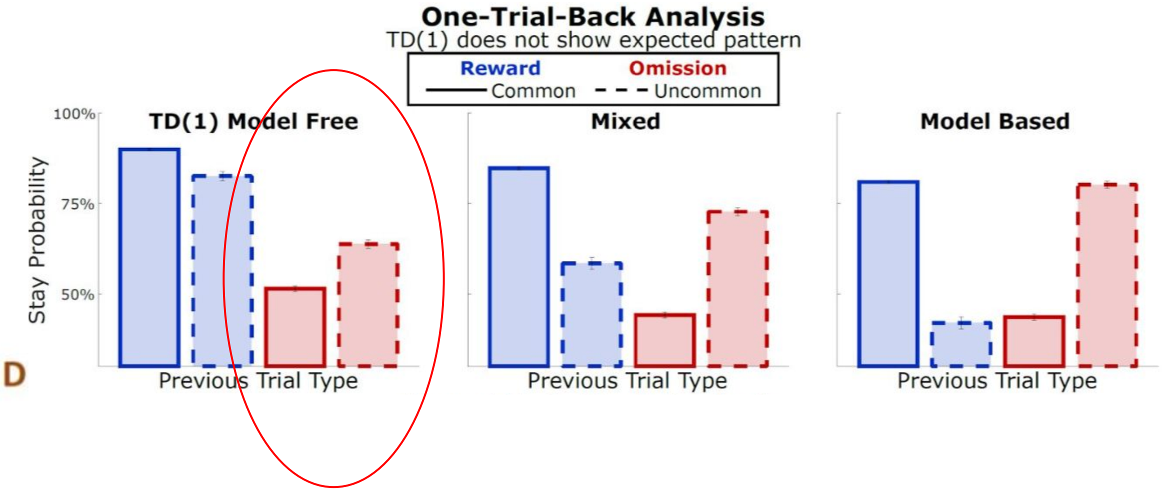
n-back to the future



(miller et al 2016)

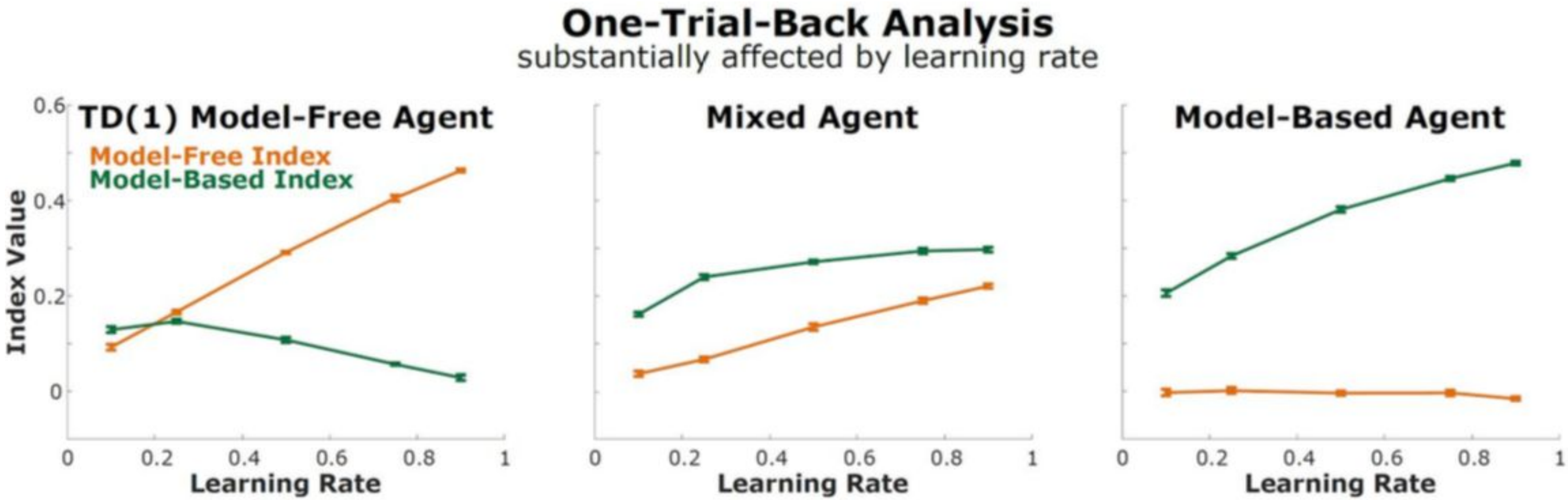
model-free looks model-based in less-stochastic task





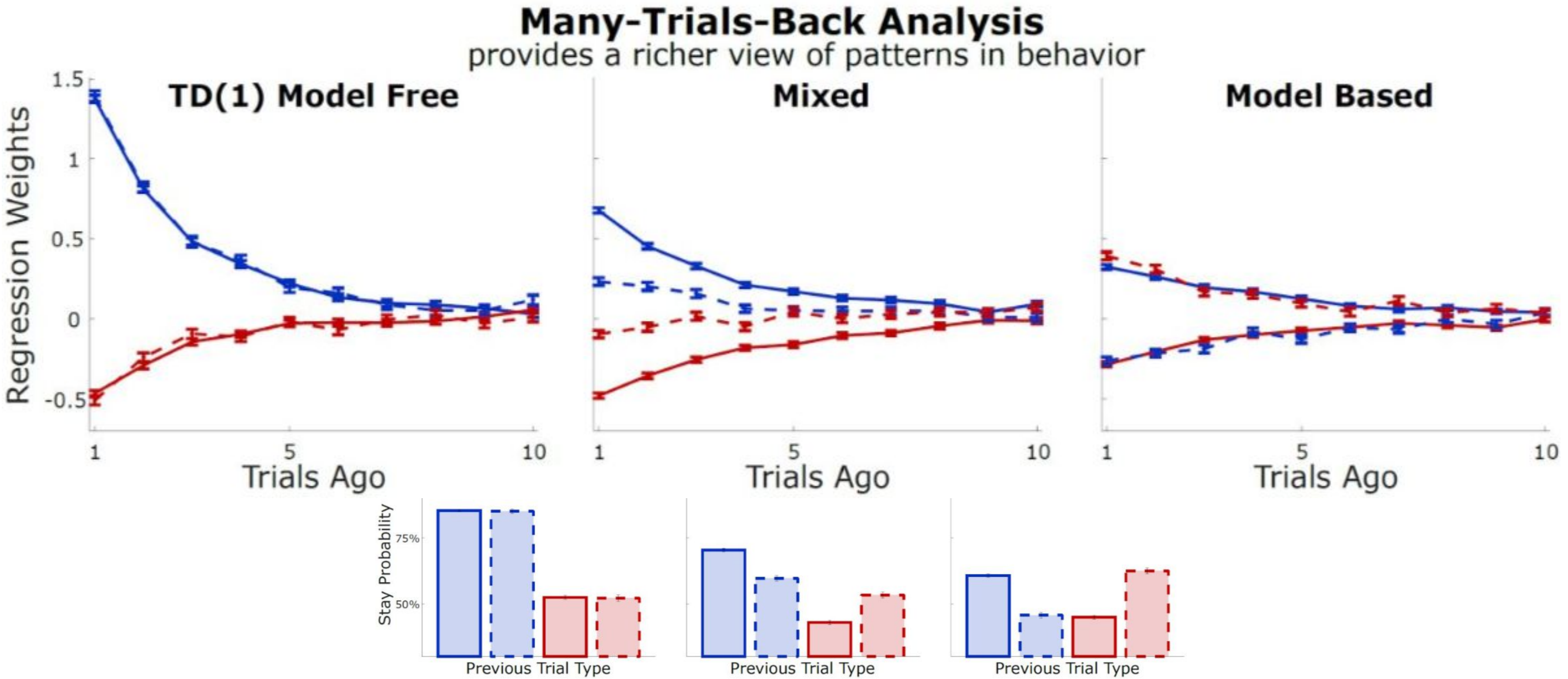
(miller et al 2016)

“slow” model-free can look model-based



(miller et al 2016)

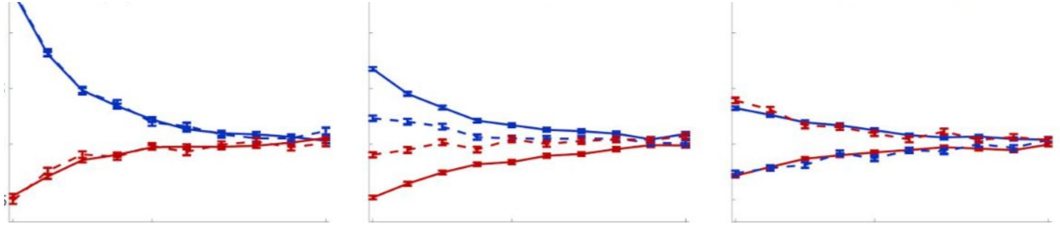
n-back to the future



(miller et al 2016)

n-back to the future

$$log\left(\frac{P_{left}(t)}{P_{right}(t)}\right) = \sum_{\tau=1}^{T} \beta_{RC}(\tau) * RC(t-\tau)$$



$$+ \sum_{\tau=1}^{T} \beta_{RU}(\tau) * RU(t-\tau)$$

$$+ \sum_{\tau=1}^{T} \beta_{OC}(\tau) * OC(t-\tau)$$

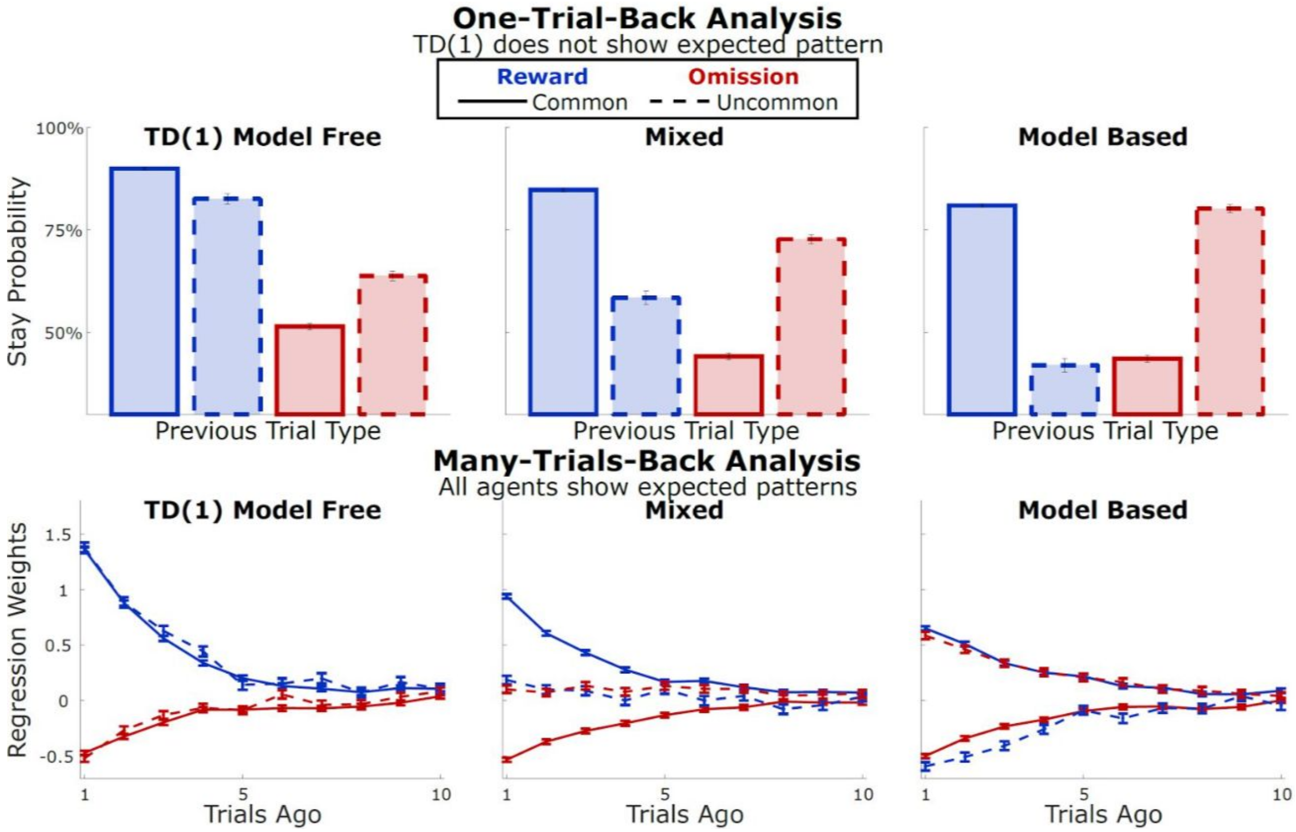
$$+ \sum_{\tau=1}^{T} \beta_{OU}(\tau) * OU(t-\tau)$$

$$ModelFreeIndex = \sum_{\tau=1}^{T} [\beta_{RC}(\tau) + \beta_{RU}(\tau)] - \sum_{\tau=1}^{T} [\beta_{OU}(\tau) + \beta_{OC}(\tau)]$$

$$PlanningIndex = \sum_{\tau=1}^{T} [\beta_{RC}(\tau) - \beta_{RU}(\tau)] + \sum_{\tau=1}^{T} [\beta_{OU}(\tau) - \beta_{OC}(\tau)]$$

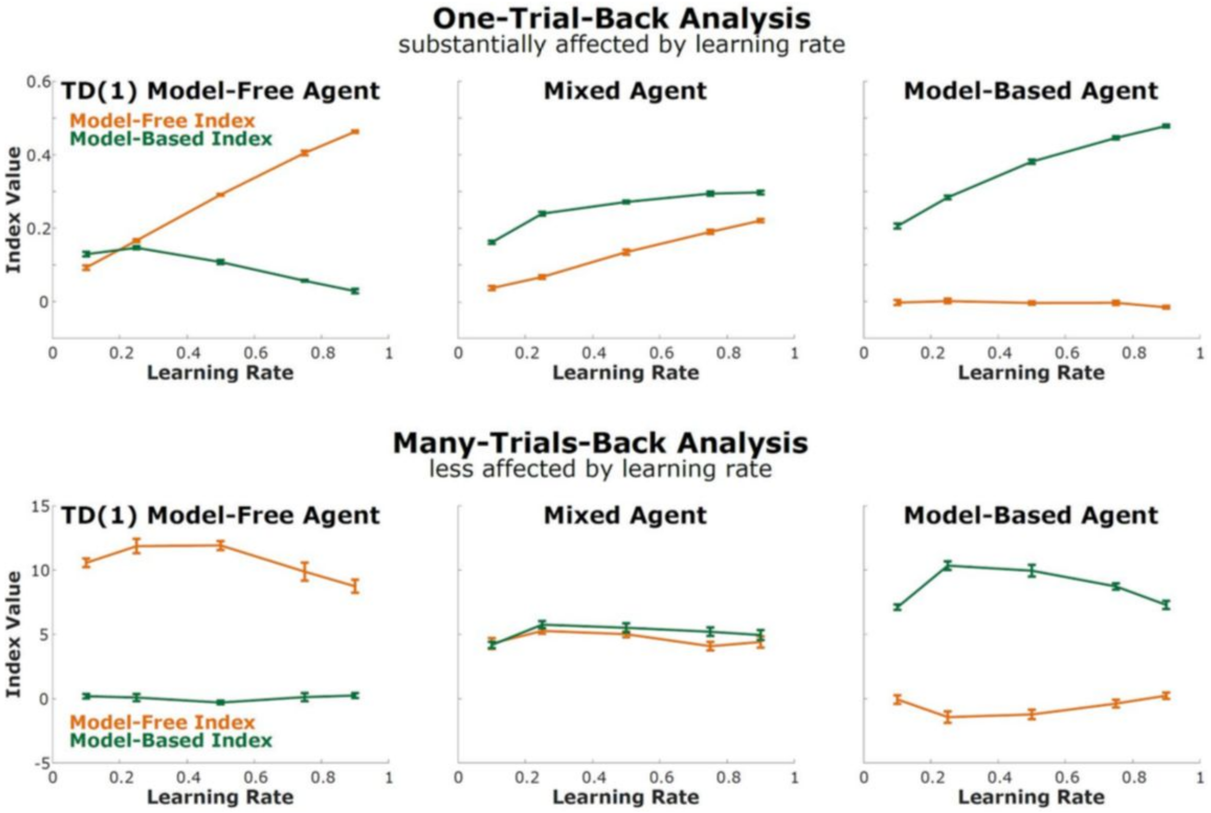
(miller et al 2016)

model-free can look model-based



(miller et al 2016)

“slow” model-free can look model-based

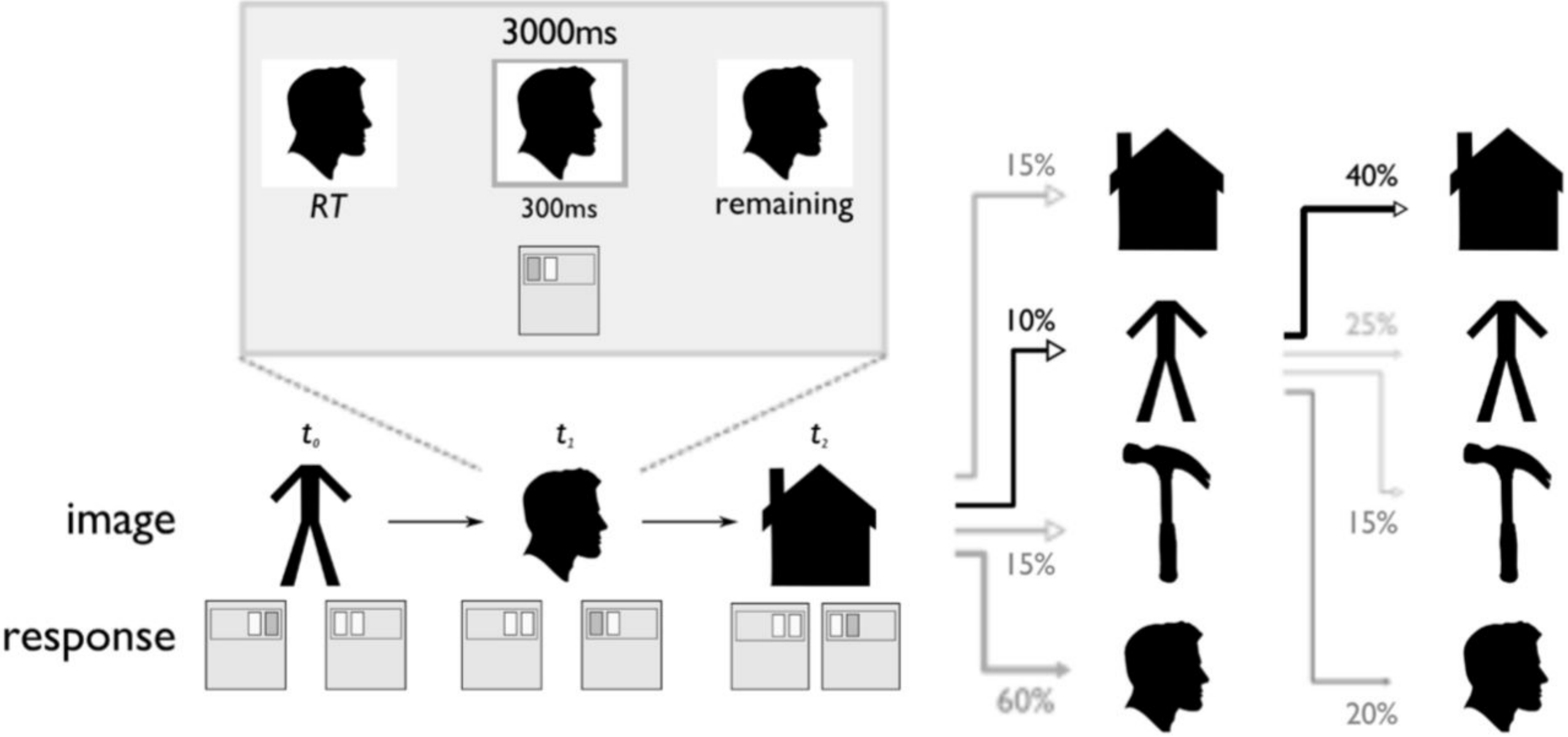


(miller et al 2016)

“latent” learning

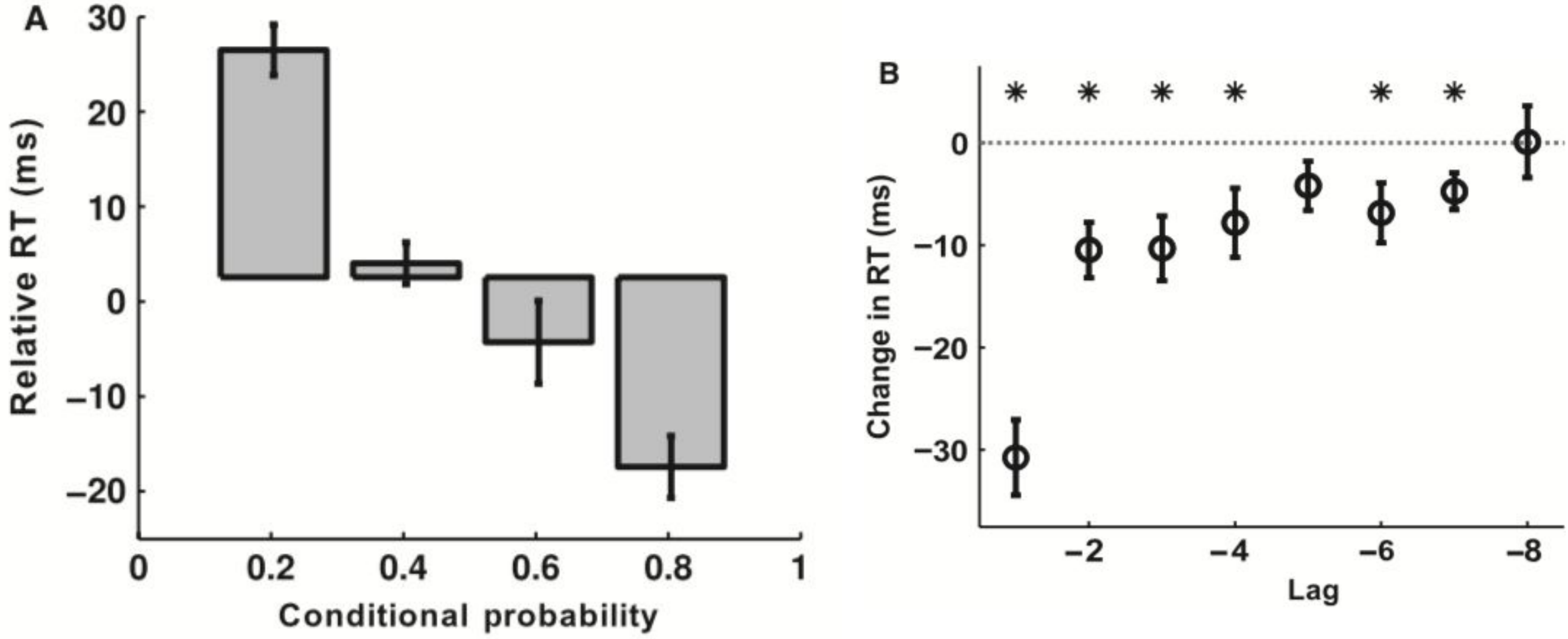
(bornstein & daw 2012, 2013)

“latent” learning



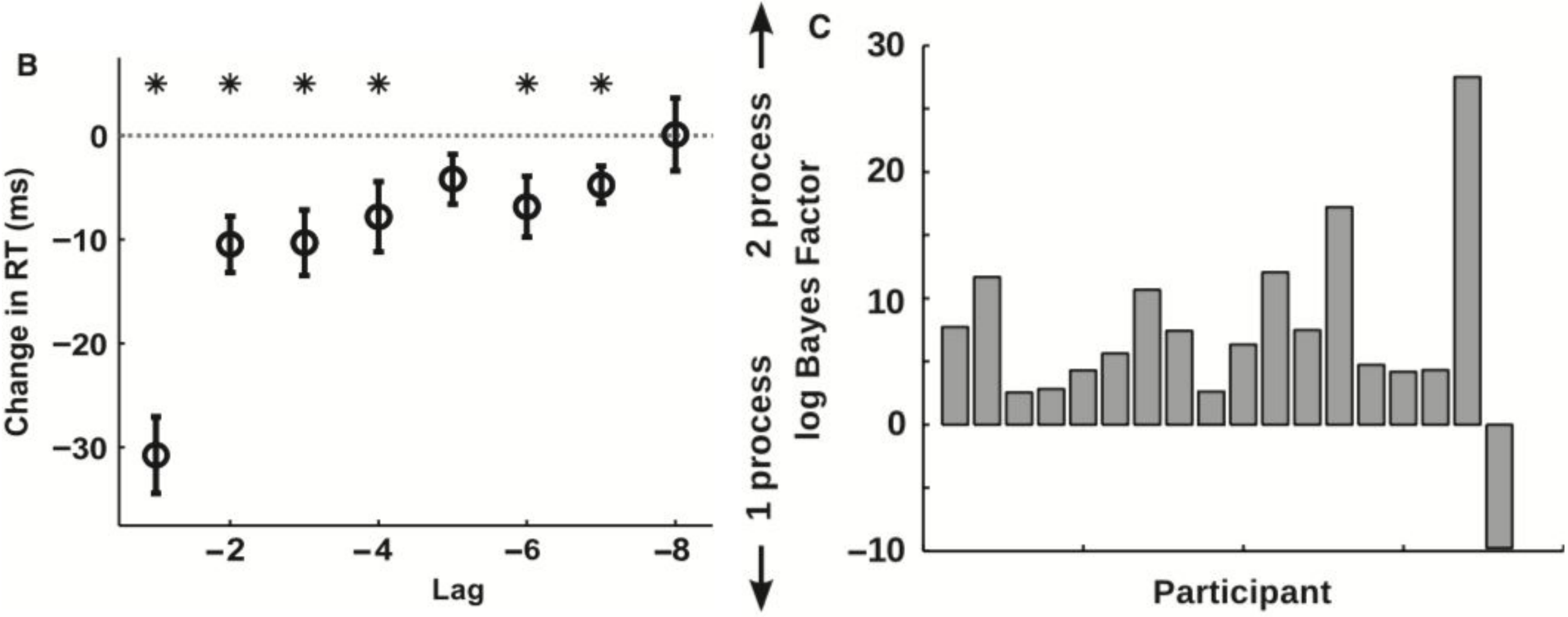
(bornstein & daw 2012, 2013)

“latent” learning



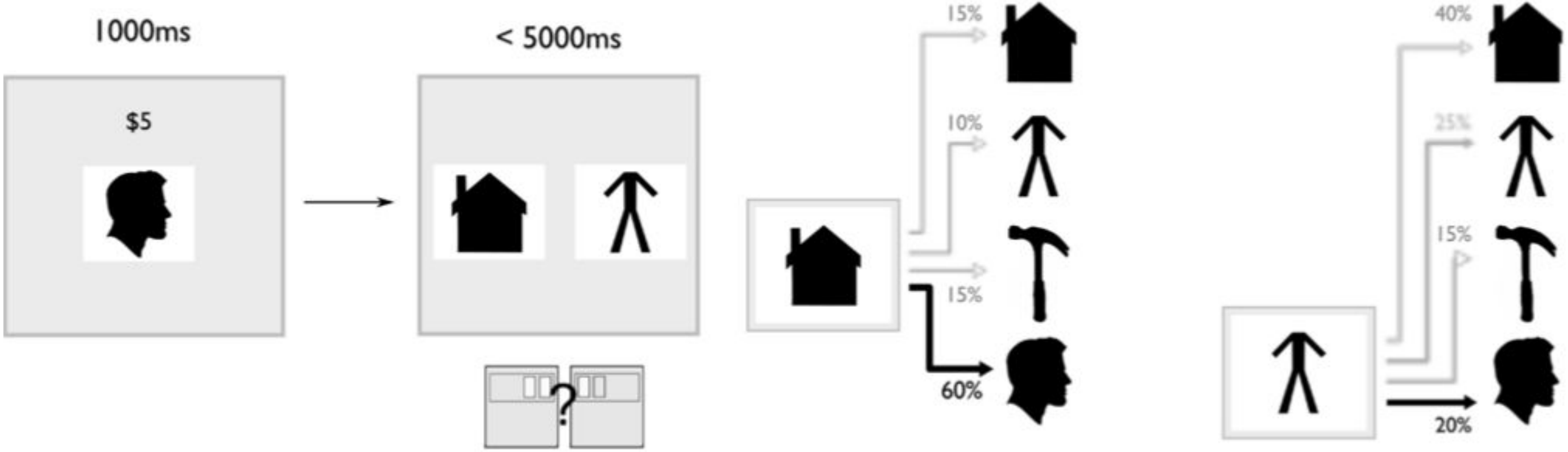
(bornstein & daw 2012, 2013)

“latent” learning



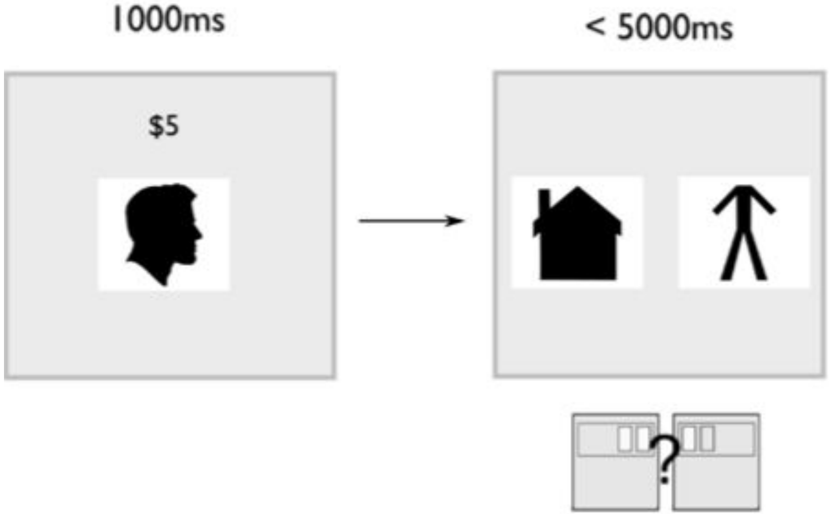
(bornstein & daw 2012, 2013)

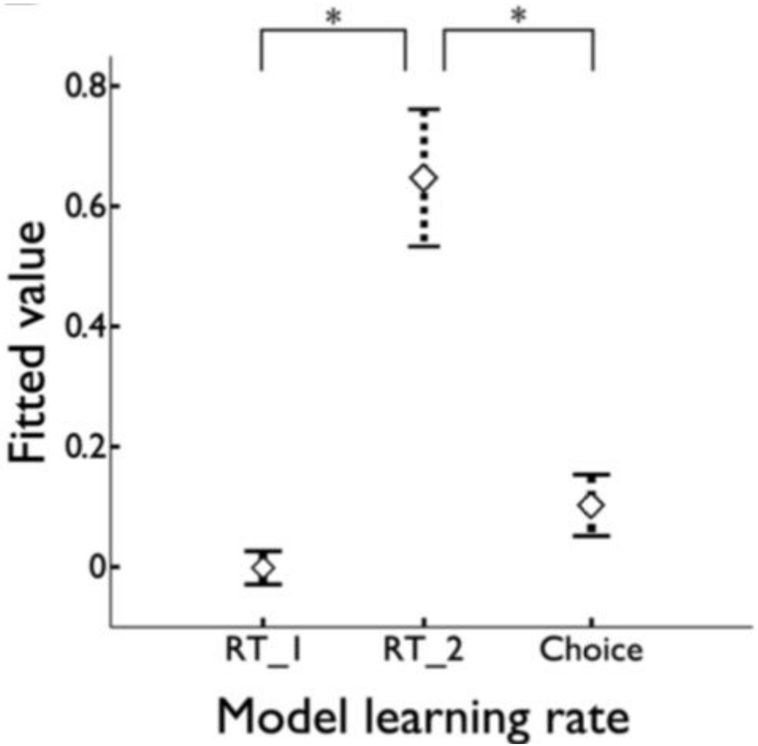
“latent” learning



(bornstein & daw 2012, 2013)

“latent” learning





(bornstein & daw 2012, 2013)

interim summary

- “model-based” behavior can be distinguished by:
  a. **outcome-sensitivity**: quick response to outcome devaluation

  b. **offline updating**: value function updates that reflect knowledge of transition structure

  c. **online evaluation**: use of transition function to make online decisions with novel rewards

- model-based and model-free behavior can “trade off” based on computational demands of the current task
    - model-free: simple structure, lots of experience
    - model-based: complex structure, little experience

outline

I. motivations

II. behavioral signatures

III. neural substrates

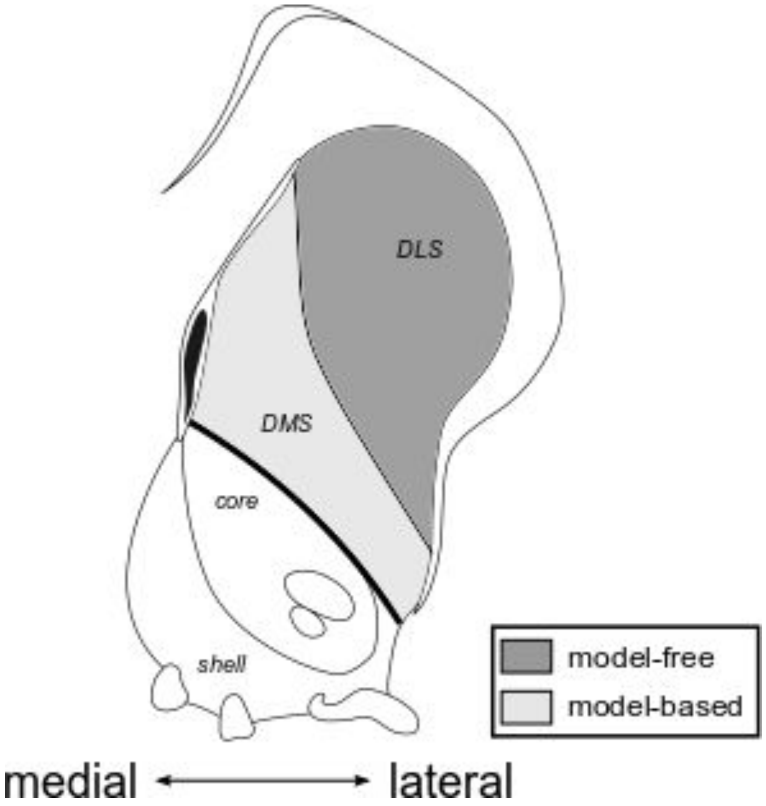
IV. if time: open questions

neural substrates: striatal subdivisions

- muscimol inactivations to dorsolateral striatum impair overtraining (yin et al 2004)

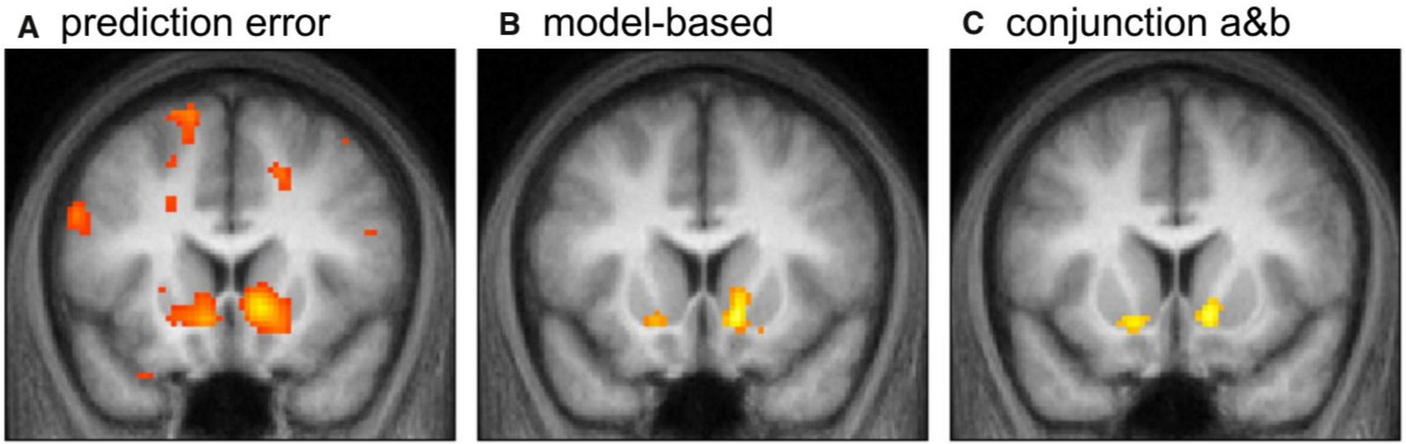
- inactivations to dorsomedial striatum enhance devaluation-insensitivity (yin et al 2005)

- interpretation:
    - neural ensembles in DLS reflect “stimulus-response” (S-R) associations
    - in DMS, “action-outcome” (A-O) associations



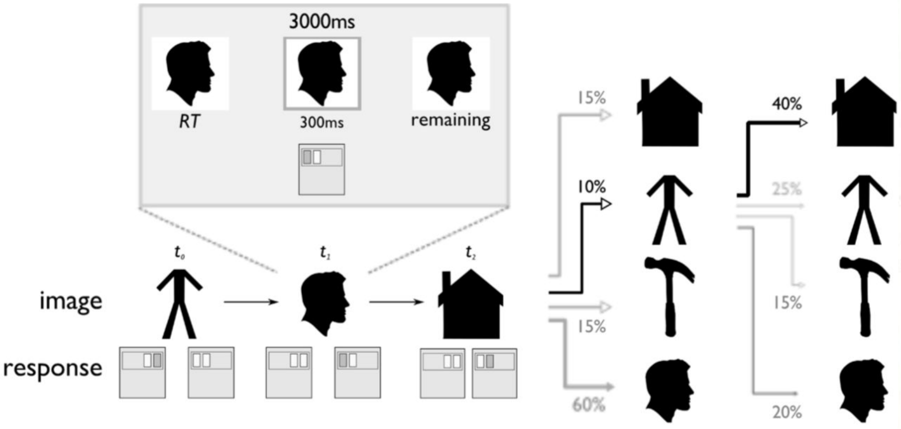
neural substrates: RPE

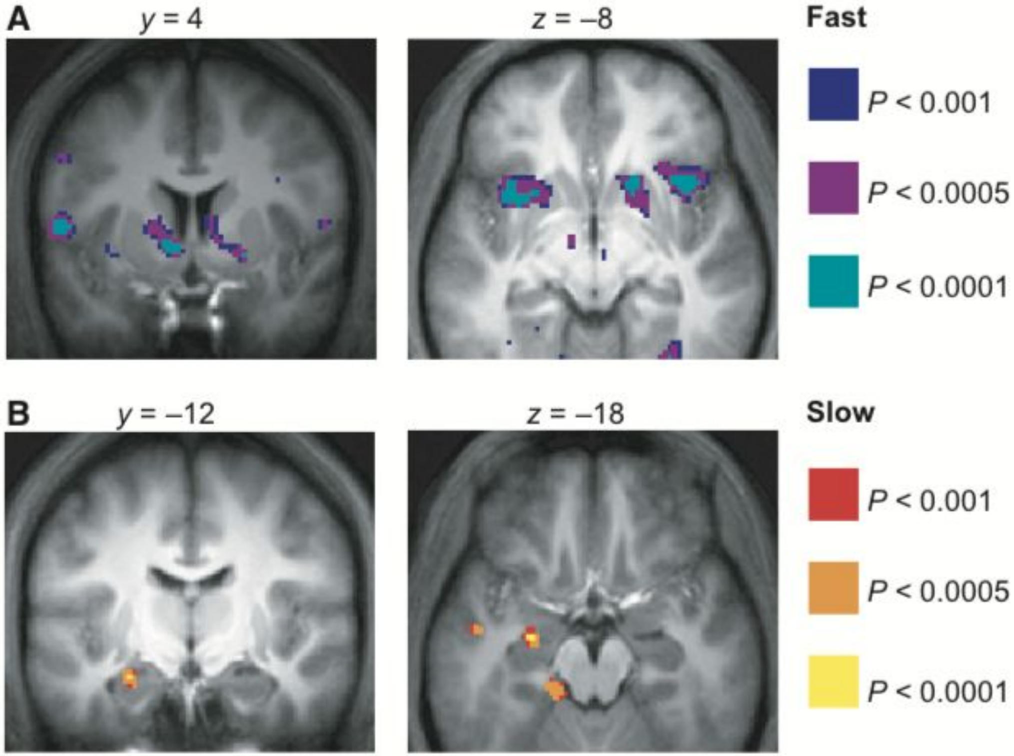
- ventral striatum is a primary target of the midbrain dopaminergic nuclei
- BOLD signal in vStr tracks RPE (mcclure et al 2004; daw, o'doherty et al 2006)
- in *repeated* choice tasks, RPE reflects a mixture of model-based and model-free influence (daw et al 2011; simon & daw 2011)
- in *planning-based* tasks, RPE reflects solely model-based influence (bornstein & daw 2013)



(daw et al 2011)

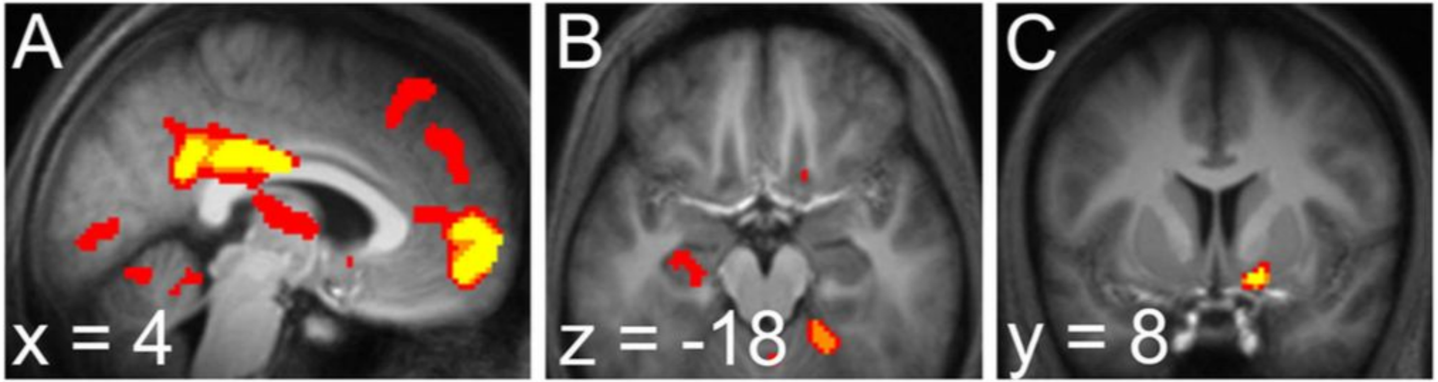
multiple model-based

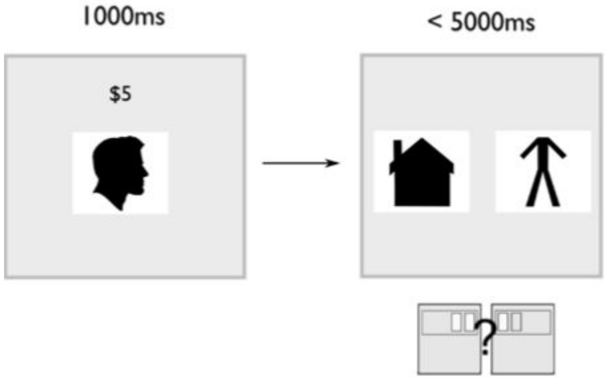




(bornstein & daw 2012, 2013)

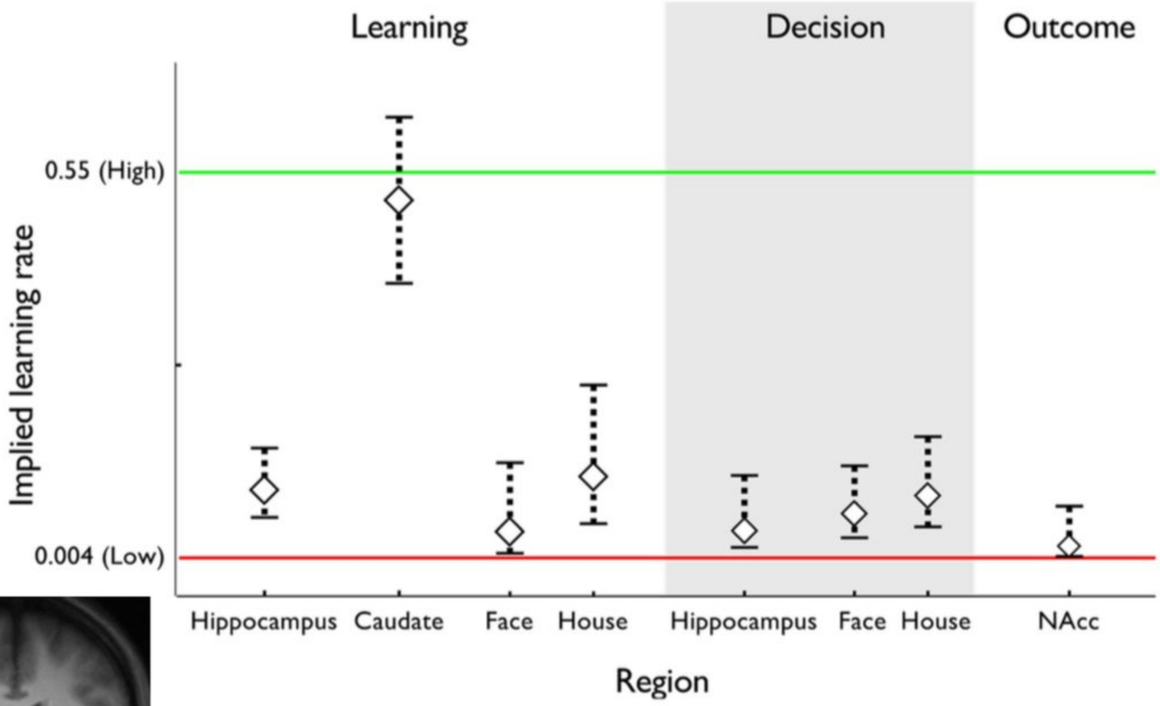
only hippocampus predicts planning for reward

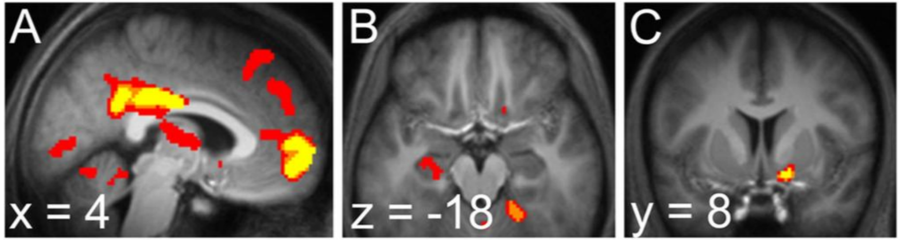




(bornstein & daw 2012, 2013)

only hippocampus predicts planning for reward





(bornstein & daw 2012, 2013)

where ~~is~~ are the model**s**?

- cerebellum (doya et al 2002)
- lateral PFC/prelimbic (PL) cortex:
  - inactivations impair A-O learning (balleine et al 1998)
  - “state prediction errors” (glascher et al 2010)
  - muscimol inactivation impairs transitive reward inference (pan et al 2018)
- dorsomedial striatum/SMA:
  - inactivation impairs sensitivity to outcome-devaluation (yin et al 2005)
  - “ramping” predicts decisions (ding, gold 2010)
  - fast-timescale S-S transition learning (bornstein & daw 2012, 2013)
- MTL/hippocampus:
  - (right, but not left) MTL lesion patients are “model-free” in 2-step task (vikbladh et al 2018a)
  - slow-timescale S-S transition learning (bornstein & daw 2012, 2013)
  - “cognitive map” / replay (foster & wilson 2006; johnson & redish 2007; pfeiffer & foster 2013)

summary

- “model-based” **planning**…
  - allows fast adaptation to changes in both rewards and contingencies
  - relies on a **value function**, just like “model-free” methods
  - but augments this with a **model** that can be used to update the value function via simulated experience

- multiple **representations** can be used to make decisions
  - these reflect various physical (motor, sensory) and latent (cognitive) structure(s)
  - the influence of each representation may depend on the uncertainty in that representation

- model use can be “online” or “offline”
  - these can be mutually beneficial

outline

I. motivations

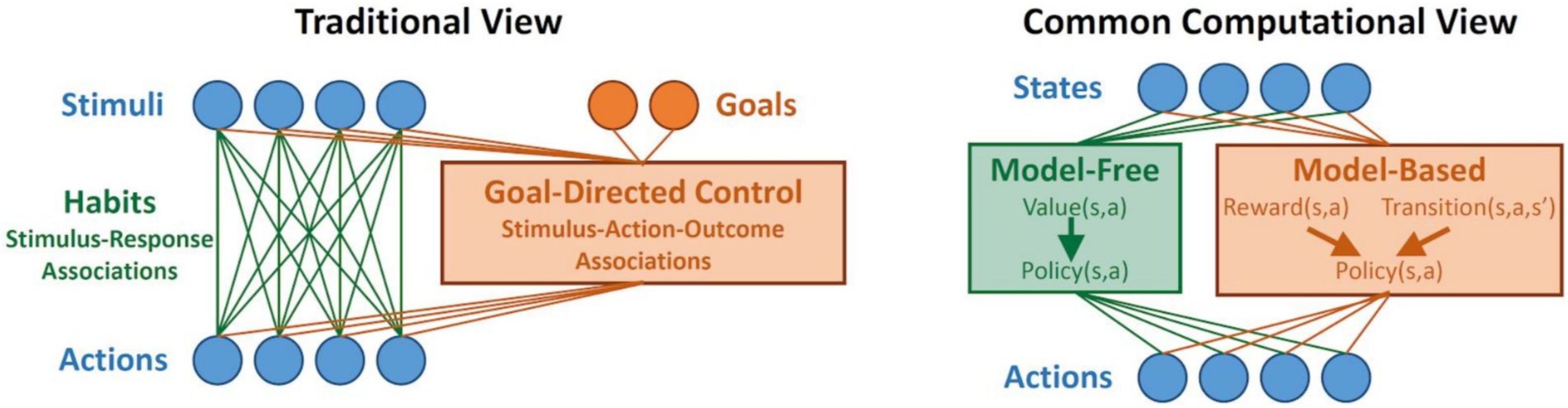
II. behavioral signatures

III. neural substrates

IV. open questions

open q: is anything truly "model-free"?

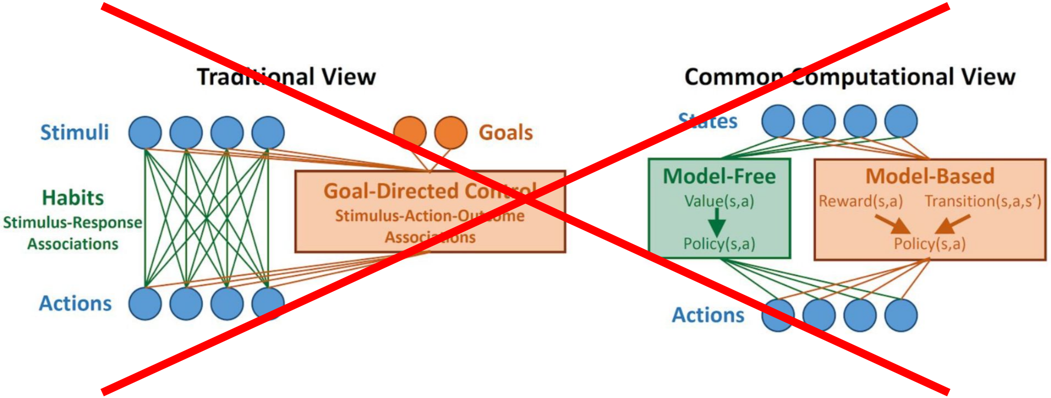
- pretty much every healthy behavior/neural signal reflects model use (doll et al 2012)
- model-based/model-free ≠ goal-directed/habit?

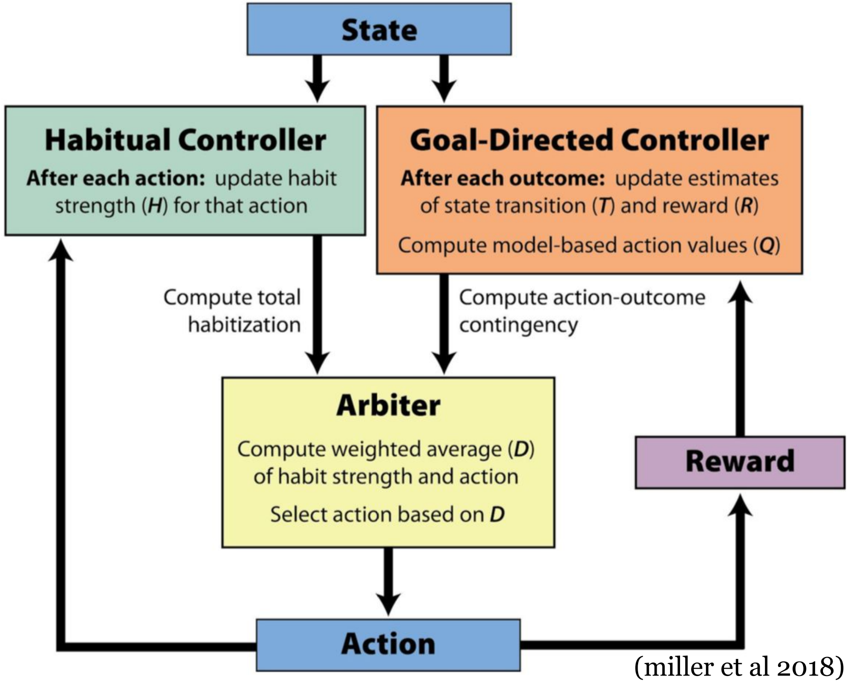


(doll et al 2012; miller et al 2018)

open q: is anything truly “model-free”?

- pretty much every behavior/neural signature model-based (doll et al 2012)
- model-based/model-free ≠ goal-directed/habit?
    - habits may be “value-free” (miller et al 2018)





(miller et al 2018)

open q: underlying representations

- computational RL: many varieties of “model”
  - *sample* models v *distribution* models

open question: sample updates or expected updates?
  - distribution models can be used to generate samples, or to compute entire expectations
  - this can be difficult to distinguish experimentally, at the level of aggregate behavior
  - can even be difficult to distinguish at the level of neural activity! (beck et al 2008; berkes et al 2010)

- neuroscience: a continuum of representations
  - full state-space (daw et al 2005; glascher et al 2010; smittenar et al 2013; wilson et al 2016)
  - flexible action sequences (doya et al 2002; bornstein & daw 2012)
  - flexible stim-stim sequences (dayan 1993; bornstein & daw 2012, 2013)
  - episodes (lengyel & dayan 2008; bornstein & daw 2013; bornstein et al 2017a,b; vikbladh et al 2018a,b; ritter et al 2018)
- further frontiers
  - not just states or plans (e.g. categories — http://www.j-paine.org/dobbs/why_be_interested_in_categories.html)
  - general principles apply across representations: learning incrementally, by experience, direct or simulated

open q: trajectory sampling?

- no one has yet decoded *multi*-step decisions, either offline or online

- thus it’s an open question whether planning is trajectory sampling, or single-step value-function updates

tomorrow

- state inference
- decisions by sampling (from memory)
- the episodic memory route to model-based planning

further reading

- all cited papers are at: http://aaron.bornstein.org/ccnss/
  - plus some others i think are worth reading

- 2nd edition of sutton & barto book (latest update 2018.**07.03**): http://incompleteideas.net/book/the-book-2nd.html

- forthcoming book: “goal-directed decision making: computations and neural circuits” — ask for pdfs in a couple months
  - table of contents: http://aaron.bornstein.org/cv/pubs/2018_gdcnc/

- happy to talk about research any time ⟹ aaron@bornstein.org