Reinforcement Learning

Woo-Young (Young) Ahn
Department of Psychology
Seoul National University
ccs-lab.github.io

Reinforcement Learning (RL)

- *What is RL?*

- *RL models (algorithms for prediction and control)*
  - *Classical conditioning*
    - *Rescorla-Wagner (R-W) model*
    - *(Bayesian or non-Bayesian) extension of R-W models*
  - *Instrumental conditioning*
    - *Model-free vs Model-based learning*
  - *Pavlovian control vs Instrumental control*
  - *Action selection models*
    - *Softmax + others*
    - *Exploration/exploitation*

- *Limitations & Future directions*

Learning objectives

Participants will…

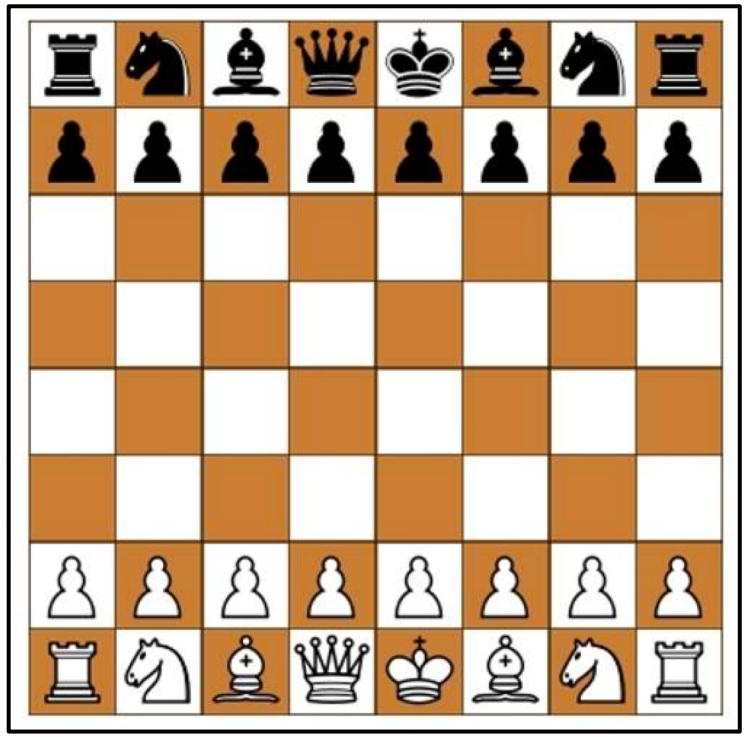
- *Understand the key concepts and notations of RL*
- *Know (some of) popular RL models (& references)*
  - *Simple to complex models*
- *Understand their limitations*

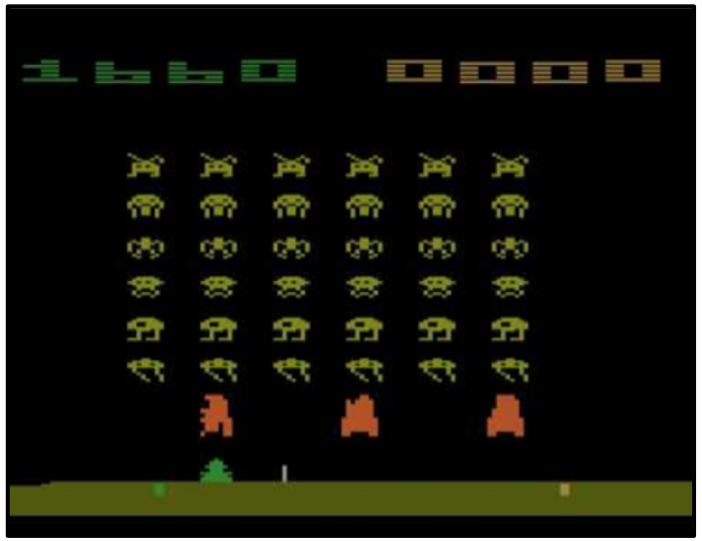
What is RL?

"Learning what to do" …
based on rewards and punishments

Sutton & Barto (1998) Reinforcement Learning
Dayan & Labott (2000) Theoretical neuroscience

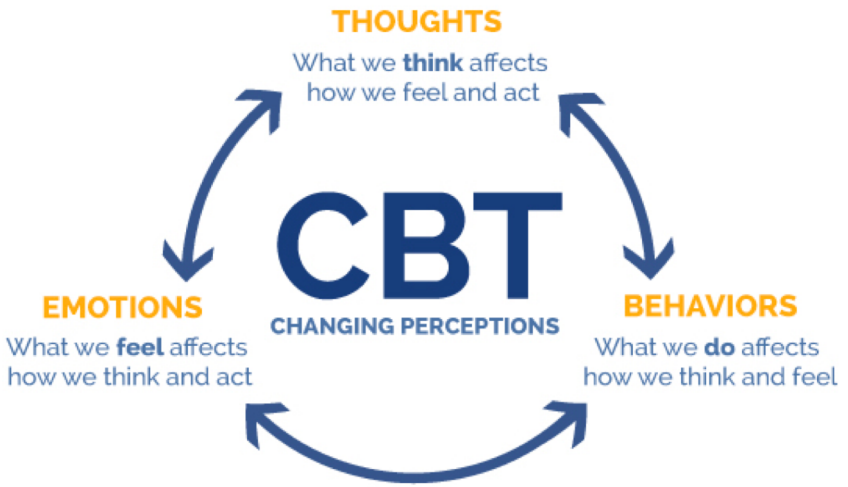
"Learn optimal ways to make decisions"
in an uncertain environment





Mnih et al (2015) Nature

Silver et al (2016) Nature



RL is a type of Machine Learning

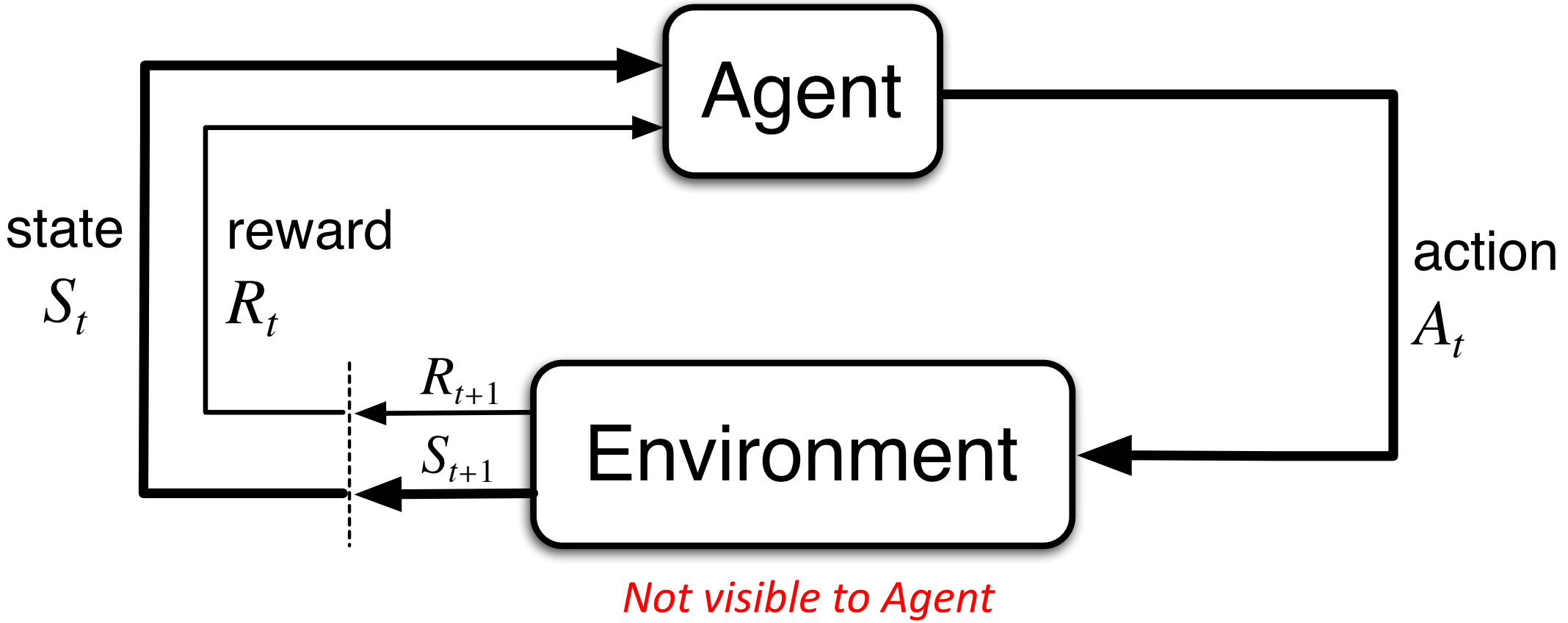
- *Supervised Learning*
- *Unsupervised Learning*
- *Reinforcement Learning*

Q) How is RL different from other ML paradigms?

- *No external supervisor (“minimally supervised”)*
- *Reward signals (learn from trials and errors)*
- *Interaction with environment*
- *Closely tied to action selection (e.g., exploration/exploitation)*

Agent-Environment Interface

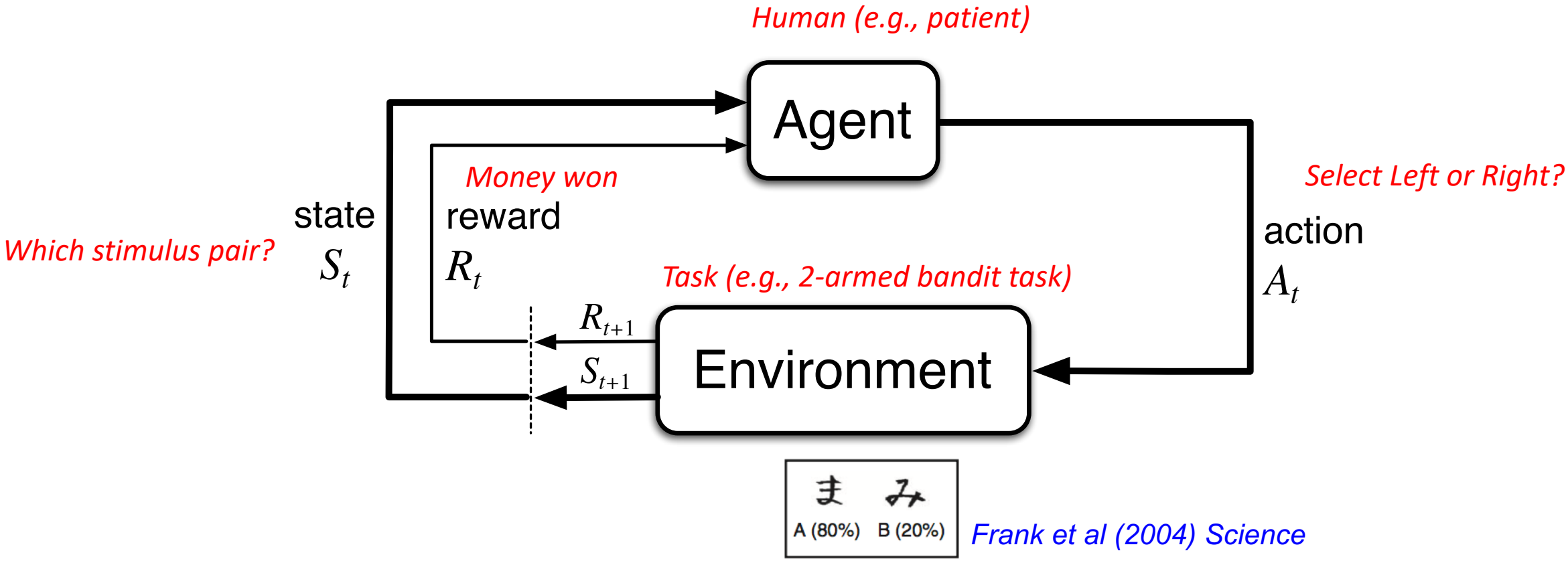
e.g., Maze task, Tree search, N-armed Bandit



Sutton & Barto (1998) Reinforcement Learning

Typically in Computational Psychiatric research settings..

Model parameters → Psychologically meaningful processes/constructs



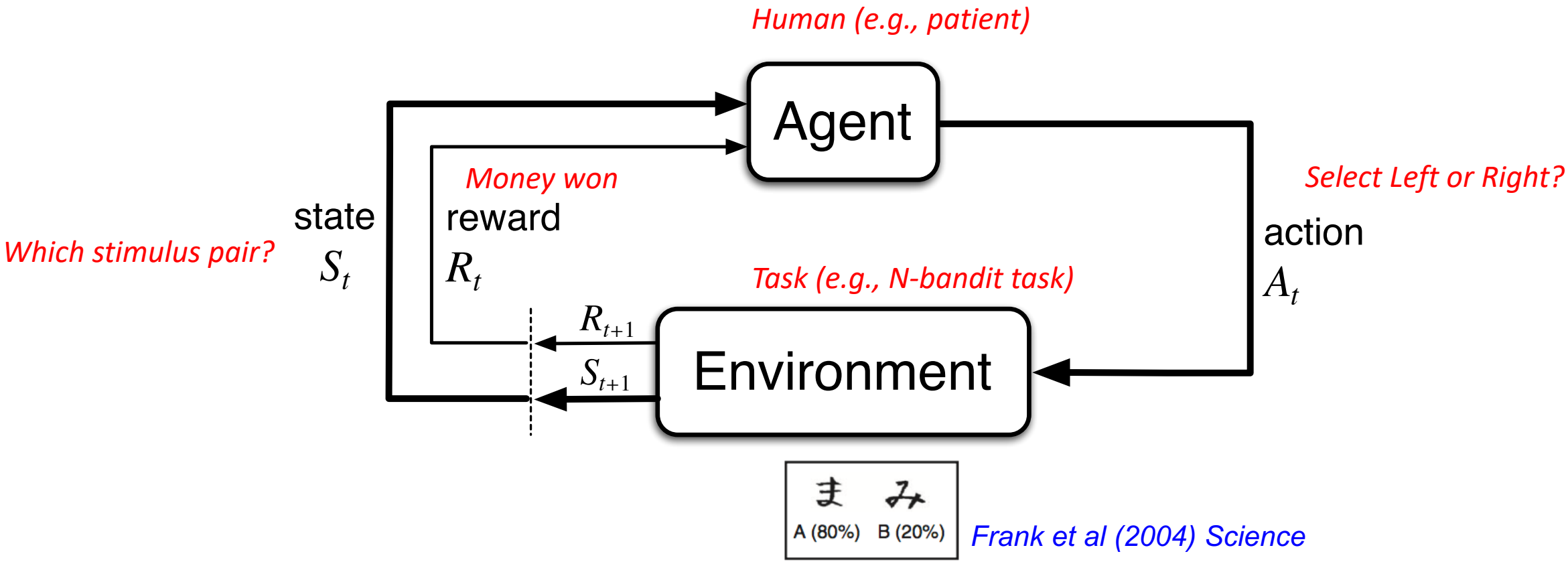
Frank et al (2004) Science

S_t: State value on time (trial) t

A_t: Action value on time (trial) t

R_t: Reward on time (trial) t

$\pi_t(a_t, s_t)$: Policy on time (trial) t → mapping from states to actions



RL models (algorithms for prediction)

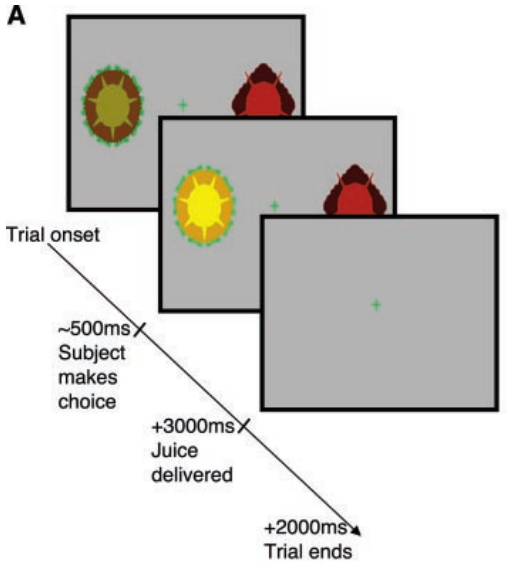
Two experimental set-ups
(Not a distinction of learning mechanisms)

Classical conditioning
(No action required)

CS+ CS- CS Neut

e.g., O'Doherty et al (2003) Neuron

Instrumental conditioning
(Action required)



e.g., O'Doherty et al (2004) Science

Classical conditioning

CS+ CS- CS Neut

e.g., O'Doherty et al (2003) Neuron

Rescorla-Wagner (R-W) model

$$V_t = V_{t-1} + \alpha(R_t - V_{t-1})$$

Learning rate

Stimulus value (t) *Stimulus value (t-1)* *Outcome* *Stimulus value (t-1)*

Prediction error

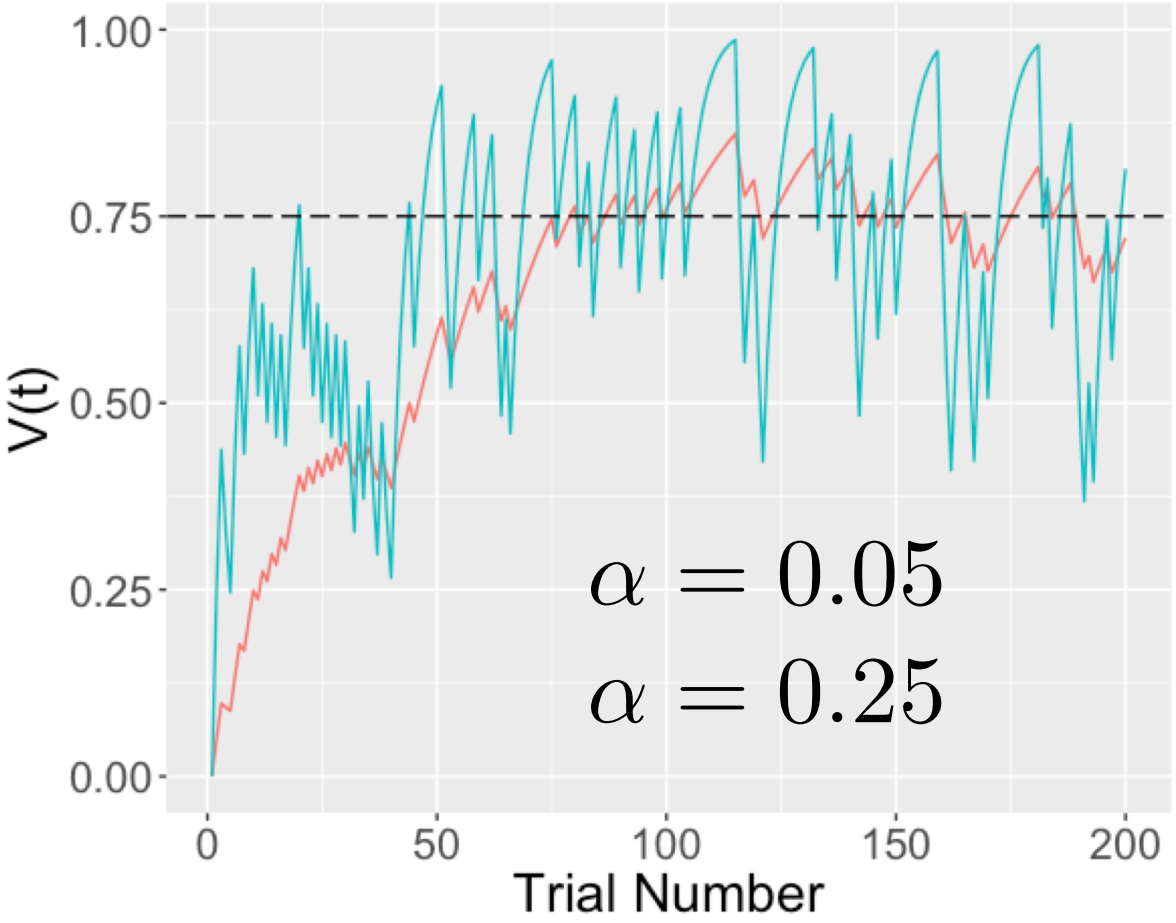
→ Point estimates of V_t

Classical conditioning

CS+ CS- CS Neut

e.g., O'Doherty et al (2003) Neuron

e.g., Reward rate = 0.75



* *Rescorla-Wagner (R-W) model*
→ *Point estimates of* V_t

* *Bayesian generalization of R-W*
→ *Kalman filter*

Dayan et al (2000); Kakade & Dayan (2002)
Daw et al (2006); Kruschke (2008)

http://haines-lab.com/post/2017-04-04-choice_rl_1/

Classical conditioning

CS+ CS- CS Neut

e.g., O'Doherty et al (2003) Neuron

Temporal Difference (TD) Learning model

→ *Generalization of R-W (real-time model)*
→ *To account for within-trial and between-trial relationships among stimuli*

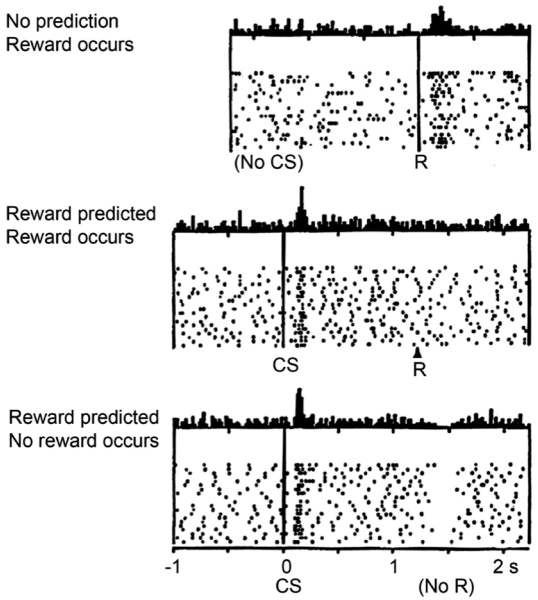
Reward Prediction Error
TD learning model

Computational roles for dopamine in behavioural control

P. Read Montague[1,2], Steven E. Hyman[3] & Jonathan D. Cohen[4,5]

[1]*Department of Neuroscience and* [2]*Menninger Department of Psychiatry and Behavioral Sciences, Baylor College of Medicine, 1 Baylor Plaza, Houston, Texas 77030, USA (e-mail: read@bcm.tmc.edu)*
[3]*Harvard University, Cambridge, Massachusetts 02138, USA (e-mail: seh@harvard.edu)*
[4]*Department of Psychiatry, University of Pittsburgh and* [5]*Department of Psychology, Center for the Study of Brain, Mind & Behavior, Green Hall, Princeton University, Princeton, New Jersey 08544, USA (e-mail: jdc@princeton.edu)*

Montague et al (2004) Nature



Temporal difference (TD) learning model

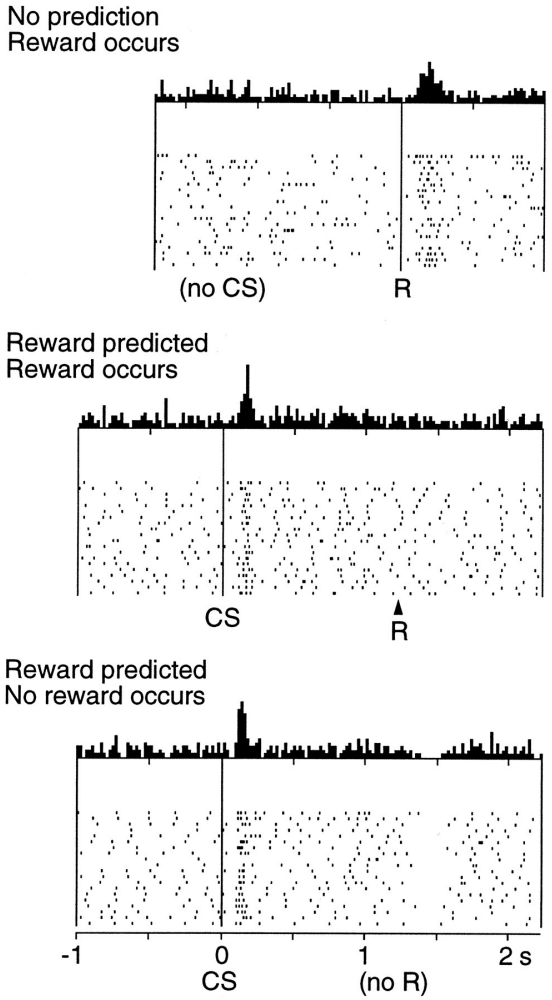
$$\delta(t) = \text{prediction error}\ (t) = E[r_t] + \gamma \cdot \hat{V}(s_{t+1}) - \hat{V}(s_t)$$
$$\approx \text{current reward} + \gamma \cdot \text{next prediction} - \text{current prediction}$$

Sutton & Barto (1998) Reinforcement Learning

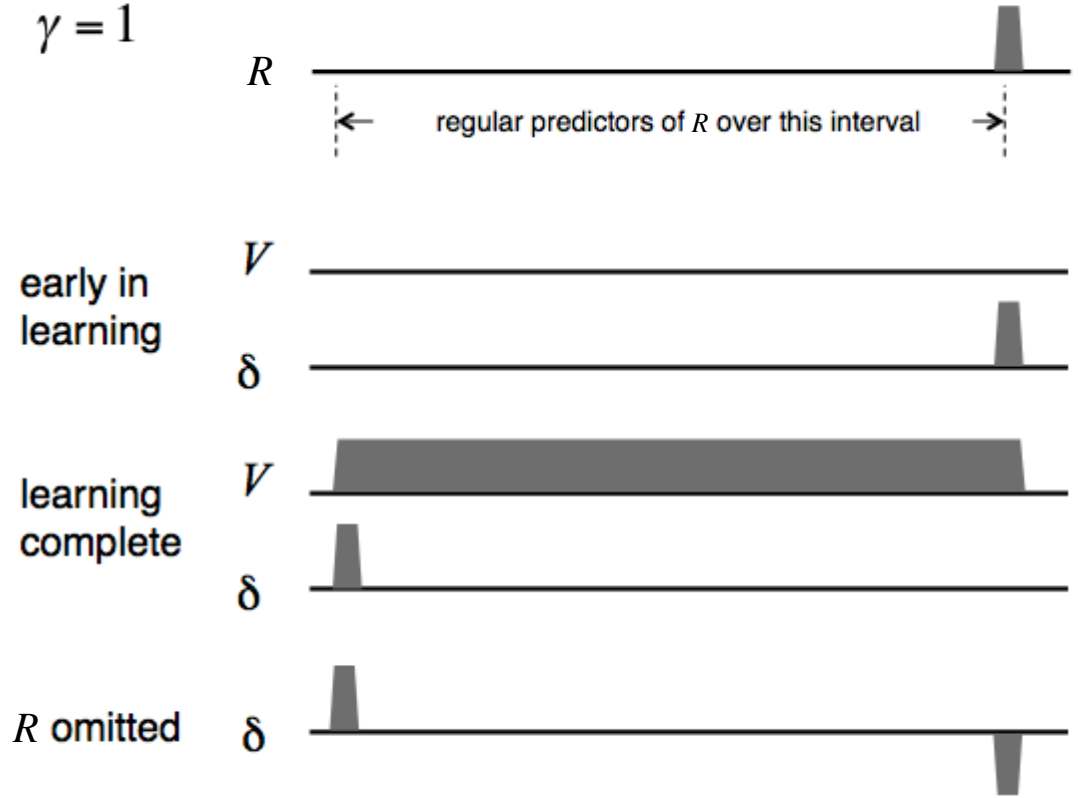
Q) How TD learning accounts for the phasic response of a dopamine neuron?

Sutton & Barto (2017) Reinforcement Learning, 2nd Ed., Chapter 15

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



$\gamma = 1$



Reward onset

$$\delta_t = R_t + V_t - V_{t-1} = R_t + 0 - 0 = R_t$$

Cue onset

$$\delta_t = R_t + V_t - V_{t-1} = 0 + R_t - 0 = R_t$$

Reward onset

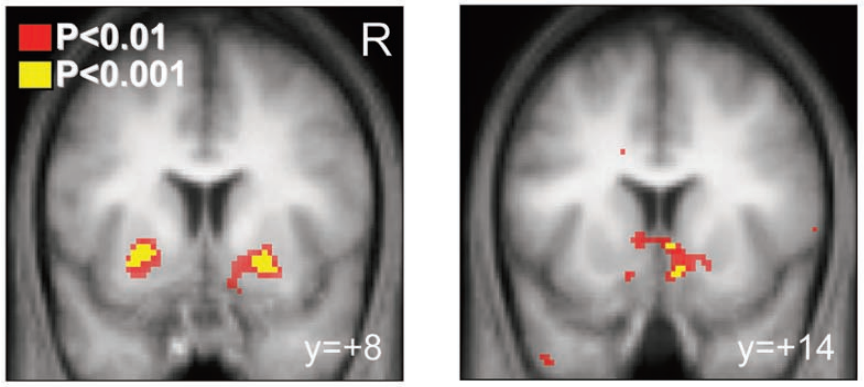
$$\delta_t = R_t + V_t - V_{t-1} = 0 + 0 - R_t = -R_t$$

Dissociable Roles of Ventral and Dorsal Striatum in Instrumental Conditioning

John O'Doherty,[1]* Peter Dayan,[2] Johannes Schultz,[1]
Ralf Deichmann,[1] Karl Friston,[1] Raymond J. Dolan[1]

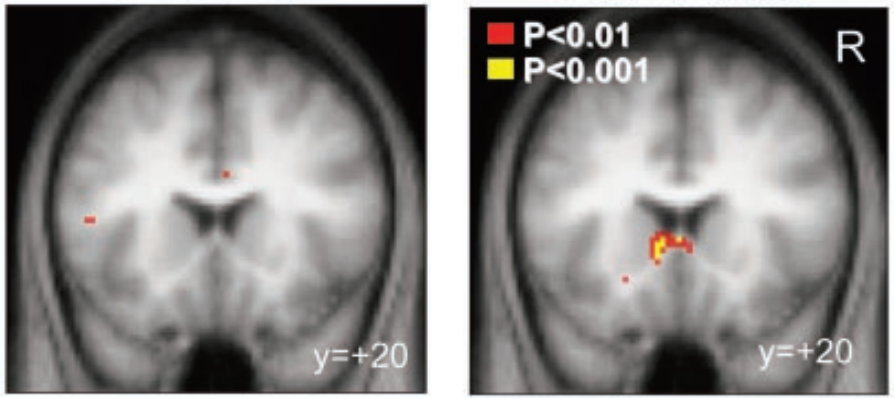
O'Doherty et al (2004) Science

Pavlovian Cond. *Instrumental Cond.*



ventral Str → both Pav. and Instr.

Pavlovian Cond. *Instrumental Cond.*



dorsal Str → only instrumental

Instrumental learning

Model-based vs Model-free

Model-based vs Model-free

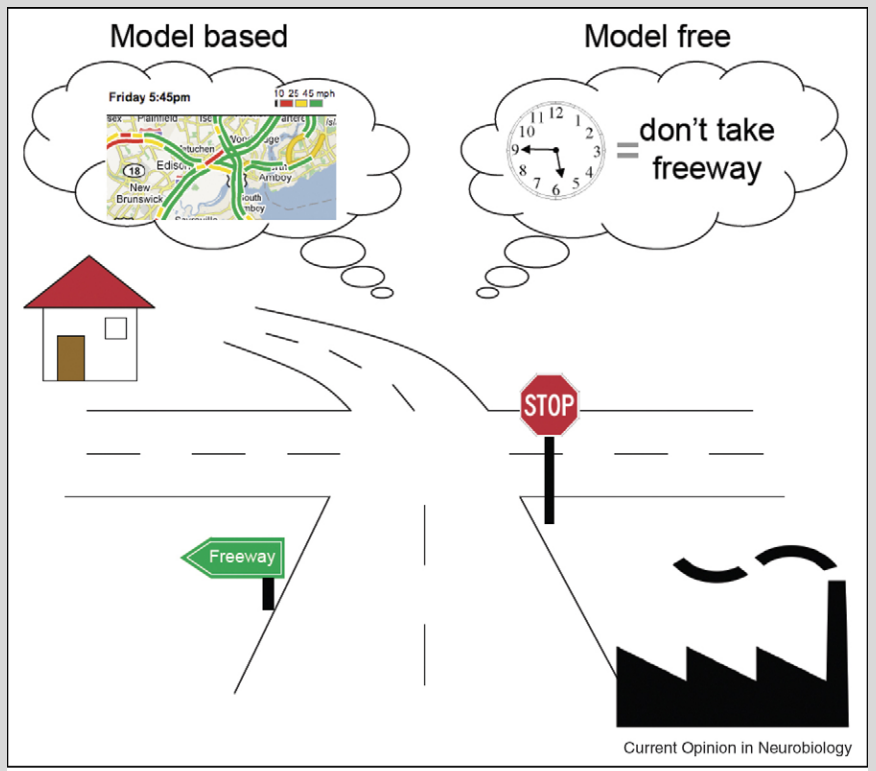


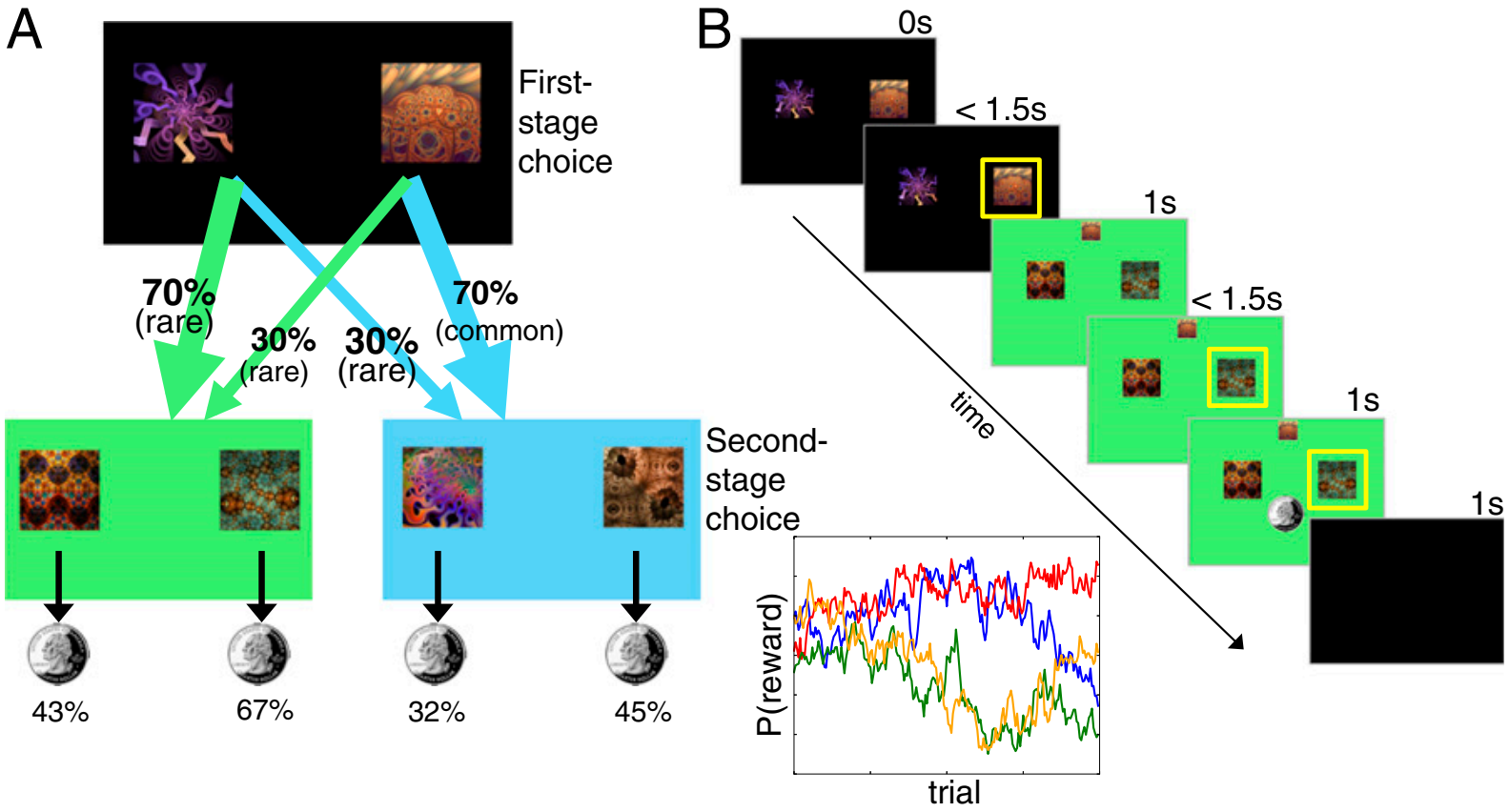
Figure 1: Two ways to choose which route to take when traveling home from work on friday evening.

Dayan & Niv (2008)

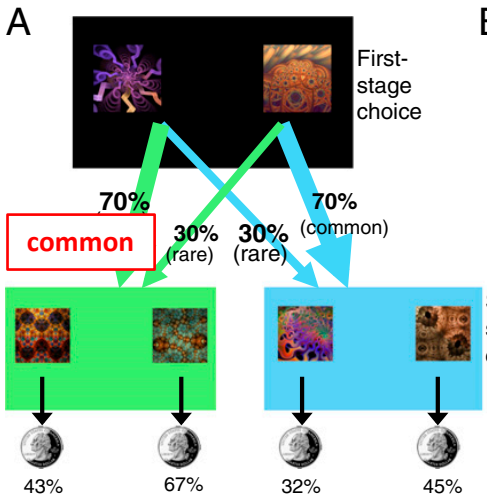
- *Model-based (goal-directed) learning: build a model of an environment. Effortful but flexible.*
- *Model-free (habitual) learning: relies on trials-and-errors. Efficient but inflexible.*
- *(Clinical) examples: compulsive behaviors, etc.*

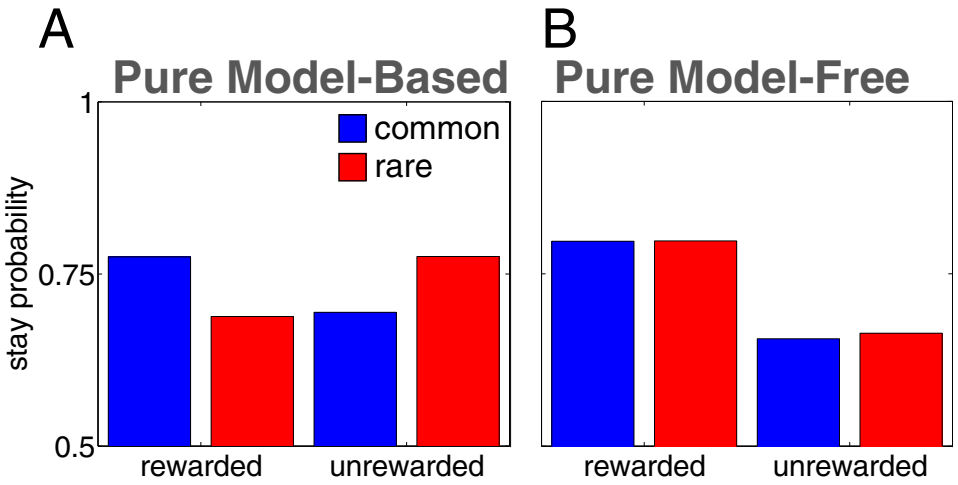
Two-Step task

Daw et al (2011) Neuron



Competing predictions





Scenario 1 (model-based individual)

- Step1: choose Left
- Common transition (70%) to green
- Step2: choose Left and won!

Next trial

- To choose the same 2nd level stimulus,
- In Step 1, I will choose Left (=stay)

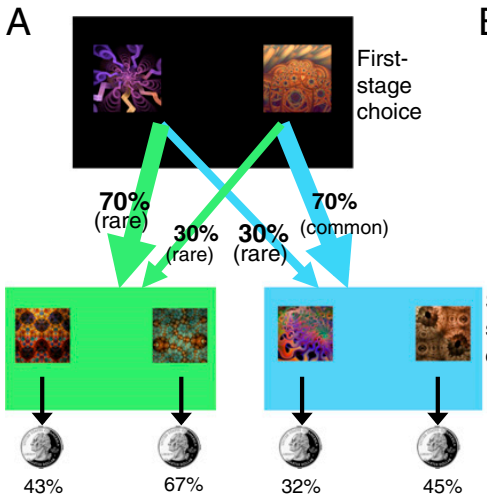
Scenario 2 (model-based individual)

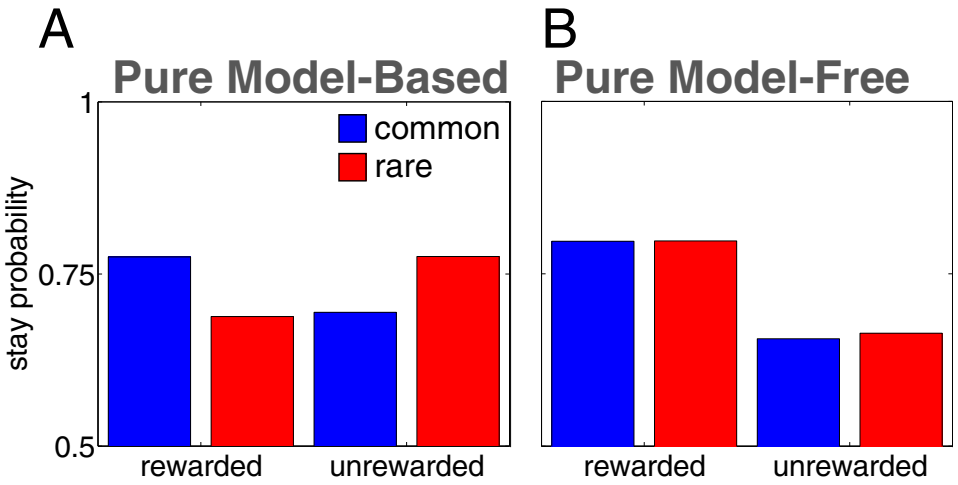
- Step1: choose Left
- Rare transition (30%) to blue
- Step2: choose Left and won!

Next trial

- To choose the same 2nd level stimulus,
- In Step 1, I will choose Right (=switch)

More scenarios





Scenario 1 (model-free individual)

- Step1: choose Left
- Common transition (70%) to green
- Step2: choose Left and won!

Next trial

- To choose the same 2nd level stimulus,
- In Step 1, I will choose Left (=stay)

Scenario 2 (model-free individual)

- Step1: choose Left
- Rare transition (30%) to blue
- Step2: choose Left and won!

Next trial

- To choose the same 2nd level stimulus,
- In Step 1, I will choose **Left (=stay)**

Computational model

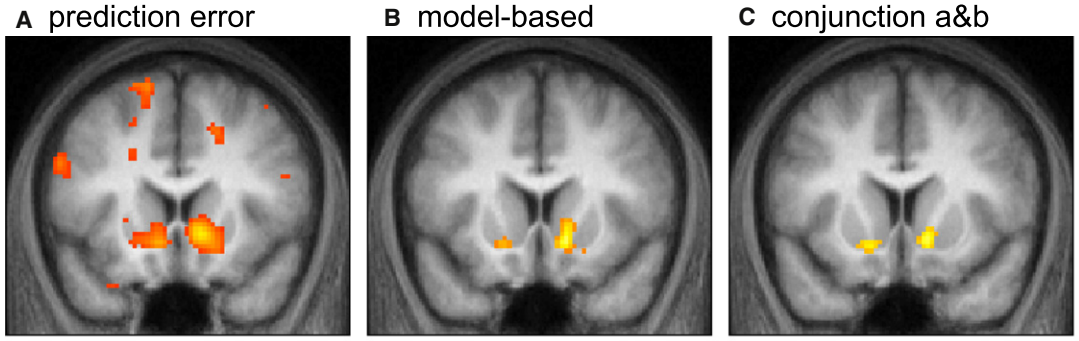
Daw et al (2011) Neuron
Wunderich et al (2012) Neuron

- *Separately calculate V^{MF} and V^{MB} (assuming full knowledge of the environment).*

- *Omega (ω): weight for model-based (MB)*

  – *0 (completely model-free) ≤ ω_{MB} ≤ 1 (completely model-based)*

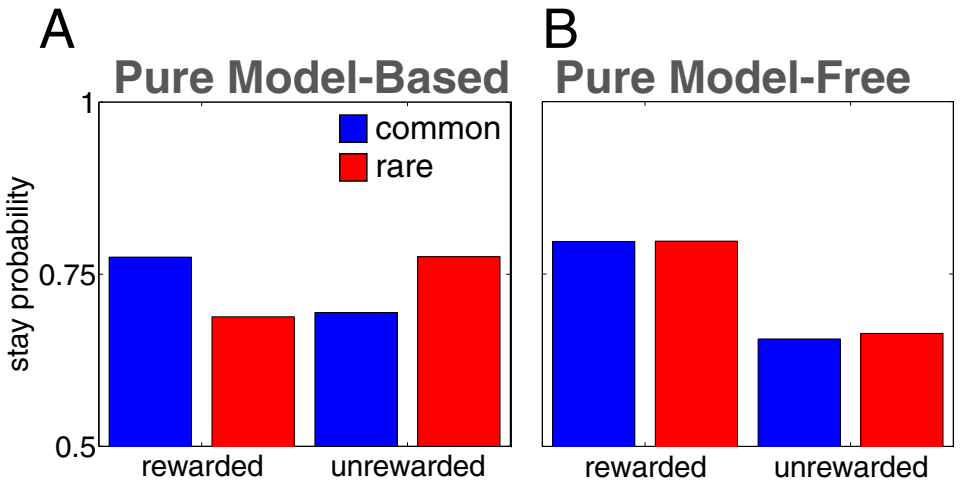
$$V^{Hybrid} = \omega \cdot V^{MB} + (1 - \omega) \cdot V^{MF}$$



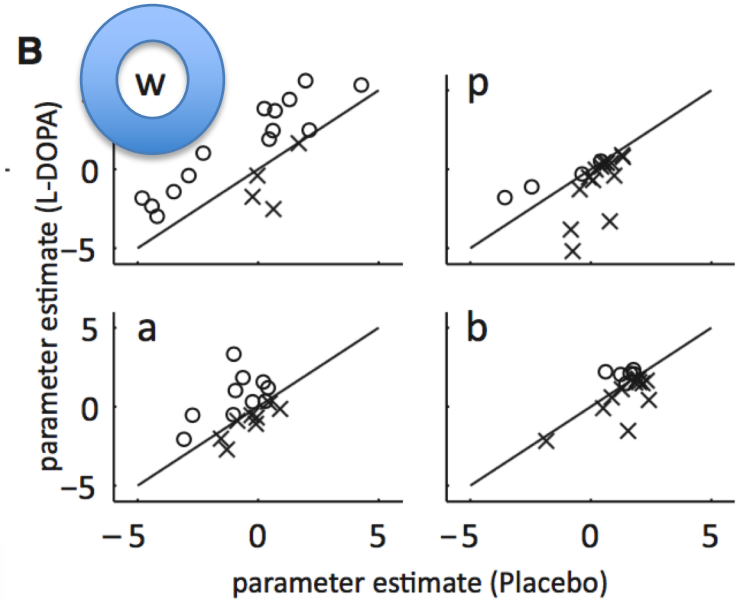
Daw et al (2011) Neuron

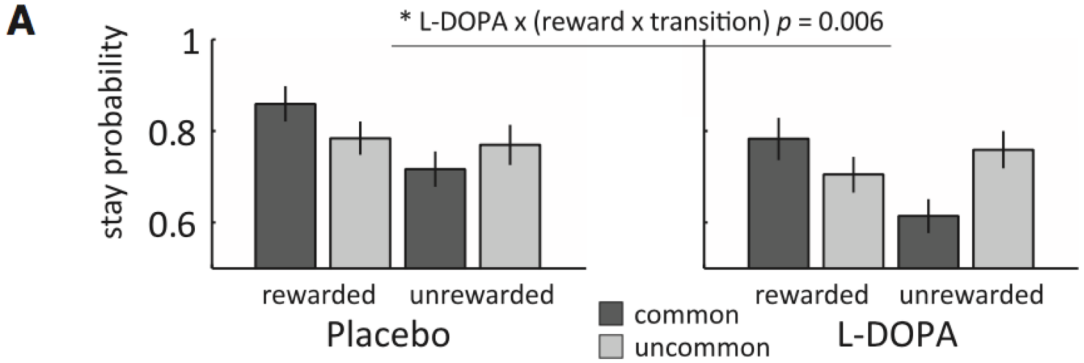
DA enhances model-based behavior

Wunderich et al (2012) Neuron



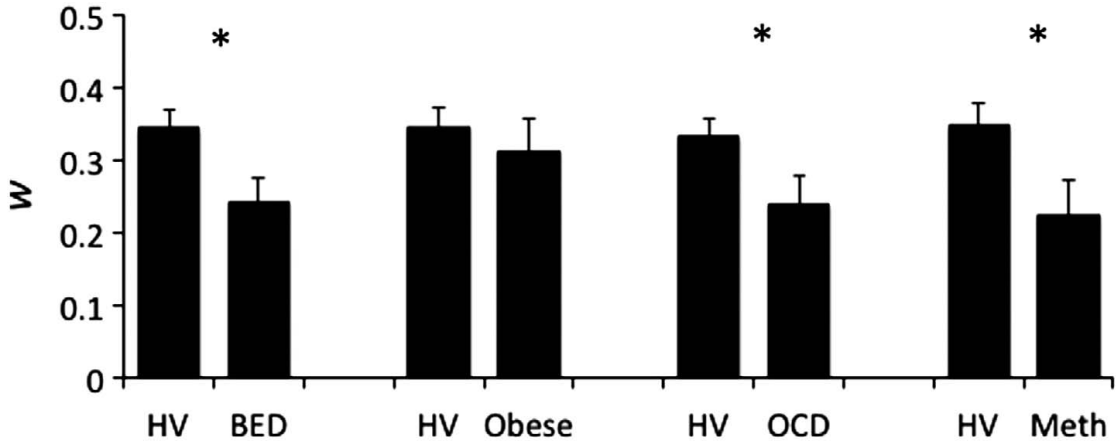
Measure of model-based over model-free



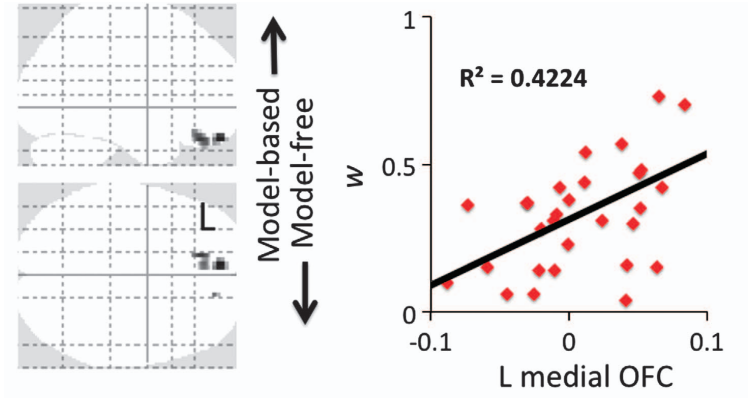


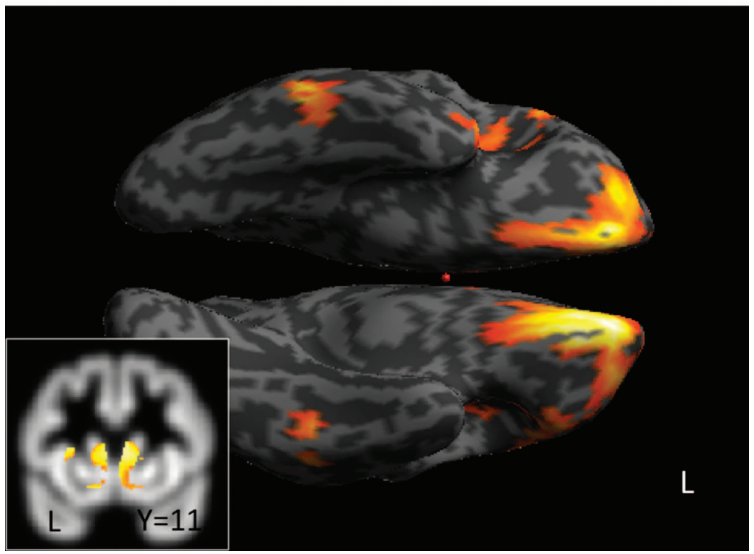
Reliance of model-based control → Disorders of compulsivity

Voon et al (2014) Molecular Psych



$$V^{Hybrid} = \omega \cdot V^{MB} + (1 - \omega) \cdot V^{MF}$$

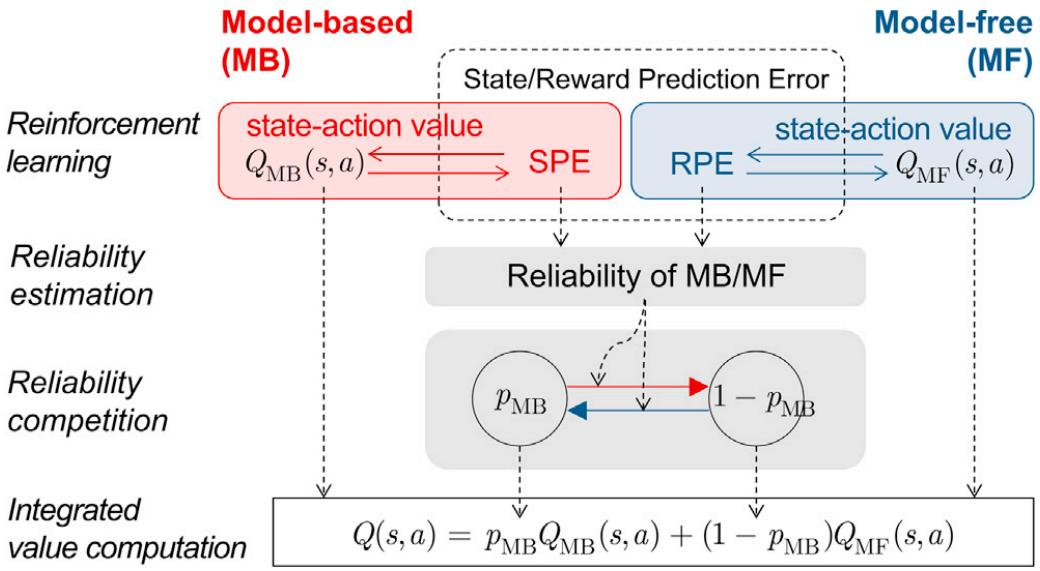


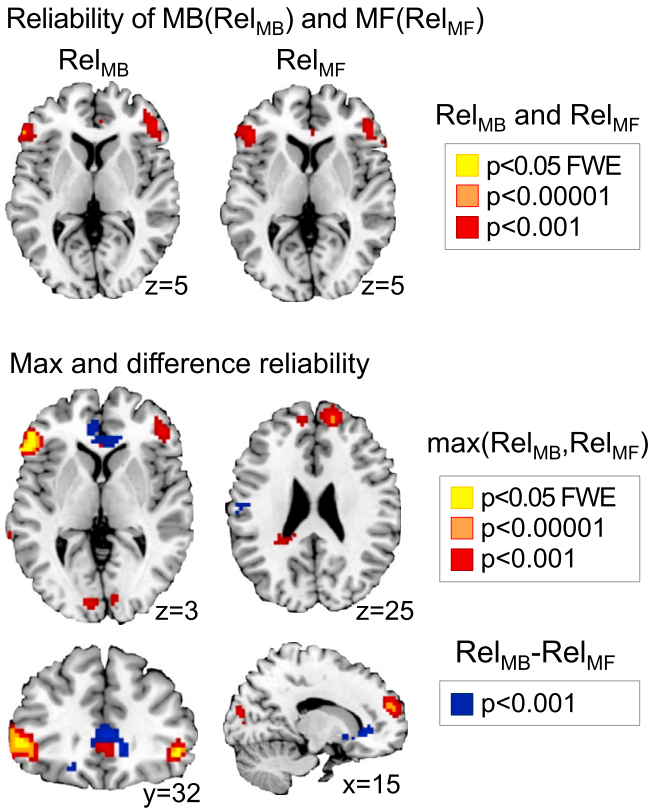


Gray-matter volume

Correlated with model-based weight among healthy controls

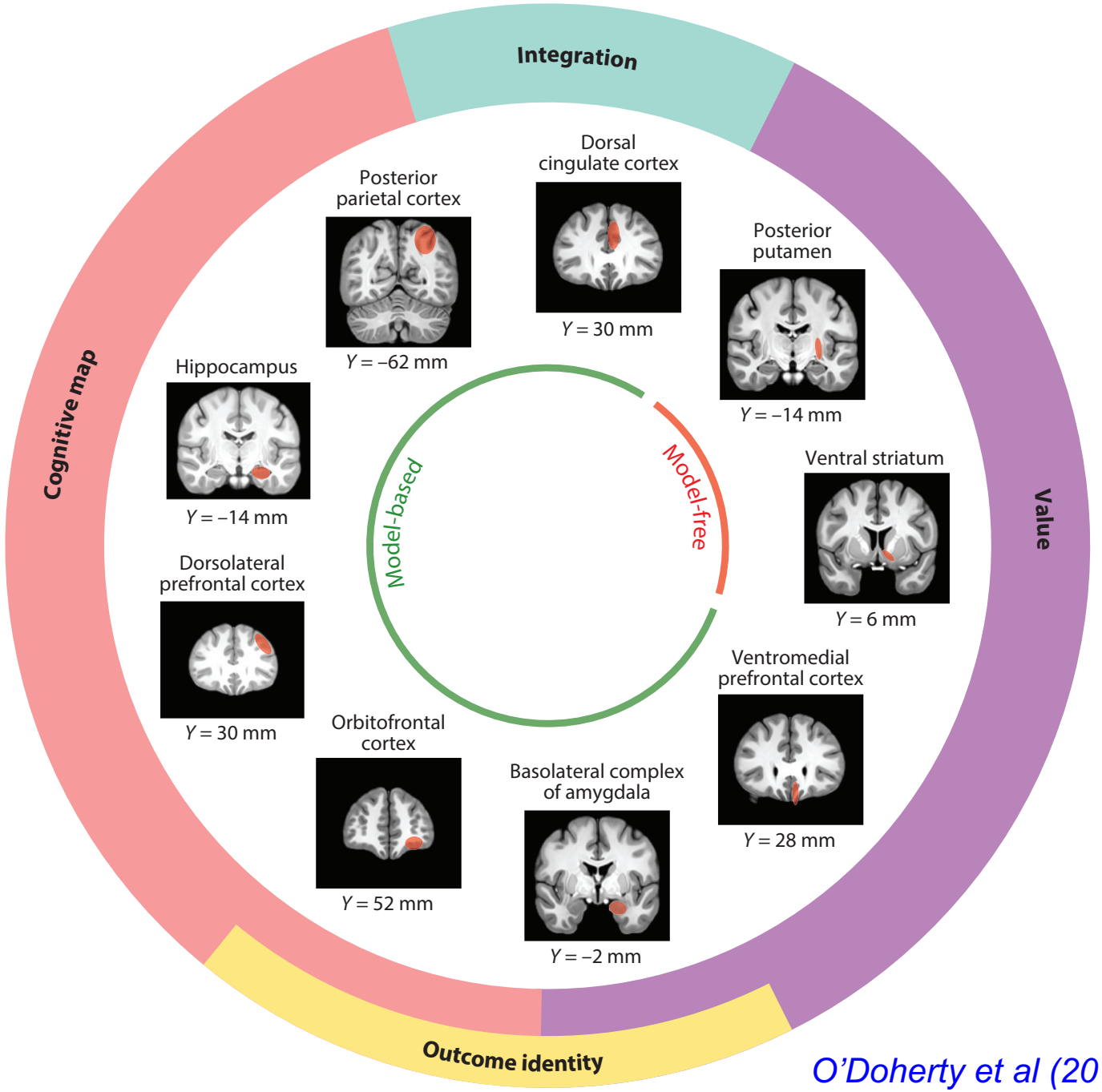
Reliability-based arbitration between model-based and model-free





Lee et al (2014) Neuron
Daw et al (2005) Nature Neuroscience
Wang et al (2018) Brain & Neuro. Advances

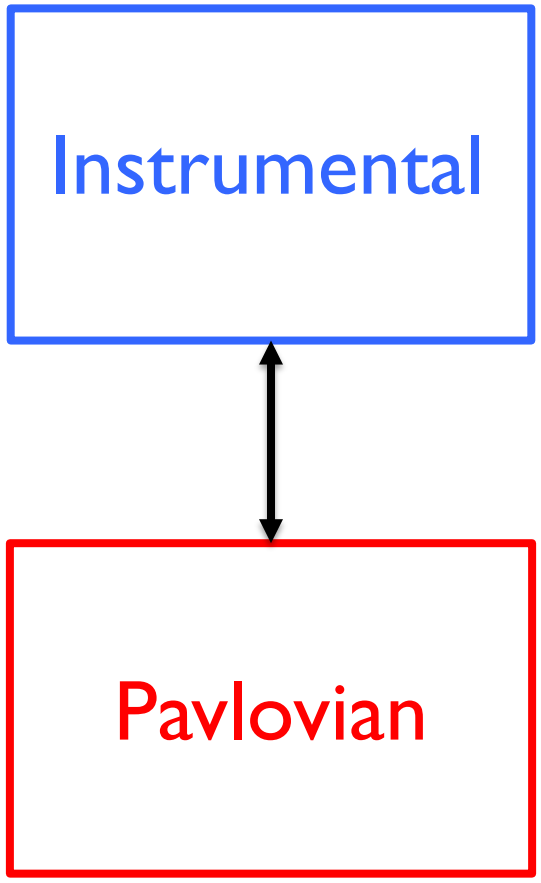
Inferior lateral prefrontal and frontopolar cortex



O'Doherty et al (2017) Annu. Rev. Psychology

Pavlovian vs Instrumental control

Pavlovian vs Instrumental control



"Learn what's best"

Approach for (+)
Withdraw from (-)

"Pavlovian bias"

Opinion

CellPress

Action versus valence in decision making

Marc Guitart-Masip[1,2], Emrah Duzel[3,4,5], Ray Dolan[2], and Peter Dayan[6]

[1] Aging Research Centre, Karolinska Institute, SE-11330 Stockholm, Sweden
[2] Wellcome Trust Centre for Neuroimaging, Institute of Neurology, University College London, London WC1N 3BG, UK
[3] Institute of Cognitive Neuroscience, University College London, London WC1N 3AR, UK
[4] Otto von Guericke University Magdeburg, Institute of Cognitive Neurology and Dementia Research, D-39120 Magdeburg, Germany
[5] German Center for Neurodegenerative Diseases, D-39120 Magdeburg, Germany
[6] Gatsby Computational Neuroscience Unit, University College London, London W1CN 3AR, UK

Balleine & O'Doherty (2010); Dayan et al (2006); Dayan (2013); Dayan & Niv (2008); Dolan & Dayan (2013); Dayan & Berridge (2014); Rangel et al (2008)

Hungry Chicken

Food!

Hershberger (1986)

Hungry Chicken

Food!

Hershberger (1986)

Pavlovian-Instrumental competition

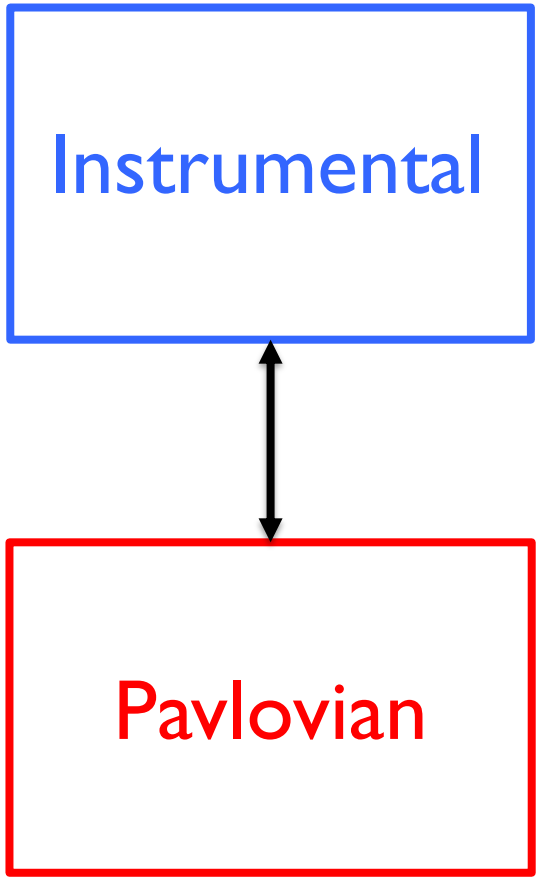
Hershberger (1986)

Impulse control

www.ignitermedia.com

Orthogonalized Go/Nogo task
Pavlovian-Instrumental competition

Guitart-Masip et al (2012) Neuroimage
Also, see Huys et al (2011) Plos Comp Biology



"Learn what's best"

Approach for (+)
Withdraw from (-)

"Pavlovian bias"

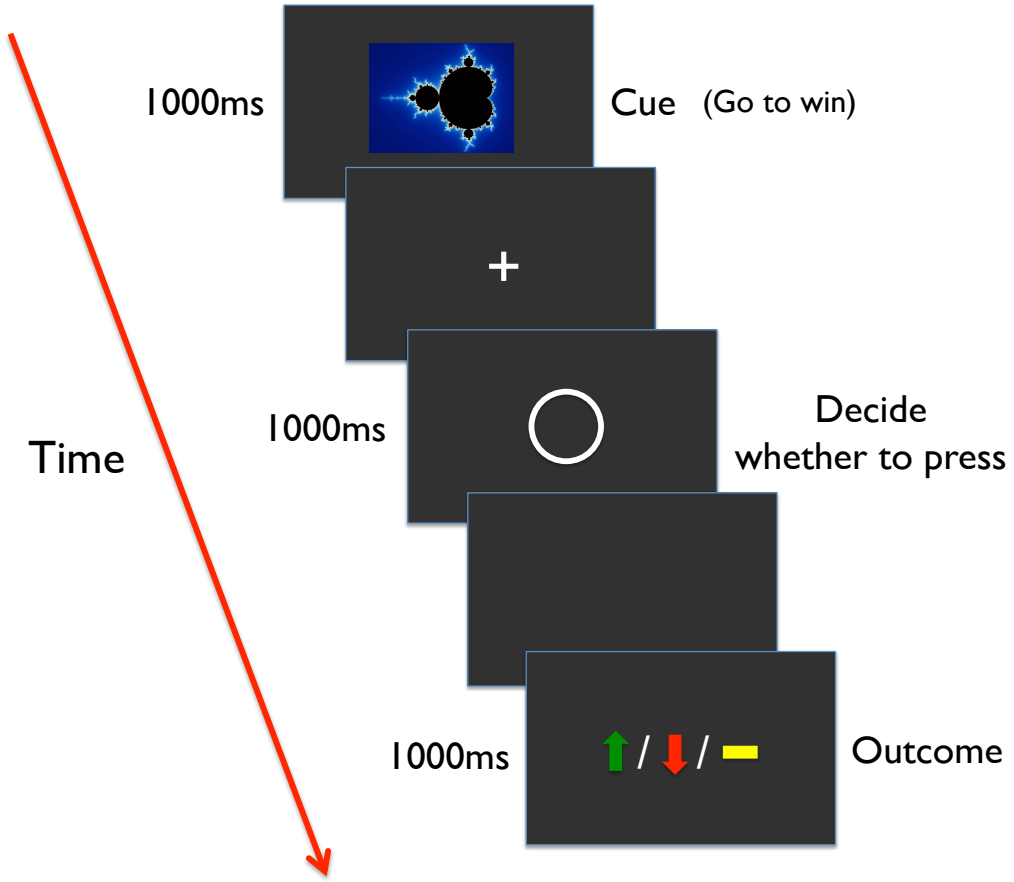
Balleine & O'Doherty (2010); Dayan et al (2006); Dayan (2013); Dayan & Niv (2008); Dolan & Dayan (2013); Dayan & Berridge (2014); Rangel et al (2008)

Orthogonalized Go/Nogo task

	Loss	Gain
Go	Go to avoid	Go to win
Nogo	Nogo to avoid	Nogo to win

Pavlovian congruent

Pavlovian incongruent



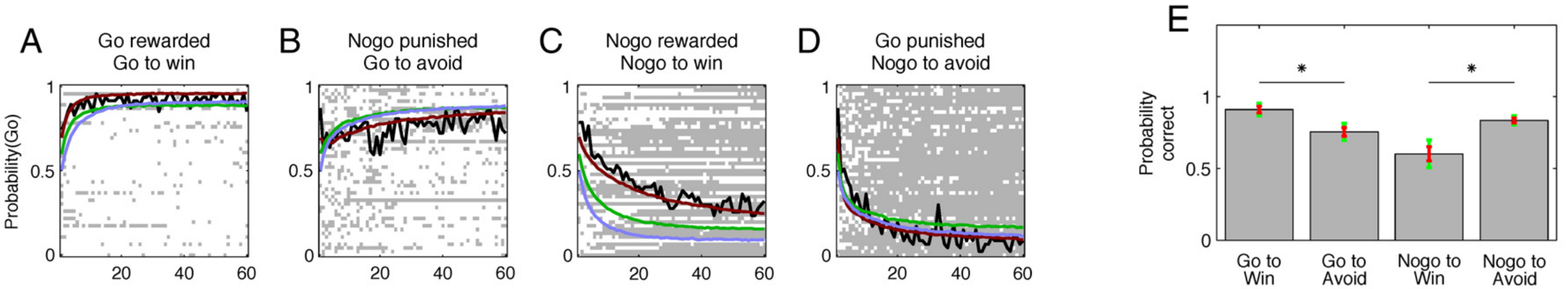
- *4 cues (conditions)*

  2 actions (Go / Nogo) x
  2 valence (Gain / Loss)

Guitart-Masip et al (2012) Neuroimage
Cavanagh et al (2013) J Neuro

Orthogonalized Go/Nogo task

Guitart-Masip et al (2012) Neuroimage



Reward sensitivity

$$Q_t(a_t, s_t) = Q_{t-1}(a_t, s_t) + \epsilon \cdot (\rho r_t - Q_{t-1}(a_t, s_t))$$

Q value

Modified R-W rule

Go bias

$$W_t(Go_t, s_t) = Q_t(a_t, s_t) + b + \pi V_t(s_t)$$

Pavlovian bias

Arbitration between Pavlovian and instrumental systems

- *Much less is known*
- *Appraisal implemented via PFC, down-regulation of the amygdala and vStr?* *(O'Doherty et al., 2017, Annu Rev Psychol)*
- *The nature of computations mediating this arbitration?*

Action selection models (algorithms for control)

Q) How to select actions or stimuli?

30 25

- *Greedy choice*
- *ε-Greedy (random choice with ε)*
- *Softmax*
  - *Variants (lapse rate, perseverance…)* Guitart-Masip et al (2012); Moutoussis et al (2011)

$$Pr_j[C_{t+1}] = \frac{e^{\theta \cdot \pi_j(t)}}{\sum_{k=1}^{N} e^{\theta \cdot \pi_k(t)}}$$

θ *: Exploration/exploitation parameter*

(a.k.a. inverse temperature)

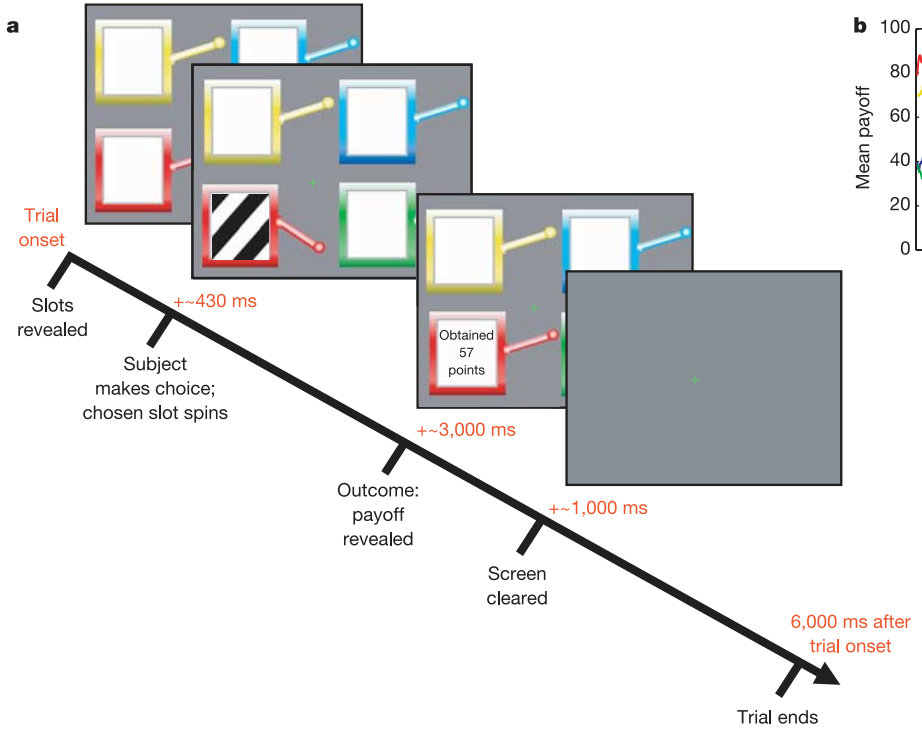
- ***Sequential sampling models***

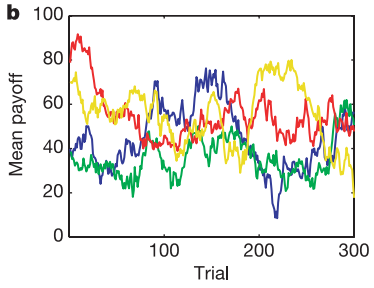
Q) How to select actions or stimuli?

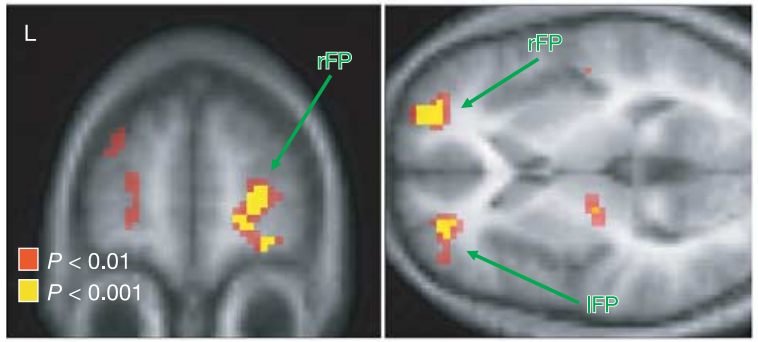
Cortical substrates for exploratory decisions in humans

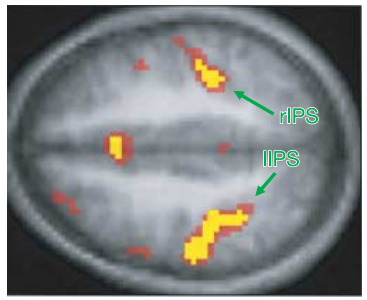
e.g., Daw et al (2006) Nature
Badre et al (2012) Neuron;
Jepma & Nieuwenhuis (2011)

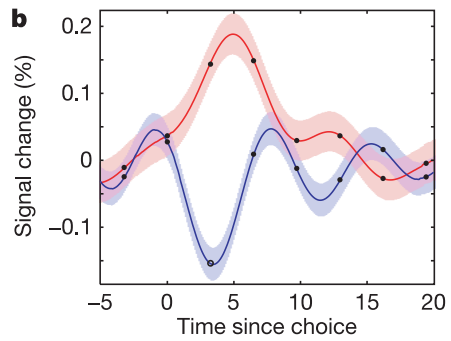
Nathaniel D. Daw[1]*, John P. O'Doherty[2]*†, Peter Dayan[1], Ben Seymour[2] & Raymond J. Dolan[2]











Some tips?

- *Other DM variables (effort, uncertainty)*

- *Stimulus value vs. Action value*

- *Only 1 action for each stimulus*
  → Stimulus value = action value

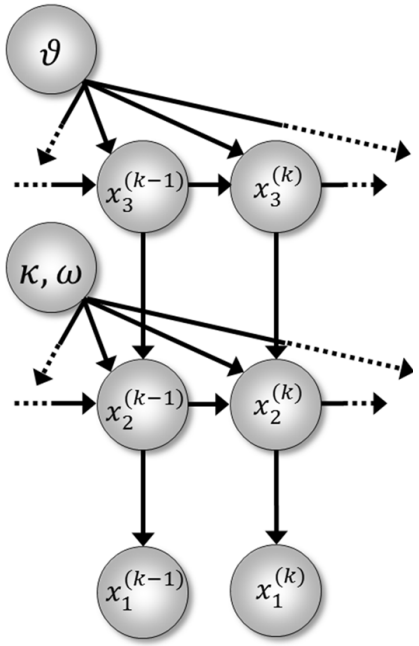
- No within-trial events → TD model is not necessary

Limitations & Future directions

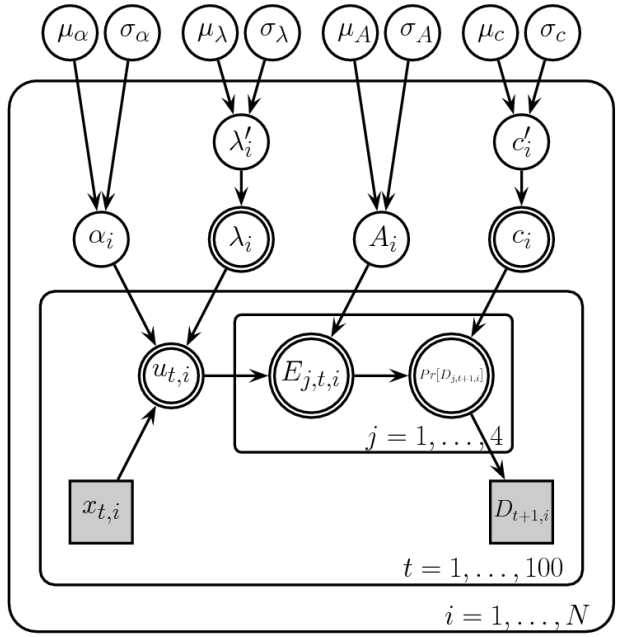
- *Overly simplified "toy" problems*
  - *Violated assumptions (e.g., discrete space/action & Markov..) , (Gershman & Daw, 2016, Annu Rev Psych)*
  - *Predict real-life DM? (Mobb et al., 2018, Nat Rev Neuro)*
- *One-shot learning with sparse data*
  - *Episodic memory (hippocampus) (Gabrieli, 1998; Eichenbaum et al., 1999)*
- *Adaptive Design Optimization (ADO)*
- *Modeling of even toy problems is hard for many people*

I like the idea of modeling

But…



Mathys et al (2011) Frontiers

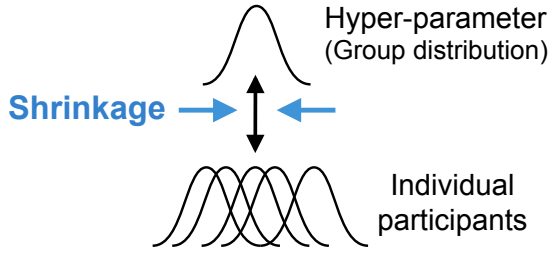


Ahn et al (2011) JNPE

hBayesDM *(hierarchical Bayesian modeling of Decision-Making tasks) Package*

- *Models for "many" tasks/paradigms (next slide)*
- *Single-line of coding in R*
  - *Model fitting, visualization, model comparisons*
- *Based on the advanced Bayesian software, Stan (https://mc-stan.org).*
- *Hierarchical Bayesian modeling*
- *All codes are publicly available*

https://github.com/CCS-Lab/hBayesDM



What tasks and models are available?

in 2017

Ahn & Busemeyer (2016) Curr Opin Behav Sci

- *Choice reaction time → sequential sampling models*
- *Delay Discounting (e.g., Mazur, 1987)*
- *Iowa Gambling (Bechara et al, 1994)*
- *(Orthogonalized) Go/Nogo (Guitart-Masip et al, 2012)*
- *Two-armed Bandit (Experience-based) including Reversal Learning (e.g., Erev et al, 2010)*
- *Four-armed Bandit (Experience-based) (e.g., Seymour et al, 2012)*
- *Two-choice Description-based (e.g., Sokol-Hessner et al, 2009; Tom et al, 2007)*
- *Ultimatum Game (e.g., Xiang et al, 2013)*
- **Two-Step (Daw et al, 2011)*

**Version 0.4.1. or later*

What tasks and models are available in 2018?

Ahn & Busemeyer (2016) Curr Opin Behav Sci

- Balloon Analogue Risk Task (BART) (Wallsten et al, 2005)
- Choice under Risk and Ambiguity (Levy et al, 2010)
- Choice reaction time → sequential sampling models
- Delay Discounting (e.g., Mazur, 1987)
- Iowa Gambling (Bechara et al, 1994)
- (Orthogonalized) Go/Nogo (Guitart-Masip et al, 2012)
- Peer influence task (Chung et al, 2015)
- Probabilistic Selection task (e.g., Frank et al, 2007)
- Two-armed Bandit (Experience-based) including Reversal Learning (e.g., Erev et al, 2010)
- Four-armed Bandit (Experience-based) (e.g., Seymour et al, 2012)
- Two-choice Description-based (e.g., Sokol-Hessner et al, 2009; Tom et al, 2007)
- Ultimatum Game (e.g., Xiang et al, 2013)
- Two Step task (Daw et al, 2011)
- Wisconsin Card Sorting task (Bishara et al, 2010)

** hBayesDM in Python **

Reinforcement Learning (& Decision Making) with the hBayesDM package

"(Cutting-edge) Modeling Can Be as Easy as Doing a T-Test"

Woo-Young (Young) Ahn

Practical Session D

Friday 1:30pm – 5:00pm

Thank you!