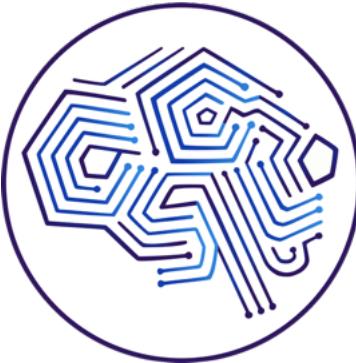


Reinforcement Learning



*Woo-Young (Young) Ahn
Computational Clinical Science Lab
Department of Psychology
Seoul National University
ccs-lab.github.io*



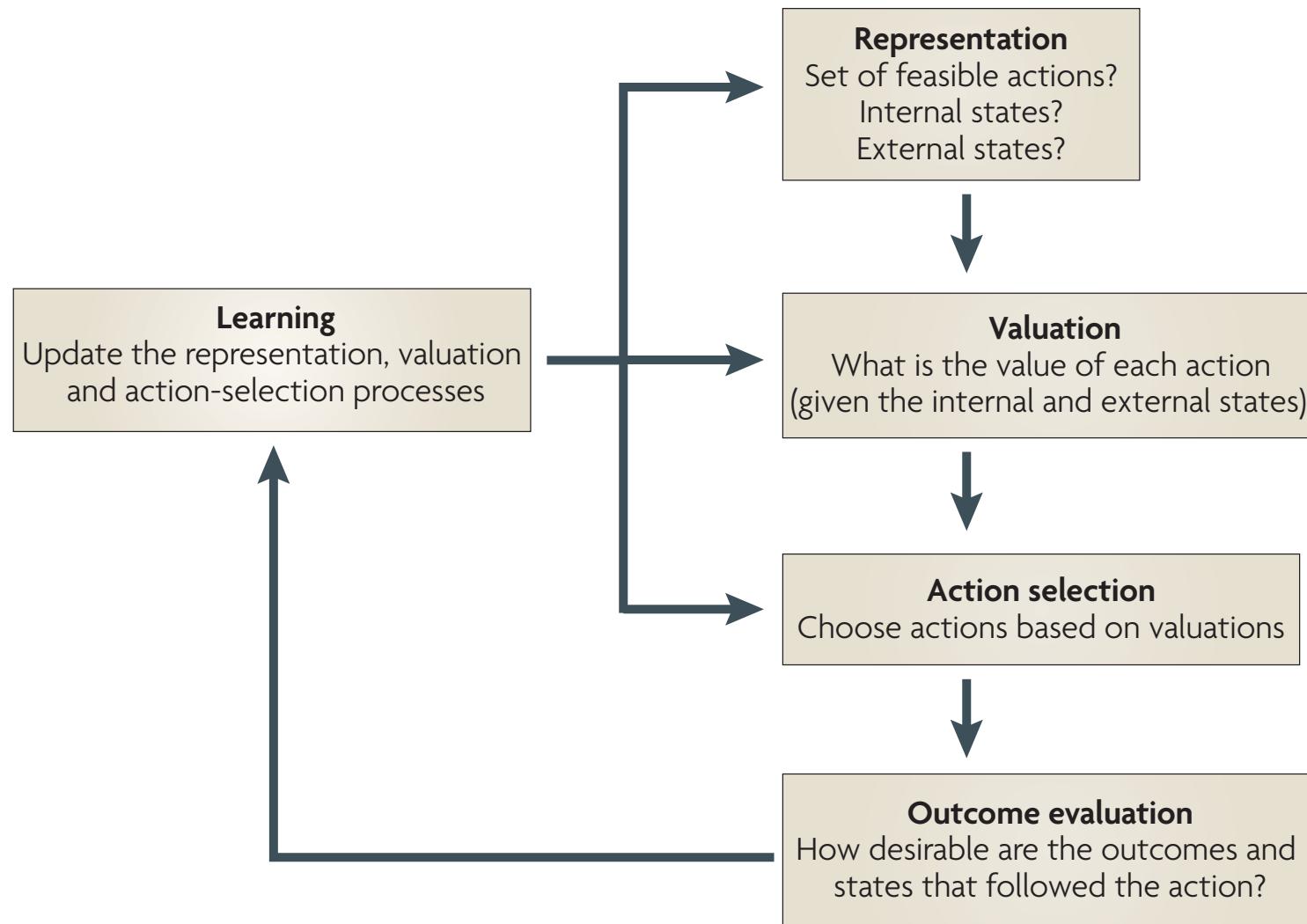
Reinforcement Learning (RL)

- *What is RL?*
 - *RL in human research vs RL in AI*
- *RL models (algorithms for prediction and control)*
 - *Classical conditioning*
 - *Rescorla-Wagner (R-W) model*
 - *(Bayesian or non-Bayesian) extension of R-W models*
 - *Operant (instrumental) conditioning*
 - *Model-free vs Model-based learning*
 - *Pavlovian control vs Instrumental control*
- *Adaptive Design Optimization within the RL framework*
- *Limitations & Future directions*

Learning objectives

Participants will...

- *Understand the key concepts and notations of RL (in multiple fields)*
- *Know (some of) popular RL models (& references)*
 - *Simple to complex models*
- *Limitations of RL and some new approaches*



What is RL?

*“Learning what to do” ...
based on (an incomplete history of)
rewards and punishments*

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

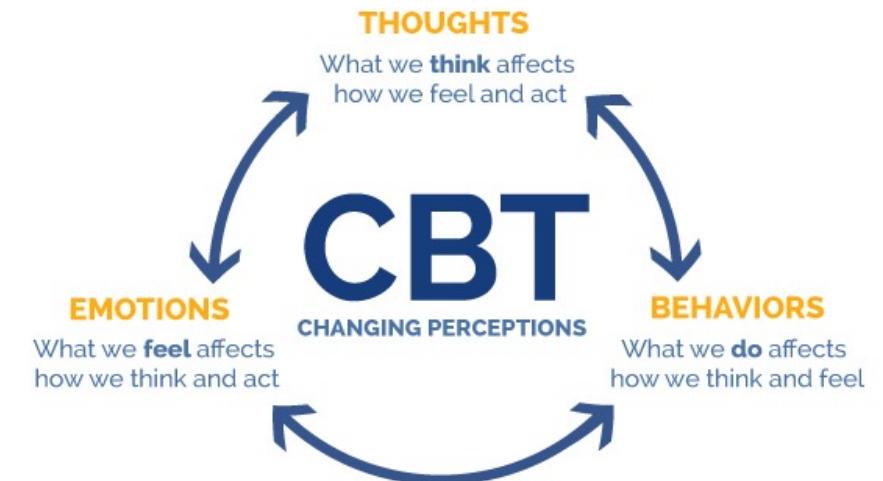
*“Learn optimal ways to make decisions”
in an uncertain environment*



Mnih et al (2015) Nature



Silver et al (2016) Nature



RL is a type of Machine Learning

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

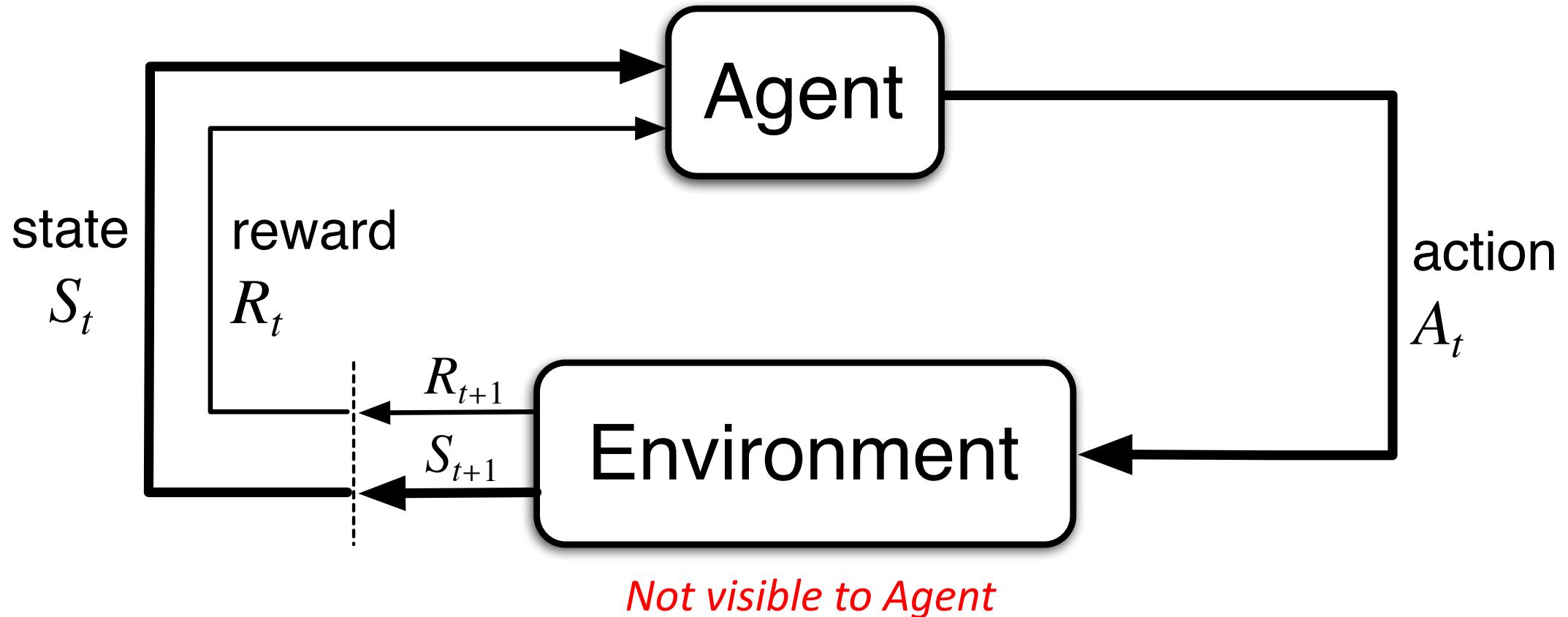
Q) How is RL different from other ML paradigms?

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

“Learn optimal ways to make decisions” in an uncertain environment

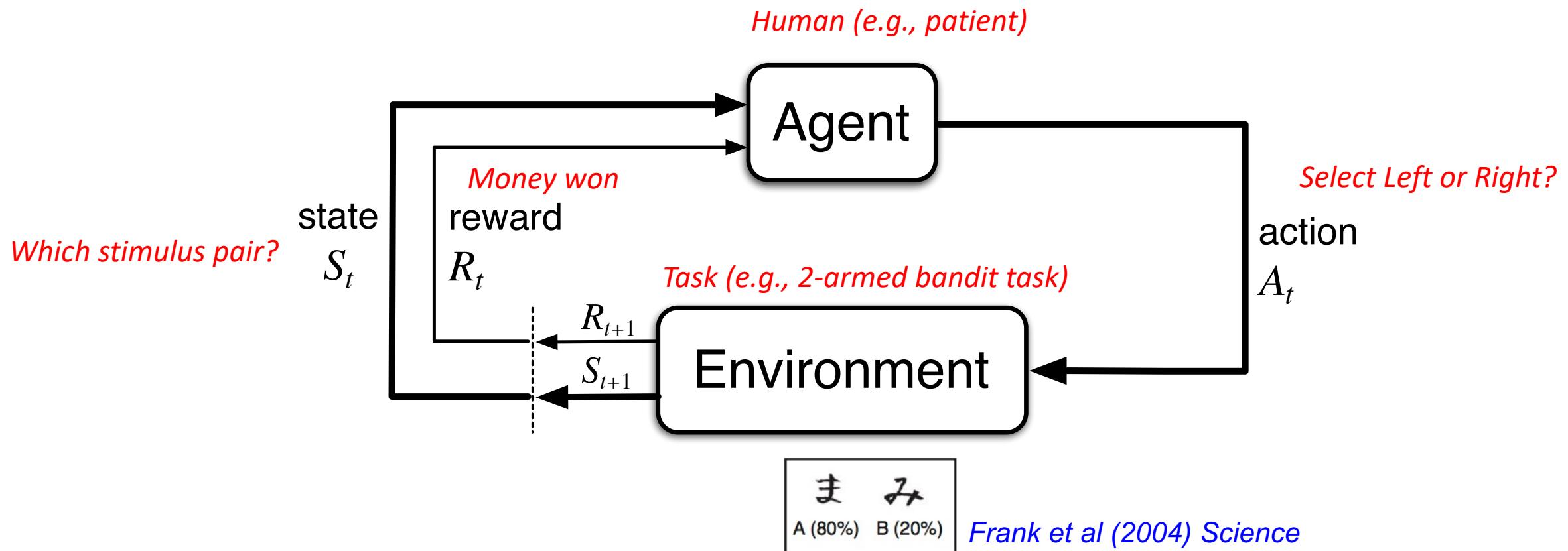
	<i>RL in human research</i>	<i>RL in AI</i>
<i>Goal</i>	<i>Characterize individual differences</i>	<i>Generate optimal solution</i>
<i>Amount of data</i>	<i>Small</i>	<i>Very large</i>
<i># parameters</i>	<i>Typically < 10</i>	<i>A lot</i>
<i>Parameter estimation</i>	<i>Important</i>	<i>Estimate? Often fixed</i>

Agent-Environment Interface
(check MDP lecture by Matthew Nassar)
e.g., Maze task, Tree search, N-armed Bandit



Typically in Computational Psychiatric research settings..

Model parameters → Psychologically meaningful processes/constructs

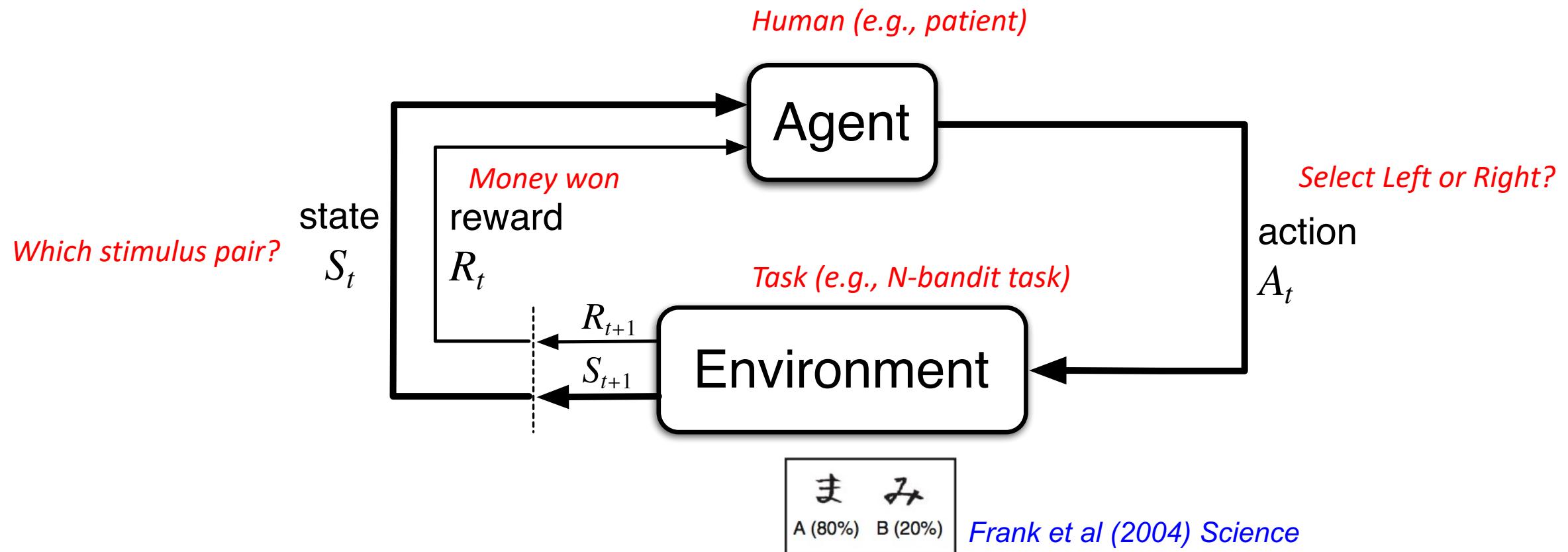


S_t : State value on time (trial) t

A_t : Action value on time (trial) t

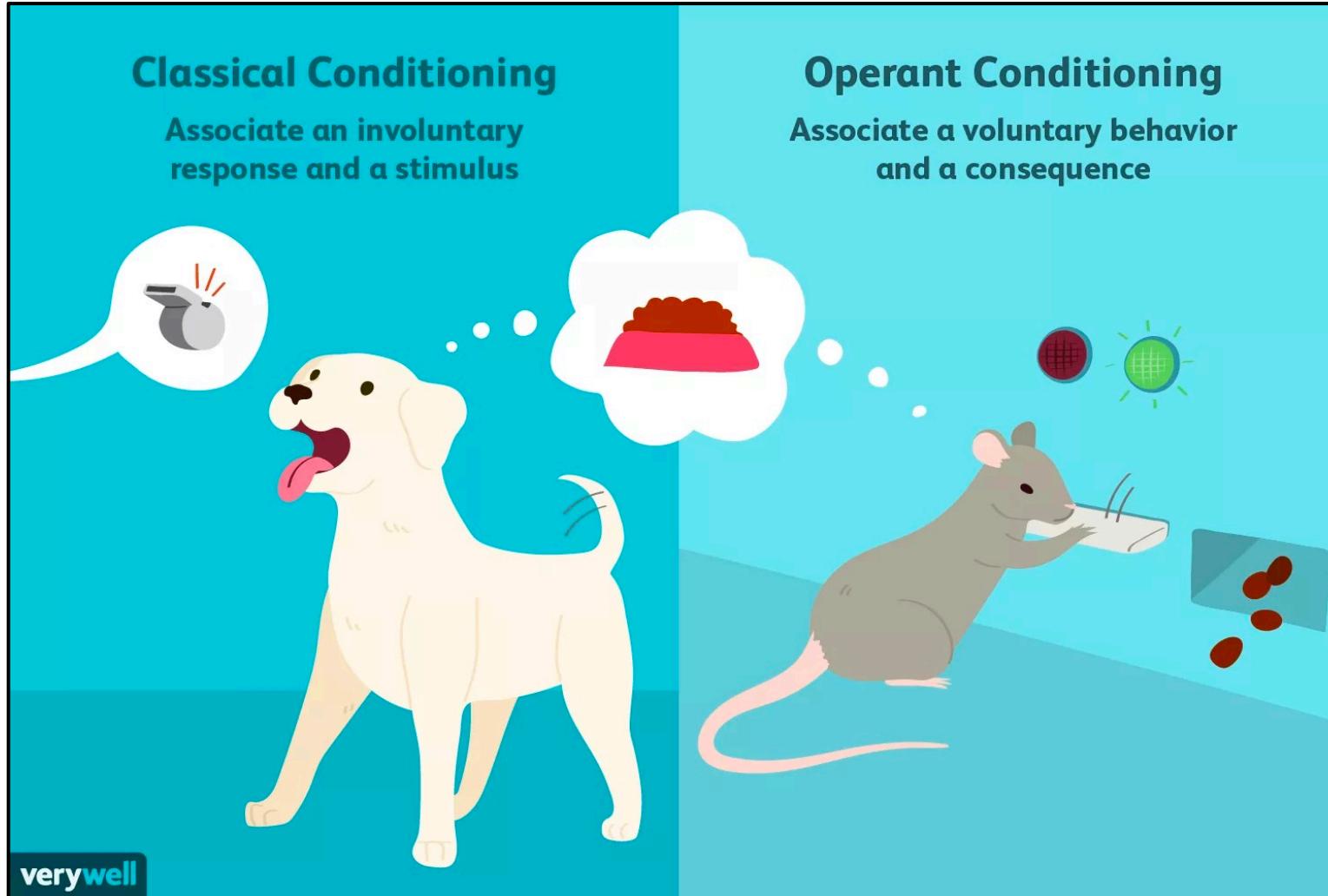
R_t : Reward on time (trial) t

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



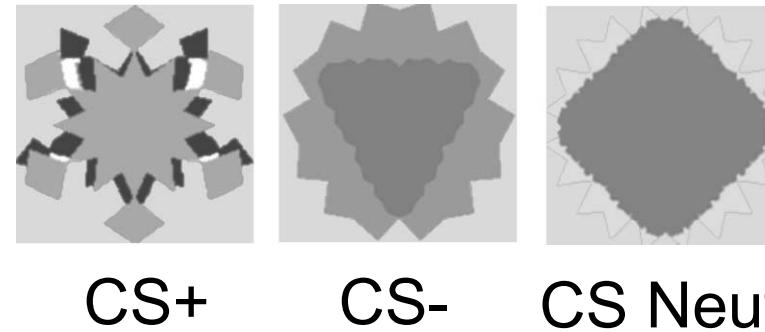
RL models (algorithms for prediction)

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



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

*Classical conditioning
(No action required)*



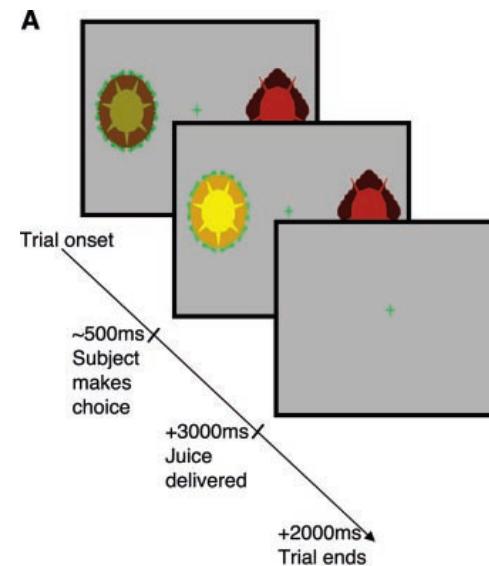
CS+

CS-

CS Neut

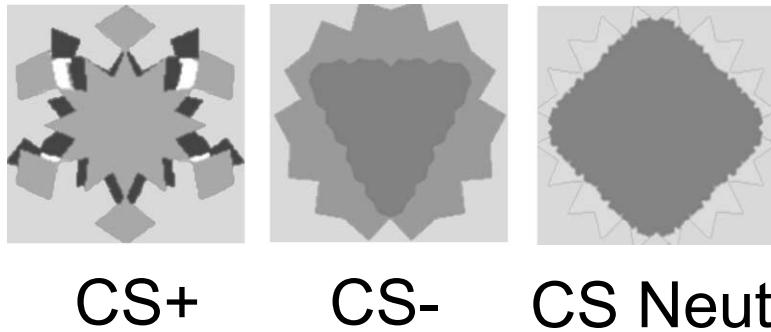
e.g., O'Doherty et al (2003) *Neuron*

*Operant (Instrumental)
Conditioning (Action required)*



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

Classical conditioning



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

Rescorla-Wagner (R-W) model

→ Point estimates of V_t

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

Stimulus
value (t)

Stimulus
value (t-1)

Outcome

Stimulus
value (t-1)



Prediction error

Classical conditioning



CS+ CS- CS Neut

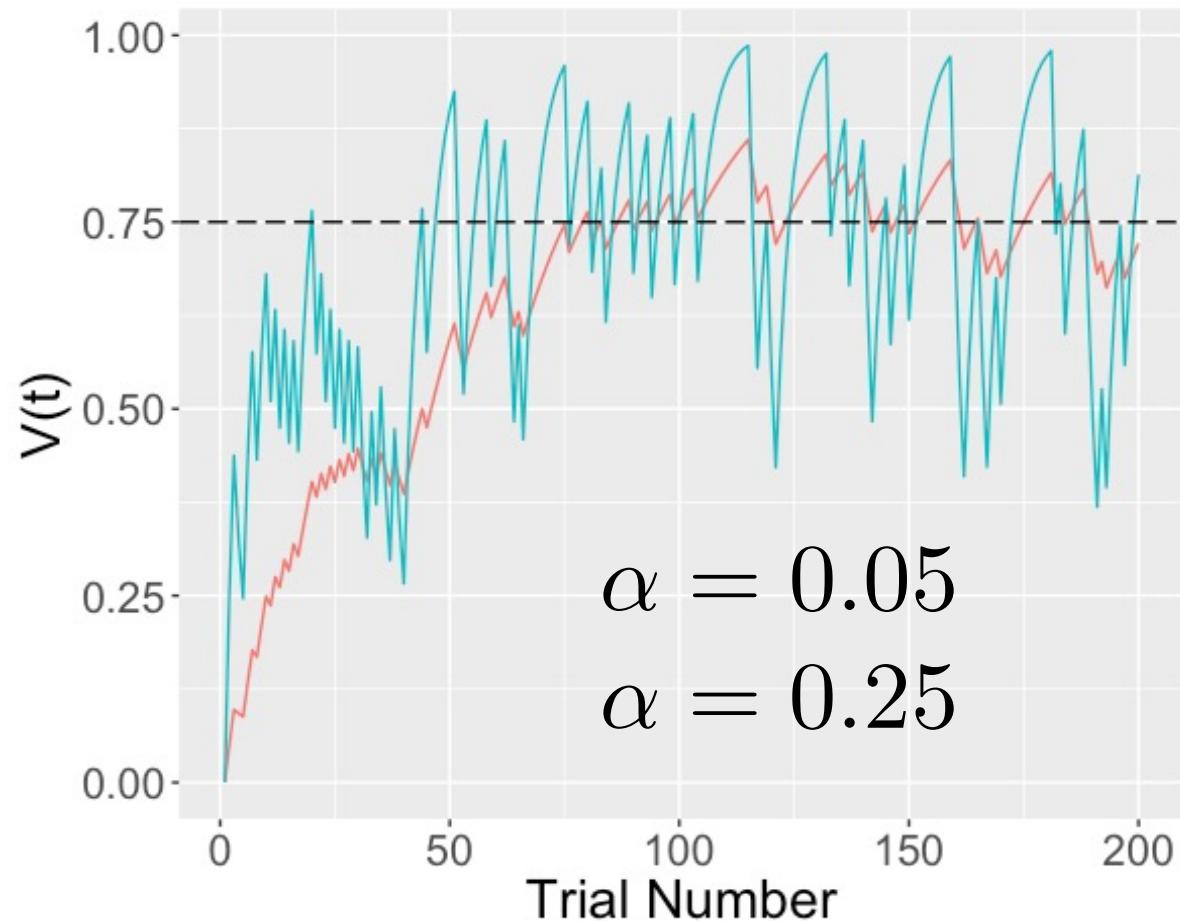
e.g., O'Doherty et al (2003) *Neuron*

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

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

Dayan et al (2000); Kakade & Dayan (2002)
Daw et al (2006); Kruschke (2008); Mathys et al (2011; 2014)

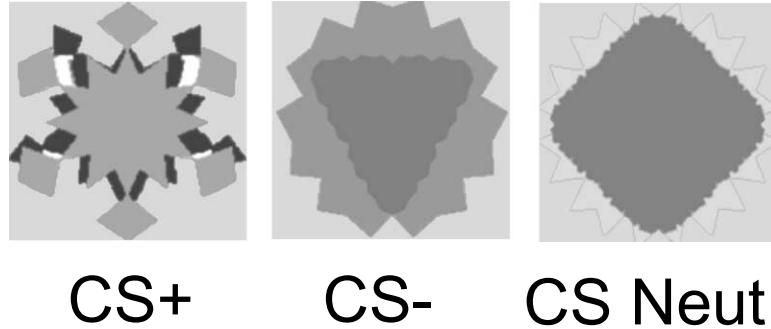
e.g., Reward rate = 0.75



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

Also see Maaten Speekenbrink's blogs
<https://speekenbrink-lab.github.io/blog/>

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

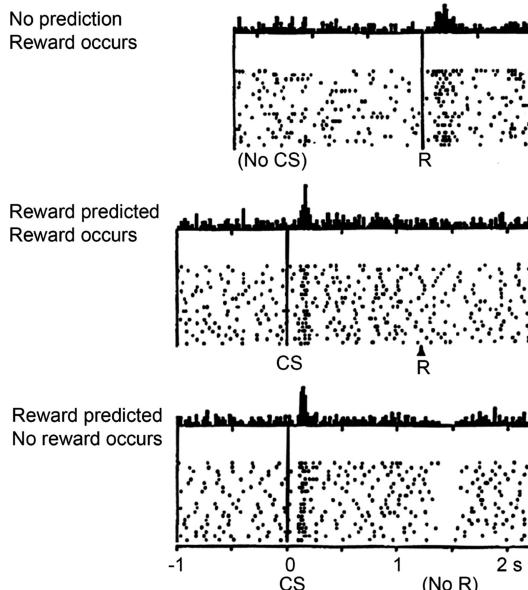
P. Read Montague^{1,2}, Steven E. Hyman³ & Jonathan D. Cohen^{4,5}

¹Department of Neuroscience and ²Menninger Department of Psychiatry and Behavioral Sciences, Baylor College of Medicine, 1 Baylor Plaza, Houston, Texas 77030, USA (e-mail: read@bcm.edu)

³Harvard University, Cambridge, Massachusetts 02138, USA (e-mail: seh@harvard.edu)

⁴Department of Psychiatry, University of Pittsburgh and ⁵Department 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

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

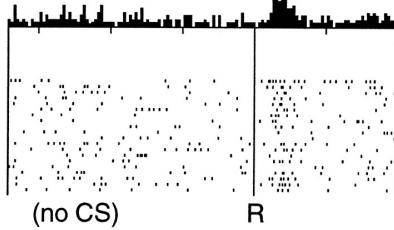
$\approx \text{current reward} + \gamma \cdot \text{next prediction} - \text{current prediction}$

Sutton & Barto (1998) Reinforcement Learning

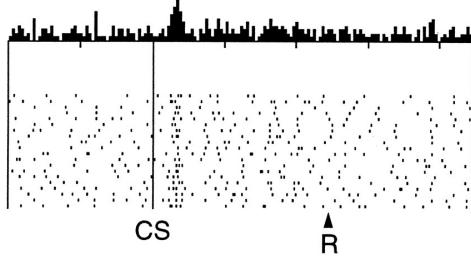
Q) How TD learning accounts for the phasic response of a dopamine neuron?

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

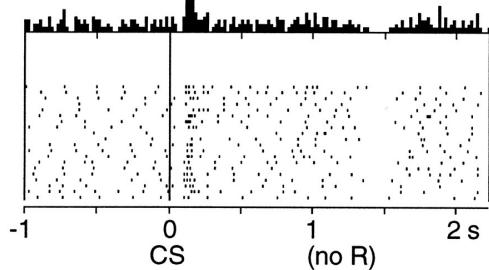
No prediction
Reward occurs



Reward predicted
Reward occurs



Reward predicted
No reward occurs



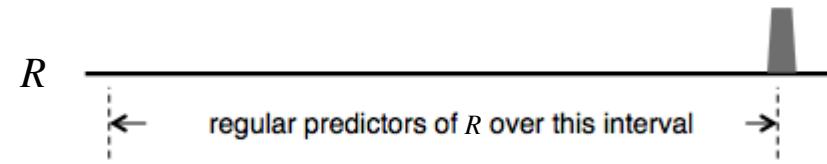
$$\gamma = 1$$

early in learning

learning complete

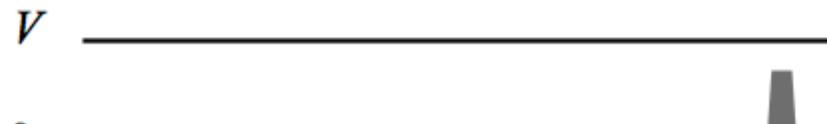
R omitted

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



Reward onset

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



Cue onset

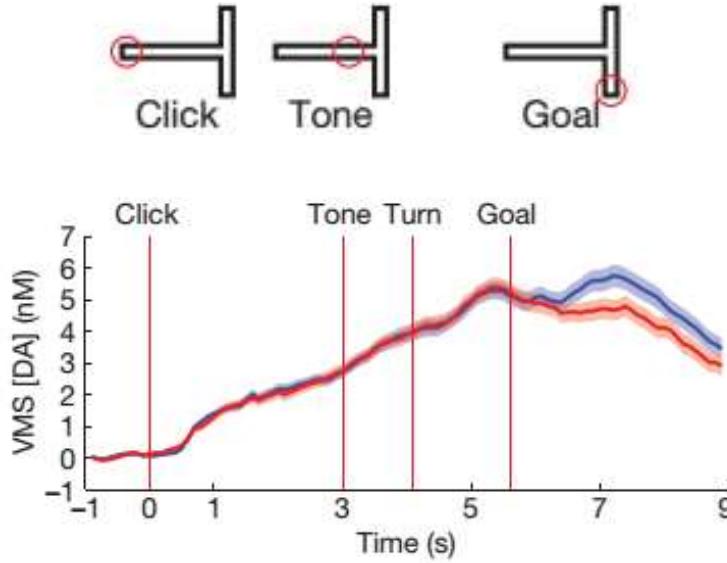
$$\delta_t = R_t + V_t - V_{t-1} = 0 + R_t - 0 = R_t$$



Reward onset

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

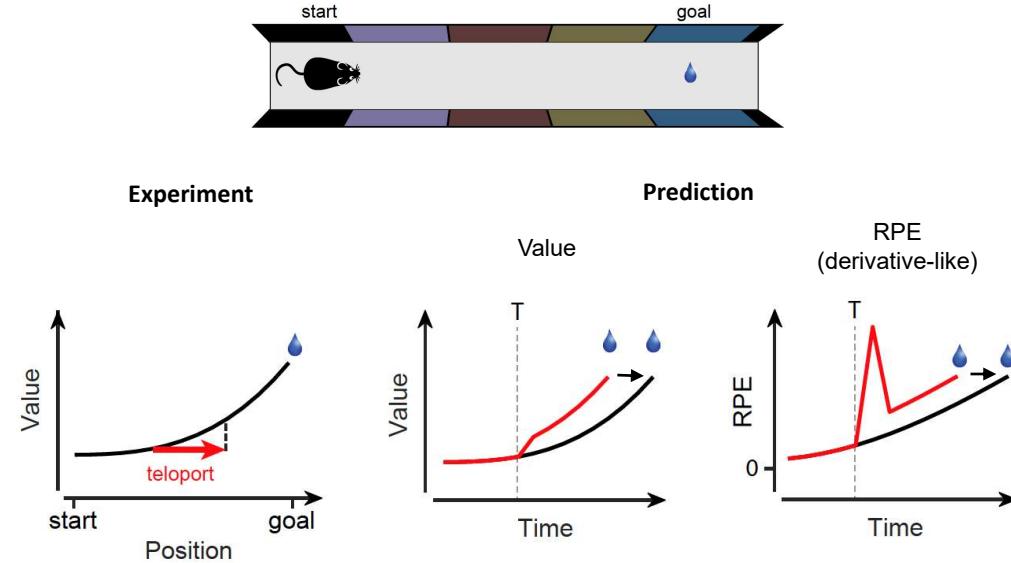
Continued discussion about reward prediction error



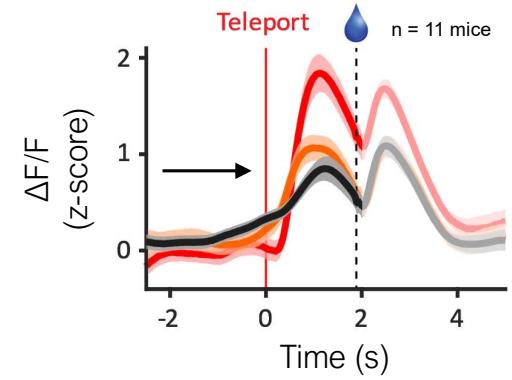
Howe et al (2013) Nature

State value $V(t)$?
Or moment-by-moment TD error?

Gershman (2014) Neural Computation



Kim et al (2020) Cell



Instrumental learning

Model-based vs Model-free

Model-based vs Model-free

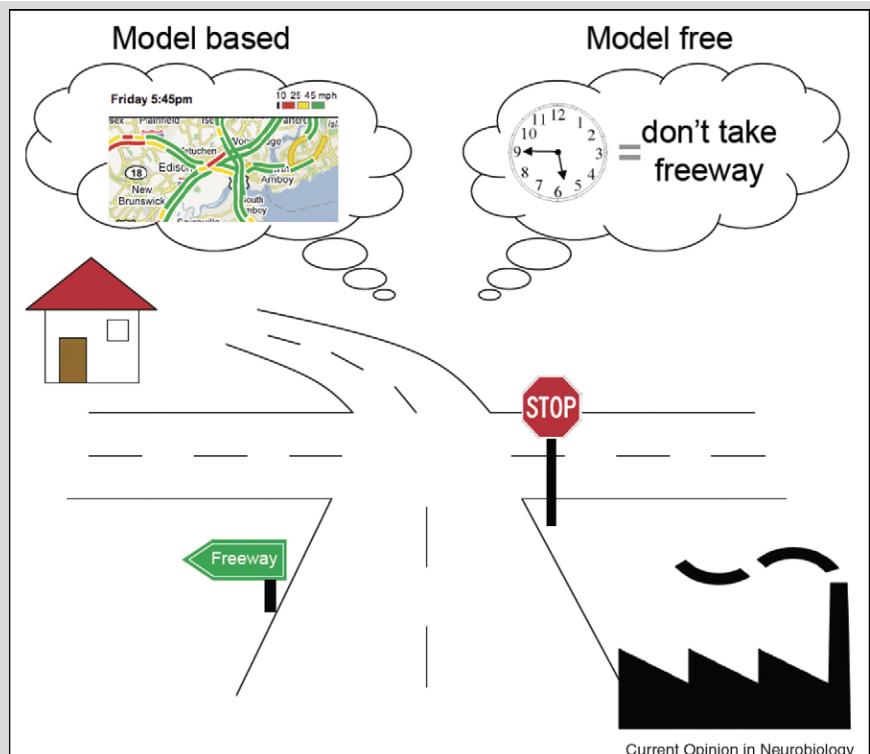


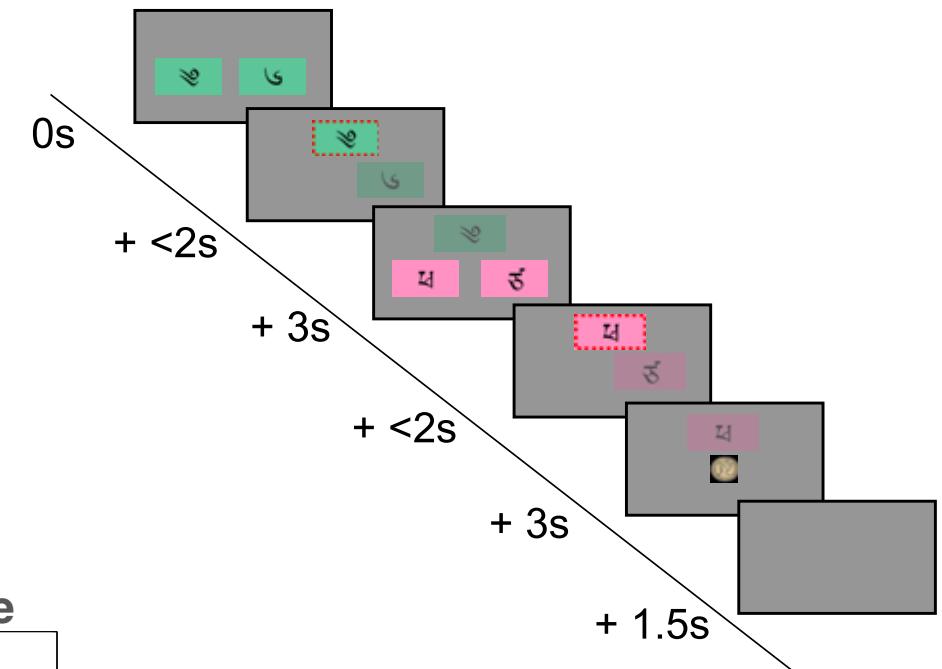
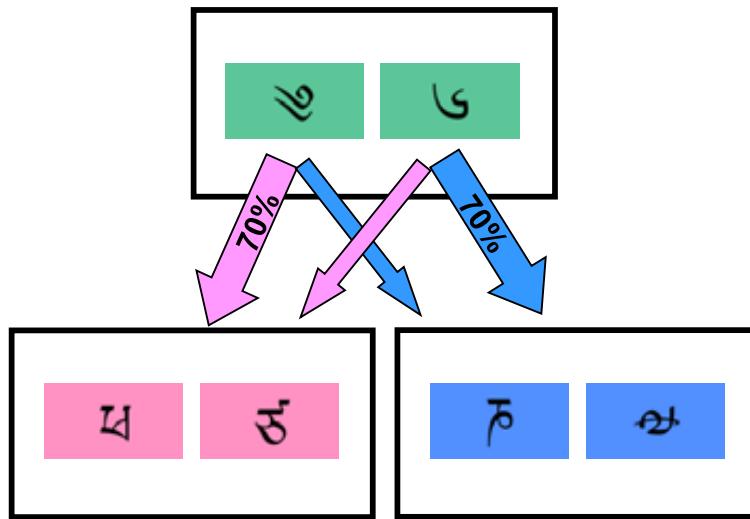
Figure 1: Two ways to choose which route to take when traveling home from work on friday evening.

Dayan & Niv (2008)

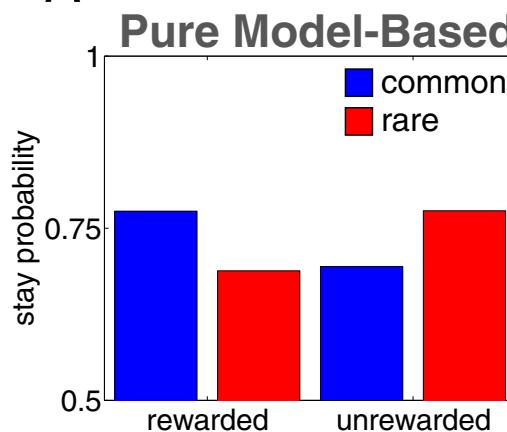
- *Model-based (goal-directed) learning: build a model of an environment. Effortful but flexible.*
- *Model-free (habitual) learning: relies on trials-and-errors. Efficient but inflexible.*
- *(Clinical) examples: compulsive behaviors, etc.*

Two-Step task

Daw et al (2011) Neuron

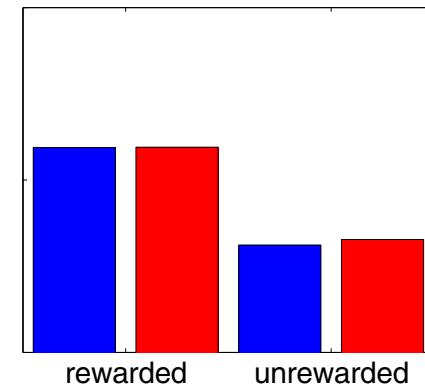


A



B

Pure Model-Free

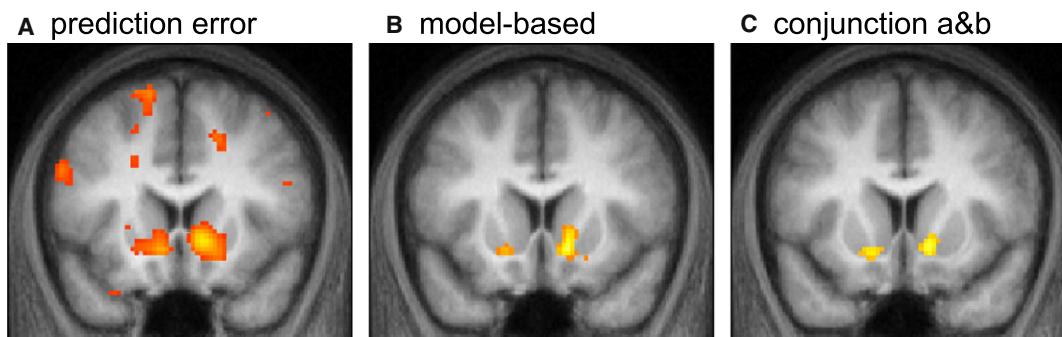


Computational model

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) $\leq \omega_{MB} \leq 1$ (completely model-based)

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

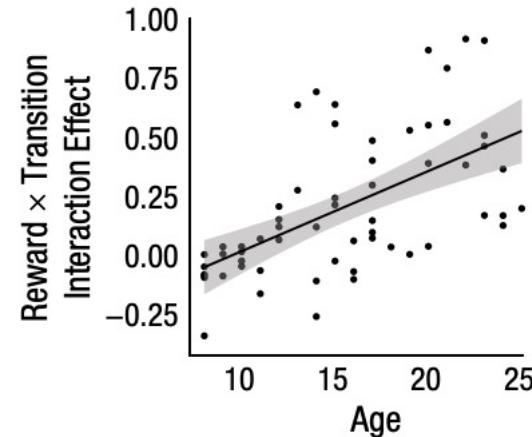


Daw et al (2011) Neuron

Individual differences in MB-MF

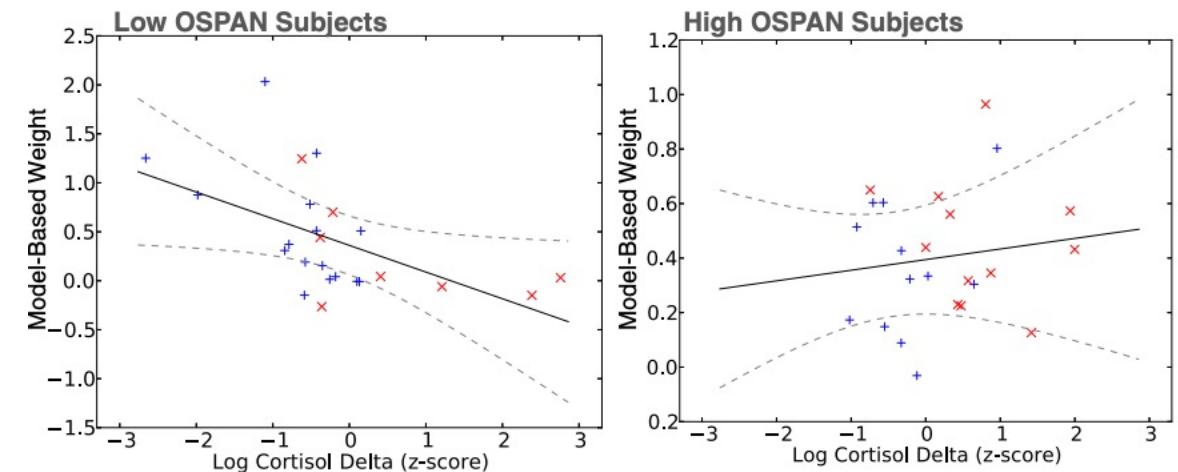
- *Age*
- *WM and stress*

Otto et al (2013) PNAS



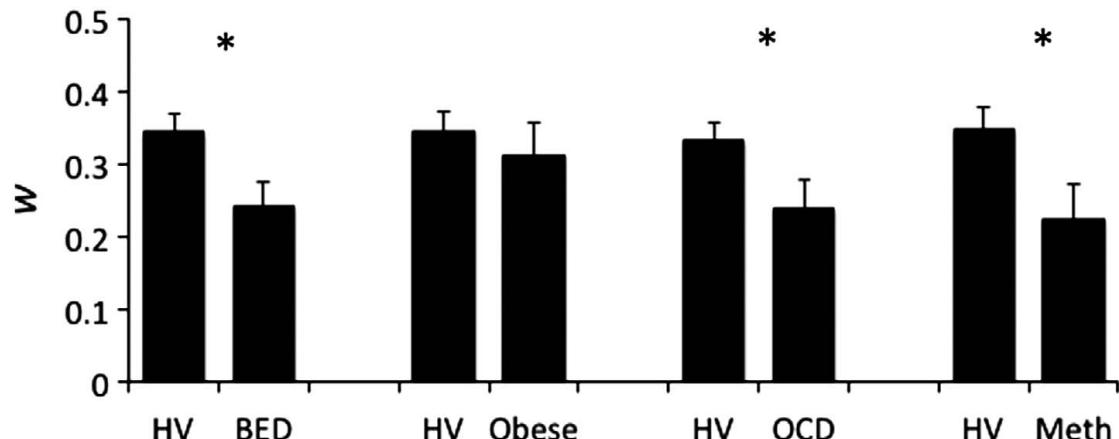
Decker et al (2016) Psych Sci

- *Psychiatric disorders*

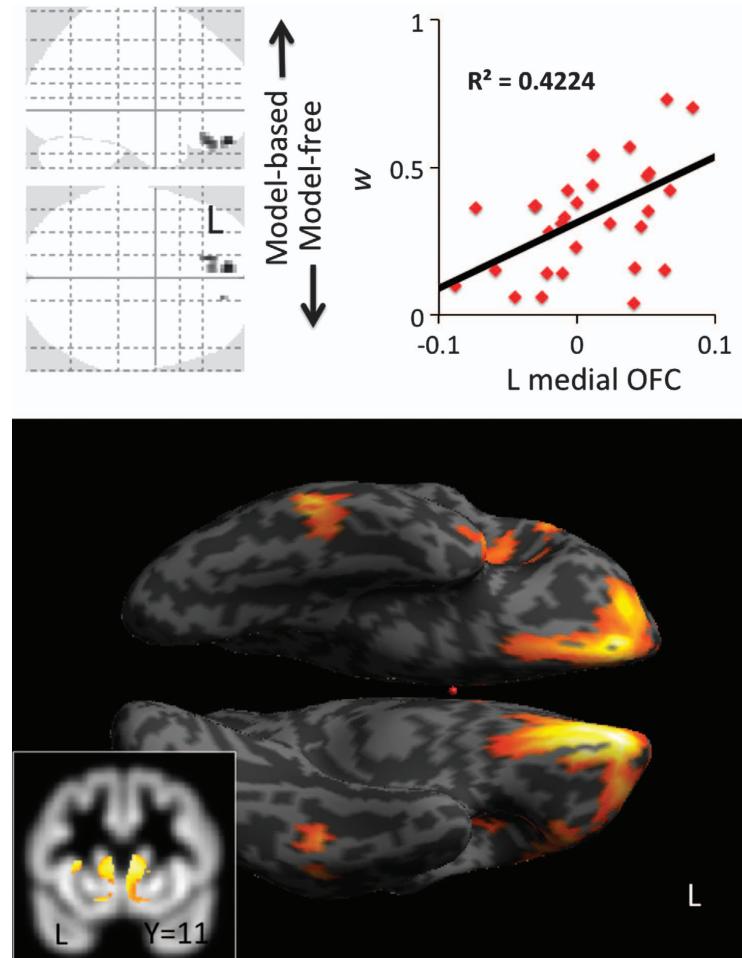


Reliance of model-based control → *Disorders of compulsion*

Voon et al (2014) Molecular Psych



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



RESEARCH ARTICLE

When Does Model-Based Control Pay Off?

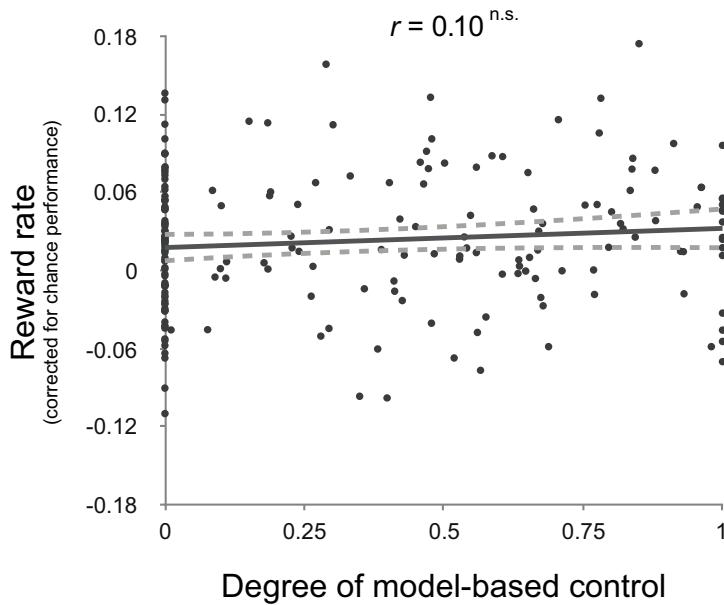
Wouter Kool^{1*}, Fiery A. Cushman¹✉, Samuel J. Gershman^{1,2}✉

1 Department of Psychology, Harvard University, Cambridge, Massachusetts, United States of America,

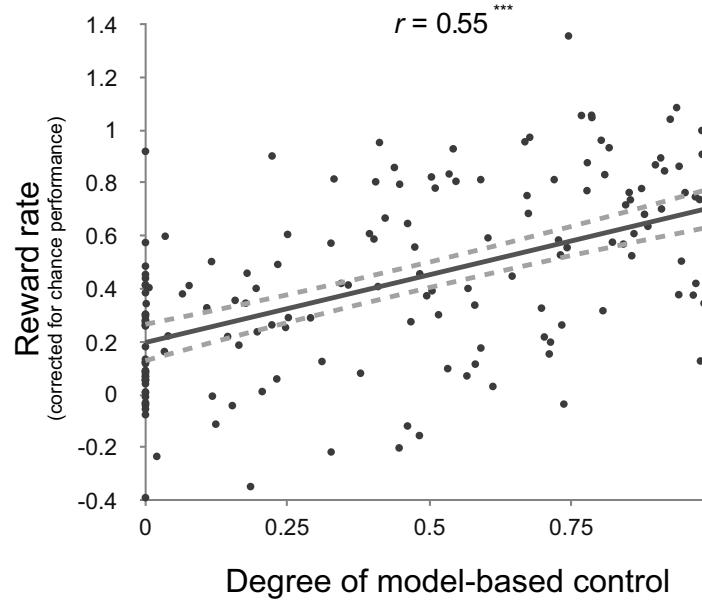
2 Center for Brain Science, Harvard University, Cambridge, Massachusetts, United States of America

✉ These authors contributed equally to this work.

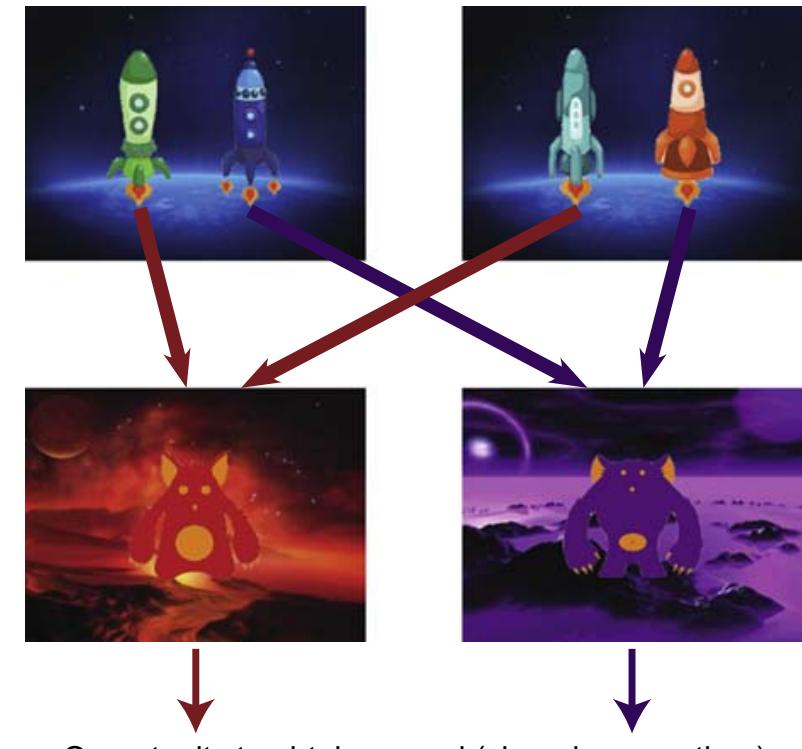
* wkool@fas.harvard.edu



Daw Two-Step Task

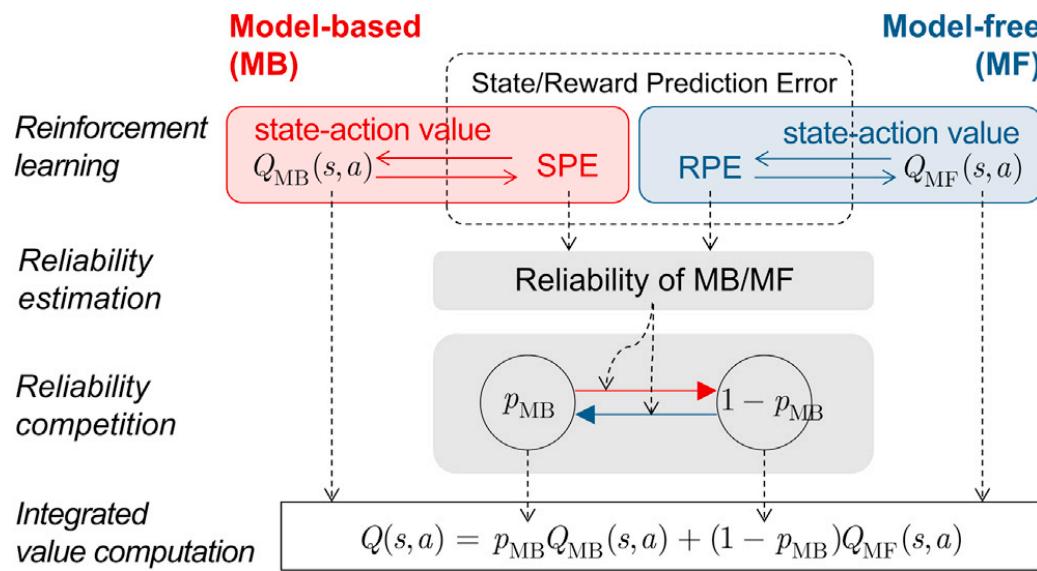


Kool Two-Step Task



Kool et al (2016) PLoS Comput Biol

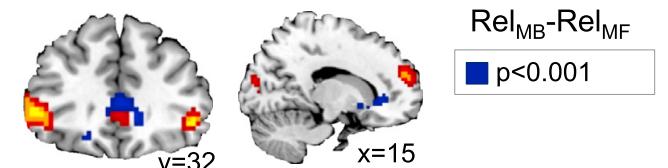
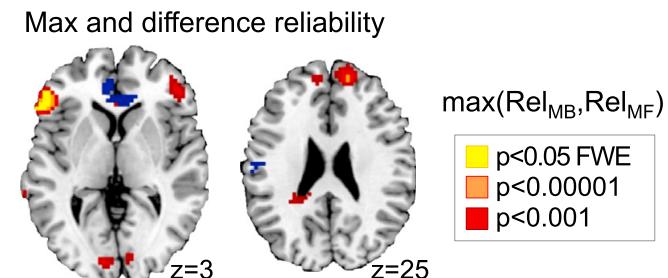
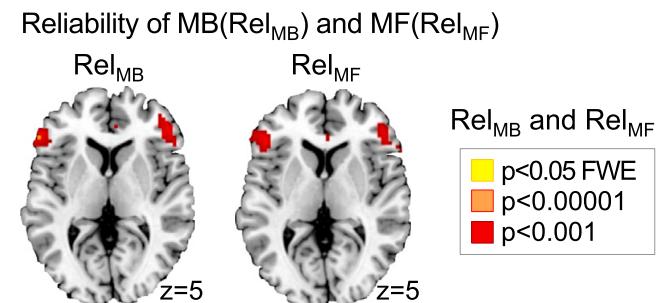
Reliability-based arbitration between model-based and model-free



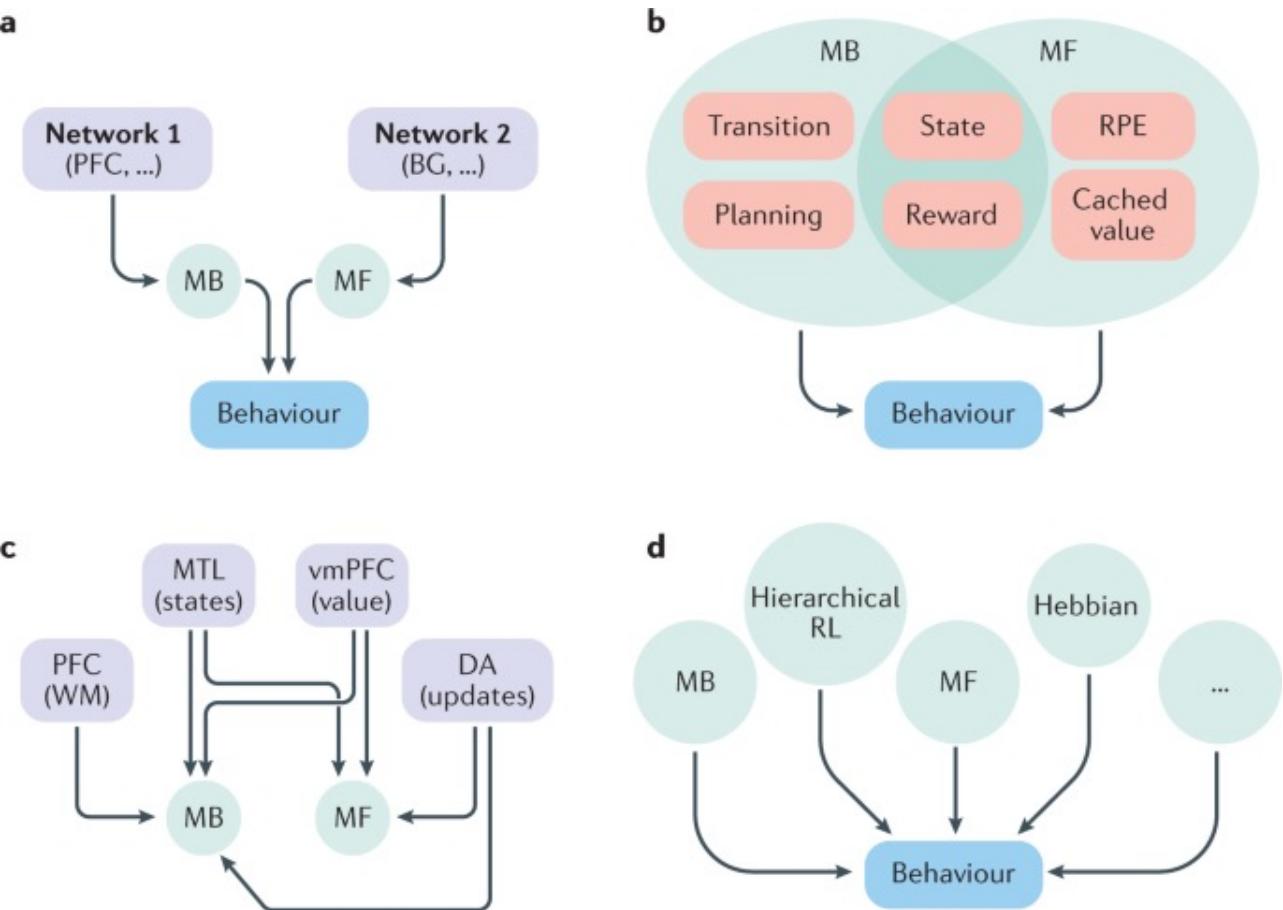
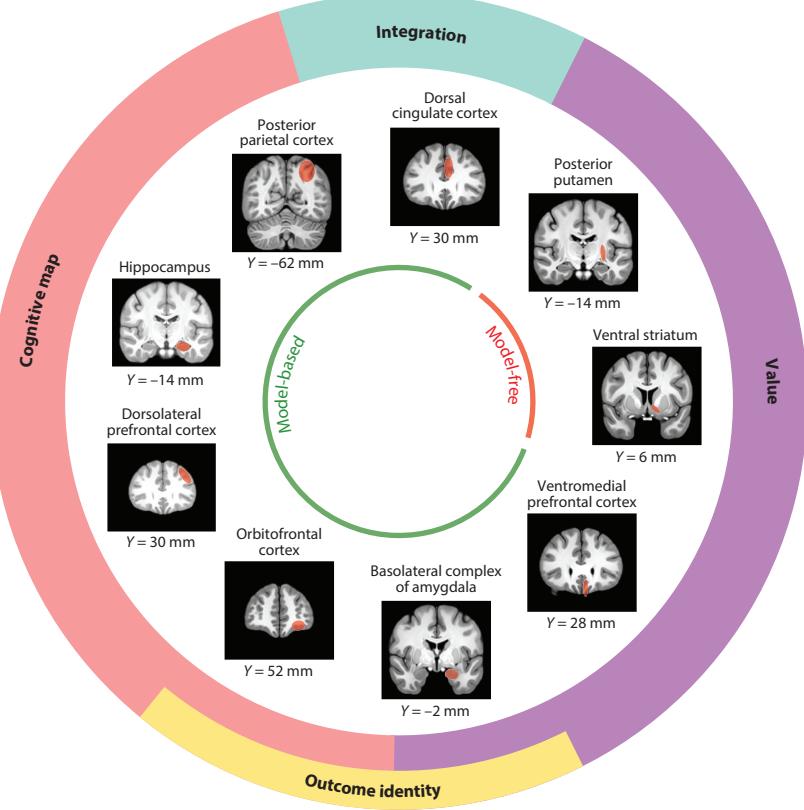
Lee et al (2014) Neuron

Daw et al (2005) Nature Neuroscience

Wang et al (2018) Brain & Neuro. Advances



Inferior lateral prefrontal and frontopolar cortex

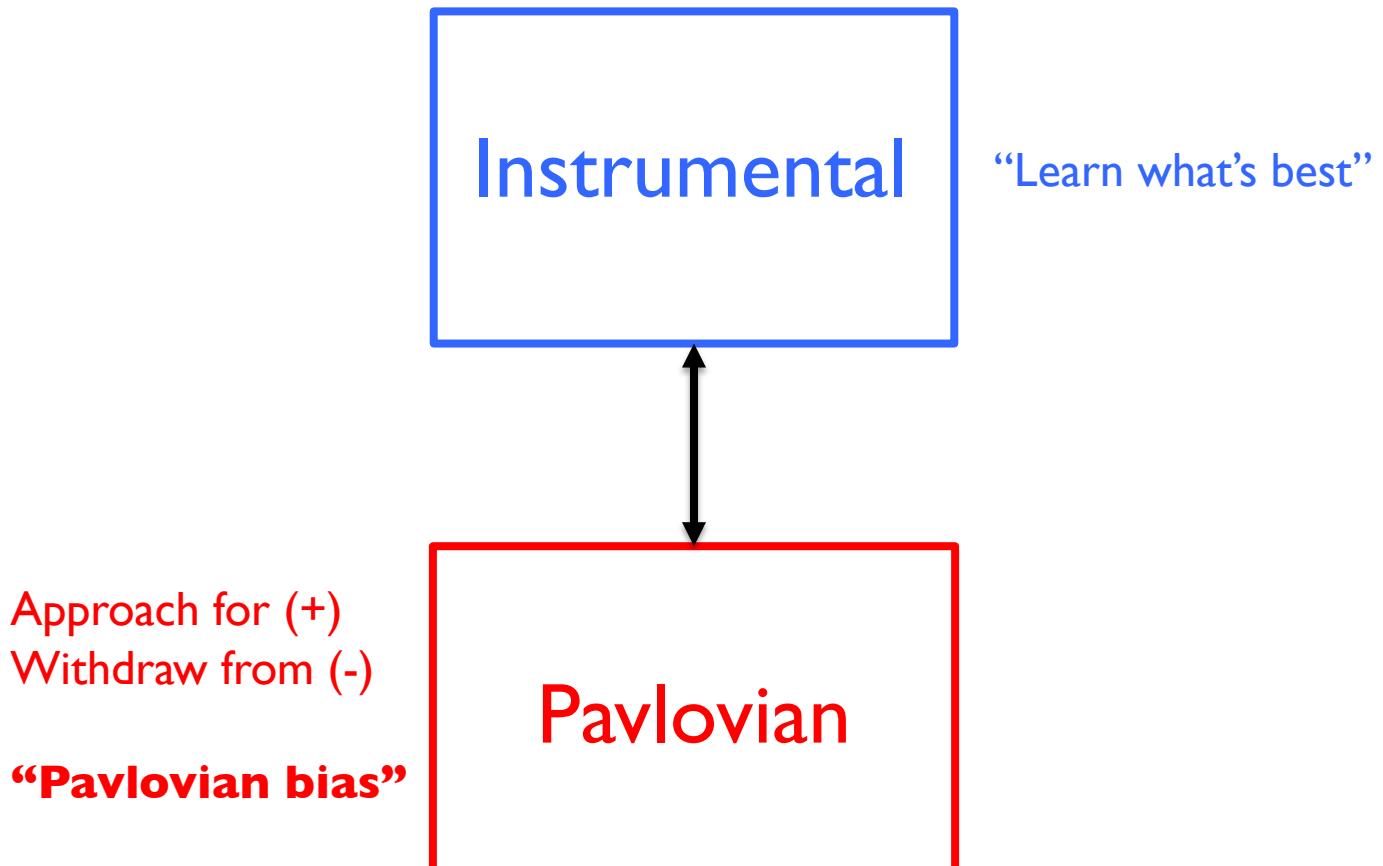


O'Doherty et al (2017) Annu. Rev. Psychology

Collins & Cockburn (2020) Nature Reviews Neuroscience

Pavlovian vs Instrumental control

Pavlovian vs Instrumental control



Opinion

CellPress

Action versus valence in decision making

Marc Guitart-Masip^{1,2}, Emrah Duzel^{3,4,5}, Ray Dolan², and Peter Dayan⁶

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

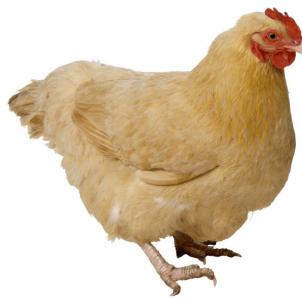
⁴ Otto von Guericke University Magdeburg, Institute of Cognitive Neurology and Dementia Research, D-39120 Magdeburg, Germany

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

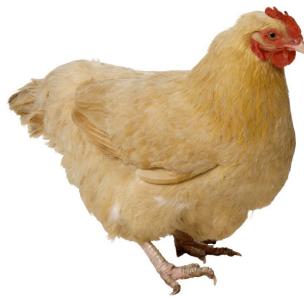


Food!



Hershberger (1986)

Hungry
Chicken

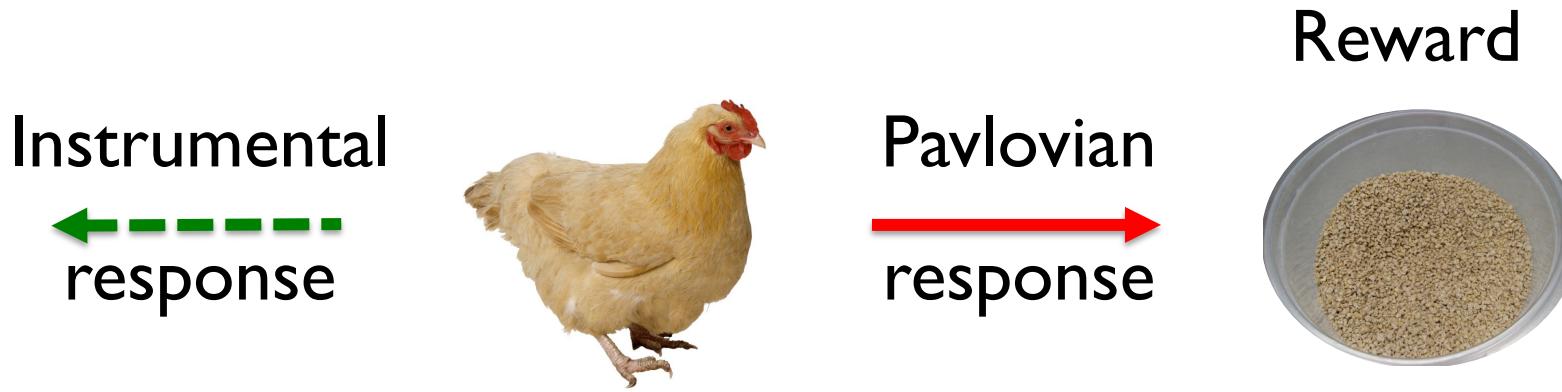


Food!



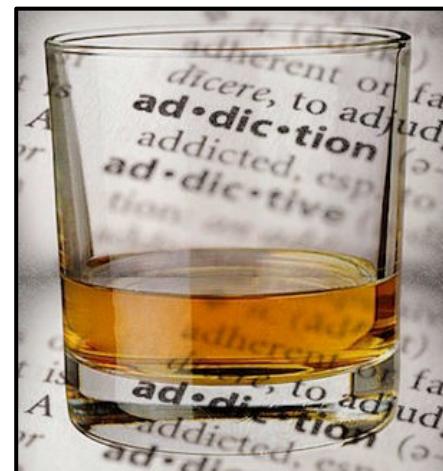
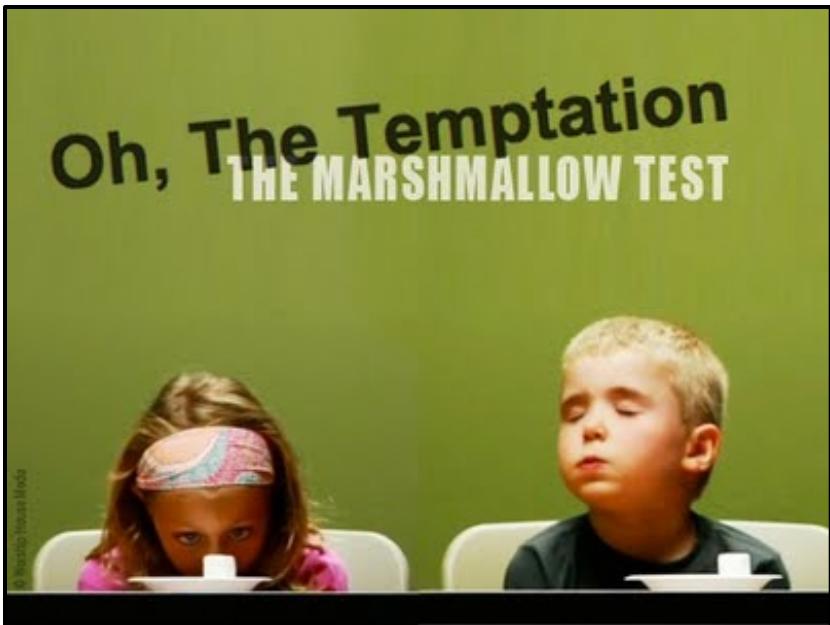
Hershberger (1986)

Pavlovian-Instrumental competition



Hershberger (1986)

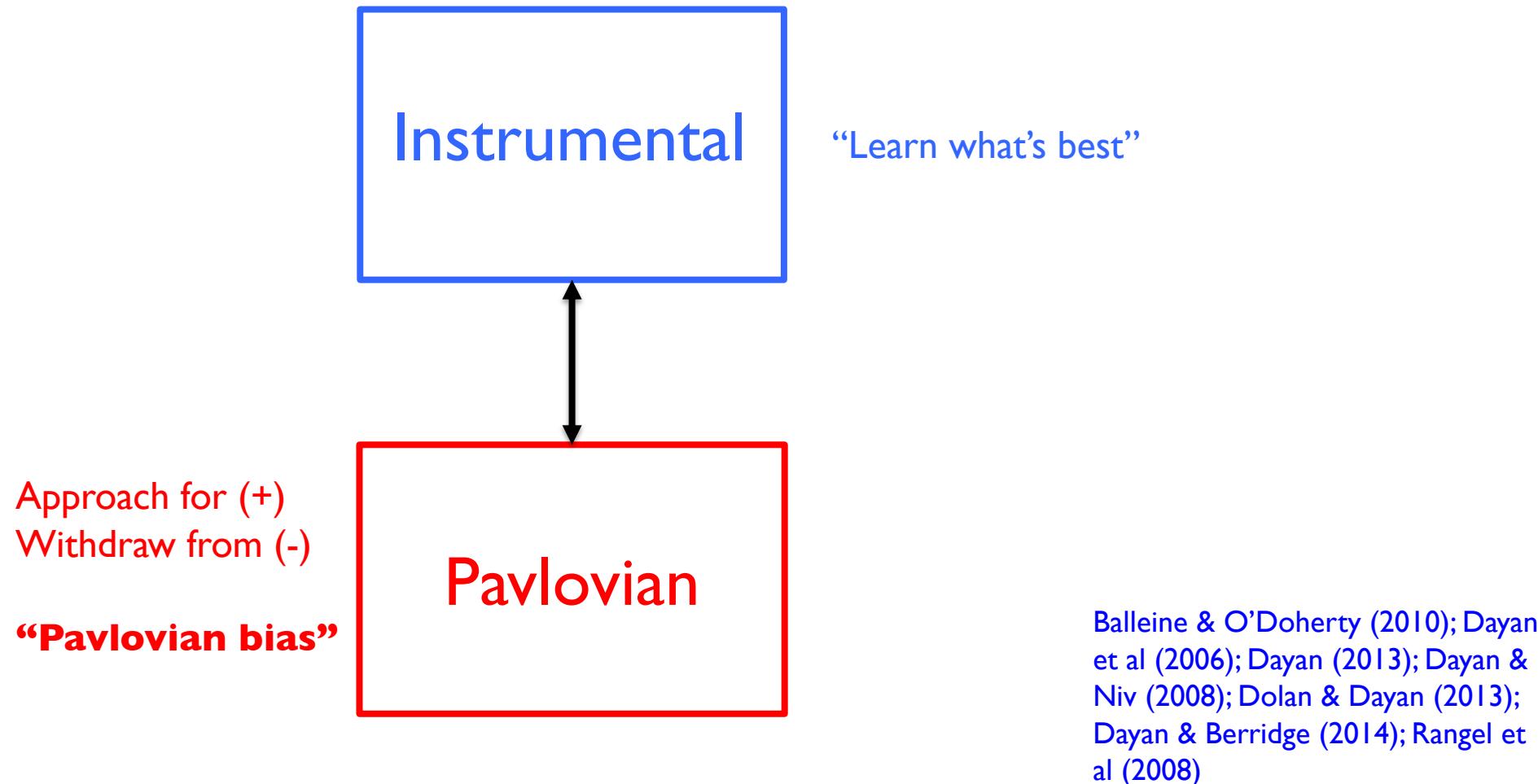
Impulse control



Orthogonalized Go/Nogo task

Pavlovian-Instrumental competition

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



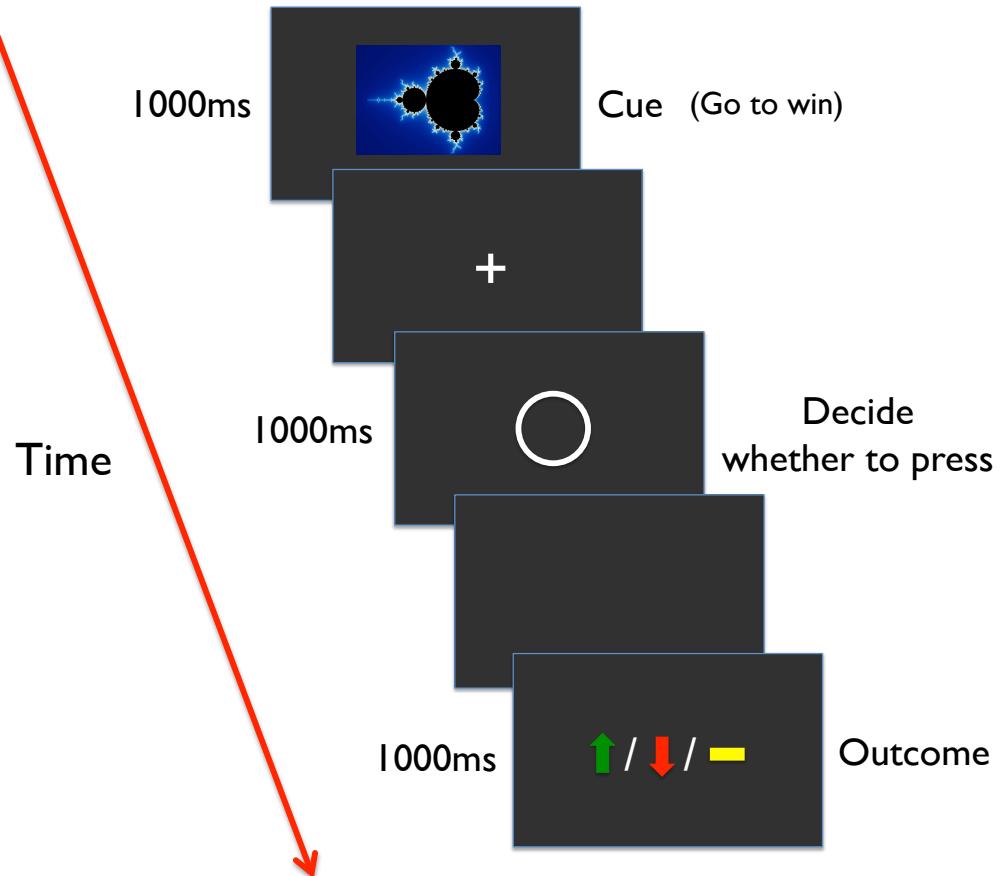
Orthogonalized Go/Nogo task

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



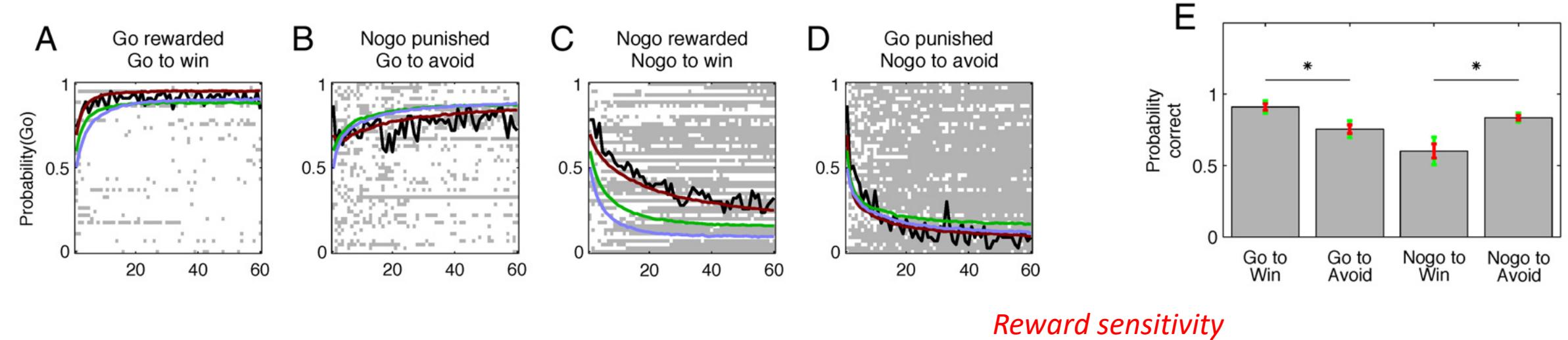
- 4 cues (conditions)

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



Orthogonalized Go/Nogo task

Guitart-Masip et al (2012) Neuroimage



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

Q value

Modified R-W rule

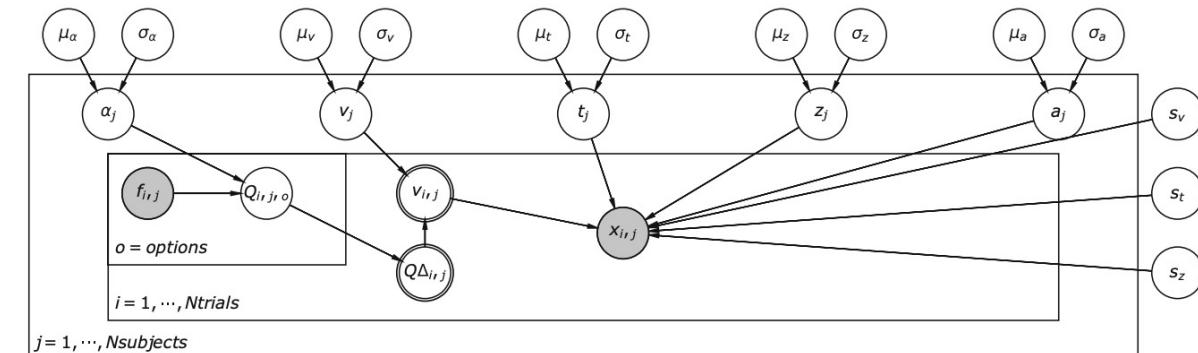
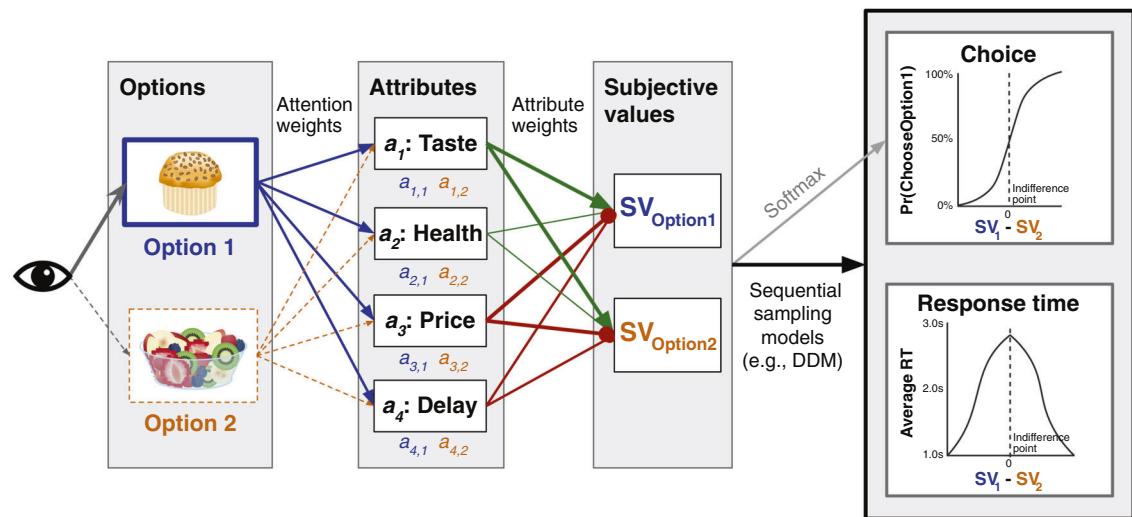
$$W_t(Go_t, s_t) = Q_t(a_t, s_t) + b + \pi V_t(s_t)$$

Go bias

Pavlovian bias

More...

RL + sequential sampling models

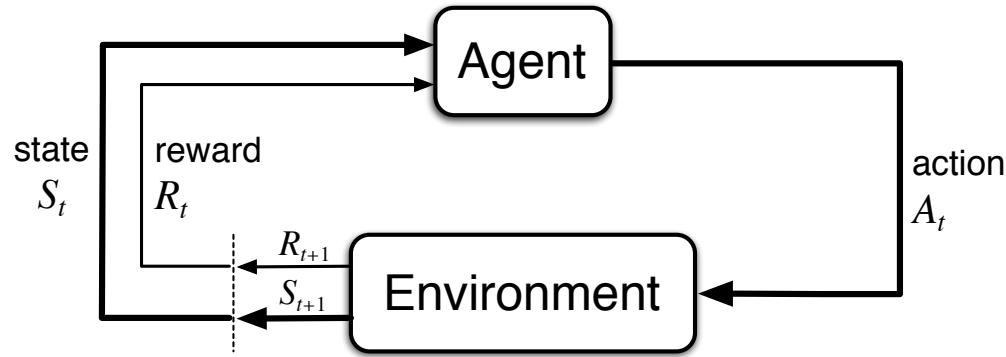


Pedersen & Frank (2020) Computational Brain & Behavior

Collins & Shenhav et al (2021) Neuropsychopharmacology

*Adaptive Design Optimization
within the RL framework*

Optimize experiments on the fly!



$$P(\theta|y) = \frac{P(y|\theta) P(\theta)}{P(y)}$$

Bayesian updating

Update the current state of knowledge with observed response via Bayes rule

Adaptive Design Optimization

Design optimization

Find the most informative design for next experimental trial

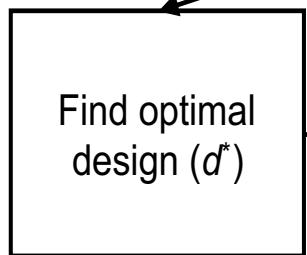
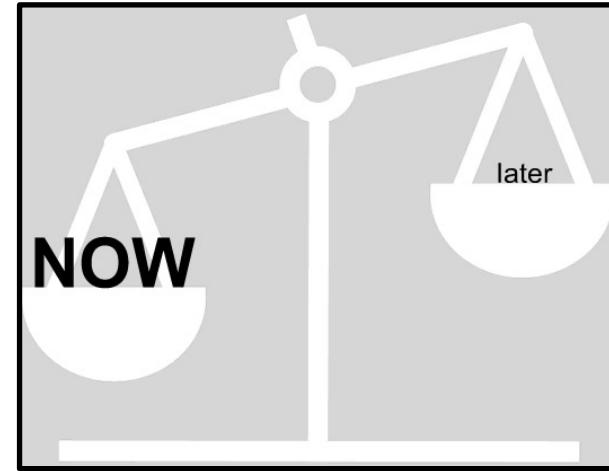
Experiment

Present the optimal design on next trial and record observed response

$$d^* = \operatorname{argmax}_d \iint u(d, \theta, y) P(\theta) P(y|\theta, d) d\theta dy$$

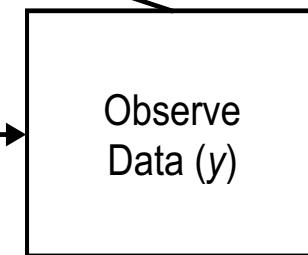
$$p(\theta|y, d) = \frac{p(y|\theta, d)p(\theta)}{p(y|d)}$$

*Bayesian
updating of
model
parameters (θ)*



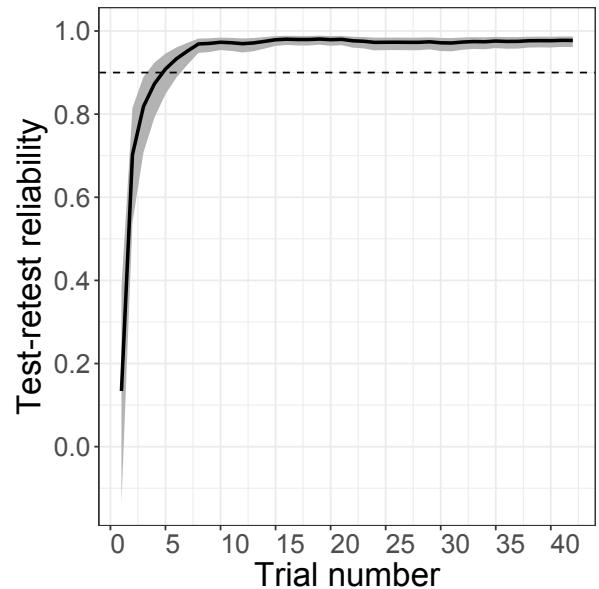
*Design & Conduct a mini-experiment on every trial
“What’s the most informative design (d^*) we should use?”*

e.g., \$320 now vs \$800 in 3 years

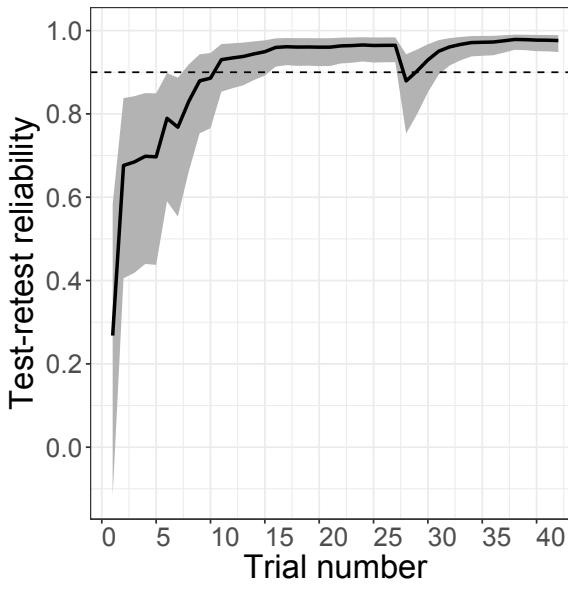


$$d^* = \operatorname{argmax}_d \int \int u(d, \theta, y) p(y|\theta, d) p(\theta) dy d\theta$$

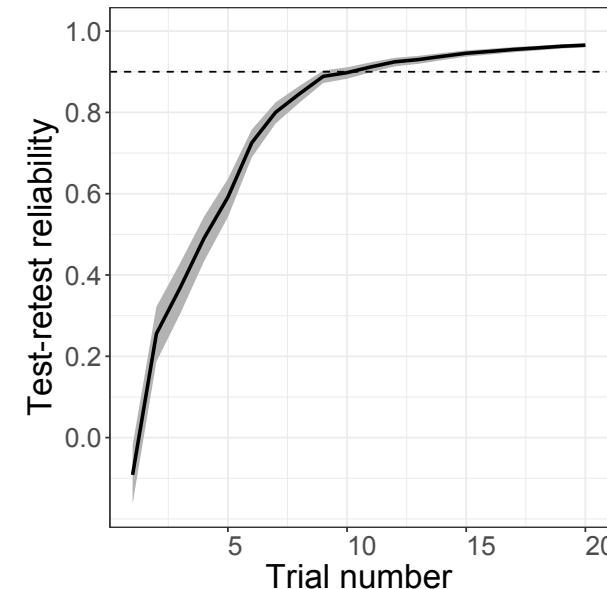
Up to 0.98 test-retest reliability within ~10 trials



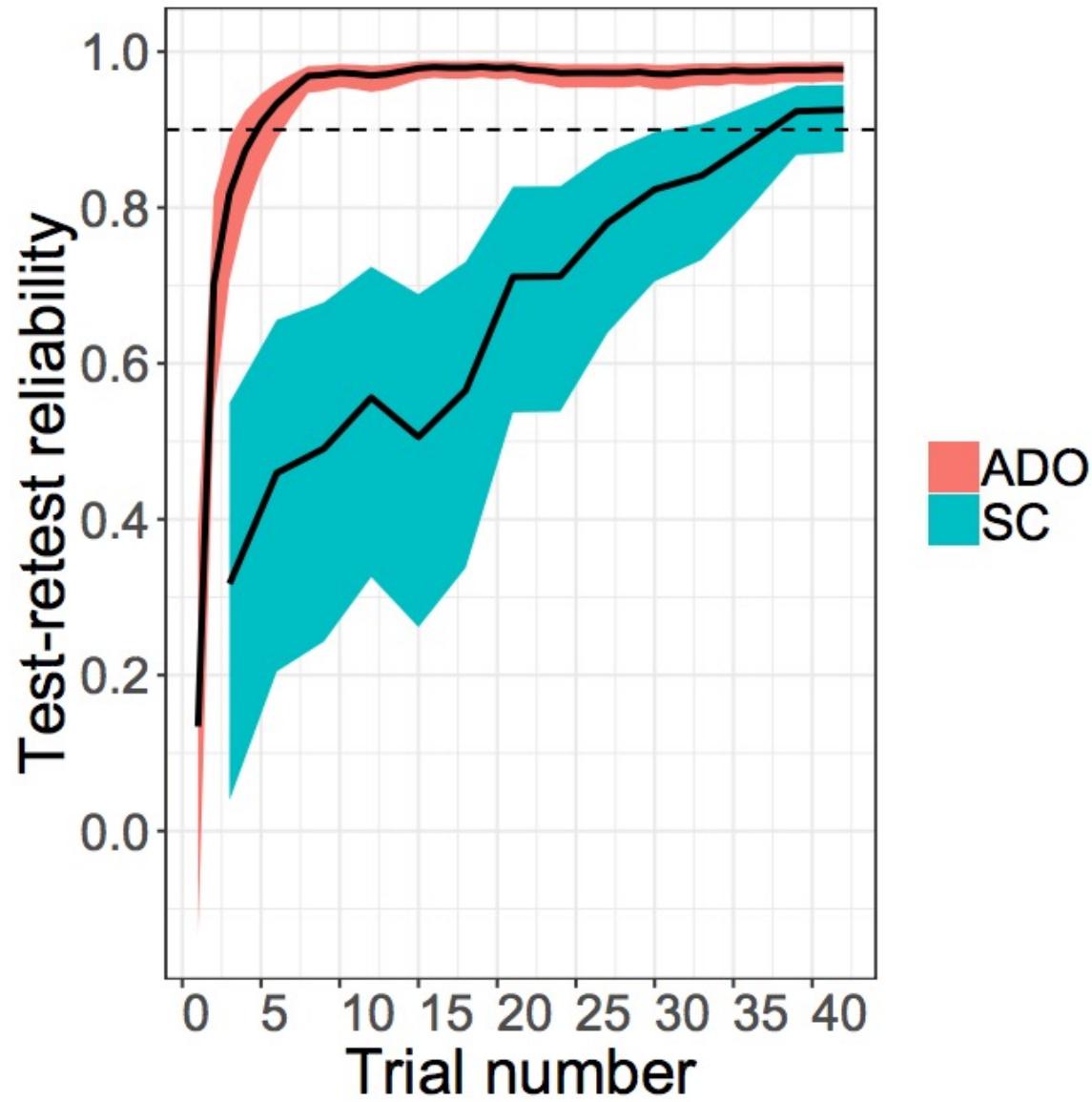
College students



Patients with SUDs



Online Amazon MTurk



3-5 times more precise

3-8 times more efficient

Lowering the barrier to ADO

Yang, Pitt, Ahn, Myung (2020) Behavior Research Methods

- *ADOpy (<https://adopy.org/>)*
- *A Python package for easily implementing ADO for any (modellable) tasks*
- *Workshop at CogSci 2019 (Montreal, Canada)*

ADOpY
Adaptive Design Optimization
for Psychological Tasks

repo status WIP build passing codecov 81% codefactor A



Jaeyeong Yang



Jay Myung



Mark Pitt

<https://github.com/adopy/adopy>

ADOpY

pypi v0.3.1

repo status Active

build passing

codecov 93%

ADOpY is a Python implementation of Adaptive Design Optimization (ADO; Myung, Cavagnaro, & Pitt, 2013), which computes optimal designs dynamically in an experiment. Its modular structure permit easy integration into existing experimentation code.

ADOpY supports Python 3.5 or above and relies on NumPy, SciPy, and Pandas.

Features

- Grid-based computation of optimal designs using only three classes: `adopy.Task`, `adopy.Model`, and `adopy.Engine`.
- Easily customizable for your own tasks and models
- Pre-implemented Task and Model classes including:
 - Psychometric function estimation for 2AFC tasks (`adopy.tasks.psi`)
 - Delay discounting task (`adopy.tasks.ddt`)
 - Choice under risk and ambiguity task (`adopy.tasks.cra`)
- Example code for experiments using PsychoPy ([link](#))

Resources

- [Getting started](#)
- [Documentation](#)
- [Bug reports](#)

<https://adopy.org/>

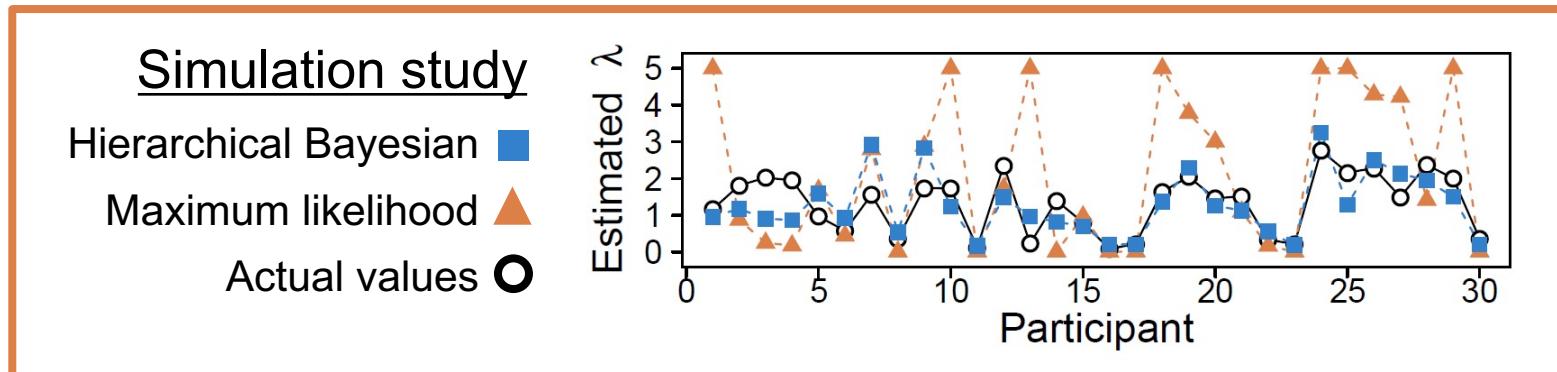
ADOpY

Adaptive Design Optimization
for Experimental Tasks

Tips for RL (in human research?)

Use a hierarchical approach across subjects when estimating model parameters

- *Human data are noisy*
- *Hierarchical approaches lead to more reliable estimates*



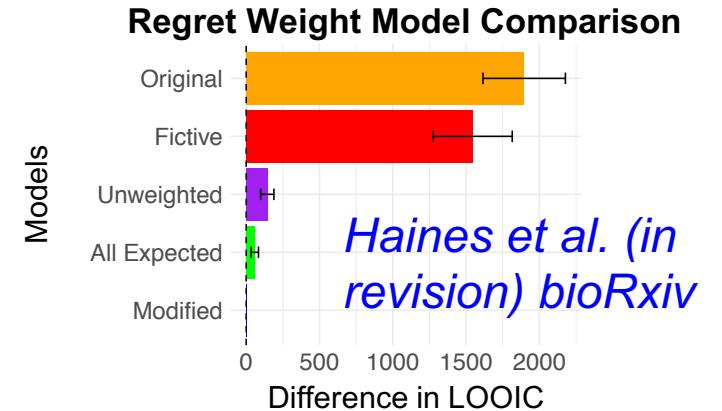
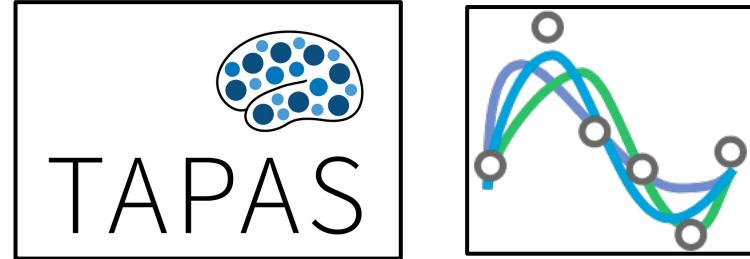
Ahn et al. (2011, JNPE)

Limitations

- *Overly simplified “toy” problems*
 - *Violated assumptions (e.g., discrete space/action & Markov..)* ,
(Gershman & Daw, 2016, Annu Rev Psych)
- *One-shot learning with sparse data*
 - *Episodic memory (hippocampus)* *(Gabrieli, 1998; Eichenbaum et al., 1999)*
- *Modeling of even toy problems is hard for many people*

Future directions

- *Incorporating multi-modal inputs into computational models* ([Haines et al., in revision, bioRxiv 2020](#))
- *Predict real-life DM?* ([Mobb et al., 2018, Nat Rev Neuro](#))
- *Lowering the barrier to RL and computational modeling*



*Ahn et al. (2017)
Computational Psychiatry*

ADOp
Adaptive Design Optimization
for Experimental Tasks

*Yang et al. (2020)
Behavior Research Methods*

Thank you!



Jay Myung



Mark Pitt



*Jaeyeong
(Jayce) Yang*



Nate Haines



Computational Clinical Science Laboratory

<https://ccs-lab.github.io>