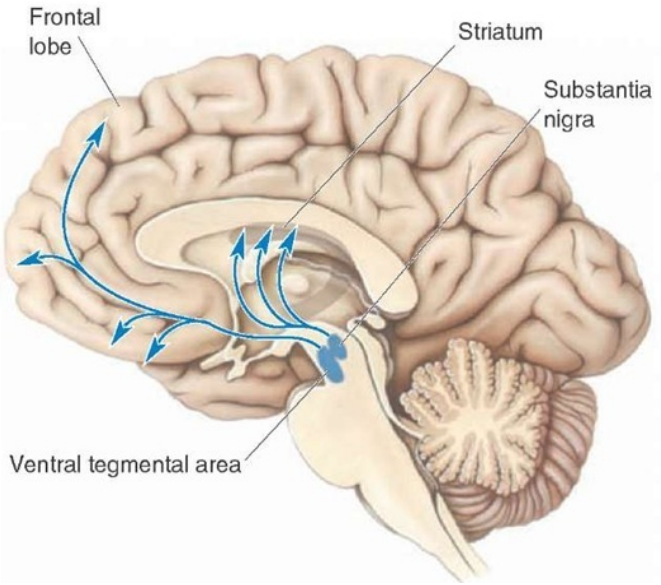
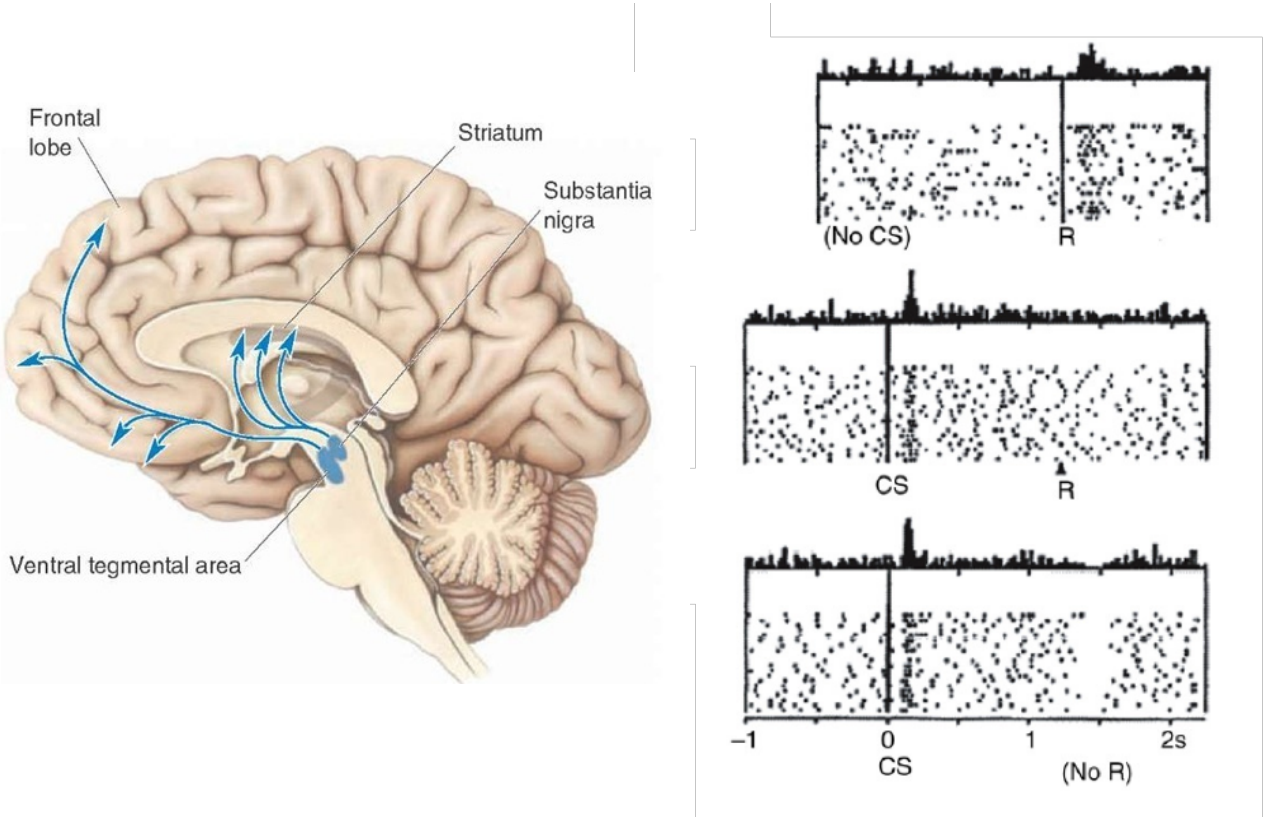
# Meta-Reinforcement Learning

Matthew Botvinick
DeepMind, London UK
Gatsby Computational Neuroscience Unit, UCL

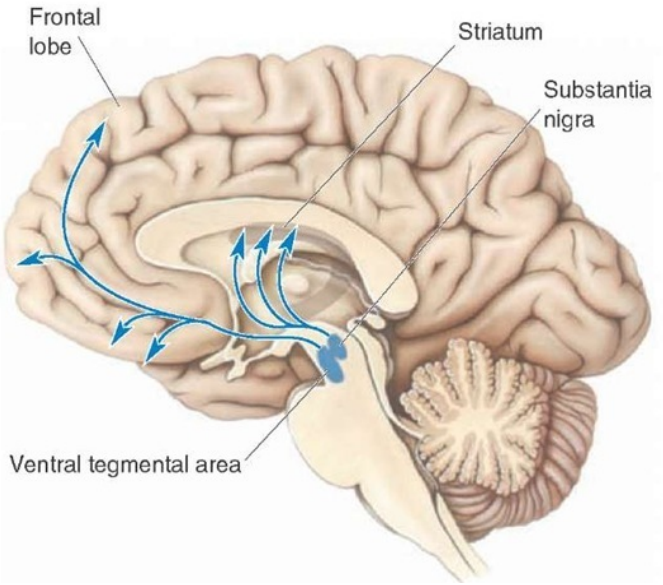
“Reinforcement learning is learning what to do — how to map situations to actions — so as to maximize a numerical reward signal. The learner is not told which actions to take...but instead must discover which actions yield the most reward by trying them.”

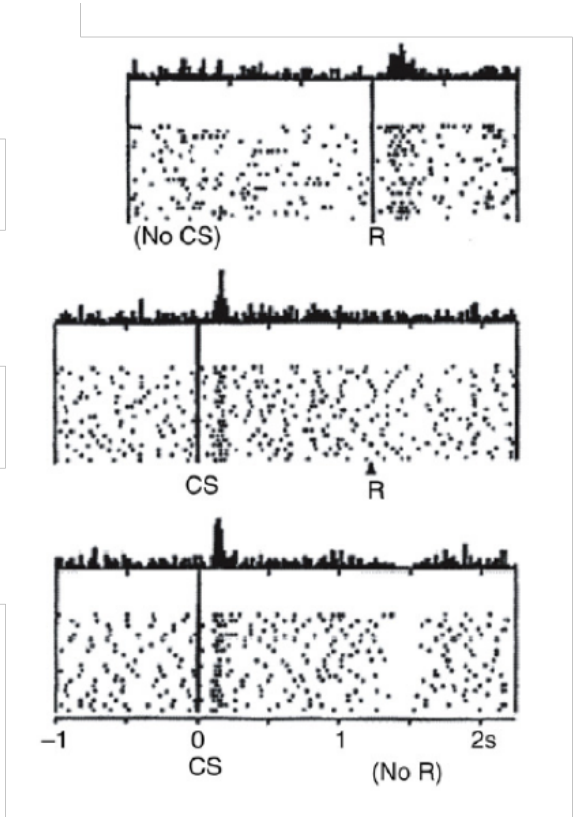
— Sutton & Barto, 1998

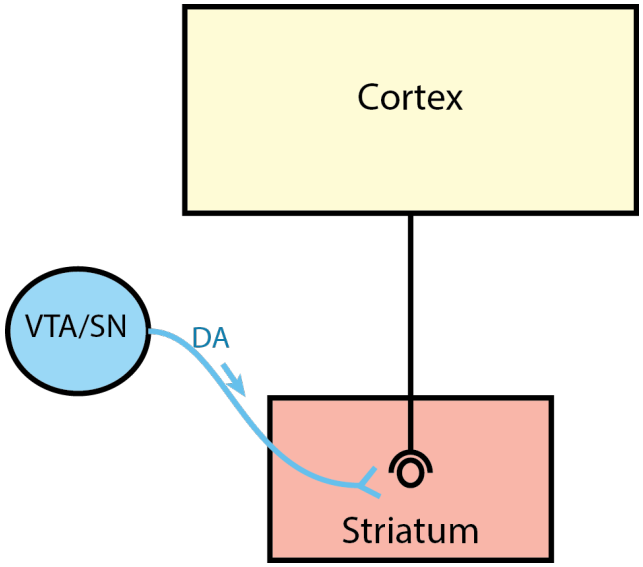

Frontal lobe
Striatum
Substantia nigra
Ventral tegmental area



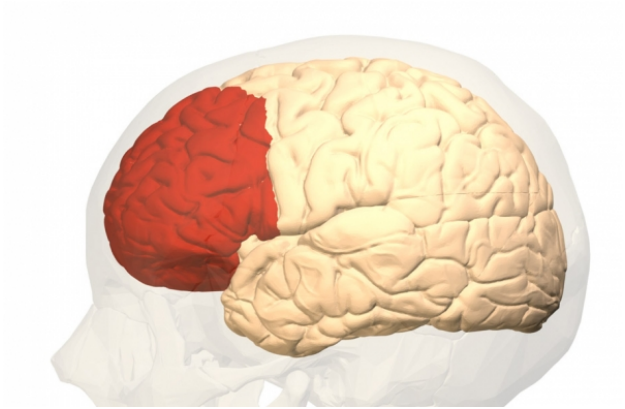

*Schultz et al, Science (1997)*

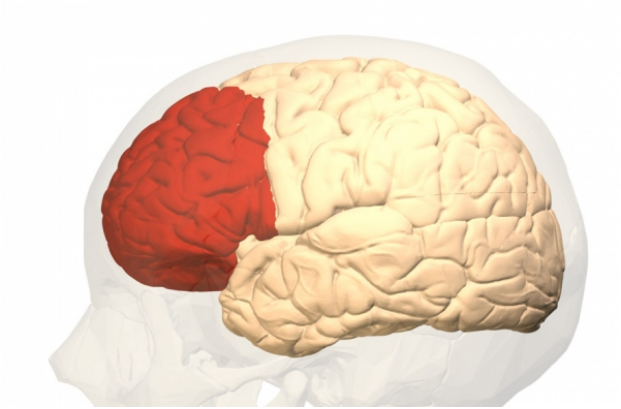


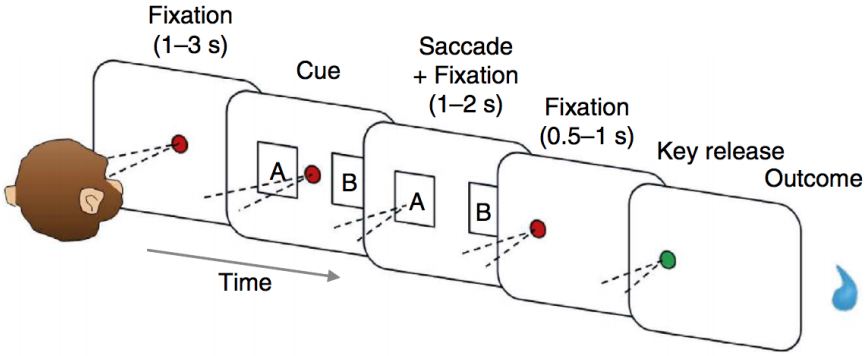




*Schultz et al, Science (1997)*

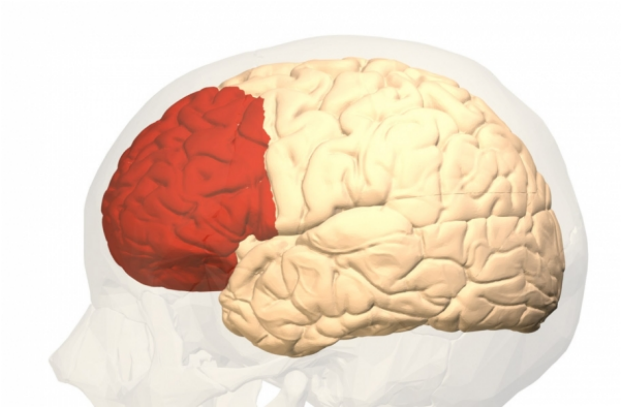


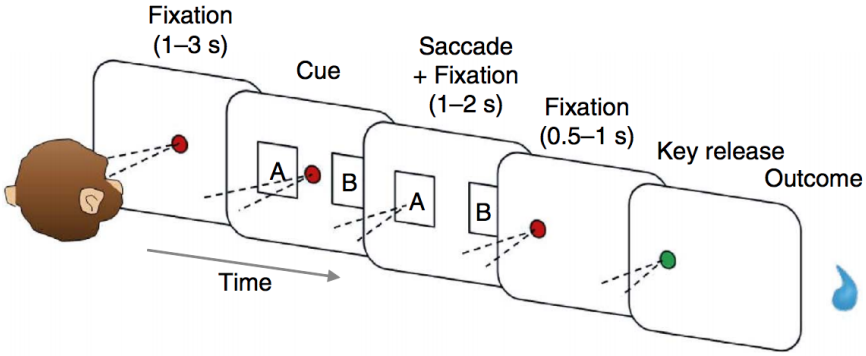


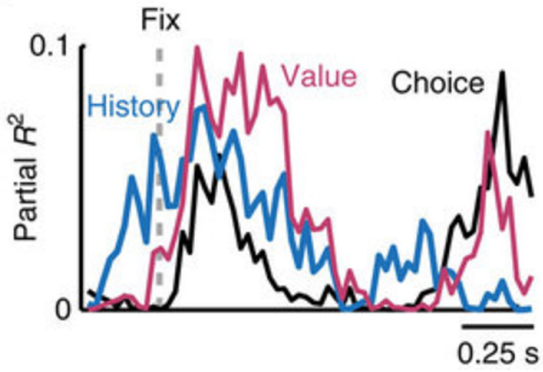



*Tsutsui et al., Nat. Comm., 2016*

*See also: Barraclough et al., Nat. Neuro. 2004; Seo & Lee, J. Neurosci 2007; Shima & Tanji, Science 1998; Matsumoto et al., Nat. Neuro. 2007*

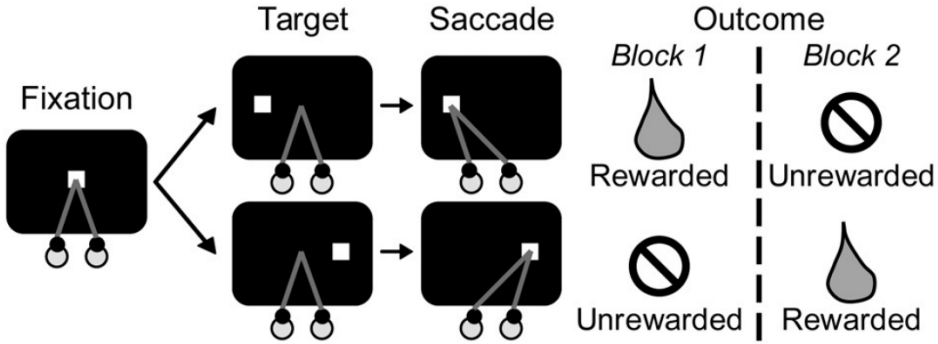




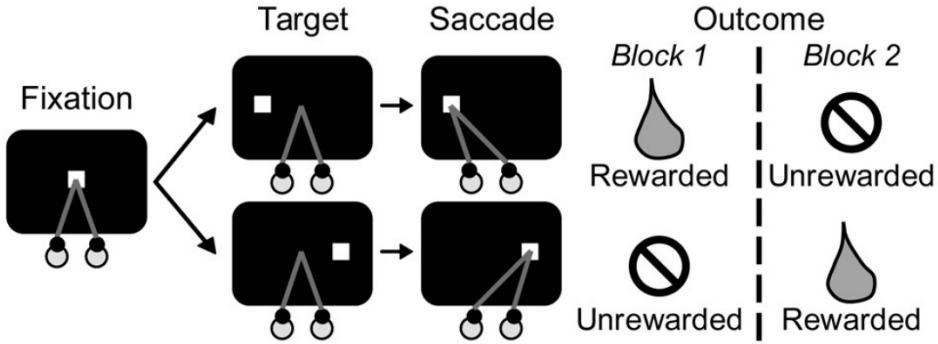




*Tsutsui et al., Nat. Comm., 2016*

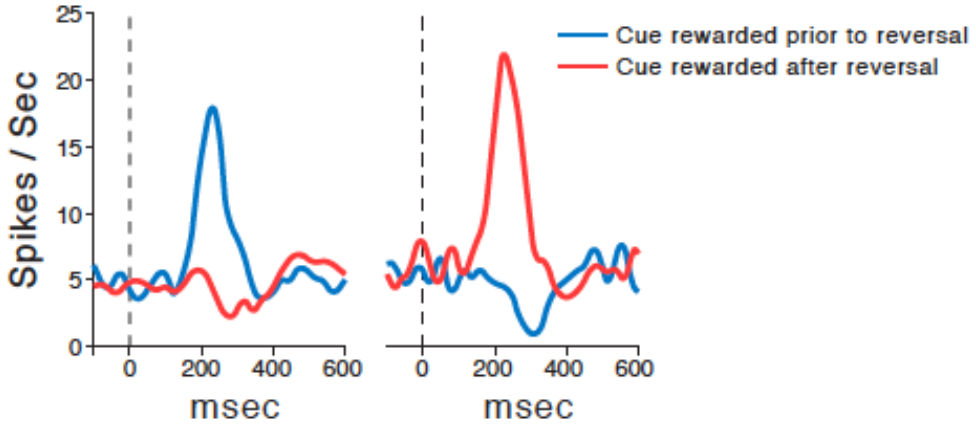
*See also: Barraclough et al., Nat. Neuro. 2004; Seo & Lee, J. Neurosci 2007; Shima & Tanji, Science 1998; Matsumoto et al., Nat. Neuro. 2007*

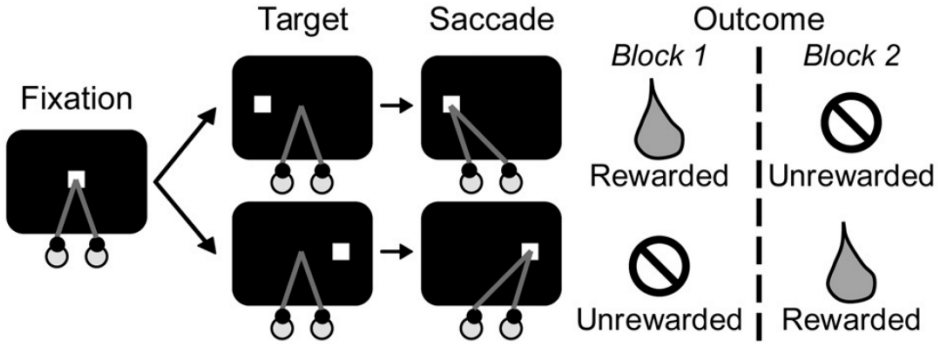



*Bromberg-Martin et al, J Neurophys, 2010*

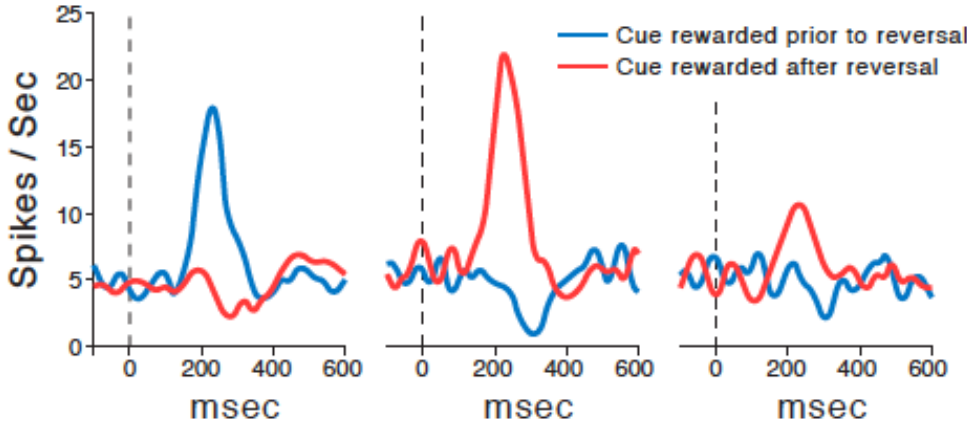


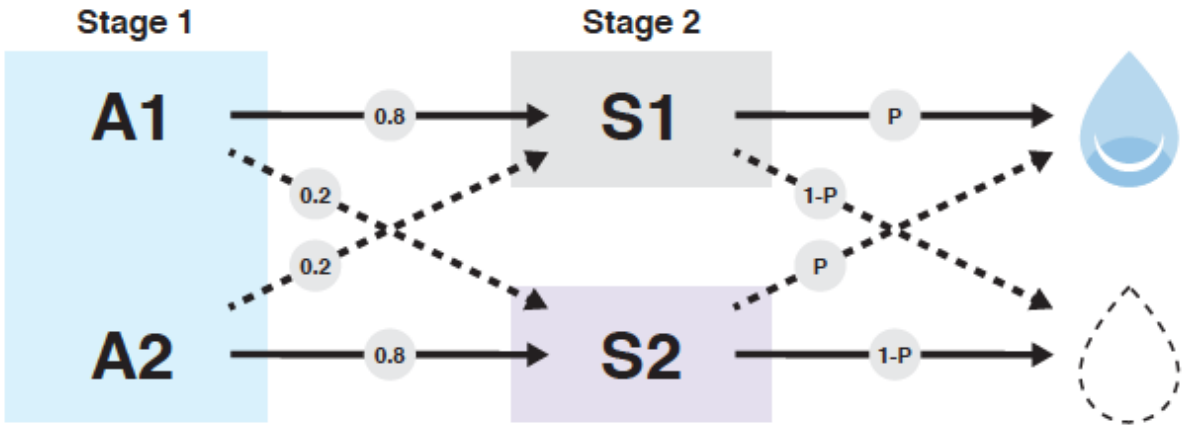

*Bromberg-Martin et al, J Neurophys, 2010*



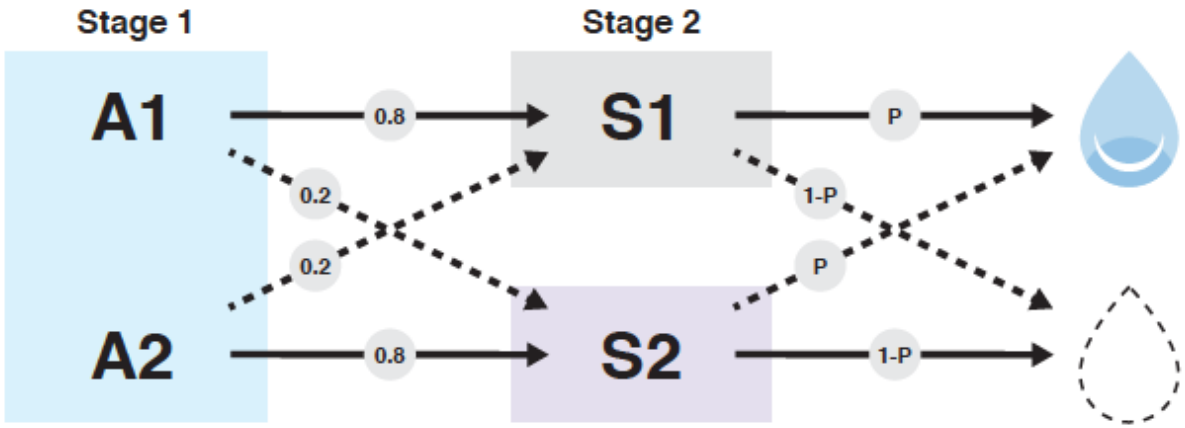



*Bromberg-Martin et al, J Neurophys, 2010*

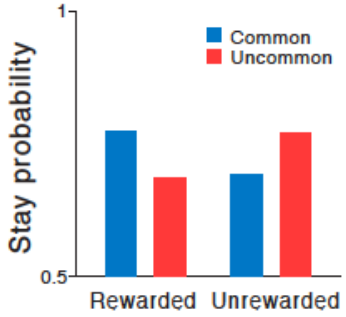




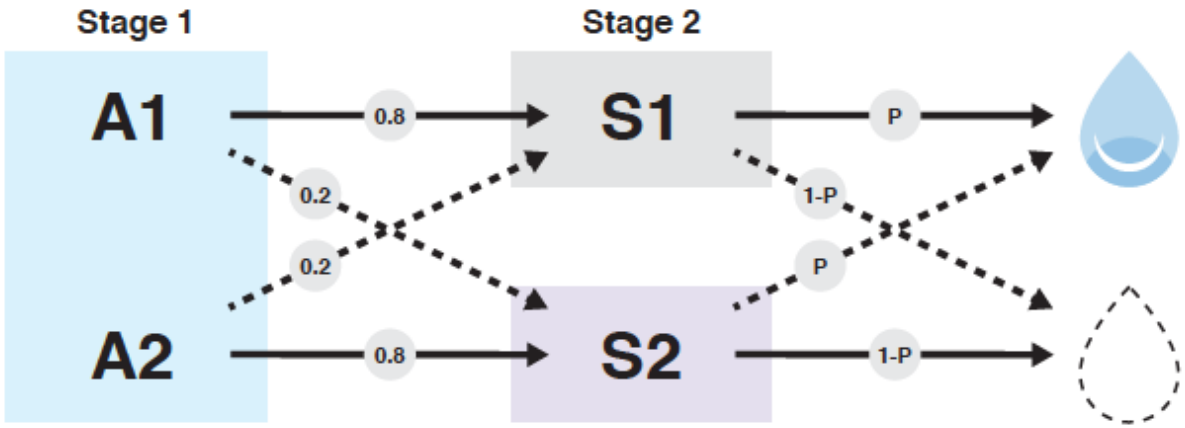

Miller, Botvinick & Brody (in press); Daw et al., *Neuron,* 2011



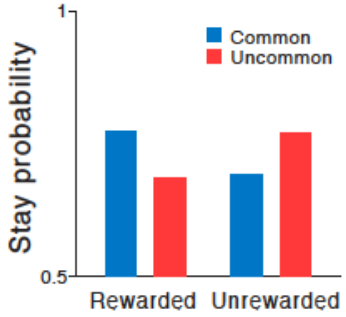


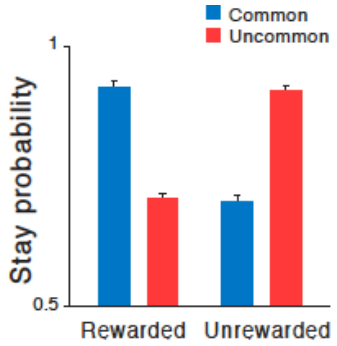



Miller, Botvinick & Brody (in press); Daw et al., *Neuron,* 2011

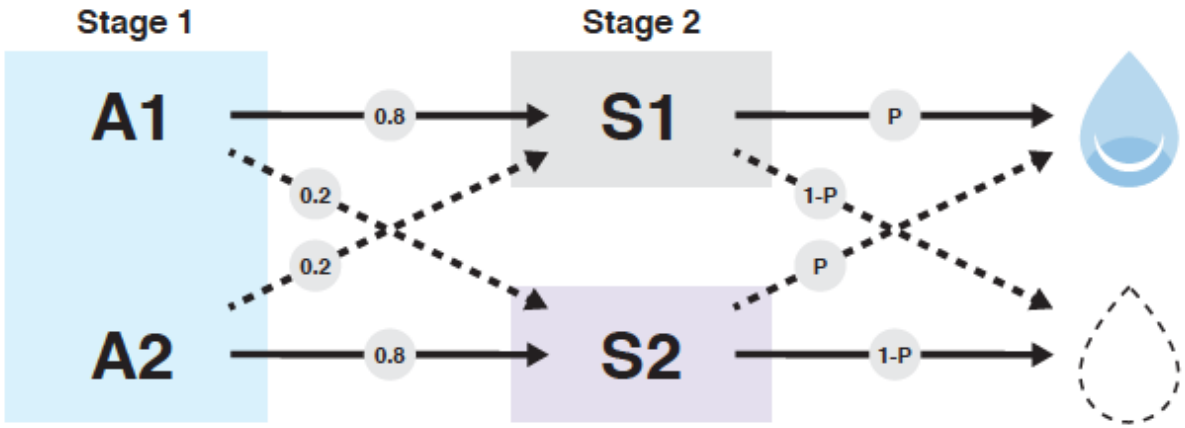


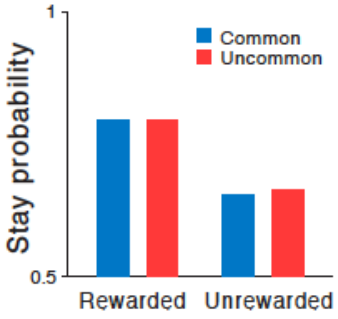


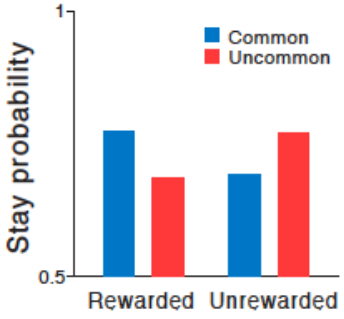


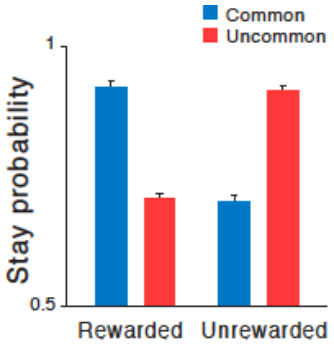



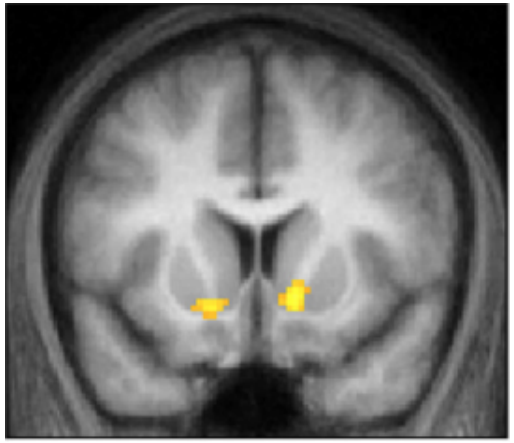

Miller, Botvinick & Brody (in press); Daw et al., *Neuron,* 2011



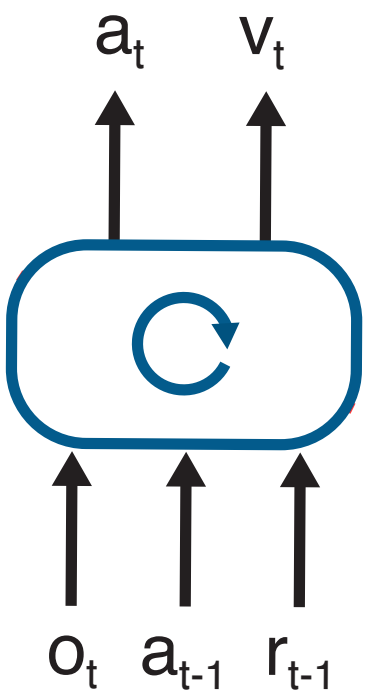
Miller, Botvinick & Brody (in press); Daw et al., *Neuron,* 2011

Key assumptions:

Key assumptions:

1. PFC circuit is a recurrent neural network

   (e.g. Mante et al, Nature 2013; O'Reilly & Frank, Neural Comput. 2006)

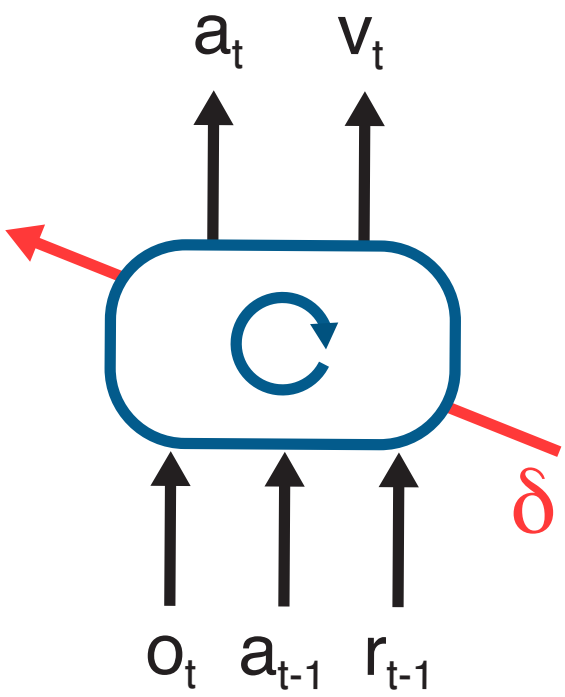



Wang *et al.,* arXiv (2016), *Cog Sci Soc* (2017)
Duan et al., arXiv (2016)

Key assumptions:

1. PFC circuit is a recurrent neural network
   (e.g. Mante et al, Nature 2013; O'Reilly & Frank, Neural Comput. 2006)

2. Synaptic weights within the PFC circuit are adjusted through model-free, dopamine-driven RL
   (e.g. O'Reilly & Frank, Neural Comput. 2006)




Wang *et al.,* arXiv (2016), *Cog Sci Soc* (2017)
Duan et al., arXiv (2016)

Key assumptions:

1. PFC circuit is a recurrent neural network
   (e.g. Mante et al, Nature 2013; O'Reilly & Frank, Neural Comput. 2006)

2. Synaptic weights within the PFC circuit are adjusted through model-free, dopamine-driven RL
   (e.g. O'Reilly & Frank, Neural Comput. 2006)

3. RL task environment is not fixed, but rather is sampled from a distribution or family
   (e.g. Rougier et al, PNAS 2005)

Emergent consequence:

Emergent consequence:

- The PFC (RNN) learns its own, autonomous RL procedure, distinct from the RL algorithm used to set the network weights (‘meta-RL’)

Emergent consequence:

- The PFC (RNN) learns its own, autonomous RL procedure, distinct from the RL algorithm used to set the network weights ('meta-RL')
  - Implemented in PFC dynamics, and can therefore execute even when synaptic weights are frozen*

* for important precedents, see Collins & Frank, 2012; O'Reilly & Frank, 2006, Nakahara & Hikosaka, 2012

Emergent consequence:

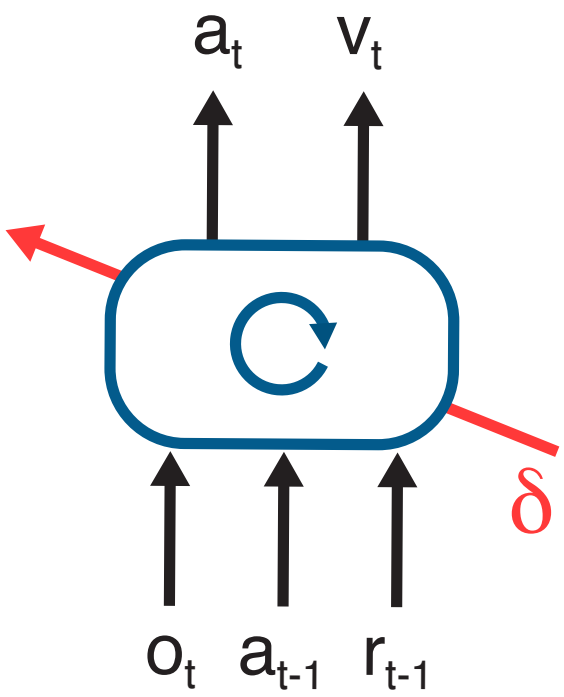
- The PFC (RNN) learns its own, autonomous RL procedure, distinct from the RL algorithm used to set the network weights ('meta-RL')
  - Implemented in PFC dynamics, and can therefore execute even when synaptic weights are frozen*
  - Differs arbitrarily from the primary RL algorithm (different hyperparameters, model-based profile, etc.)

* for important precedents, see Collins & Frank, 2012; O'Reilly & Frank, 2006, Nakahara & Hikosaka, 2012

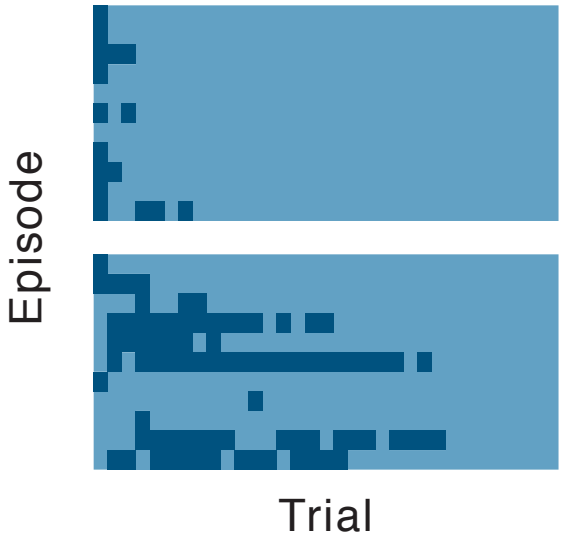
Emergent consequence:

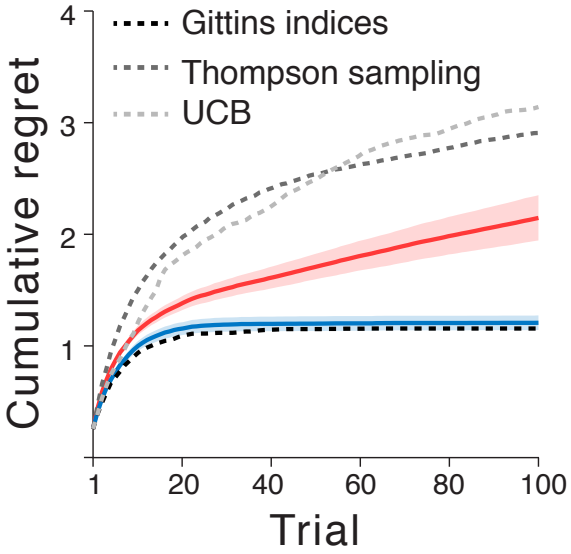
- The PFC (RNN) learns its own, autonomous RL procedure, distinct from the RL algorithm used to set the network weights ('meta-RL')
  - Implemented in PFC dynamics, and can therefore execute even when synaptic weights are frozen*
  - Differs arbitrarily from the primary RL algorithm (different hyperparameters, model-based profile, etc.)
  - Sculpted by the task environment, therefore exploits consistent task structure to learn faster

* for important precedents, see Collins & Frank, 2012; O'Reilly & Frank, 2006, Nakahara & Hikosaka, 2012

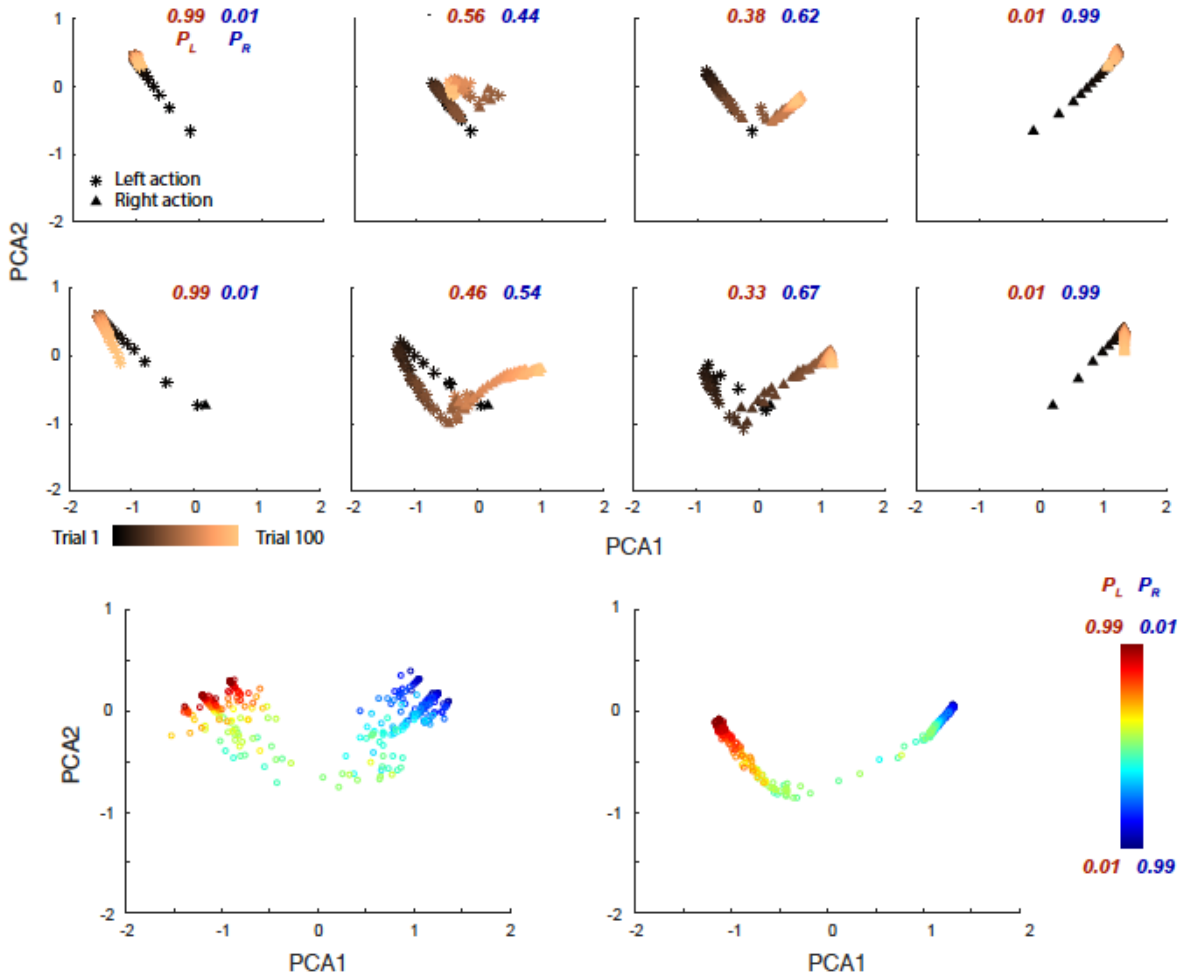



Wang *et al.,* arXiv (2016), *Cog Sci Soc* (2017)
Duan et al., arXiv (2016)

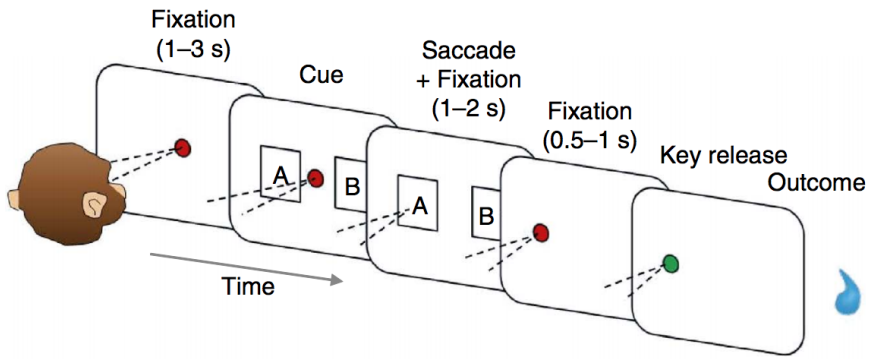


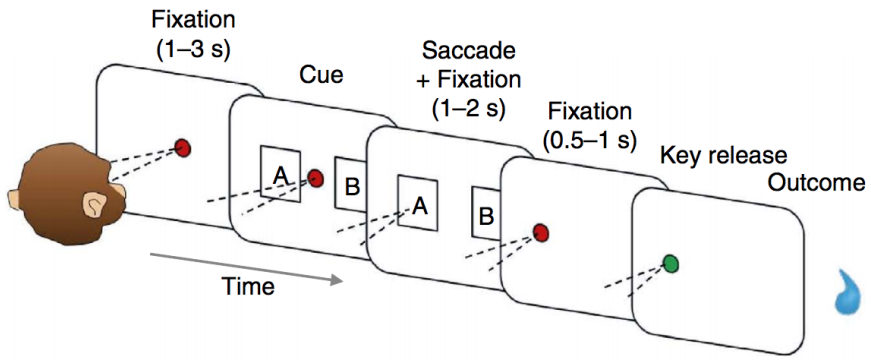

Wang *et al.,* arXiv (2016), *Cog Sci Soc* (2017)
Duan et al., arXiv (2016)

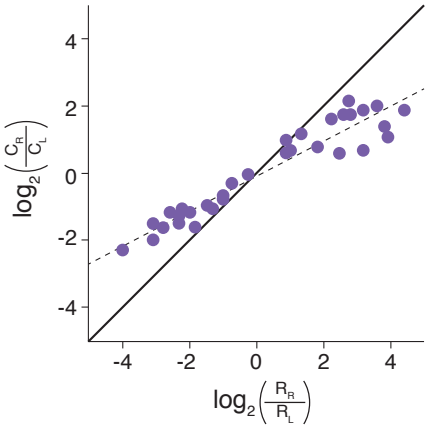


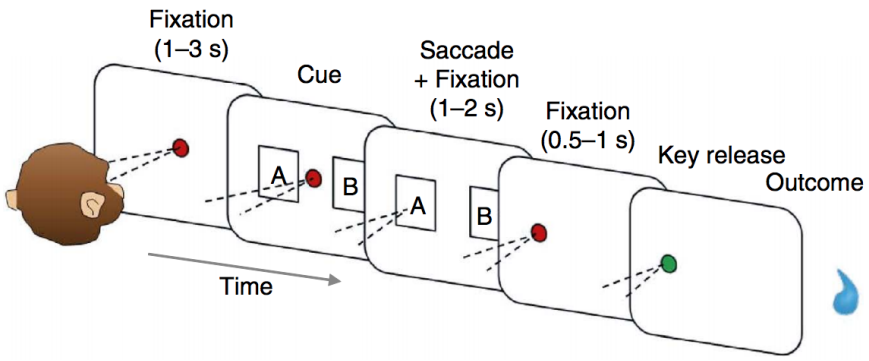

Wang et al., under review.

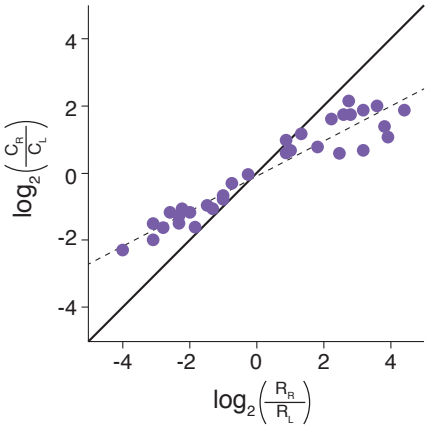


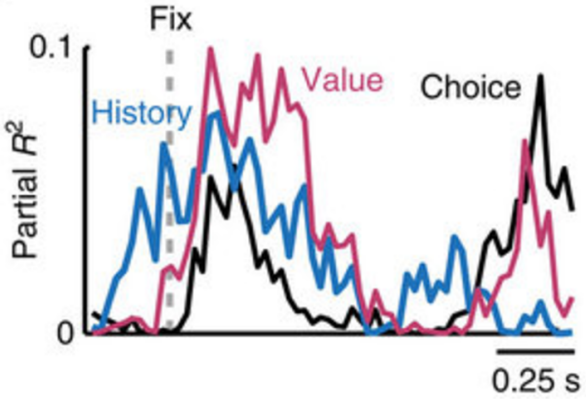

Wang *et al.* (under review)

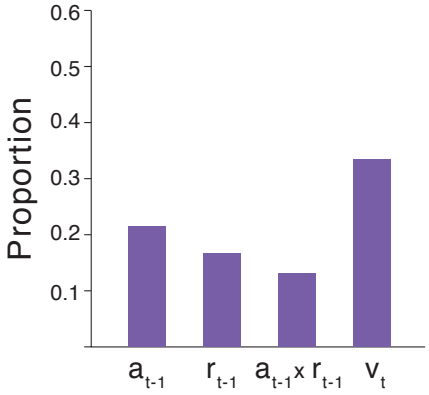

Fixation
(1–3 s)
Cue
Saccade
+ Fixation
(1–2 s)
Fixation
(0.5–1 s)
Key release
Outcome
A
B
A
B
Time


4
2
0
-2
-4
-4
-2
0
2
4
$\log_2\left(\frac{C_R}{C_L}\right)$
$\log_2\left(\frac{R_R}{R_L}\right)$

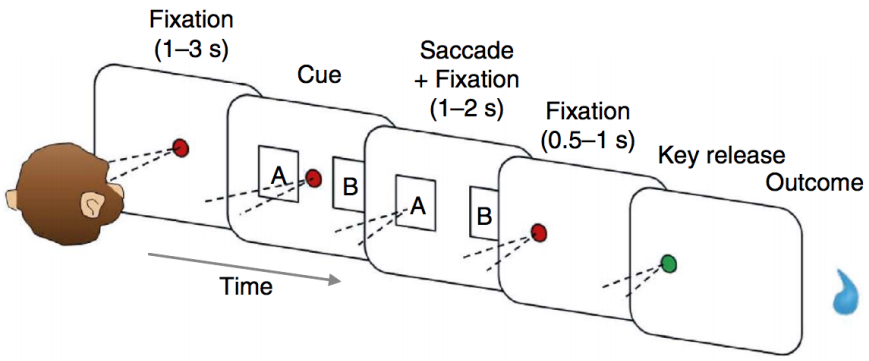


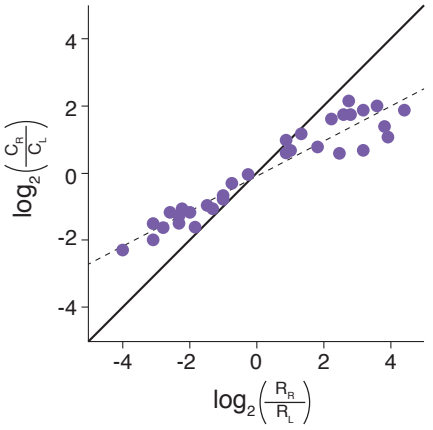



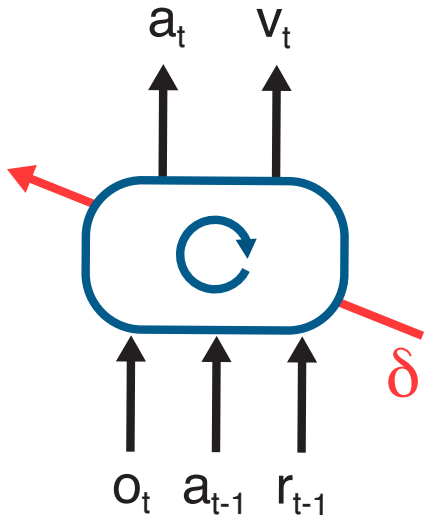



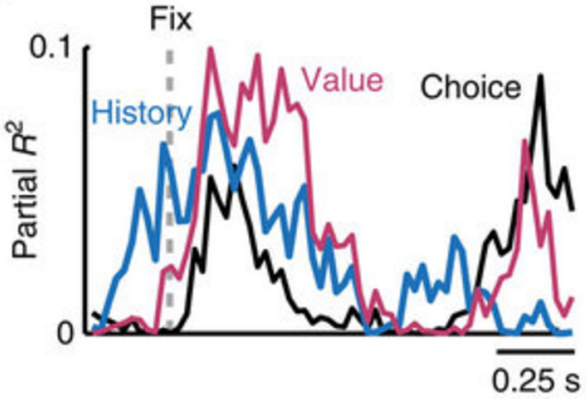



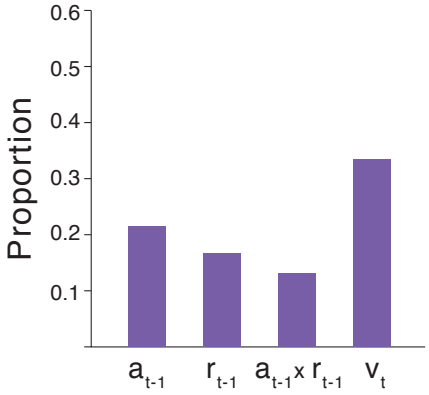

Wang *et al.* (under review)



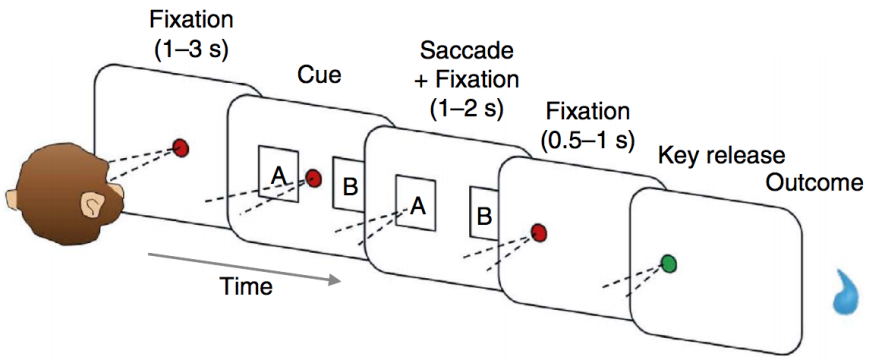



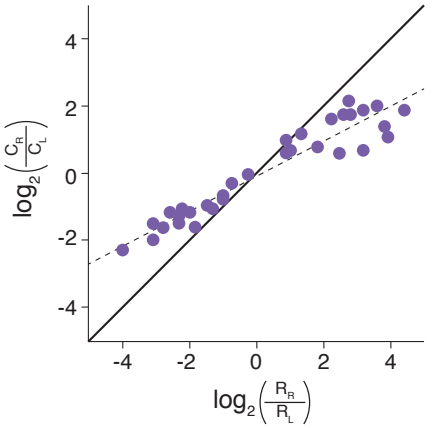



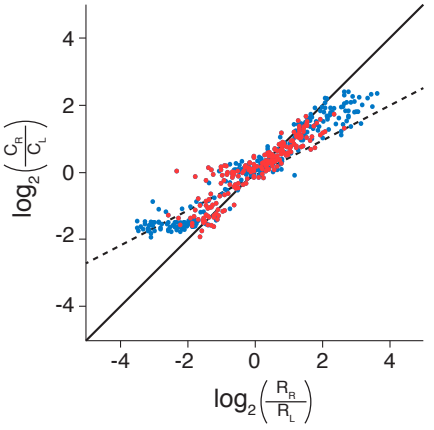



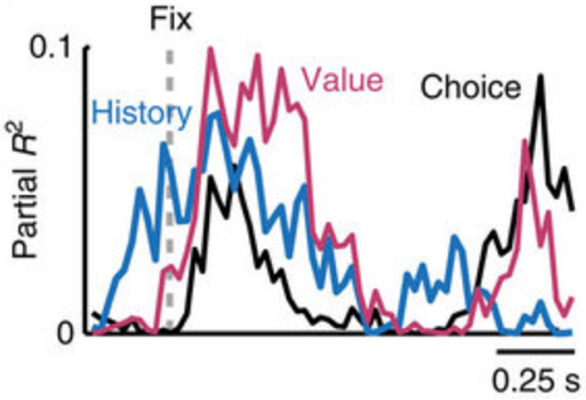



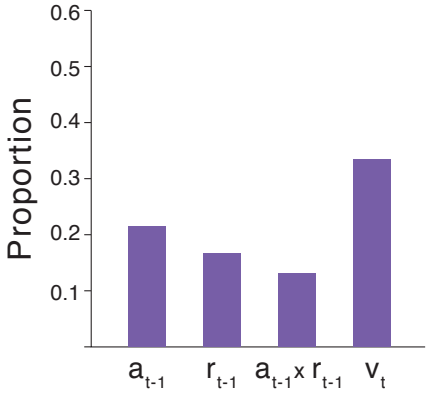

Wang *et al.* (under review)

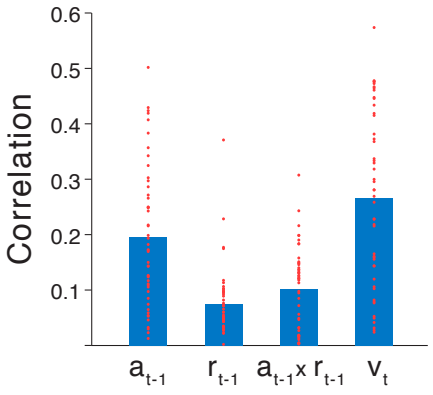


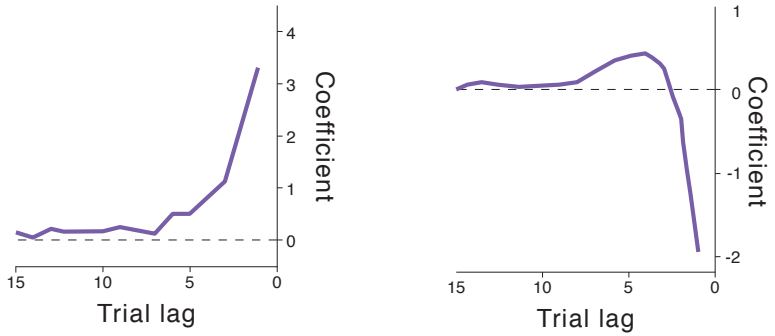



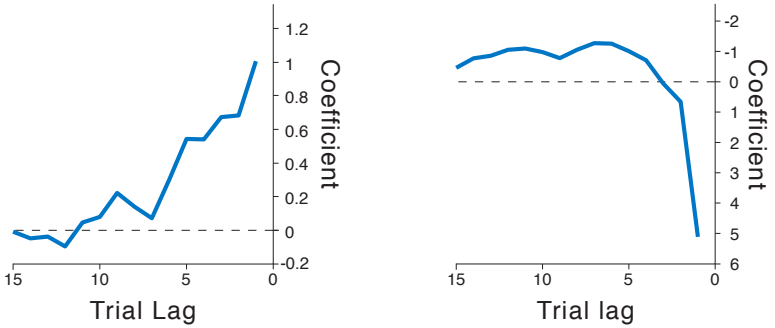



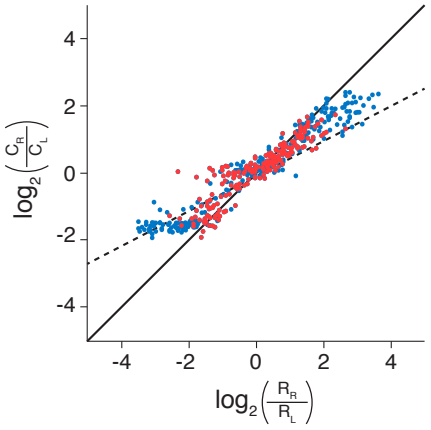

Wang *et al.* (under review)

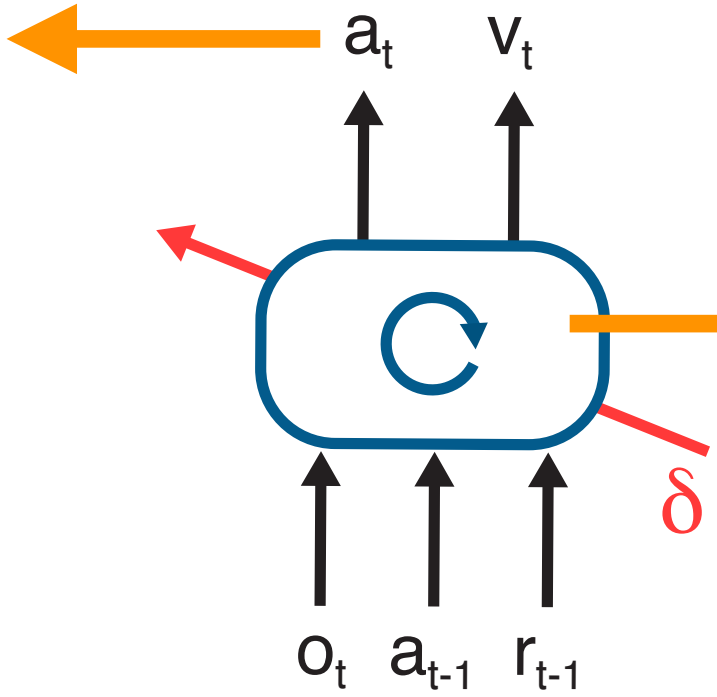


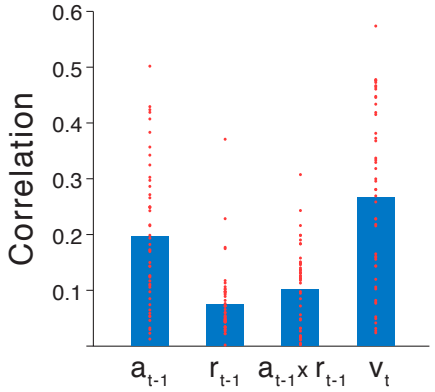

*Lau and Glimcher, J. Exp. Analysis Behavior 2005*



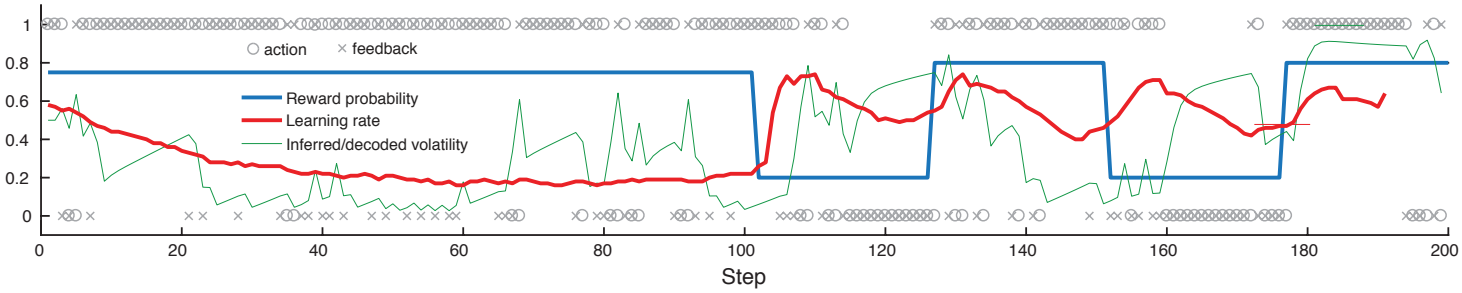

Wang *et al.* (under review)

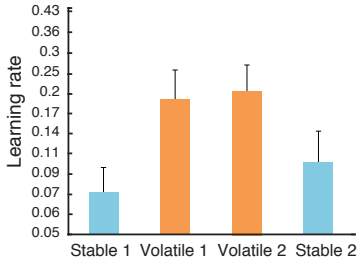


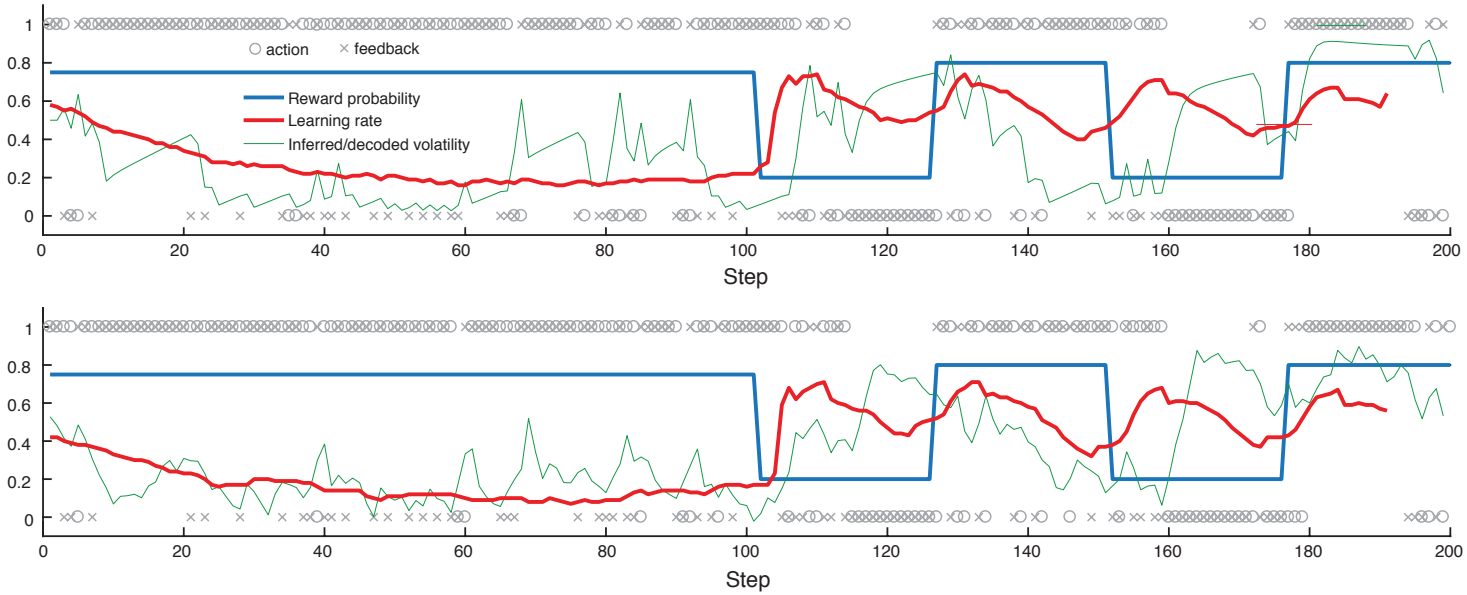



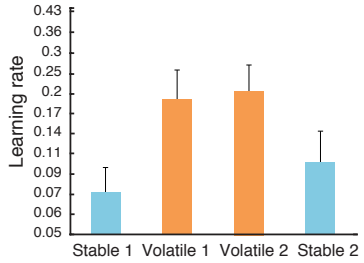

Wang *et al.,* arXiv (2016), *Cog Sci Soc* (2017)
Duan et al., arXiv (2016)



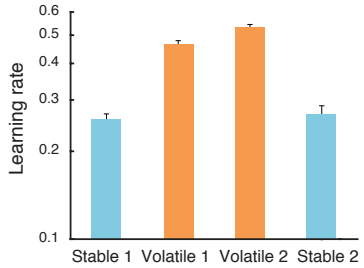

*Behrens et al., Nature Neuroscience, 2007*



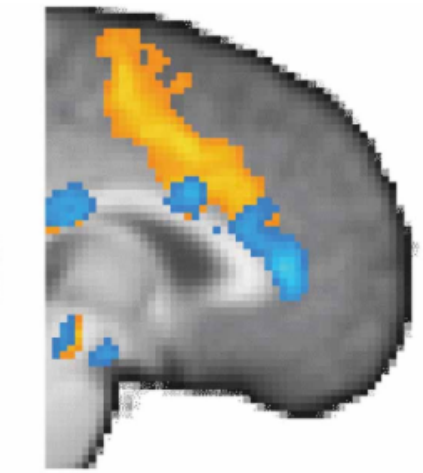



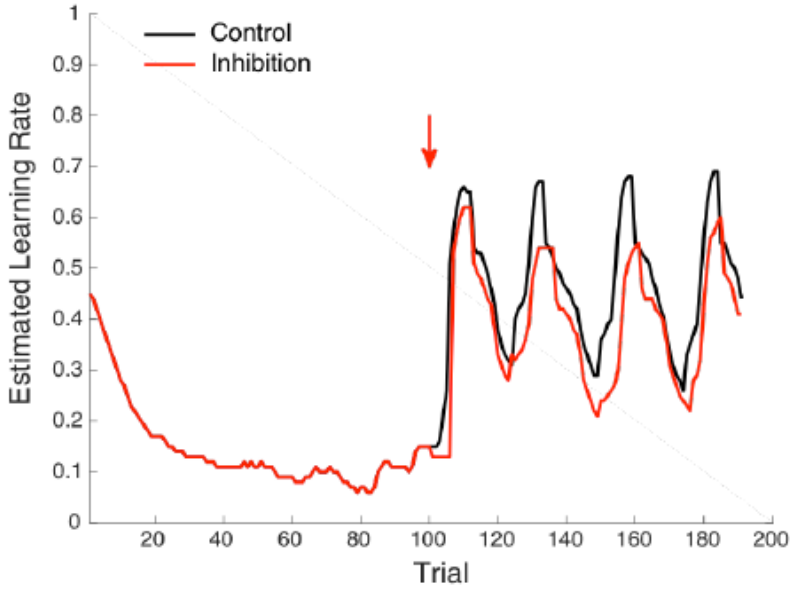






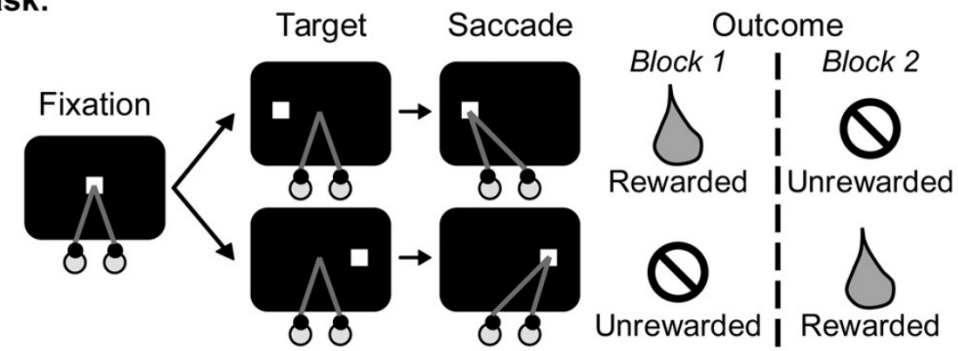

*Behrens et al., Nature Neuroscience, 2007*
Wang *et al.* (under review)



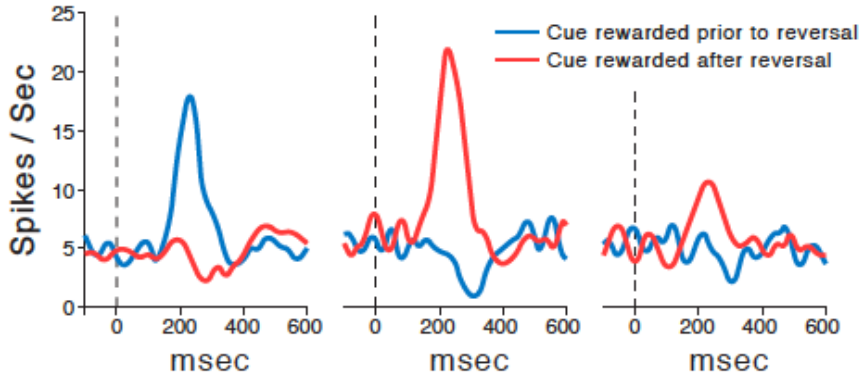



*Behrens et al., Nature Neuroscience, 2007*
Wang *et al.* (under review)

**Task:**

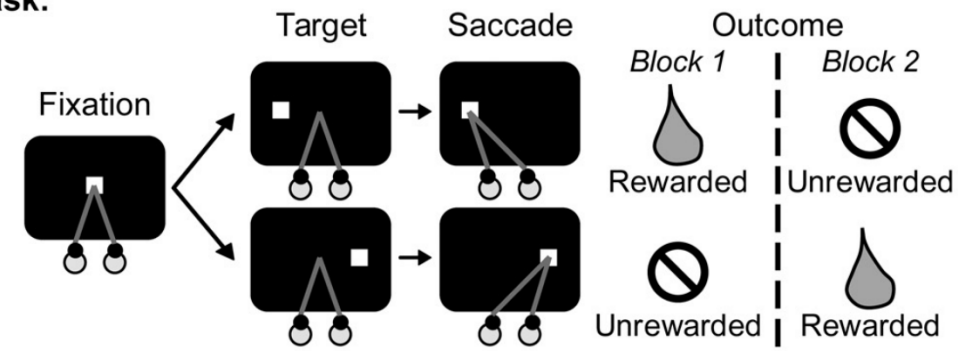



*Bromberg-Martin et al, J Neurophys, 2010*

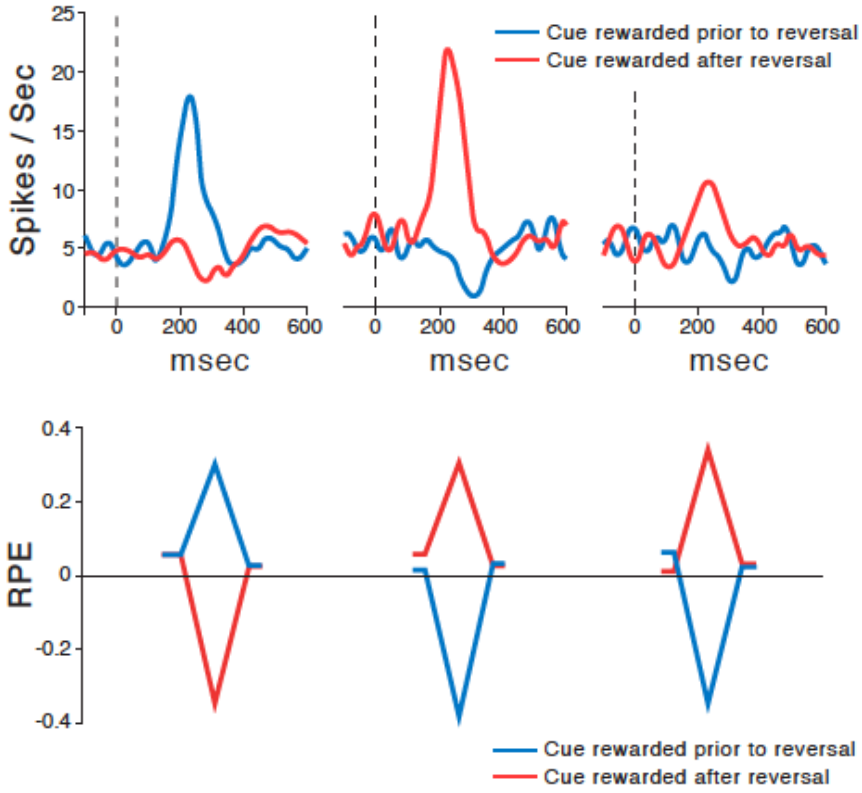



Wang *et al.* (under review)

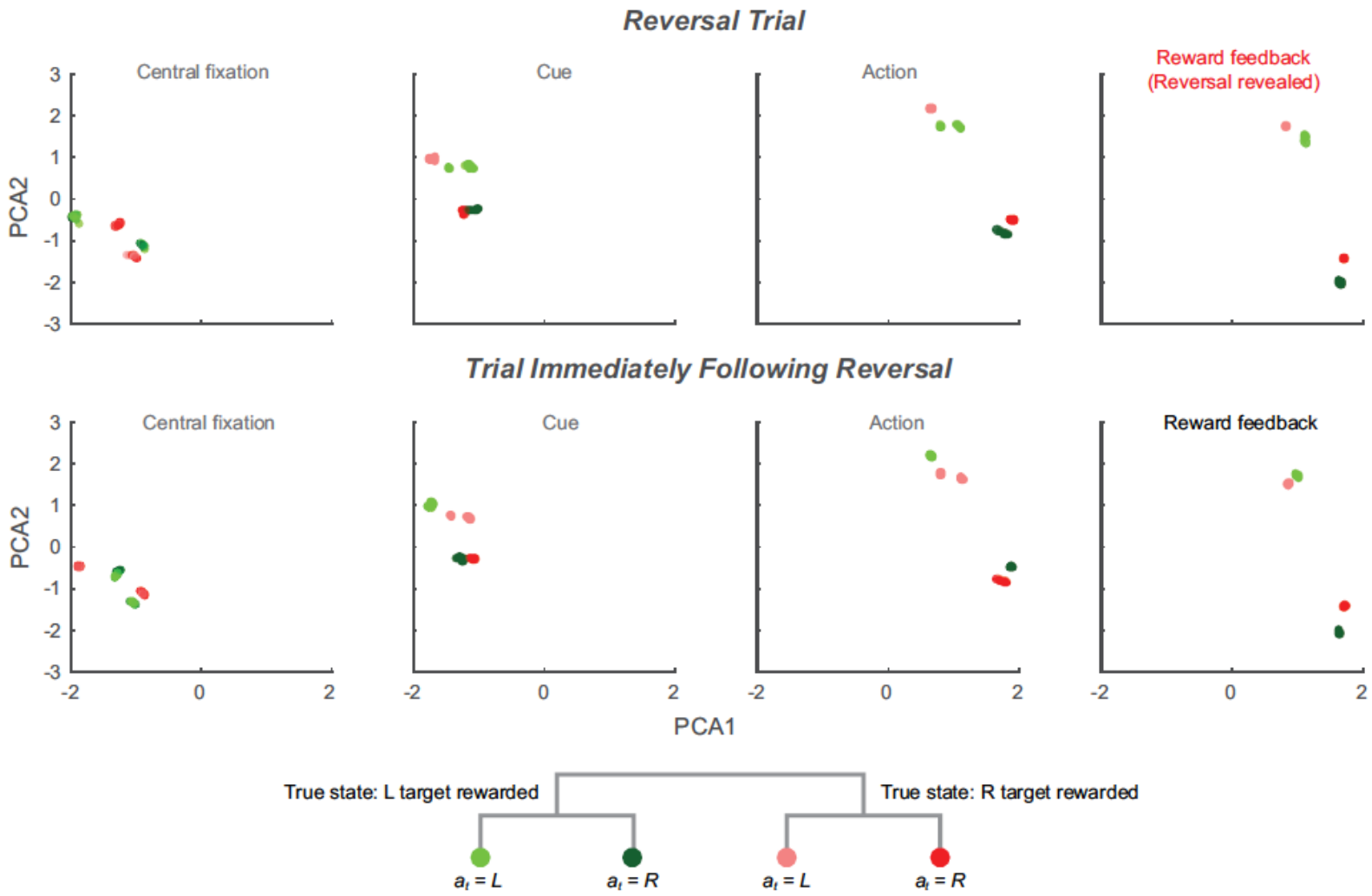
**Task:**



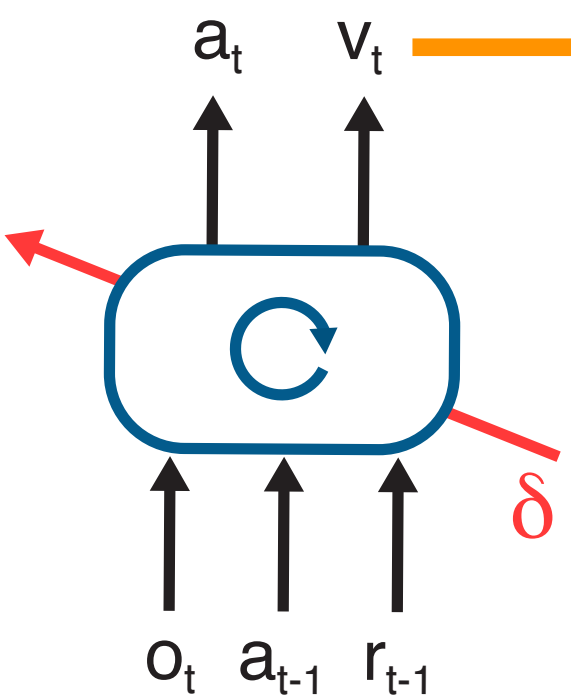

*Bromberg-Martin et al, J Neurophys, 2010*

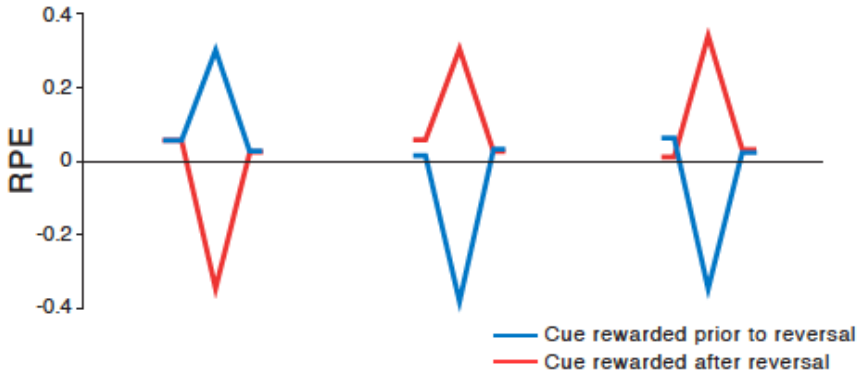



Wang *et al.* (under review)



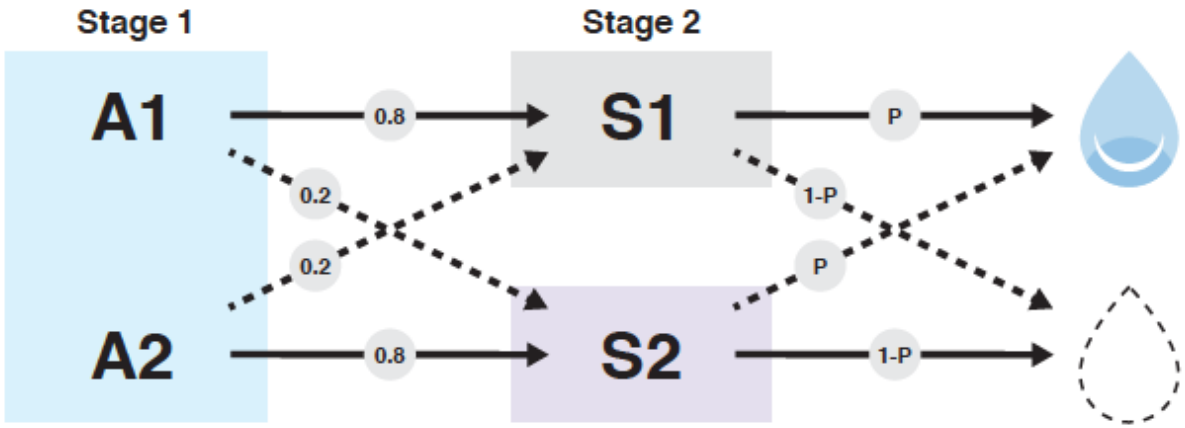

Wang et al., under review.



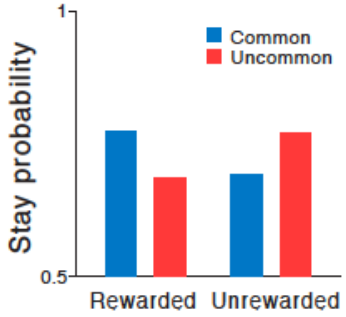



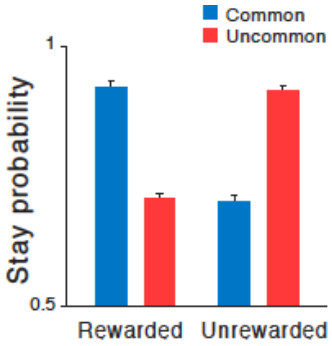

$$\delta = r_{t+1} + \gamma V_{t+1} - V_t$$

Wang *et al.*, arXiv (2016), *Cog Sci Soc* (2017)

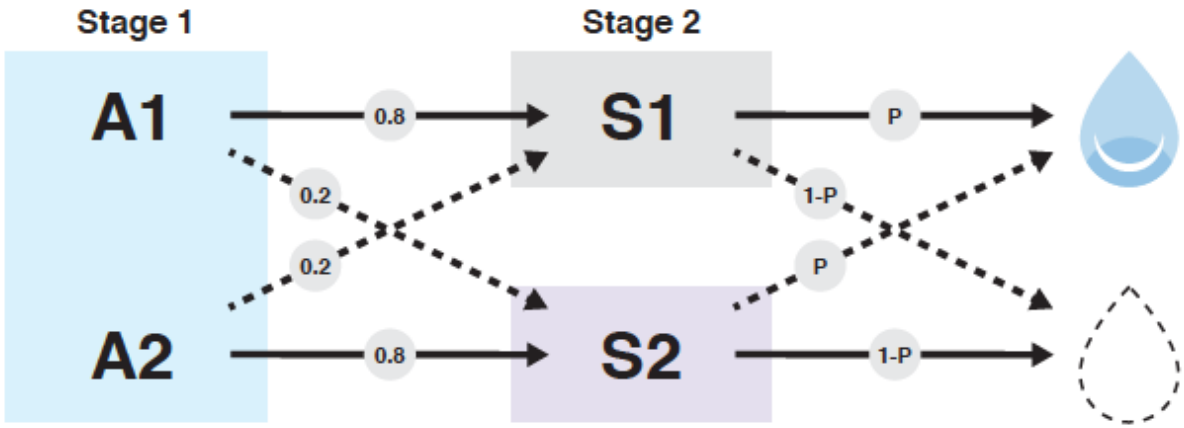


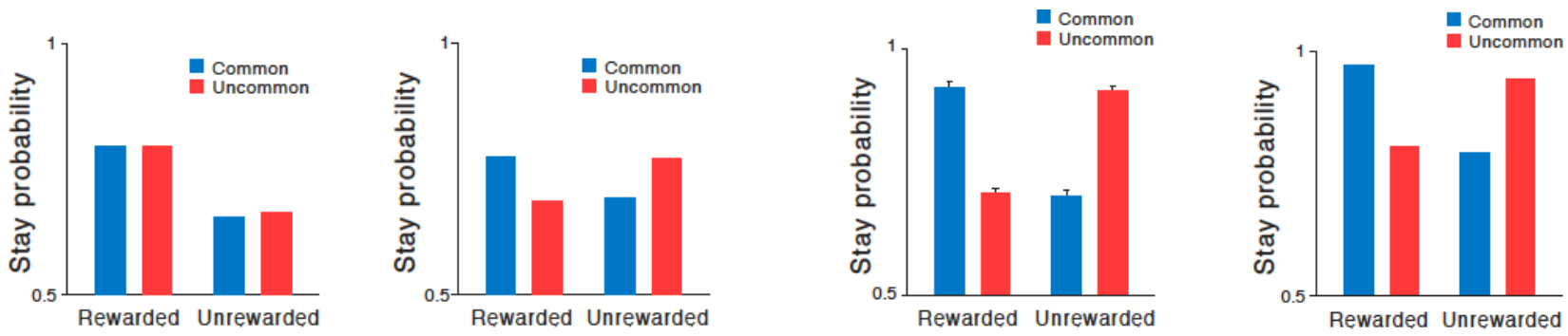




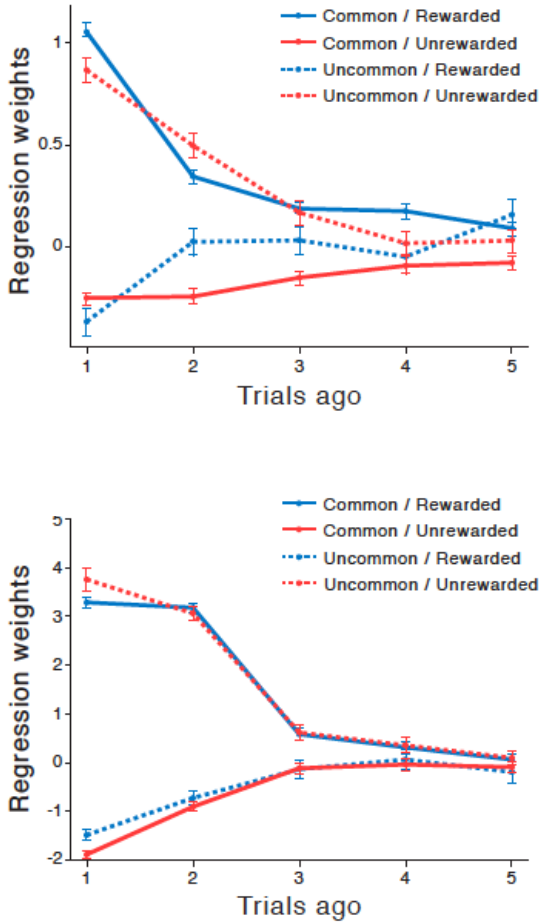




Miller, Botvinick & Brody (in press); Daw et al., *Neuron,* 2011

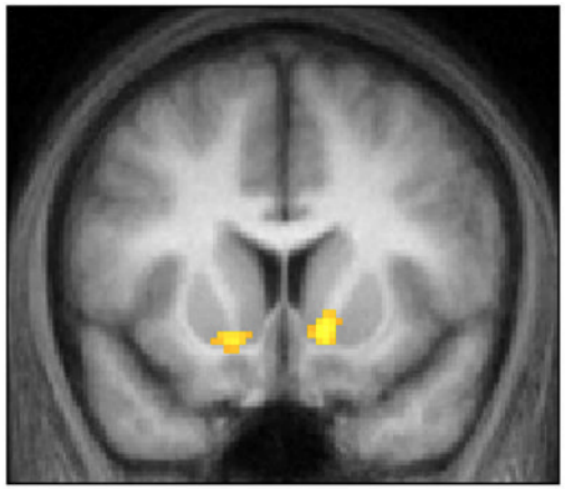


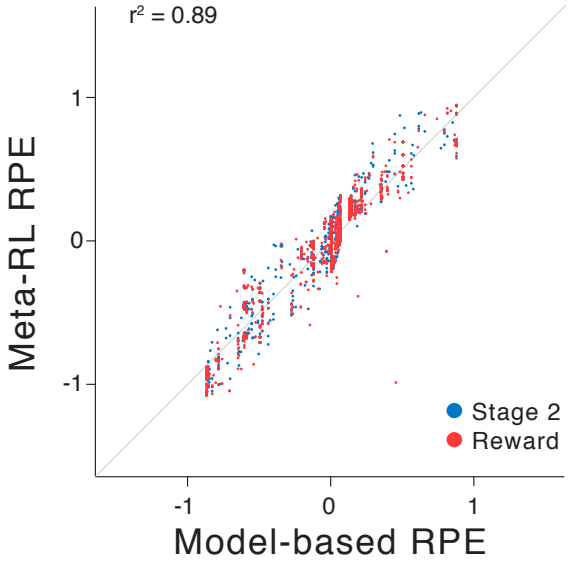




Miller, Botvinick & Brody (in press); Daw et al., *Neuron,* 2011

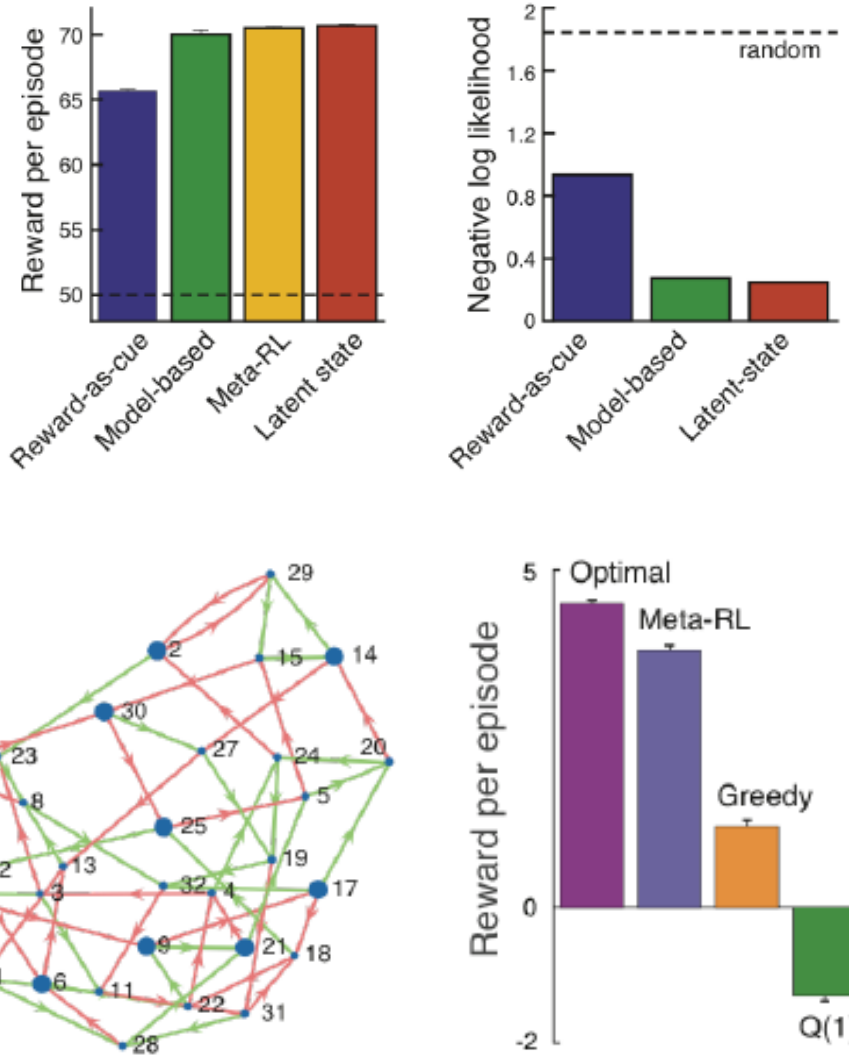



Miller, Botvinick & Brody (in press); Daw et al., *Neuron,* 2011

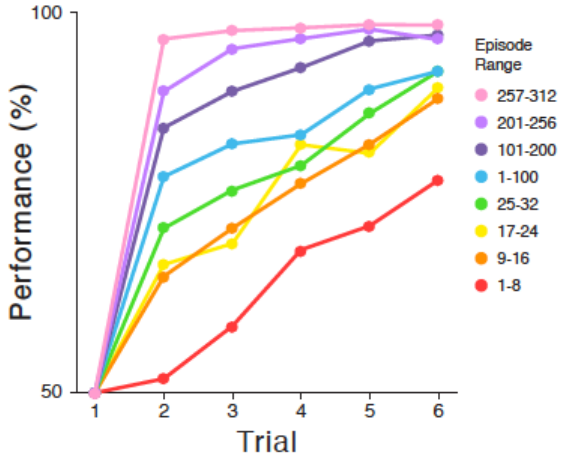





Daw et al., *Neuron,* 2011; Wang et al. (under review).

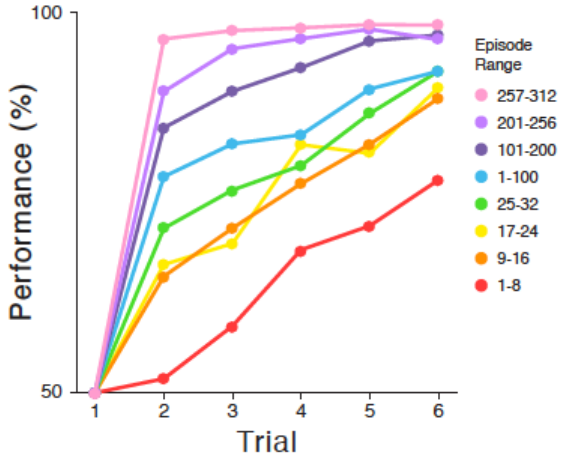



Wang et al., under review.



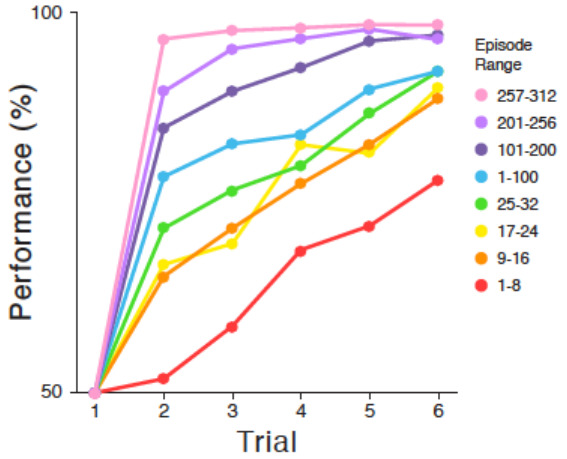

Training episodes

*Harlow, Psychological Review, 1949*

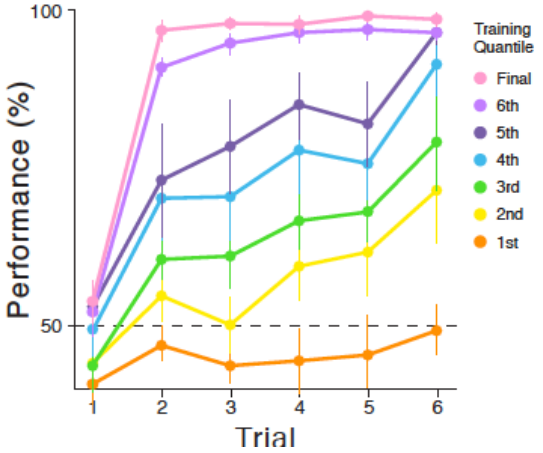



Training episodes

Wang *et al., arXiv;* under review




Training episodes

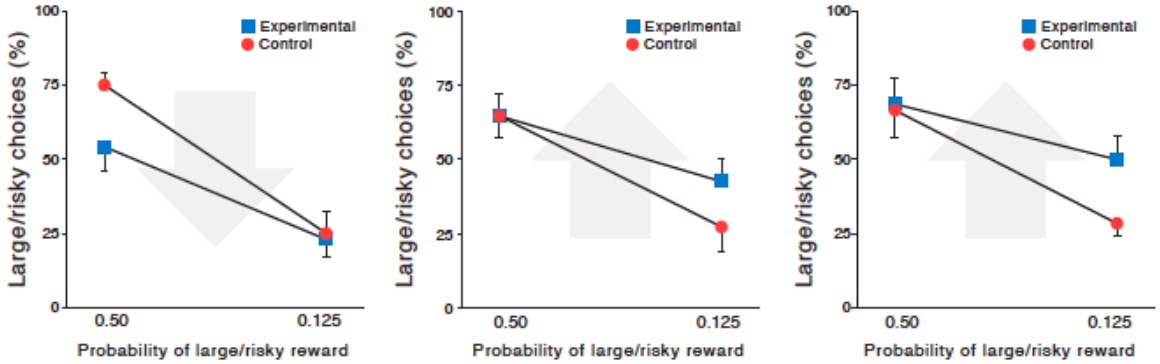



Wang *et al., arXiv;* under review

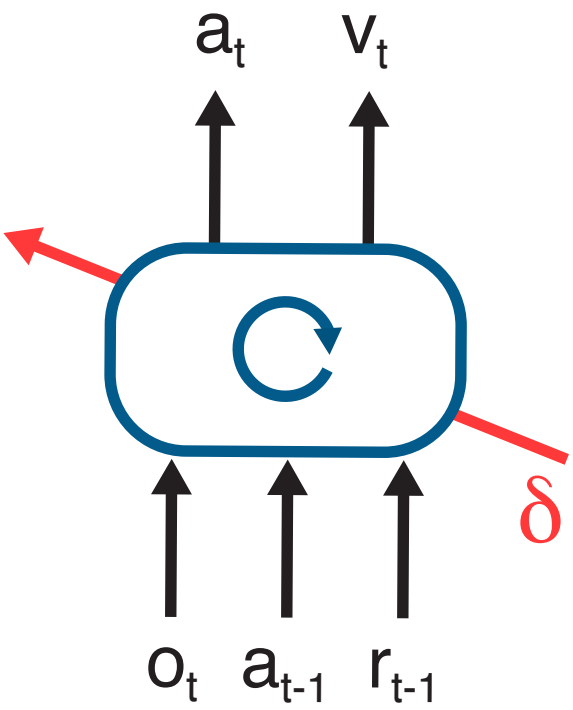
*DA blocked upon food reward from large/risky option*

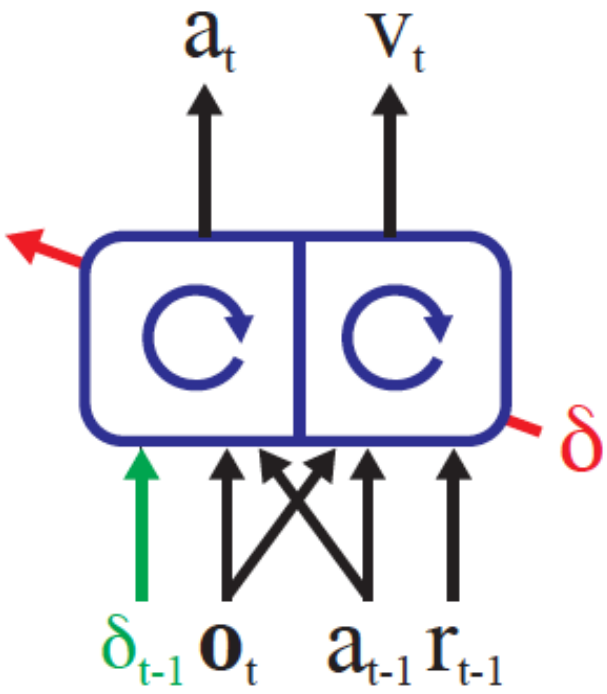
*DA blocked upon food reward from small/certain option*

*DA triggered upon food omission from large/risky option*

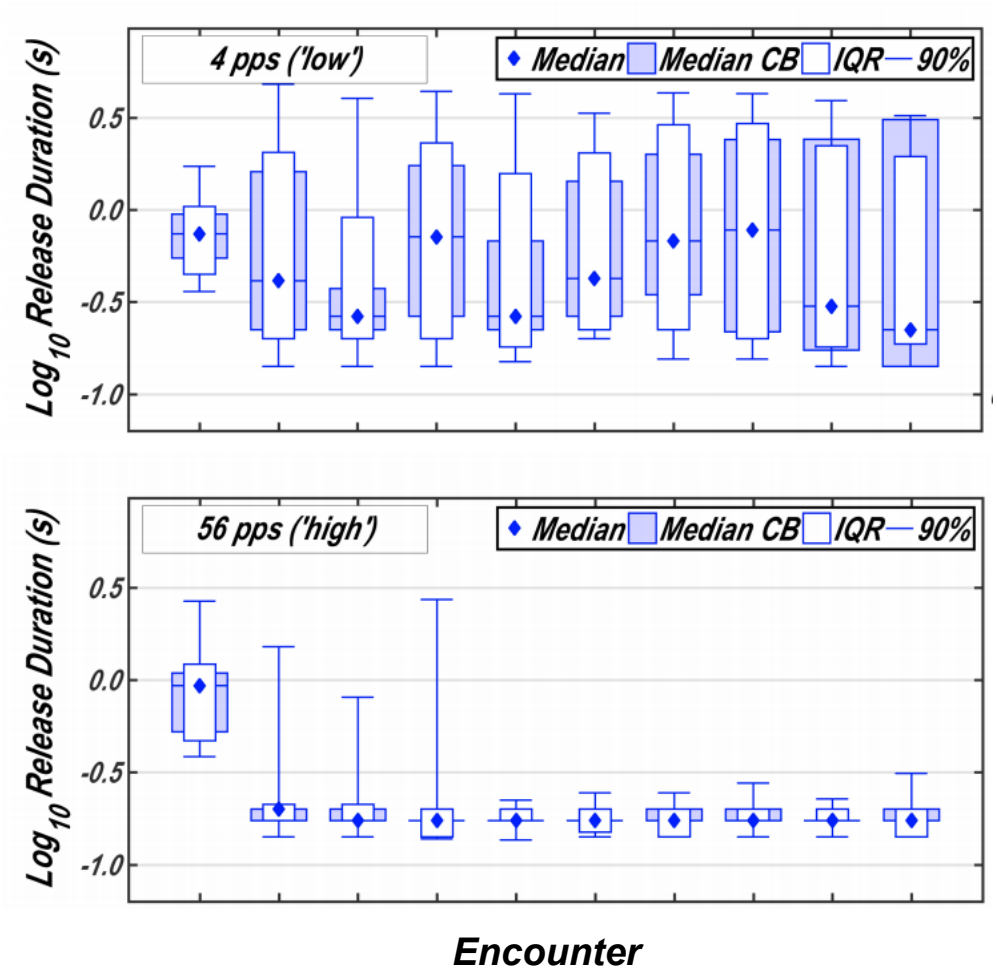



*Stopper et al., Neuron, 2014*




Wang *et al., arXiv;* under review

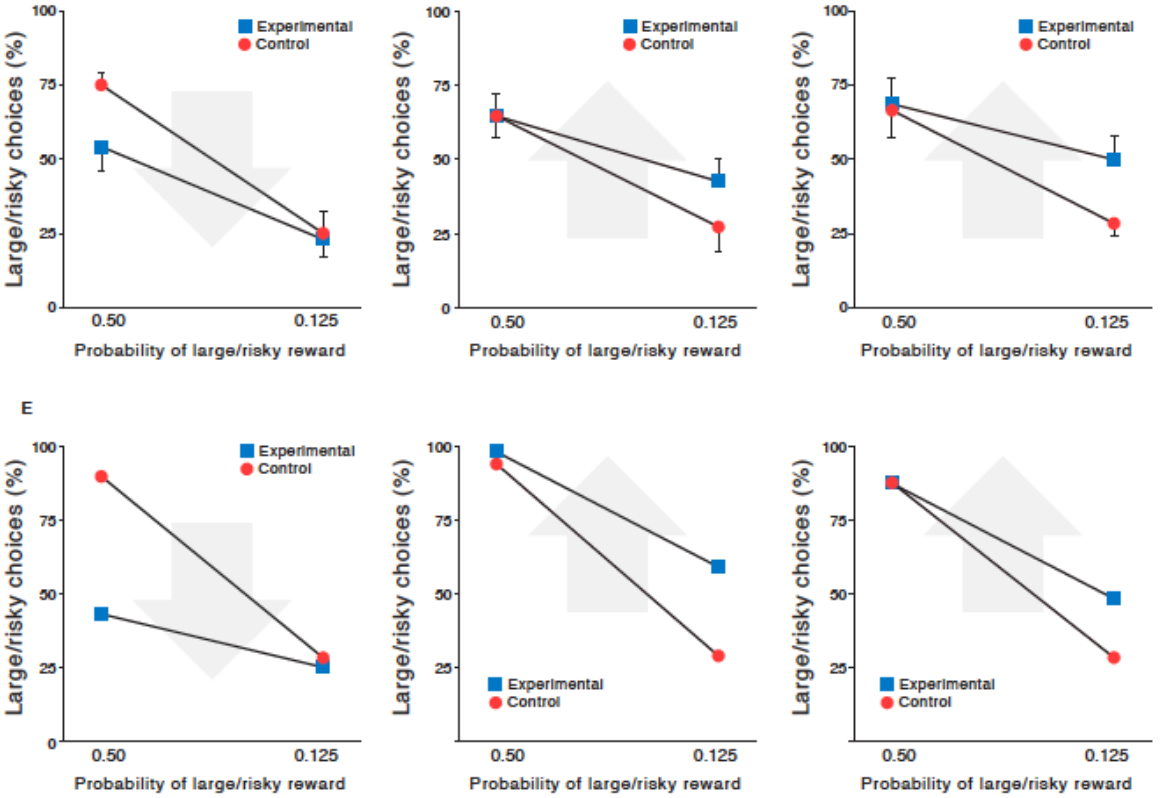



*Trujillo-Pisanty et el., SFN, 2016*

*DA blocked upon food reward from large/risky option*

*DA blocked upon food reward from small/certain option*

*DA triggered upon food omission from large/risky option*




Wang *et al., arXiv;* under review

- Meta-reinforcement learning: a new framework recasting roles of DA and recurrent dynamics of PFC within reward-driven learning
- Three key requirements:
  - PFC recurrent dynamics integrating past reward, history, and observations
  - Primary DA-based RL algorithm that uses reward prediction error to adjust weights
  - Multi-environment task drawn from a distribution
- Emergent, learned RL algorithm implemented by PFC activity dynamics exploits correlations and task/reward structure

# Collaborators

Jane Wang
Zeb Kurth-Nelson
Dharshan Kumaran
Chris Summerfield
Hubert Soyer
Joel Leibo
Sam Ritter

Adam Santoro
Tim Lillicrap
David Barrett
Dhruva Tirumala
Remi Munos
Charles Blundell
Demis Hassabis

DeepMind, London UK
Gatsby Computational Neuroscience Unit, UCL