Prefrontal cortex as a meta-reinforcement learning system

Wang et al.

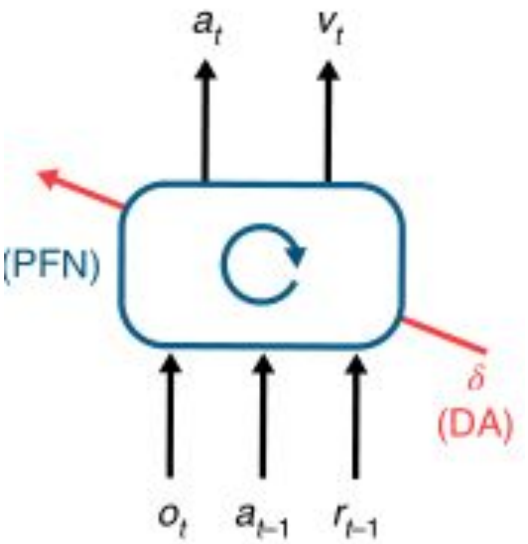
CS330 Student Presentation

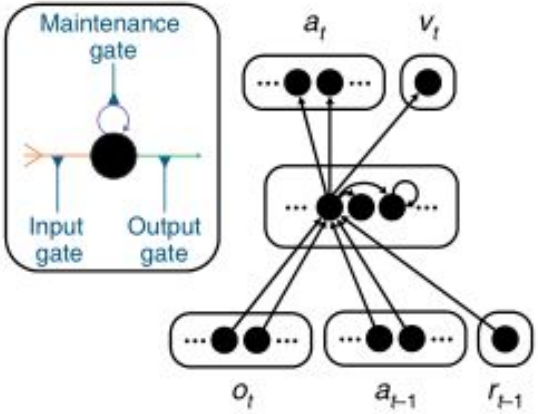
Motivation

- Computational Neuro: **AI <> Neurobio Feedback Loop**
  - Convolutions and the eye, SNNs and Learning Rules, etc.
- **Meta Learning to Inform Biological Systems**
  - Canonical Model of Reward-Based Learning
    - dopamine 'stamps in' associations between situations, actions and rewards by modulating the strength of synaptic connections between neurons.
  - Recent findings have placed this standard model under strain.
    - neural activity in PFC appears to reflect a set of operations that together constitute a self-contained RL algorithm
- **New model of Reward Based Learning** - proposes a insights from Meta-RL that explain these recent findings
  - 6 simulations - tie experimental neuroscience data to matched Meta-RL outputs

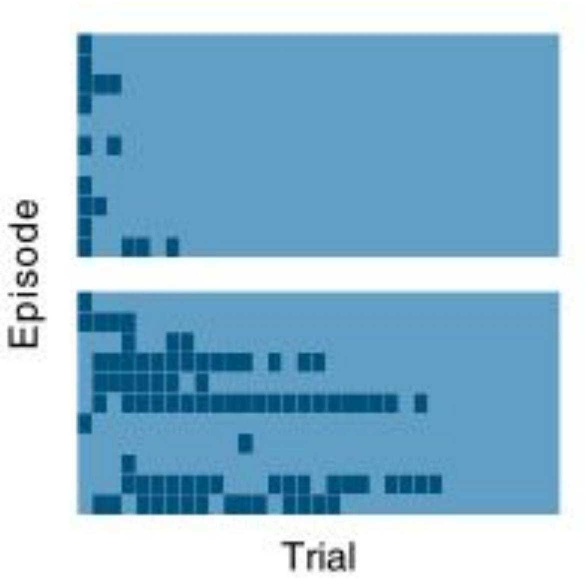
Modeling Assumptions

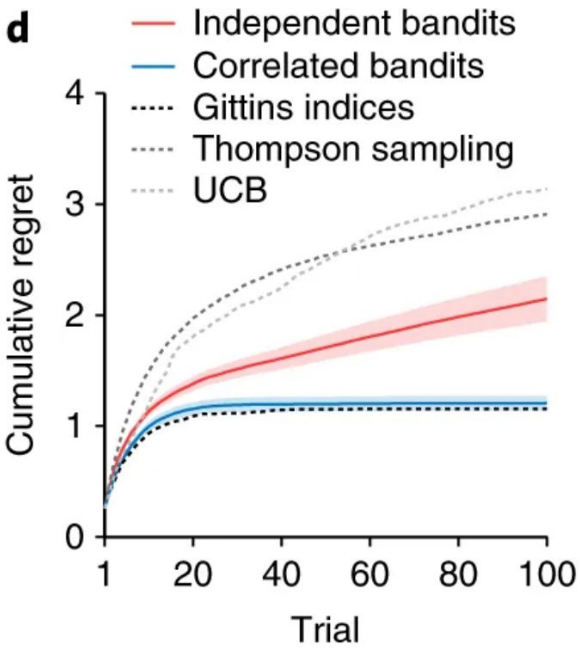
- System Architecture
  - PFC (and basal ganglia, thalamic nuclei) as an RNN
  - **Inputs**: Perceptual data with accompanying information about actions and rewards
  - **Outputs**: triggers for actions, estimates of state value
- Learning
  - **DA - RL system for synaptic learning (meta train)**
    - Modified to provide RPE, in place of reward, as input to the network
  - **PFC - RL system for activity based representations (meta-test)**
- Task Environment
  - RL takes place on a series of interrelated tasks
  - Necessitating ongoing inference and behavioral adjustment





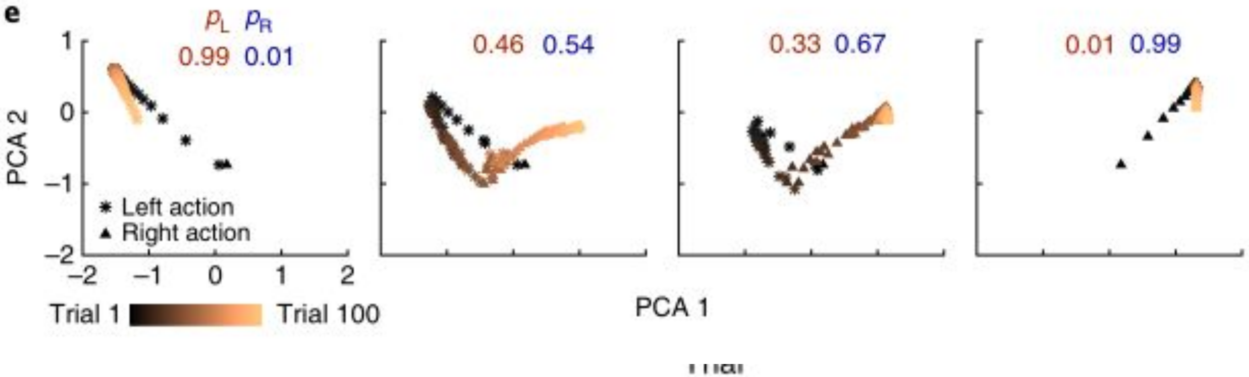
Model Performance - Two Armed Bandit task

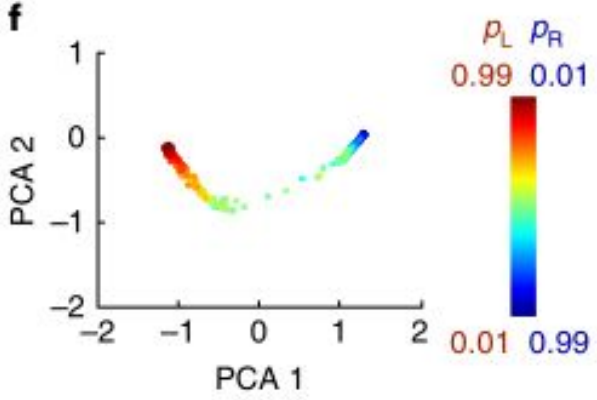




Exploration -> Exploitation
0.25, 0.75 (top)
0.6, 0.4 (bottom)

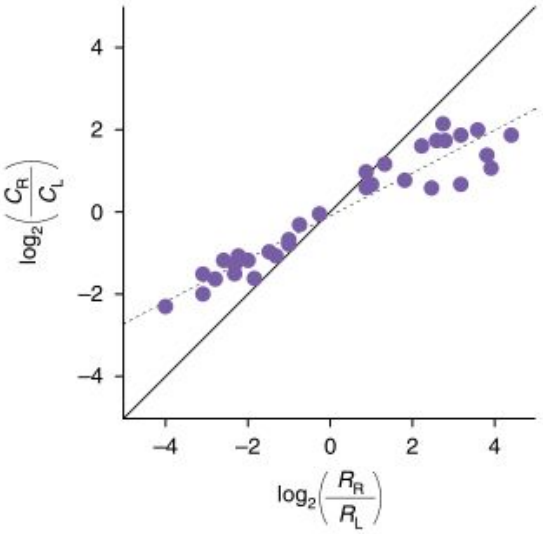
Model Performance - Two Armed Bandit task



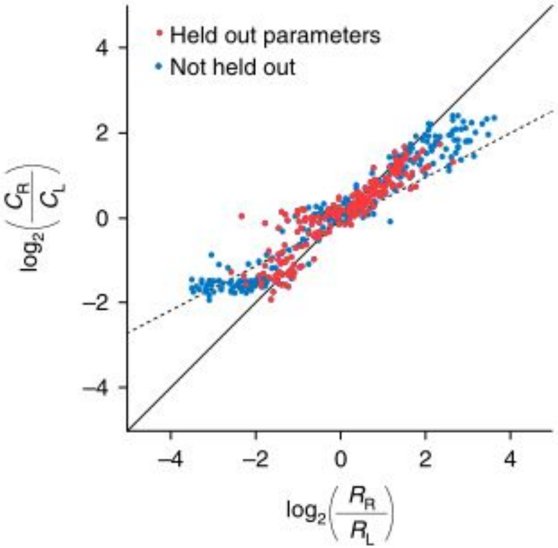


Simulation 1 -

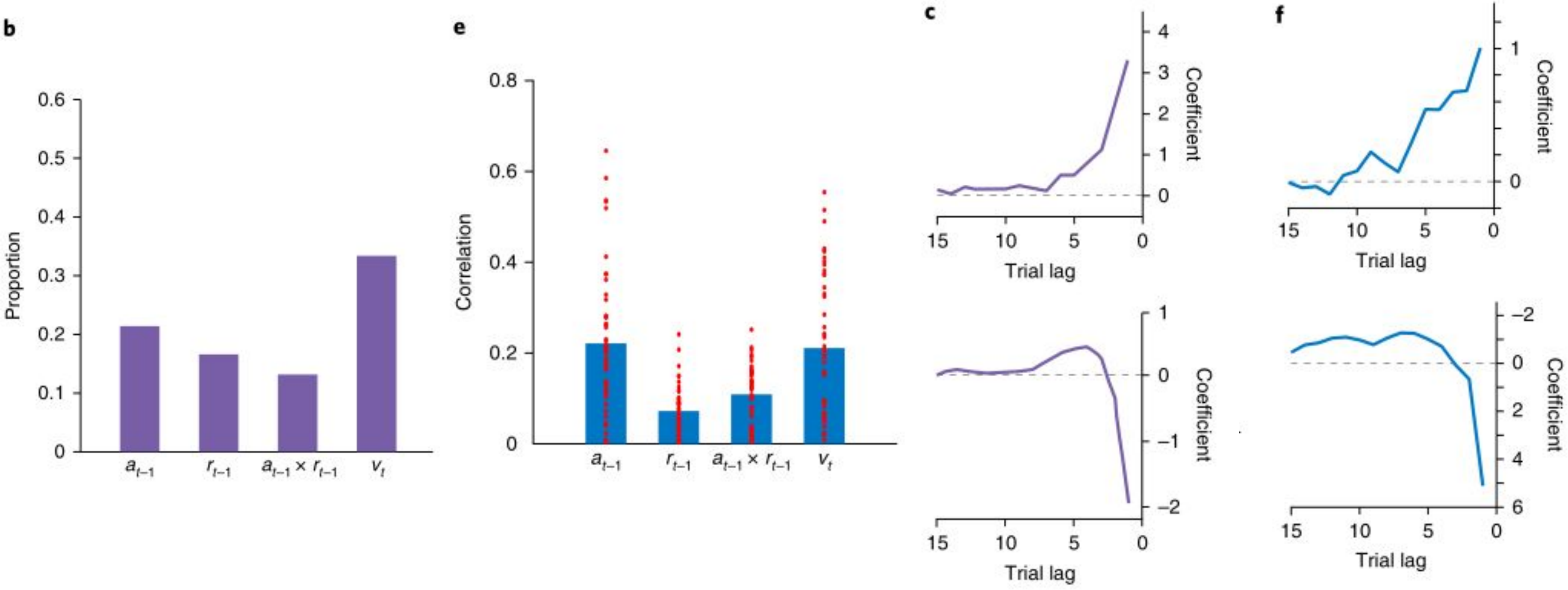
a



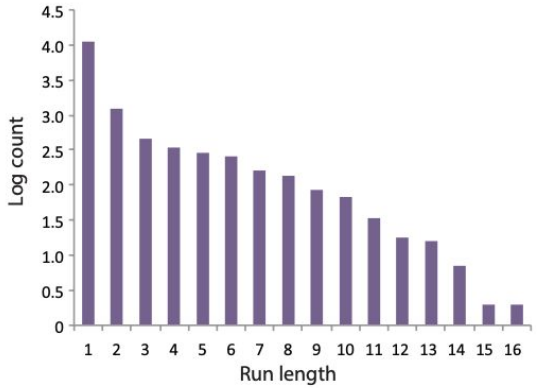
d

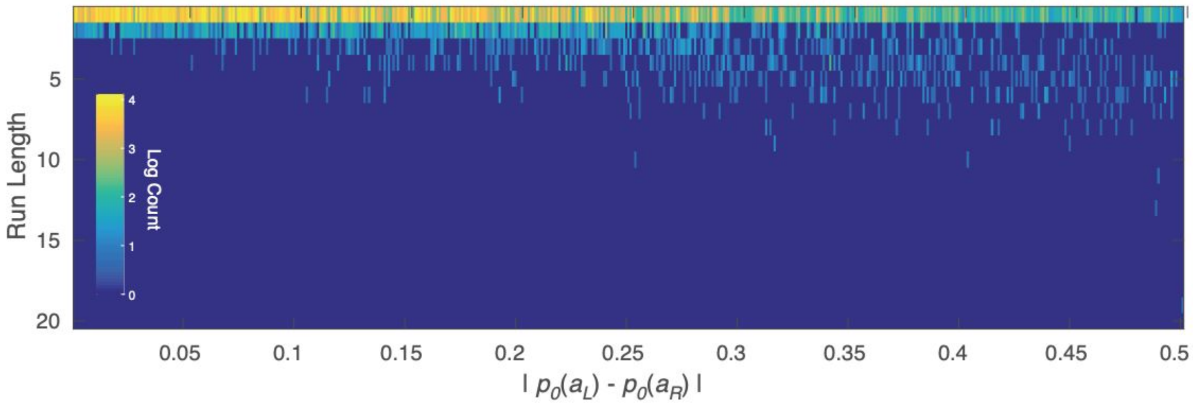


Simulation 1 -



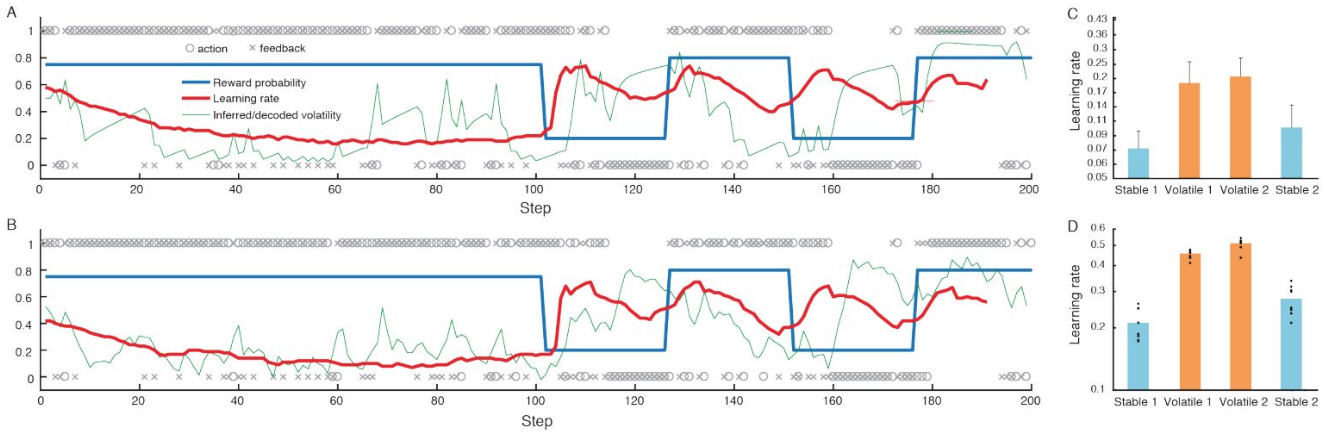
Simulation 1 -





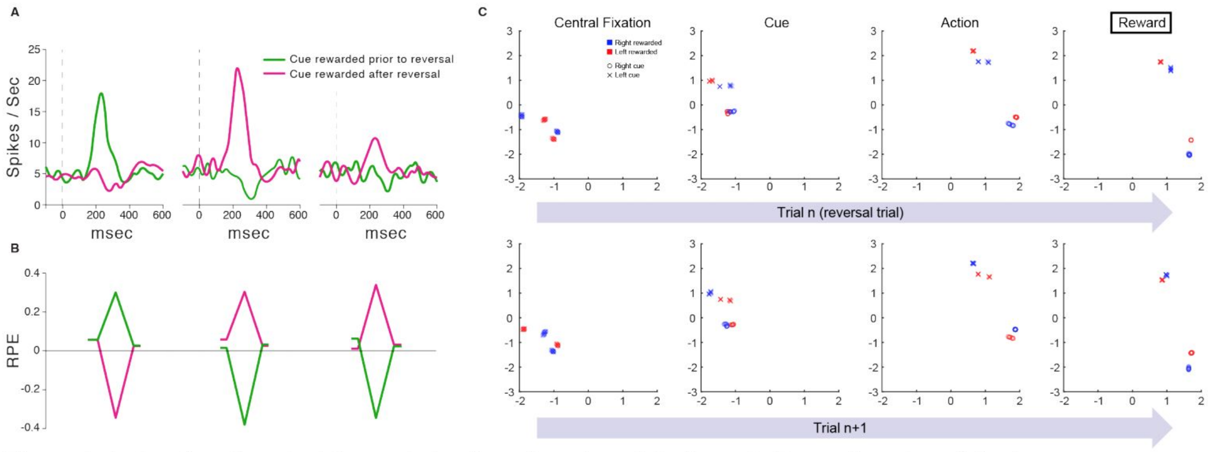
Simulation 2

- Meta Learning on the learning rate
  - Treated as a two-armed bandit task
  - Stable periods vs volatile periods (re: pay-off probabilities)
- Different environment structures will lead to different learning rules



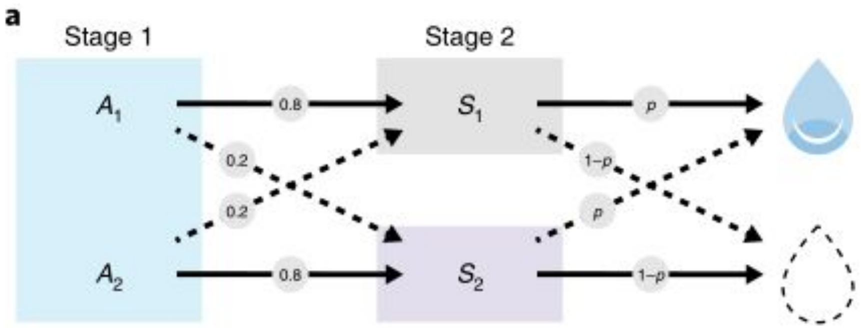
Simulation 3

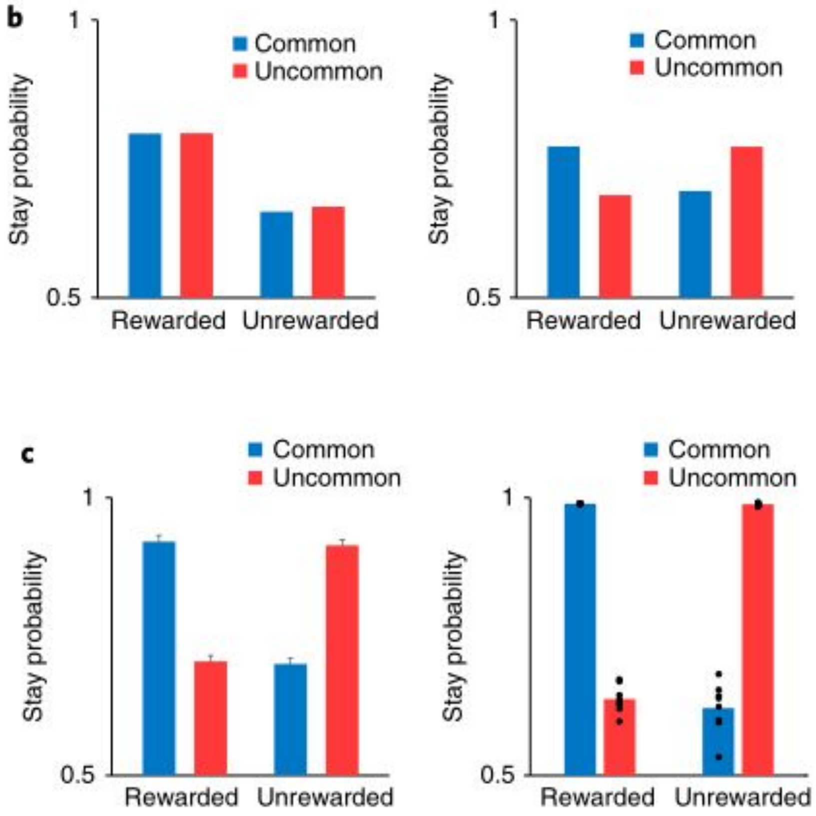
- Visual target appeared to the left or right of a display
- Left or right targets yielded juice rewards and sometimes the roles reversed
  - Whenever the rewards reversed, the dopamine response changed to the other target also changed which show that the hippocampus encodes abstract latent-state representations



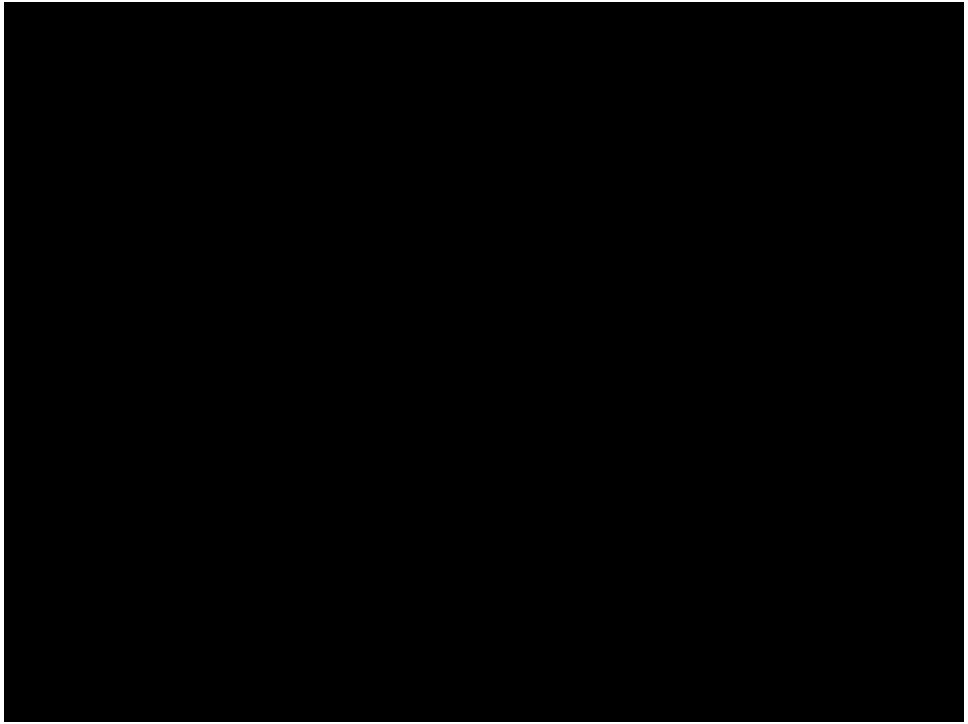
Simulation 4

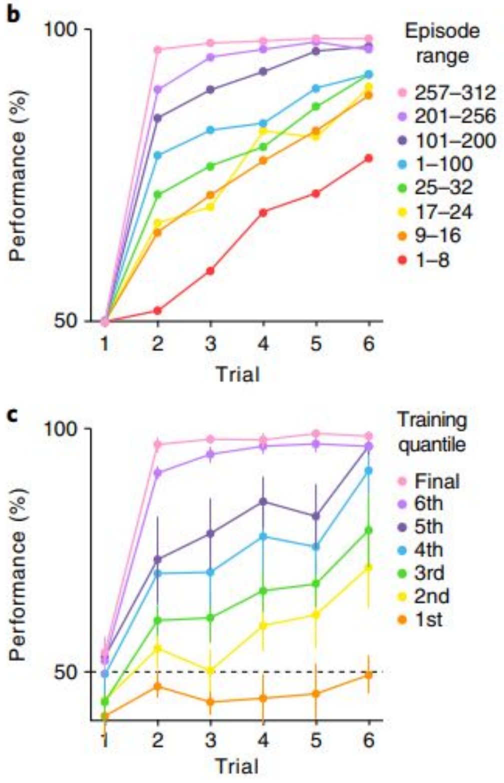
Two step task





Simulation 5



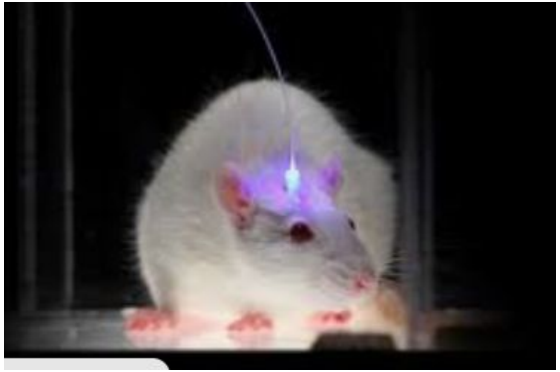


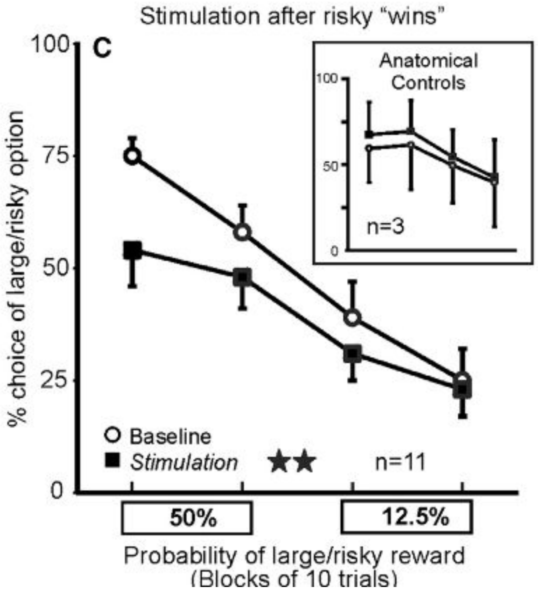
Simulation 6 - Experimental

Setup: **Overriding phasic dopamine** signals **redirects action selection** during risk/reward decision making. *Neuron*

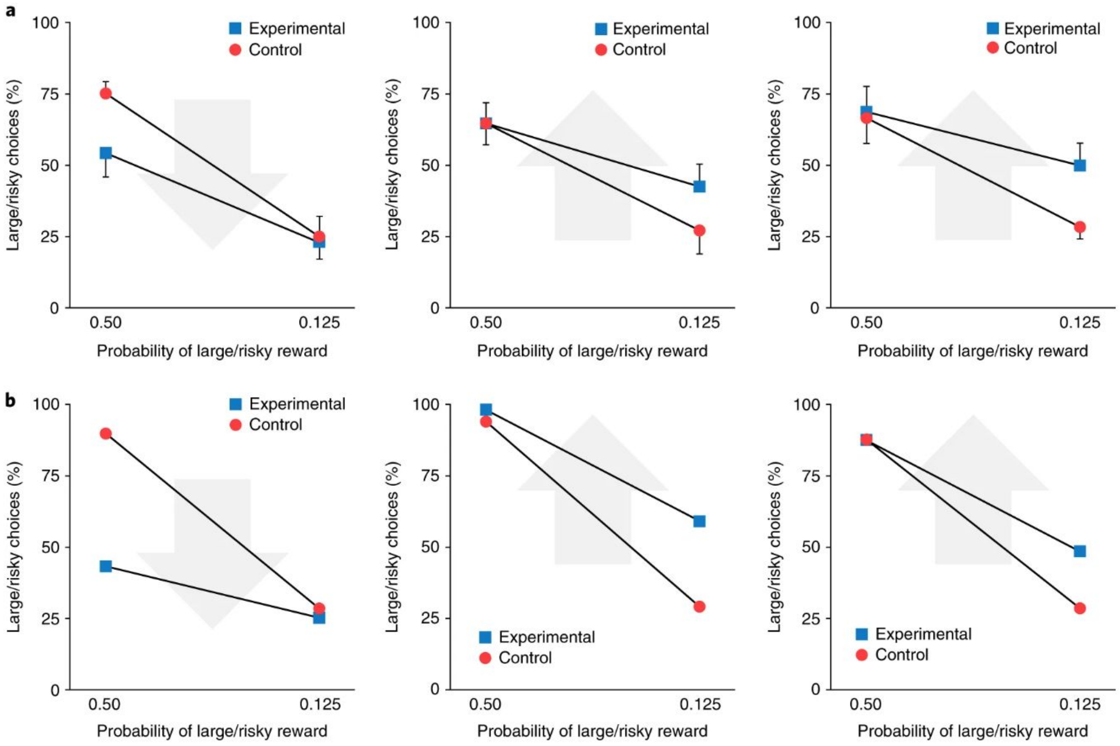
Probabilistic risk/reward task (mice/optogen.)

- Choice: ‘safe’ arm that always offered a small reward (rS = 1) or a ‘risky’ arm that offered a large reward (rL = 4) p = 0.125
- 5 forced pulls each of the safe and risky arms (in randomized pairs), followed by 20 free pulls.





Simulation 6 - Results



Simulate optogenetic stimulation <> **manipulating the value of the reward prediction error** fed into the actor

Same performance across a range of payoff parameters and dopamine interference

Extensions + Criticisms

- Analyses in the paper mostly intuition based - “these charts match up”
    - Ideally should have stronger correlative evidence beyond this
- Observation/end results based, not much to do with physical/inner mechanisms of PFC/DA
    - Results are compared to high level aggregated behaviors
    - Not much exploration/variation into reference architecture used

Overall Conclusions

- Simulations demonstrate comparisons between meta-RL and RL algorithms with human and animal tests
- Various roles of the brain and associated chemicals in creating model-based learning
- Leverage findings from neuroscience/psychology and existing AI algorithms to help explain learning