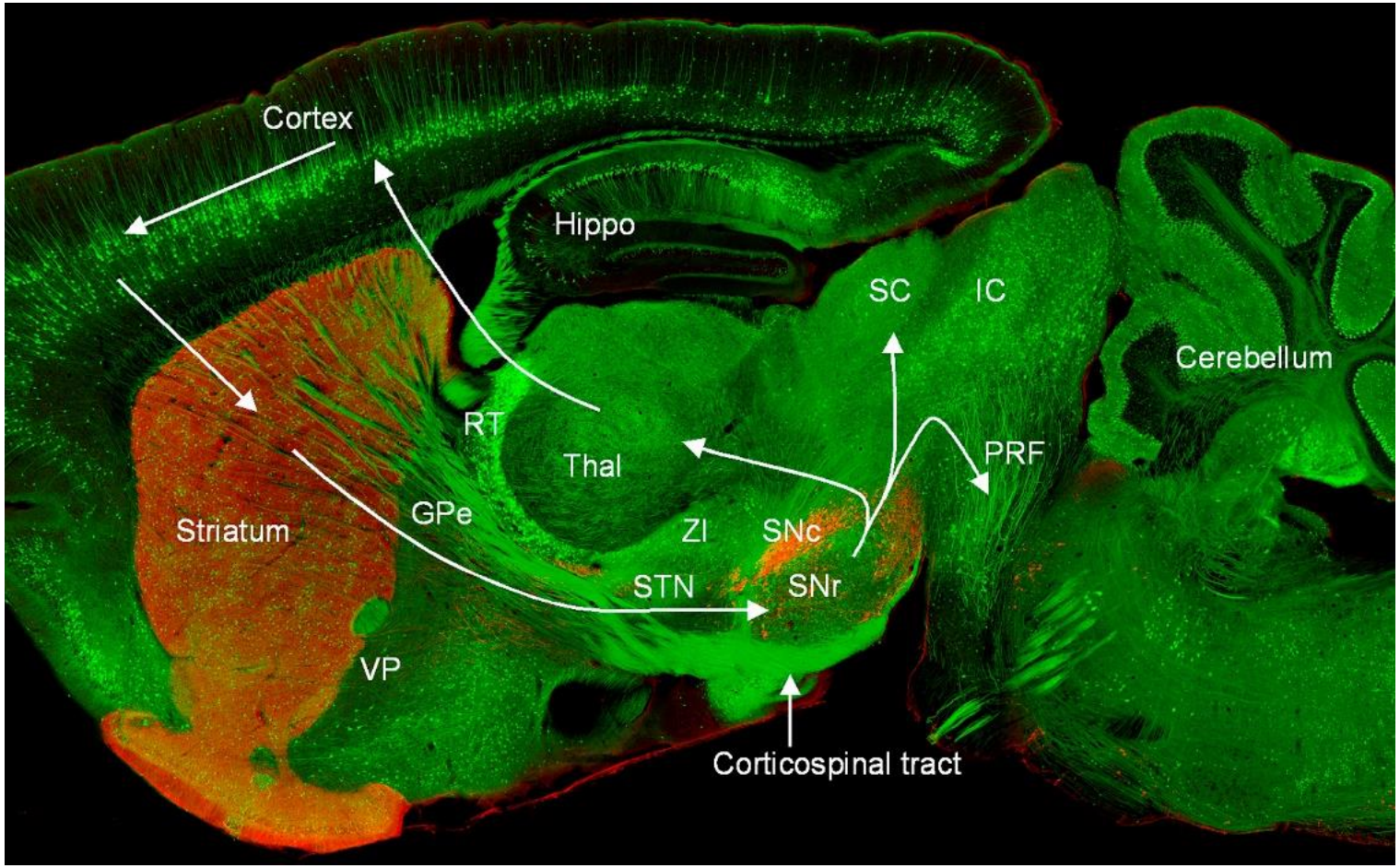
Brain architecture for adaptive behaviour



Thomas Akam
RLDM Tutorial
11/06/2025

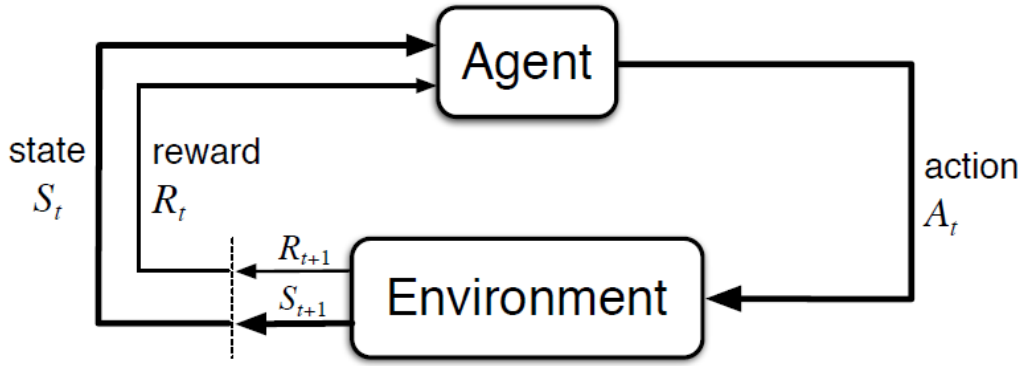
Behavioural neuroscience and Reinforcement learning

Behavioural neuroscience:

How do brains generate (evolutionarily) adaptive behaviour?

Reinforcement learning:

How can agents learn to maximise long run reward?



Information content of human genome:

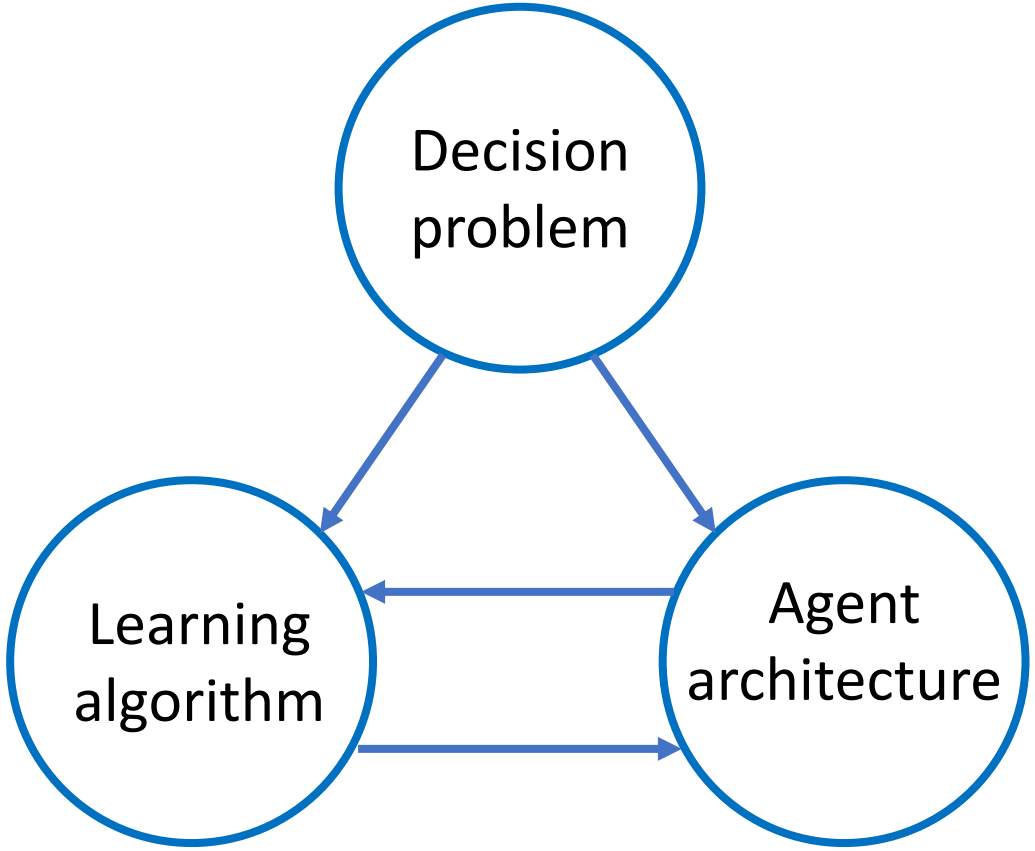
~3 billion base pairs, 2 bit/pair → ~6×10^9 bits

Information content of human cortex:

~1.6×10^{14} synapses, > 1bit/synapse → >1.6×10^{14} bits

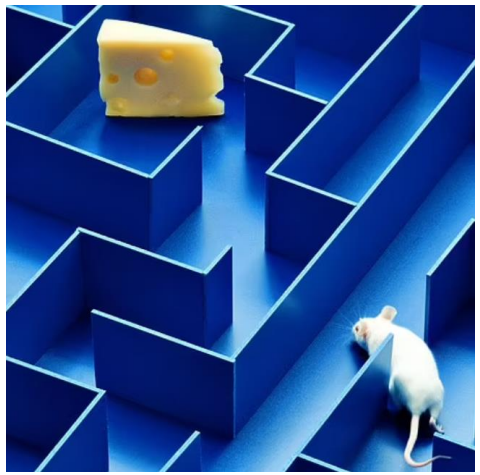
Tang et al. Synapse 2001

Behavioural neuroscience and Reinforcement learning



The brain's control problem

Delayed rewards

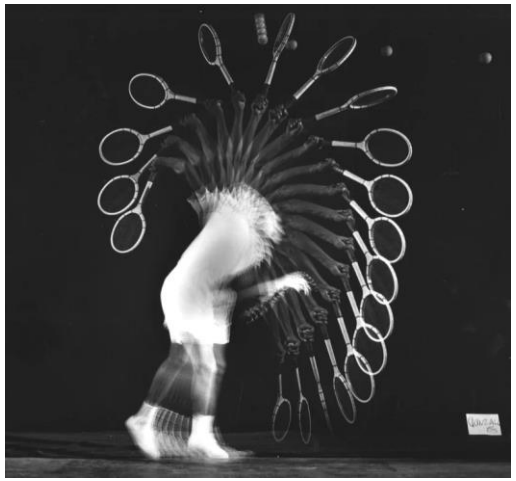


© Getty Images

Complex, non-stationary, state space

© Getty Images

High-dimensional observation, action space



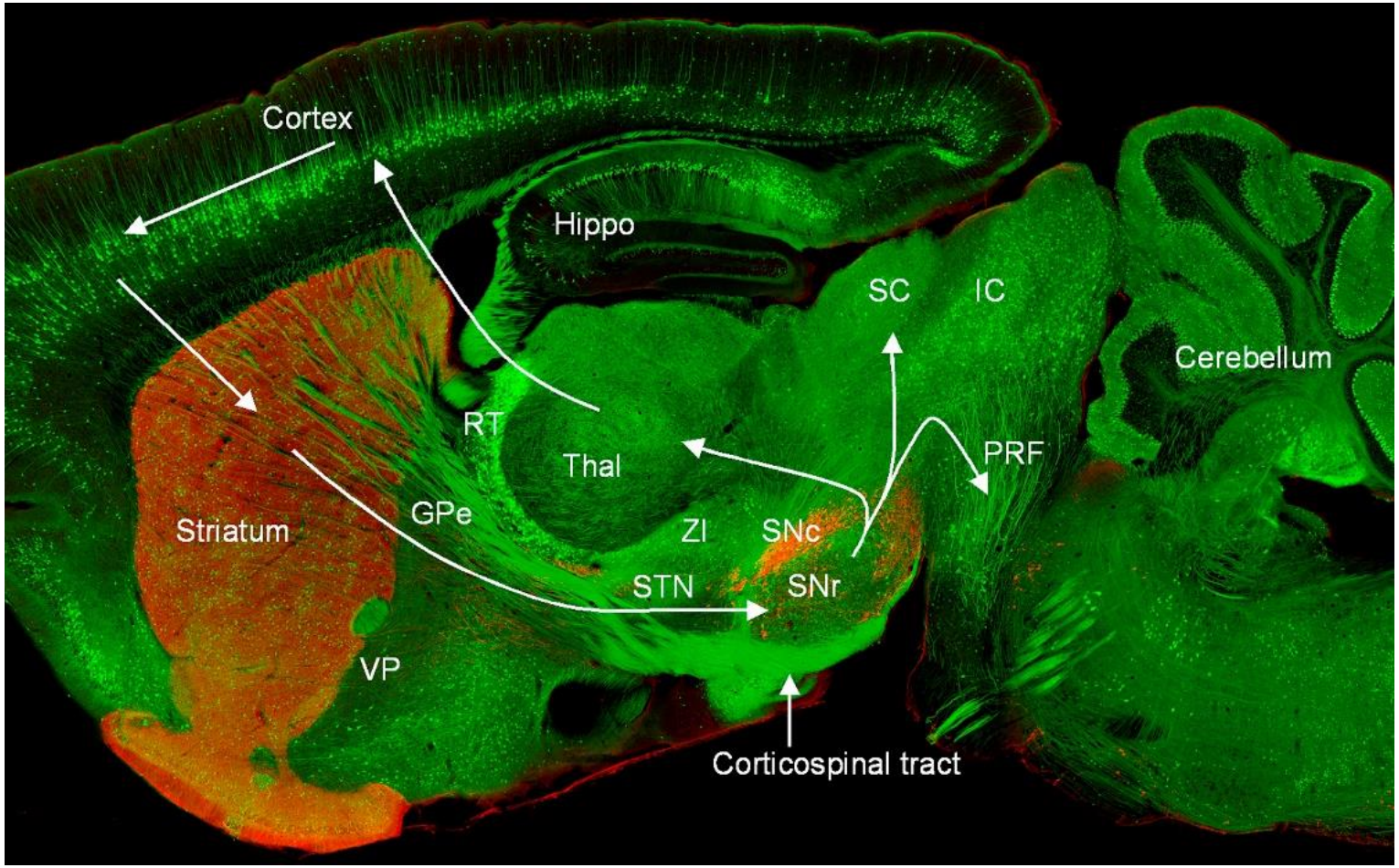
© Harold Edgerton

Partial observability

Lion behind tree

© Nevit Dilmen

What would a satisfying answer look like?



- What learning algorithm(s) does the brain use?
- How do these map onto the architecture of brain circuits?

Zhou 2016

Talk aim and overview

Motivating Question: How do learning algorithms map onto brain structure to solve the biological control problem?

Talk outline:

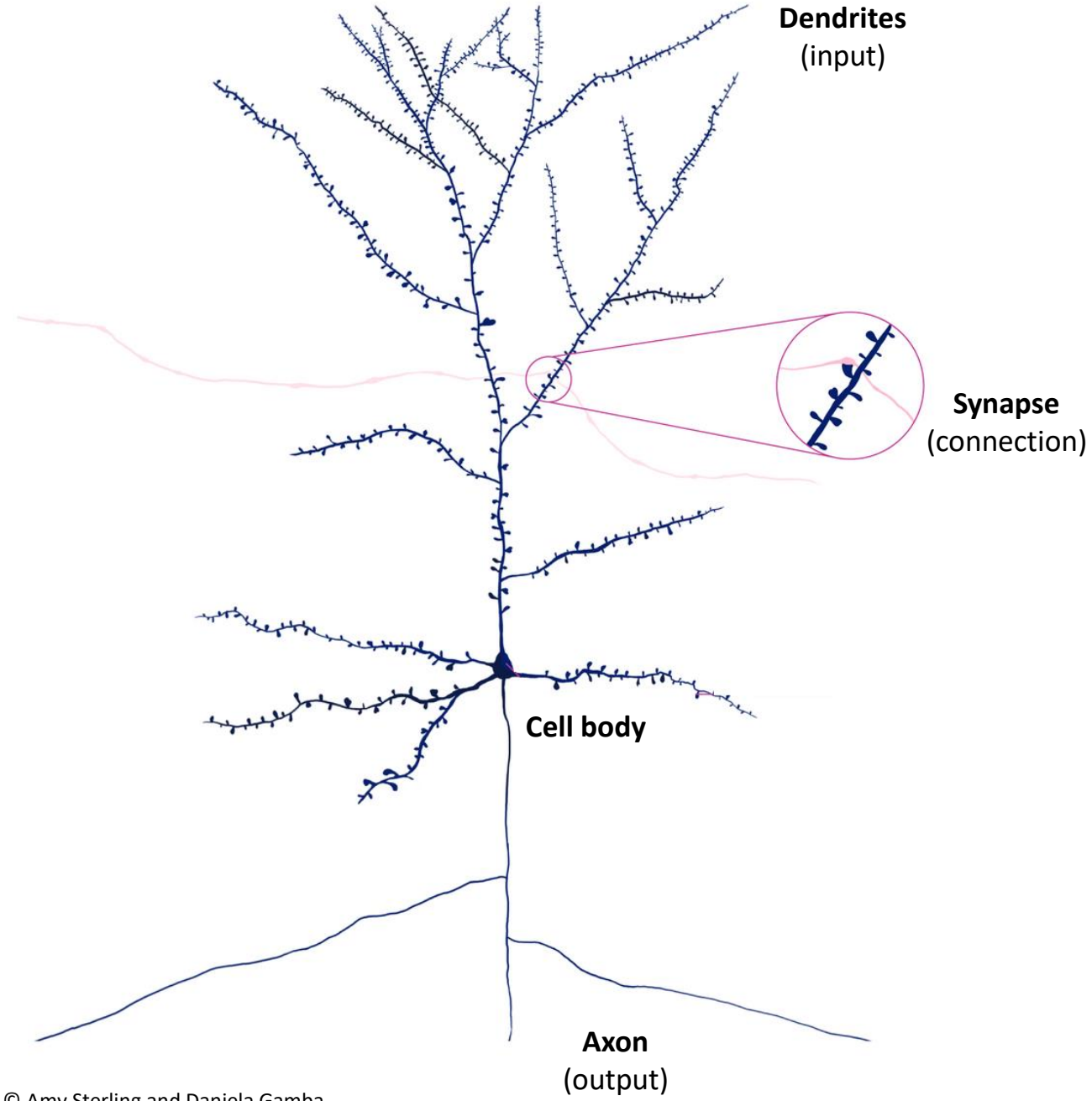
1. Striatum and dopamine: The brain's temporal difference reinforcement learning system?

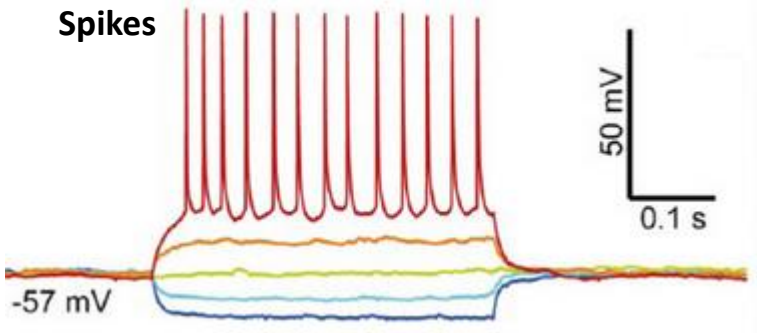
2. Basal ganglia outputs and control of action selection.

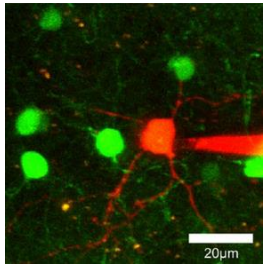
3. Cortex: State representation and beyond

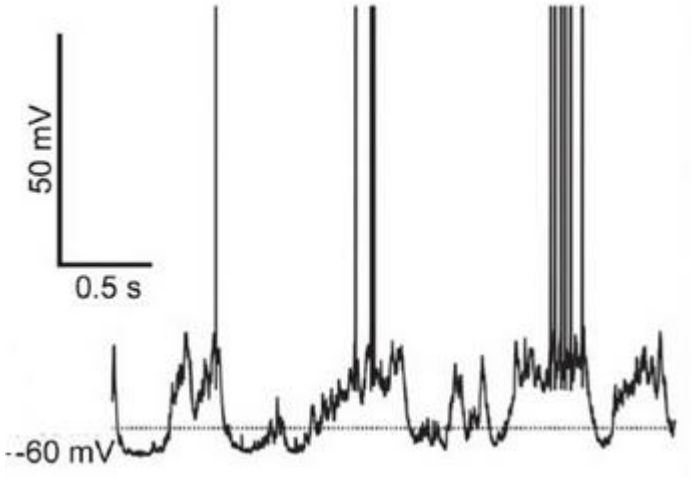
4. Hippocampus: Sequence generation and model-based control

Neuro 101



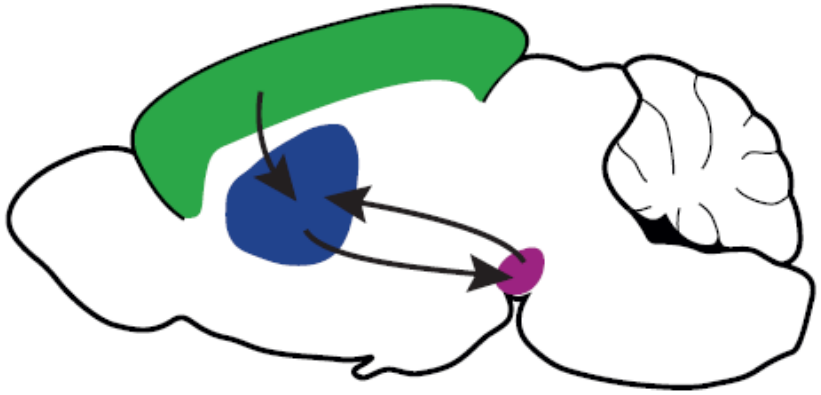


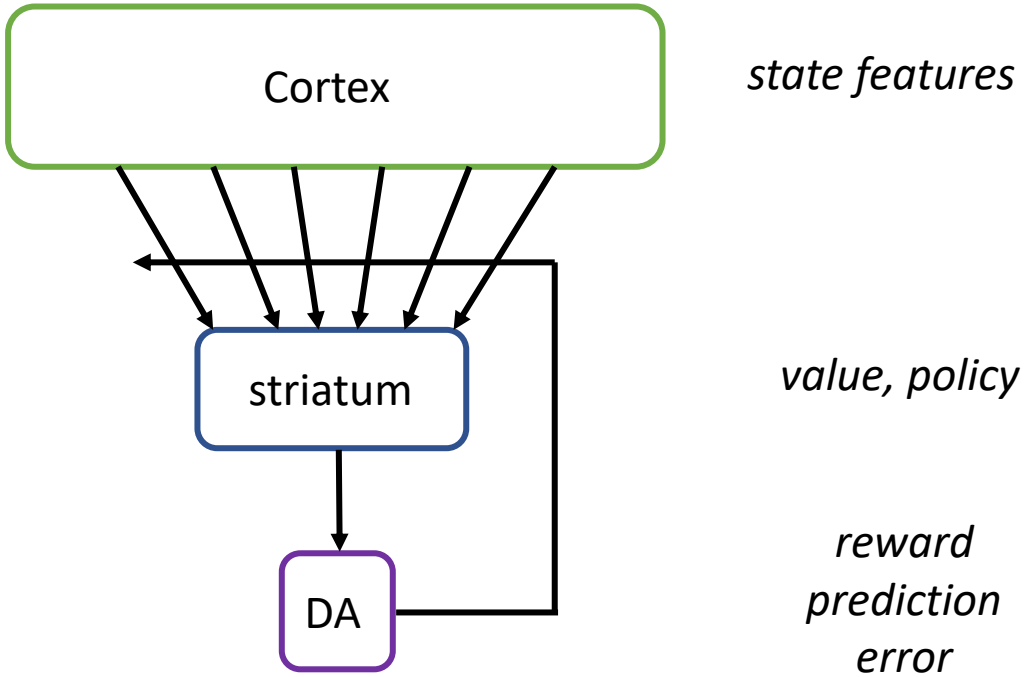




Annecchino & Schultz 2017

1. Striatum and dopamine: The brain's temporal difference (TD) learning system?

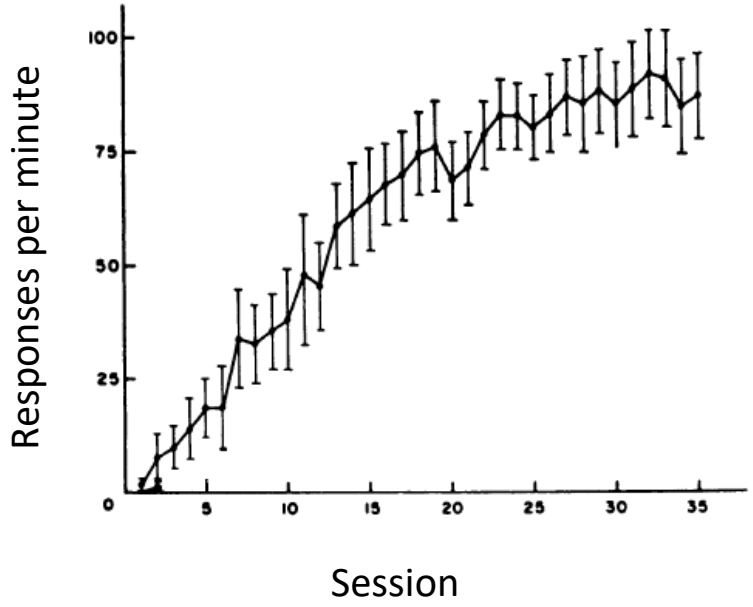


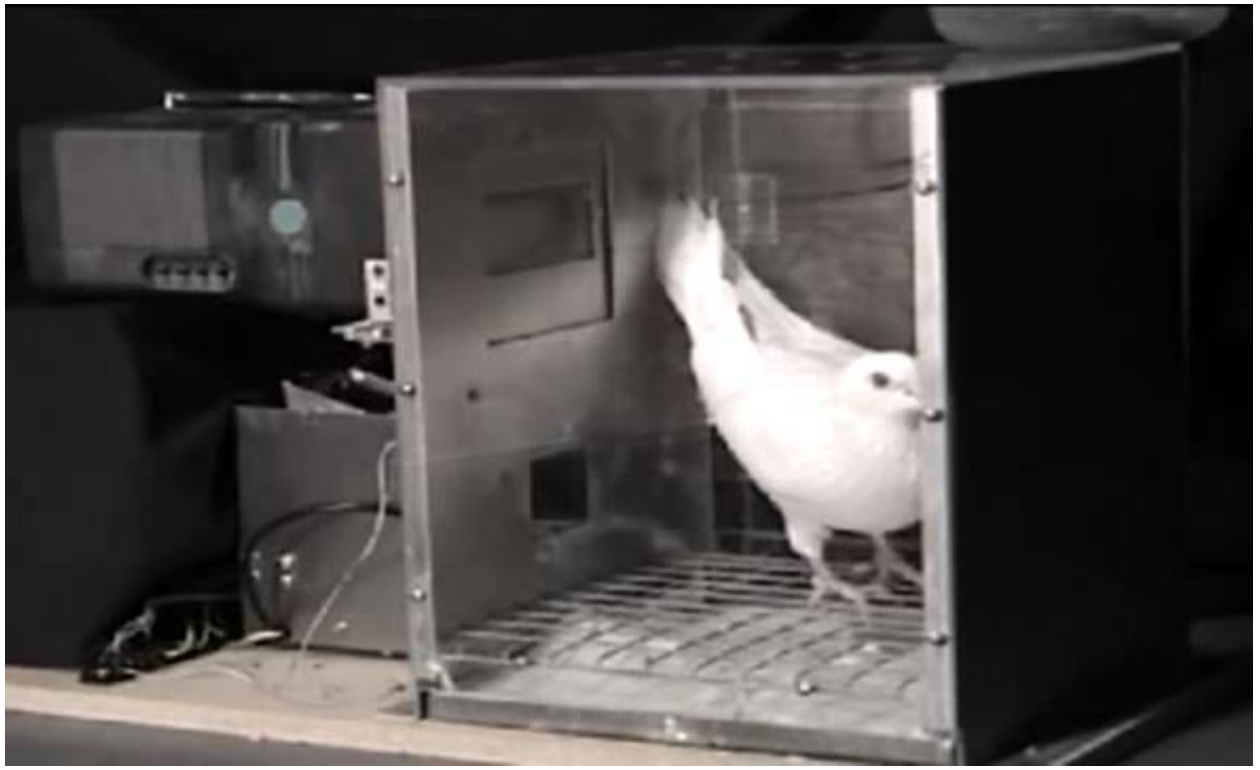


Pavlovian conditioning

Conditioned stimulus (CS)

Unconditioned stimulus (US)



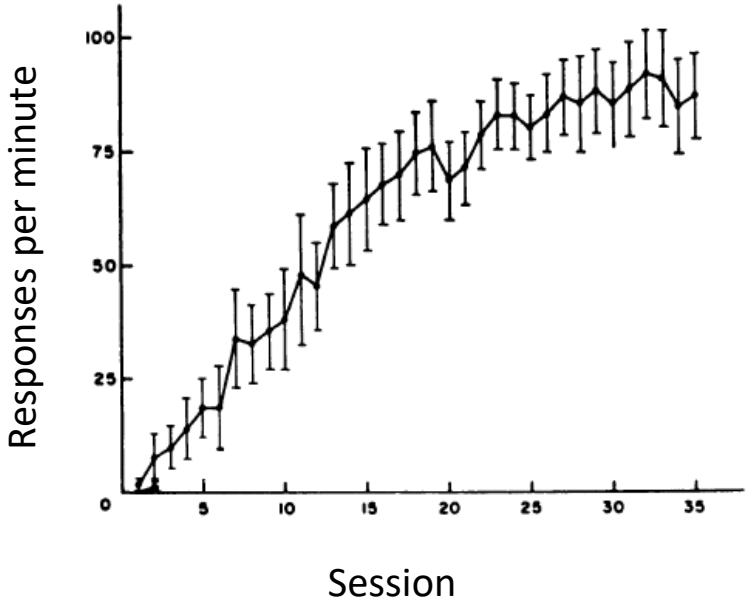


Gamzu and Williams 1973

Pavlovian conditioning

Conditioned stimulus (CS)

Unconditioned stimulus (US)



Rescorla – Wagner model (1972)

Prediction: $V = \sum_i w_i$

Learning: $\Delta w_i = \alpha(R - \sum_i w_i)$

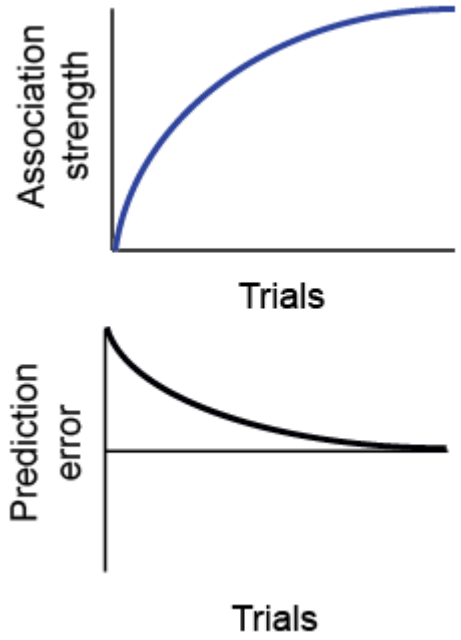
- Δw_i: change in association strength
- α: learning rate
- R: strength of US
- $\sum_i w_i$: sum of association strengths for all CS

Gamzu and Williams 1973

Pavlovian conditioning

Conditioned stimulus (CS)

Unconditioned stimulus (US)



Rescorla – Wagner model (1972)

Prediction: $V = \sum_i w_i$

Learning: $\Delta w_i = \alpha(R - \sum_i w_i)$

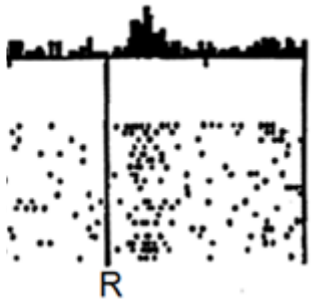
change in association strength

learning rate

prediction error

Dopamine in Pavlovian conditioning

No prediction
Reward occurs



Activity burst to unexpected rewards

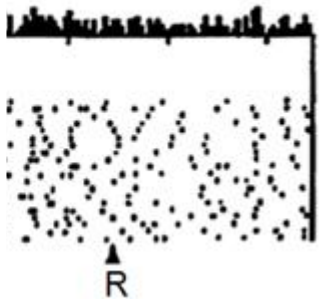
$$\Delta w_i = \alpha(R - \sum_i w_i)$$

change in association strength

learning rate

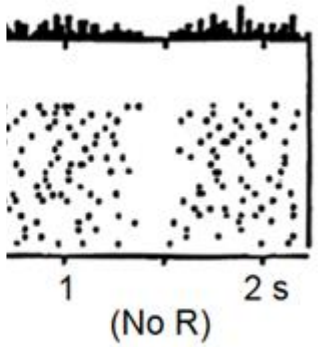
prediction error

Reward predicted
Reward occurs



No response to predicted reward

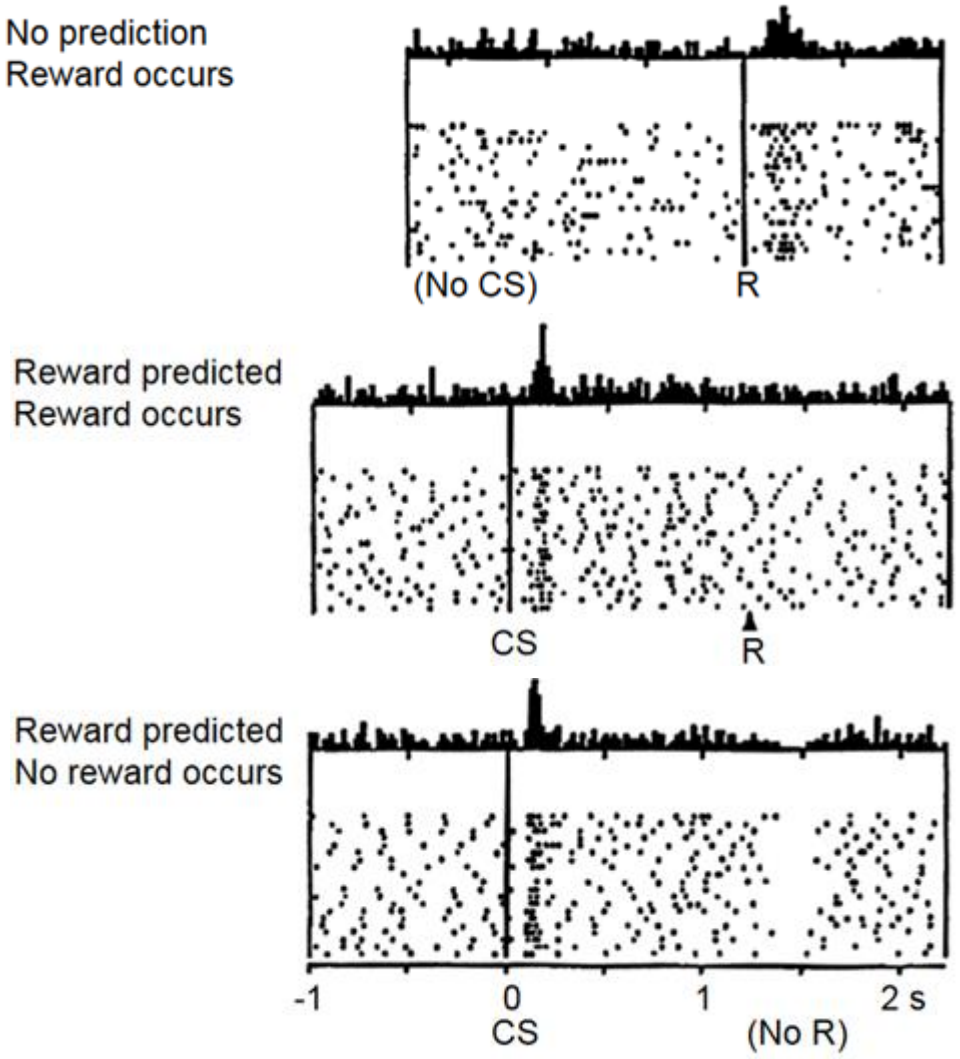
Reward predicted
No reward occurs



Pause to omission of predicted reward

Schultz, Dayan, Montague, 1997

Dopamine in Pavlovian conditioning



Activity burst to reward predicting cue

$$\Delta w_i = \alpha(R - \sum_i w_i)$$

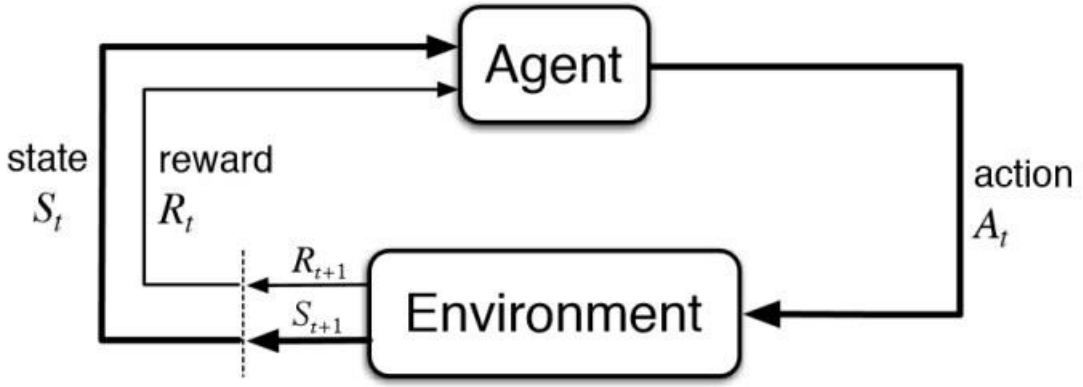
change in association strength

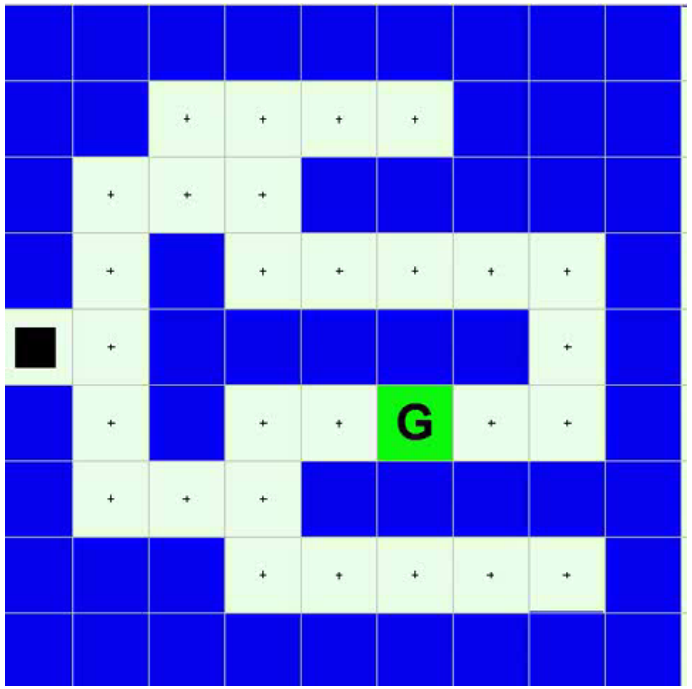
learning rate

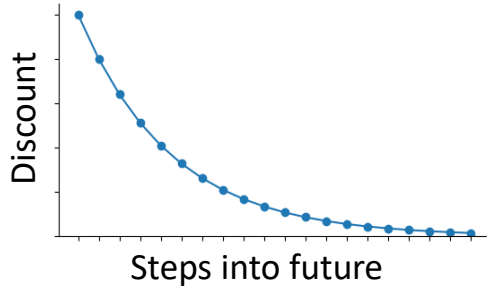
prediction error

Schultz, Dayan, Montague, 1997

Delayed rewards & temporal difference reward prediction errors







$$V(s_t) = E\left[\sum_{k=0}^{\infty} \gamma^k R_{t+k+1}\right]$$

long run value of state s

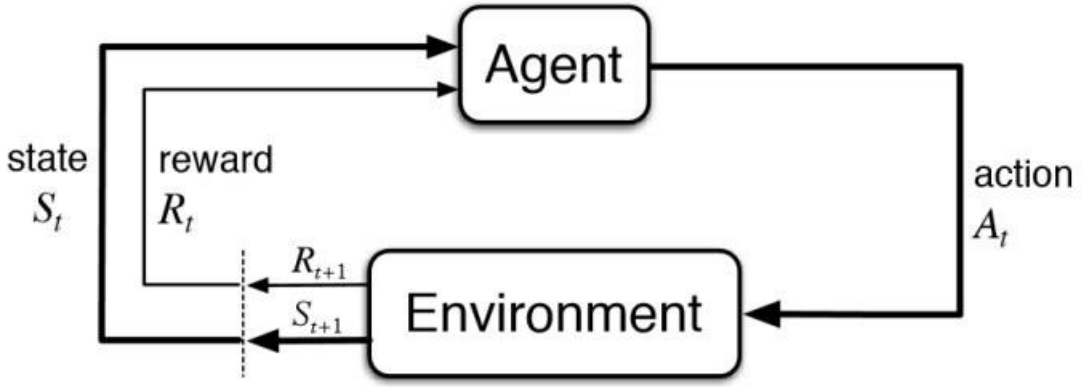
sum of discounted future rewards

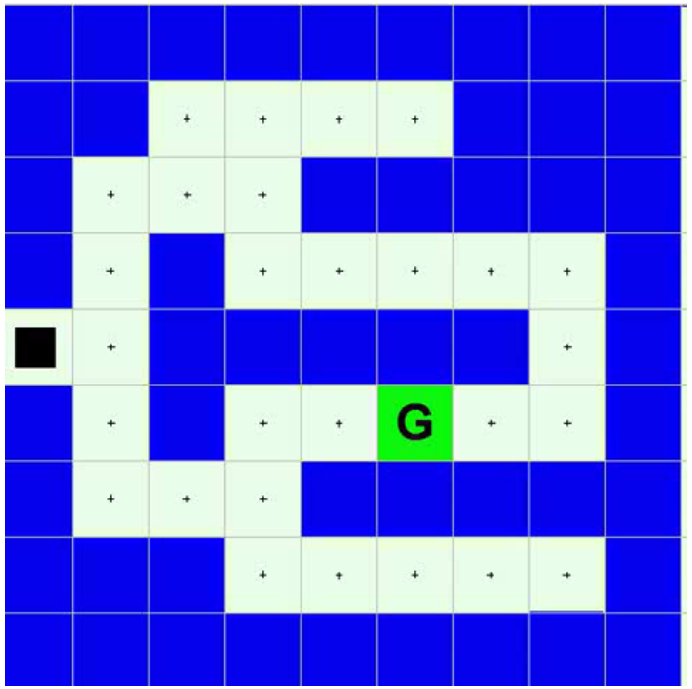
$$V(s_t) = R_{t+1} + \gamma E[V(s_{t+1})]$$

Immediate reward

Discounted value of next state

Delayed rewards & temporal difference reward prediction errors





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

reward prediction error

Update target

prediction

$$V(s_t) \leftarrow V(s_t) + \alpha\delta_t$$

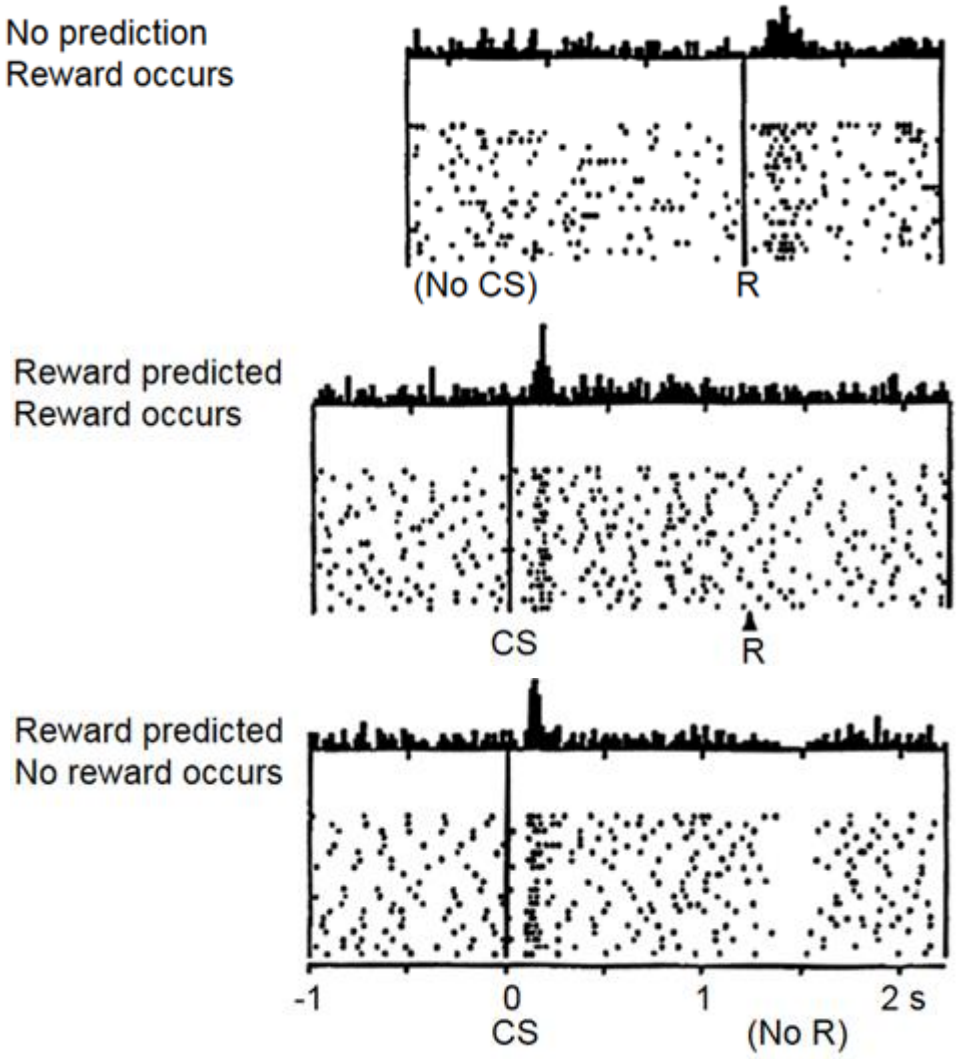
Updated value estimate

old value estimate

learning rate

reward prediction error

Dopamine as a reward prediction error signal

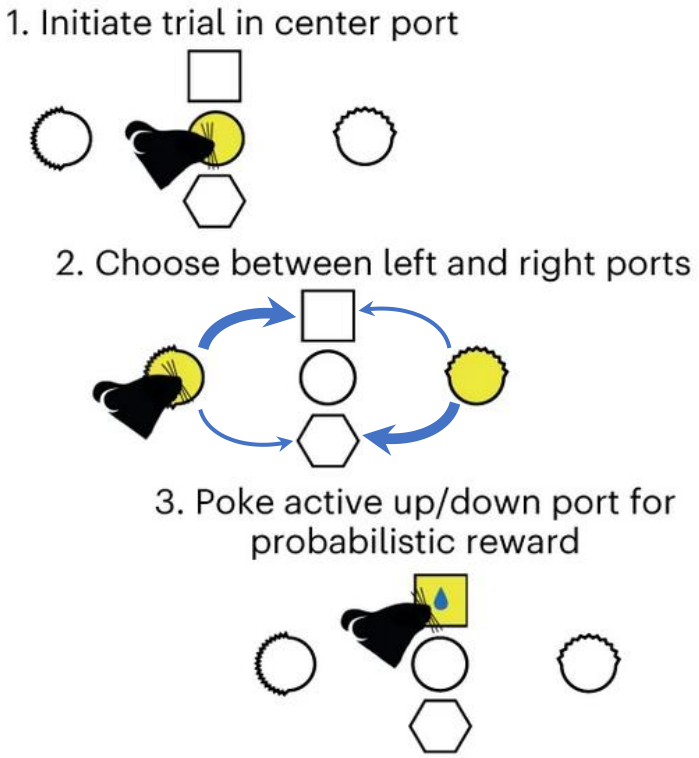


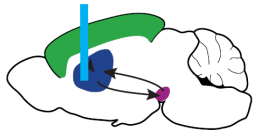
$$\delta_t = R_{t+1} + \gamma V(s_{t+1}) - V(s_t)$$

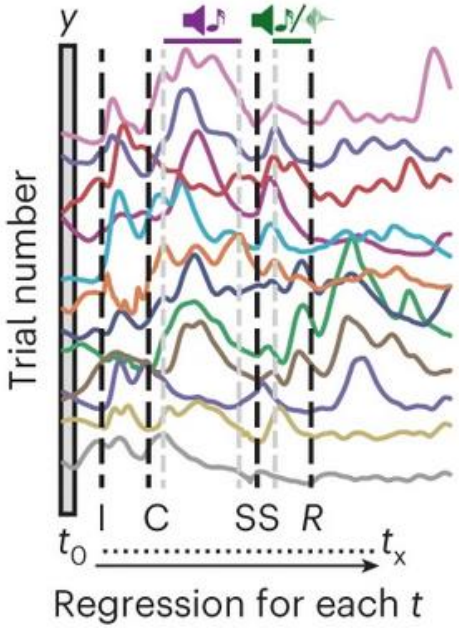
- δ_t: reward prediction error
- R_{t+1}: immediate reward
- $V(s_{t+1})$: value of new state
- $V(s_t)$: value of previous state

Schultz, Dayan, Montague, 1997

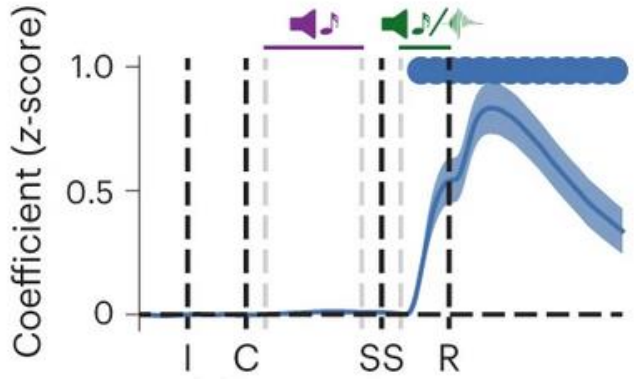
Dopamine as a reward prediction error signal



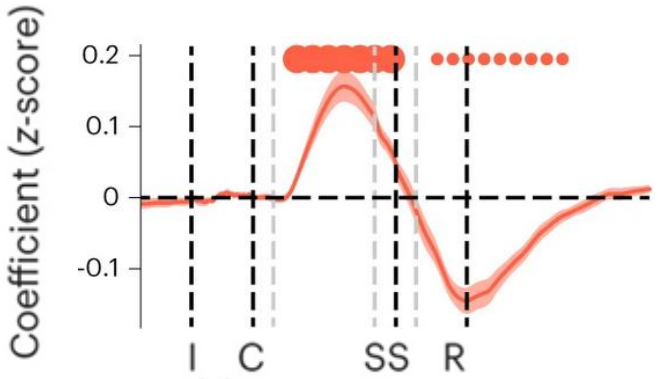


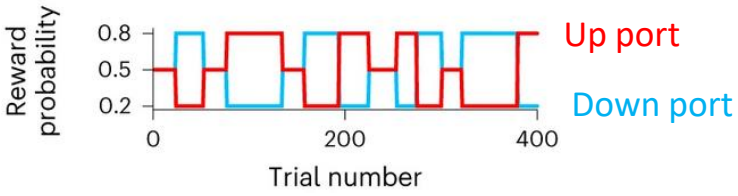


Trial outcome (rewarded or not)



Value of up / down port





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

reward prediction error

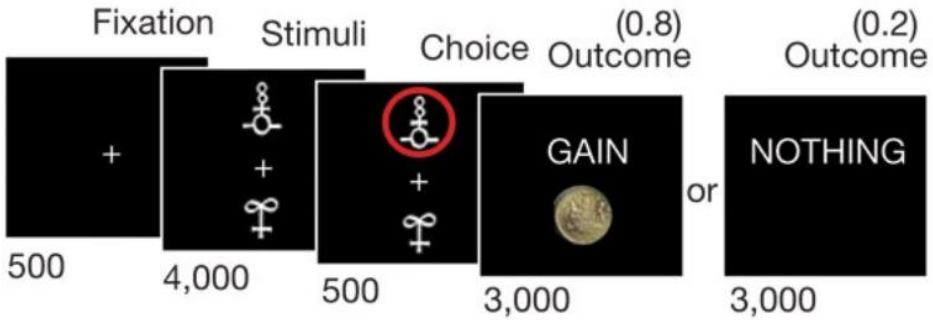
immediate reward

value of new state

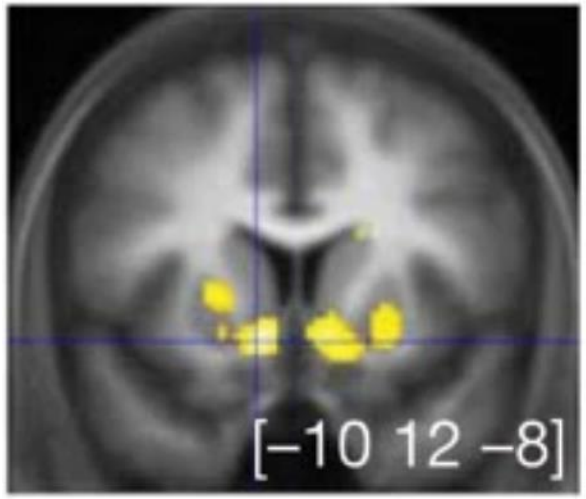
value of previous state

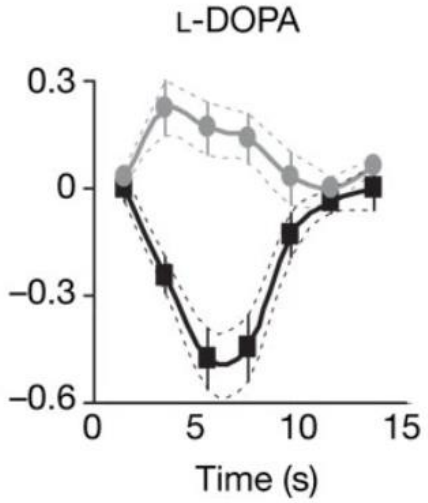
Blanco-Pozo et al. 2024

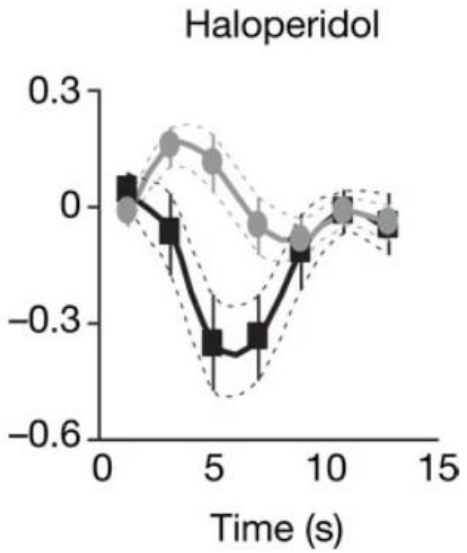
Dopamine as a reward prediction error signal



Voxels correlated with RPE





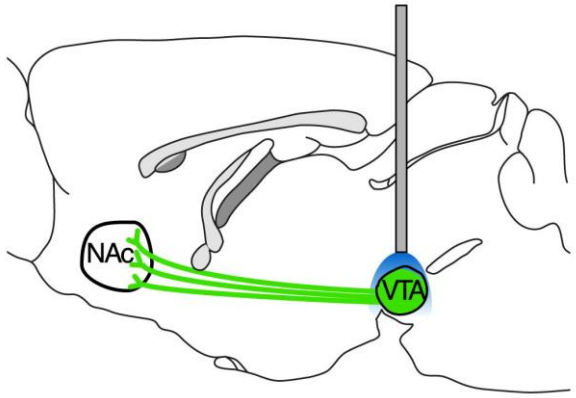


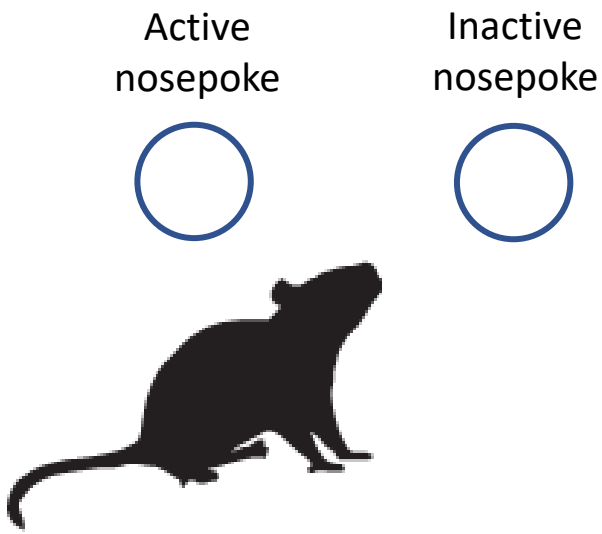
- Brain activity in human ventral striatum correlates with RPE.
- Boosting / blocking dopamine transmission modulates RPE signal size

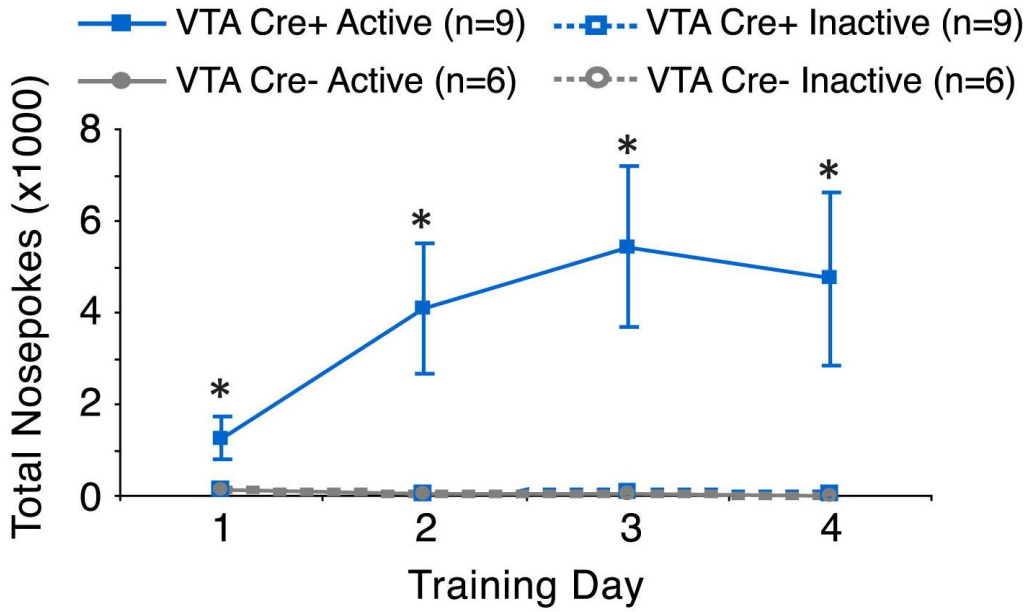
Pessiglione et al. 2008

Causal manipulation of dopamine and reinforcement

Optical stimulation





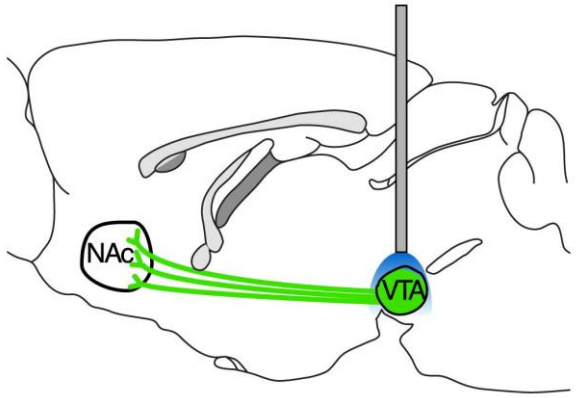


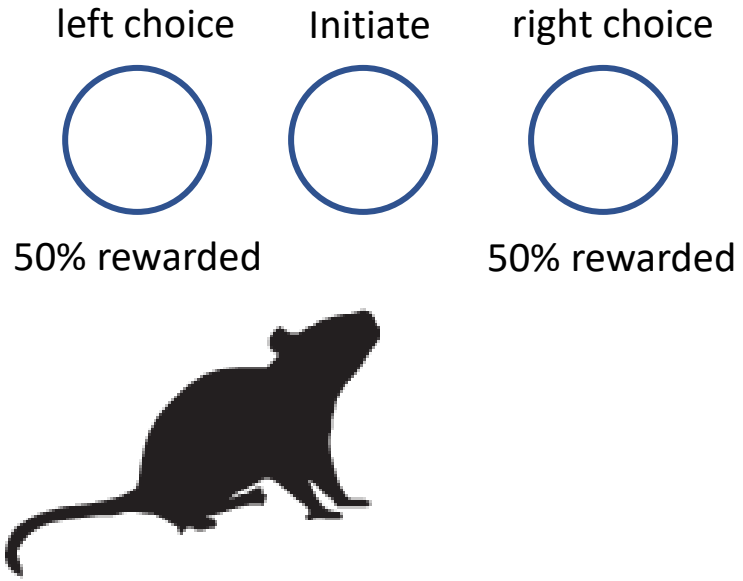
- Optogenetic stimulation of dopamine neurons contingent on action is reinforcing.

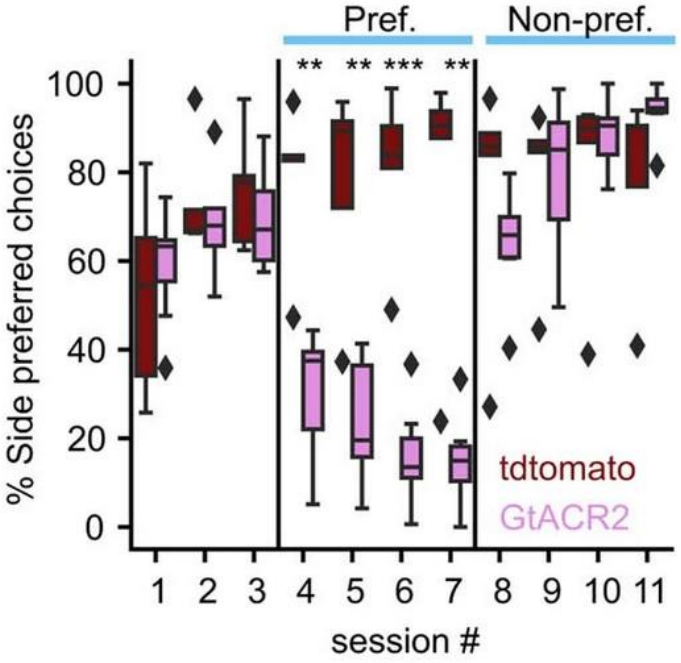
Witten et al. Neuron 2011

Causal manipulation of dopamine and reinforcement

Optical Inhibition



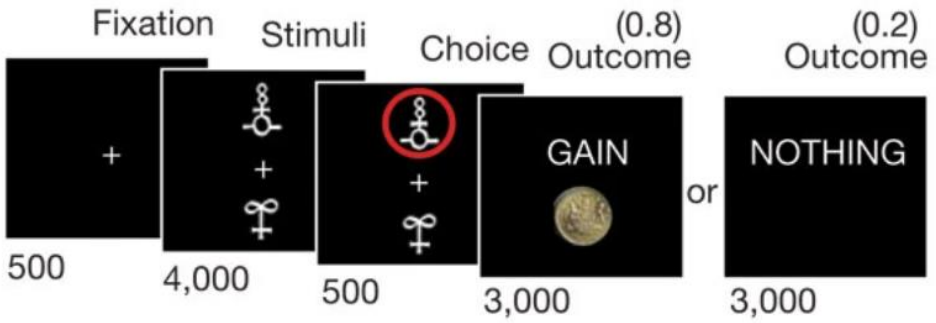


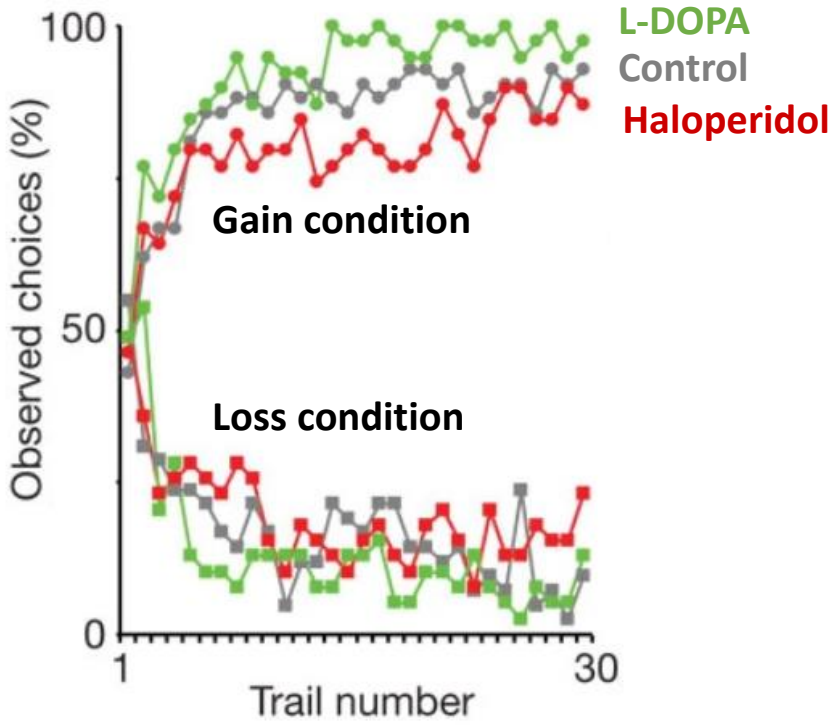


- Optogenetic inhibition of dopamine neurons supresses preceding actions

Blanco-Pozo et al. 2024

Causal manipulation of dopamine and reinforcement





- Boosting / blocking dopamine transmission modulates human learning

Pessiglione et al. 2008

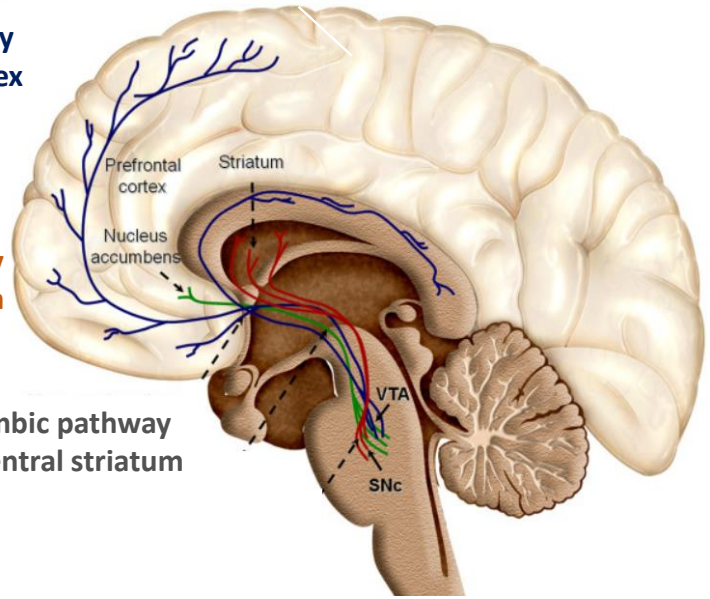
Dopamine projections

Human

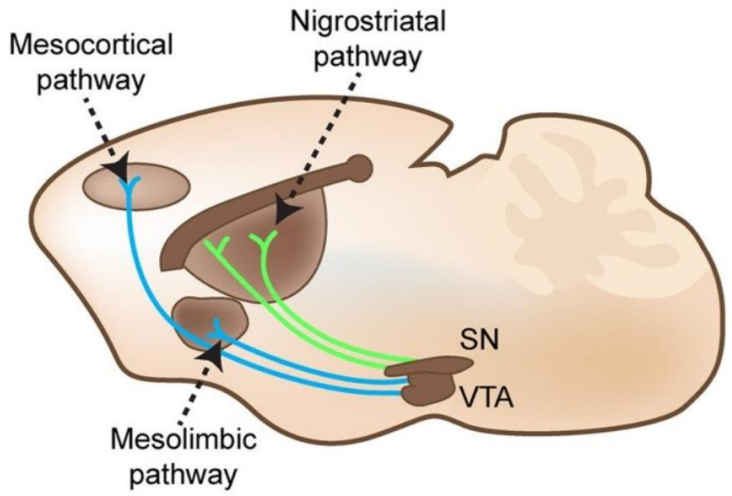
Meso-cortical pathway
VTA → Prefrontal cortex

Nigro-striatal pathway
SNc → dorsal striatum

Meso-limbic pathway
VTA → ventral striatum



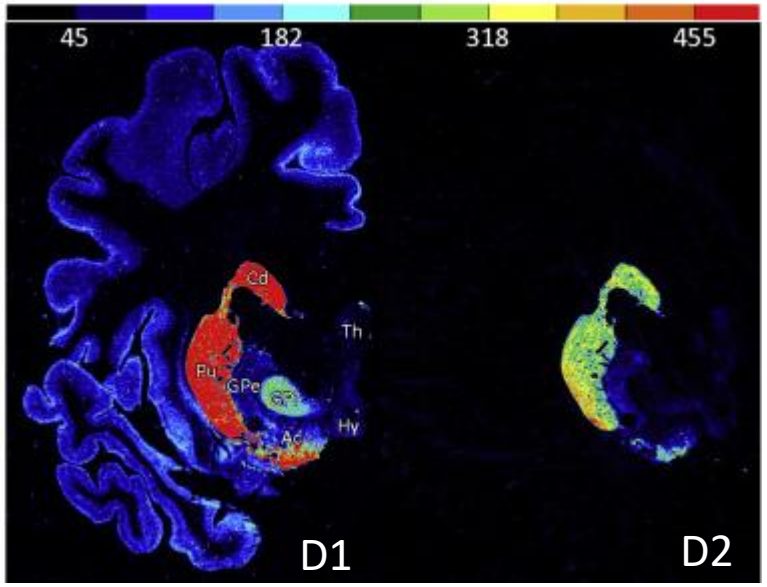
Mouse



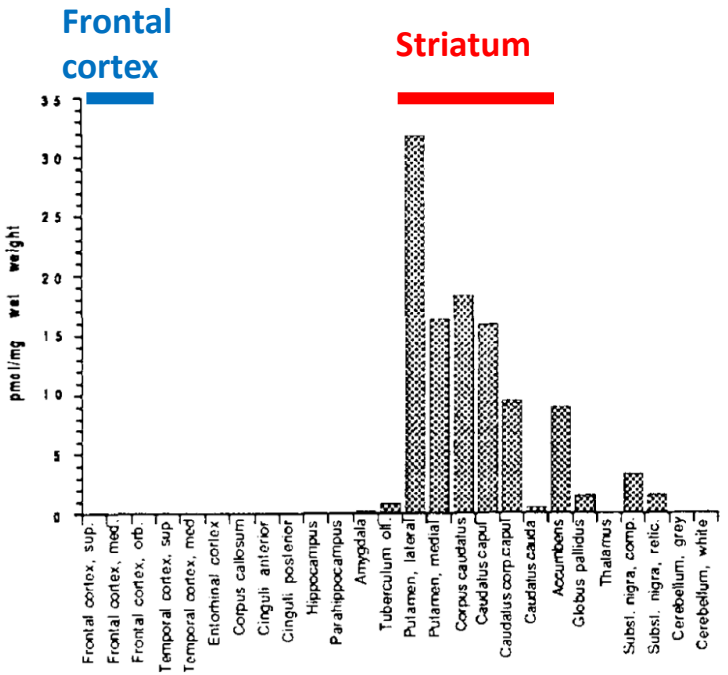
Arias-Carrión et al. 2010

Dopamine projections

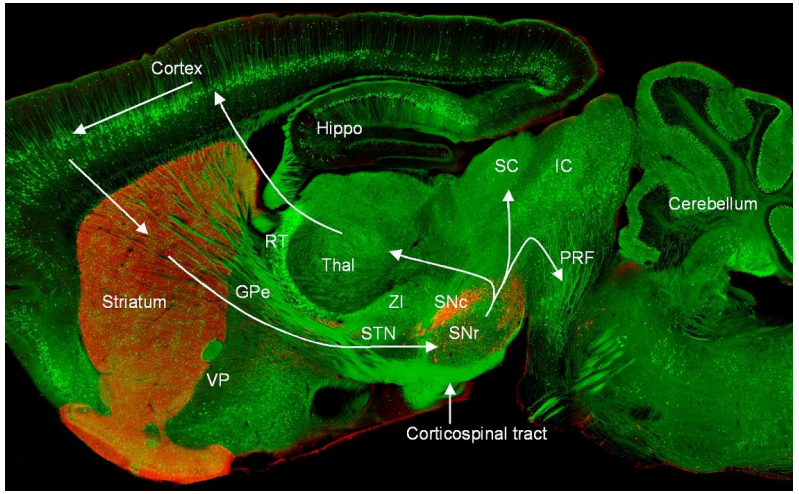
Human brain dopamine receptor density



Human brain dopamine concentration



Mouse brain TH stain (dopamine neurons)



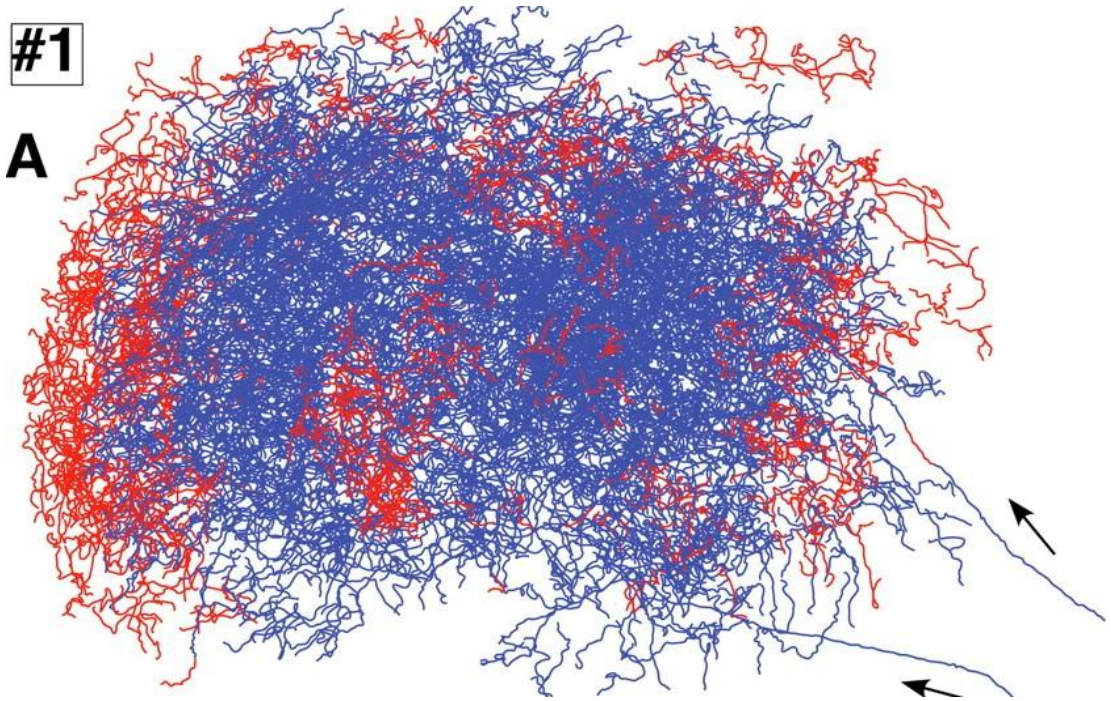
- Dopamine receptor density and dopamine neuron innervation are much greater in striatum than frontal cortex

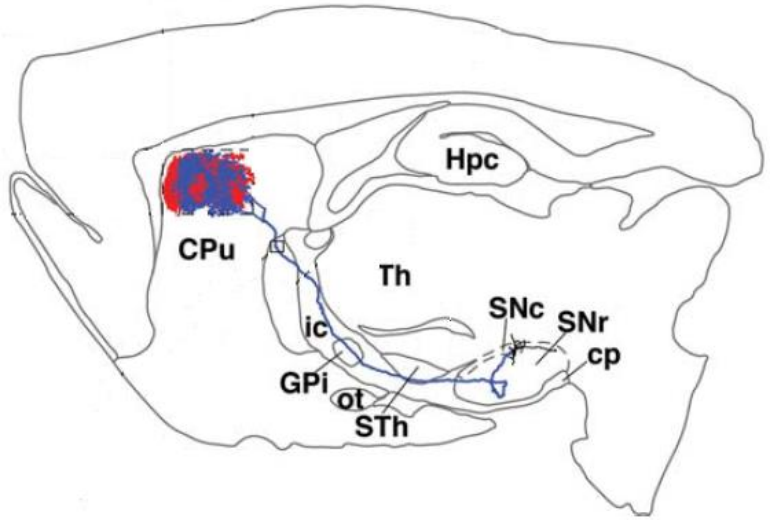
Dagher & Palomero-Gallagher 2020

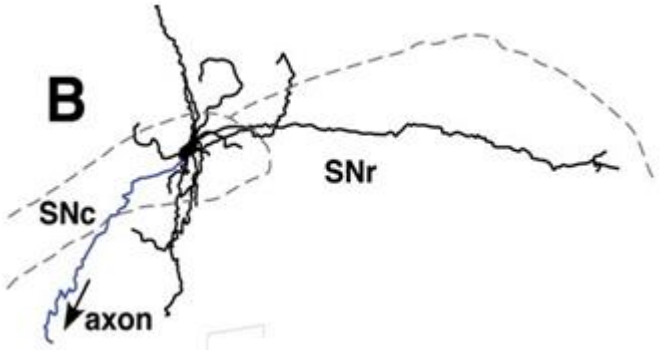
Hall et al. 1994

Zhou 2016

Dopamine projections





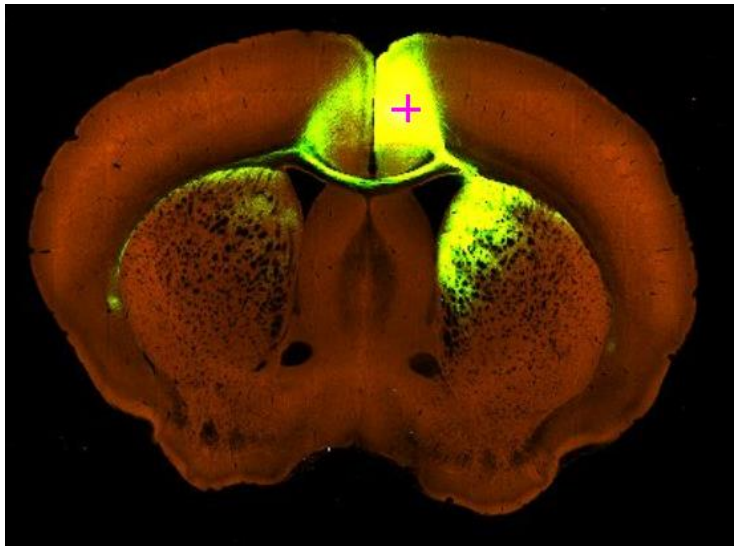


- Individual dopamine neurons densely innervate a large volume of striatum with 250,000 – 400,000 synapses per cell.

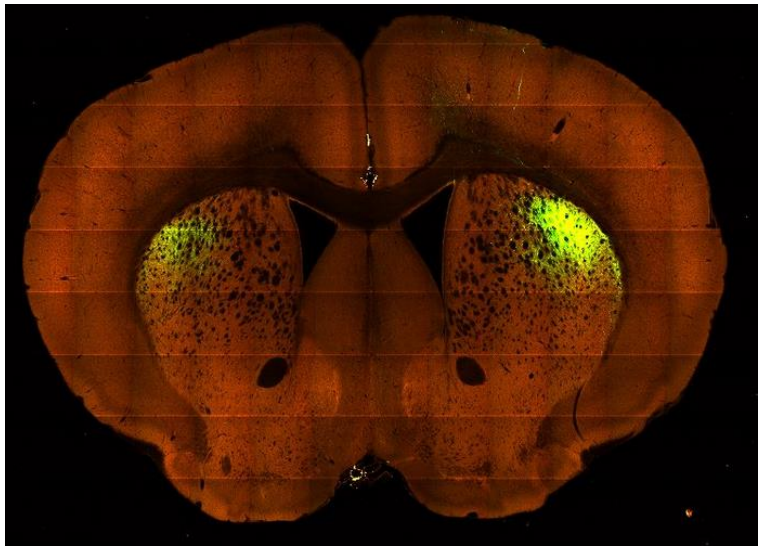
Matsuda et al. 2009

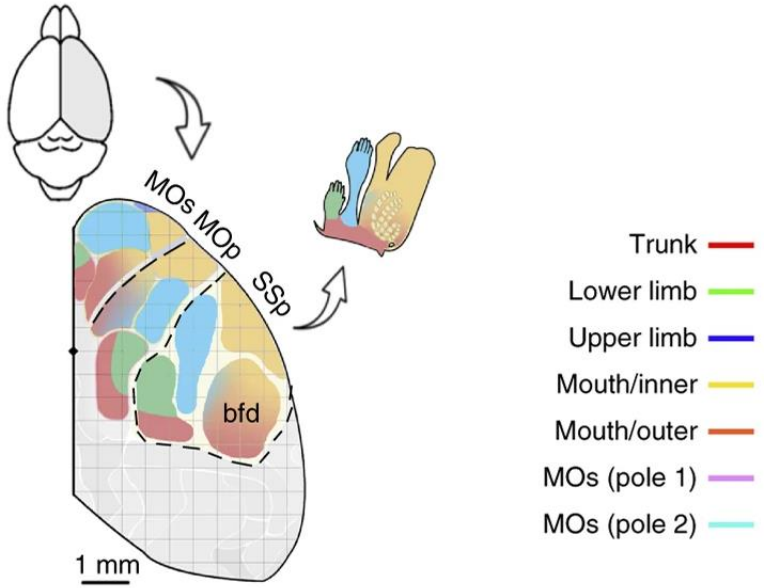
Cortical input to striatum

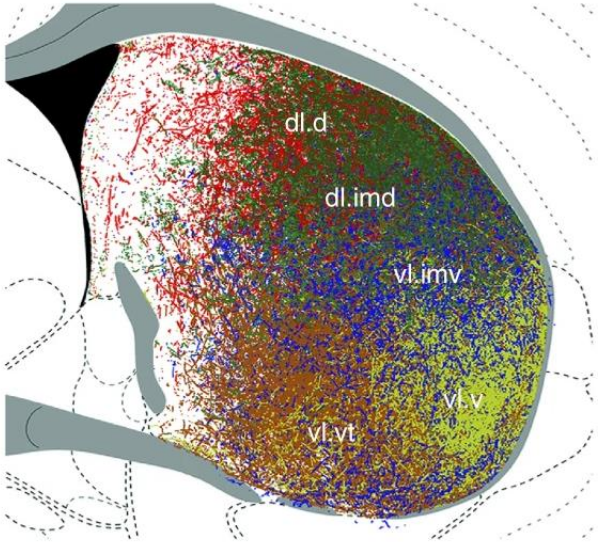
Anterior cingulate cortex → dorsomedial striatum



Primary somatosensory cortex → dorsolateral striatum





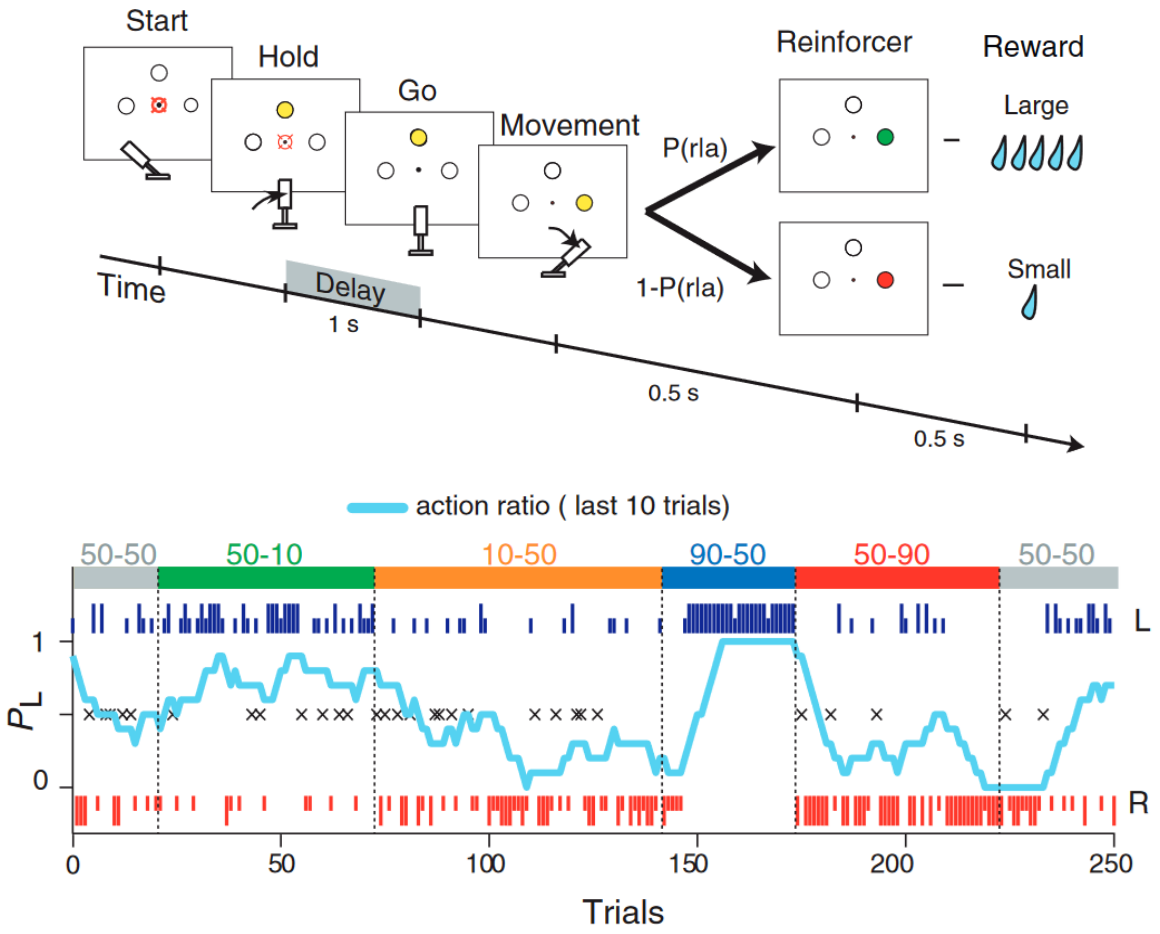


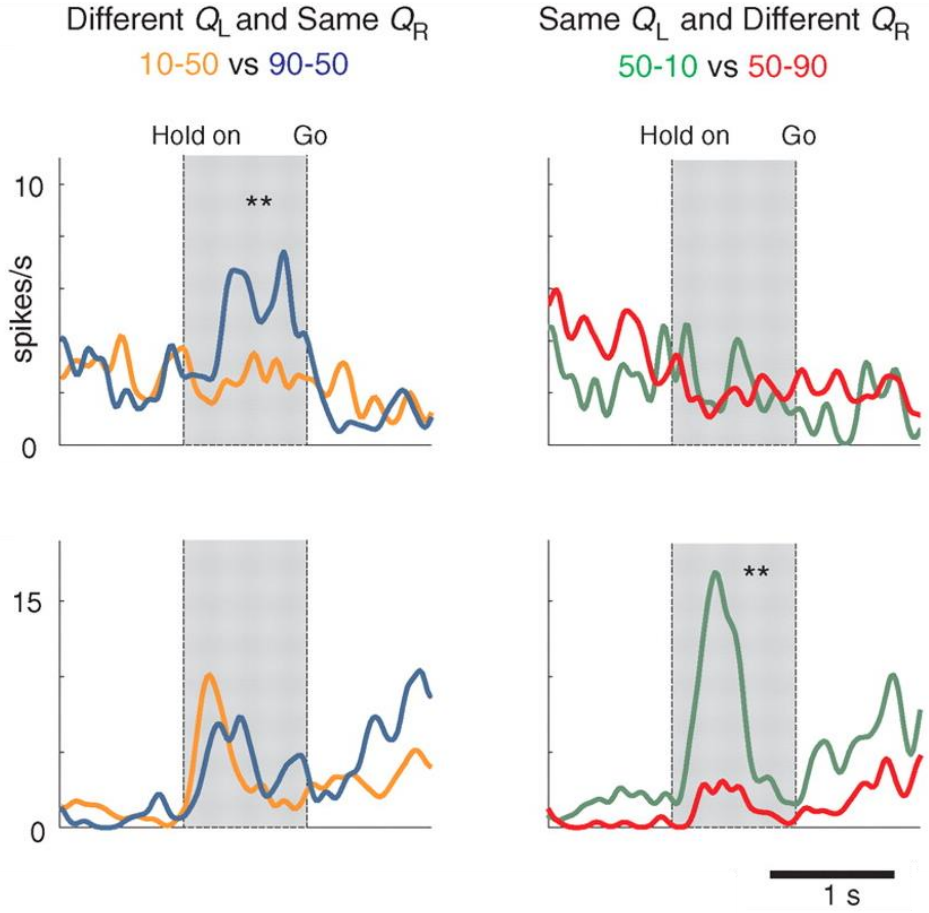
- Massive, highly structured, projection from cortex to striatum

Allen Brain Atlas

Hintiryan et al. 2016

Striatal action value signals

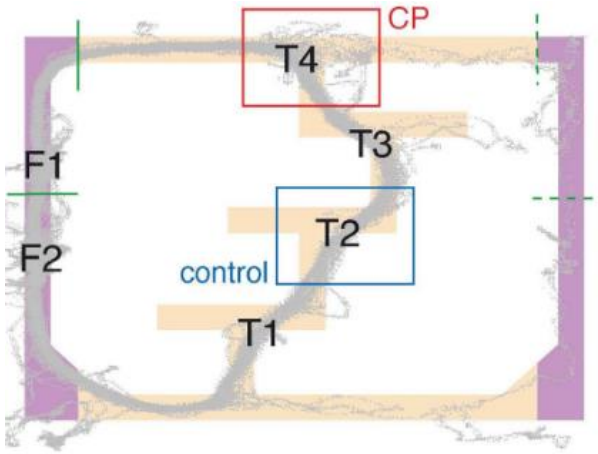


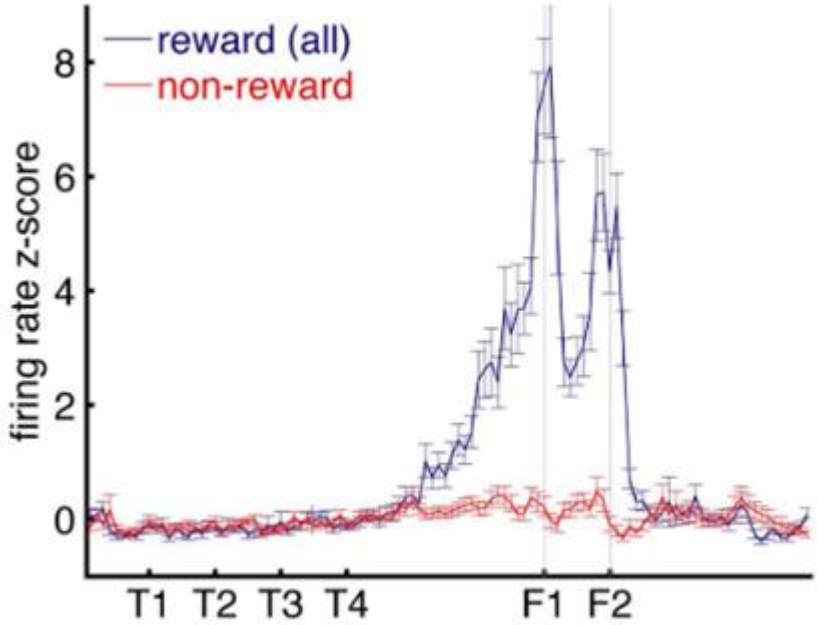


- Striatal neuron activity tracks the value of specific motor actions in a dynamic reward guided decision task

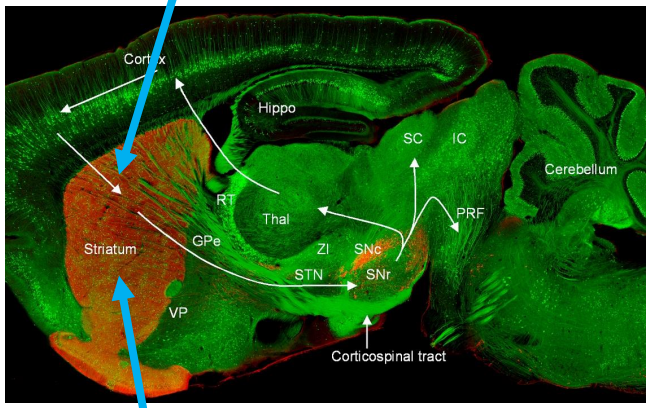
Samejima et al. 2005

State-value like signals in ventral striatum





Dorsal striatum

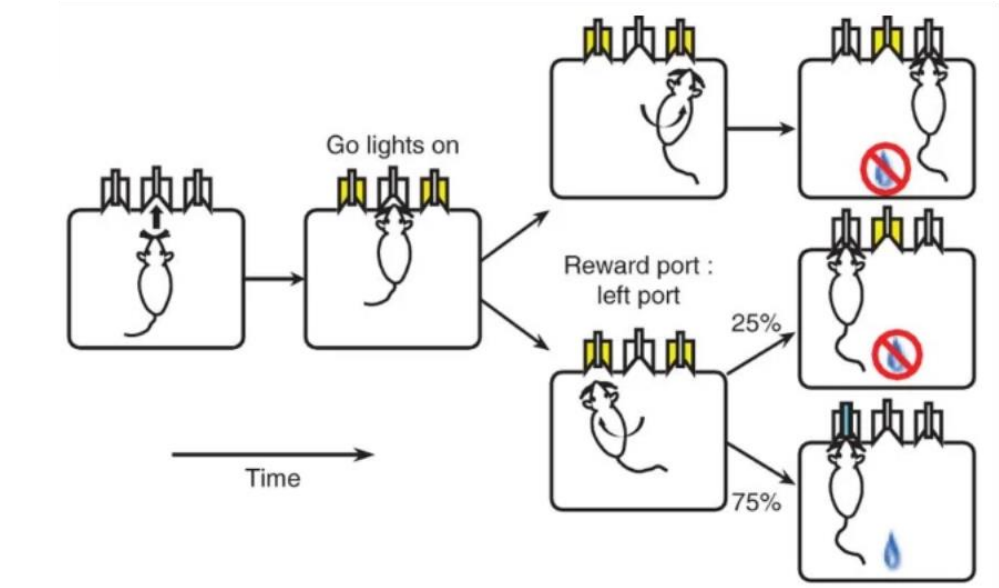


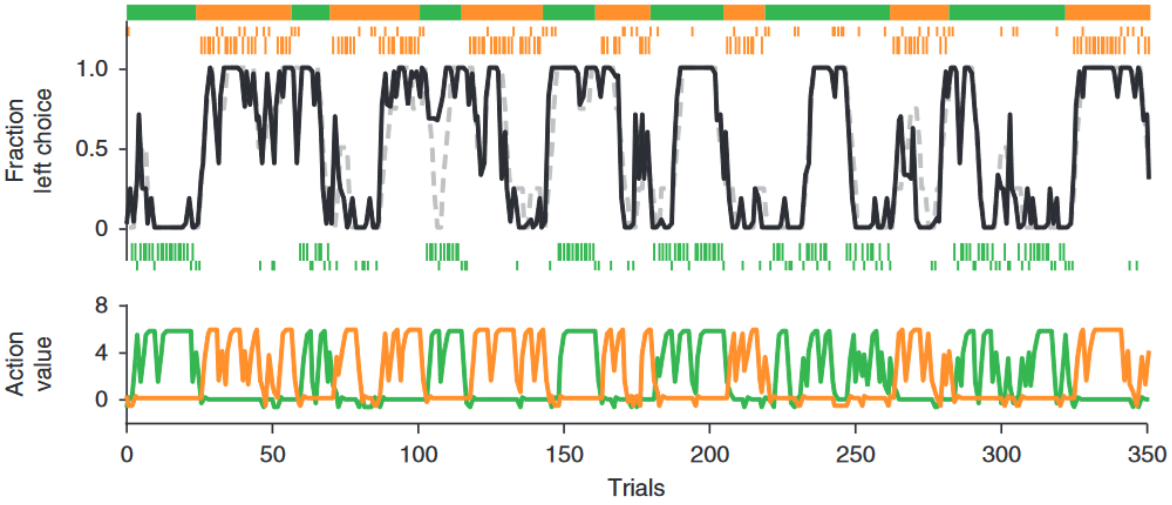
Ventral striatum

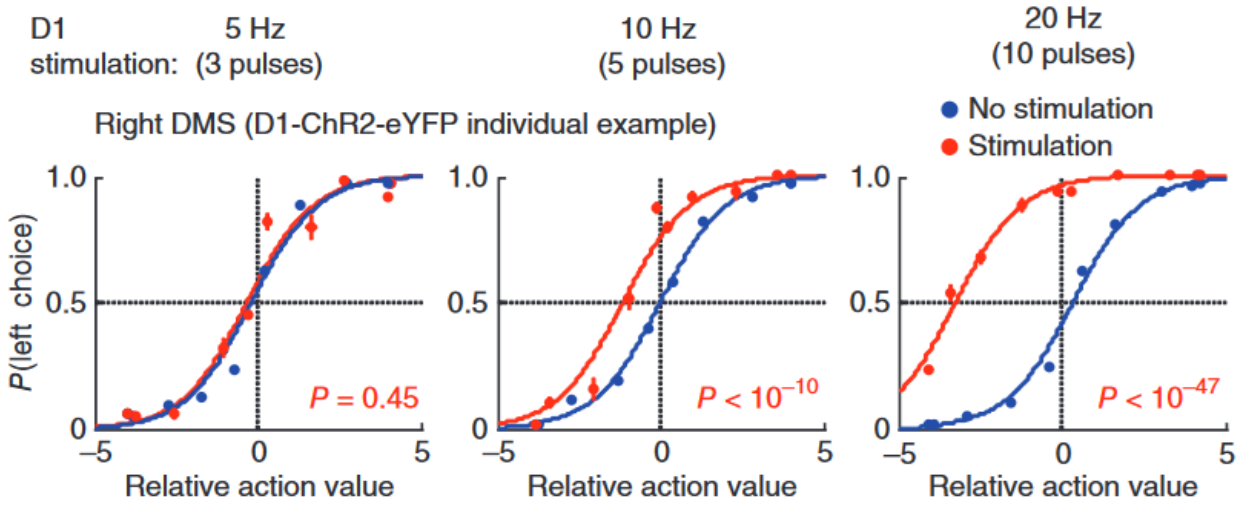
- Ramping activity in ventral striatum during approach to reward resembles state value.
- Differential involvement of ventral/dorsal striatum in state value / action value or policy also seen in human fMRI.

Van der Meer & Redish 2009

Striatal Causal manipulations



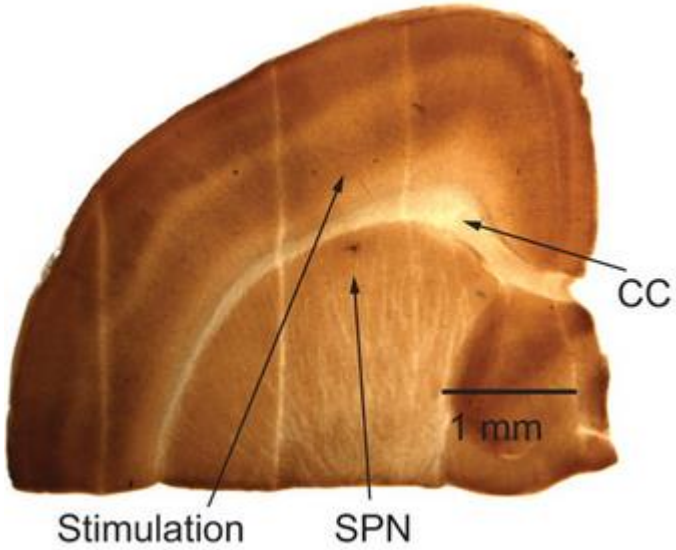


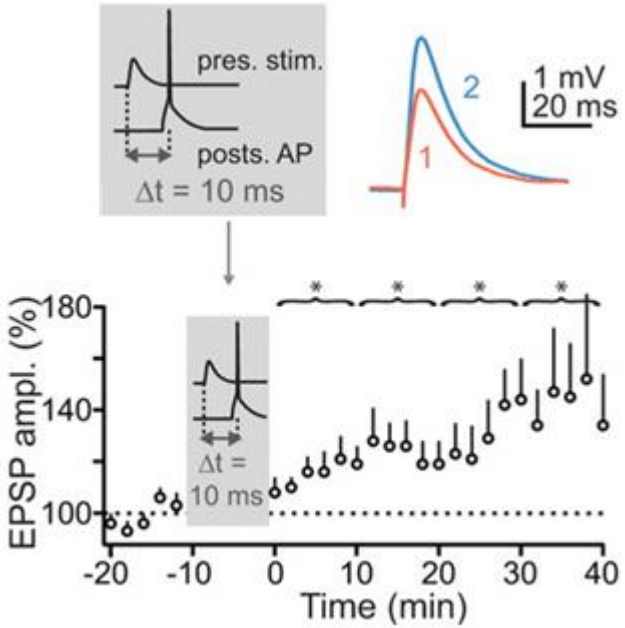


- Stimulation of dorsal striatal neurons biases action selection, consistent with modifying action values

Tai et al. 2012

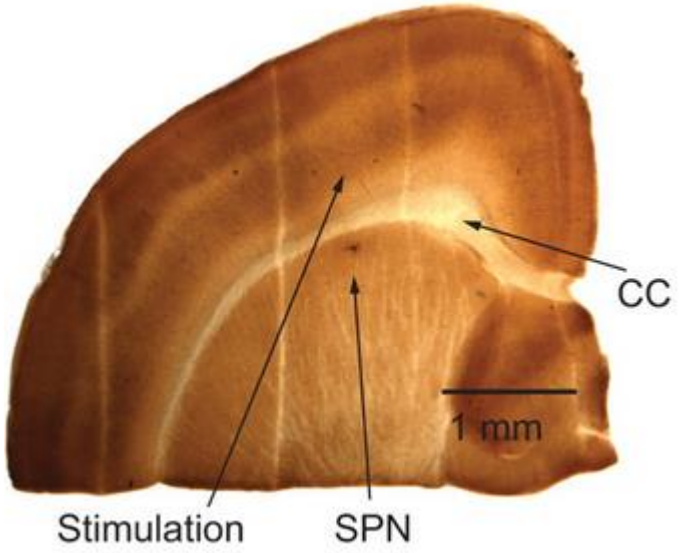
Dopaminergic modulation of cortico-striatal plasticity





[Pawlak & Kerr 2008]

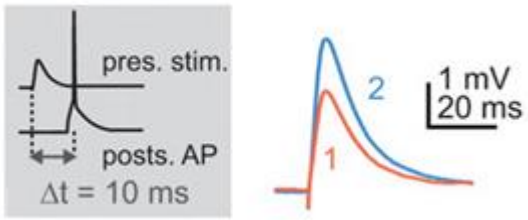
Dopaminergic modulation of cortico-striatal plasticity

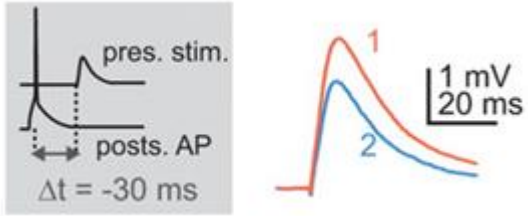


Cortex stim **before** striatum spike

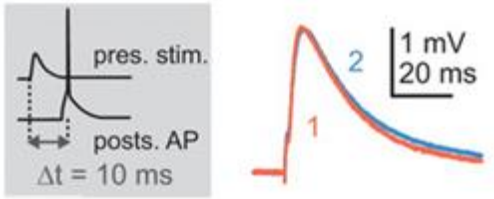
Cortex stim **after** striatum spike

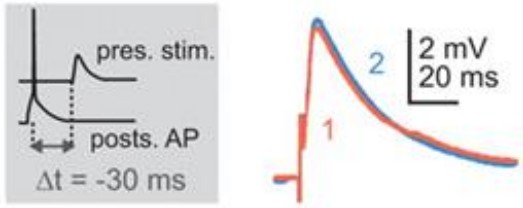
Control





Dopamine D1 receptors blocked





- Plasticity at cortico-striatal synapse depends on pre- and post- synaptic activity + dopamine signalling

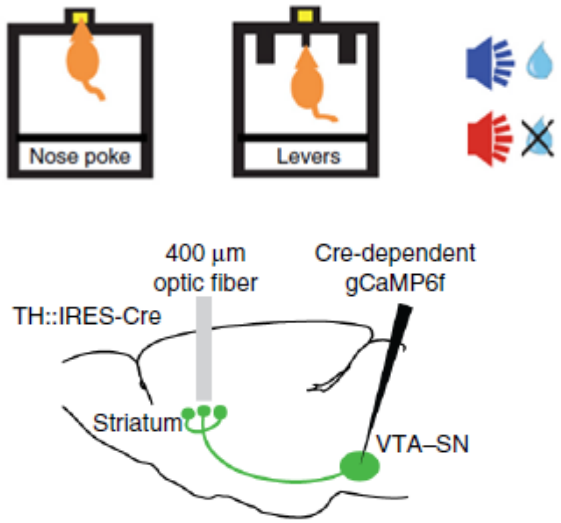
Pawlak & Kerr 2008

Interim summary:

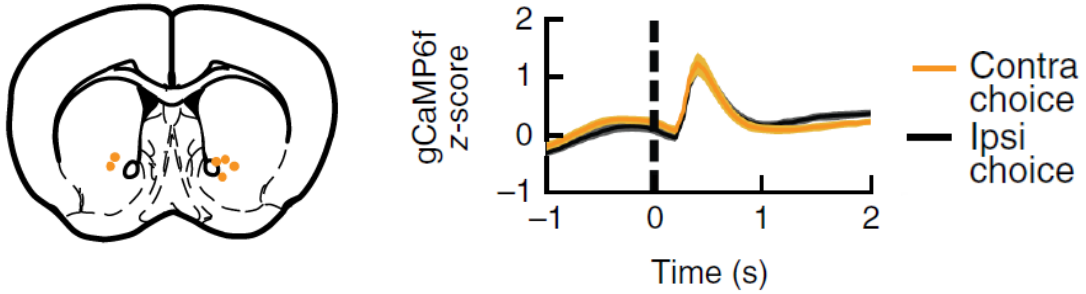
Findings broadly consistent with RPE theory:

- Dopamine activity resembles a TD RPE in many behavioural task and species.
- Dopamine manipulations can reinforce / supress behaviour.
- Value-like signals are observed in striatum, the primary target of the dopamine system.
- Dopamine potently modulates plasticity at cortico-striatal synapses

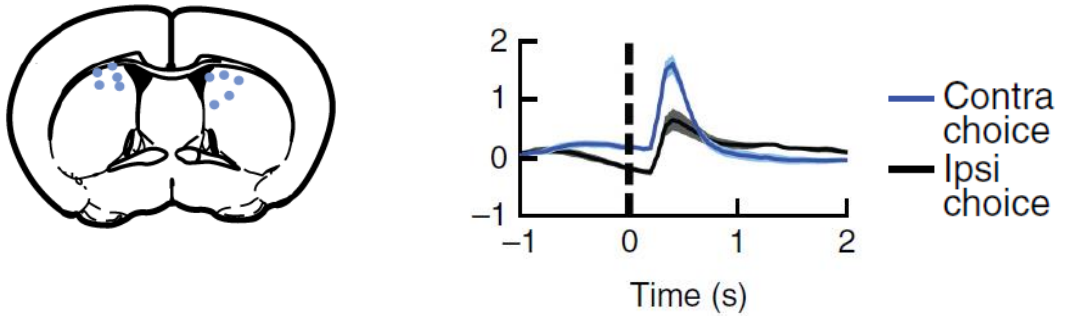
Heterogeneity of dopamine signals



Ventral striatum



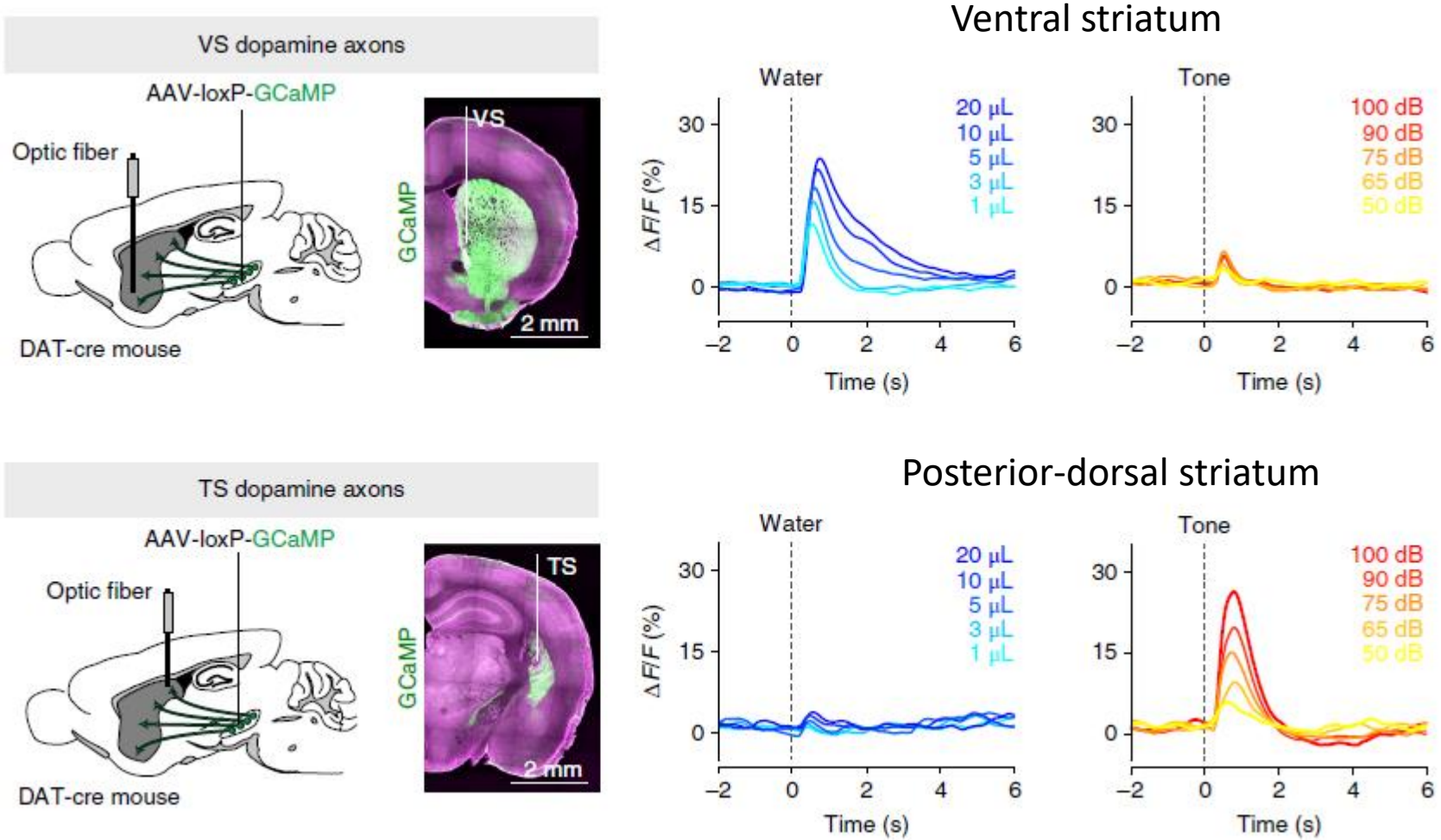
Dorsomedial striatum



- Dopamine neuron terminals in dorsomedial striatum but not ventral striatum respond to contralateral choices.

Parker et al. 2016

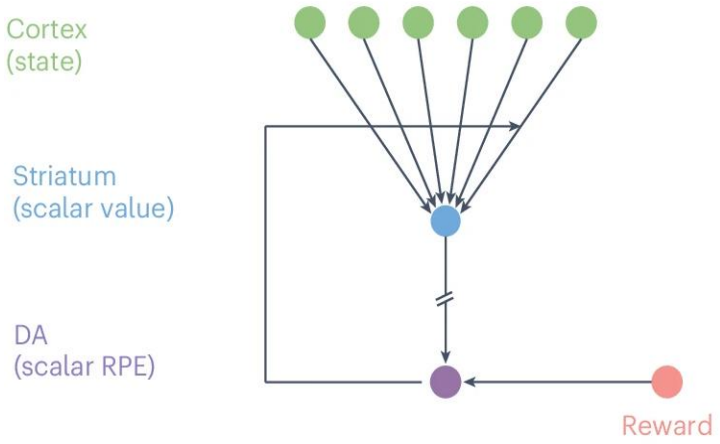
Heterogeneity of dopamine signals



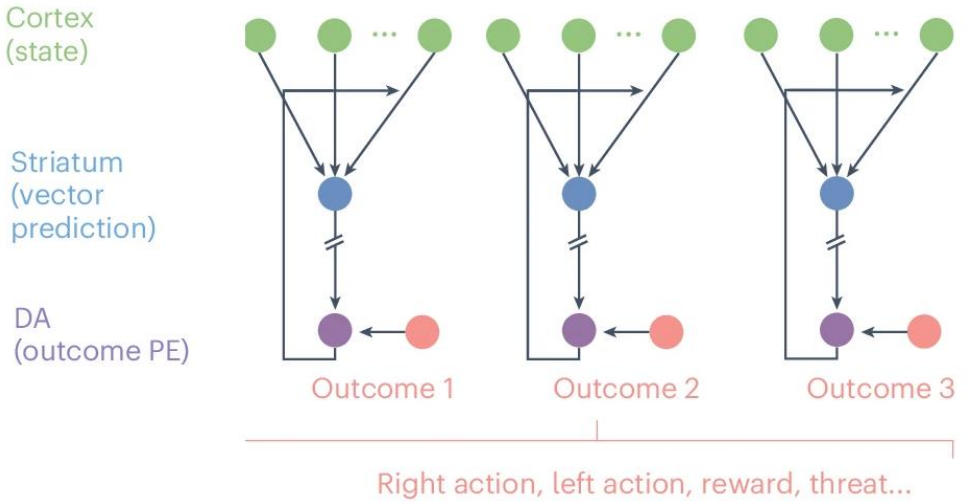
- Dopamine neuron axons in the posterior striatum respond to aversive load noise and not to rewards.

Heterogeneity of dopamine signals

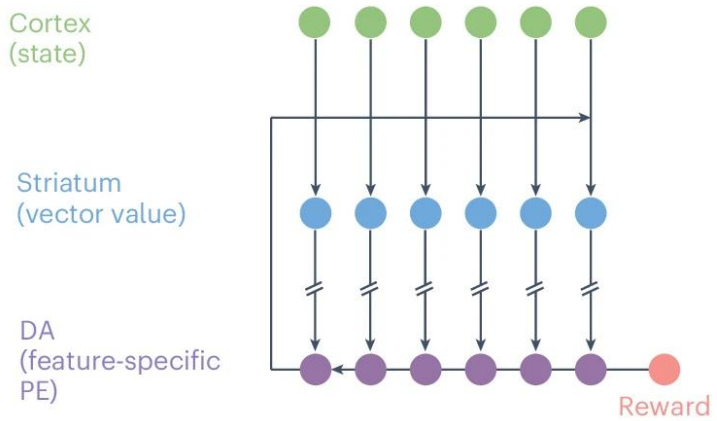
Classic scalar RPE model



Outcome specific PE model



Feature specific PE model



- Heterogeneity is inconsistent with a scalar RPE signal.

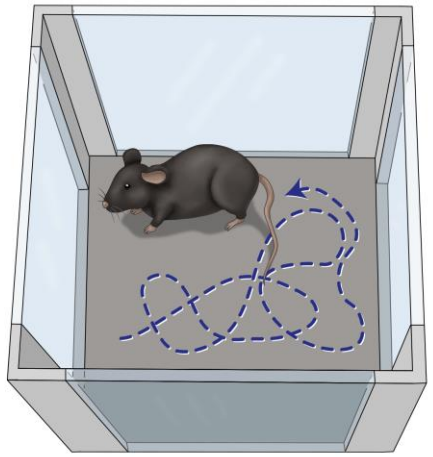
Possible reasons for heterogeneity:

- Outcome specific PE: Prediction of different outcomes (e.g. reward, threat, action) by different cortico-basal ganglia loops.

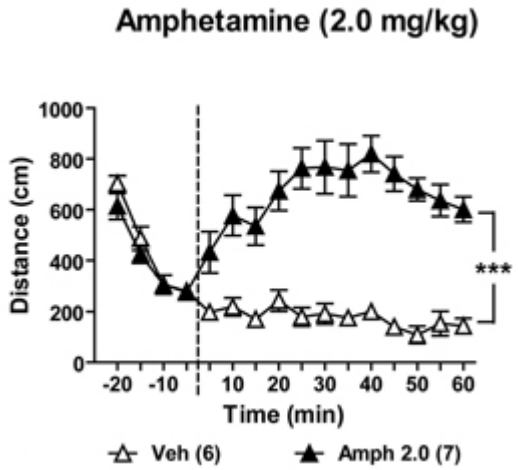
- Feature specific PE: Different cortico-basal ganglia loops predict reward using different state features.

Lee et al. 2024

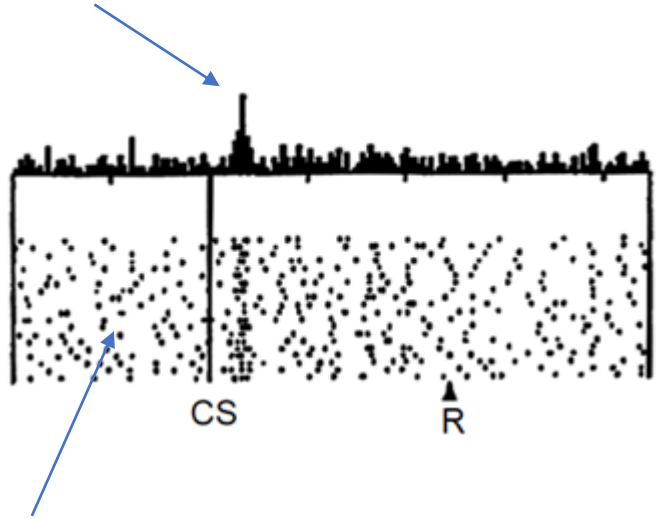
Dopamine, vigor, and reward rate



© Keri Jones



"phasic" burst of spikes in response to cue



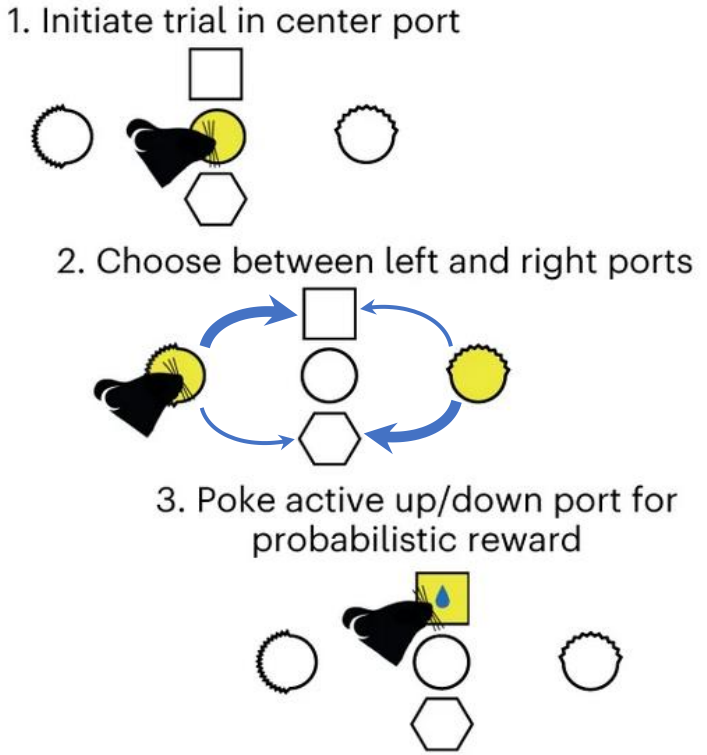
"tonic" activity at low firing rate in absence of external events

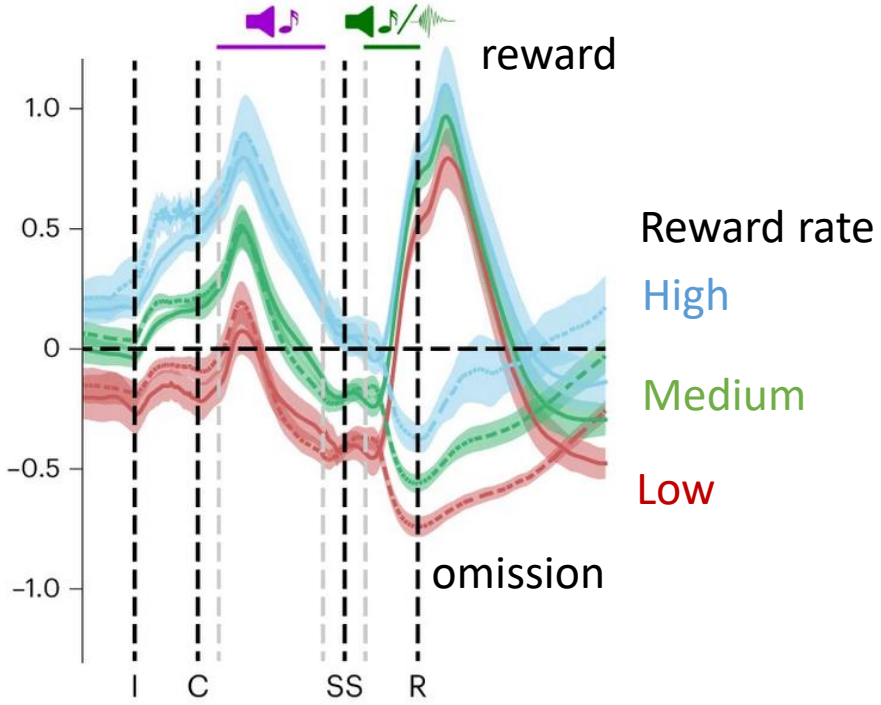
- Pharmacologically boosting dopamine levels induces behavioural activation and increased response vigor.
- Rapid time-course appears inconsistent with RPE-driven learning.
- Niv et al. (2007) propose that "tonic" and "phasic" activity as separate information channels:
  - "Phasic" bursts: RPE → learning
  - "tonic" activity: average reward rate → opportunity cost of time that controls motivation/vigor

Yen et al. 2013

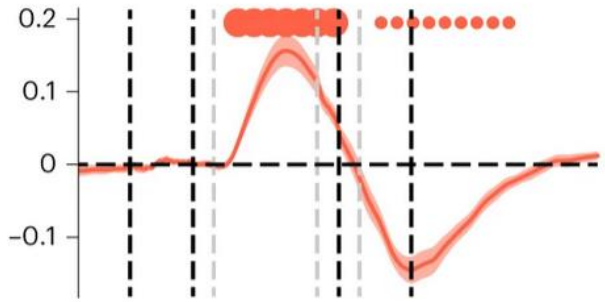
Schultz, Dayan, Montague, 1997

Dopamine, vigor, and reward rate

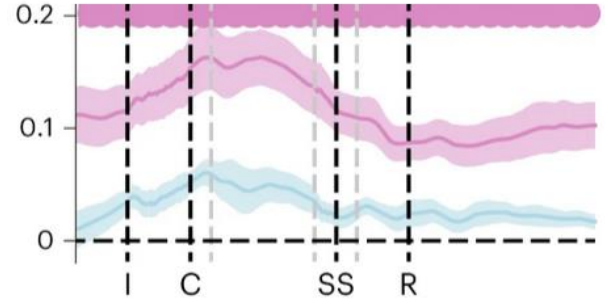




Value of up/down port



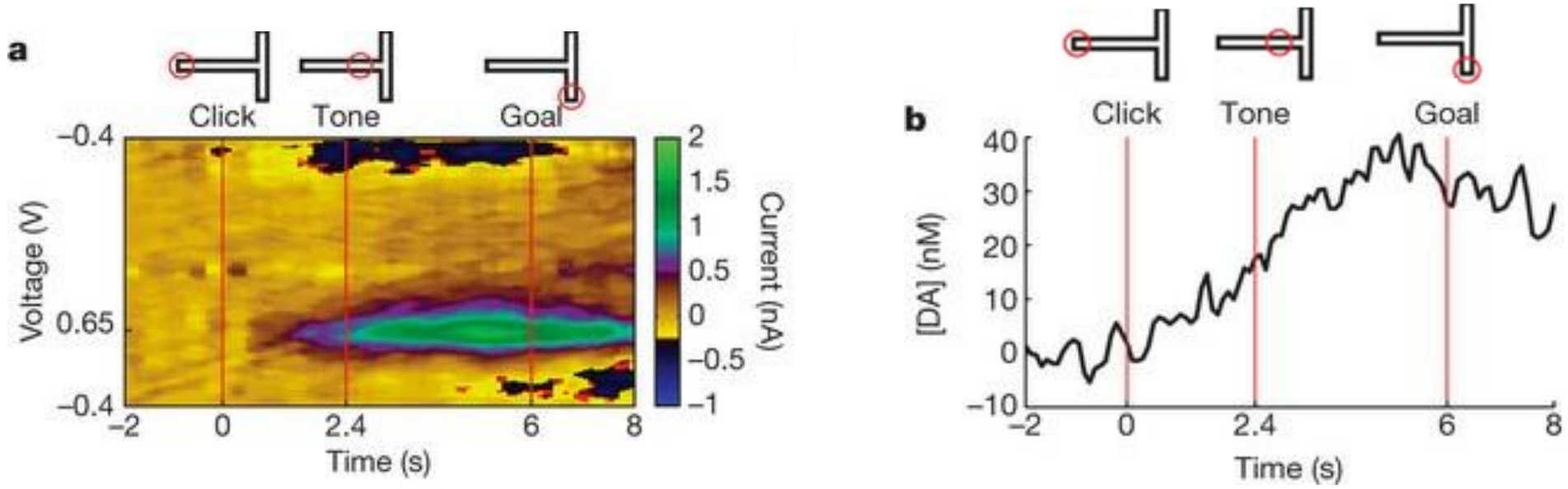
Average reward rate



- Average reward rate drives sustained changes in dopamine neuron activity, dissociable from fast phasic RPE responses.

Blanco-Pozo et al. 2024

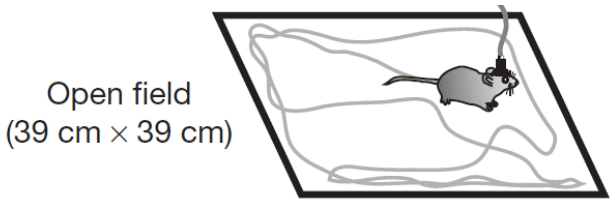
Dopamine ramps

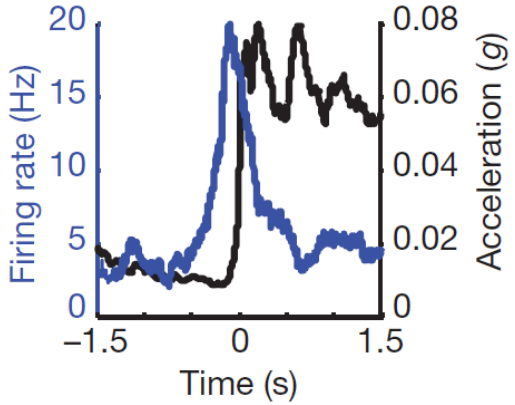


- Striatal dopamine during navigation to reward show ramping that resembles value more than RPE.
- Many theoretical accounts…

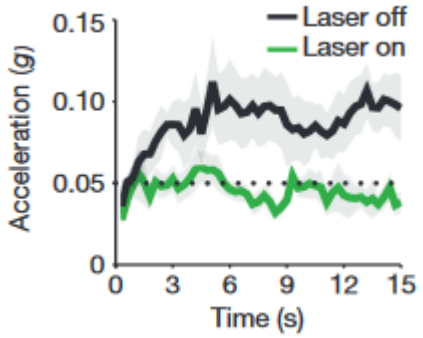
Howe et al. 2013

Dopamine and movement initiation

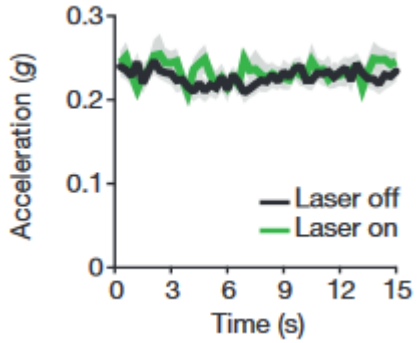




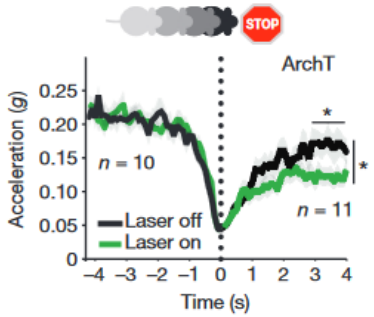
Stationary



Moving



Stopping



- SNc dopamine neurons are modulated before spontaneous movement initiation.
- Inhibiting SNc supresses movement initiation but does not stop stopping ongoing movement.

Alves da Silva et al. 2018

Dopamine and RPE summary

Findings broadly consistent with RPE theory:

- Dopamine activity resembles a TD RPE in many behavioural task.
- Dopamine manipulations can reinforce / supress behaviour.
- Value-like signals are observed in striatum, the primary target of the dopamine system.
- Dopamine potently modulates plasticity at cortico-striatal synapses

Findings not straightforwardly predicted by RPE theory:

- Direct effects on motivation/vigor
- Involvement in action initiation
- Ramping ‘value-like’ activity.

Talk aim and overview

Question: How do learning algorithms map onto brain structure to solve the biological action control problem?

Talk outline:

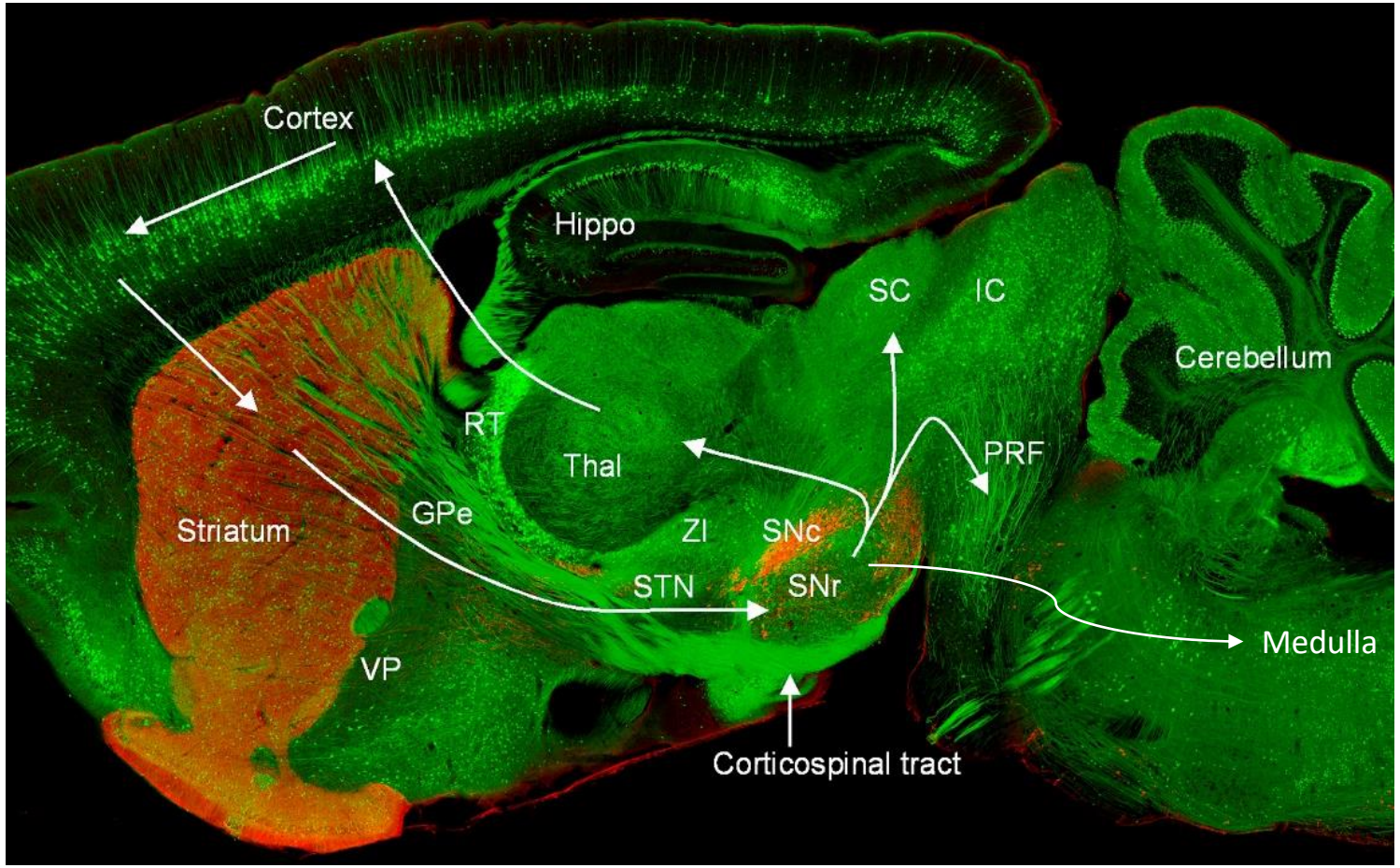
1. Striatum and dopamine: The brain's temporal difference reinforcement learning system?

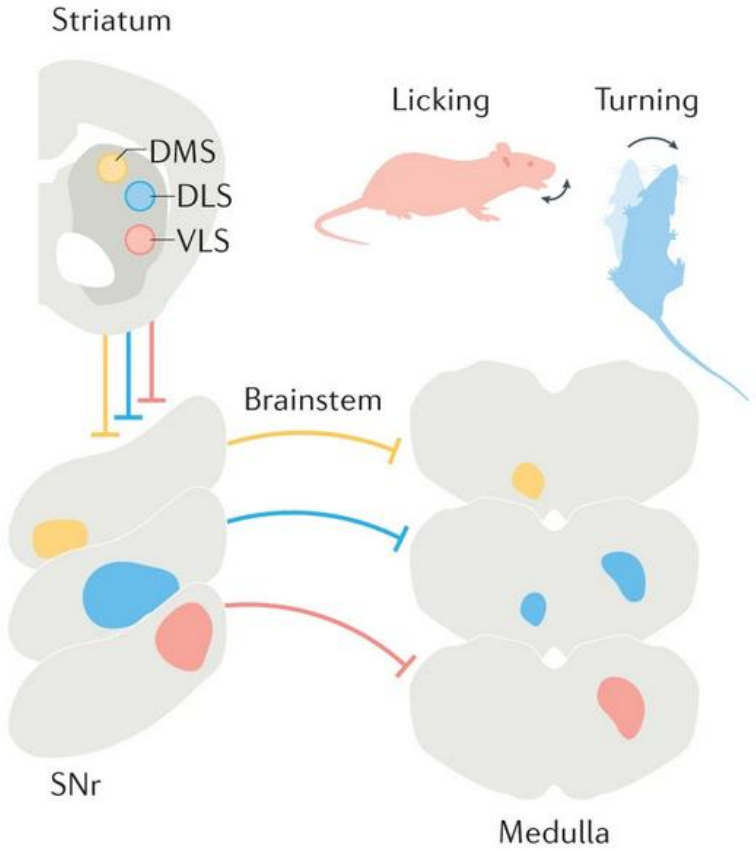
2. Basal ganglia outputs and control of action selection.

3. Cortex: State representation and beyond

4. Hippocampus: Sequence generation and model-based control

Basal ganglia outputs to brainstem for control of movement



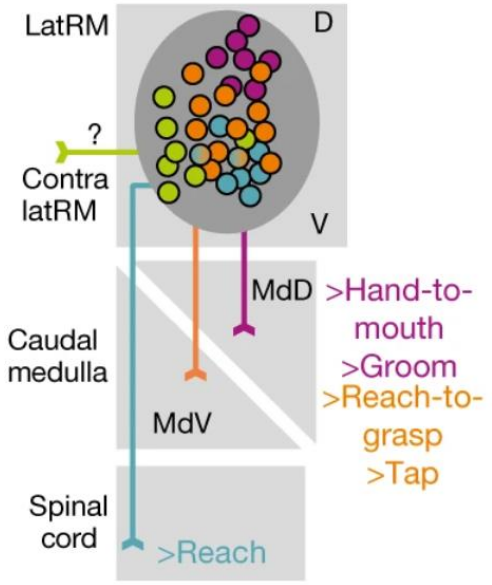


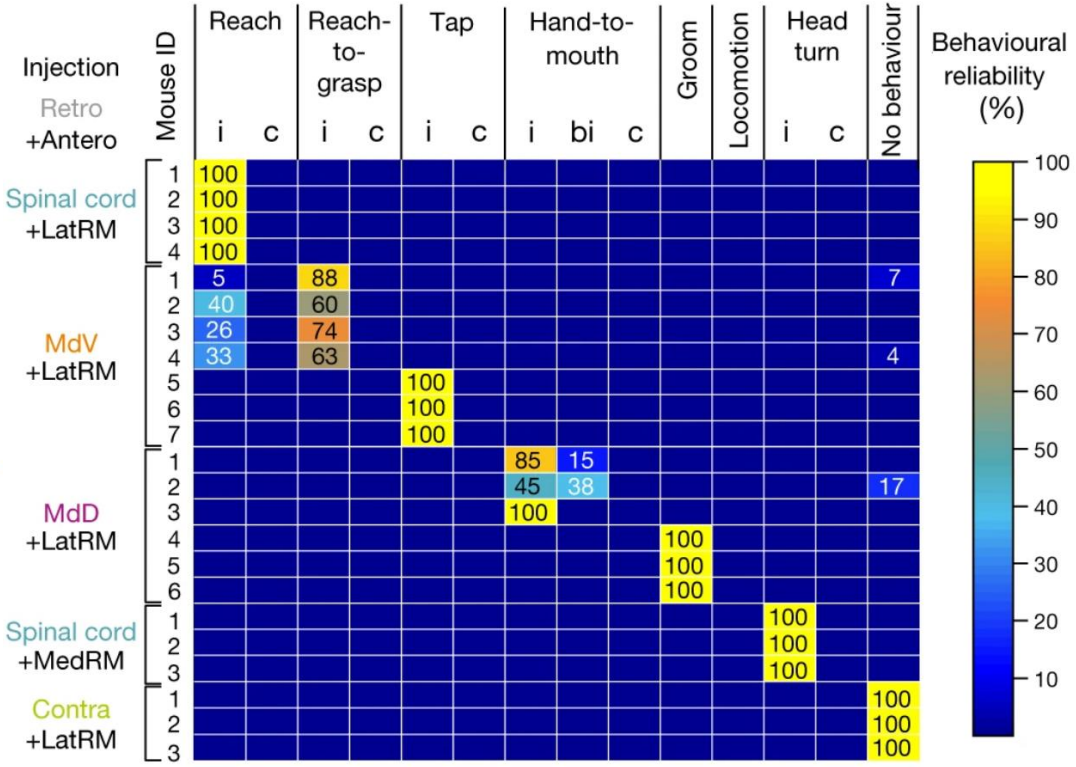
- Basal ganglia projects to motor control nuclei in the brainstem: Striatum → SNr → Brainstem

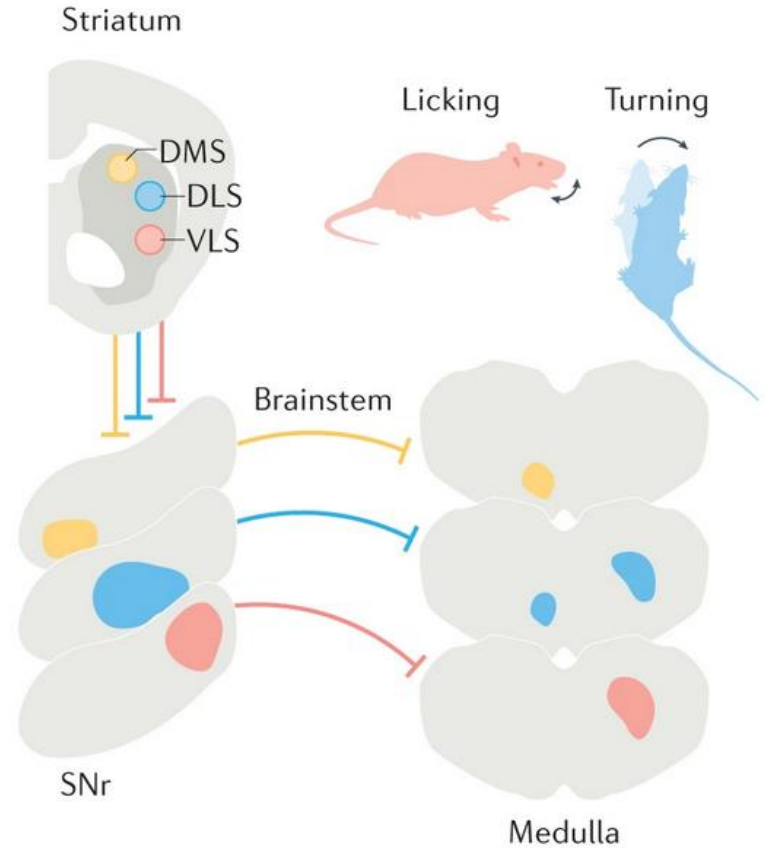
- Output to brainstem is topographically organised into parallel pathways

Arber & Costa 2022

Basal ganglia outputs to brainstem for control of movement







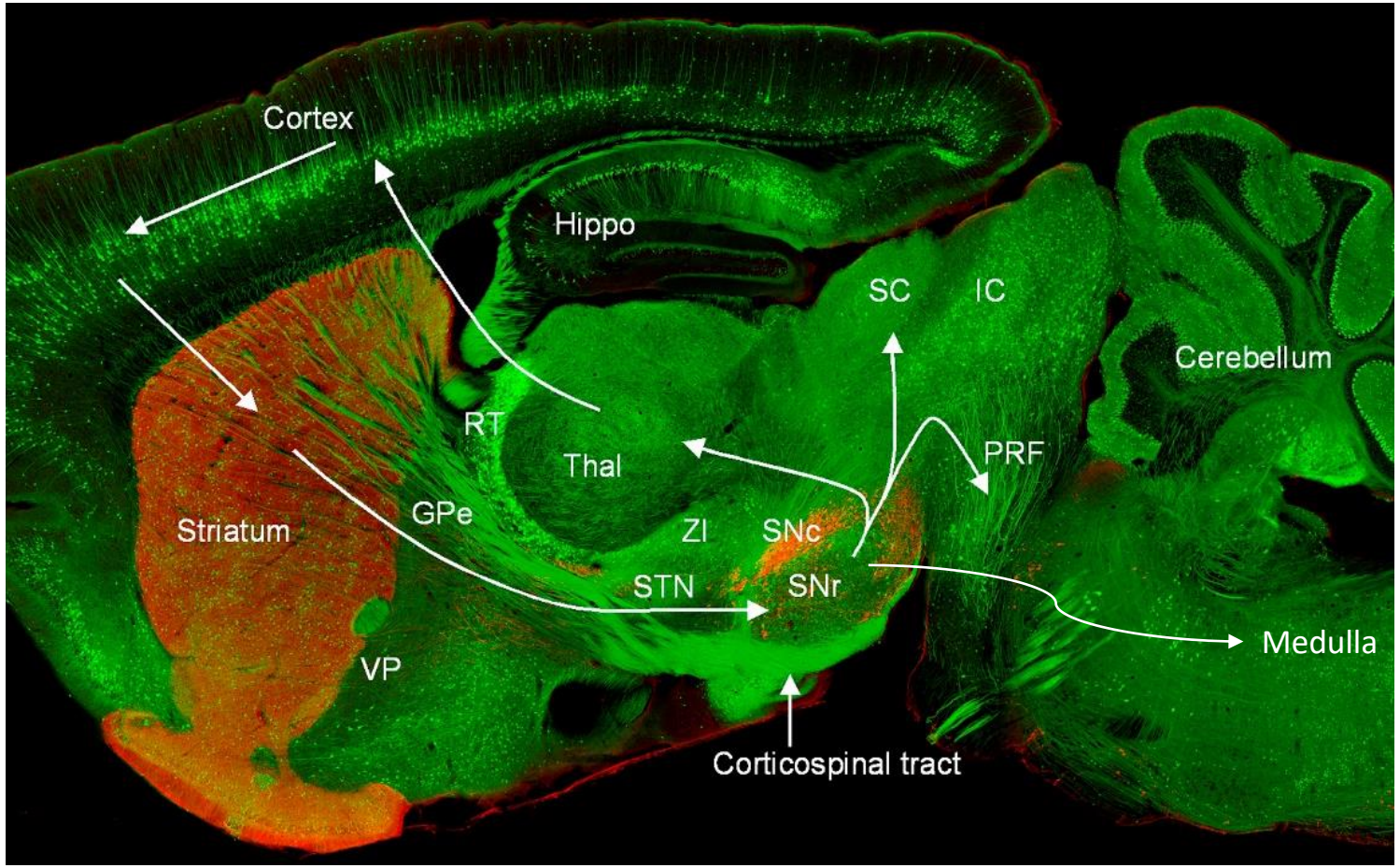
- Different populations of brainstem neurons control distinct motor actions

→ Striatum is upstream of motor control nuclei consistent with striatal value / policy signals controlling action

Ruder et al. 2021

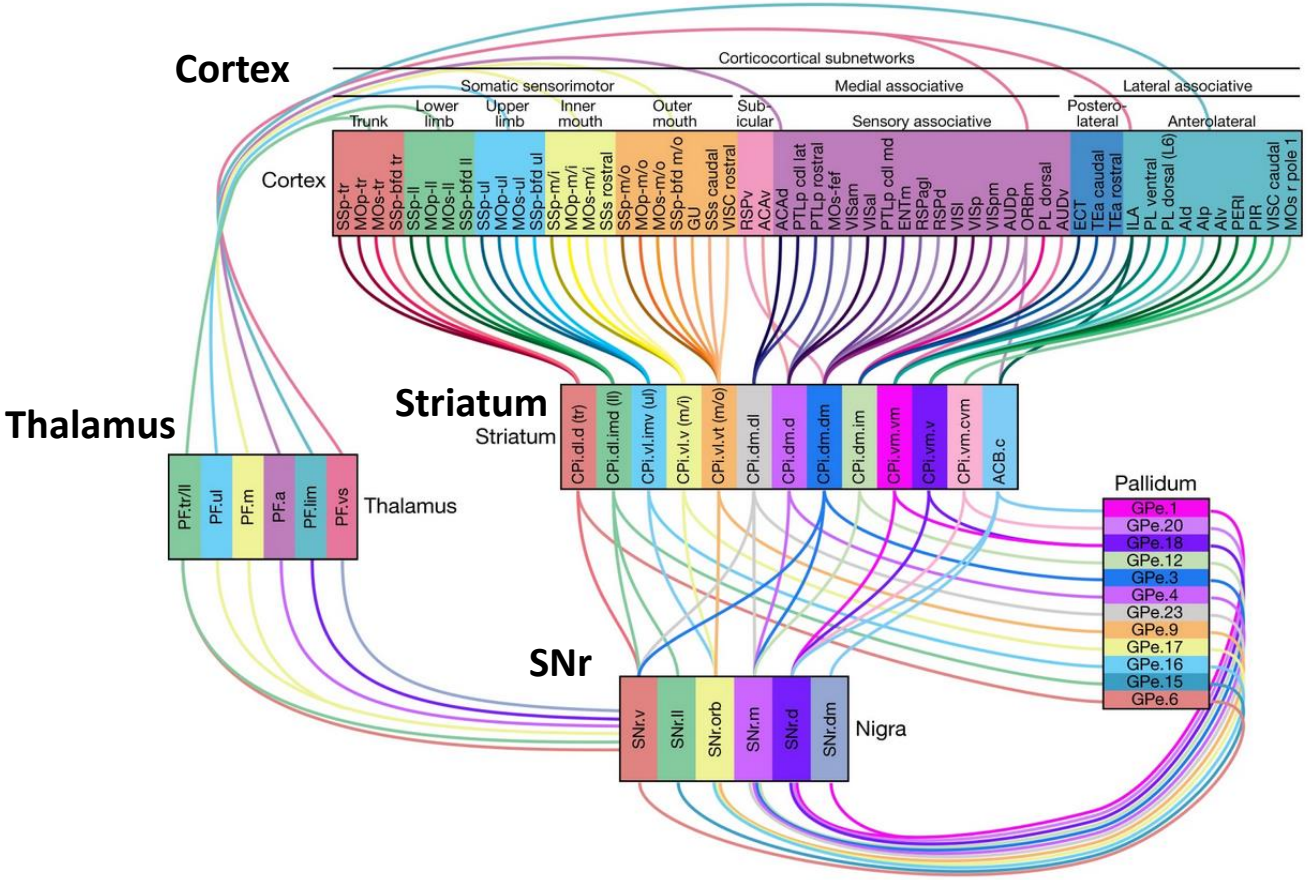
Arber & Costa 2022

Basal ganglia outputs: Projections to thalamus and control of cortical activity



- Basal ganglia output neurons send collaterals to thalamus which has an excitatory projection to cortex

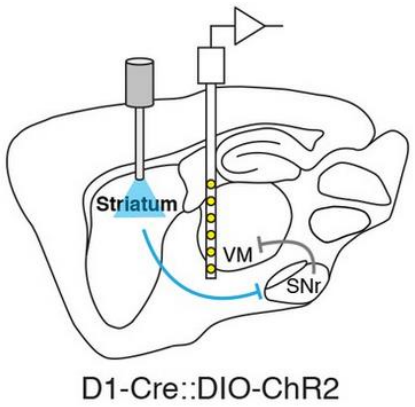
Basal ganglia outputs: Projections to thalamus and control of cortical activity



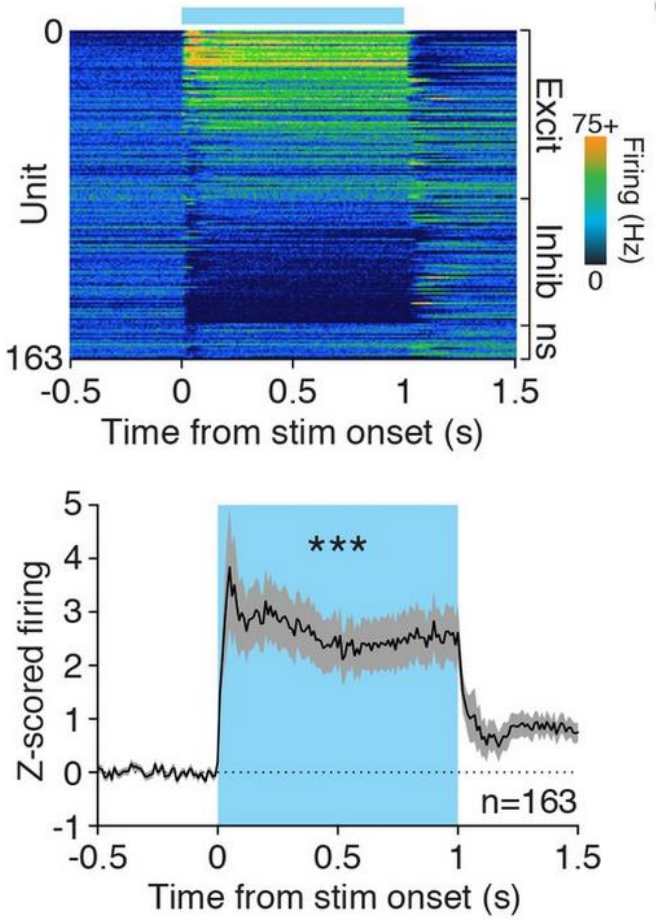
- Basal ganglia output neurons send collaterals to thalamus which has an excitatory projection to cortex
- Parallel loops architecture maps BG outputs channels back to the cortical inputs that drive them

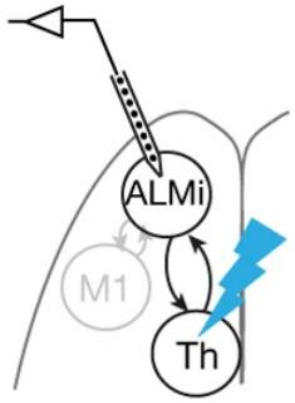
[Foster et al. 2021]

Basal ganglia outputs: Projections to thalamus and control of cortical activity

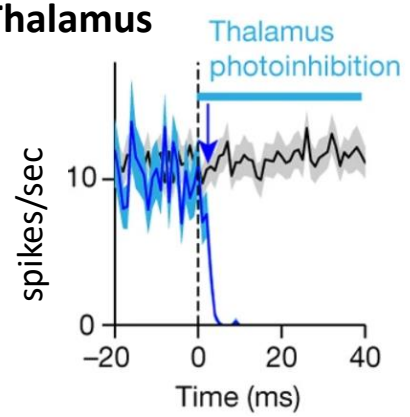


D1-Cre::DIO-ChR2

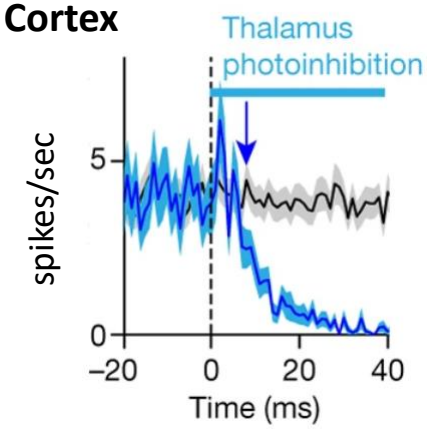




Thalamus



Cortex

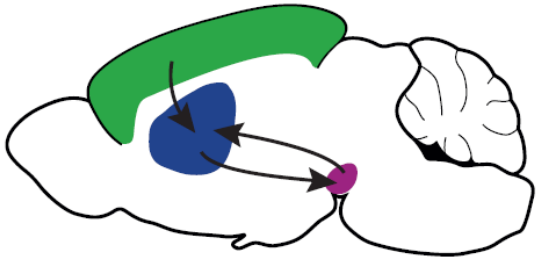


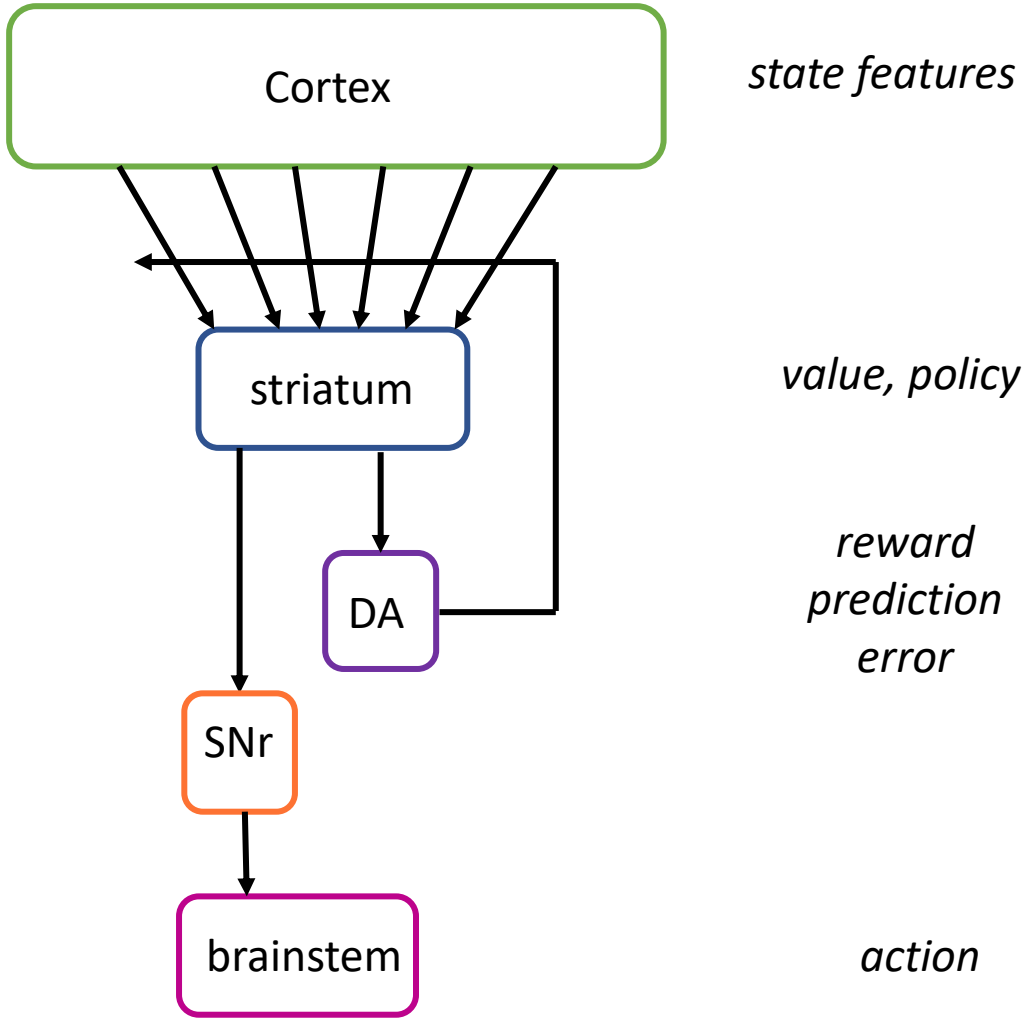
- Basal ganglia output neurons send collaterals to thalamus which has an excitatory projection to cortex
- Parallel loops architecture maps BG outputs channels back to the cortical inputs that drive them
- Striatum can potently modulate thalamic activity and hence cortical activity

Lalive et al. 2018

Guo et al. 2017

Where are we at with the boxes and arrows..





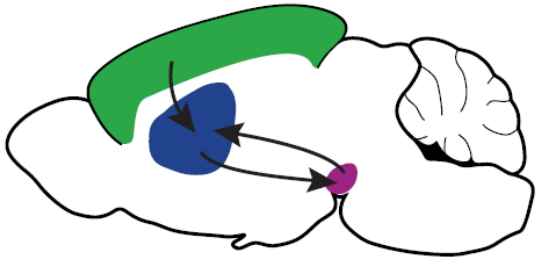
state features

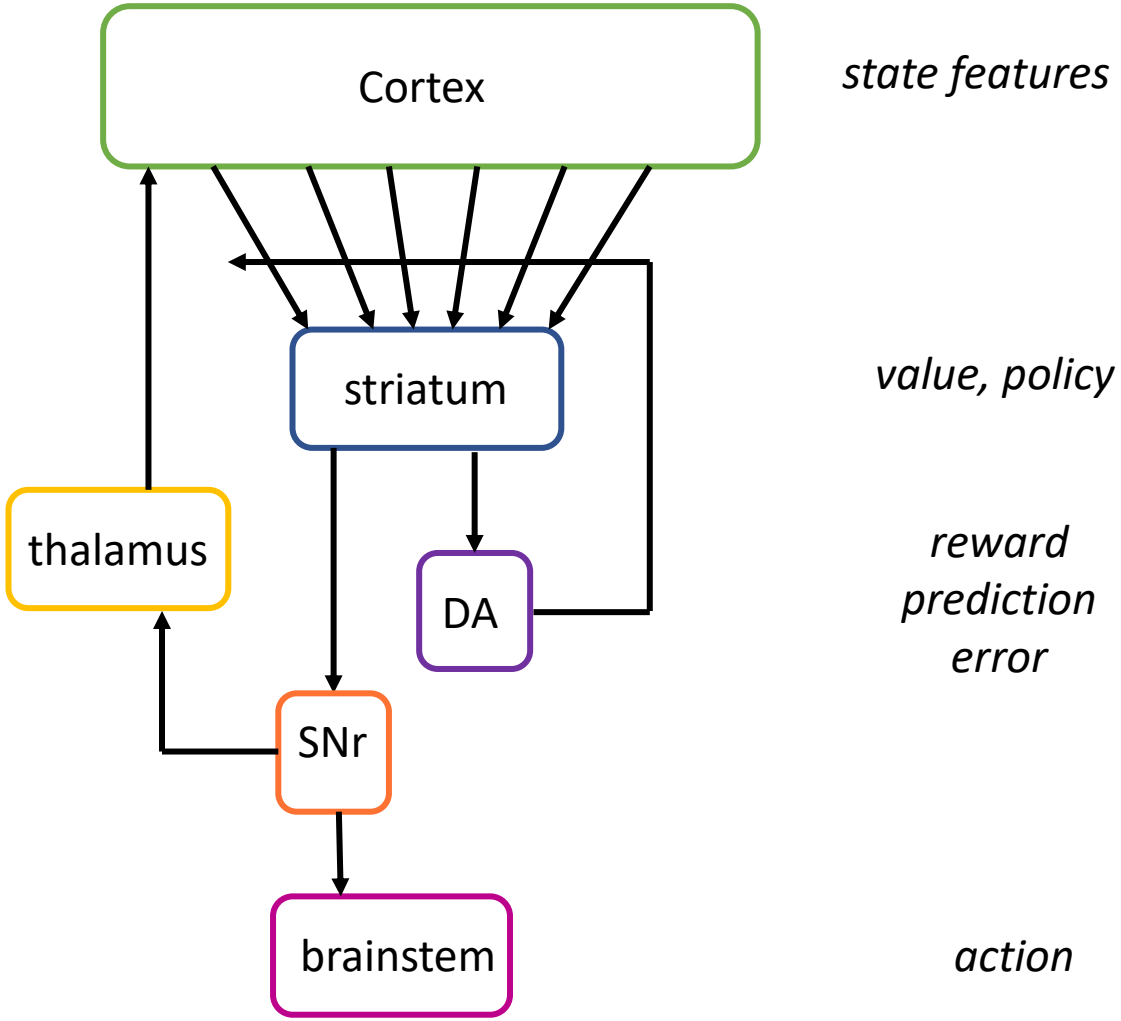
value, policy

reward prediction error

action

Where are we at with the boxes and arrows..





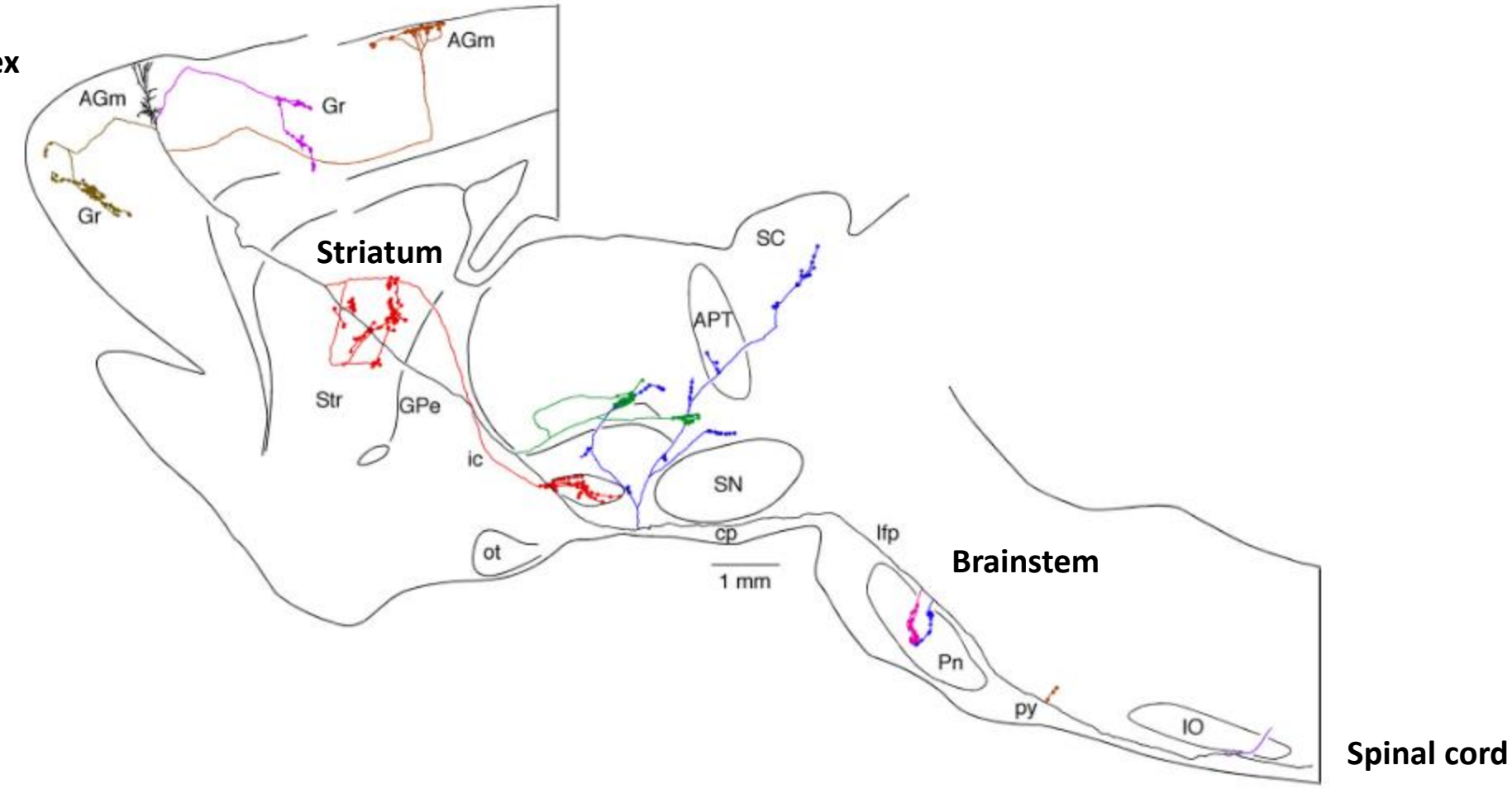
state features

value, policy

reward prediction error

action

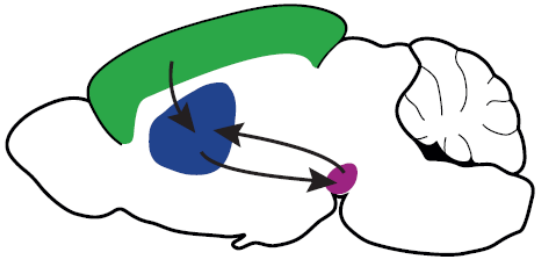
Further complications

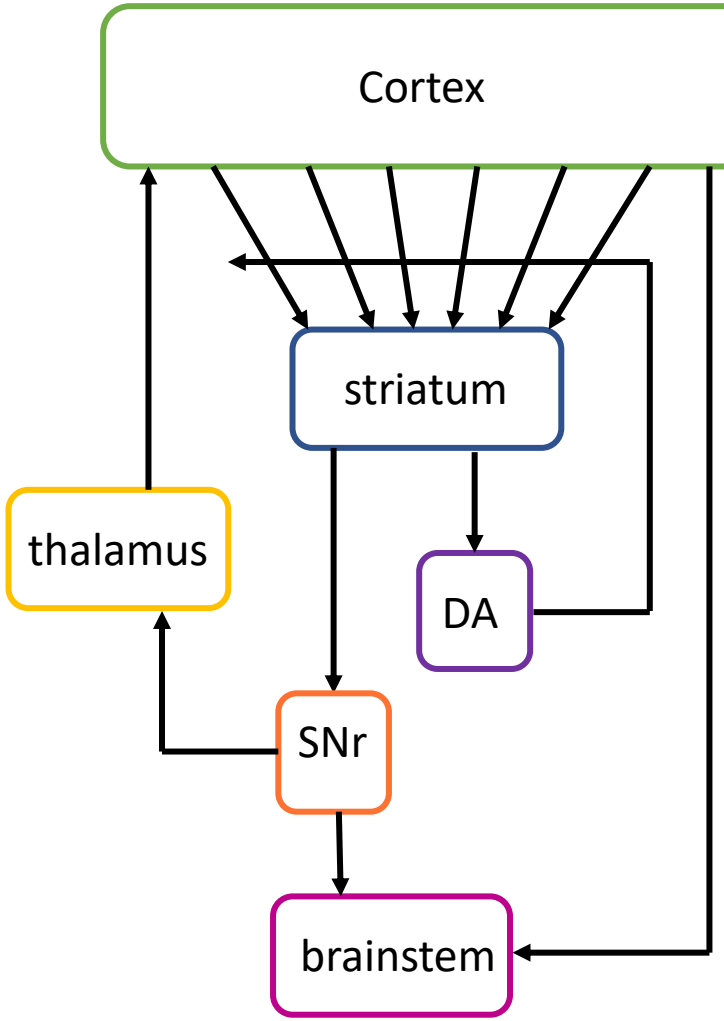


- Frontal and motor cortex make direct projections to brainstem and spinal cord.

Kita & Kita 2012

It’s complicated





state features/action/value/policy???

value, policy

reward prediction error

action

Talk aim and overview

Question: How do learning algorithms map onto brain structure to solve the biological action control problem?

Talk outline:

1. Striatum and dopamine: The brain's temporal difference reinforcement learning system?

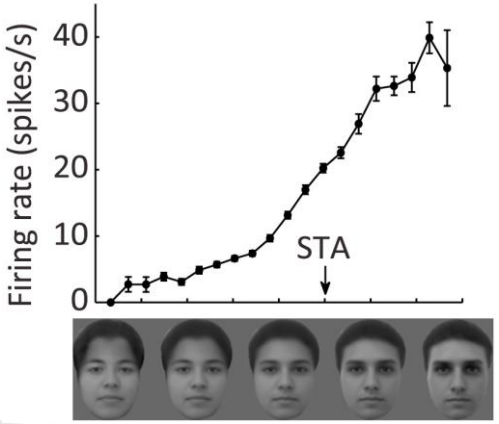
2. Basal ganglia outputs and control of action selection.

3. Cortex: State representation and beyond

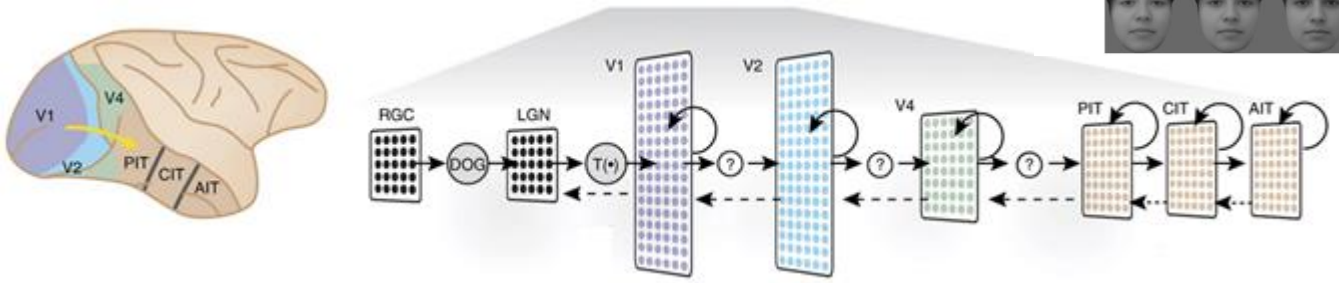
4. Hippocampus: Sequence generation and model-based control

Cortex and state representation

IT face tuning



V1 receptive field



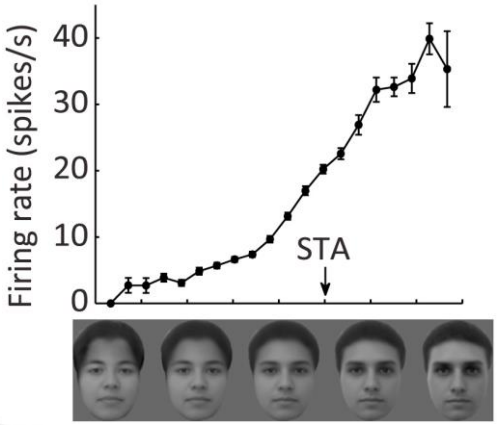
- Sensory cortex extracts behaviourally relevant features (e.g. what and where) from high dimensional sensory input.

Yamins and DiCarlo 2016 Liu et al. 2016 Chang & Tsao 2017

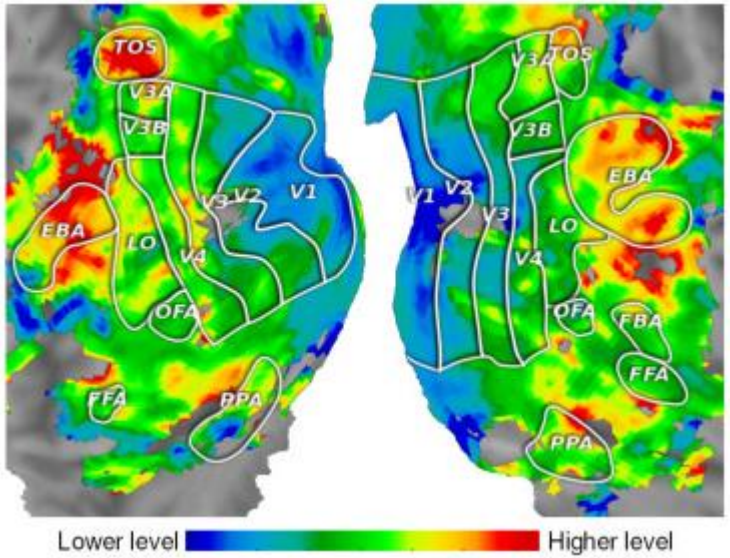
Cortex and state representation

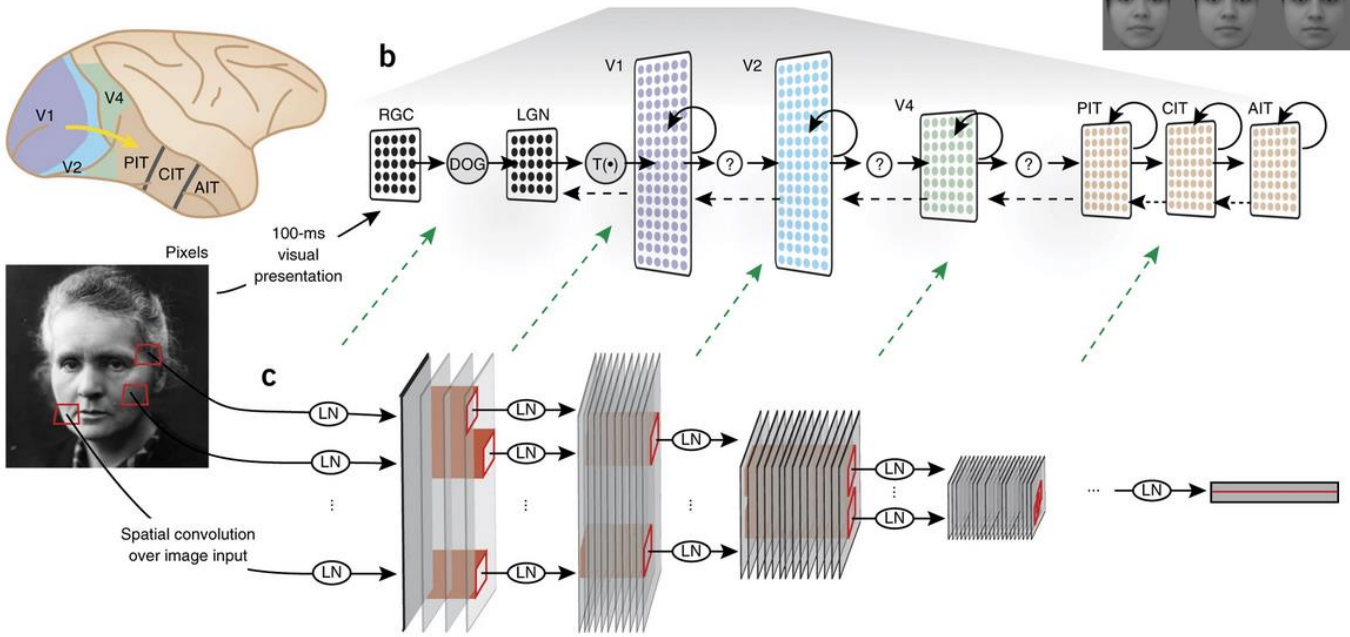
V1 receptive field

IT face tuning



Linear prediction from CNN layer to brain activity

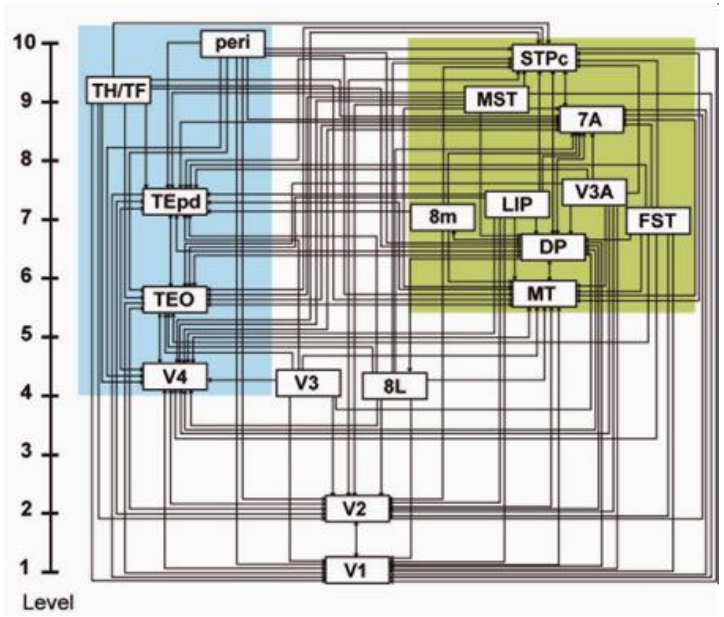


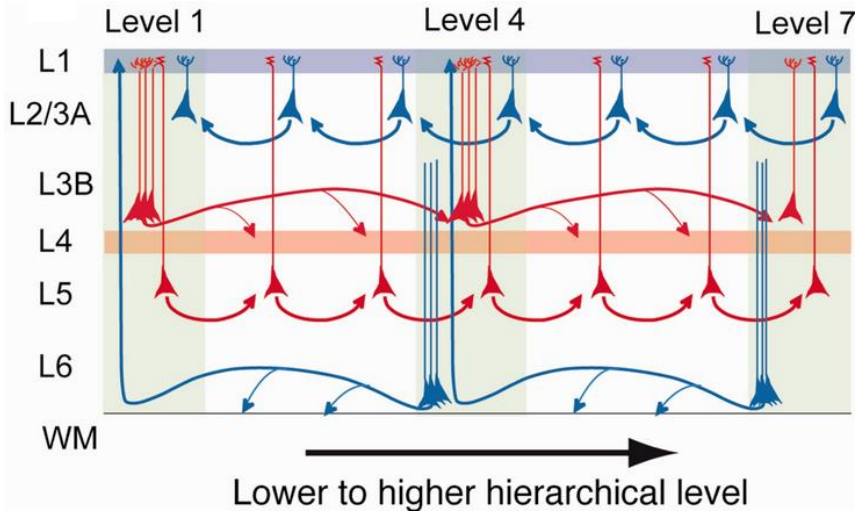


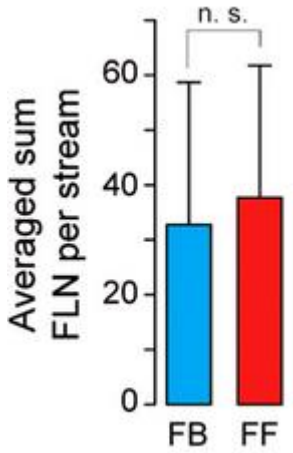
- Sensory cortex extracts behaviourally relevant features (e.g. what and where) from high dimensional sensory input.

Yamins and DiCarlo 2016 Liu et al. 2016 Chang & Tsao 2017 Eickenberg et al. 2017

Cortex as a deep, hierarchical, recurrent network



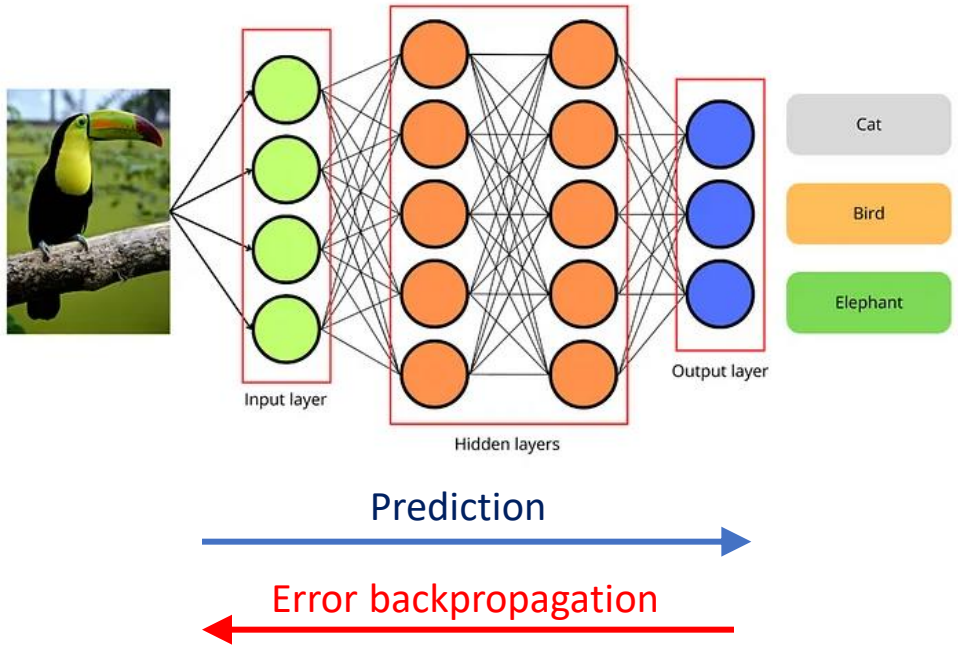


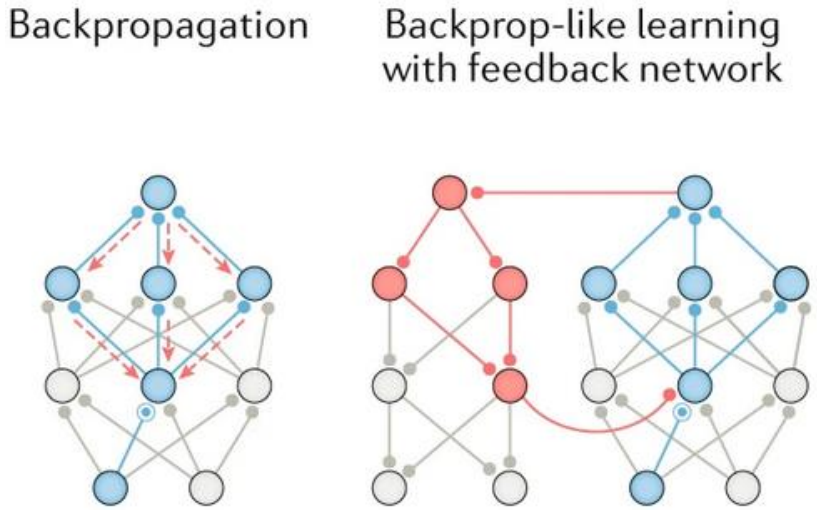


- Cortex is hierarchically organised, with defined feed-forward and feed-back pathways.
- Majority of inter-regional connections are reciprocal.
- Approximately equal numbers of neurons participate in feed-forward and feed-back pathways.

Markov et al. 2012, 2013

Reciprocal connectivity and learning in biological networks

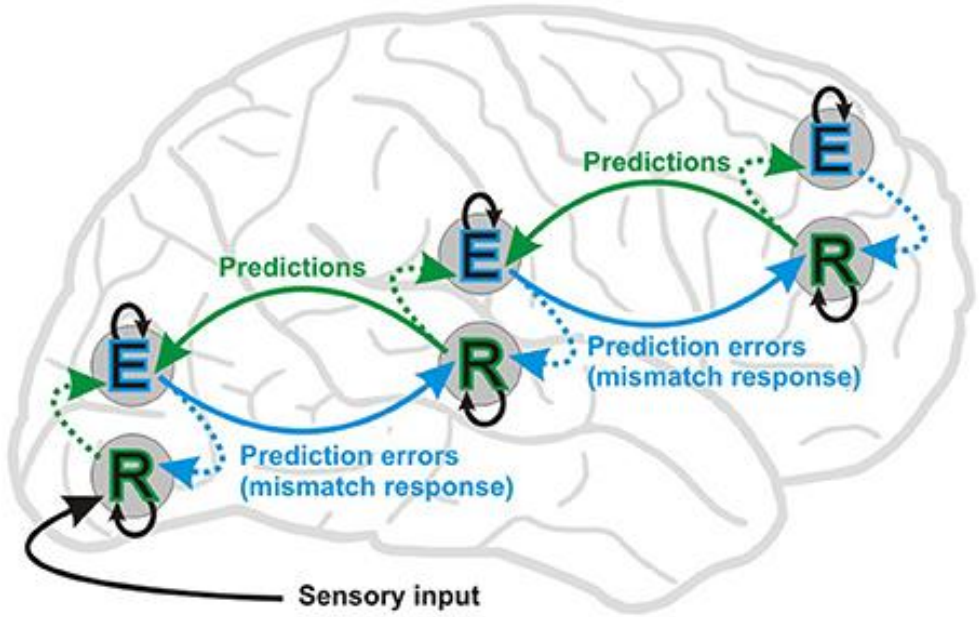


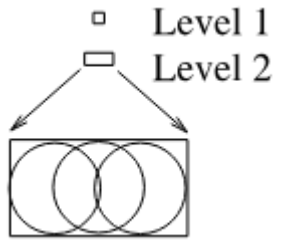


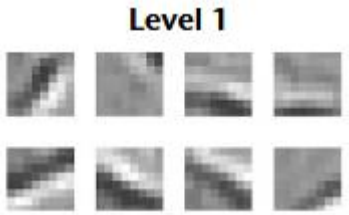
- Deep artificial neural networks learn mappings from input to output by backpropagating errors.
- Backpropagation use non-local information to compute weight updates → not biologically plausible
- Feedback connections can enable backprop-like updates using only local learning rule.

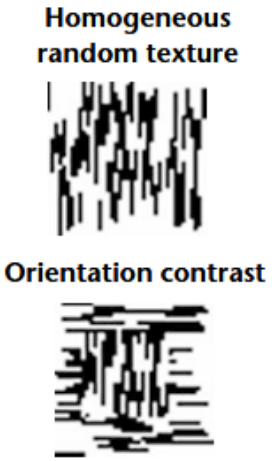
Lillicrap et al. 2020

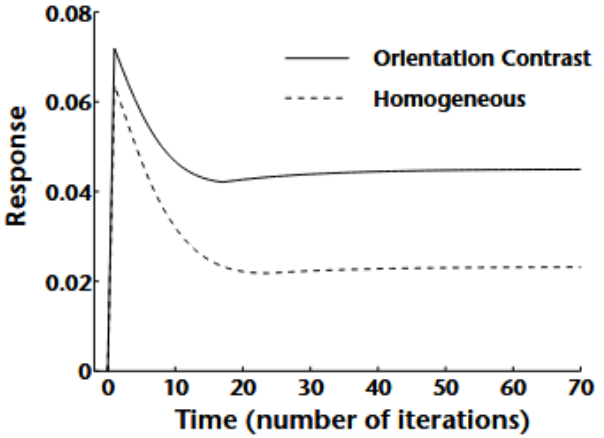
Hierarchical predictive coding (HPC)









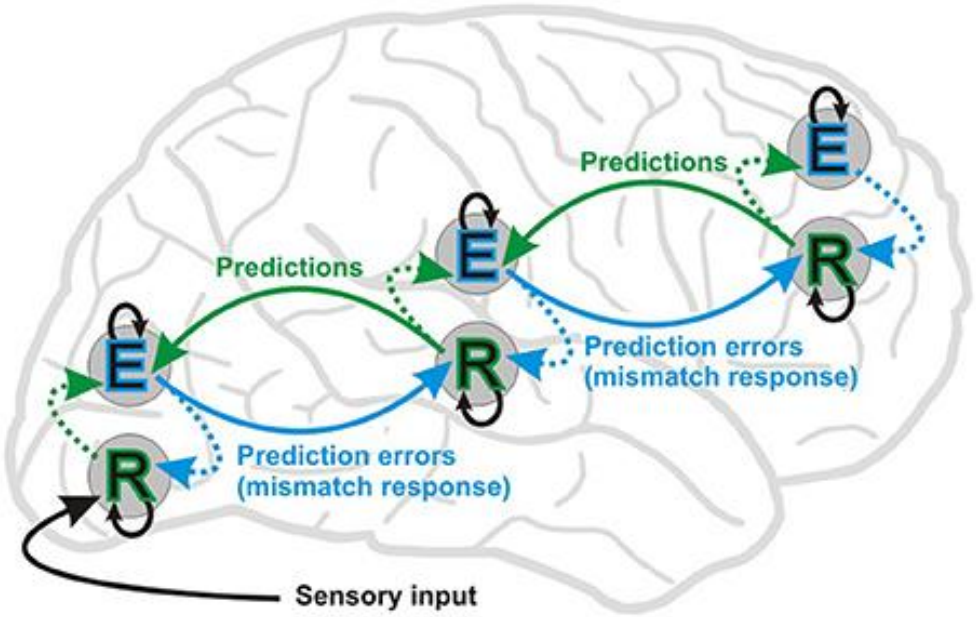


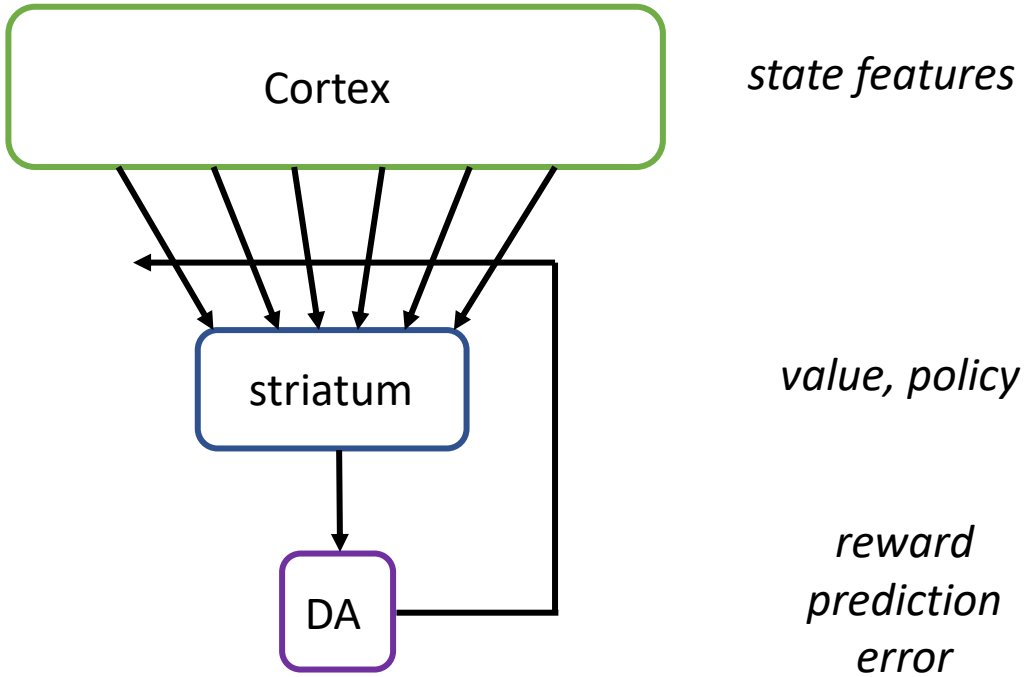
- In HPC, each level of the hierarchy tries to predict activity in the level below, sends prediction errors to the level above.
- HPC learns a generative model of sensory input, using only local learning rules.
- Trained on natural image data, HPC learns neuron-like responses.

Stefanics et al. 2014

Rao and Ballard 1999

Prediction and error in cortex and basal ganglia

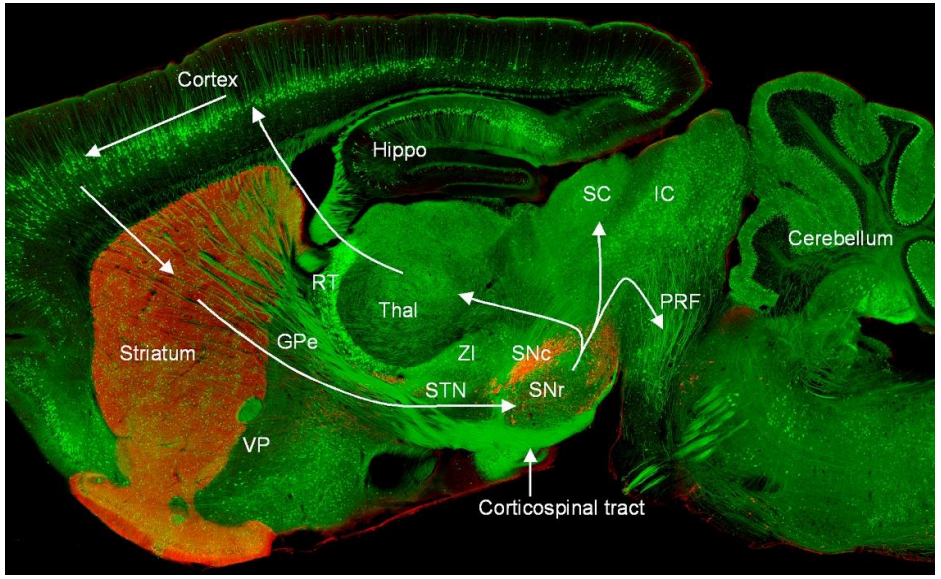


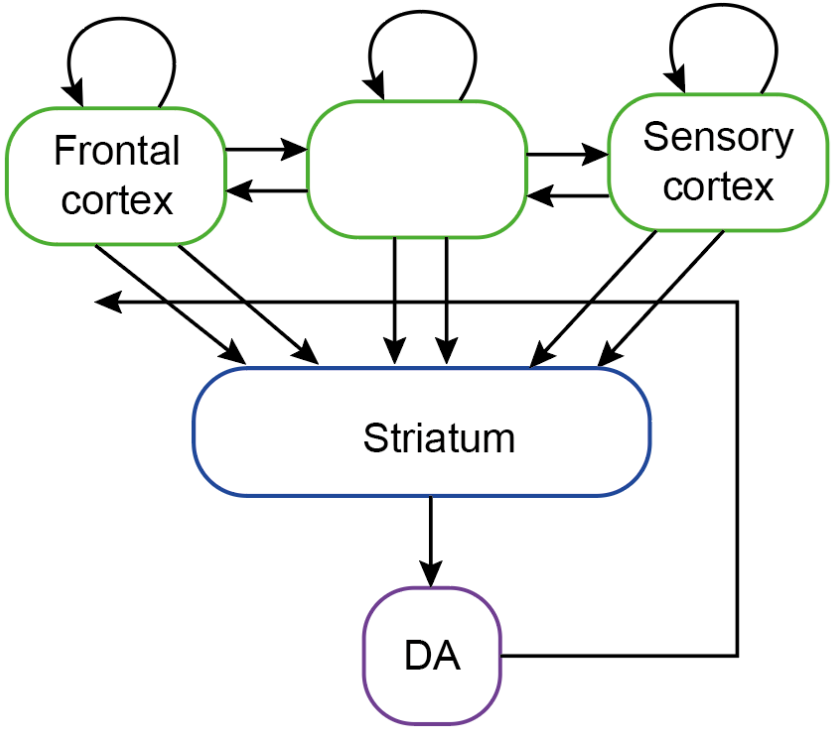


- In HPC models of sensory cortex the prediction target is the next sensory observation → un-/self-supervised learning
- In TD models of basal ganglia function the prediction target is long-run reward → reinforcement learning

Stefanics et al. 2014

Why this architecture?





Self-supervised

RL

- Cortex: Hierarchical, recurrent, largely symmetric projections.
- Striatum: Shallow, no recurrent excitation, no direct projection back to cortex

Bootstrapping and stability in TD learning

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

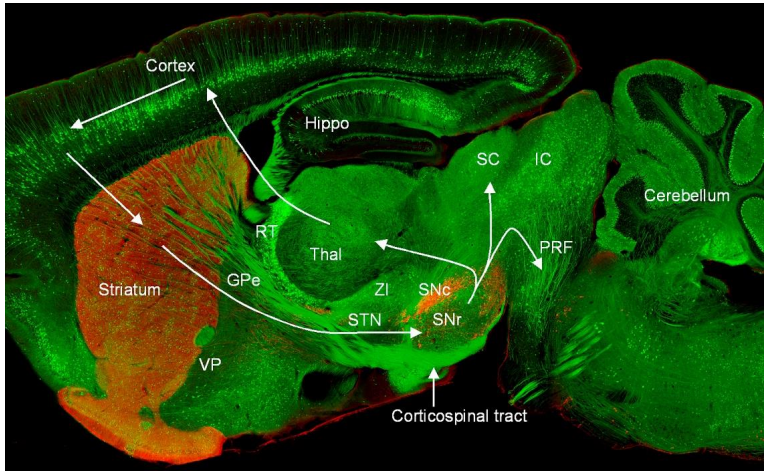
reward prediction error

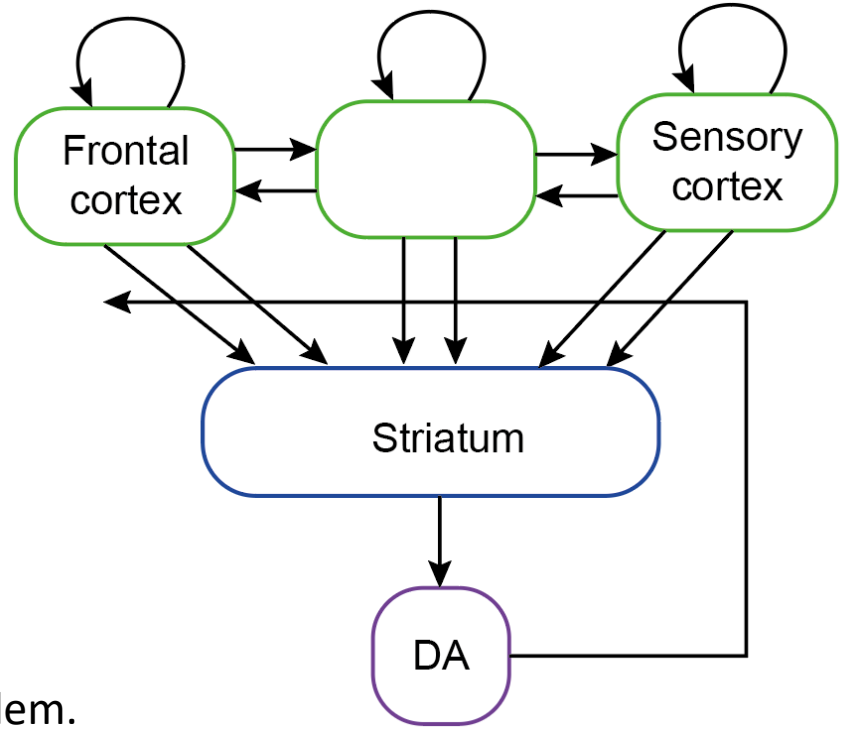
Update target

previous estimate

- Estimated value of next state contributes to update target (bootstrapping).
- Bootstrapping + neural network function approximators is notoriously unstable.
- Stabilizing methods from deep RL don't appear biologically plausible:
  - Use of future rewards in computing weight updates (A2C, PPO)
  - Duplication of network weights (Deep Q)
  - Experience replay (Deep Q)

Why this architecture?





Self-supervised

RL

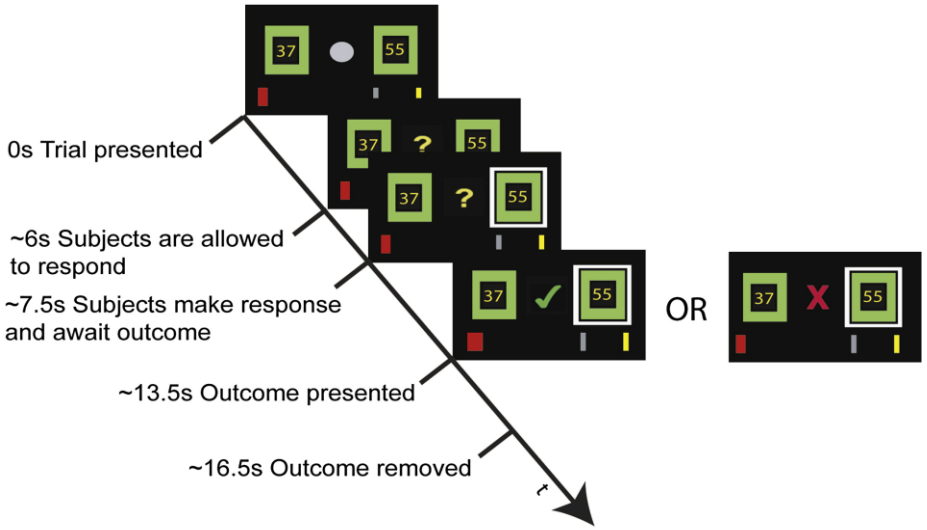
Constraints:

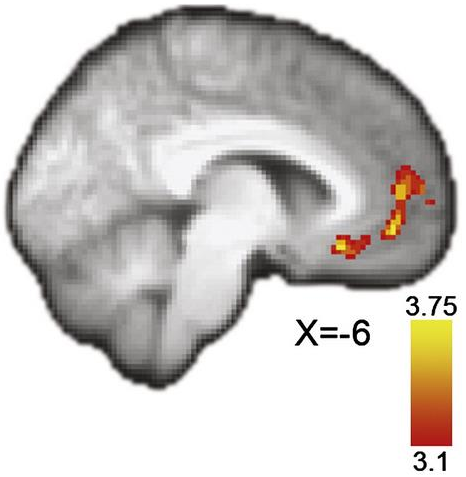
- Need TD learning to solve delayed reward problem.
- Need deep, recurrent, neural networks to handle high dimensional observations and partial observability.

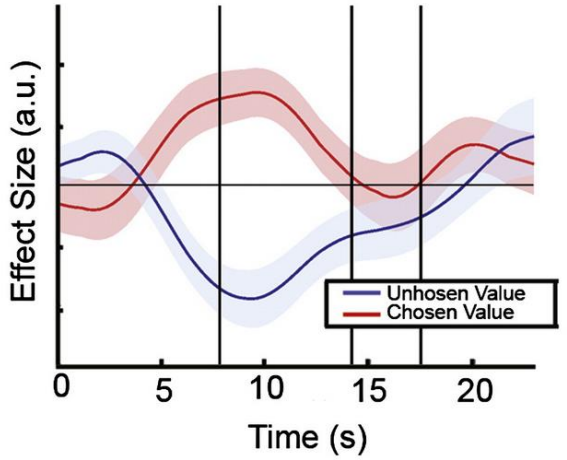
Possible solution:

- Separate representation learning from TD learning.
- Learn representations in the deep recurrent cortical network using self-supervised learning
- Learn long-run values in the shallow, feed-forward, cortico-striatal network.

Cortex as an active player in valuation and action selection



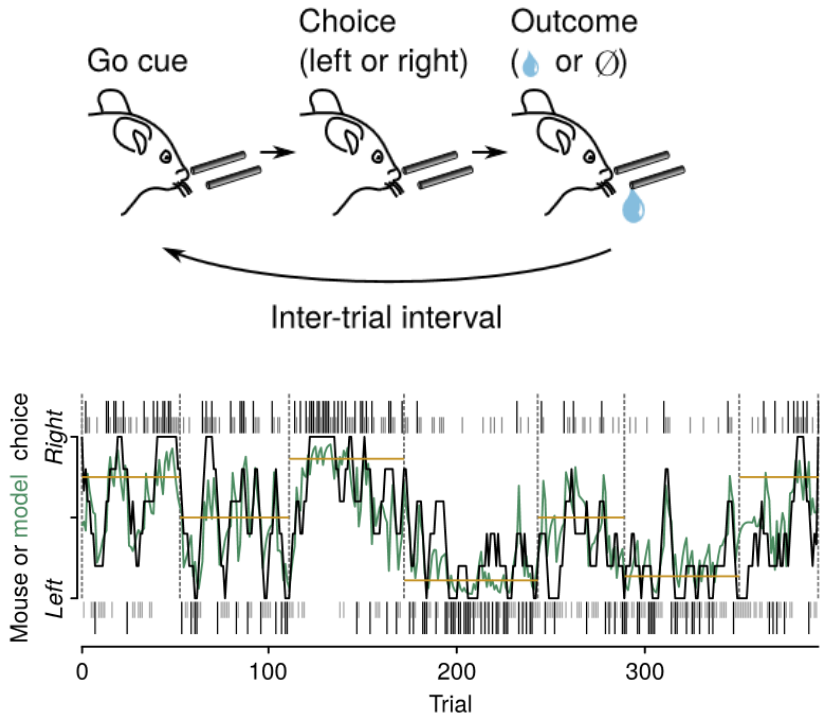


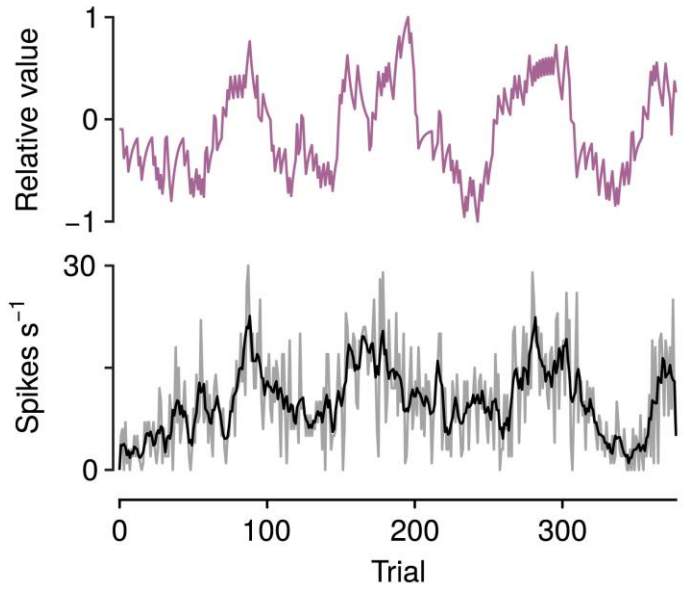


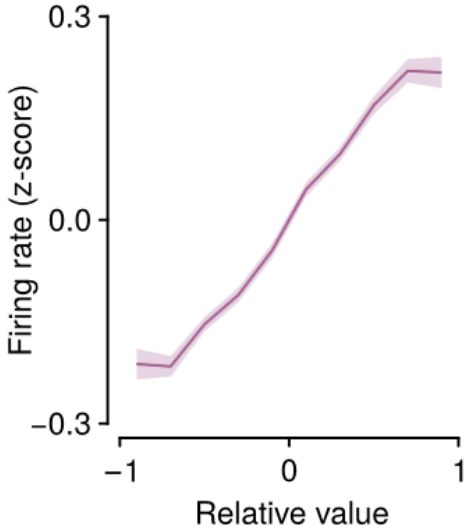
- Value signals are seen in frontal cortex during reward-guided decision making

Boorman et al. 2011

Cortex as an active player in valuation and action selection



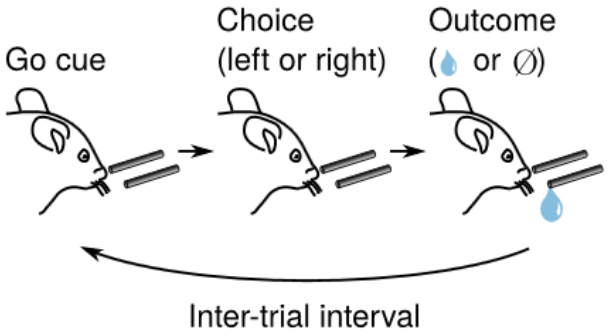




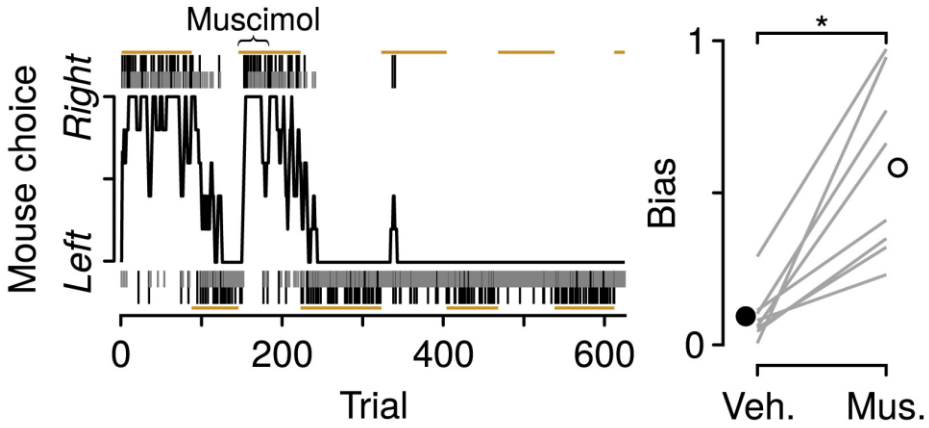
- Value signals are seen in frontal cortex during reward-guided decision making

Bari et al. 2019

Cortex as an active player in valuation and action selection



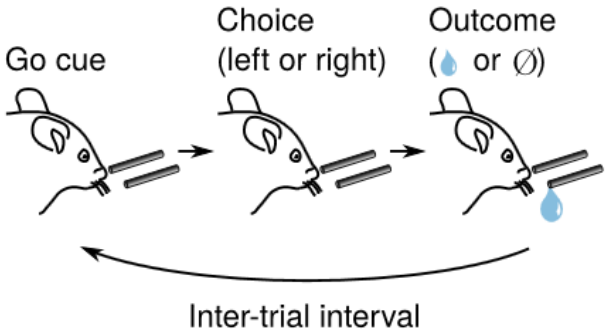
Frontal cortex inactivation

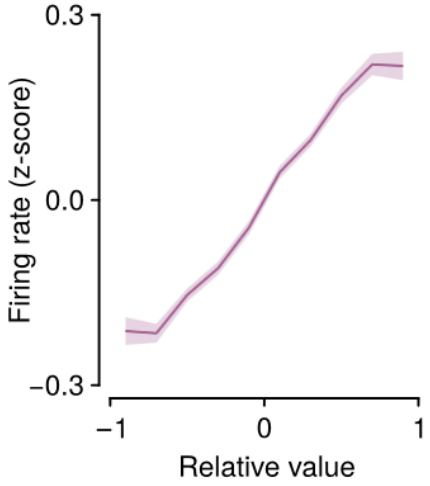


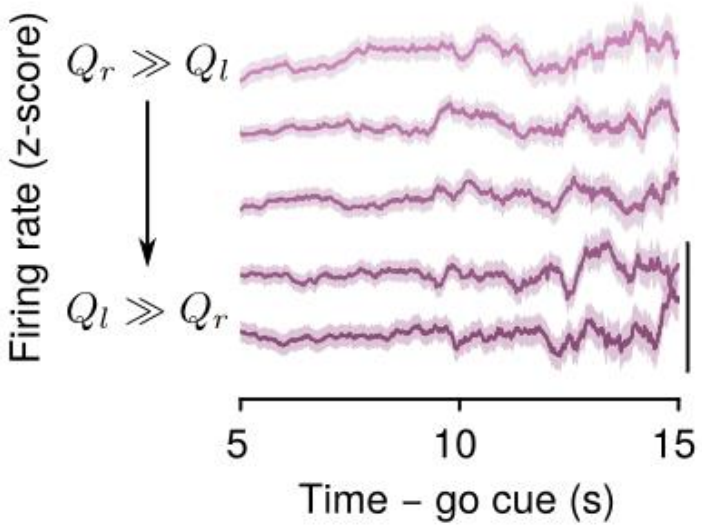
- Value signals are seen in frontal cortex during reward-guided decision making
- Inactivation of frontal cortex profoundly disrupts reward guided decision making

Bari et al. 2019

Cortex as an active player in valuation and action selection



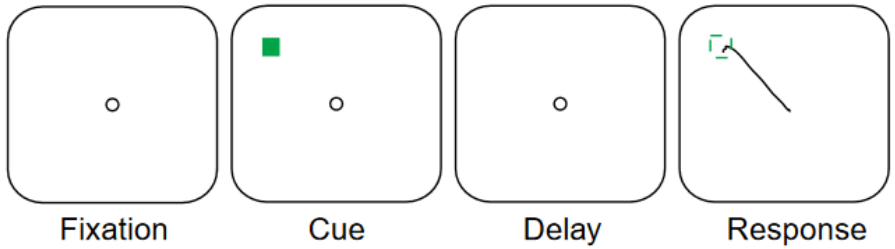


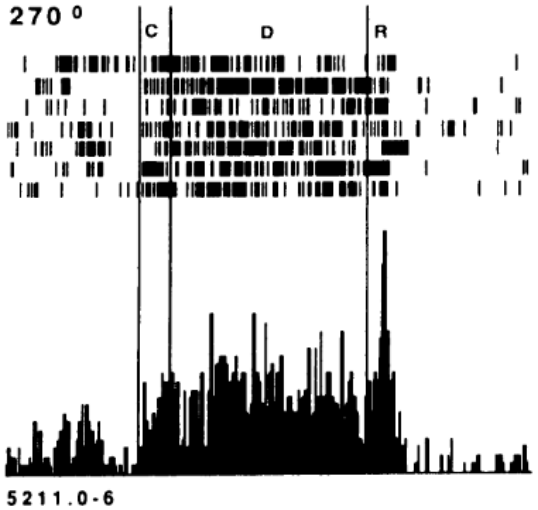


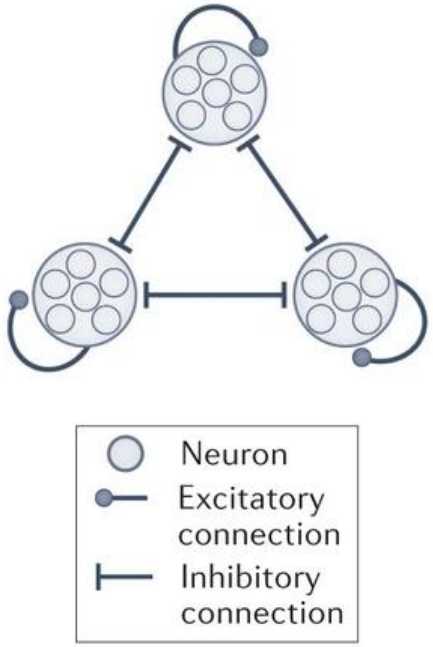
- Value signals are seen in frontal cortex during reward-guided decision making
- Inactivation of frontal cortex profoundly disrupts reward guided decision making
- Value information is present in persistent activity across long inter-trial intervals.

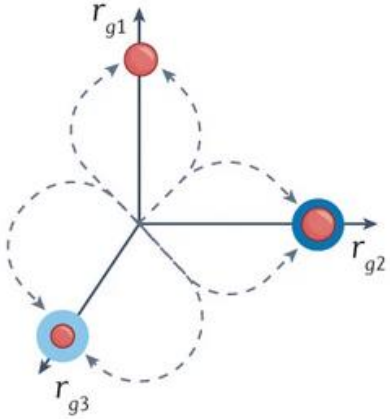
Bari et al. 2019

Recurrent excitation and working memory









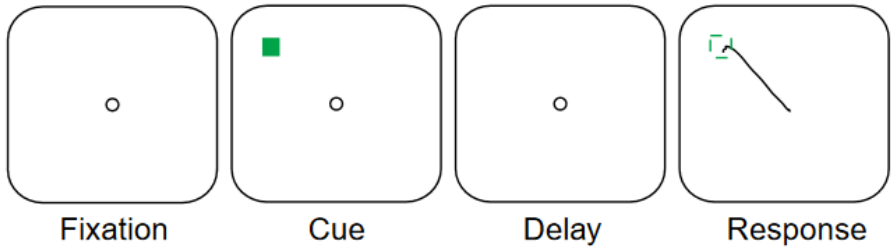
- Frontal cortex neurons famously show sustained persistent activity during the delay period of working memory tasks

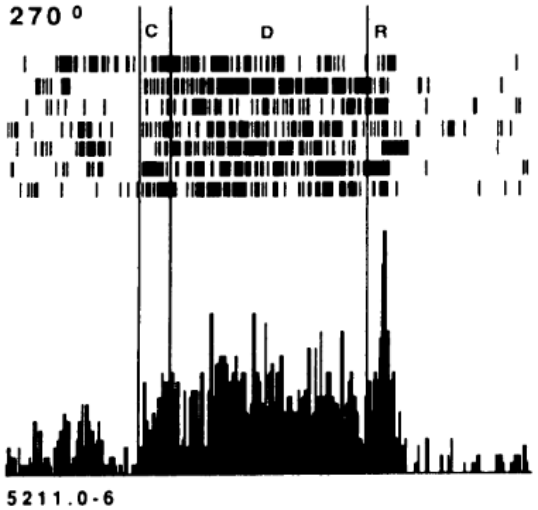
- Networks with attractor dynamics have multiple stable fixed-points in their activity space, allowing information to be stored as patterns of recurrent activity.

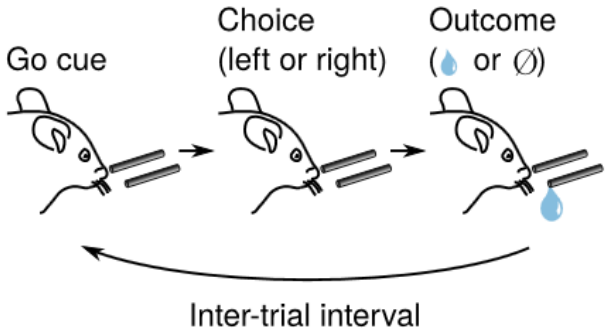
Funahashi et al. 1989, Curtis & D'Esposito 2003

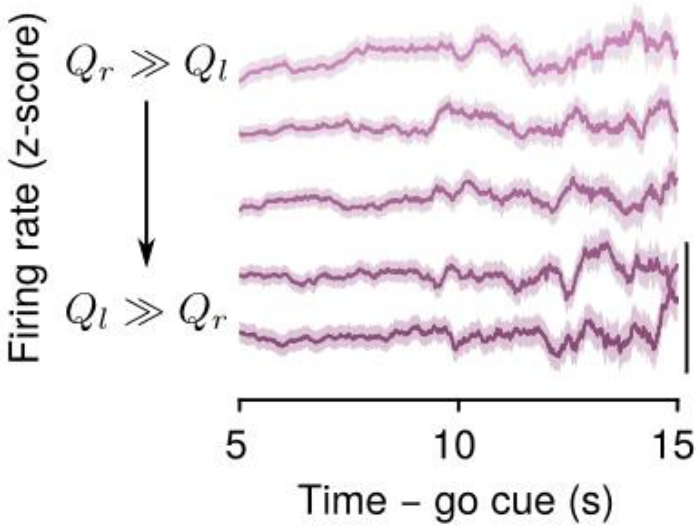
Khona & Fiete 2022

Reward guided decision making and working memory







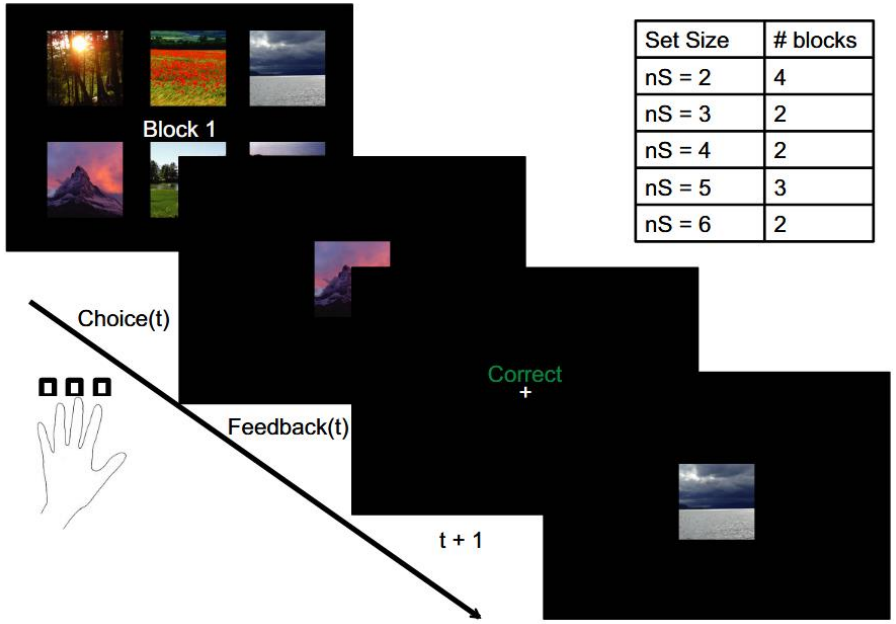


- Both working memory tasks and reward guided decision tasks are partially observable.
- Persistent value coding suggests recent reward history is stored by attractor dynamics not synaptic weight changes.

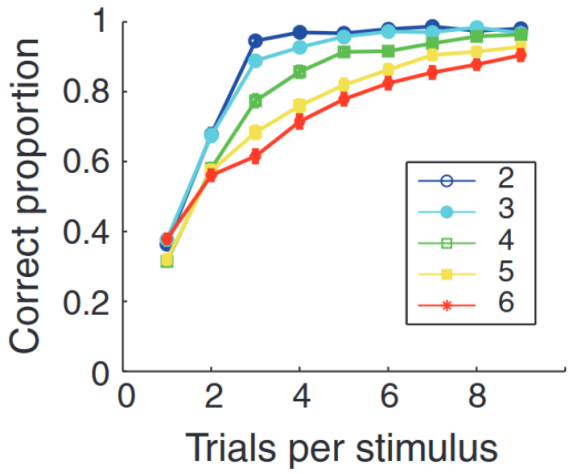
Funahashi et al. 1989, Curtis & D'Esposito 2003

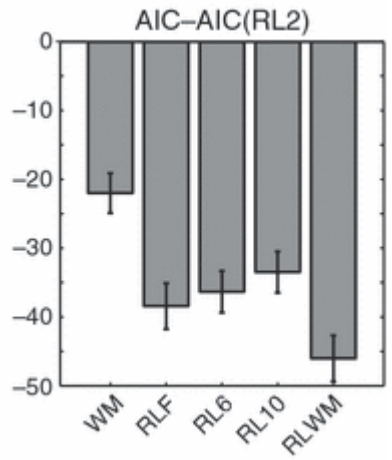
Bari et al. 2019

Working memory in human action learning



Set Size	# blocks
nS = 2	4
nS = 3	2
nS = 4	2
nS = 5	3
nS = 6	2



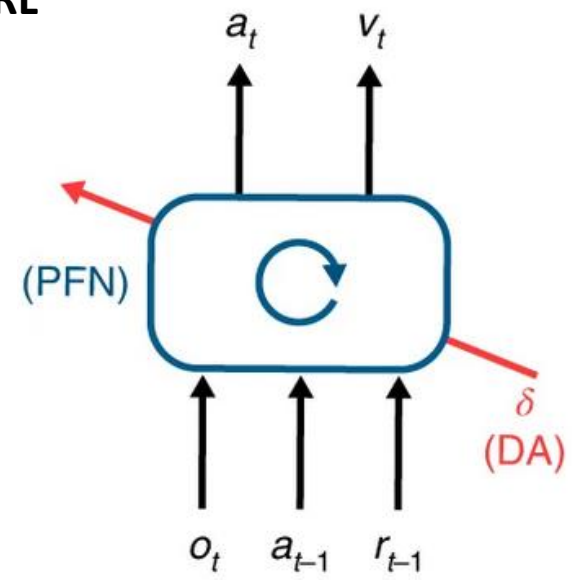


- Human trial-and-error learning shows strong set size effects consistent with limited working memory capacity.
- Data best fit by model combing fast capacity limited working memory with slow capacity unlimited RL.

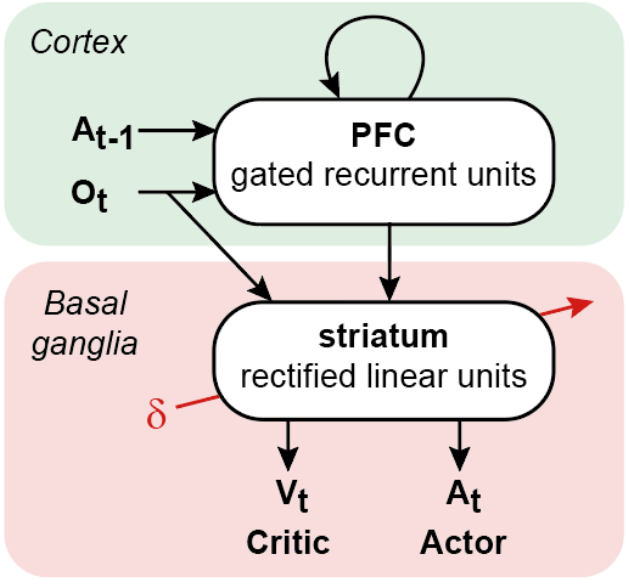
Collins & Frank 2012

Learning recurrent dynamics for reward guided decision making

Meta RL



PFC-BG model



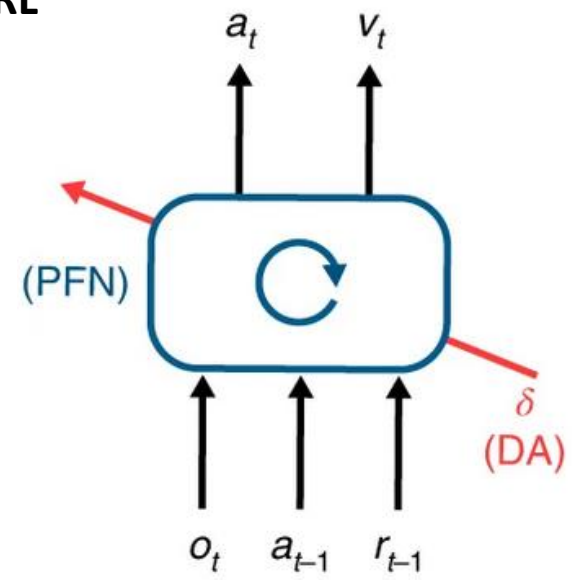
- Meta RL:
  - Recurrent network (LSTM) trained end-to-end using RL (A2C) to predict long run reward and select actions.
- PFC-BG model:
  - Recurrent “frontal cortex” network trained with self-supervised learning to predict the next observation.
  - Feed-forward “basal-ganglia” network trained with RL (A2C) to predict long run reward and select actions.

Wang et al. 2018

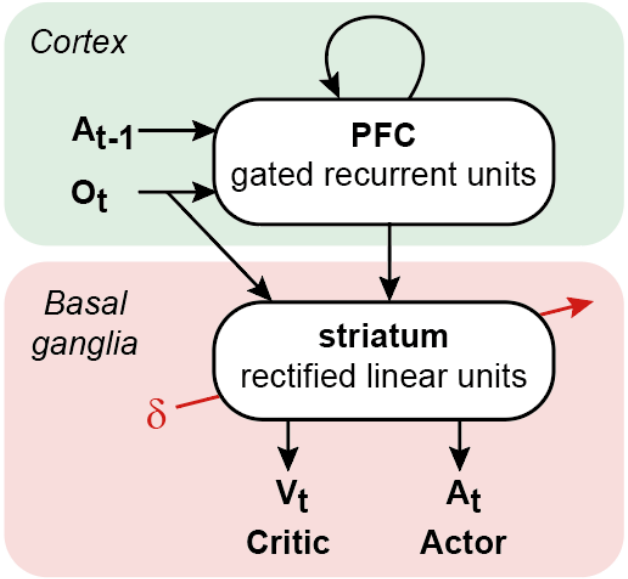
Blanco-Pozo et al. 2024

Learning recurrent dynamics for reward guided decision making

Meta RL



PFC-BG model

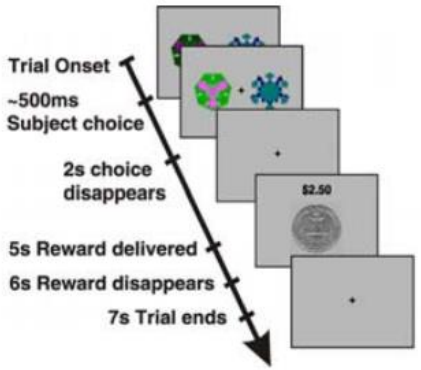


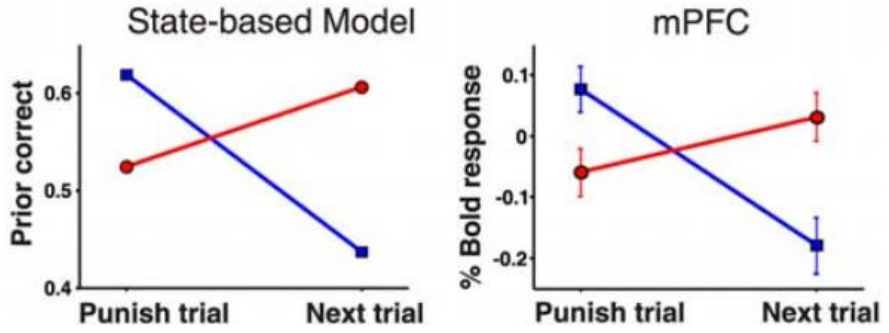
- In both models, synaptic plasticity on slow timescale (task acquisition) sculpts recurrent dynamics that drive behavioural adaptation on a fast timescale (trial-to-trial).

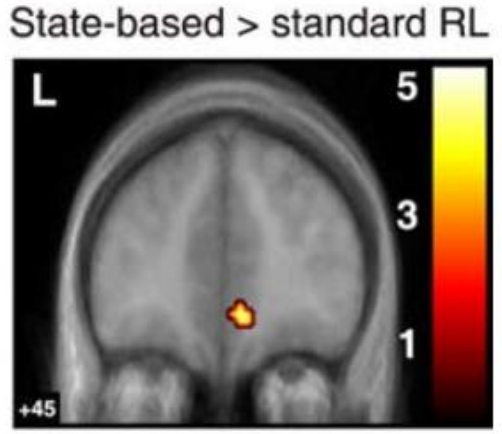
Wang et al. 2018

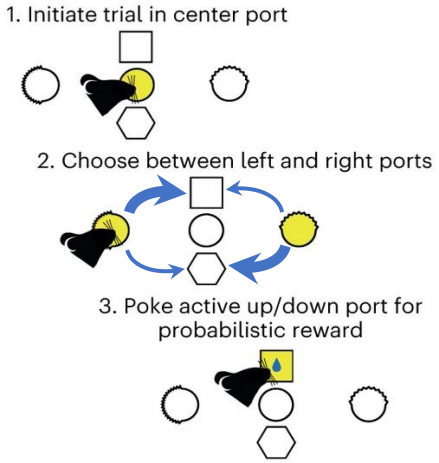
Blanco-Pozo et al. 2024

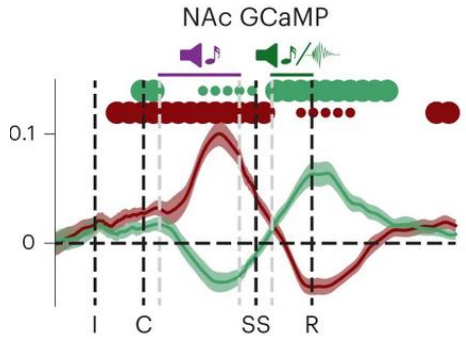
Recurrent network models and inferred value updates











- In reversal learning tasks, evidence one option is good is also evidence the other option is bad.

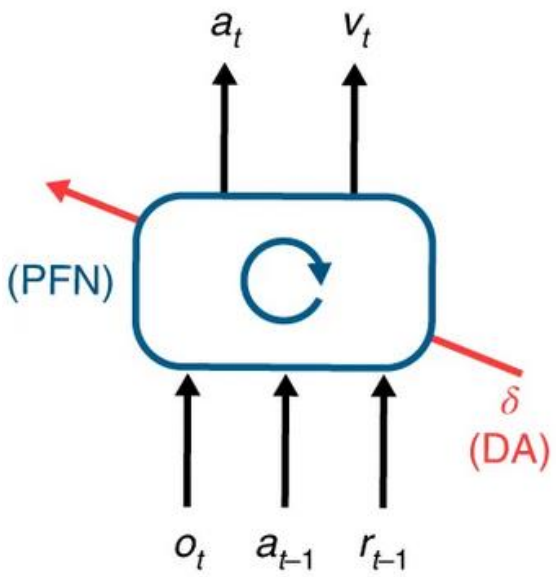
- Brain signals reflect such inferred value updates, both in humans and mice, frontal cortex and dopamine.

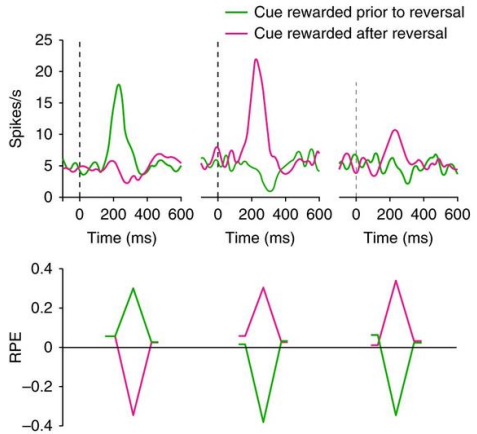
Hampton et al. 2006

Blanco-Pozo et al. 2024

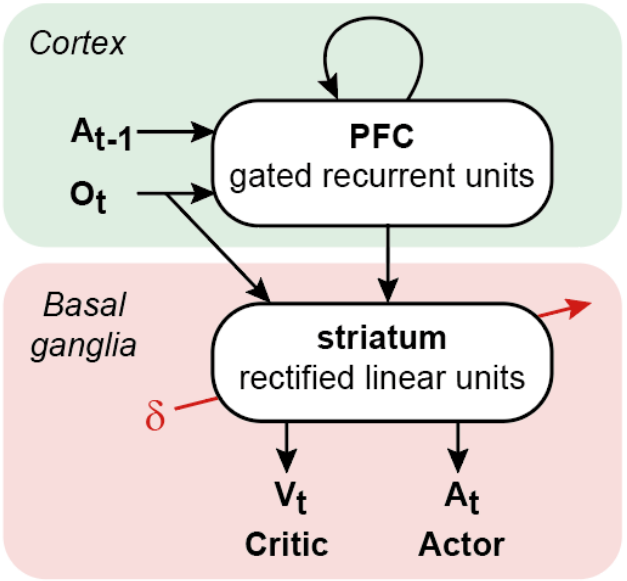
Recurrent network models and inferred value updates

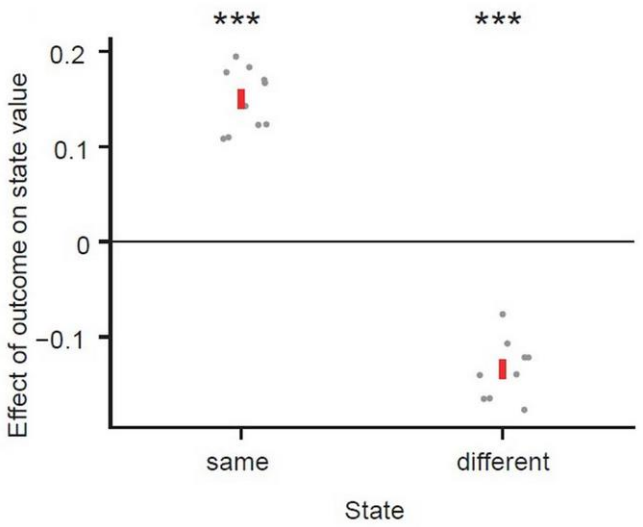
Meta RL





PFC-BG model



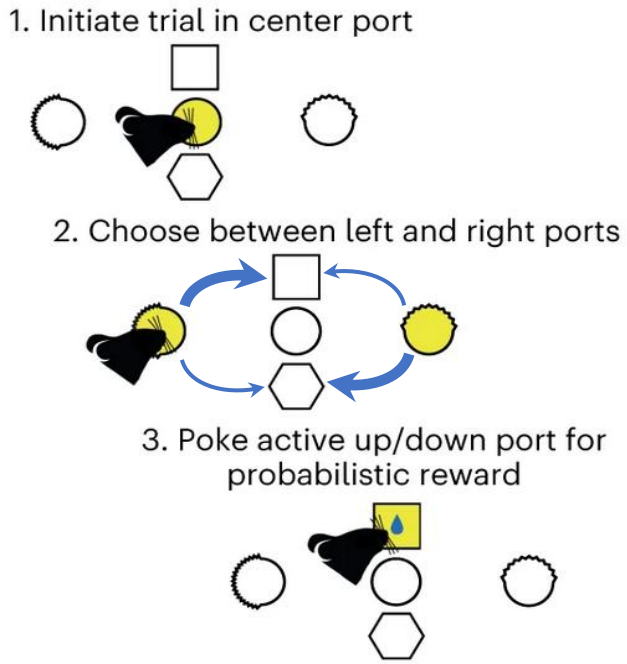


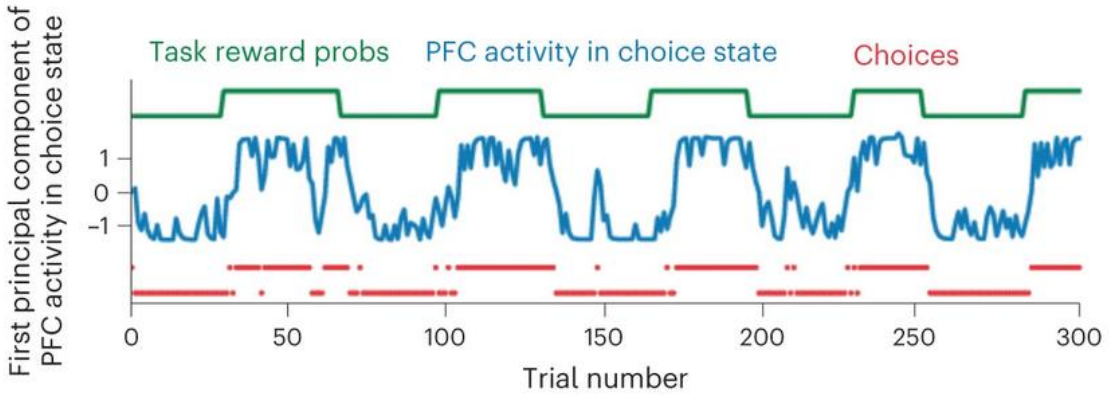
- Both models reproduce inferred value updates when reward probabilities are anticorrelated.

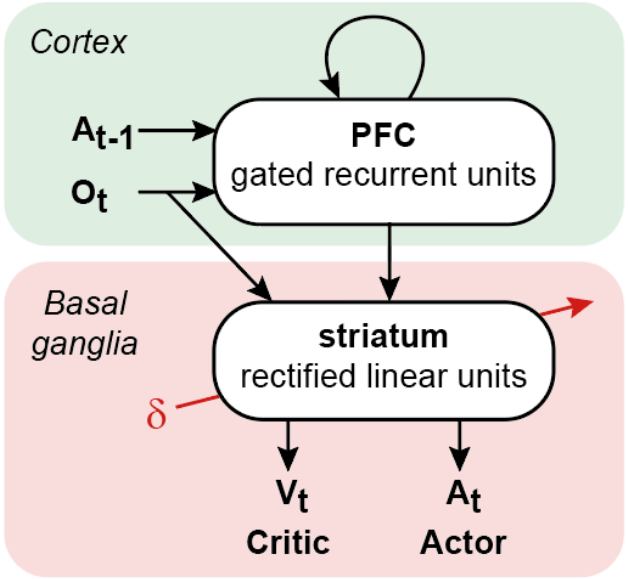
Wang et al. 2018

Blanco-Pozo et al. 2024

Frontal value signals or state representations?







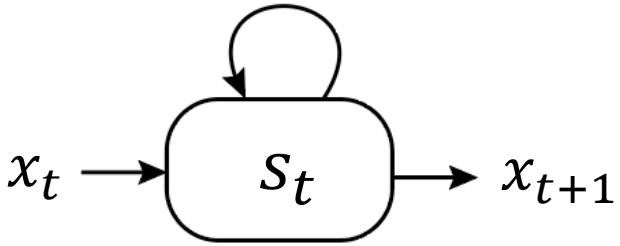
- Recurrent network activity tracks task reward probabilities despite being trained only to predict the next observation.
- Not RL “value” (long run reward) as recurrent network trained only with self-supervised learning.

Blanco-Pozo et al. 2024

Frontal value signals or state representations?

$$\underbrace{x_0,\ x_1,\ x_2,\ \dots,\ x_{t-1},\ x_t}_{\text{Past observation history}},\ \underbrace{x_{t+1}, x_{t+2}, \dots}_{\text{Future observations}}$$

$$\underbrace{s_t}_{\text{Current state}}$$

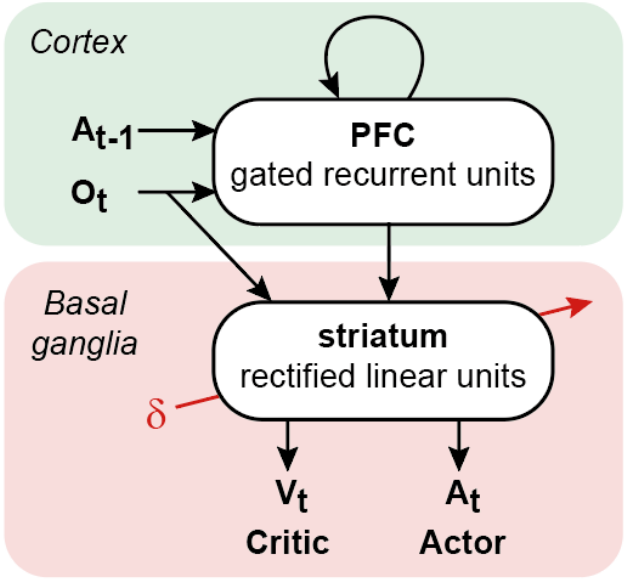


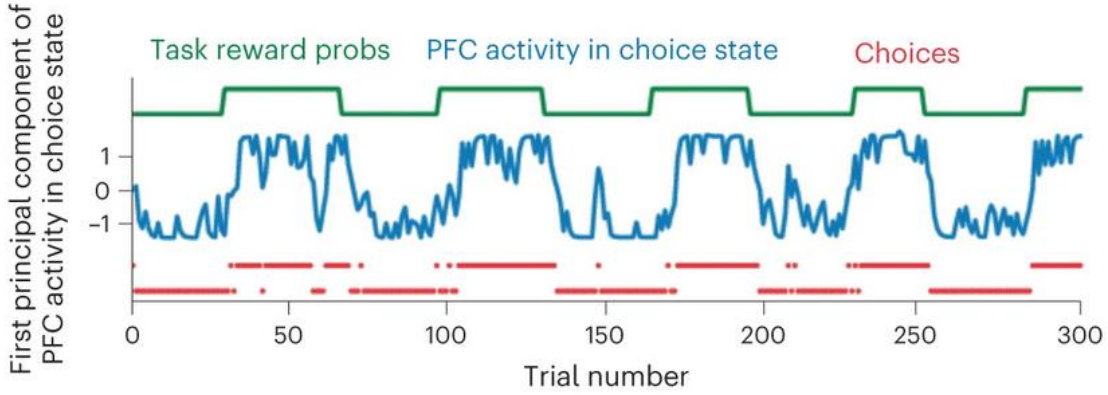
- A “Markov” state summarizes all information in the observation history necessary for making any prediction.

A common strategy for finding a Markov state is to look for something compact that is recursively updatable and enables accurate short-term predictions. In fact, it is only necessary to make accurate one-step predictions.

Sutton and Barto

Frontal value signals or state representations?

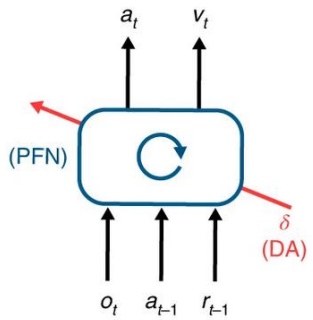




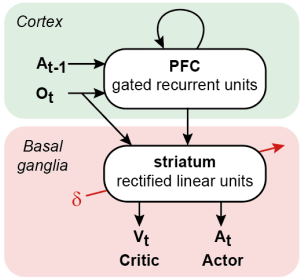
- PFC recurrent network learns a Markov state representation by predicting observations.
- BG network learns values and policy over observable and learned state features.
- “Value” signals in frontal cortex are in in fact beliefs about latent task state.

Blanco-Pozo et al. 2024

Frontal value signals or state representations?



Frontal cortex as value/policy

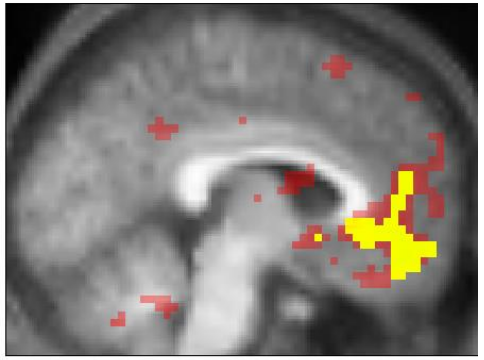


Frontal cortex as state representation

[Wang et al. 2018]

[Blanco-Pozo et al. 2024]

Frontal value signals or state representations?

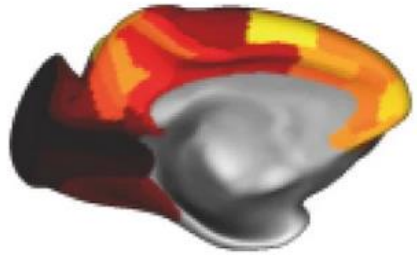


Money –
value predicted

Faces -
attractiveness

Coke or Pepsi –
behavioral preference

Dopamine D1 receptor density



First principal component of 14
receptor densities

- Ventromedial frontal cortex consistently responds to valued stimuli across diverse domains
- Gradients in receptor densities and many other neuronal properties across cortical hierarchy

→ Reward does appear special in frontal cortex, learning algorithm remain poorly constrained.

Daw et al 2006 McClure et al. 2004 O'Doherty et al 2003 Froudist-Walsh et al. 2023

Talk aim and overview

Question: How do learning algorithms map onto brain structure to solve the biological action control problem?

Talk outline:

1. Striatum and dopamine: The brain's temporal difference reinforcement learning system?

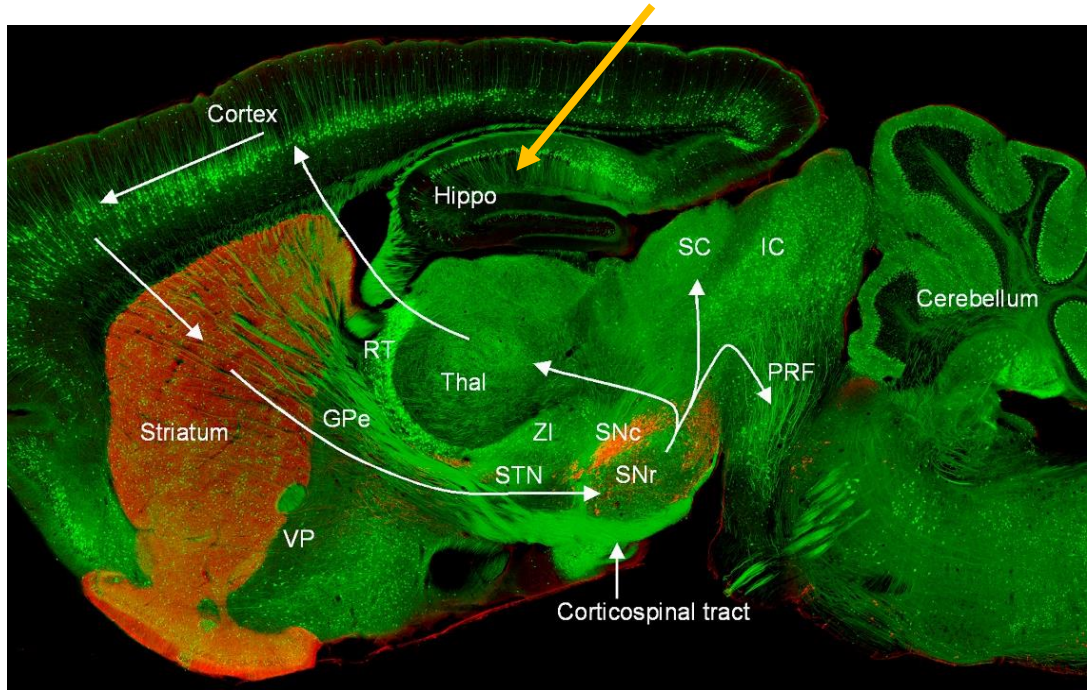
2. Basal ganglia outputs and control of action selection.

3. Cortex: State representation and beyond

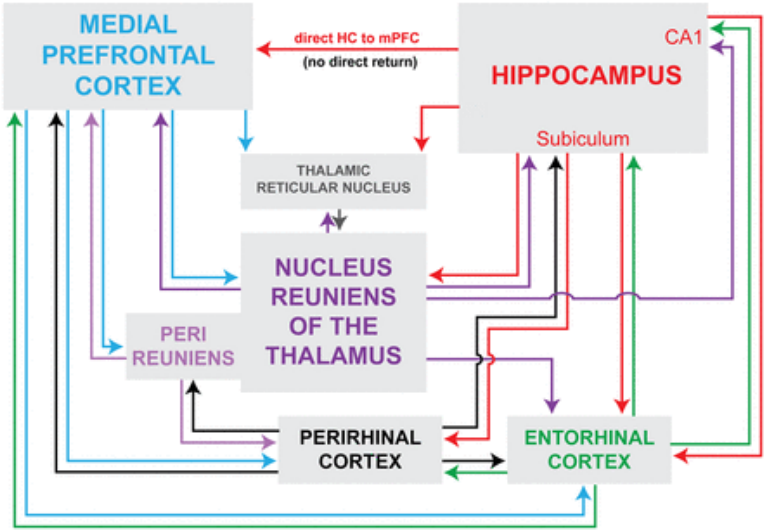
4. Hippocampus: Sequence generation and model-based control

Hippocampus: Sequence generation and model-based control

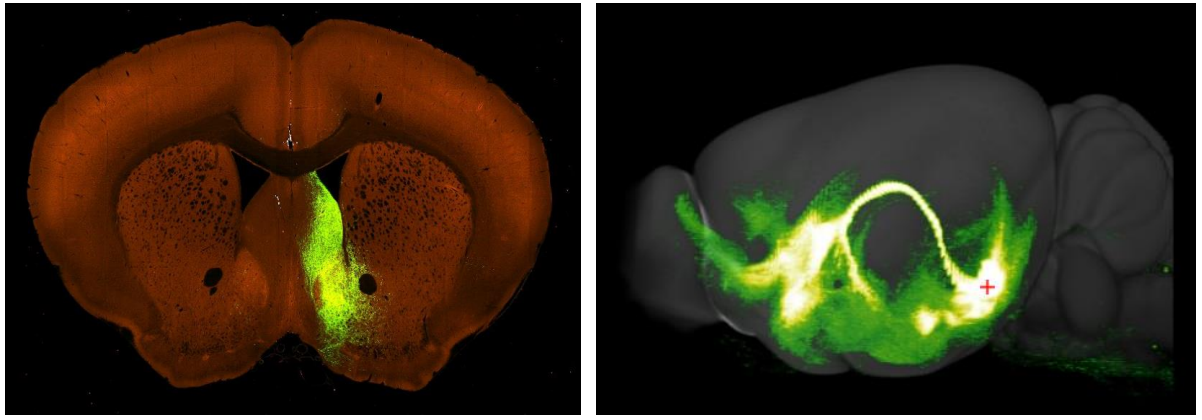
Hippocampus



Connections with frontal cortex



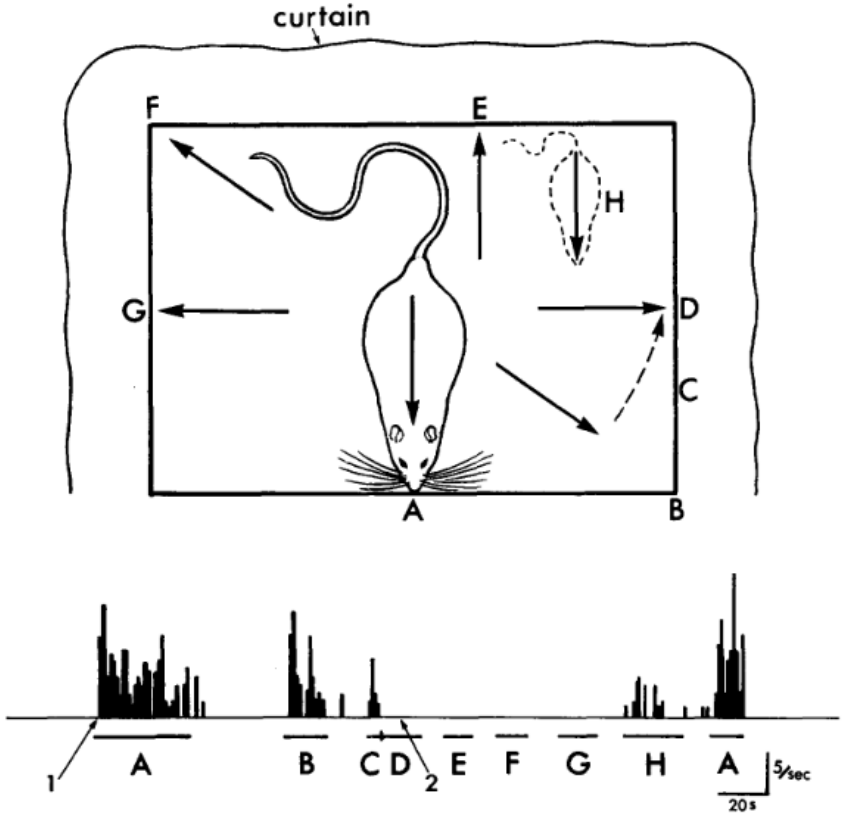
Projection to ventral striatum

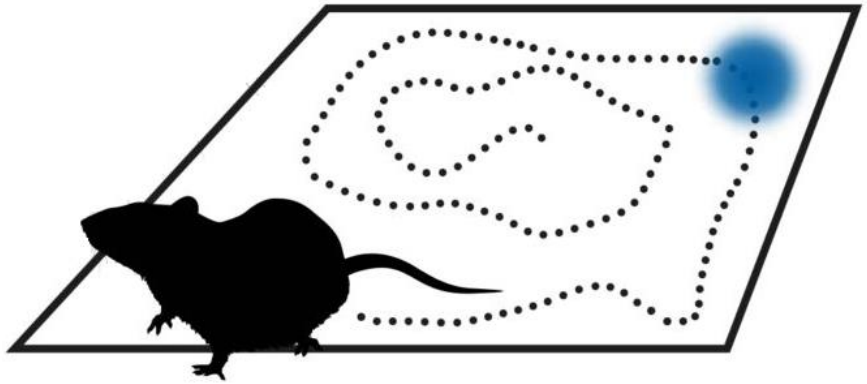


Allen Brain Atlas

Dolleman-van der Weel et a. 2022

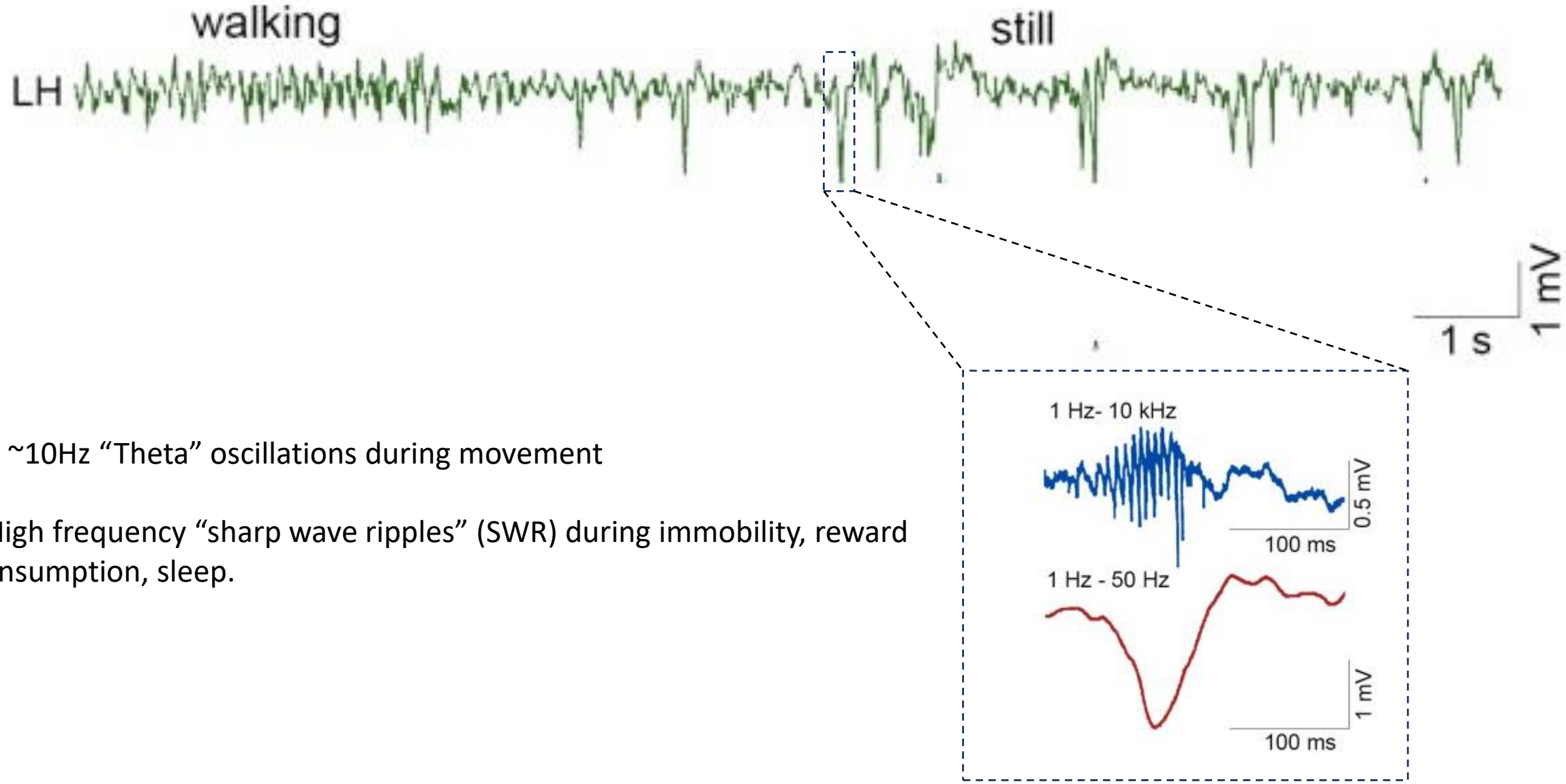
Hippocampal place cells





O'Keefe & Dostrovsky 1971

Hippocampal sequences

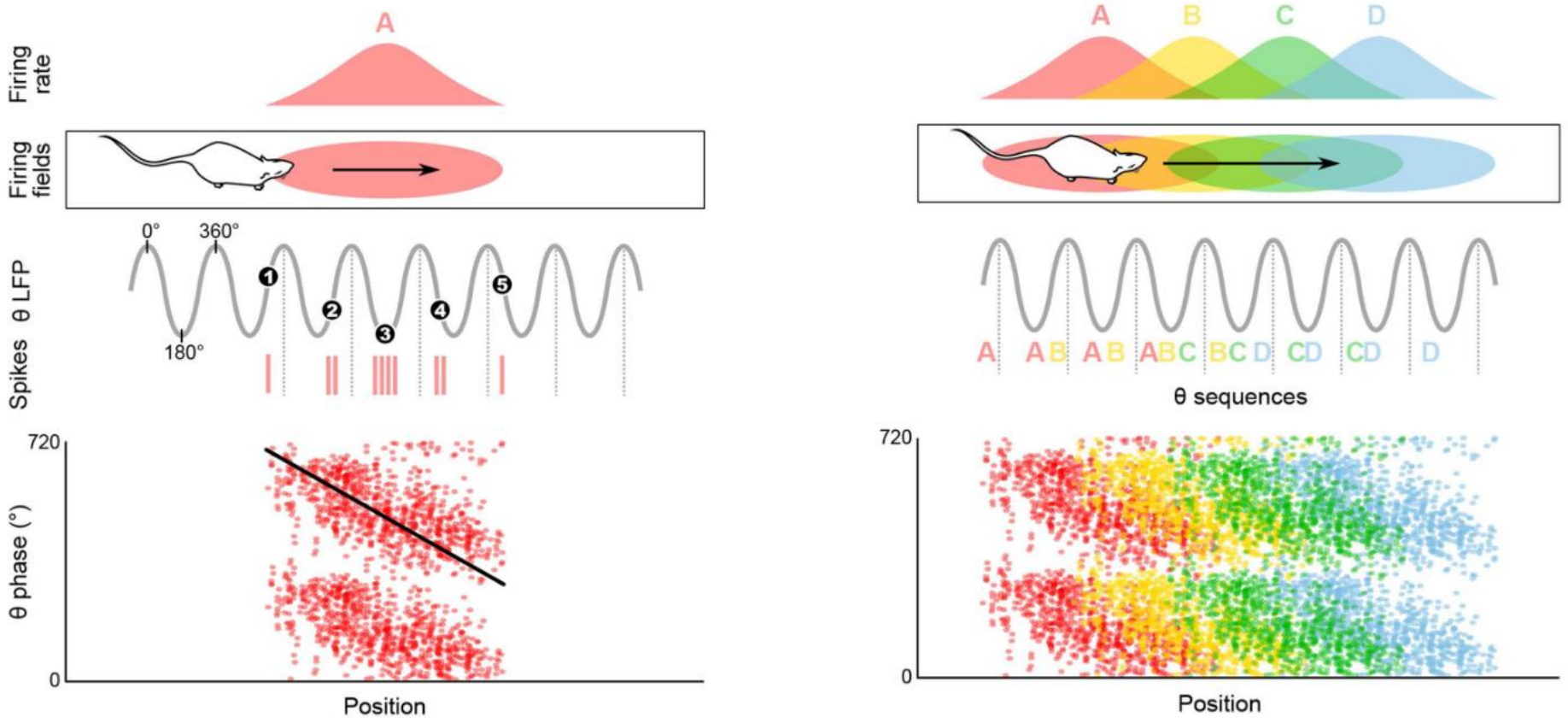


- ~10Hz “Theta” oscillations during movement

- High frequency “sharp wave ripples” (SWR) during immobility, reward consumption, sleep.

Buzsáki 2015

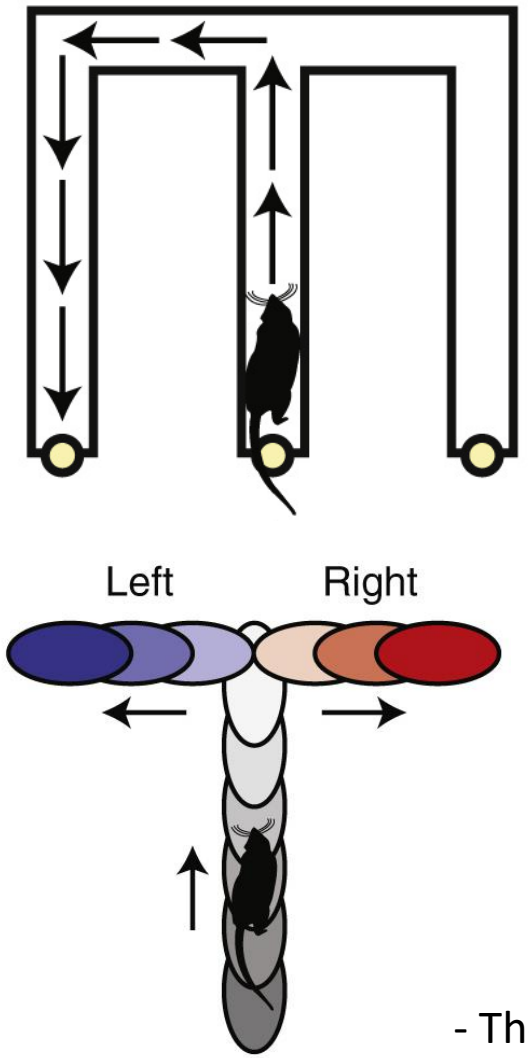
Hippocampal sequences: Theta oscillations

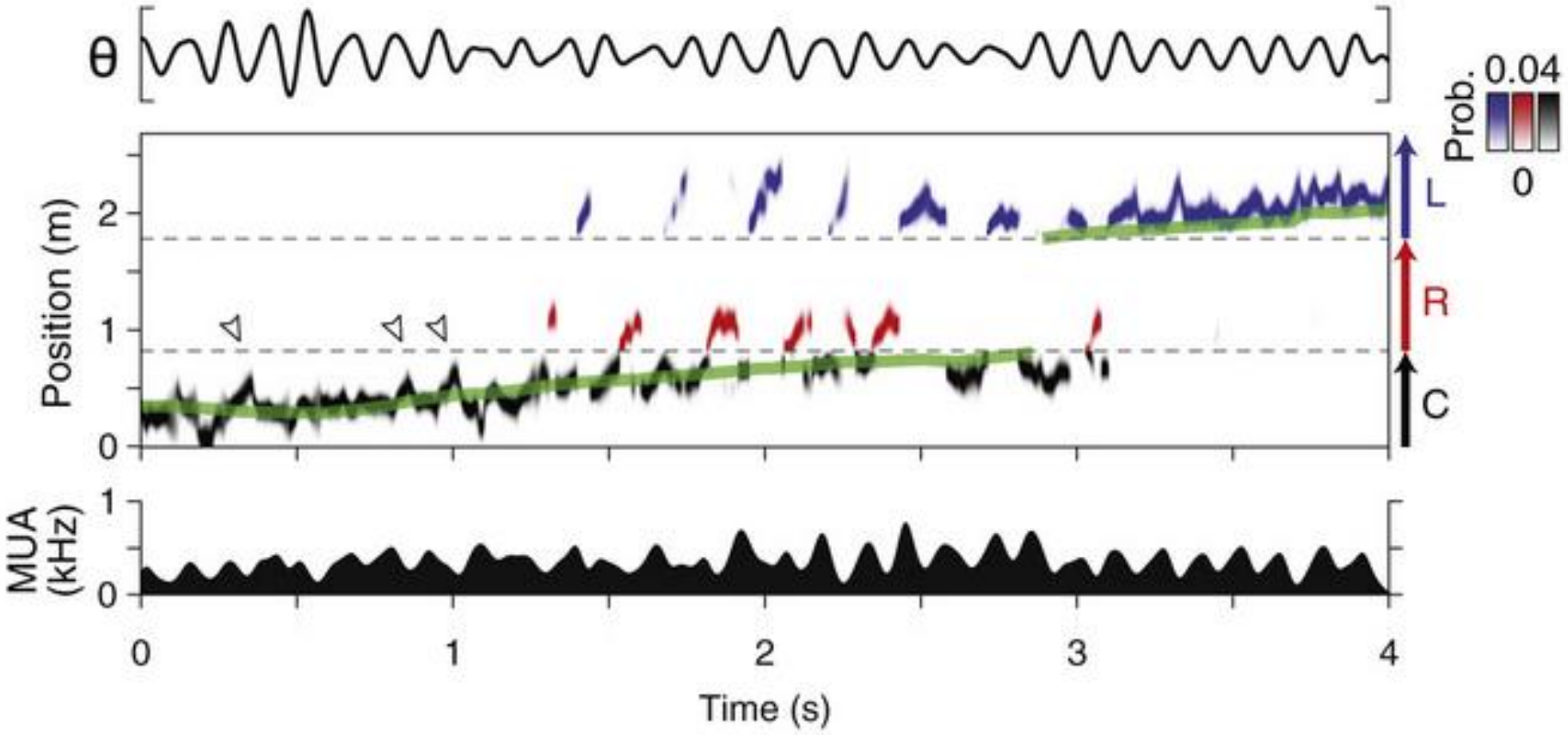


- Firing phase shifts earlier relative to theta in cycle during movement through place field
- Place cells along trajectory fire in a compressed sequence on each theta cycle.

Drieu & Zugaro 2019

Hippocampal sequences: Theta oscillations

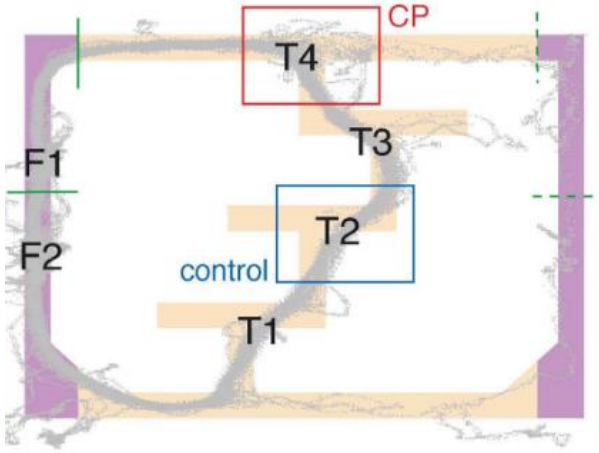


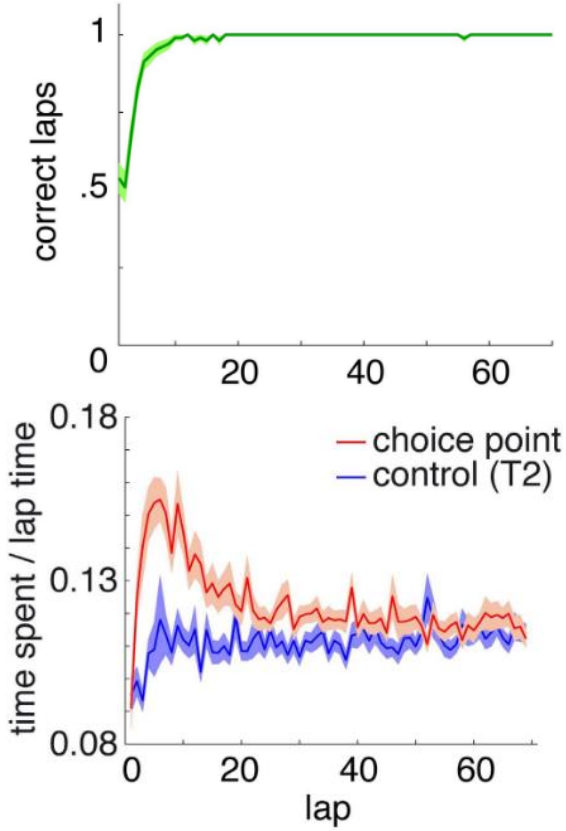


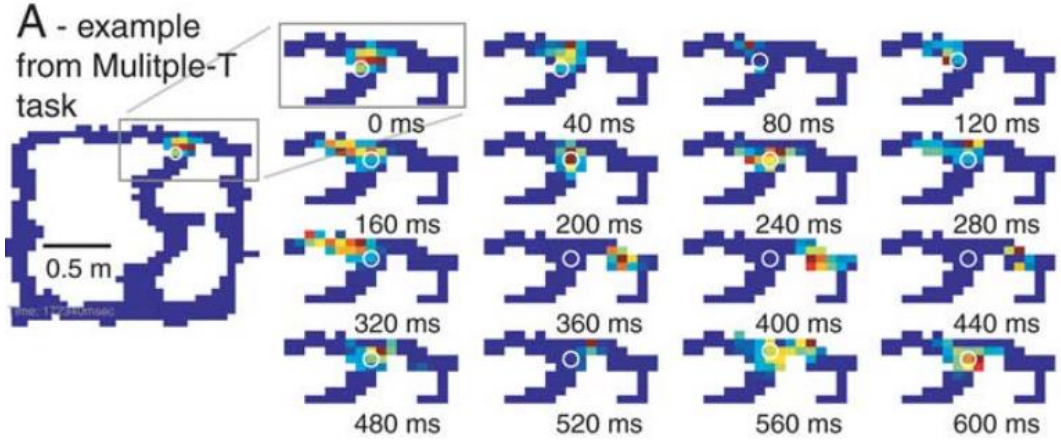
- Theta sweeps alternate between possible future trajectories at decision points.

Kay et al. 2020

Hippocampal sequences: Theta oscillations





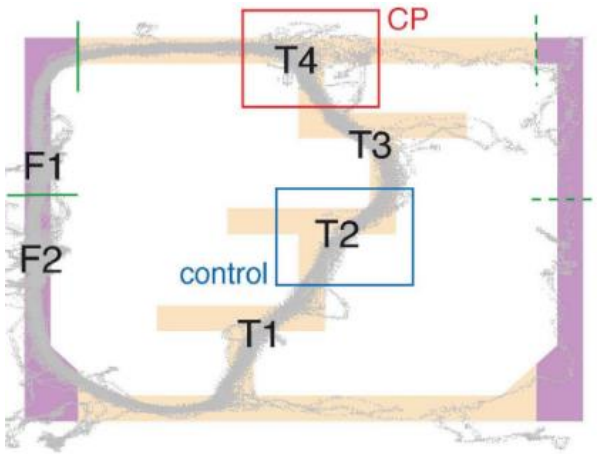


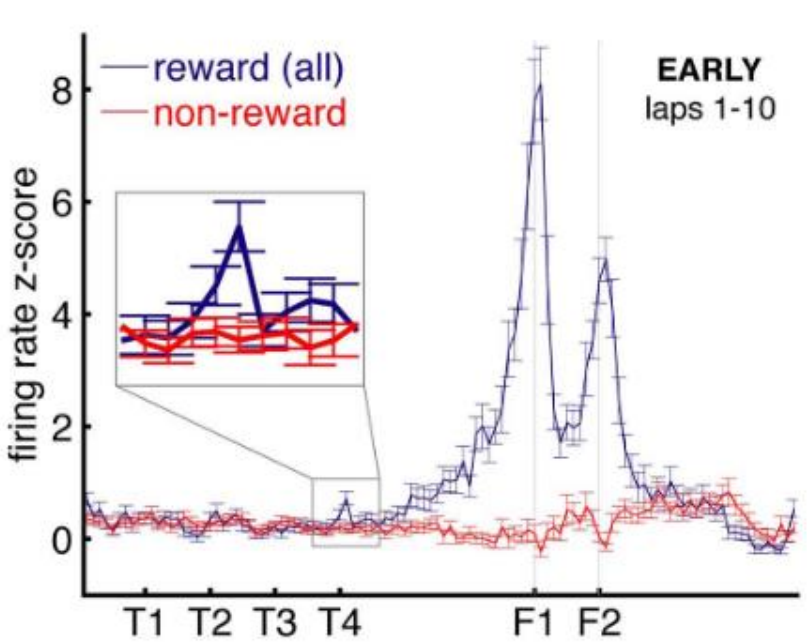
- “Deliberation” at important decision point during initial learning accompanied by theta sweeps down choice arms.

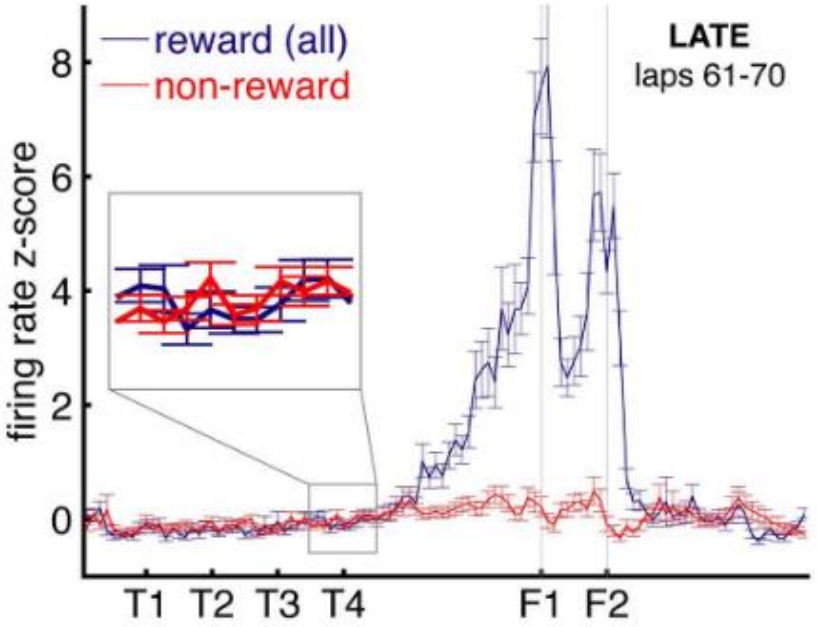
Van der Meer & Redish 2009

Johnson & Redish 2007

Hippocampal sequences: Theta oscillations



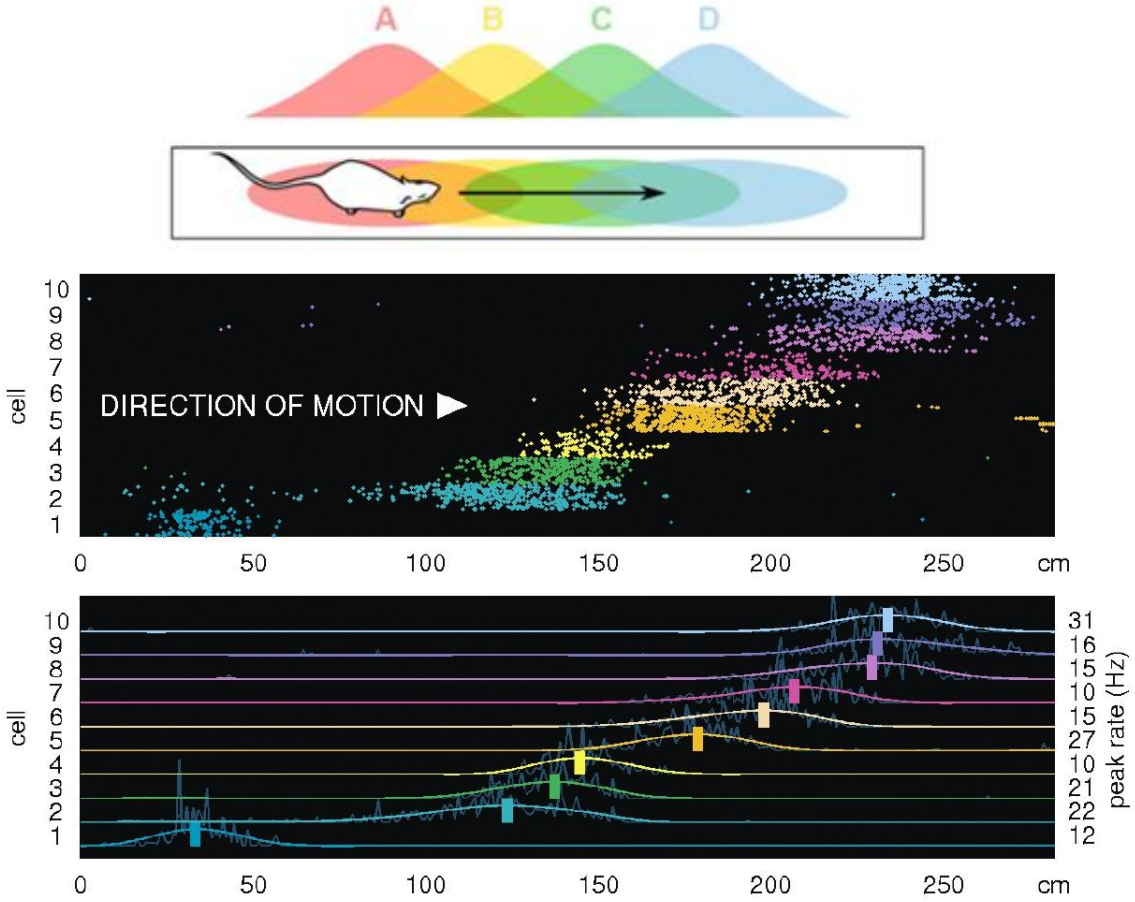


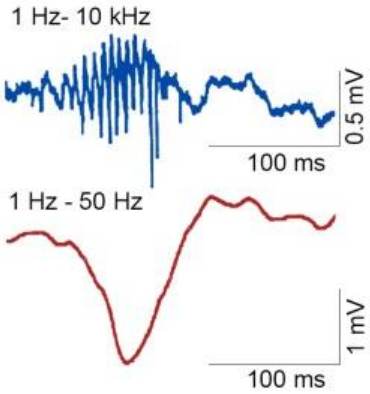


- "Deliberation" at important decision point during initial learning accompanied by theta sweeps down choice arms.
- Activation of "value" neurons in ventral striatum at decision point during early learning.
- Theta sweeps as online planning?

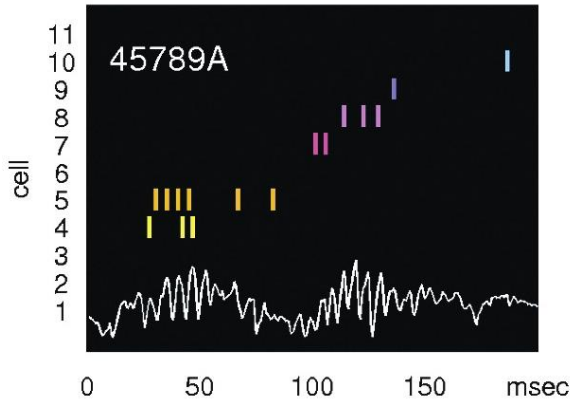
Van der Meer & Redish 2009

Hippocampal sequences: Sharp wave ripples





Sharp wave ripple during sleep

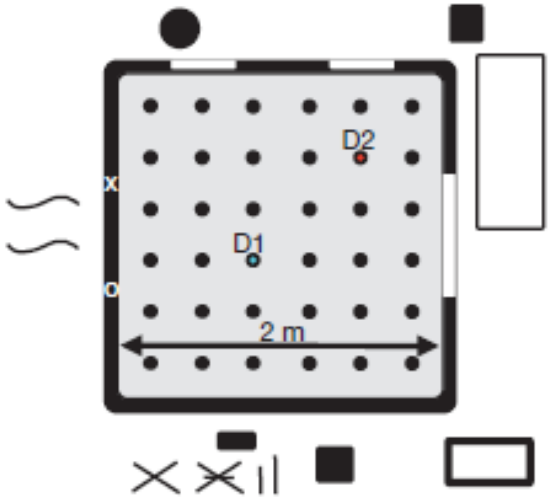


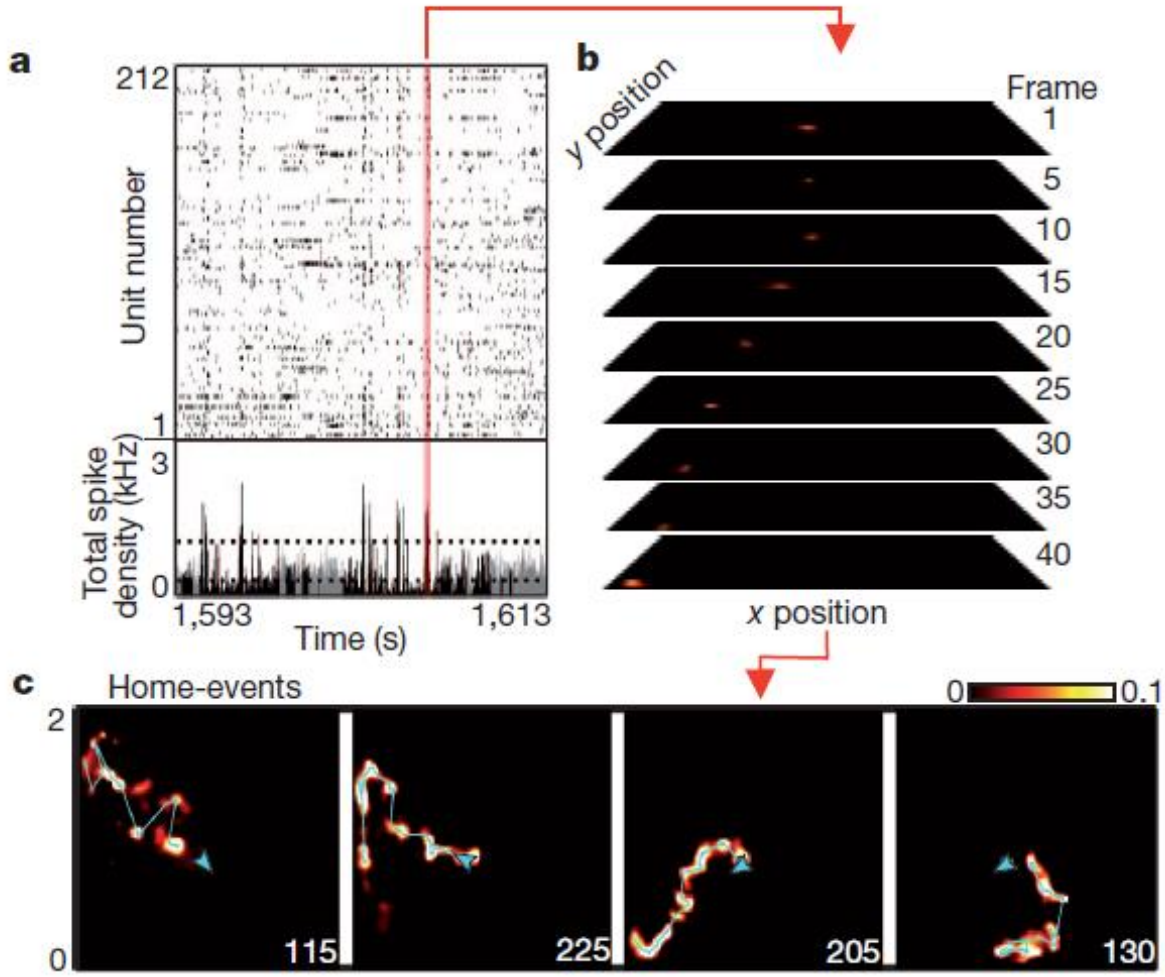
- “Replay” of behavioural sequences during sleep sharp wave ripples

Lee & Wilson 2002

Buzsáki 2015

Hippocampal sequences: Sharp wave ripples

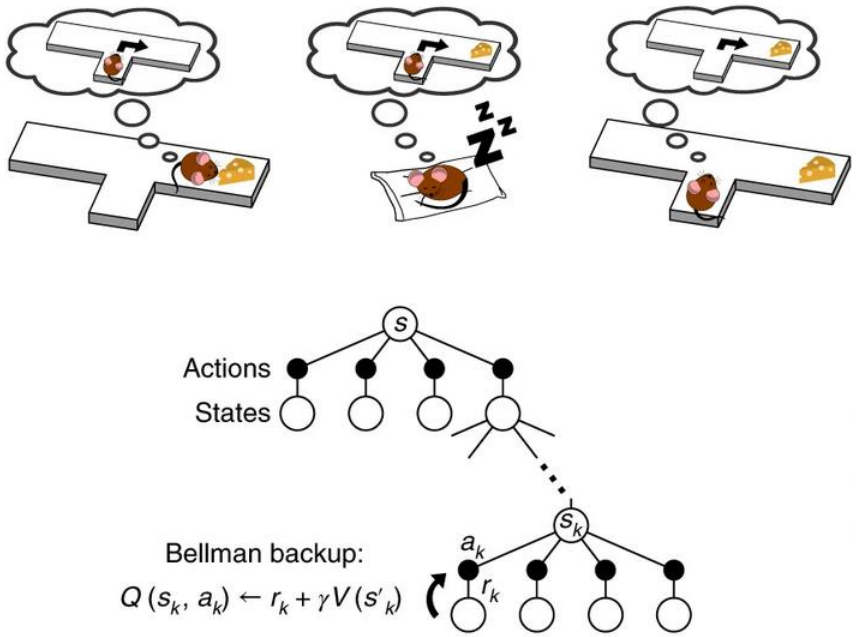


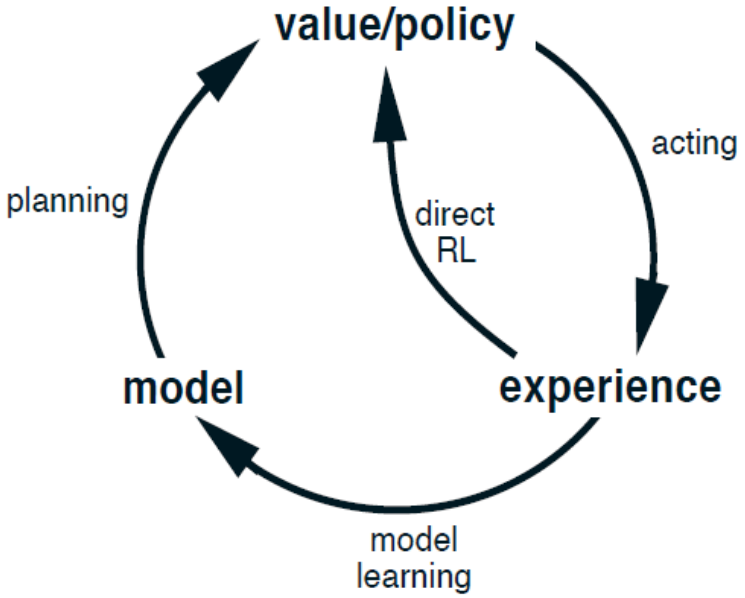


- Awake SWR sequences play our trajectories towards remembered goal locations before navigating to them.

Pfeiffer & Foster, 2013

Sharp wave ripples as DYNA

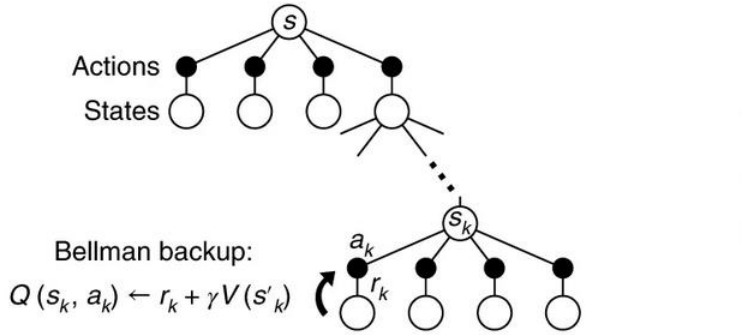




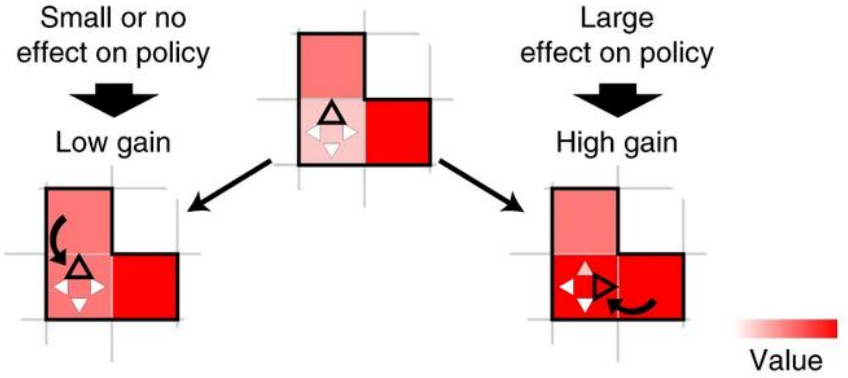
What state transitions should you replay if you are using them to do DYNA-like action-value updates?

[Mattar & Daw 2018]

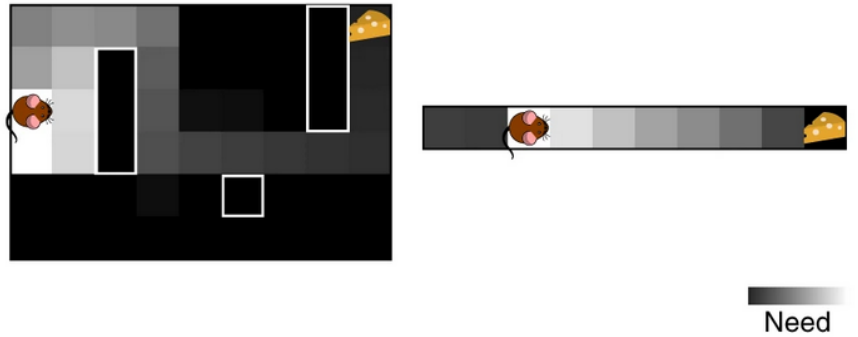
Sharp wave ripples as DYNA



Gain term



Need term

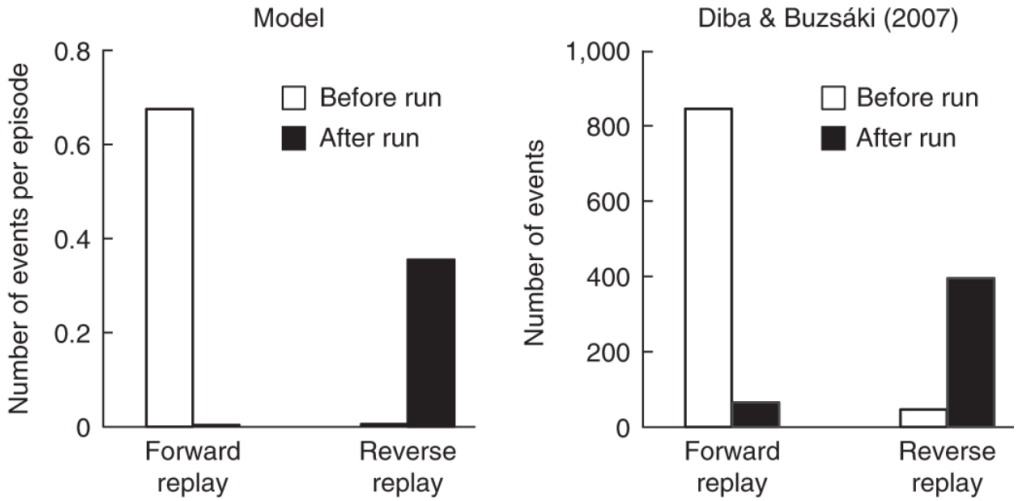


What state transitions should you replay if you are using them to do DYNA-like action-value updates?

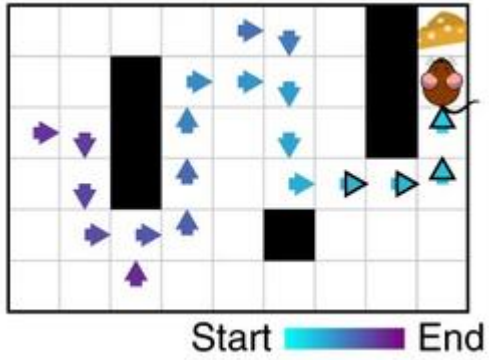
- Gain term: Replay transitions that will drive large value updates.
- Need term: Replay transitions that you are likely to take in the near future.

Mattar & Daw 2018

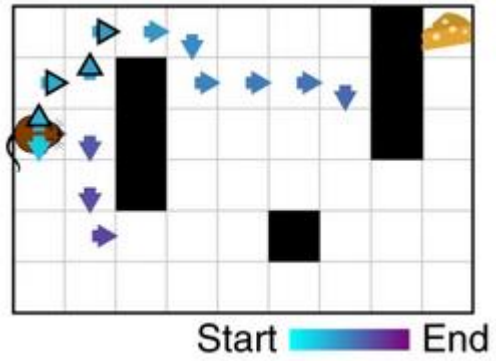
Sharp wave ripples as DYNA



Reverse replay after reward driven by Gain term



Forward replay before navigation driven by Need term

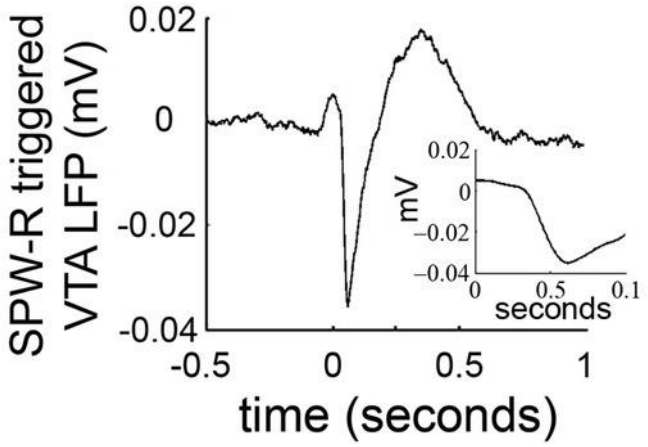


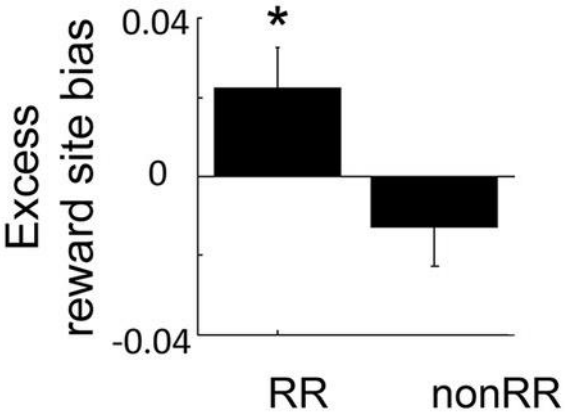
What state transitions should you replay if you are using them to do DYNA-like action-value updates?

- Gain term: Replay transitions that will drive large value updates.
- Need term: Replay transitions that you are likely to take in the near future.

Mattar & Daw 2018

Sharp wave ripples as DYNA

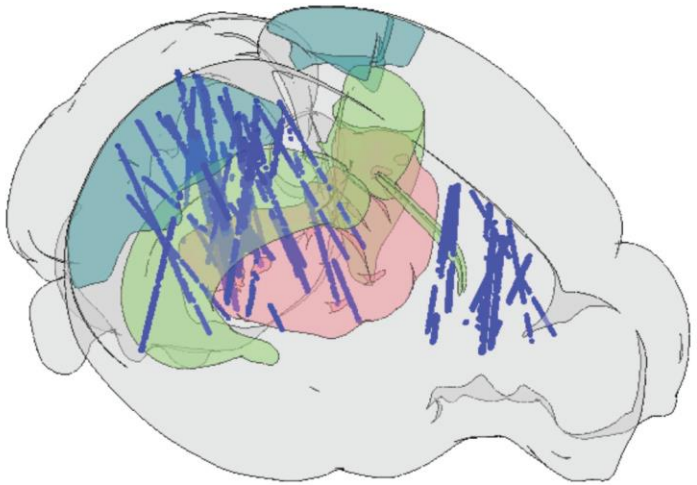


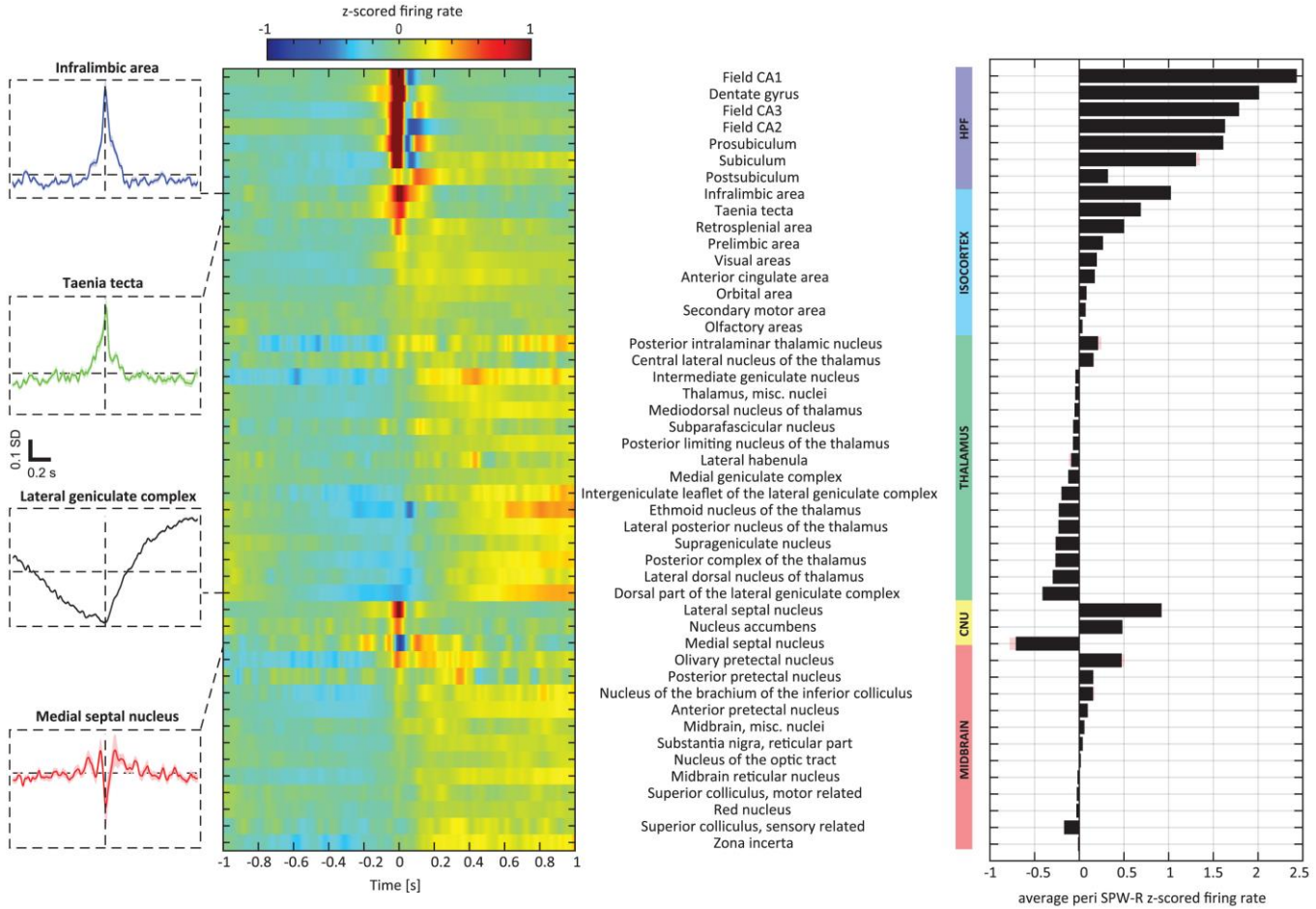


- Reward responsive neurons in VTA (putative dopamine neurons) are activated during SWRs

Gomperts et al. 2015

Sharp wave ripples engage brain-wide networks



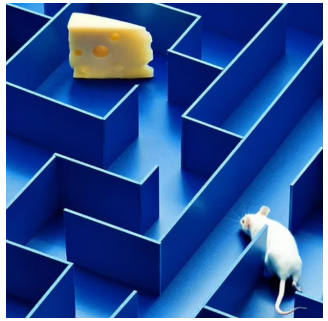


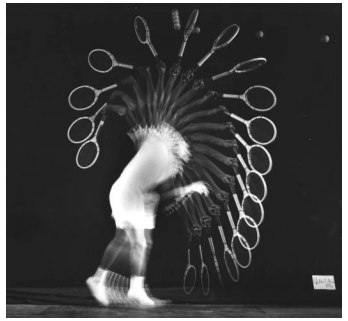
- Brain-wide activity is modulated around sharp-wave ripples.
- Ripples are thought to be important for consolidation of episodic memories from hippocampus into cortex.

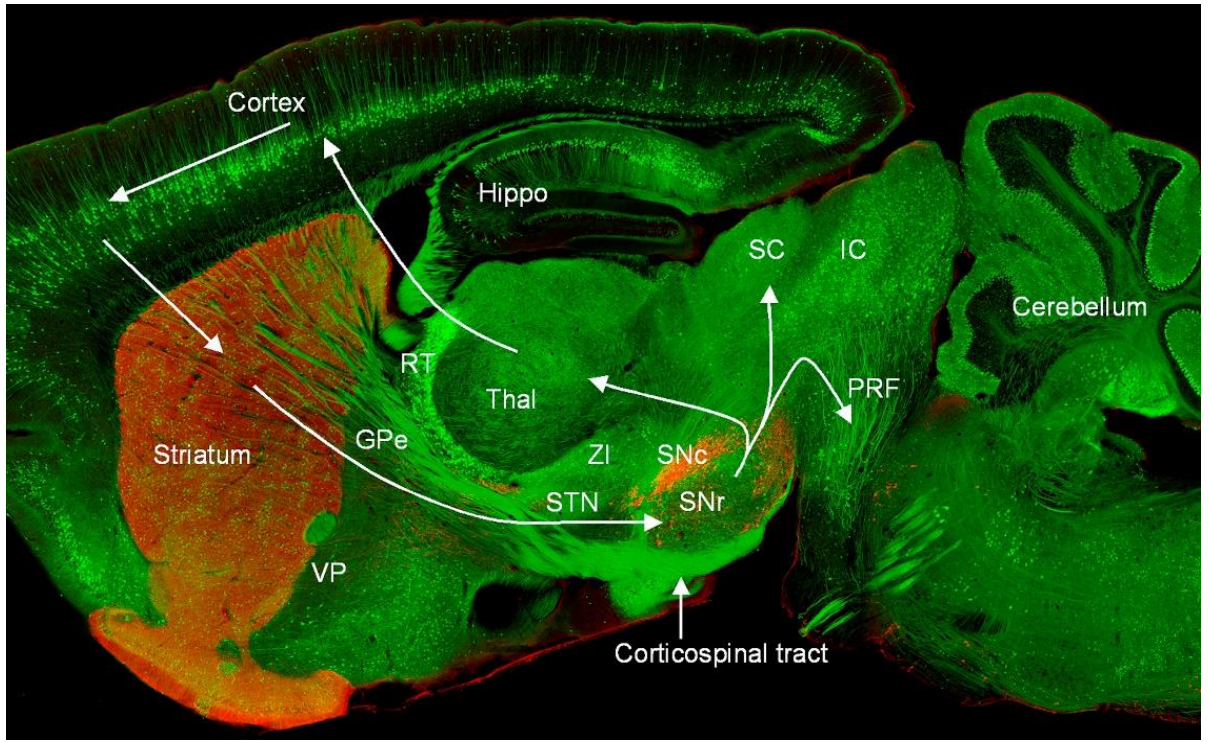
  → DYNA-like value learning is likely part of the story but not the whole picture.

[Nitzan et al. 2022]

Summary







- Brains evolved to solve a challenging control problem characterised by delayed rewards and a complex state, action and observation space.

- The brain’s solution likely combines aspects of temporal difference value learning, self-supervised representation learning, and model-based planning, distributed across a set of specialised brain networks implementing different algorithms.

Thank you for listening.