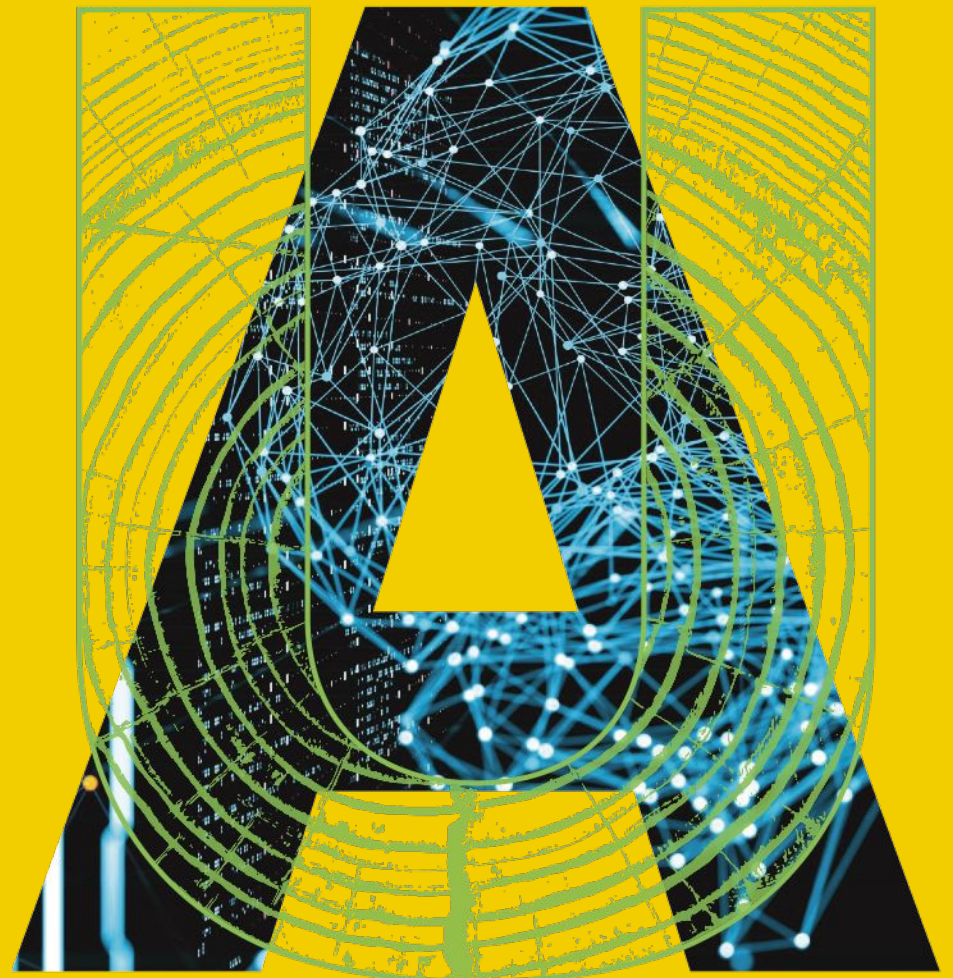
MACHINE LEARNING & THE BRAIN

Decision Making / Planning & Reinforcement Learning

Thursday 31 October 2023

Alex Murphy

UNIVERSITY OF ALBERTA



Today

05 September 2023: **Introduction to Neuroscience and Machine Learning**

12 September 2023: **The Visual System & CNNs**

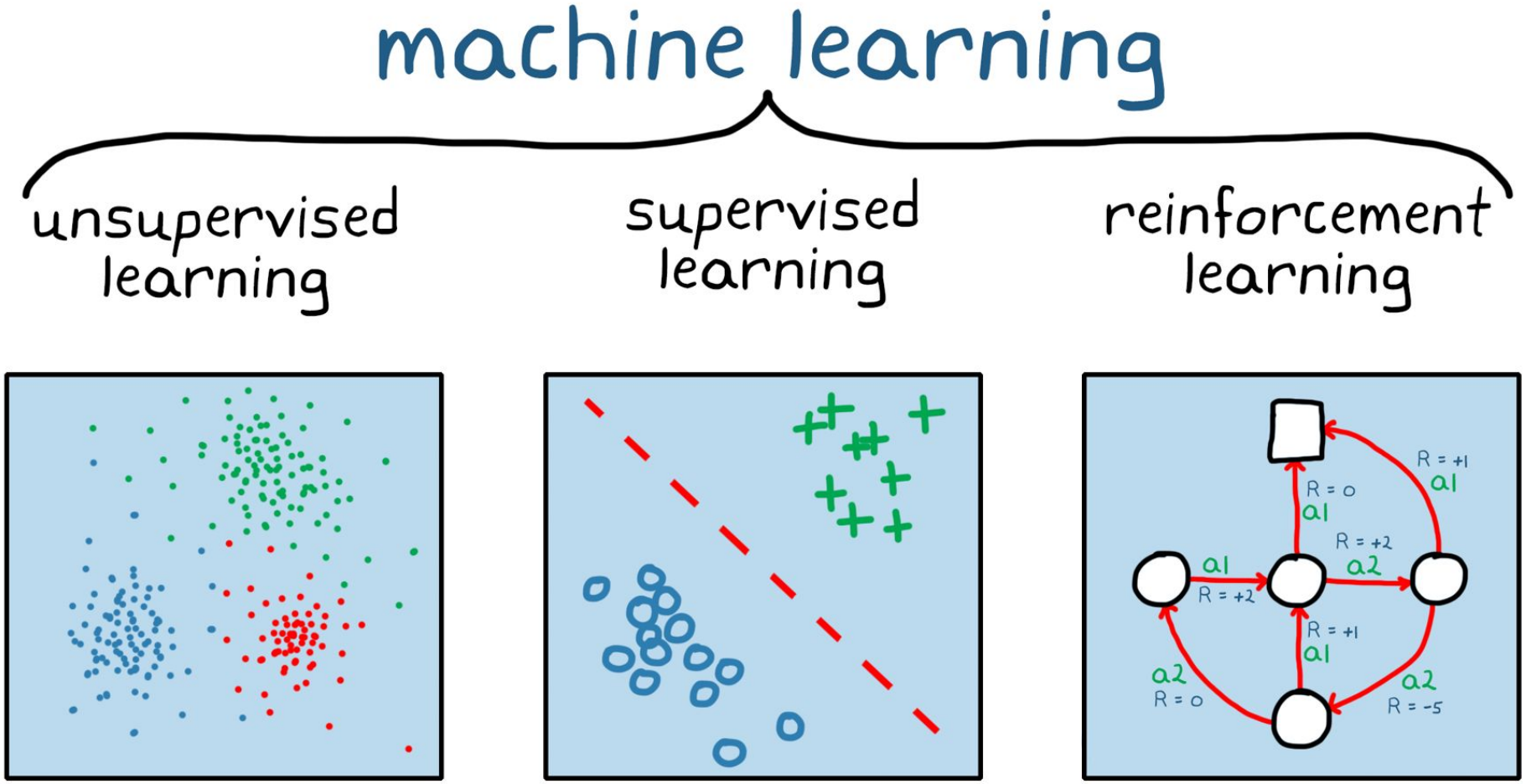
28 September 2023: **Coding Workshop**

05 October 2023: **Language Models & Language Neuroscience**

31 October 2023: **Decision Making / Planning & Reinforcement Learning**

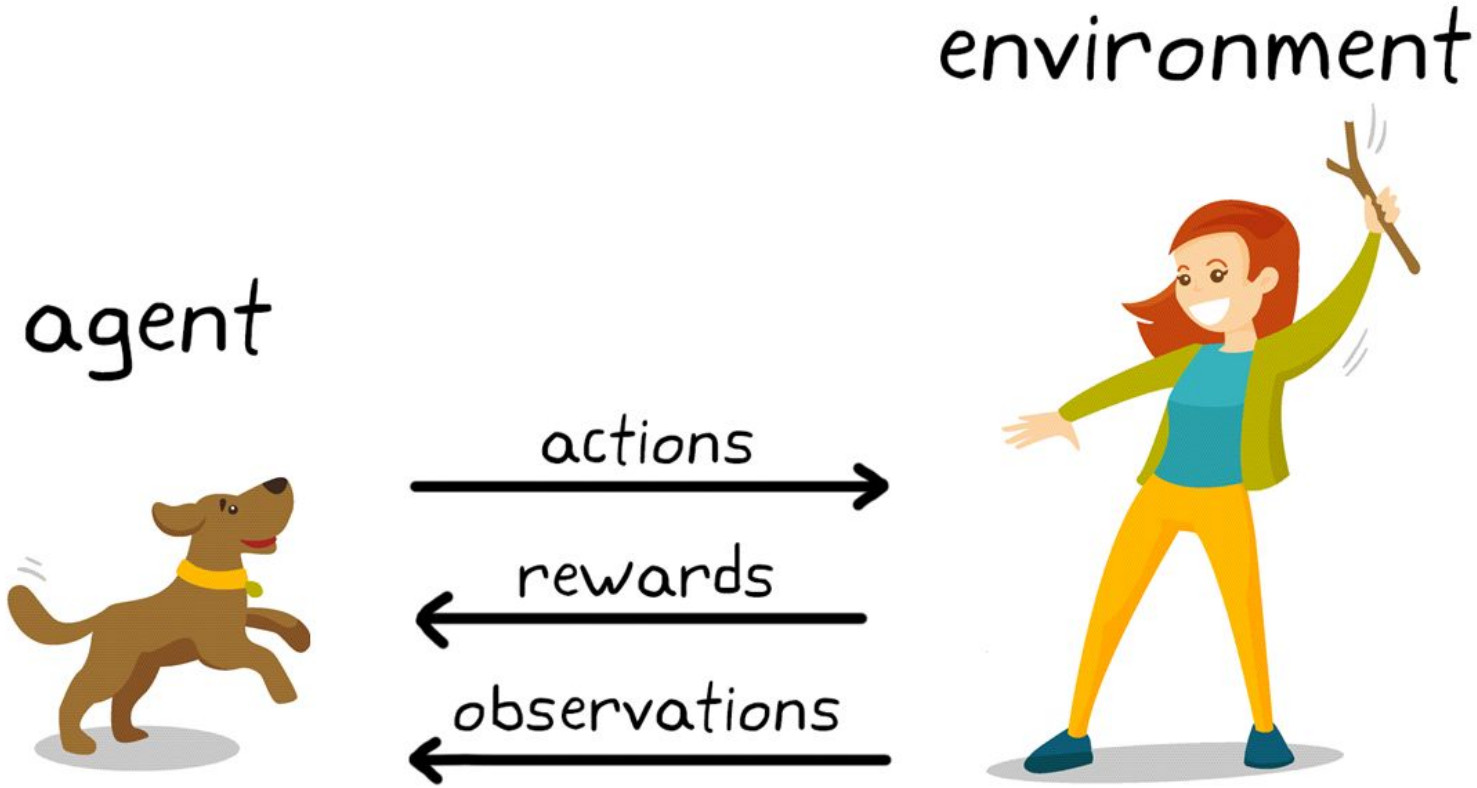
But First: RECAP

Reinforcement Learning



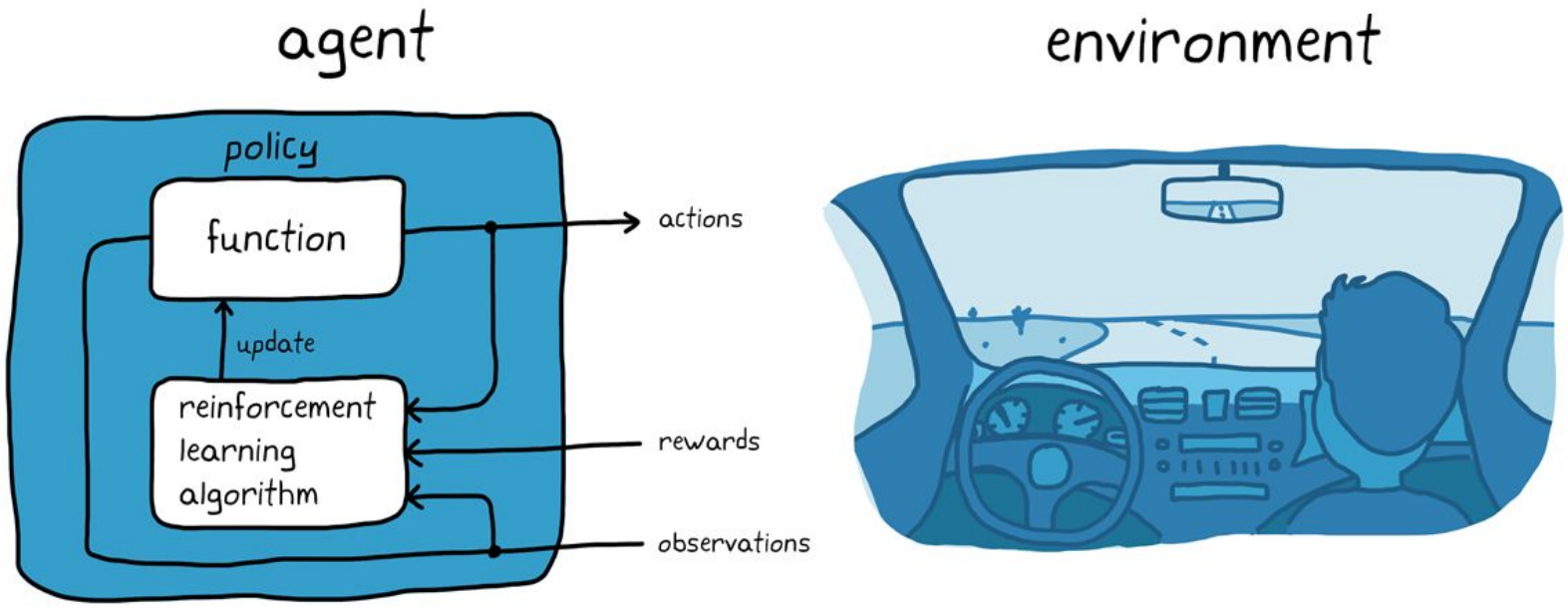
Great Tutorial: -> https://www.mathworks.com/discovery/reinforcement-learning.html (image source)

Reinforcement Learning



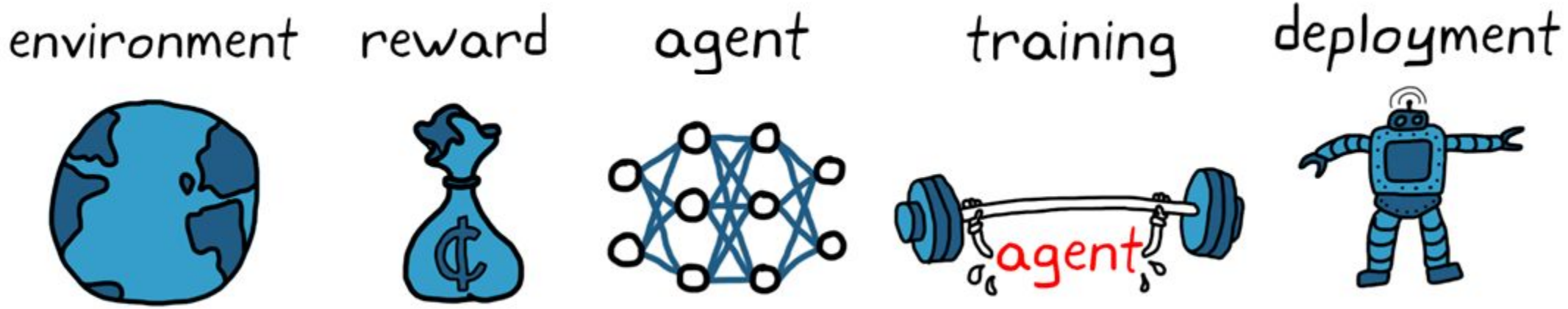
Great Tutorial: -> https://www.mathworks.com/discovery/reinforcement-learning.html (image source)

Reinforcement Learning



Great Tutorial: -> https://www.mathworks.com/discovery/reinforcement-learning.html (image source)

Reinforcement Learning



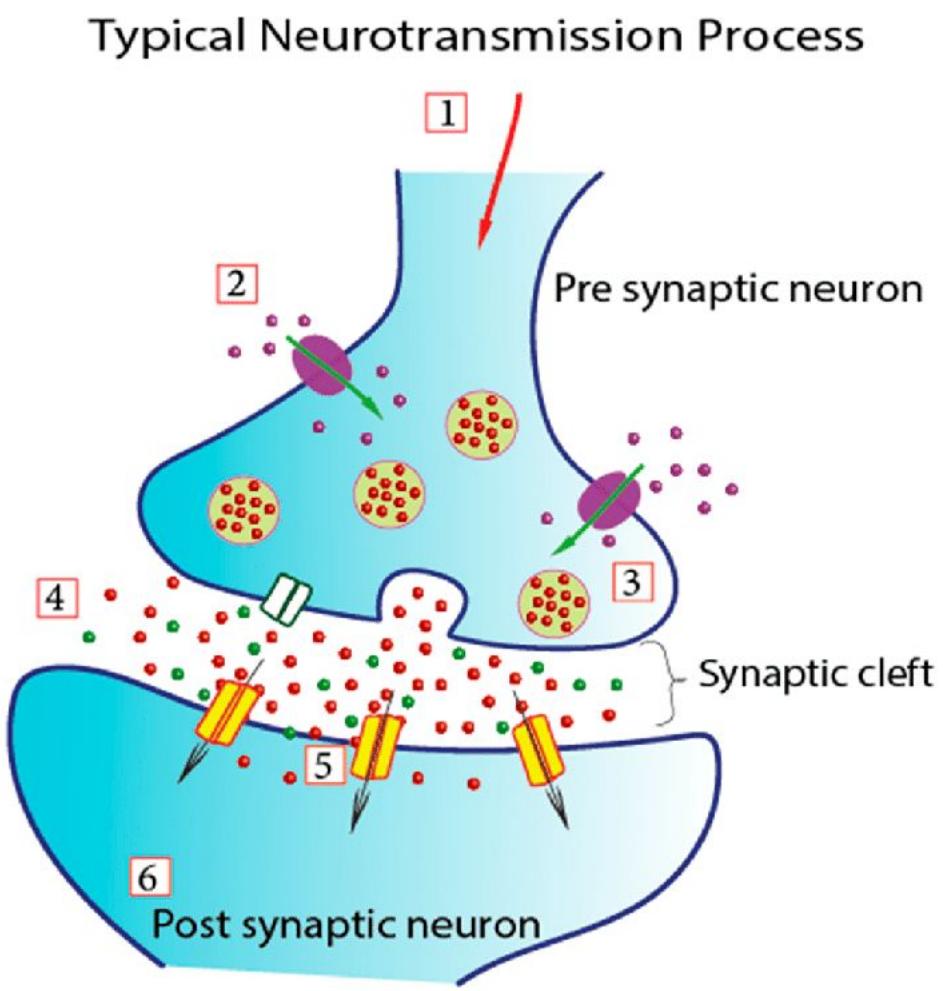
Great Tutorial: -> https://www.mathworks.com/discovery/reinforcement-learning.html (image source)

Reinforcement Learning

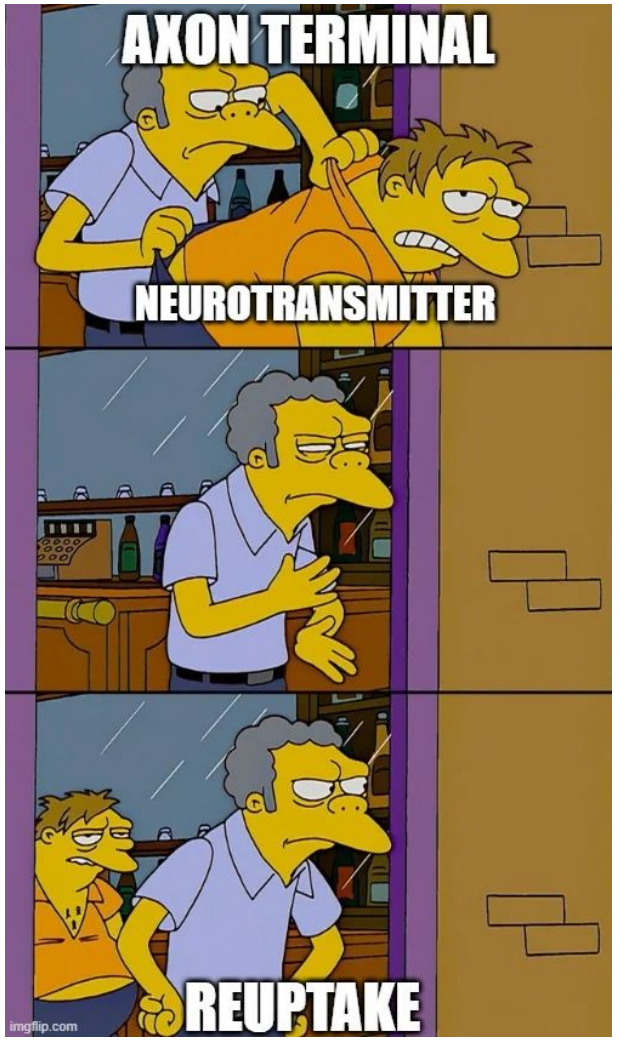
- Today we will talk about
  - How the brain learns
  - Dopamine reward system
  - ERN (EEG)
  - Explore / Exploit (fMR)
  - Some Brain-RL papers
  - Future of brain-inspired RL

Neurotransmitters

- Adrenaline
- Dopamine
- Serotonin
- Glutamate
- Endorphins
- GABA
- Acetylcholine
- Cortisol
- Noradrenaline



https://byjus.com/biology/neurotransmitter/



r/sciencememes

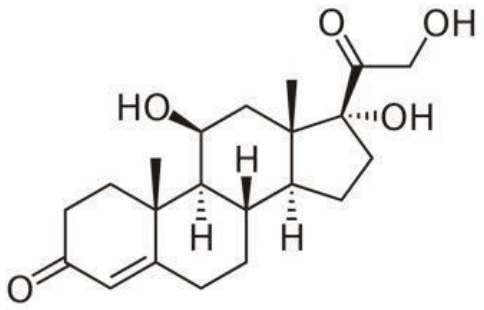
Neurotransmitters

Neurotransmitters relate to **specific functions**

- **Adrenaline**: fight or flight
- **Noradrenaline**: concentration
- **Dopamine**: pleasure
- **Serotonin:** mood
- **GABA**: calming
- **Glutamate:** memory
- **Cortisol**: stress
- **Acetylcholine:** learning / motion
- …

SEROTONIN

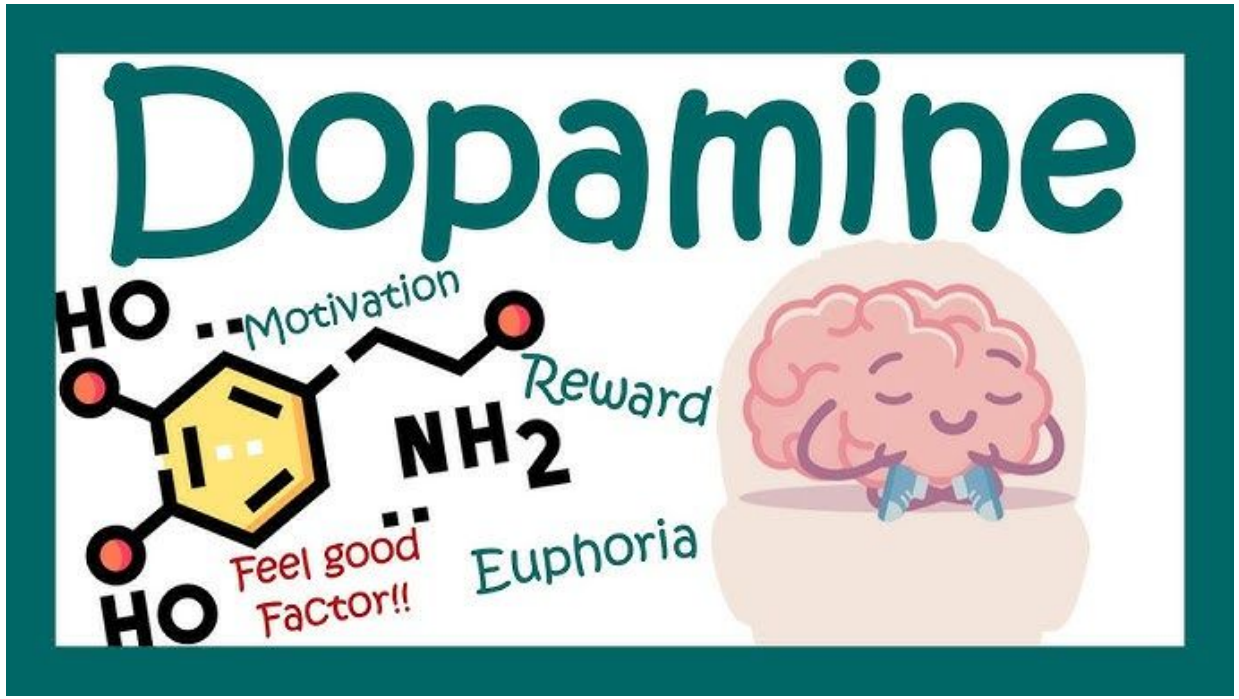
CORTISOL



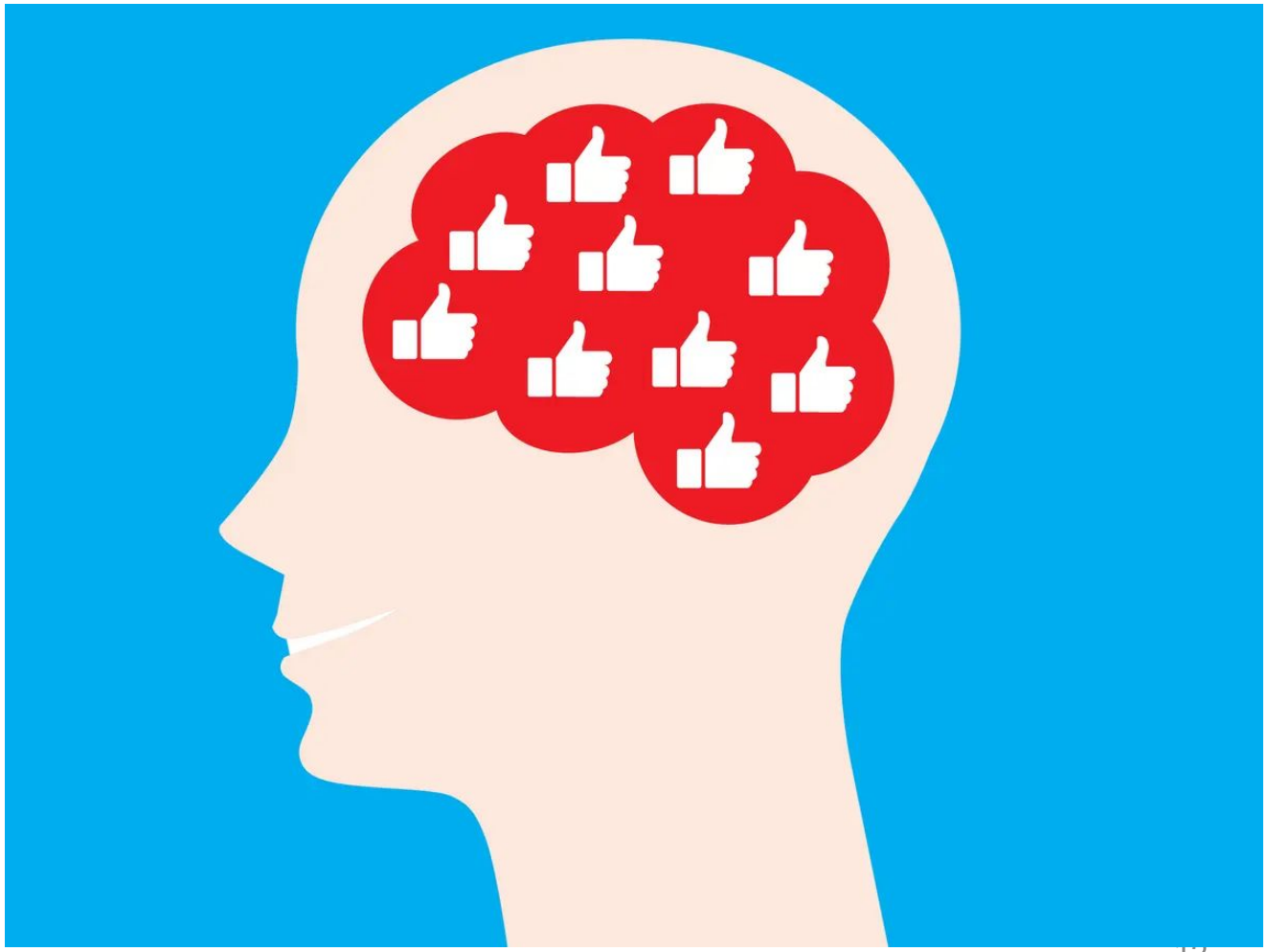
Neurotransmitters

Neurotransmitters relate to **specific functions**

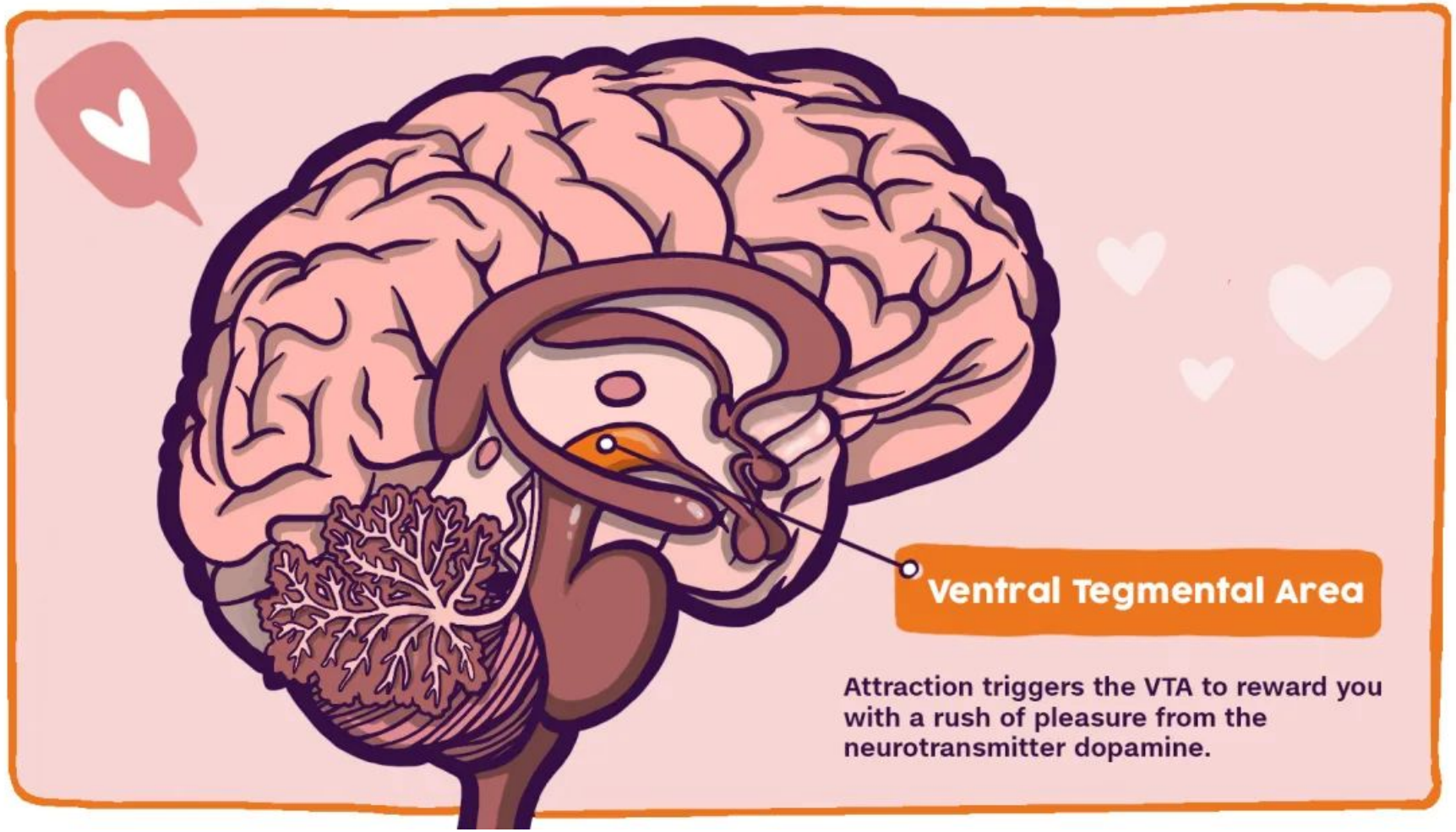
- **Adrenaline**: fight or flight
- **Noradrenaline**: concentration
- **Dopamine**: pleasure
- **Serotonin:** mood
- **GABA**: calming
- **Glutamate:** memory
- **Cortisol**: stress
- **Acetylcholine:** learning / motion
- …



Dopamine Loops



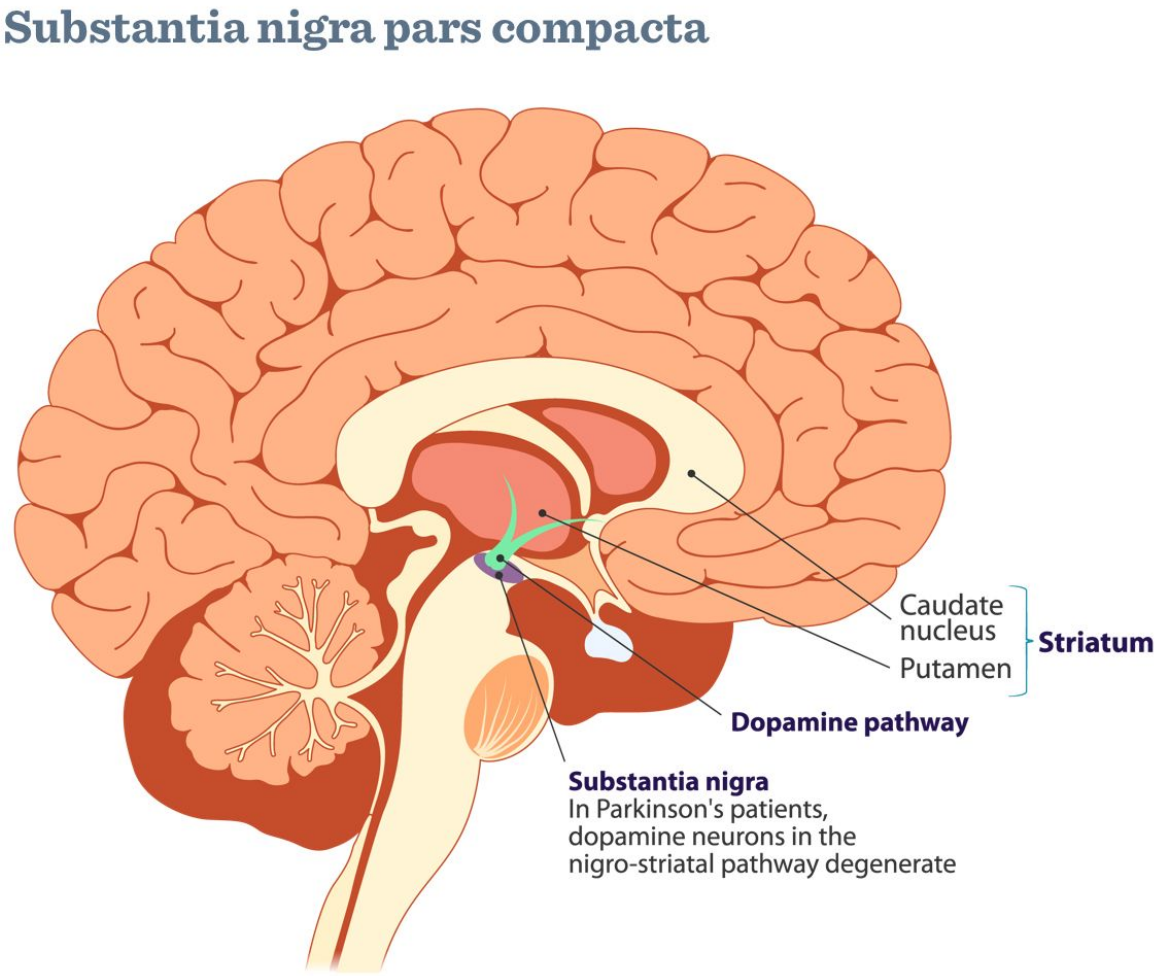
Ventral Tegmental Area



https://www.cnn.com/2020/02/14/health/brain-on-love-wellness/index.html

Substantia Nigra

- The second dopamine centre of the brain
- Means “Black Substance”
- Degeneration here results in Parkinson’s Disease
- Like VTA, sends dopamine (reward) signal to other parts of the brain



https://parkinsonsdisease.net/basics/pathophysiology-what-is-it

Reward Signals: Why Dopamine?

- Dopamine rules the brain's reward pathway
- The **mesolimbic reward system** increases dopamine
- Brain's : learn behaviours which stimulate secretion of dopamine
- Addiction is the reward system **gone into overdrive**
- Also implicated in **OCD**
- Complex reward prediction system connected increasing likelihood of taking actions for higher reward
- **Reinforcement Learning is inspired by the brain's reward system**

Reward System: History

Reward System: History

- Olds & Milner (1954)
- **Dopamine Hedonia Hypothesis**
- Proposed by Roy Wise
- Long-since rejected, but the idea of dopamine as the actual reward signal is widely understood in the general population
- 2nd quote in 1997, same year as the Schultz study
- RPE hypothesis proposed in 1996 by Montague et al.
  - See 15.3 in Sutton & Barto for more

> "dopamine junctions represent a synaptic way station…where sensory inputs **are translated into the hedonic messages we experience as pleasure,** euphoria, or 'yumminess'," (p. 94) (Wise, 1980).

> "I no longer believe that the amount of pleasure felt is proportional to the amount of dopamine floating around in the brain," (p. 35) (Wickelgren, 1997)

> "pleasure is not a necessary correlate of dopamine elevations" (p. 179) (Wise, 2008)

Quotes taken from Berridge & Kringelbach (2015), ***Pleasure Systems in the Brain***

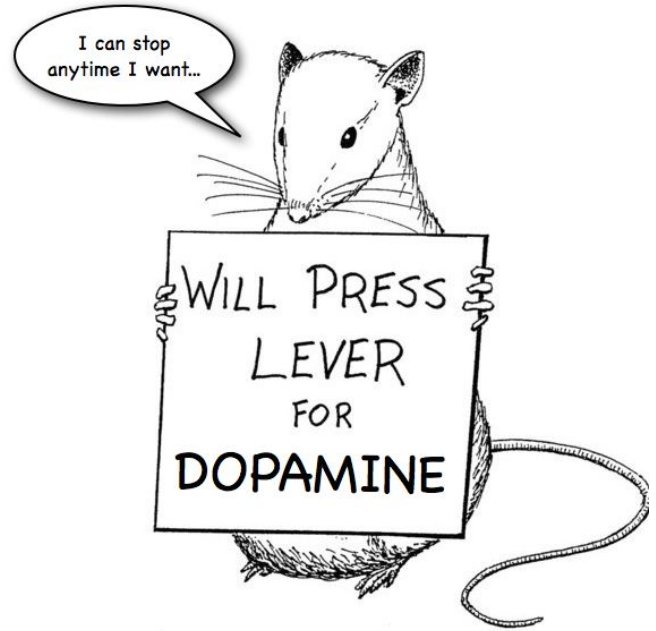
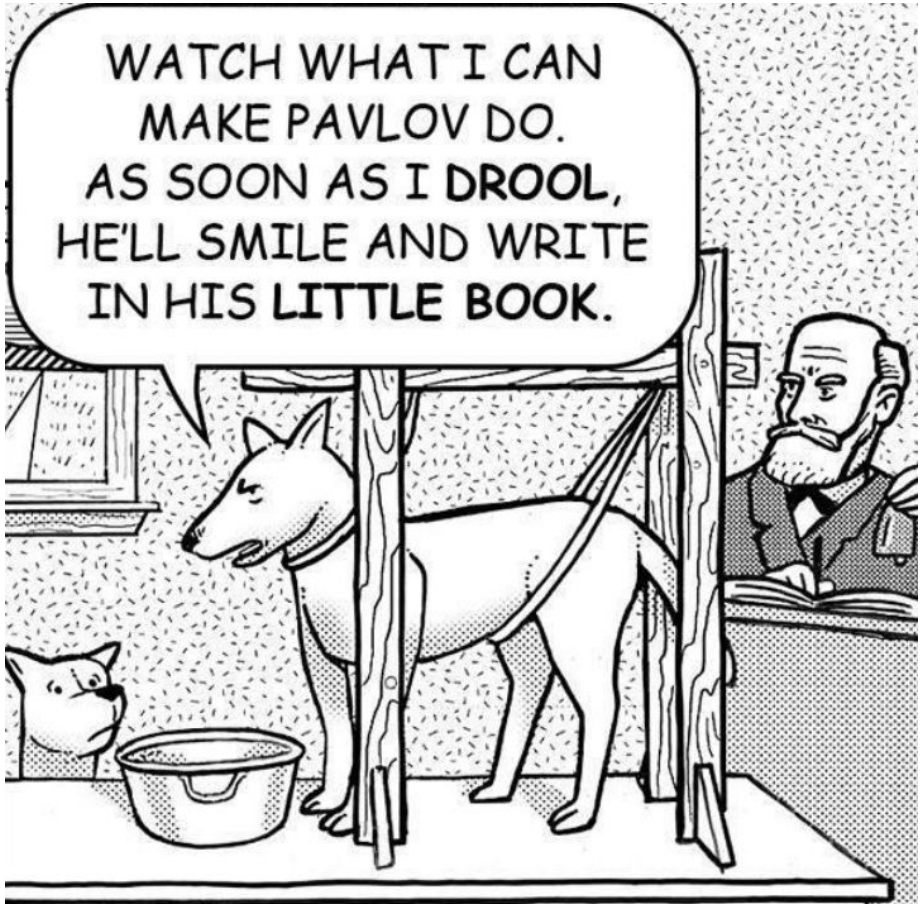


image credit: Craig Swanson; modified by Talia Lerner

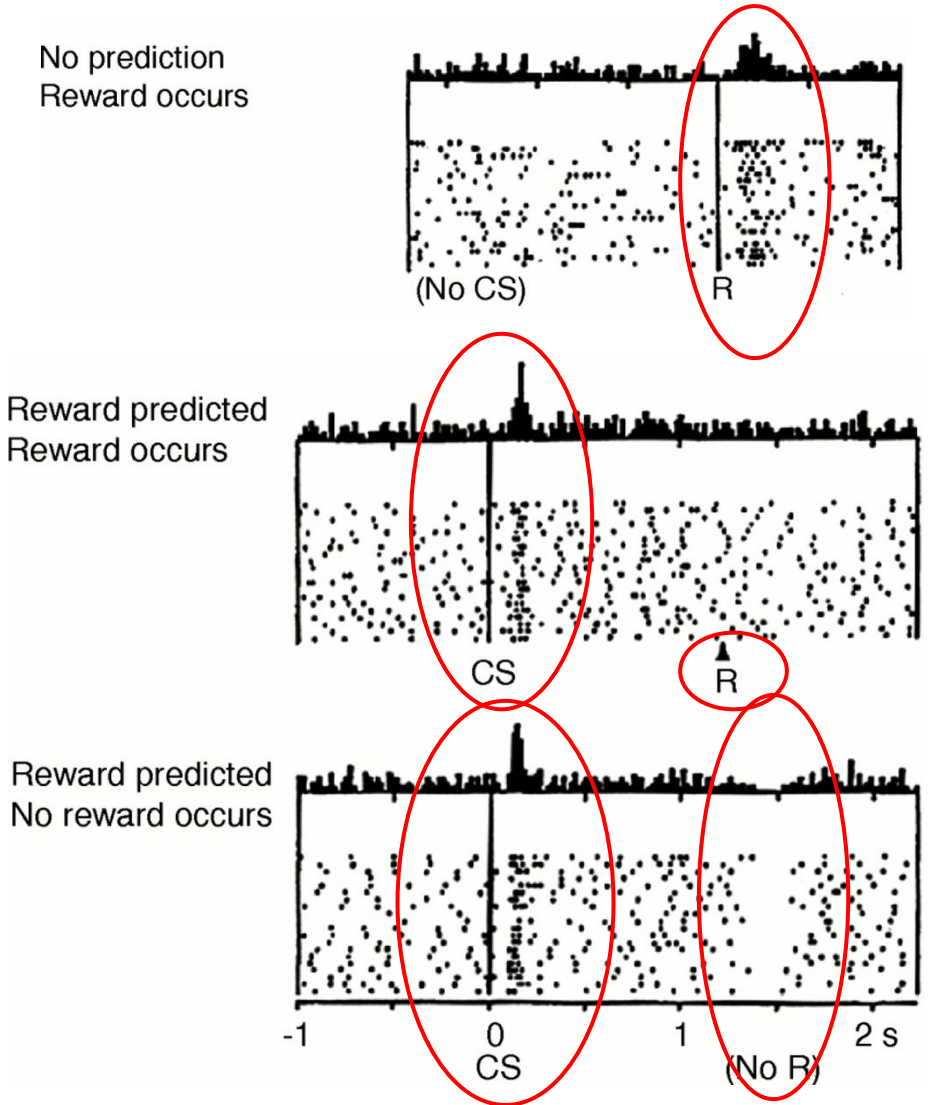
Reward System: History

Hypothesis: Dopamine release is concurrent with reward timing and signals pleasure reward

Shultz et al. in the 1990s performed some experiments where they introduced a reward cue.



Reward System: History

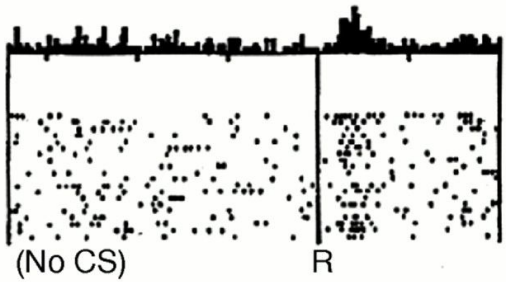


Reward-Prediction Error

Schultz et al. 1997

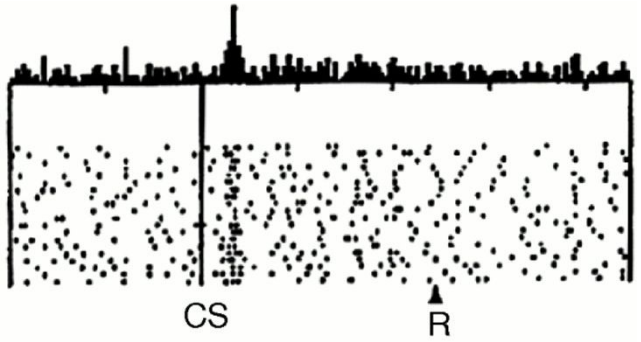
Reward System: History

No prediction
Reward occurs



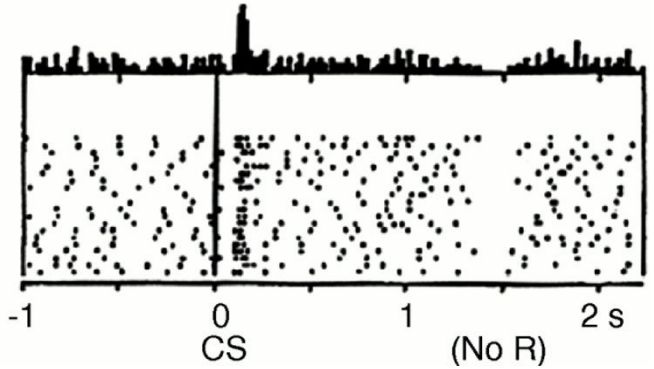
1 - 0 = 1

Reward predicted
Reward occurs



1 - 1 = 0

Reward predicted
No reward occurs



0 - 1 = -1

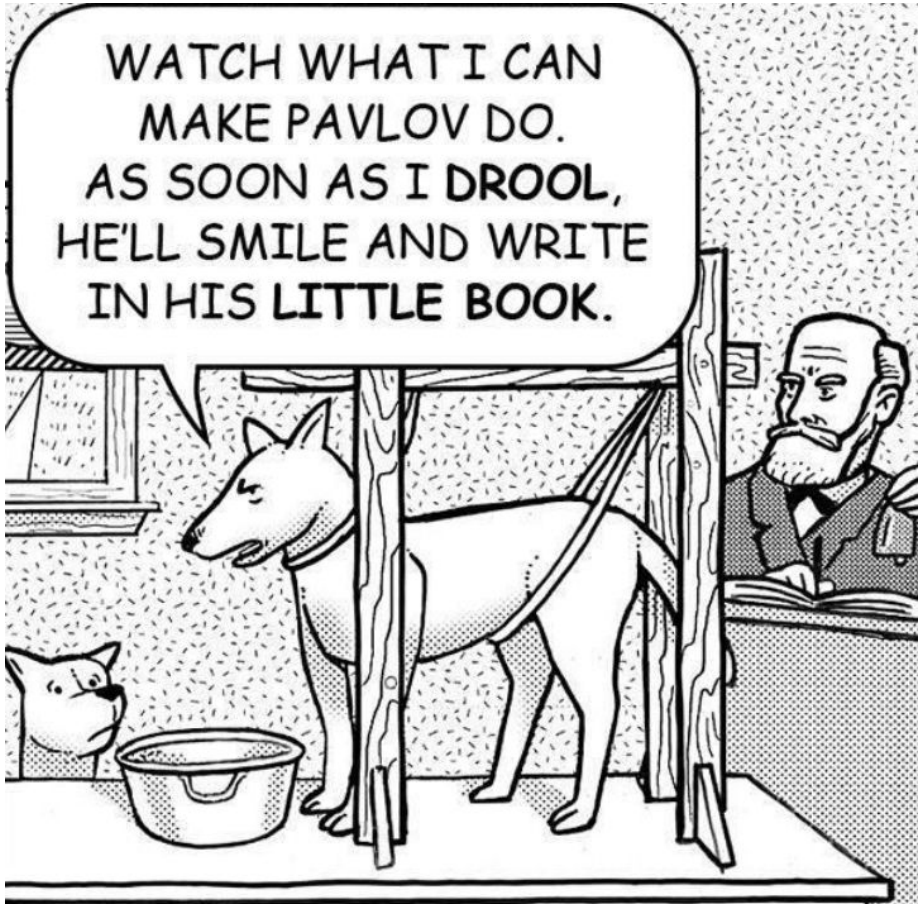
Reward Prediction Error (RPE) =
Actual Reward - Expected Reward

Reward System: History

Hypothesis: Dopamine release is concurrent with reward timing and signals pleasure reward

Hypothesis: Dopamine release is concurrent with reward prediction errors

Note: when we don't make a prediction but receive a reward, we have a negative prediction error so this shows us that our system needs to be updated so that in future, we are more likely to predict rewards similar to the one that just surprised us.



Reward System: TD Learning

- Paper points out dopamine's similarity to the method of **temporal differences** (TD)
- NN models use an **adaptive critic model** to output a TD error when own prediction changes
- TD error **positive** when reward **is better** than expected
- TD error **negative** when reward **is worse** than expected
  - i.e. *absence of expected reward*

- **Actor model** selects actions to reinforce behaviours that elicit reward
- TD errors **propagate back** from moment of reward to **conditioned stimulus during learning**
- The parallel to dopamine function is strong and became an influential hypothesis
  - though not universally accepted

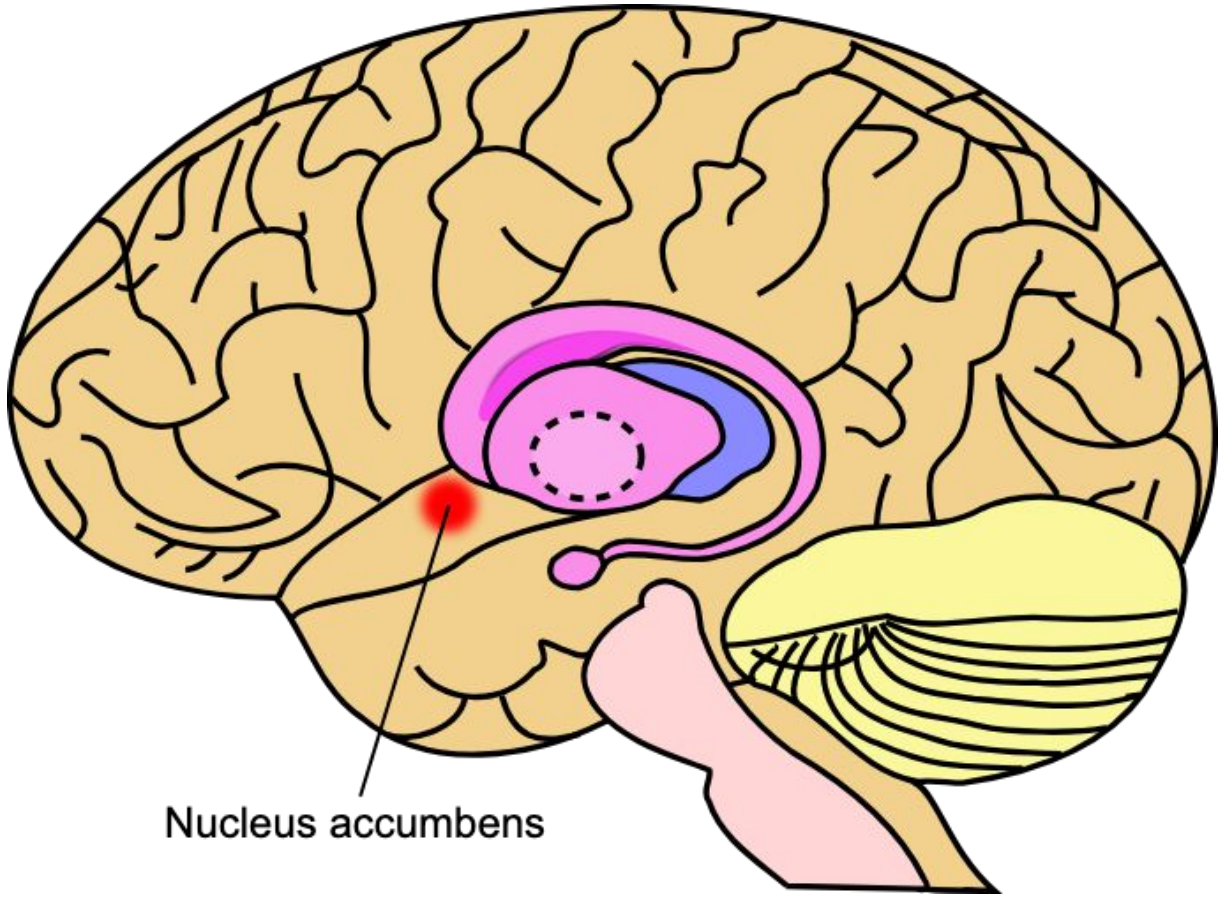
We'll come back to this …

Nucleus Accumbens

Nucleus Accumbens

"The nucleus accumbens is considered as the neural interface **between motivation and action**, playing a key role on feeding, sexual, reward, stress-related, drug self-administration behaviors, etc."

(Fernández-Espejo E. Cómo funciona el nucleus accumbens? [How does the nucleus accumbens function?]. Rev Neurol. 2000 May 1-15;30(9):845-9. Spanish. PMID: 10870199)



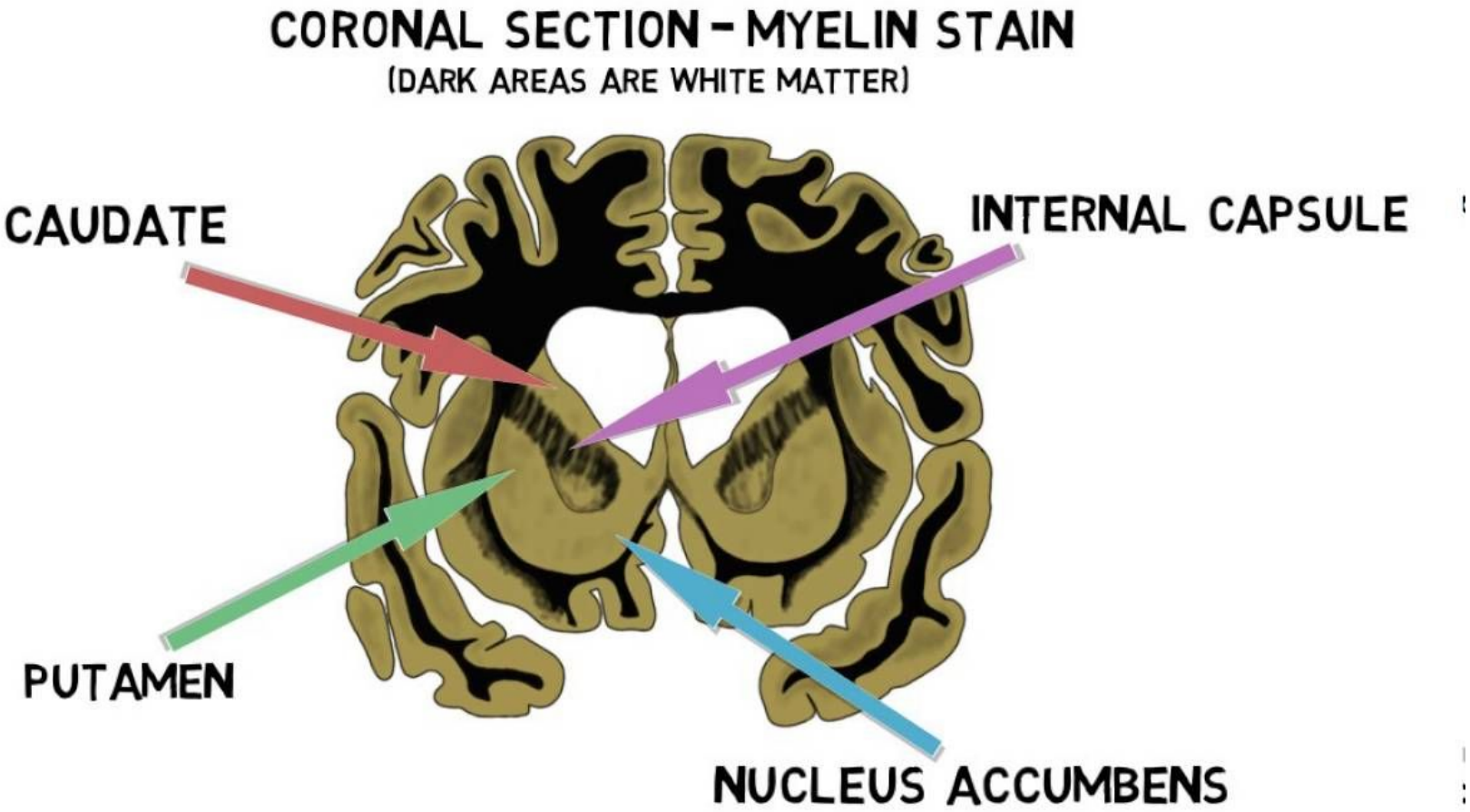
https://en.wikipedia.org/wiki/Nucleus_accumbens

Striatum (Value Calc. & Action Sel.)

- Many dopamine neurons have forward pathways to the striatum
- Cortex (high-level functions and motor planning info etc.) project to the striatum as well

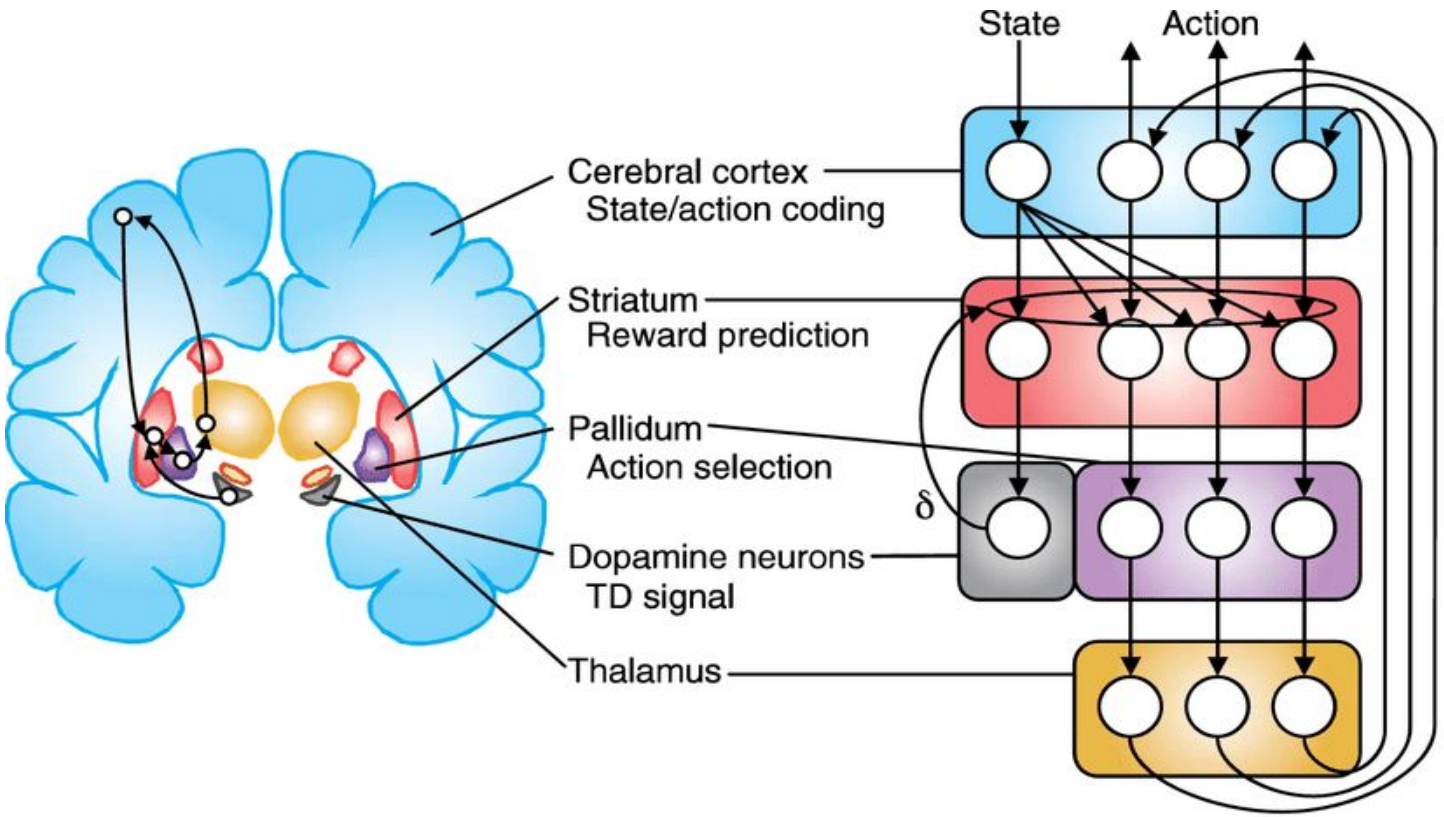
Role in RL

- Keeps track of **state and action values**
- Combines predictions and current motor / action plans from the cortex with the reward prediction error from the dopamine inputs and **integrates** the information together
- **Selects best action** according to this information and **sends output back to the cortex**



https://www.youtube.com/watch?v=EEUxKFmIUiI

Reinforcement Learning (Brain)



(*) Some of the most recent Brain-RL research says the Striatum is actually the “Action Selection” region, **not the Pallidum.**

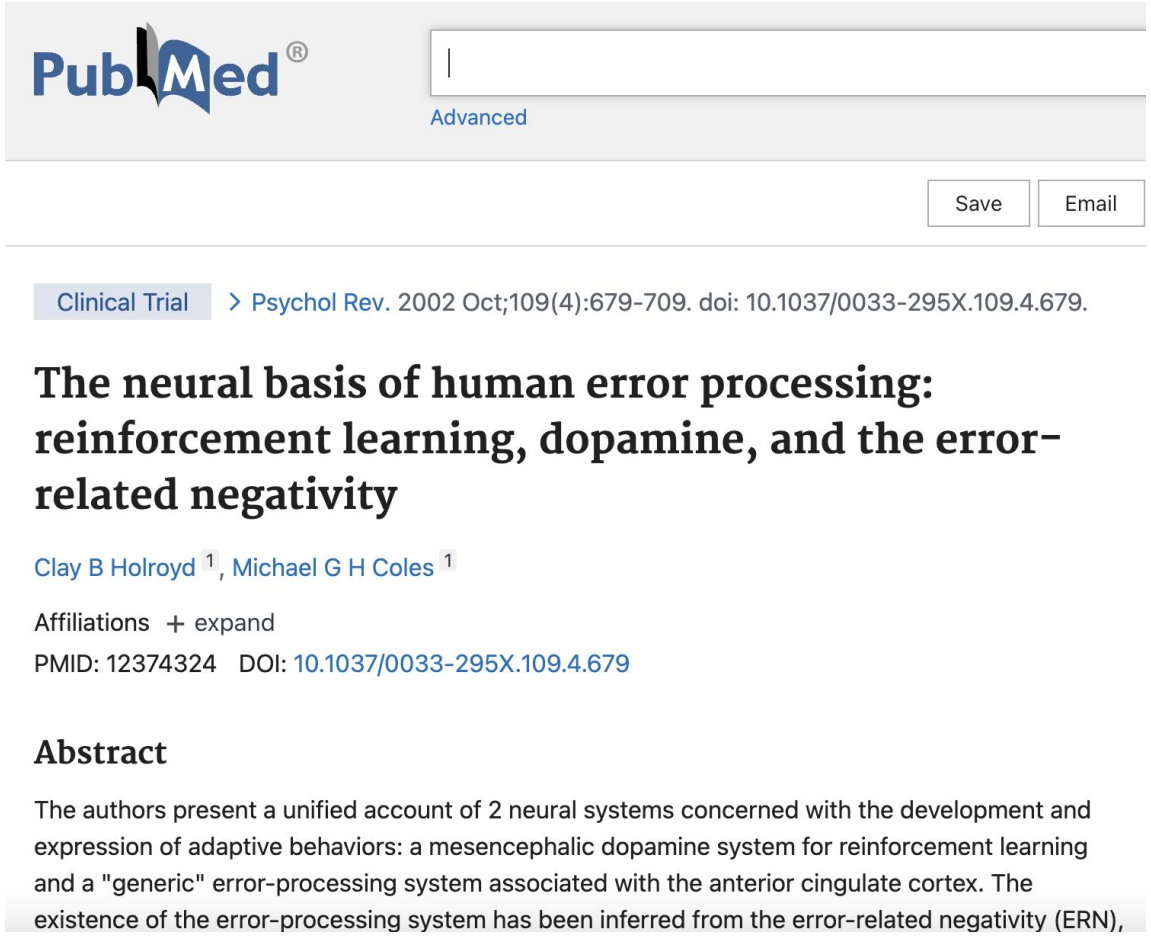
Figure 2 A hypothetical model of realization of reinforcement learning in the cortex–basal ganglia network[2]. Left, coronal section of the brain. Right, functional model, where δ denotes the reward prediction error carried by the midbrain dopamine neurons.

Complementary roles of basal ganglia and cerebellum in learning and motor control (Doya, 2000)

Neural Basis of Human Error Processing

5 years after Schulz et al. (1997)'s observations about dopamine being connected to reward prediction errors. An attempt to unify two components of human error processing that had emerged in the literature.

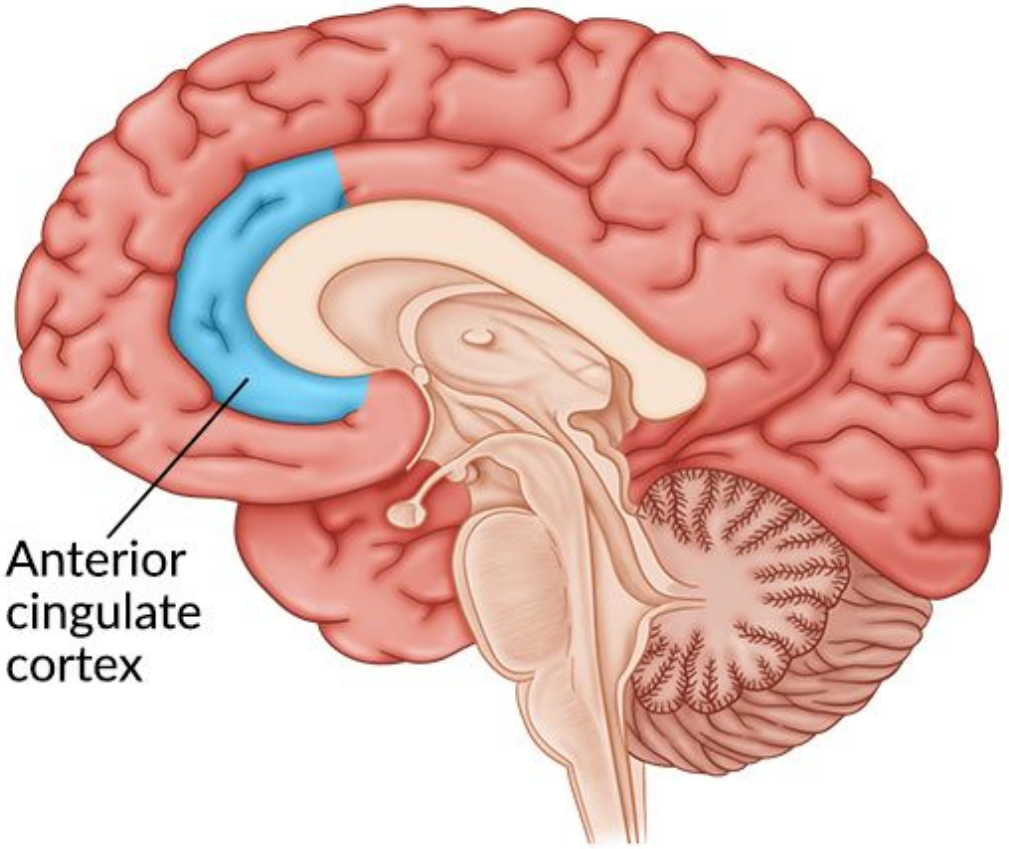
> *In this article, we propose a hypothesis that unifies the two accounts by explicitly linking the generation of the ERN to the activity of the mesencephalic dopamine system. Specifically, we suggest that when human participants commit errors in reaction time tasks, the mesencephalic dopamine system conveys a negative reinforcement learning signal to the frontal cortex, where it generates the ERN by disinhibiting the apical dendrites of motor neurons in the anterior cingulate cortex.*



Anterior Cingulate Cortex (ACC)

- Processing of reward and monitoring feedback
- Anticipation of reward
- Where motor intentions are transformed into action

We will come back to this shortly, this is just to give you an idea where the ACC is in the brain.



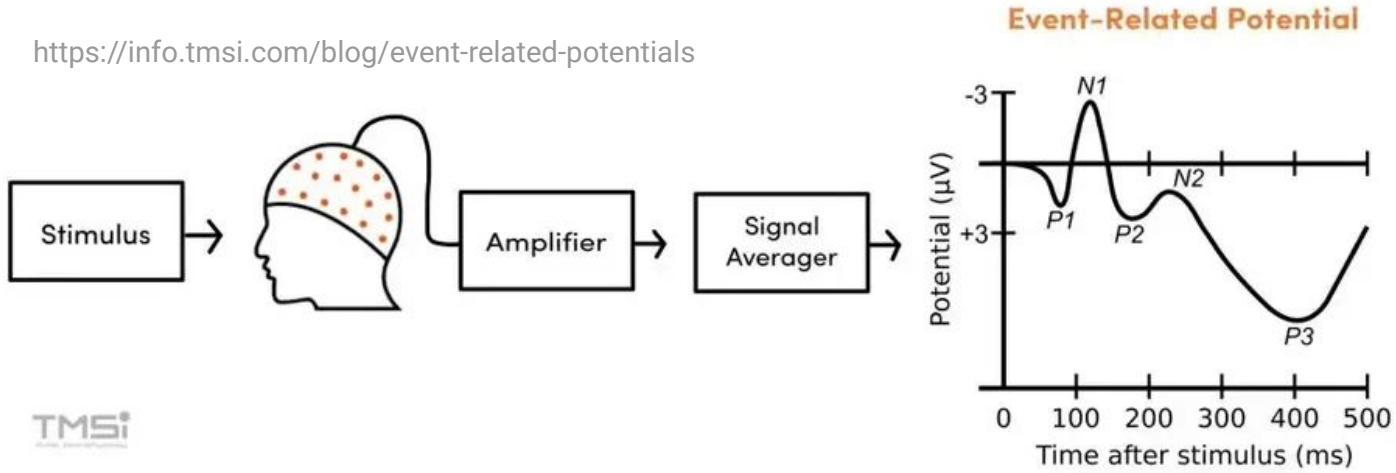
https://www.flintrehab.com/anterior-cingulate-cortex-damage/

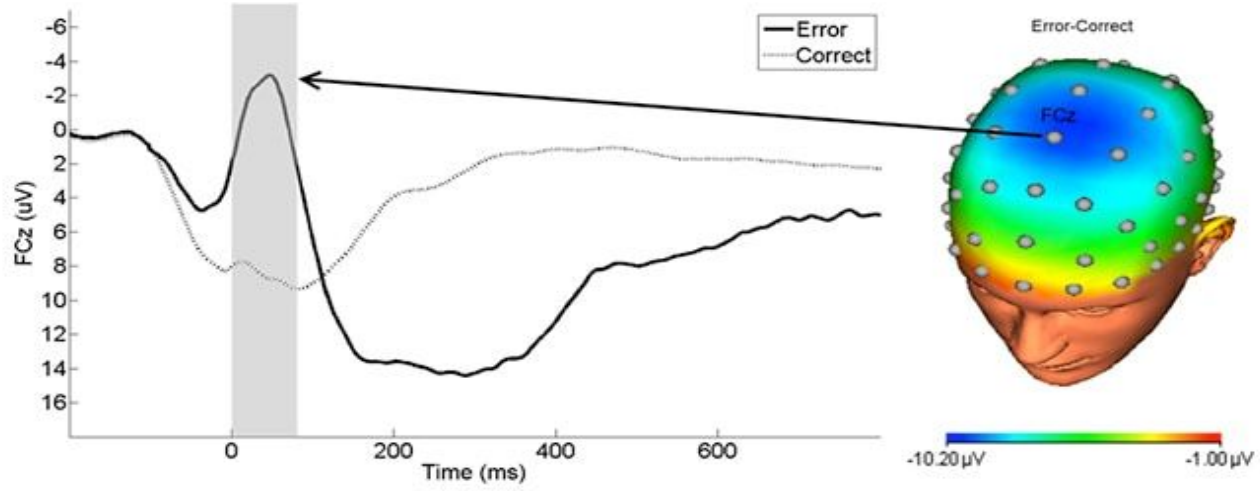
Error Related Negativity (ERN)

- Remember ERPs from coding workshop
- … and from the language lecture
  - N170 (faces)
  - N400 (semantics)
  - P600 (syntax[ish])

- Time-lock EEG recordings with stimulus
- ERN hypothesised to be **generated in ACC**
- 50-80 ms after error occurred

https://info.tmsi.com/blog/event-related-potentials





https://www.youthvoices.live/the-relationship-between-anxiety-and-error-related-negativity-across-development/

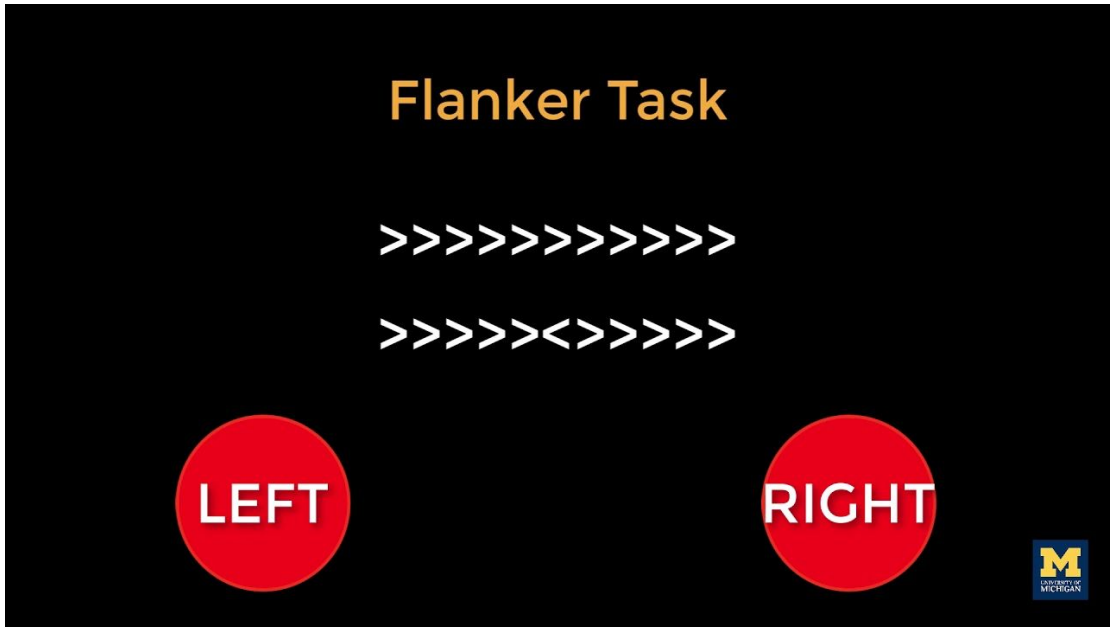
Error Related Negativity (ERN)

How do you elicit the ERN in an experiment?

"When the participant presses the incorrect button, the ERN is present, ***even when the individual is not aware they have made a mistake****. The size of the ERN is a direct reflection of the cost, or value, of the error."*

"When errors are more costly and when performance is being evaluated, the ERN is larger" (Olvet & Hajcak, 2008)

Task: focus on the **middle arrow**. The middle arrow is flanked by stimuli designed to confuse the participant. **Press the button that states the direction of the middle arrow.**



https://www.youtube.com/watch?v=SKuYVFgyWf8

Olvet DM, Hajcak G. The error-related negativity (ERN) and psychopathology: toward an endophenotype. Clin Psychol Rev. 2008

Error Related Negativity (ERN)

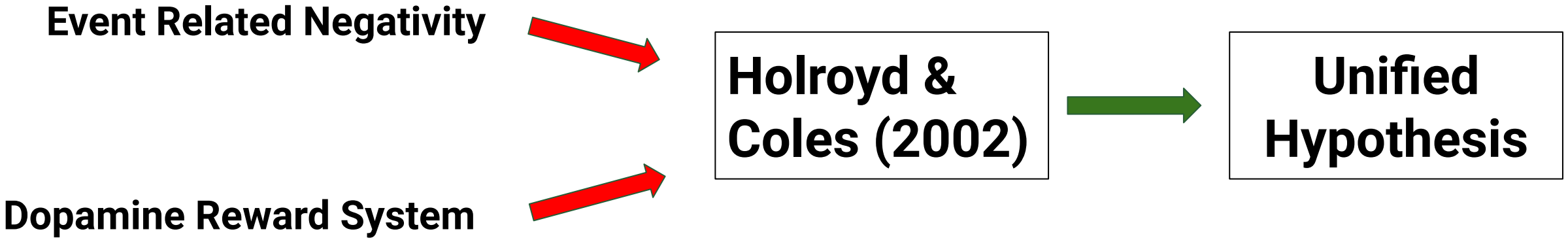
- ERN is generic and high-level
  - **generic**: applies in various tasks where errors are committed
  - **high-level:** relates to cognitive control and executive function

- ERN is sensitive to reward error
  - participants were differentially paid for their accuracy on tasks and **when the errors costed them more money, their ERN responses were greater**

- ERN is the same whether participant action given by hands, eyes, feet, verbally and also when error feedback is given back to participant visually, aurally or via somatosensory feedback (i.e. touch)

- Also observed in monkey research as far back as the 1970s (15 years before the advent of fMRI!)

Neural Basis of Human Error Processing



Hypothesis:

- mesencephalic dopamine system carries predictive error signals
- these signals are used by the brain to carry out reinforcement learning
- specifically, dopamine signals are passed to ACC in order to update future predictions with reward prediction error, which causes ACC to generate ERN

Holroyd & Coles (2002)

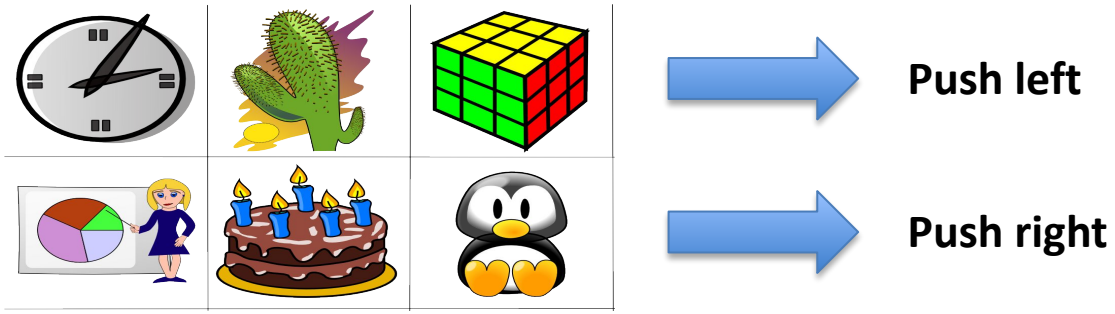
Experiment 1: ERN as a probabilistic learning task

ERN elicited by:

- negative feedback
- error commission

No prior effect had studied both sources of the ERN signal in the same task.

- 6 stimuli; 10 blocks, each block with 300 trials
- In each block, 3 of the 6 stimuli
- must either push left or right button
- feedback given to participant (reward/punishment) [$$]
- correct mapping not given, learned by trial & error
- Various stimuli have various mappings:
-
    - stable [reward given if correct response given]
    - random [50% of trials rewarded, 50% punished]
    - fixed [consistent reward/punishment] n.m.w. pressed



Trial structure in each block:

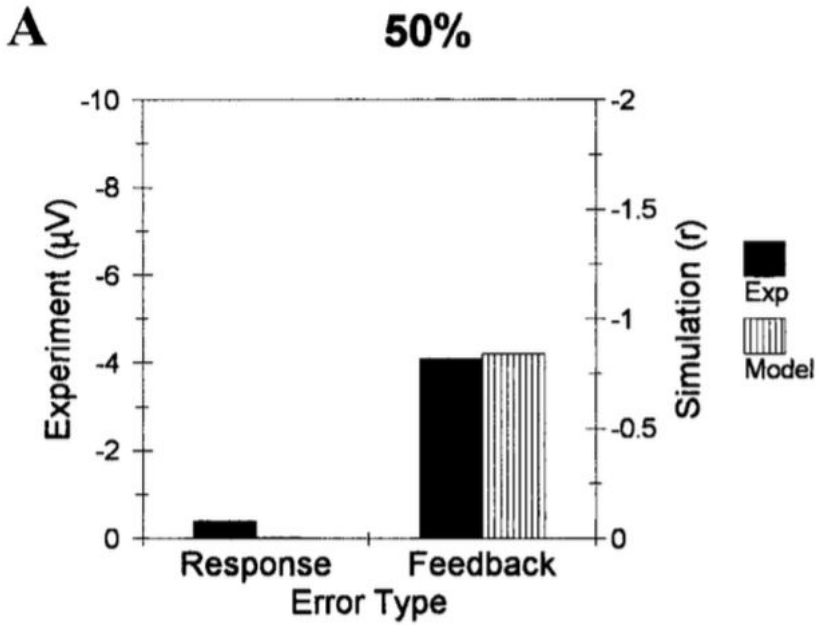
- one stimulus mapped to “left”
- one stimulus mapped to “right”
- one stimulus will give 50% reward, 50% punishment
- one stimulus will give 50% reward, 50% punishment
- one stimulus will always give reward
- one stimulus will always give punishment

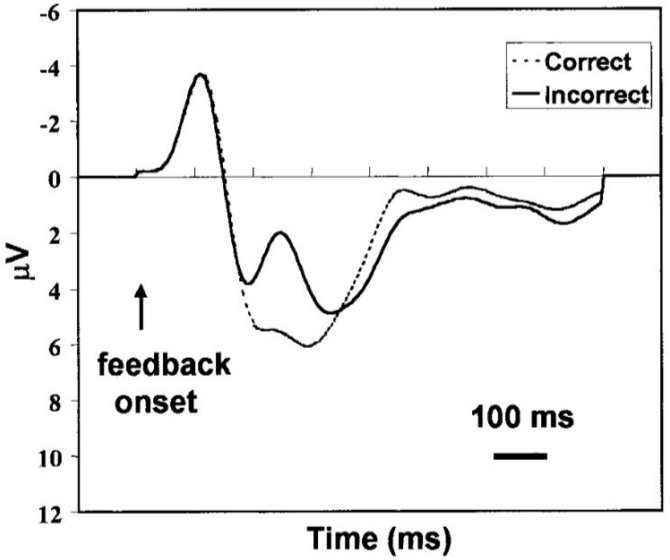
100% mapping (first two)
50% mapping (second two)
always correct/incorrect (last two)

Holroyd & Coles (2002)

*"In the 50% mapping condition, **the system must wait for the feedback** to determine the outcome of the trial. Therefore, we predicted that **negative feedback** stimuli in this condition would continue to elicit the ERN throughout the course of each block."*

(Nothing to be learned, feedback can't be predicted)

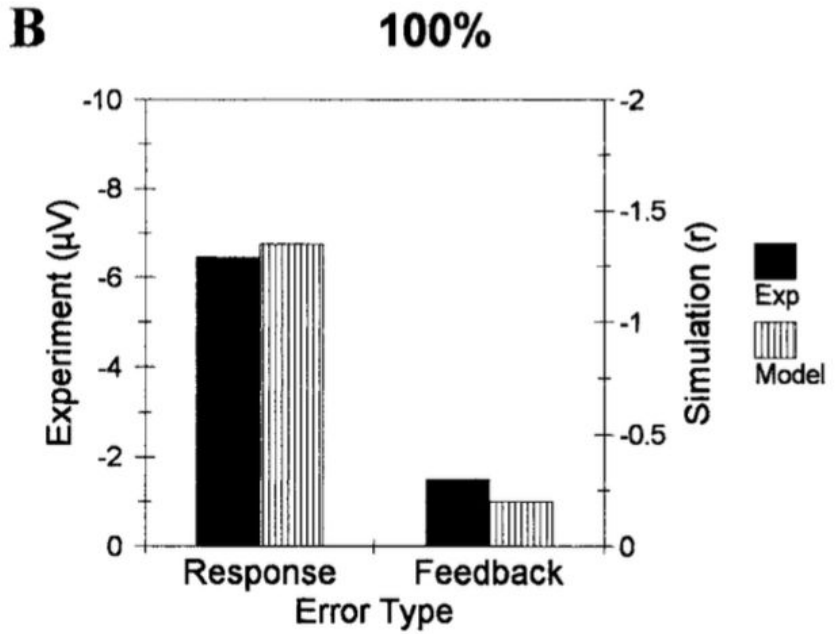


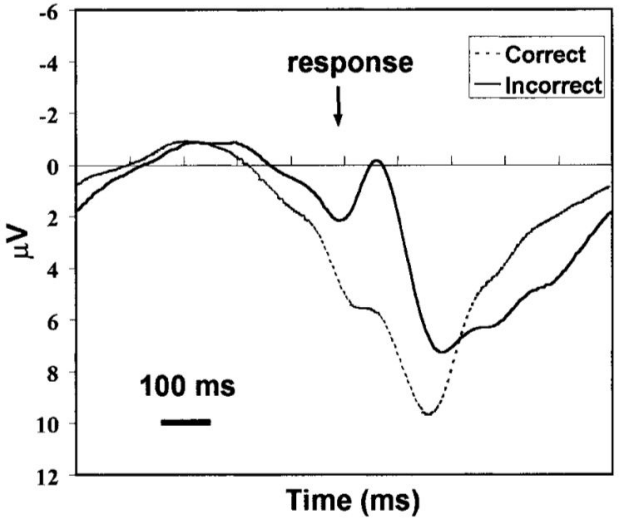


Holroyd & Coles (2002)

"In contrast, in the 100% mapping condition, ***the response itself determines the outcome of the trial****. Therefore, we predicted that as the system learned the associations between response and feedback,* ***the ERN associated with the response would increase, whereas the ERN associated with the feedback would decrease.****"*

(Something to be learned)

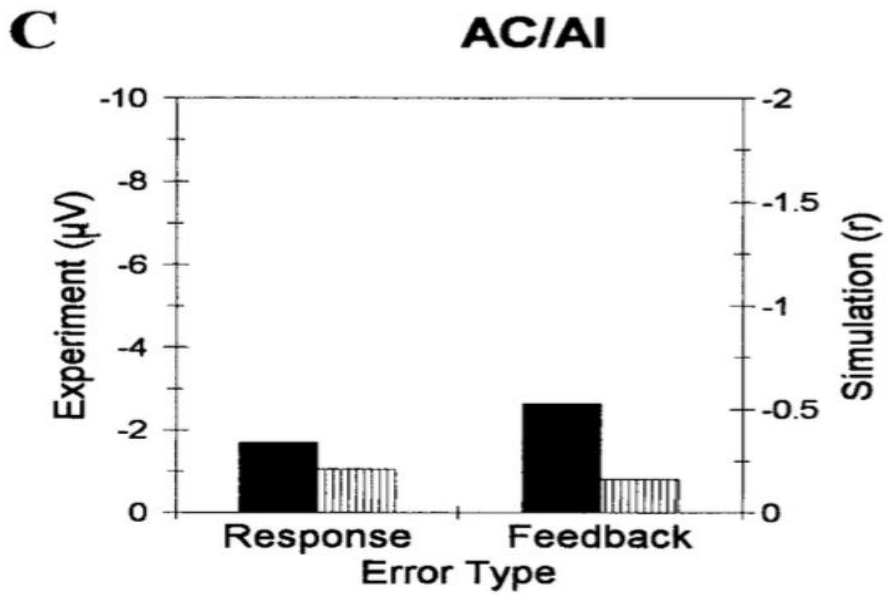




Holroyd & Coles (2002)

"Finally, in the always correct/always incorrect mapping condition, ***the imperative stimulus determines the outcome of the trial****. Therefore, we predicted that as the system learned the associations between the imperative and feedback stimuli,* ***neither the response nor the feedback would elicit the ERN****."*

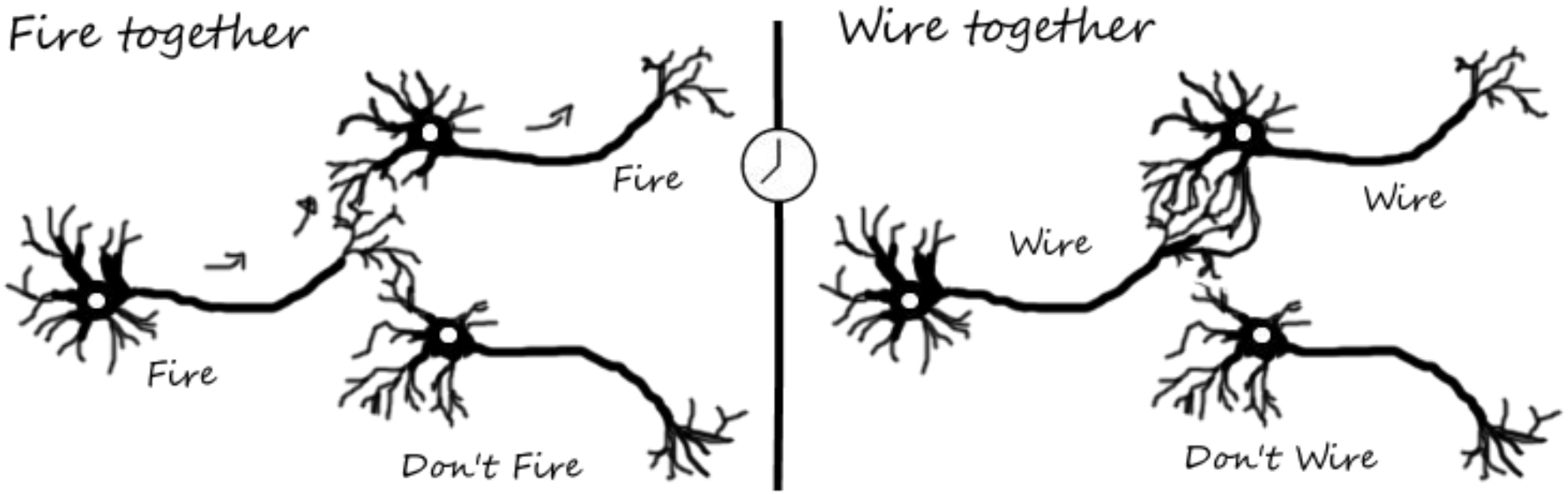
(Nothing to be learned, feedback ***can be*** *predicted)*



Holroyd & Coles (2002)

Our work on the ERN has revealed one such signal in normal human participants, has highlighted the important role it appears to play in executive control, and has suggested a neural area where it might be generated (Coles et al., 1998). This article contributes to that account by grounding the observations in a computational framework that (a) ***formalizes a general set of statements about the error signal into a coherent model with increased explanatory powe****r and (b) makes predictive statements about the error processing system that can be tested in future experiments. Furthermore, the model makes specific claims about the neural systems that implement the computations and thus* ***provides an avenue for testing the integrity of the theory using patient populations****. Although the model requires future elaboration, our initial simulations and experiments have yielded positive results that provide motivation for further research.* ***If valid, this account of the system that gives rise to the ERN may provide insight into how large-scale neural networks for motor control are trained and modulated in biological systems.***

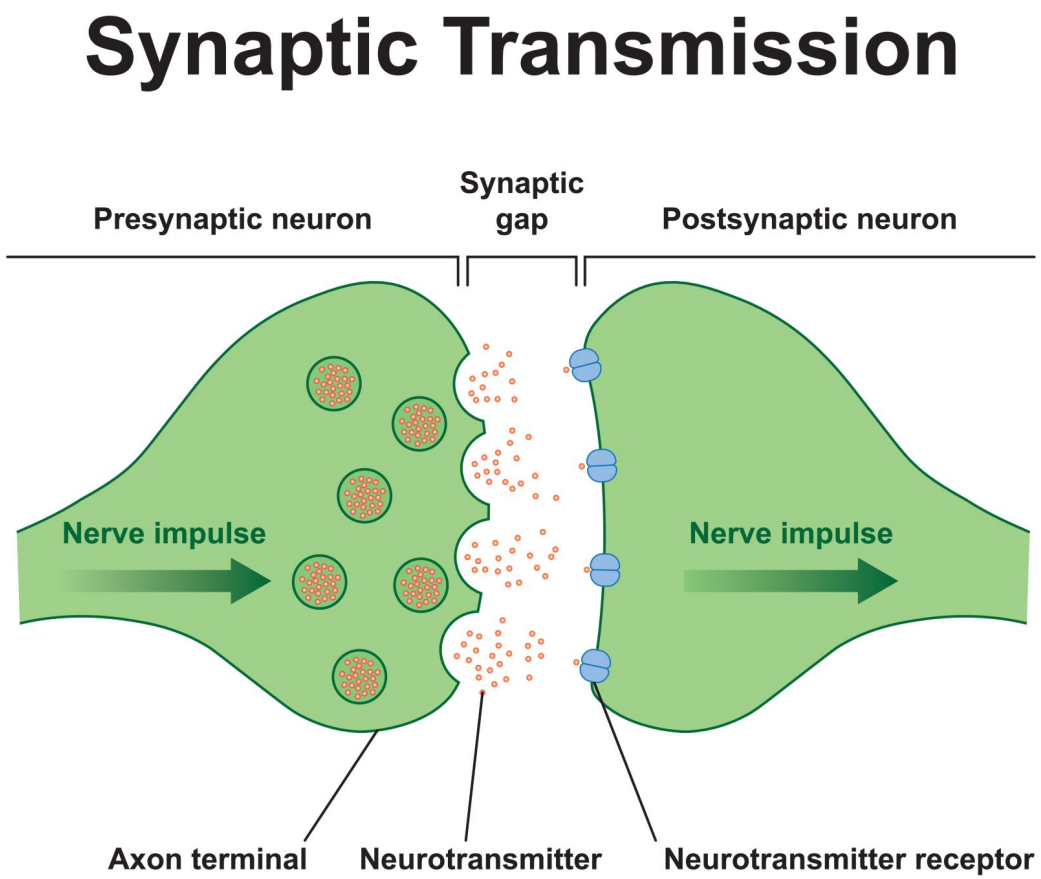
Learning in the Brain



"***Cells that fire together, wire together***" - Donald Hebb

https://togetherwefiretogetherwewire.wordpress.com/2018/06/20/primeiro-post-do-blog/

Learning in the Brain

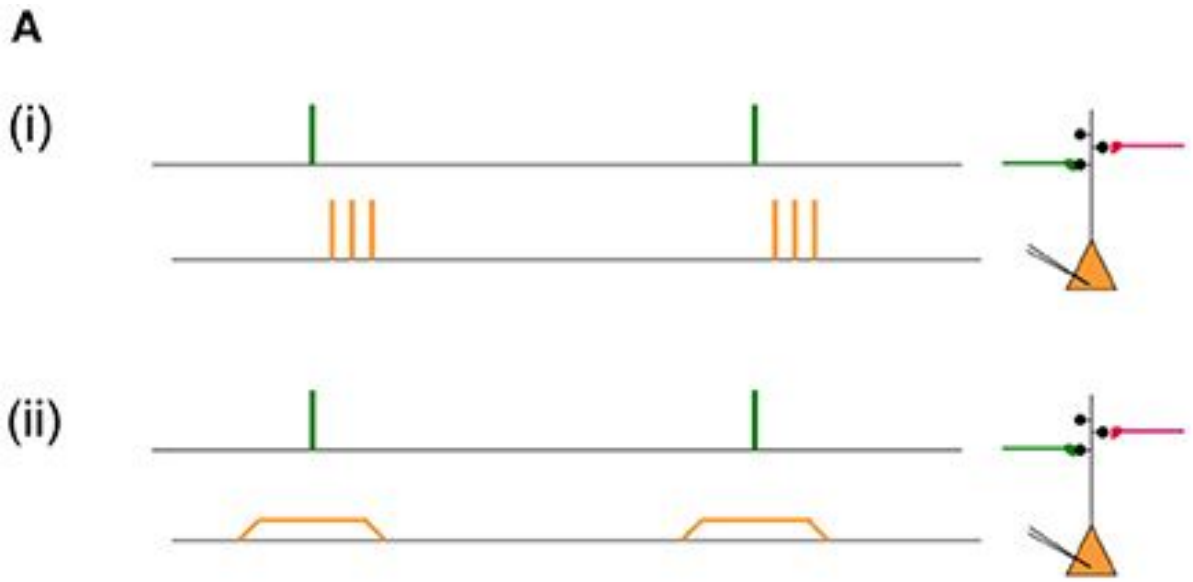


https://www.simplypsychology.org/synapse.html

Two Factor Learning Rules

Presynaptic firing
+
Postsynaptic bursts

Presynaptic firing
+
Postsynaptic sustained activity



$$\frac{dw_{ij}}{dt} = HEBB(w_{ij}; \mathrm{PRE}_j\ , \mathrm{POST}_i\ )$$

Gerstner, W., Lehmann, M., Liakoni, V., Corneil, D., & Brea, J. (2018). Eligibility Traces and Plasticity on Behavioral Time Scales: Experimental Support of NeoHebbian Three-Factor Learning Rules. Frontiers in Neural Circuits, 12, 53. https://doi.org/10.3389/fncir.2018.00053

Two Factor Learning Rules

Hebb's Rule is Incomplete

"A wealth of evidence supports Hebb's rule, but researchers realize that the ***rule is incomplete if the aim is to select appropriate actions****, because the rule is* ***ignorant about the usefulness of the network's output.*** *In animals, rewards and punishments influence learning such that behaviours that lead to reward are reinforced and behaviours that result in aversive outcomes are inhibited."*

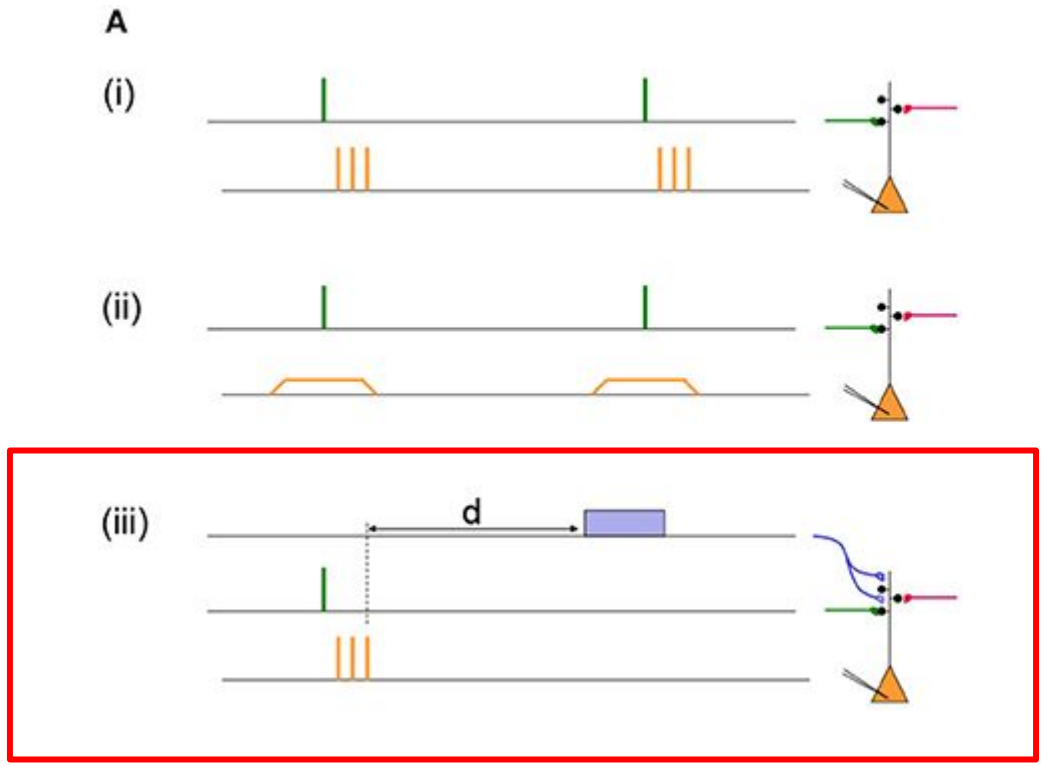
"*The influence of theories of reinforcement learning* ***increased tremendously when it became clear that neuromodulatory systems, such as the dopaminergic system, code for unexpected reward****. In reinforcement learning theory, unexpected rewards and punishments give rise to RPEs."*

Roelfsema, P. R., & Holtmaat, A. (2018). Control of synaptic plasticity in deep cortical networks. Nature Reviews Neuroscience, 19(3), 166–180.

Three Factor Learning Rules

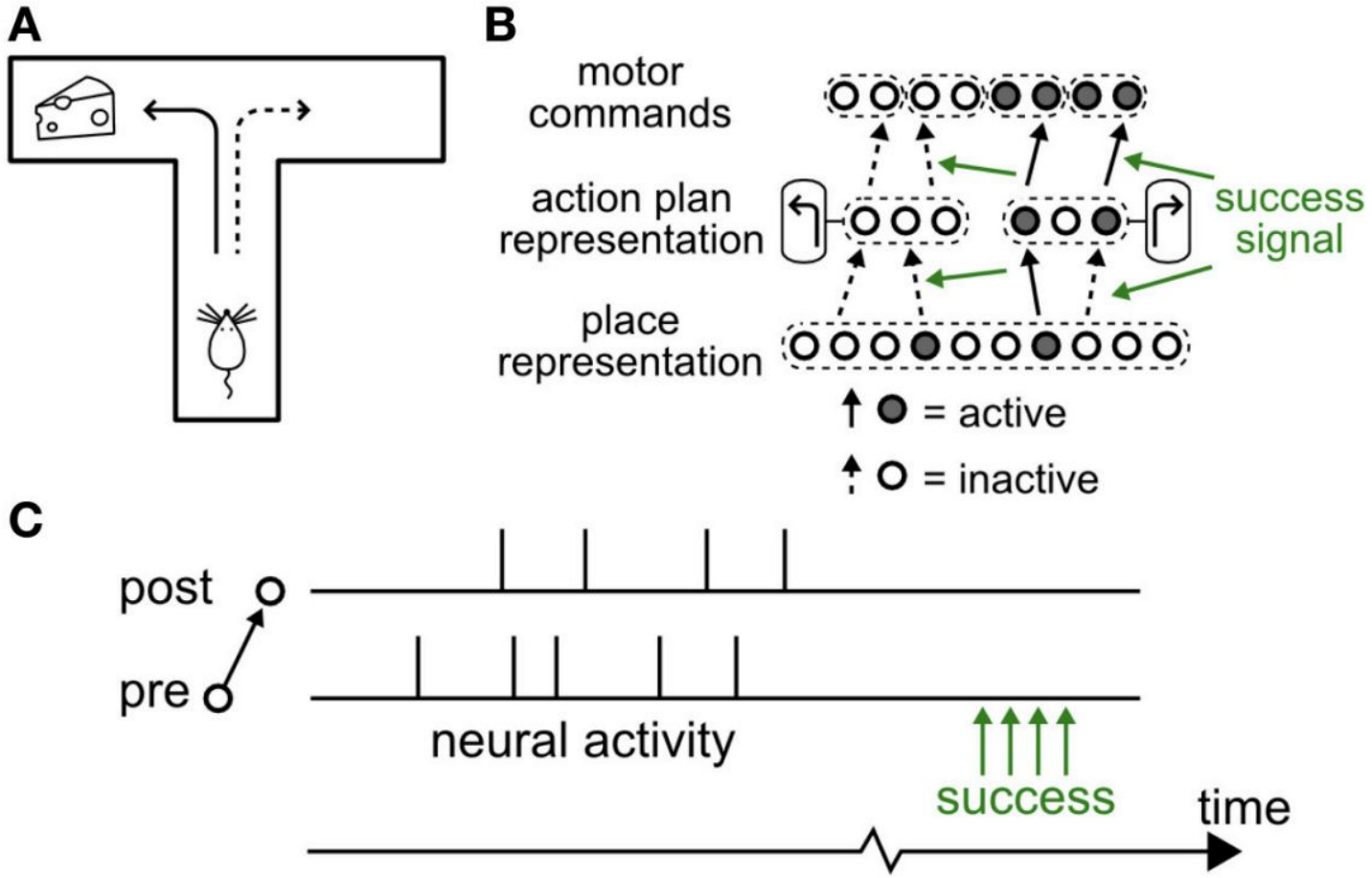
Three factors:

- presynaptic firing
- postsynaptic firing
- error signal



Gerstner, W., Lehmann, M., Liakoni, V., Corneil, D., & Brea, J. (2018). Eligibility Traces and Plasticity on Behavioral Time Scales: Experimental Support of NeoHebbian Three-Factor Learning Rules. Frontiers in Neural Circuits, 12, 53. https://doi.org/10.3389/fncir.2018.00053

Eligibility Traces in the Brain



$$e_t(s) = \begin{cases} \gamma\lambda e_{t-1}(s) & s_t \neq s \\ \gamma\lambda e_{t-1}(s) + 1 & s_t = s \end{cases}$$

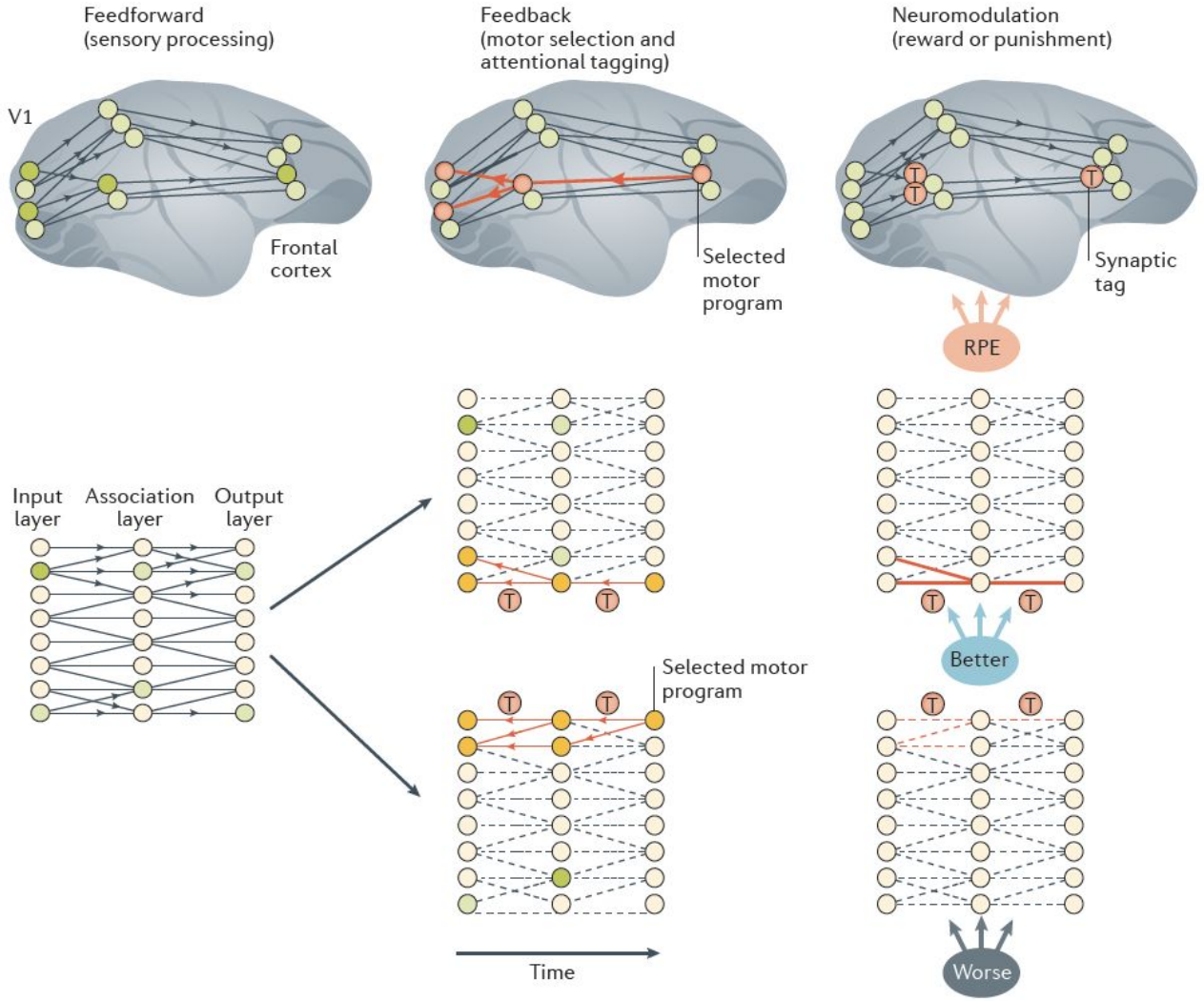
A mouse is in a maze, with hippocampus place cells signalling where it believes it is.

Two potential movement action choices at next layer of the network (go left / go right).

Each action selection links to motor cortex to carry out the action.

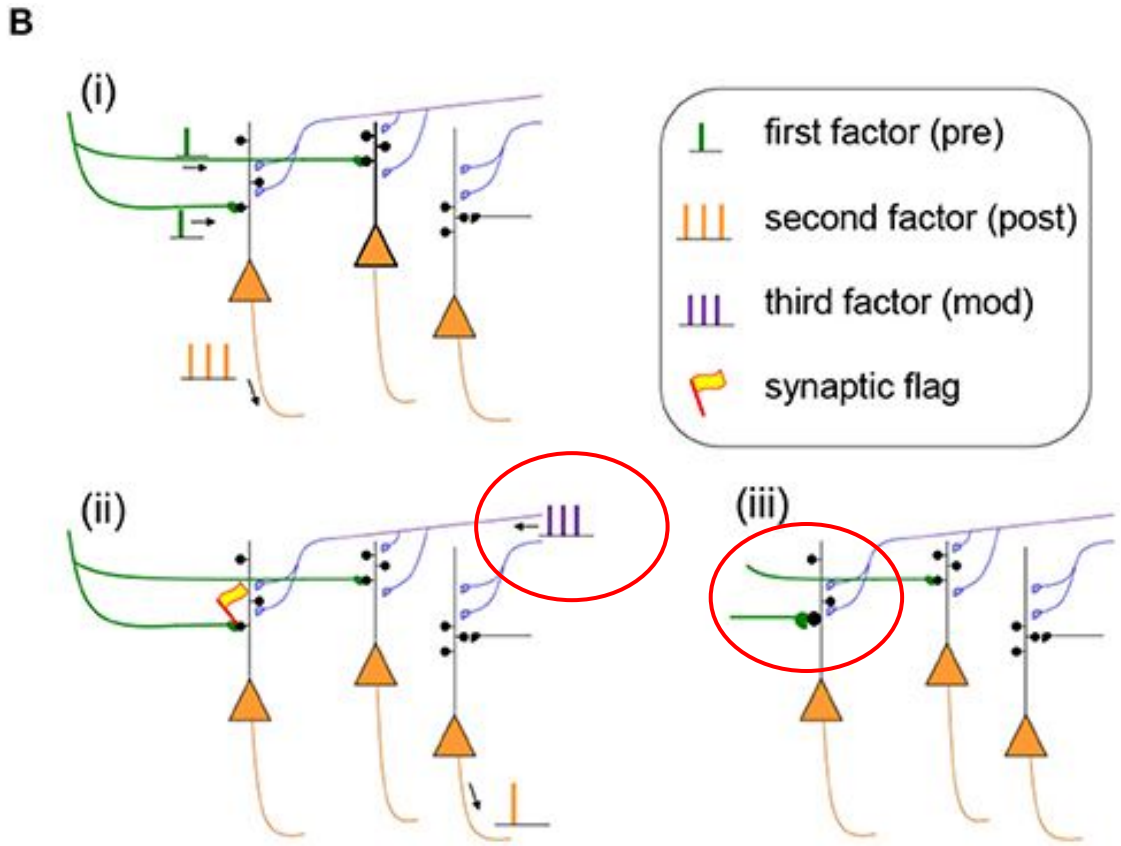
Frémaux, N., & Gerstner, W. (2016). Neuromodulated Spike-Timing-Dependent Plasticity, and Theory of Three-Factor Learning Rules. Frontiers in Neural Circuits, 9.

Eligibility Traces in the Brain



Eligibility Traces in the Brain

postsynaptic neuron not active in 2nd stage in (i) so no updates here, but is active in first



coincident firing sets synaptic flag (eligibility trace) at this synapse

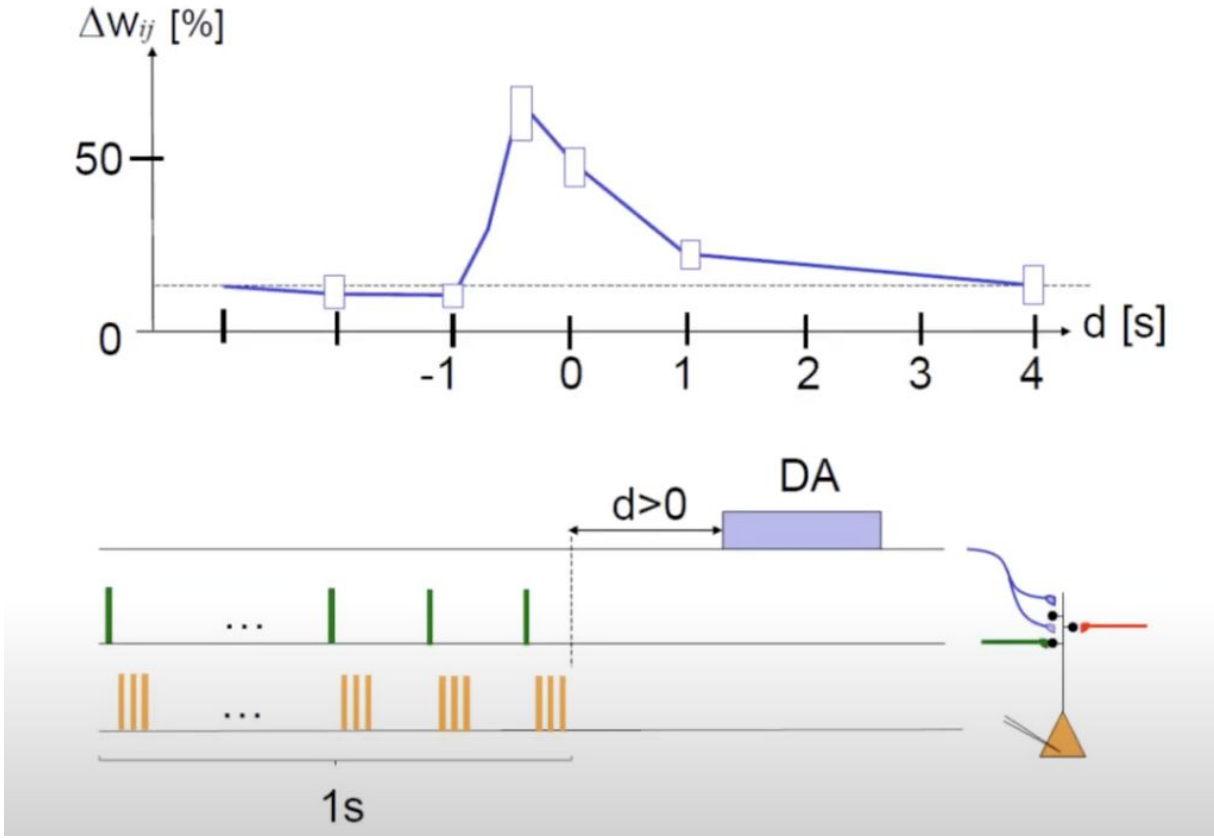
Because reward signal was given, only the first synapse gets updated

Gerstner, W., Lehmann, M., Liakoni, V., Corneil, D., & Brea, J. (2018). Eligibility Traces and Plasticity on Behavioral Time Scales: Experimental Support of NeoHebbian Three-Factor Learning Rules. Frontiers in Neural Circuits, 12, 53. https://doi.org/10.3389/fncir.2018.00053

Reward Signal Timing

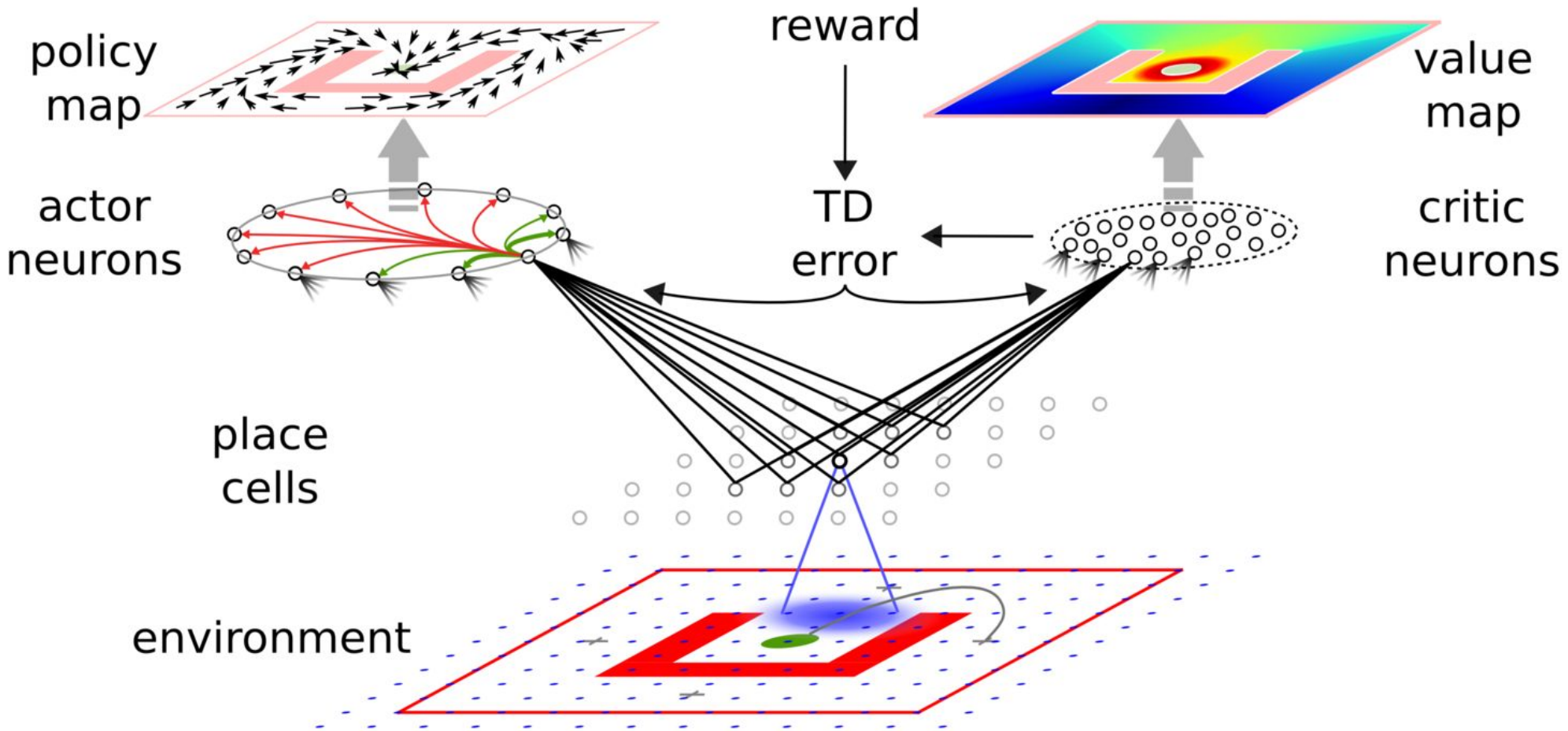
- Reward signal is third of the 3-factor learning rules
- Reward arrives after coincident activation of pre- and postsynaptic neurons
- In the brain, how long can the eligibility traces stay active?
- Also phrased: to get a a synaptic weight change, what happens as you vary the arrival of the reward signal?

- **But what about presenting the reward before the cue?**
  - Big difference between TD error behaviour
  - See **15.6** in Sutton & Barto for extra info



Yagishita, S., Hayashi-Takagi, A., Ellis-Davies, G. C. R., Urakubo, H., Ishii, S., & Kasai, H. (2014). A critical time window for dopamine actions on the structural plasticity of dendritic spines. Science, 345(6204), 1616–1620

Fremaux et al. 2013



Frémaux, N., Sprekeler, H., & Gerstner, W. (2013). Reinforcement Learning Using a Continuous Time Actor-Critic Framework with Spiking Neurons. PLoS Computational Biology, 9(4), e1003024. https://doi.org/10.1371/journal.pcbi.1003024

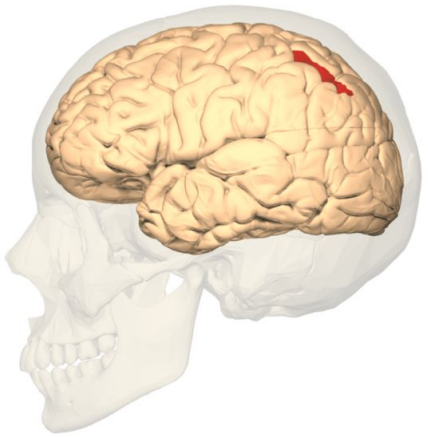
Two More Brain Areas

- **IPS** important for
  - motor planning
  - motor control
- **LPFC** important for:
  - decisions
  - planning
  - executive function

(These brain areas will be heavily referenced in Thursday's paper reading & presentation.)

IPS + PFC = **task positive network** (activation for demanding tasks)

Intraparietal Sulcus (IPS)



Lateral Prefrontal Cortex (LPFC)

Explore vs Exploit (fMRI)

"Understanding the exploration–exploitation dilemma: An fMRI study of attention control and decision-making performance" *(Laureiro-Martinez et al., 2014)*

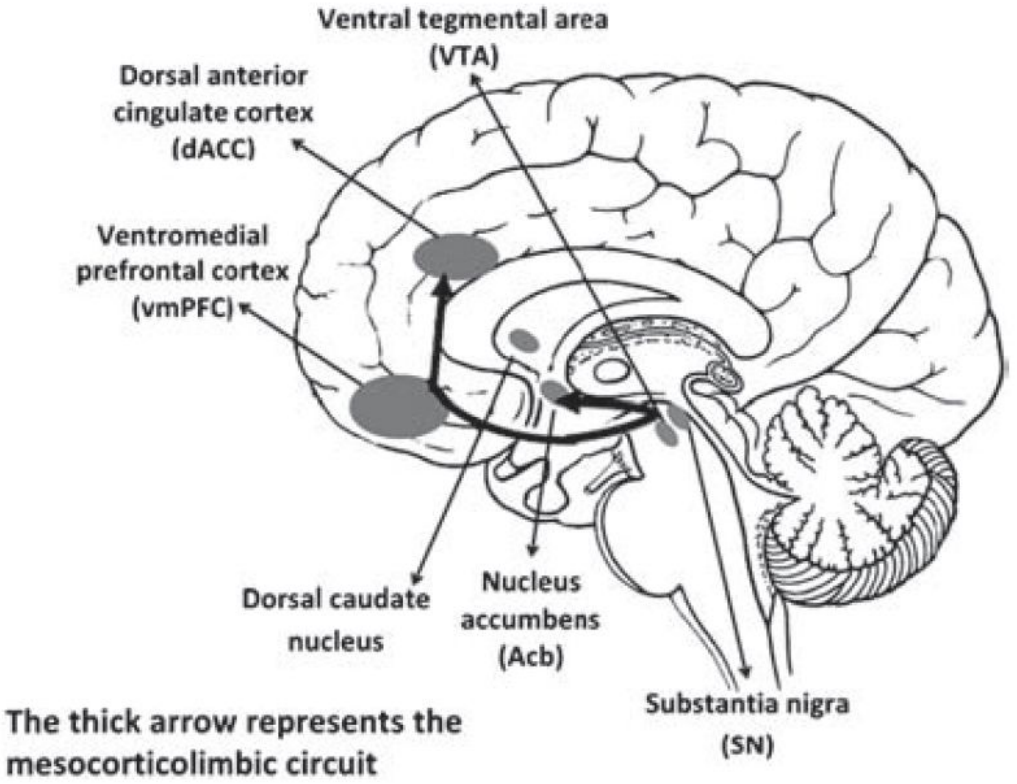
- Sample of expert decision makers (n=63; 11 females; mean age=33)
- Exploitation associated with reward-seeking
- Exploration associated with attentional control and tracking of alternative choices
- Assess intentional decision-making biases
- Exploration entails disengaging from current task to enable flexibility, while exploitation optimises task performance according to known correct assumptions at current moment in time
- What happens as task switch happens?

Explore vs Exploit (fMRI)

Hypothesis 1: Compared to exploration, exploitation will involve stronger activation of reward-related brain areas.

Hypothesis 2: Compared to exploitation, exploration will involve stronger activation of the regions involved with the assessment of reward-related uncertainty as well as cognitive and attentional control

Hypothesis 3: The stronger the activation of the attentional-control regions, the better the decision-making performance

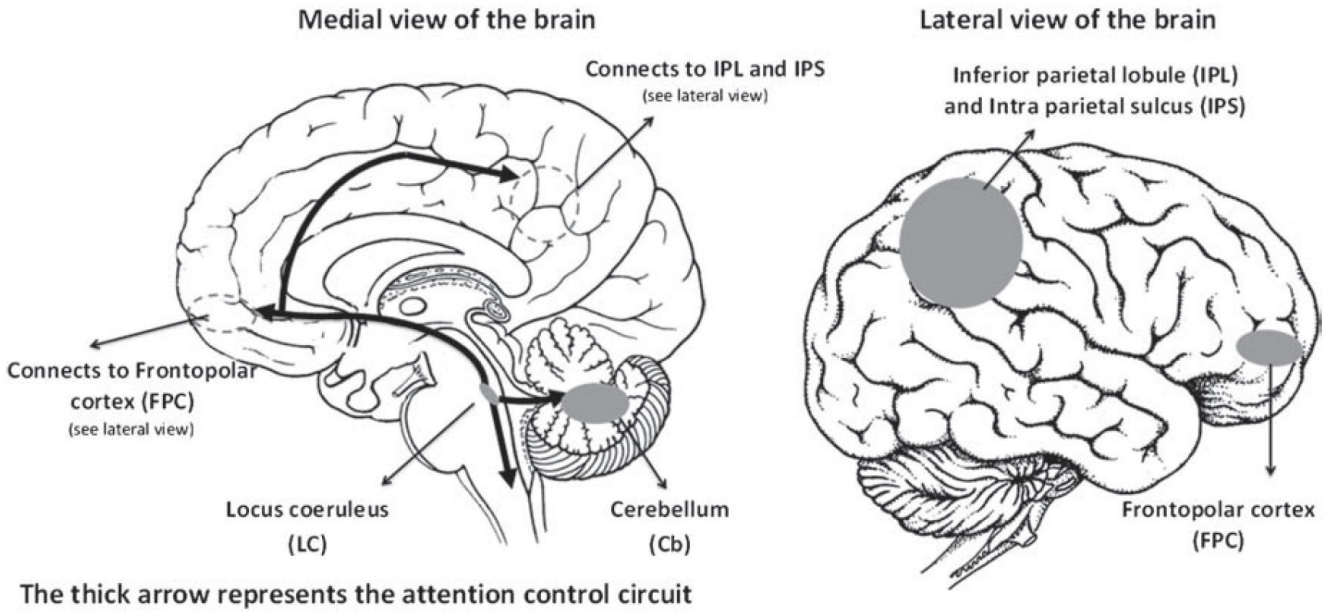


Explore vs Exploit (fMRI)

Hypothesis 1: Compared to exploration, exploitation will involve stronger activation of reward-related brain areas.

Hypothesis 2: Compared to exploitation, exploration will involve stronger activation of the regions involved with the assessment of reward-related uncertainty as well as cognitive and attentional control

Hypothesis 3: The stronger the activation of the attentional-control regions, the better the decision-making performance



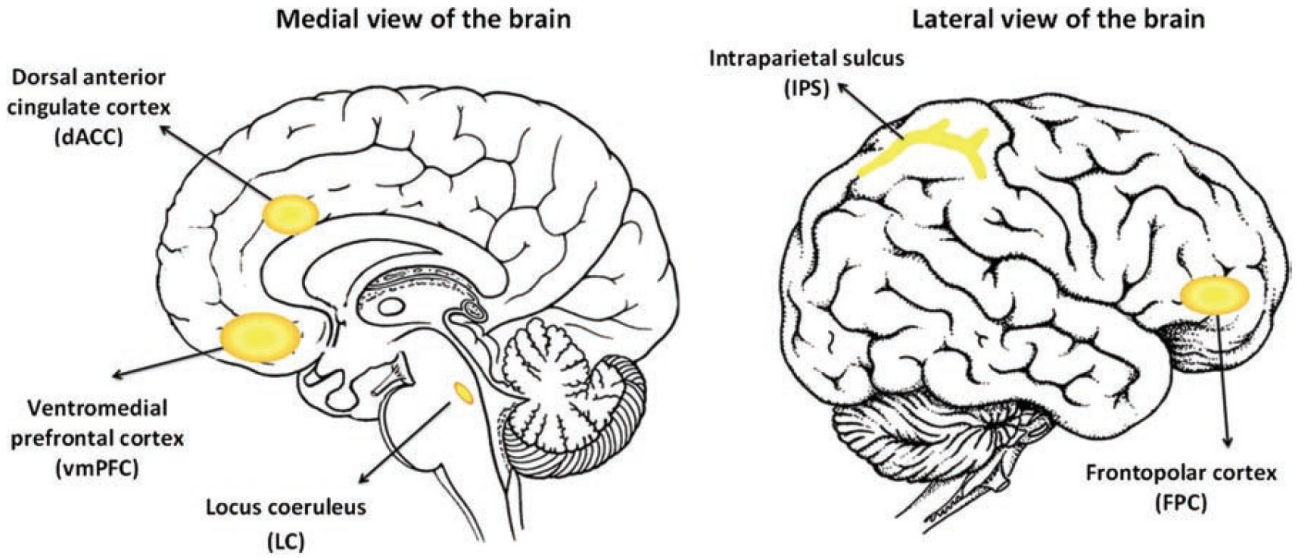
The thick arrow represents the attention control circuit

Explore vs Exploit (fMRI)

Hypothesis 1: Compared to exploration, exploitation will involve stronger activation of reward-related brain areas.

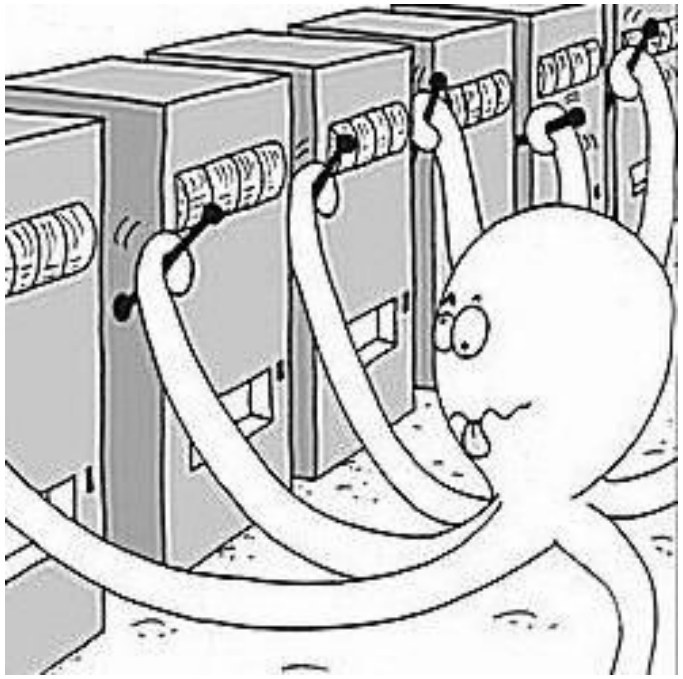
Hypothesis 2: Compared to exploitation, exploration will involve stronger activation of the regions involved with the assessment of reward-related uncertainty as well as cognitive and attentional control

Hypothesis 3: The stronger the activation of the attentional-control regions, the better the decision-making performance

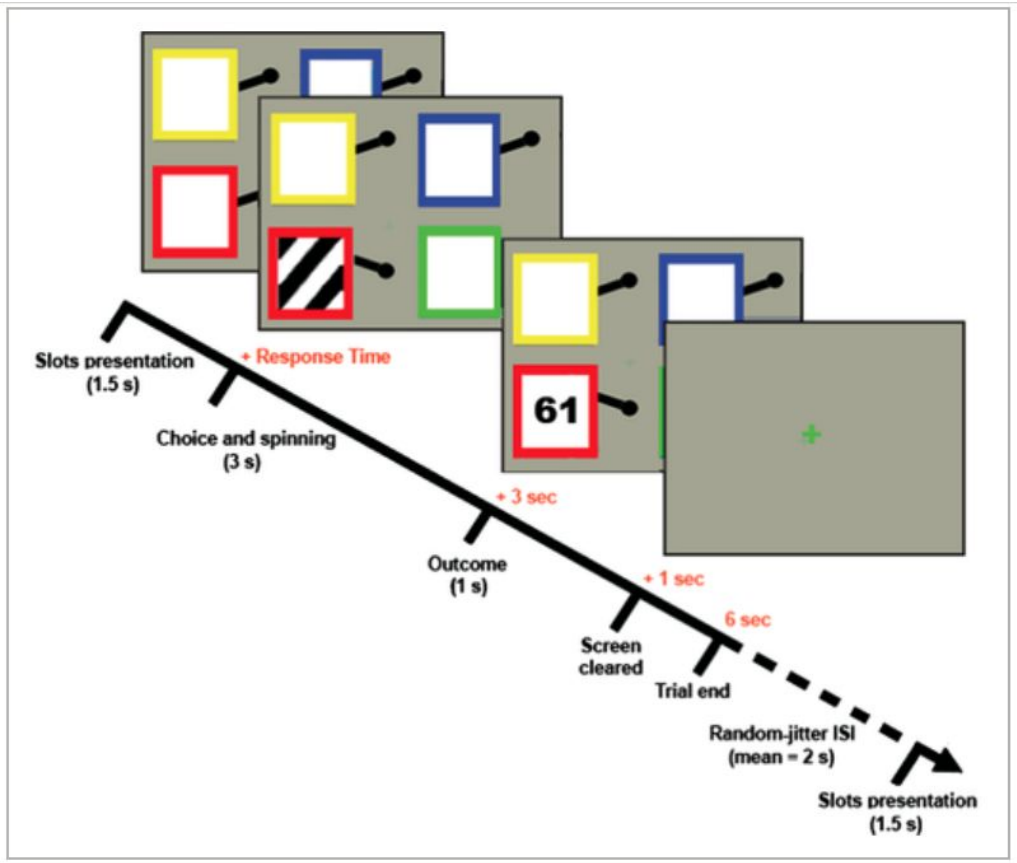


Explore vs Exploit (fMRI)

- 4-arm bandit task
- Simplified model of wider cognitive effect of interest, but which captures key ingredients of dealing with uncertainty and reward maximisation strategies
- Each bandit has a different mean reward that changes on each trial
- Participants not aware of the change
- Learn via active sampling
- Each trial classified as exploration or exploration (shift vs stay)
- Select bandit by using button keypad
- 300 trials divided into 4 sessions (1 session = 75 trials)
- Performance estimated by total points after all trials (n=300)



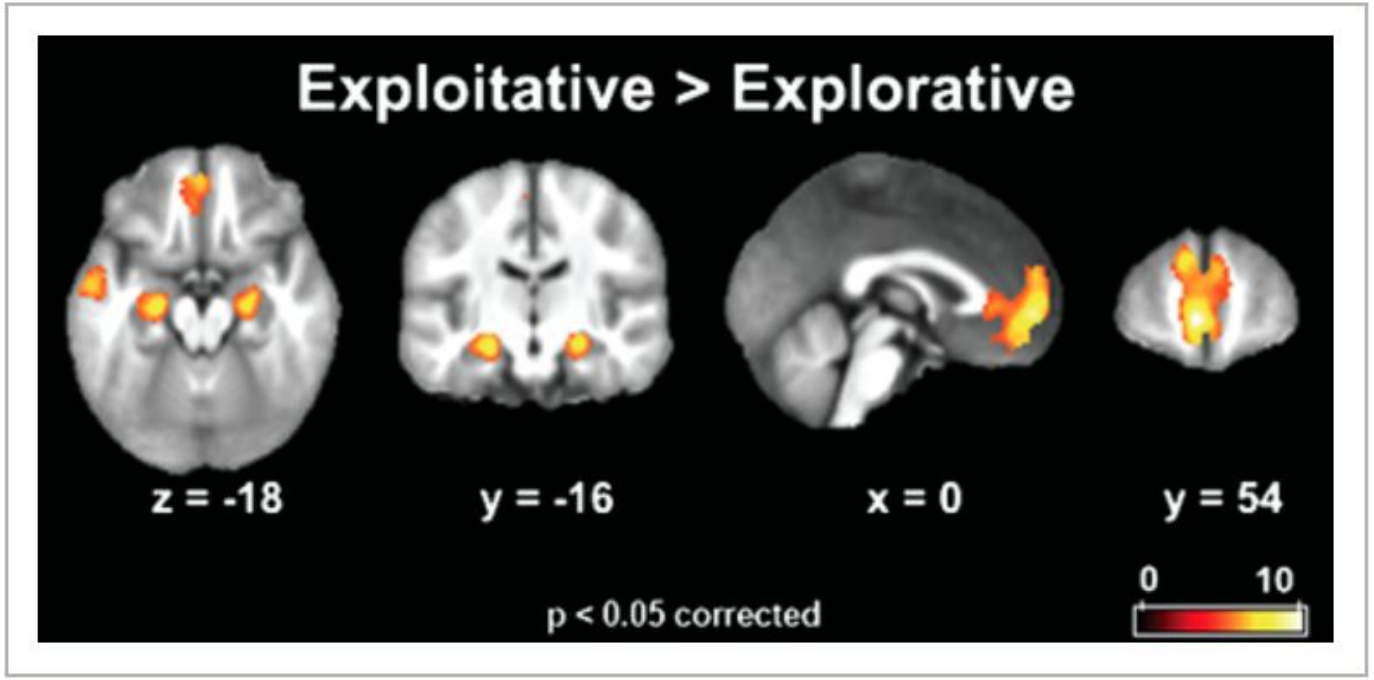
Explore vs Exploit (fMRI)



Laureiro-Martínez, D. , Brusoni, S. , Canessa, N. and Zollo, M. (2015), Understanding the exploration–exploitation dilemma: An fMRI study of attention control and decision-making performance. Strat. Mgmt. J., 36: 319-338.

Explore vs Exploit (fMRI)

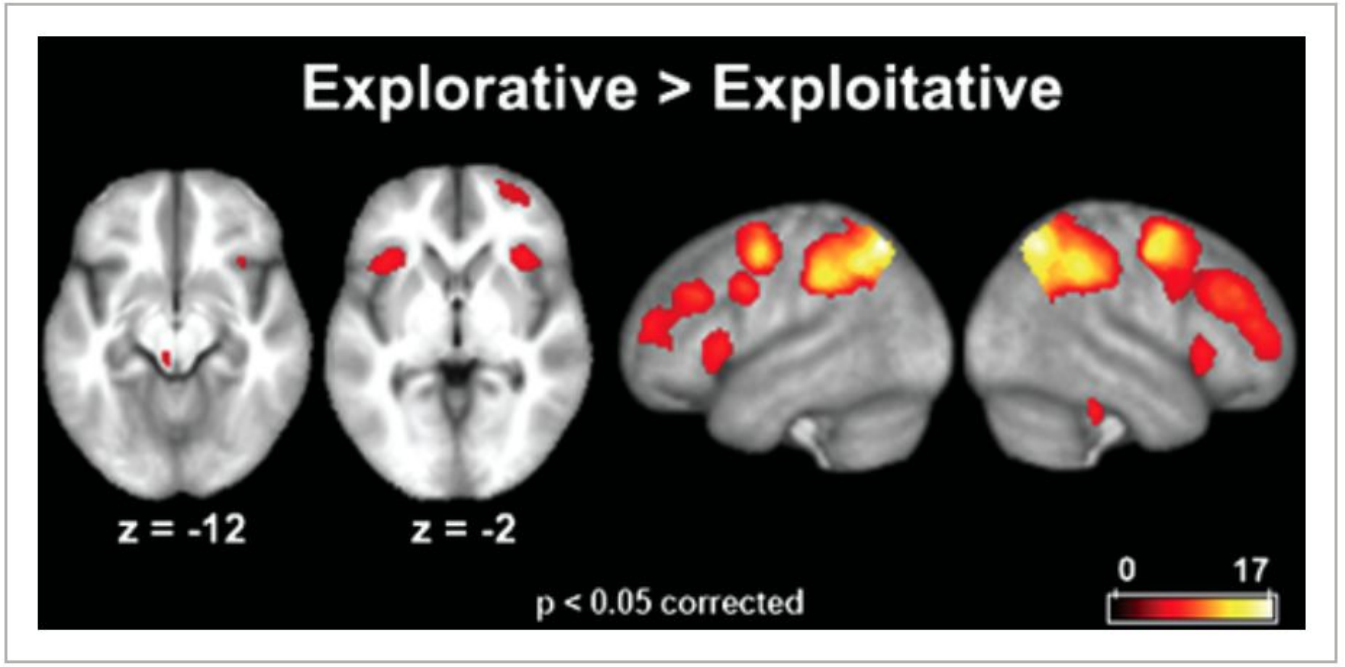
- Medial prefrontal cortex and hippocampus active
- Claimed as evidence supporting monitoring of reward exploitation
- vmPFC is target of dopamine reward system:
  - dopamine (VTA & SN)
  - -> vmPFC
- Choice to maximise current best possibility seen as a link highlighting (roughly) start and end points of the dopamine reward system



Laureiro-Martínez, D. , Brusoni, S. , Canessa, N. and Zollo, M. (2015), Understanding the exploration–exploitation dilemma: An fMRI study of attention control and decision-making performance. Strat. Mgmt. J., 36: 319-338.

Explore vs Exploit (fMRI)

- Tasks very similar in structure in terms of key presses and response selection, but resulting brain activation subtractions quite different
- A decision to “explore” driven by higher activation of attentional circuits (bilateral parietal and frontal regions)
- Insula activation (fear / anxiety)
- Authors conclude task-based anxiety / fear of moving to uncertainty at odds with task goal is responsible here, overridden by higher level understanding of the potential payoffs

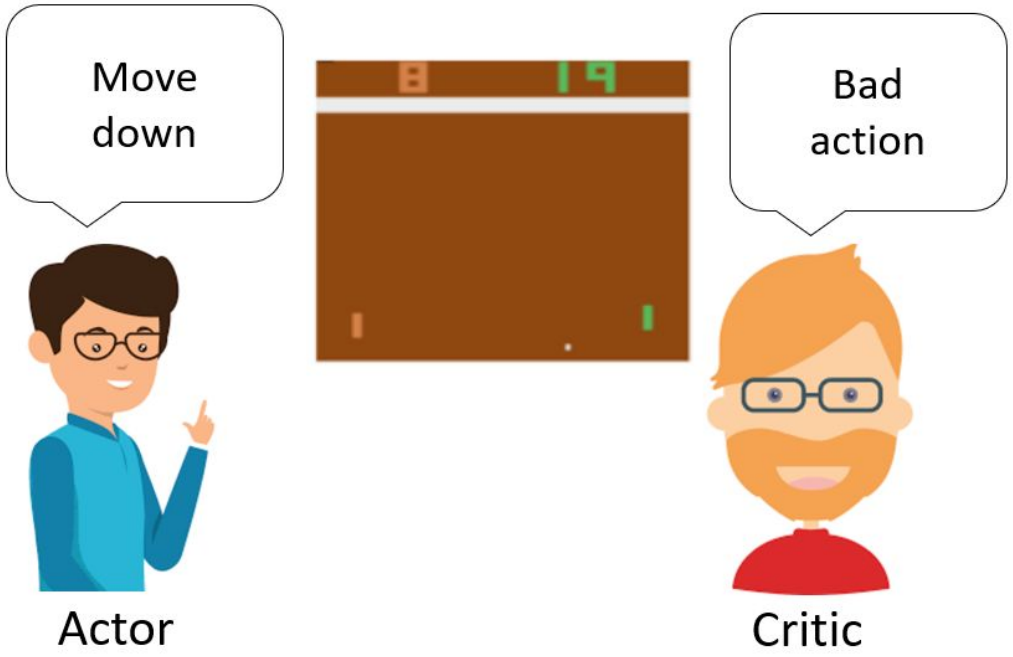


Laureiro-Martínez, D. , Brusoni, S. , Canessa, N. and Zollo, M. (2015), Understanding the exploration–exploitation dilemma: An fMRI study of attention control and decision-making performance. Strat. Mgmt. J., 36: 319-338.

Brain as Actor-Critic

"Neuroscience is still far from reaching complete understanding of the circuits, molecular mechanisms and functions of the phasic activity of dopamine neurons, but evidence supporting the reward prediction error hypothesis, along with evidence that phasic dopamine responses are reinforcement signals for learning, suggest that the brain might implement something like an actor-critic algorithm in which TD errors play critical roles."

(Sutton & Barto, 15.6)

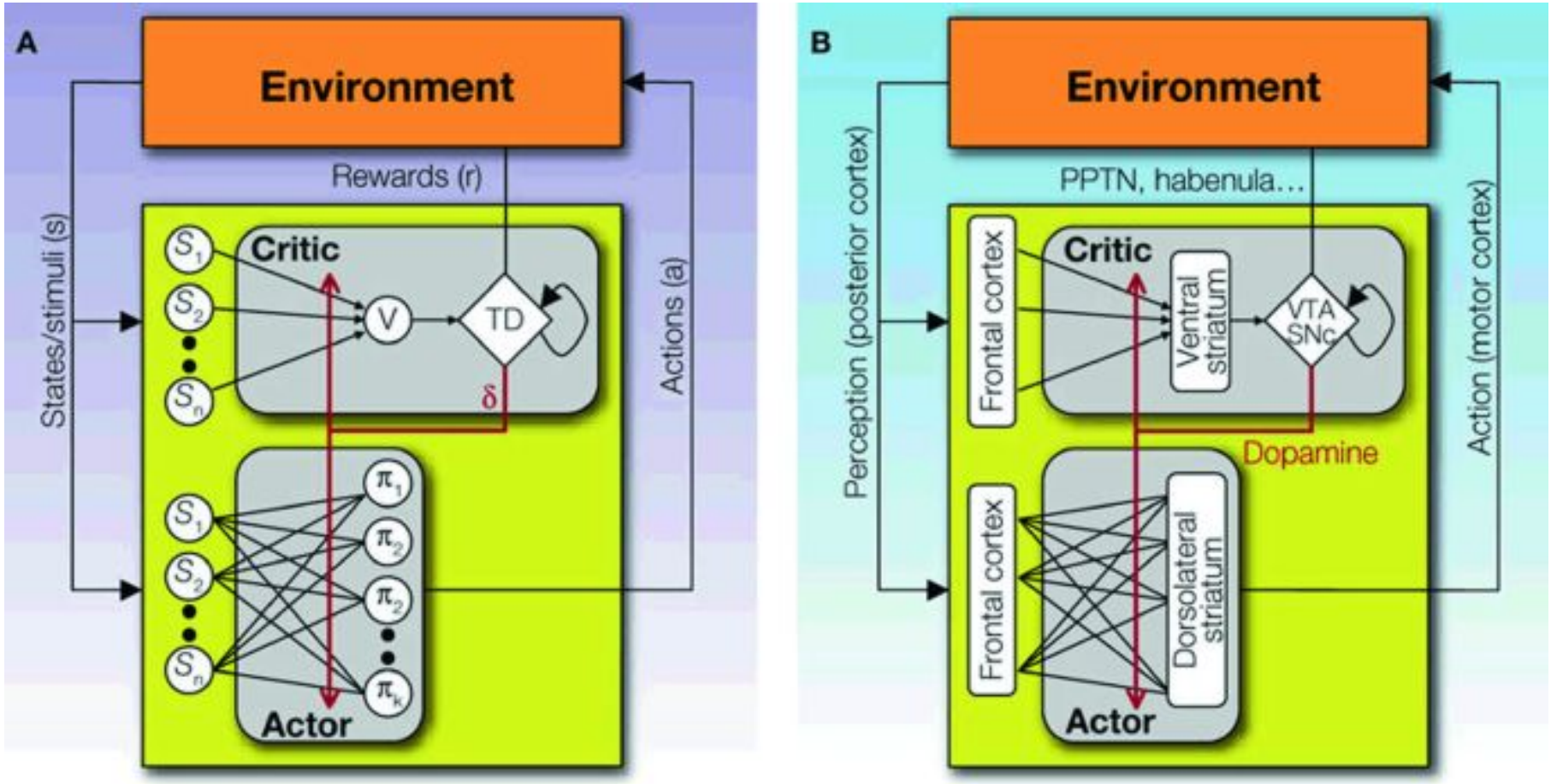


Actor = dorsal striatum

Critic = ventral striatum

(nucleus accumbens is a part of V.S.)

Brain as Actor-Critic



Enan, M. (2021). Prediction Error and Actor-Critic Hypotheses in the Brain. In: Belousov, B., Abdulsamad, H., Klink, P., Parisi, S., Peters, J. (eds) Reinforcement Learning Algorithms: Analysis and Applications. Studies in Computational Intelligence, vol 883. Springer, Cham.

Brain as Actor-Critic

- Actor-Critic algorithm is **model-free**
- Remember “actor” supposed to be in dorsal striatum?
- Experiments in rats show that:
  - dorsolateral striatum (DLS) is more like model-free RL
  - dorsomedial striatum (DMS) is more like model-based RL

Quite a bit more to the story.
Let’s wait until Thursday

Stay Tuned …

Thursday's Paper Presentation:

"States versus Rewards: Dissociable neural prediction error signals underlying model-based and model-free reinforcement learning"

Lots more about IPS, LPFC, ventral striatum, RPEs and the difference between brain's (potential) implementation of model-free vs model-based RL.

Outstanding Questions

What happens on a larger scale?

What if feedback on an action given after a week?

How readily to animal models map over to humans?

Do noninvasive (human) methods give the right level of granularity to study RL in the brain?

Resources

- 2 Minute Neuroscience (YouTube) [Videos 14-22]
- Reinforcement Learning & The Brain (YouTube)
- Reinforcement Learning (Github; resources)
- Chapter 15 of Sutton & Barto
- A lot of stuff by Matthew Botvinick! (DeepMind UK)

Bonus: Earlier Work in NHPs

- Let's expand on what was known at the time of Holroyd & Coles (2002) from work in non-human primates (NHPs)

Page 682 of paper, left column, about going back in time, conditioned stimulus, temporal difference learning signal