Neuroscience

Friday, June 26, 2020 9:55 PM

Dopamine appears to convey TD errors to brain structures where learning and decision making take place

Neuron:
- A cell specialized for processing and transmitting information using electrical and chemical signals

Dendrites:
- Structures that branch from a cell body to receive input from other neurons (or external signals in the case of sensory neurons)

Axon:
- A fiber that carries the neuron's output to other neurons (or muscles or glands)

Action Potential:
- An electrical pulse that travels along the axon
  - These are also called spikes

Axonal Arbor:
- The branching structure of a neuron's axon that connects the axon to many other neurons

Synapse:
- A structure generally at the termination of an axon branch that mediates the communication of one neuron to another

Presynaptic Neuron:
- The neuron sending a message to another neuron through a synapse

Postsynaptic Neuron:
- The neuron receiving a message from another neuron through a synapse

Synaptic Cleft:
- The very small space between the presynaptic ending and the postsynaptic neuron

Neurotransmitter:
- A chemical that transmits a signal across the synaptic cleft to excite or inhibit its spike-generating activity, or modulate its behavior in other ways
  - Dopamine is a neurotransmitter
    - There are at least five different receptor types by which dopamine can affect a postsynaptic neuron

Background Activity:
- A neuron's level of activity (usually firing rate) when the neuron doesn't appear to be driven by synaptic input related to the task of interest to the experimenter
  - This is generally due to input from the wider network or noise within the neuron or its synapses but can also be due to dynamic processes intrinsic to the neuron

Phasic Activity:
- Bursts of spiking activity usually caused by synaptic input

Tonic Activity:
- Activity that varies slowly and often in a graded manner

Synapse Efficacy:
- The strength or effectiveness by which a neurotransmitter released at a synapse influences the postsynaptic neuron
  - One way a nervous system can change through experience is to change synaptic efficacies

Neuromodulator:
- A neurotransmitter having effects other than or in addition to direct fast excitation or inhibition
  - Brains contain several different neuromodulation systems, with each system using a different neurotransmitter
  - A neuromodulatory system can distribute something akin to a scalar signal (such as a reinforcement signal) to alter the operation of synapses in widely distributed sites critical for

learning

Synaptic Plasticity:

The ability of synaptic efficacies to change

This is one of the primary mechanisms responsible for learning

Modulation of synaptic plasticity via dopamine is a plausible mechanism for how the brain might implement learning algorithms

The reward signal used in reinforcement learning R_t is like a reward signal in an animal's brain, not an object or event in the animal's environment

However, it is unlikely a single signal exists in an animal's brain, R_t is best thought of as an abstraction summarizing the effect of a multitude of neural signals generated by many systems in the brain

Reinforcement Signal:

A signal directing the changes made by a learning algorithm (e.g. in a TD method the reinforcement signal is $\delta_t = R_t + \gamma V(S_t) - V(S_{t-1})$, not just R_t)

Reward Prediction Error (RPEs):

Measures of discrepancy between the expected and received reward signal

TD errors are a special kind of RPEs

When neuroscientists refer to RPEs they usually mean TD RPEs

TD errors here usually mean one that doesn't depend on actions

Reward Prediction Error Hypothesis of Dopamine Neuron Activity:

The hypothesis that one of the functions of the phasic activity of dopamine-producing neurons in mammals is to deliver an error between an old and new estimate of expected future reward to target areas in the brain

Comparison of TD Model of Classical Conditioning with Phasic Activity of Dopamine-Producing Neurons (Montague et al 1996):

To account for TD error possibly being negative, but a negative firing rate of a neuron being impossible, dopamine-producing neuron activity is modelled as $\delta_{t-1} + b_t$ where b_t is the background activity of the neuron

The complete serial compound representation of states was chosen (see previous chapter for details)

This representation allows the TD error to mimic the fact that dopamine neuron activity not only predicts a future reward, but is also sensitive to when after a predictive cue that reward is expected to arrive

This is simply one way to get at the fact that there needs to be some way to represent how much time has passed between sensory cues and the arrival of reward

Similarities:

The phasic response of a dopamine neuron only occurs when a rewarding event is unexpected

Early in learning, neutral cues that precede a reward do not cause a phasic response but with continued learning these cues gain predictive value and elicit phasic responses

If an even earlier cue reliably precedes a cue that has already acquired predictive value, the phasic dopamine response shifts to the earlier cue; ceasing for the later cue

If after learning, the predicted rewarding event is omitted, a dopamine neuron's response decreases below its baseline level shortly after the expected time of the rewarding event

However, the choice of input representation is critical to how closely TD errors match some of the details of dopamine neuron activity, particularly the timing

There are situations where the predictions of the RPE hypothesis do not match what is observed in experiments

Dopamine:

Produced mainly by neurons whose cell bodies lie mainly in the substantia nigra pars compacta (SNpc) and ventral tegmental area (VTA)

Plays roles in processes such as motivation, learning, action-selection, most forms of addiction, schizophrenia, and Parkinson's

Neuromodulator
Role in punishment remains controversial
Early Traditional View:
- Dopamine neurons broadcast a reward signal to multiple brain regions implicated in learning and motivation
  - This followed from a paper than electrical stimulation to particular regions of a rat's brain acted as a powerful reward in controlling the rat's behavior and these regions of the brain were dopamine pathways or sites that excited dopamine pathways indirectly

Dopamine neurons have huge axonal arbors, each neuron releases dopamine at 100 to 1000 times more synaptic sites than reached by a typical neuron
- Each axon of a SNpc or VTA dopamine neuron makes roughly 500,000 synaptic contacts with neurons in targeted brain areas
- The axons of most dopamine neurons make synaptic contact with neurons in the frontal cortex and basal ganglia (areas of the brain involved in voluntary movement, decision making, learning, and cognitive functions such as planning)
  - Most ideas relating dopamine to reinforcement learning focus on the basal ganglia and connections from dopamine neurons are particularly dense there

A common belief is that dopamine neurons activate more or less identically and send the same signal to all of the sites their axons target, but modern evidence suggests that different subpopulations of dopamine neurons respond to input differently depending on the structures to which they send their signals and the different ways these signals act on their target structures
- This can be partially explained by dopamine having functions other than signaling RPEs, but it can also make sense to send different RPEs to different structures depending on the role these structures play in producing reinforced behavior
  - Sending different RPEs to different structures corresponds to an idea of vector valued RPEs
  - Sending different RPEs to different structures can address the structural version of the credit assignment problem (how to distribute credit for a success, or blame for a failure, among the many component structures that could have been involved in producing it)

Dopamine neurons respond with bursts of activity to intense, novel, or unexpected visual and auditory stimuli that trigger eye and body movement, but very little of this activity is related to the movements themselves

In experiments, dopamine neurons phasic activity shifts to the earliest stimulus that is predictive of receiving a reward
- Note that in this shift, activity stops at the point when the reward is actually received
- If a reward is predicted (a predictive stimulus is present) but none is received, dopamine neuron activity drops below the baseline activity level shortly after the time when the reward was expected
- If the wrong action is taken such that a reward is expected but none is received, dopamine neurons activity drops below baseline levels shortly after the time the reward is expected
- Summary:
  - Dopamine neurons respond to unpredicted rewards, to the earliest predictors of reward and activity drops below baseline levels if a reward or predictor of reward does not occur at its expected time
- This is very similar to TD(0) with a CSC representation
  - One key difference is that if a reward is received earlier than expected, dopamine neurons respond to the reward then go back to baseline levels but TD(0) will respond to the reward then also output a negative error when the reward was originally predicted to occur instead of just going back to 0
    - The dopamine neurons seem to be valuing waiting for the reward while the TD error values the states themselves
      - This seems like a problem with the CSC representation, and other

representations seem to do better in this regard (including the microstimulus representation)

Basal Ganglia:

A collection of neuron groups, or nuclei, lying at the base of the forebrain

Striatum:

The main input structure of the basal ganglia

Dorsal Striatum:

A region of the striatum implicated in influencing action selection

Dorsolateral Striatum (DLS):

A part of the dorsal striatum implicated in habit learning

By association this is then also implicated in model-free processes

Dorsomedial Striatum (DMS):

A part of the dorsal striatum implicated in goal-directed learning

By association this is then also implicated in model-based processes

Ventral Striatum:

A region of the striatum implicated in different aspects of reward processing, including the assignment of affective value to sensations

Medium Spiny Neurons:

The main input/output neurons of the striatum

Basal Ganglia Notes:

Essentially all of the cerebral cortex, among other structures, provides input to the striatum

The activity of cortical neurons contains information about sensory input, internal states, and motor activity

These axons of cortical neurons make synaptic contacts on the dendrites of the medium spiny neurons

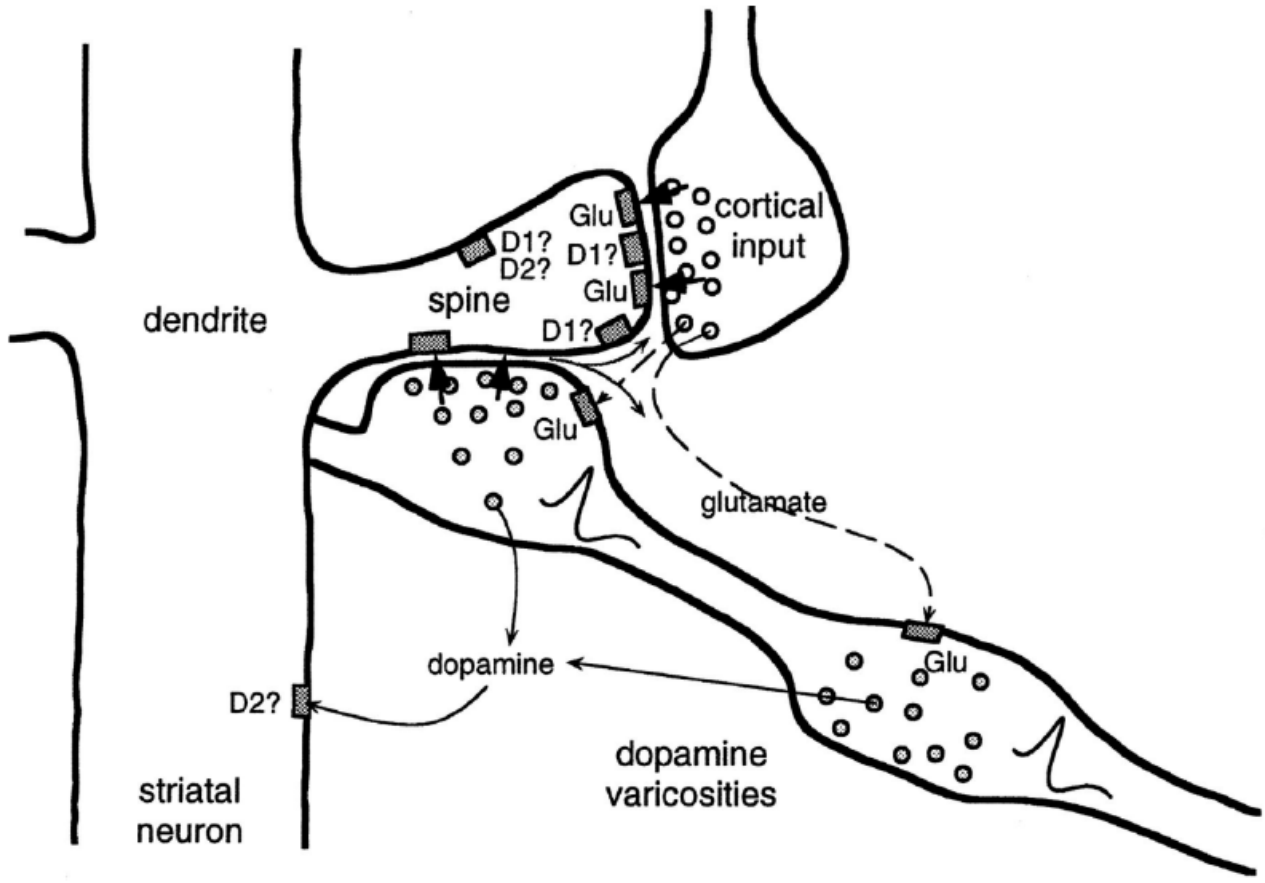
Output from the striatum loops back to the frontal areas of the cortex and to motor areas through other basal ganglia nuclei and the thalamus

The dendrites of medium spiny neurons are covered with spines on whose tips the axons of neurons of the cortex make synaptic contact

The axons of dopamine neurons make contact with these spines as well, but on the spine stem instead of the tip

This connection is called a corticostriatal synapse

Corticostriatal Synapse Diagram:



D1 and D2 here represent sites with different types of dopamine receptors that produce different effects

Glutamate is the neurotransmitter of the cortical inputs

Dopamine varicosities are sites that release dopamine at or near the spine stem in response to glutamate, which brings together presynaptic activity from the cortex, postsynaptic activity from the striatal neuron, and dopamine

These different interactions mean that several different learning rules could govern the plasticity of corticostriatal synapses

These spines may be where changes in synaptic efficacy are governed by learning rules

Earliest Reward Predicting State:

The first state in a trial that reliably predicts the trial's reward

This can also be said as an unpredicted predictor of reward, i.e. a predictor of reward that was not itself predicted

Latest Reward Predicting State:

The state immediately preceding the trial's reward state

Actor-Critic Recap:

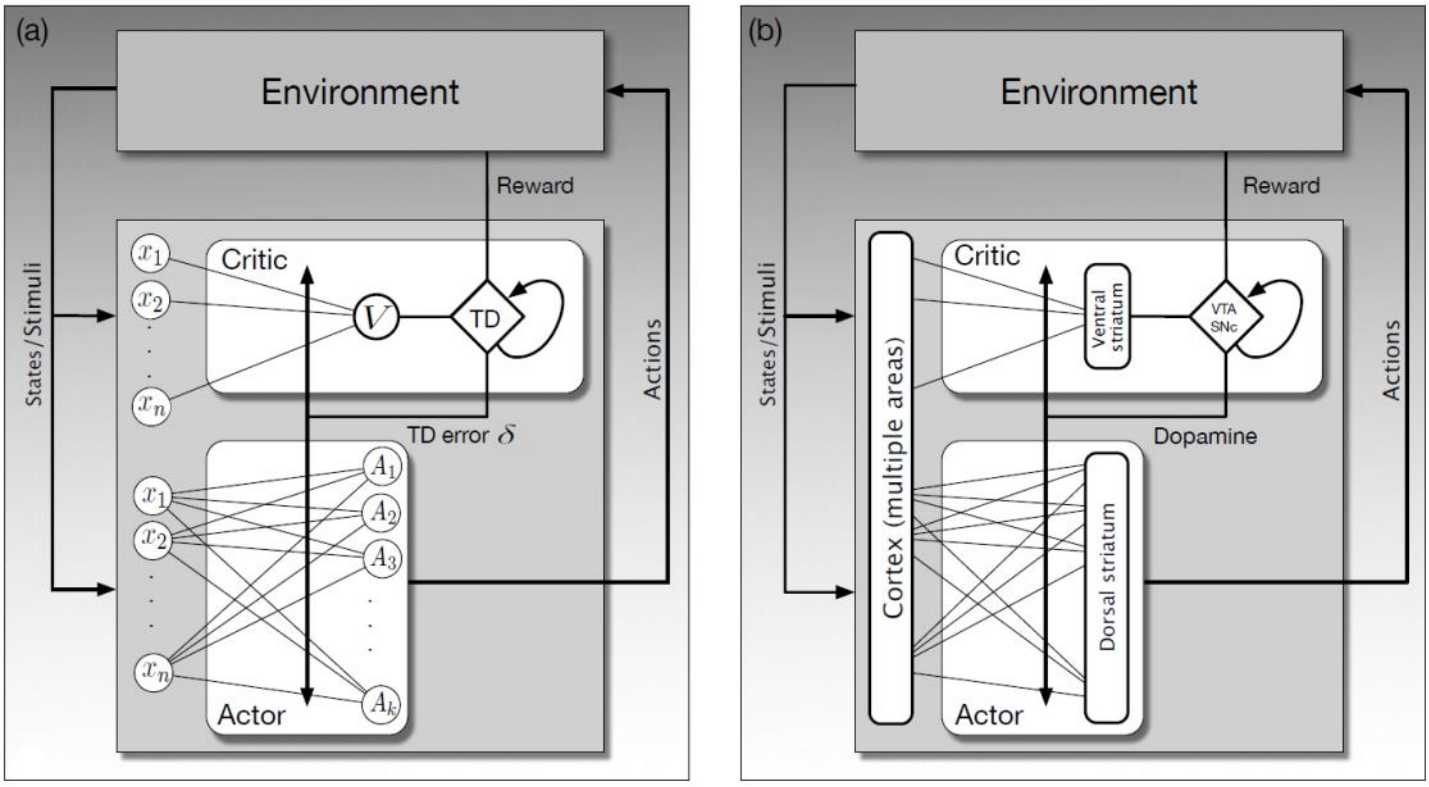
An actor learns policies and a critic uses TD learning to learn a state-value function, the critic then sends TD errors to the actor and the actor updates its policy accordingly

Features of the Brain Suggestive of an Actor-Critic Architecture:

The striatum is broken into two parts, with one being implicated in influencing action selection, and the other implicated in different aspects of reward processing, including assigning affective value to sensations (dorsal and ventral striatum respectively)

The TD error has the dual roles of being the reinforcement signal for both the actor and the critic which fits the axons of dopamine neurons targeting both the dorsal and ventral striatum, being critical for modulating synaptic plasticity in both structures, and dopamine having effects dependent on the target

Comparison Between a Neural Net Actor-Critic Implementation and a Potential Model of the Brain:



Many interconnected neural systems generate reward-related information, with different structures being recruited for different types of rewards so the input to VTA and SNpc labelled reward should be thought of as a vector of reward related information arriving from different channels

The end of this statement is reminiscent of a neural net

Contingent Eligibility Trace:

An eligibility trace that depends on postsynaptic activity

Non-Contingent Eligibility Trace:

An eligibility trace that doesn't depend on postsynaptic activity

Generally non-contingent eligibility traces are adequate for learning prediction but not learning control, which requires a contingent eligibility trace

Neural Actor-Critic:

$\delta_t = R_{t+1} + \gamma \hat{v}\left(S_{t+1}, \vec{w}\right) - \hat{v}\left(S_t, \vec{w}\right)$

$\vec{z}_t^{\vec{w}} = \gamma \vec{\lambda}^{\vec{w}} \vec{z}_{t-1}^{\vec{w}} + \nabla \hat{v}\left(S_t, \vec{w}\right)$

$\vec{z}_t^{\vec{\theta}} = \gamma \vec{\lambda}^{\vec{\theta}} \vec{z}_{t-1}^{\vec{\theta}} + \nabla \ln \pi\left(A_t \middle| S_t, \vec{\theta}\right) = \gamma \vec{\lambda}^{\vec{\theta}} \vec{z}_{t-1}^{\vec{\theta}} + \left(A_t - \pi\left(1 \middle| S_t, \vec{\theta}\right)\right) \vec{x}(S_t)$

$\vec{w} \leftarrow \vec{w} + \alpha^{\vec{w}} \delta_t \vec{z}_t^{\vec{w}}$

$\vec{\theta} \leftarrow \vec{\theta} + \alpha^{\vec{\theta}} \delta_t \vec{z}_t^{\vec{\theta}}$

$\hat{v}\left(s, \vec{w}\right) = \vec{w}^T \vec{x}(s)$

$\nabla \hat{v}\left(S_t, \vec{w}\right) = \vec{x}(s)$

Where the subscript $\vec{w}$ refers to the critic and the subscript $\vec{\theta}$ refers to the actor

This is specifically an actor-critic algorithm for the continuing case using eligibility traces

Note that the actor is working to keep the TD error as positive as possible (take actions leading to greater than expected rewards) while the critic works to reduce the TD error to 0 (accurately predict all rewards)

Note that each synapse has its own eligibility trace, which is the corresponding component of $\vec{z}$

A synapse is eligible for modification if it has a nonzero eligibility trace

The form of eligibility trace used here for the critic $\left(\vec{z}_t^{\vec{w}}\right)$ is a non-contingent eligibility trace while the eligibility trace used for the actor $\left(\vec{z}_t^{\vec{\theta}}\right)$ is a contingent eligibility trace

The postsynaptic contingency is the only difference between the learning rules for the actor and the critic

Here the critic $\hat{v}$ is in the simplest possible form, it would usually be a full neural net

The authors recommend a Bernoulli-logistic unit for the units of the actor neural net so the output of each unit is 0 or 1

The update for the critic is called a two-factor learning rule since the interaction depends on two signals, the reinforcement signal δ and non-contingent eligibility traces

The update for the actor is called a three-factor learning rule since the interaction depends on three factors, the reinforcement signal, and an contingent eligibility trace which itself depends on presynaptic and postsynaptic activity

This model and almost all others ignore activation time, which is the time delay between a signal arrival at the synapse on the presynaptic side and the firing of the postsynaptic neuron

This activation time is on the order of 10s of milliseconds

Activation time is important since it influences how contingent eligibility traces have to work

The current method works since the term $\left(A_t - \pi\left(1\middle|S_t, \vec{\theta}\right)\right)\vec{x}(S_t)$ ignores activation time and says that the presynaptic activity helps cause the postsynaptic activity but a more sophisticated version of causation here that includes activation time would be better

Hebbian Theory:

Synaptic efficacy increases as a result of a presynaptic cell's repeated and persistent stimulation of a postsynaptic cell

Spike-Timing-Dependent Plasticity (STDP):

A Hebbian-style plasticity but where changes in synaptic efficacy depend on the relative timing of presynaptic and postsynaptic action potentials

The above mentioned dependence can take many forms

The most common form is where a synapse increases in efficacy if spikes incoming via the synapse arrive shortly before the postsynaptic neuron fires and decreases in efficacy if spikes incoming via the synapse arrive shortly after the postsynaptic neuron fires

Note that this takes activation time into account

This was discovered by neuroscientists

Reward Modulated STDP:

A form of STDP where changes that would be produced by normal STDP only occur if a neuromodulatory input is present within a time window after a presynaptic spike is closely followed by a postsynaptic spike

There is mounting evidence that this occurs at the spines of medium spiny neurons in the dorsal striatum with dopamine acting as the neuromodulator

Experiments have shown the time window in the brain to be around 10 seconds (see Yagishita et al. 2014)

This points to eligibility traces having prolonged time courses

Hedonistic Neuron Hypothesis:

The idea that neurons seek to maximize the difference between synaptic input treated as rewarding and synaptic input treated as punishing by adjusting their synaptic efficacies on the basis of the rewarding or punishing consequences of their own action potentials

This was originally introduced by Klopf and included the idea that rewards and punishments were conveyed to a neuron by the same synaptic input that excites or inhibits the neuron's phasic activity

A more modern take would be to assign the reinforcing role to a neuromodulator

In this hypothesis synaptically-local traces of pre- and postsynaptic activity make neurons eligible for modification by later reward or punishment

This is implemented by a contingent eligibility trace

In this theory the shape and course of an eligibility trace reflects the durations of the feedback loops in which the neuron is embedded such that the shape of a synaptic eligibility

trace is like a histogram of these feedback loops

Lateral Inhibition:

The capacity of an excited neuron to reduce the activity of its neighbors

This can create a contrast in stimulation to allow for increased sensory perception

This occurs primarily in visual processes but also in tactile, auditory, and even olfactory processing

In multi-agent reinforcement learning, if each actor can learn effectively then the collective action of the team will improve as evaluated by a common reward signal, even when agents can't communicate with each other

However, note that this is a very difficult learning task since each agent can only partially observe the state of its environment and each agent's reward signal will be corrupted by a large amount of noise

Requirements for Collective Learning:

Each unit's learning algorithm has to have a contingent eligibility trace

This is so the relation between a team member's actions and the team reward can be estimated

There has to be variability in the actions of team members

This is so the team can explore the space of collective actions

Independent exploration by team members is the simplest case, but more sophisticated methods are possible including winner-take-all arrangements where the winners take the credit or blame for the resulting reward or punishment

Winner take all methods have an analog in the brain of lateral inhibition and in neural nets of competitive learning

A team of Bernoulli-logistic units using REINFORCE will ascend the average reward gradient when interconnected to form a neural network with only a reward signal broadcast to all units

This was done by Further and Williams (1992)

Orbitofronal Cortex (OFC):

A part of the prefrontal cortex directly above the eyes implicated in model-based (goal-directed) processes including planning and decision making

Hippocampus:

A structure of the brain critical for memory and spatial navigation also implicated in model-based (goal-directed) behavior

The hippocampus may be important in our ability to imagine new experiences

Experiments decoding the activity of neurons in the hippocampus in rats shows the representation of space in the hippocampus sweeping forward along the possible paths the rat can take from that point

Note that this is similar to a rollout algorithm

Unresolved Questions Related to Model-Based/Model-Free Processes in the Brain:

How can structurally similar areas of the brain like the DLS and DMS be essentially components of behavior that appears to be model-free and model-based respectively?

Are separate structures responsible for what we call the transition and reward components of an environmental model?

Is all planning similar to a rollout algorithm as the hippocampus experiments mentioned above suggest?

Alternatively, are models sometimes engaged in the background to refine or recompute value information such the Dyna architecture suggests?

How does the brain arbitrate between the habit and goal-directed systems?

Is there actually a clear distinction between the habit and goal-directed systems?

Some evidence points to no for this question, with dopamine signals exhibiting model-based influences

Many--through not all--drugs of abuse increase dopamine levels either directly or indirectly in regions around terminals of dopamine neuron axons in the striatum

One model for addiction in the context of TD error is to make drugs prevent δ from ever becoming

negative so the reward from drugs cannot be predicted away
Page 435 is the start of the bibliographical and historical remarks section, which gives many references for further search