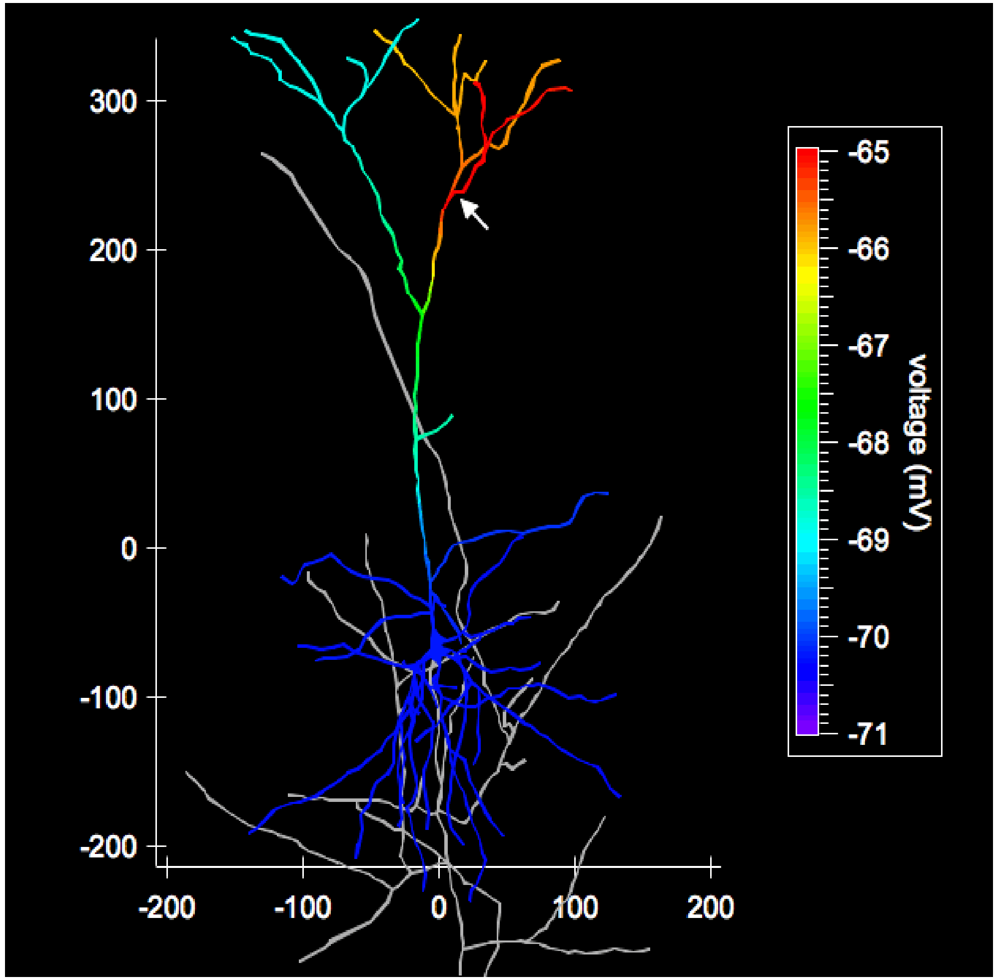
Reinforcement Learning in a Neurally Controlled Robot Using Dopamine Modulated STDP

By Richard Evans

Aug 8, 2023
Sungmin Yoon

Index

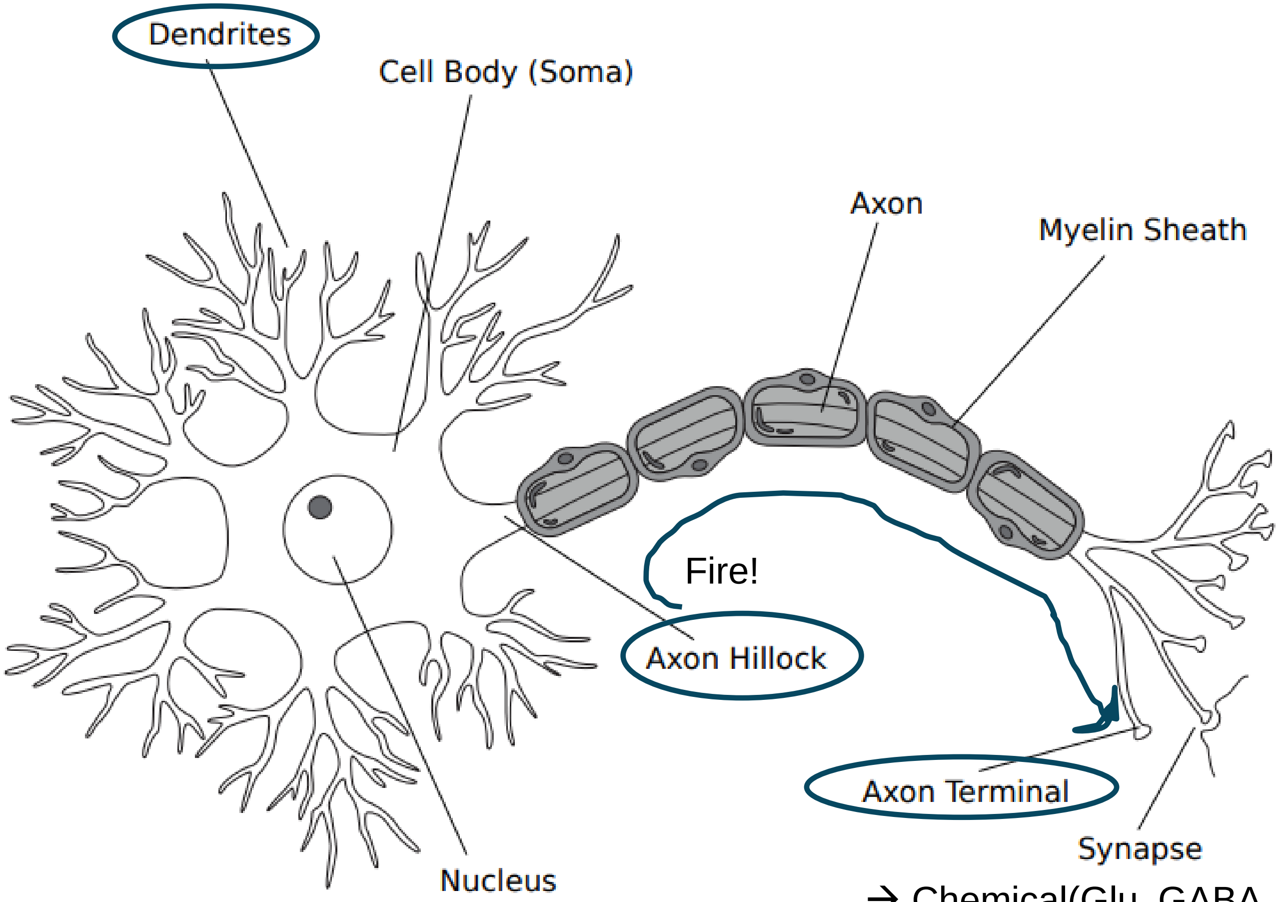
- Neurons
- RL in Brain
- SNN controlled robot
- Discussion



Neurons

- General architecture

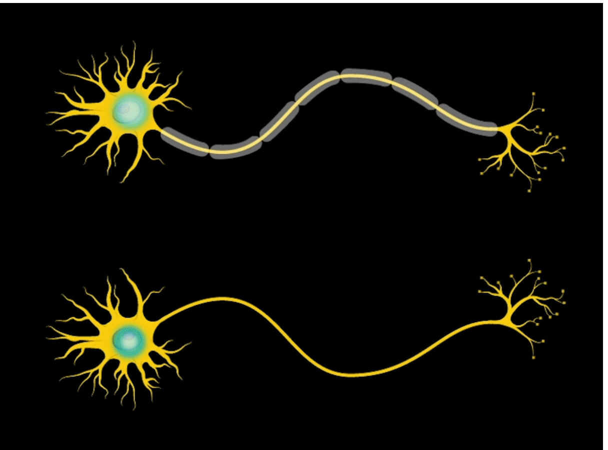
Input from several
Pre-synaptic neurons



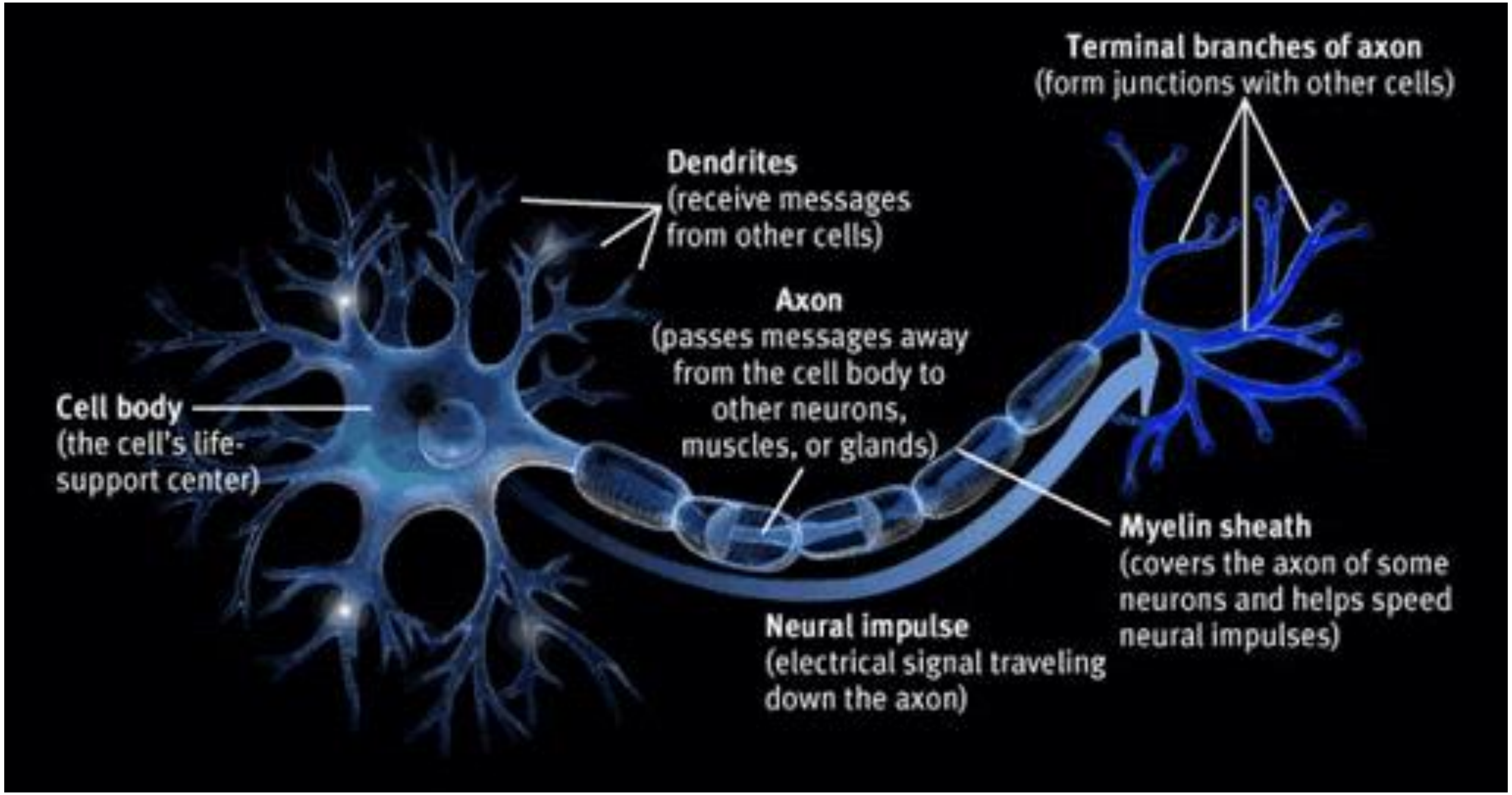
→ Chemical(Glu, GABA..)
→ Refractory period

Excitatory
or
Inhibitory

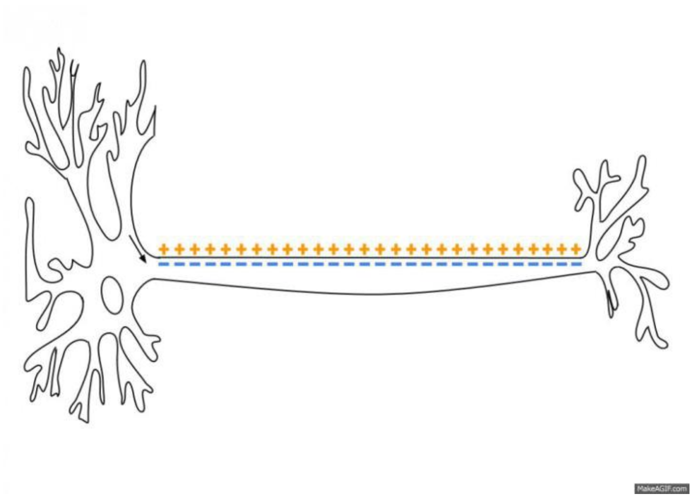
Figure 2.1: The structure of a neuron.

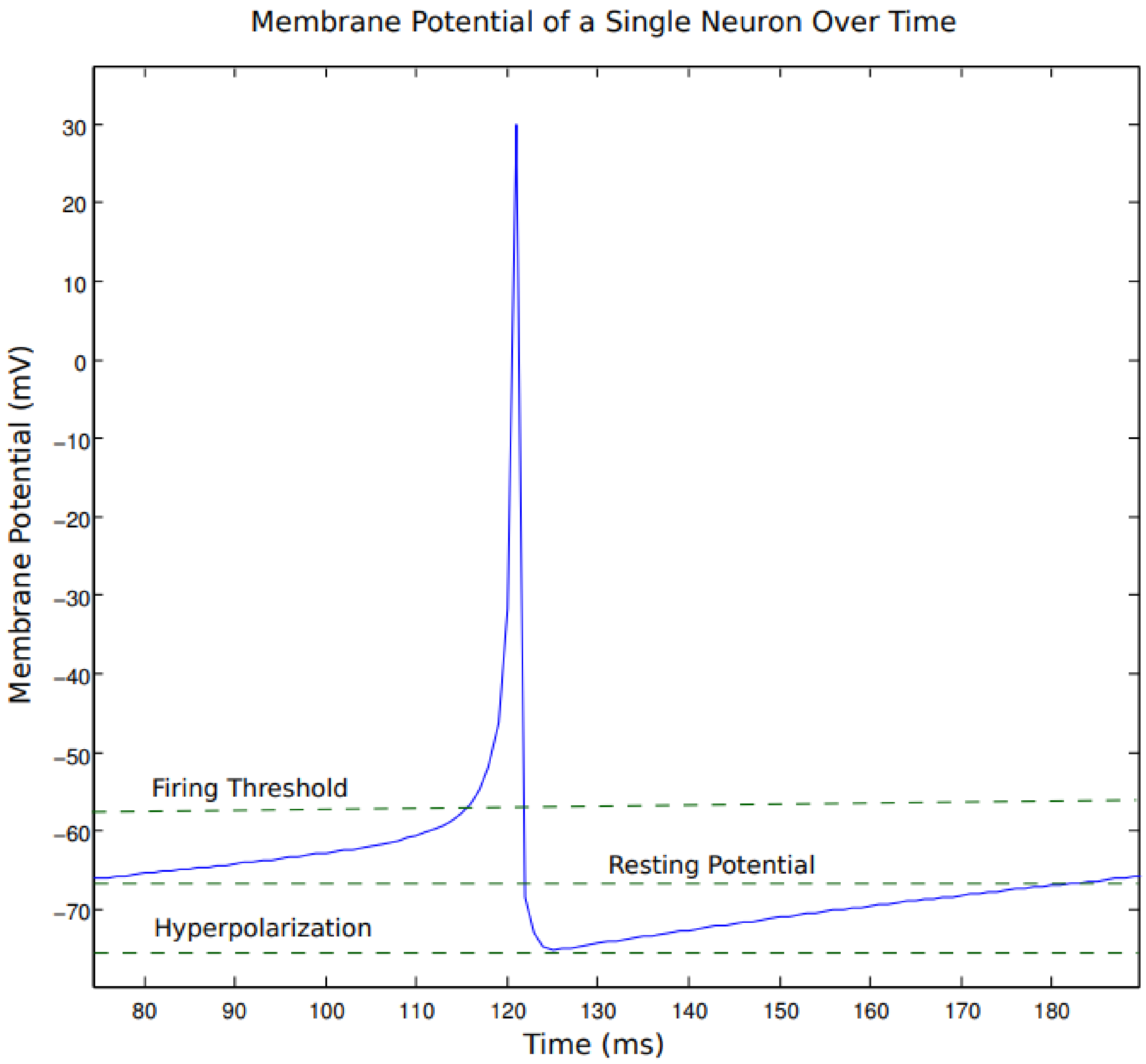


Neurons


Terminal branches of axon
(form junctions with other cells)
Dendrites
(receive messages
from other cells)
Axon
(passes messages away
from the cell body to
other neurons,
muscles, or glands)
Cell body
(the cell's life-
support center)
Myelin sheath
(covers the axon of some
neurons and helps speed
neural impulses)
Neural impulse
(electrical signal traveling
down the axon)

Neurons

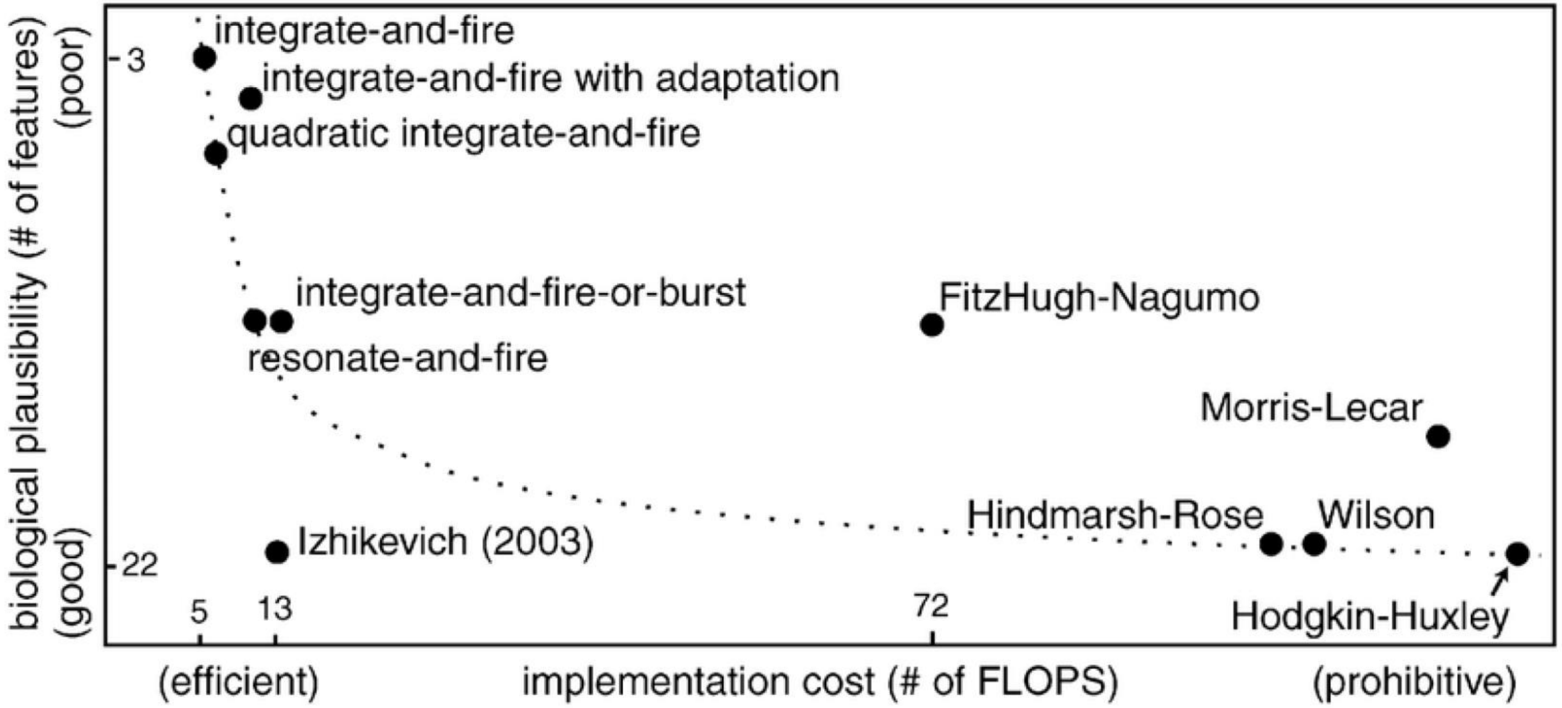




Membrane Potential of a Single Neuron Over Time
Membrane Potential (mV)
30
20
10
0
−10
−20
−30
−40
−50
−60
−70
Firing Threshold
Resting Potential
Hyperpolarization
80
90
100
110
120
130
140
150
160
170
180
Time (ms)

Neurons

- Hodgkin-Huxley Model
- Izhikevich Model
- LIF Model



Neurons

- Hodgkin-Huxley Model

$$C\frac{dv}{dt} = -\sum_k I_k + I$$

where

$$\begin{aligned} C &= \text{The capacitance of the neuron,} \\ v &= \text{The membrane potential of the neuron,} \\ I_k &= \text{The various ionic currents that pass through the cell,} \\ I &= \text{The external current coming from pre-synaptic neurons,} \\ t &= \text{Time.} \end{aligned}$$

Neurons

- Izhikevich Model

$$\frac{dv}{dt} = 0.04v^2 + 5v + 140 - u + I \tag{2.2}$$

u is the recovery variable that determines the refractory period

$$\frac{du}{dt} = a(bv - u) \tag{2.3}$$

$$\text{if } v \geq 30 \text{ then} \begin{cases} v \leftarrow c \\ u \leftarrow u + d \end{cases} \tag{2.4}$$

Neurons

- Izhikevich Model

$$\frac{du}{dt} = a(bv - u)$$

$$\geq 30 \text{ then } \begin{cases} v \leftarrow c \\ u \leftarrow u + d \end{cases}$$

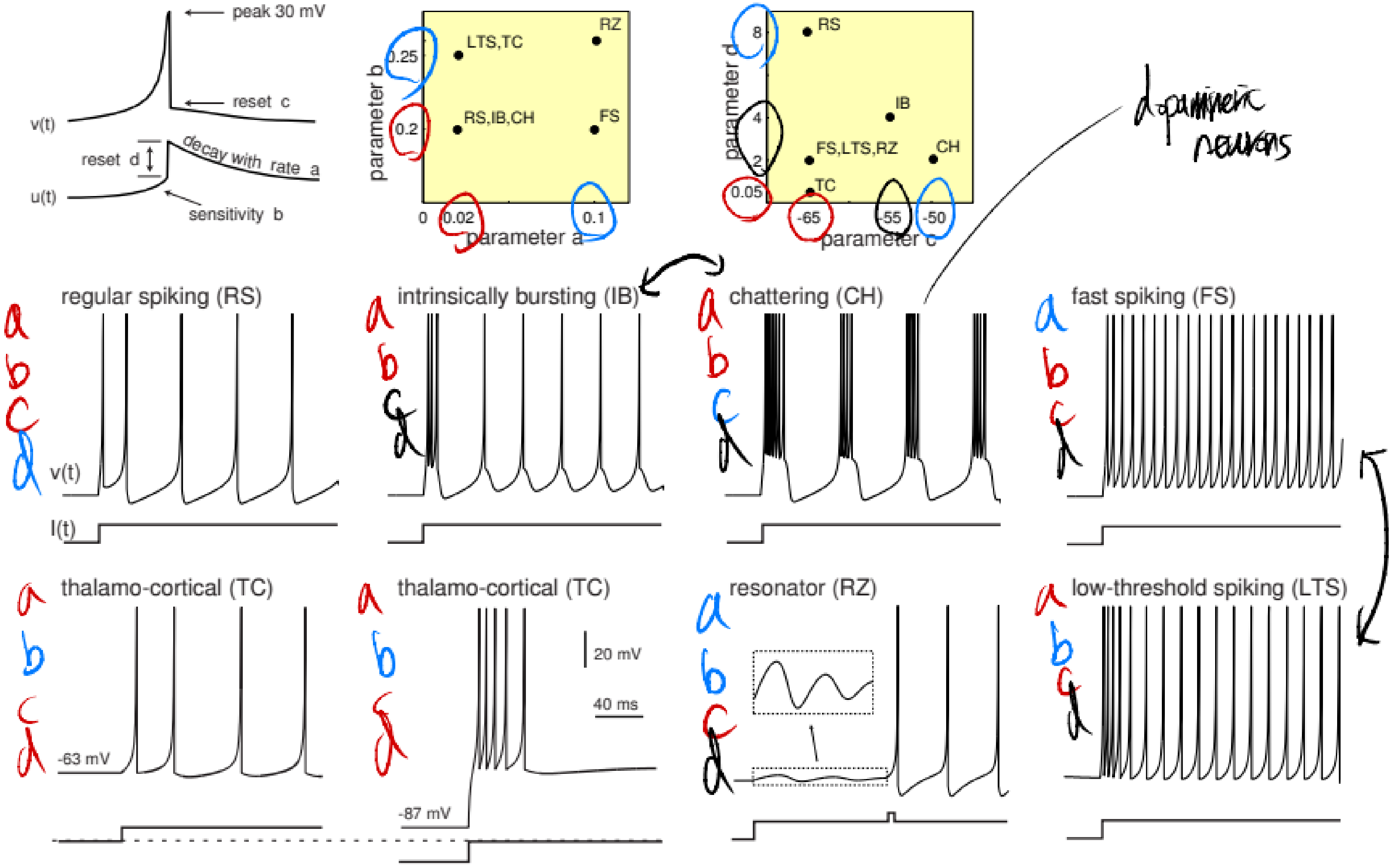


Figure 2.3: An overview of some types of neurons that can be modelled with the Izhikevich model[1].

Reinforcement Learning

- Markov property
- Q-function
- Sarsa
- Q-Learning

$$Q^{\pi}(s,a) = E[\sum_{k=0}^{\infty} \gamma^k r_{t+k+1} | \pi, s_t = s, a_t = a]$$

Algorithm 1 Sarsa

Initialize $Q(s,a)$ arbitrarily
Repeat (for each episode)
  Initialize s
  Choose a from s using policy derived from Q
  Repeat (for each step of episode):
    Take action a, observe r, s'
    Choose a' from s' using policy derived from Q
    $Q(s,a) \leftarrow Q(s,a) + \alpha[r + \gamma Q(s',a') - Q(s,a)]$
    $s \leftarrow s'; a \leftarrow a';$
  until s is terminal

Algorithm 2 Q-Learning

Initialize $Q(s,a)$ arbitrarily
Repeat (for each episode)
  Initialize s
  Choose a from s using policy derived from Q with exploration
  Repeat (for each step of episode):
    Take action a, observe r, s'
    $Q(s,a) \leftarrow Q(s,a) + \alpha[r + \gamma max_{a'} Q(s',a') - Q(s,a)]$
    $s \leftarrow s'$
  until s is terminal

Reinforcement Learning

- Eligibility Traces

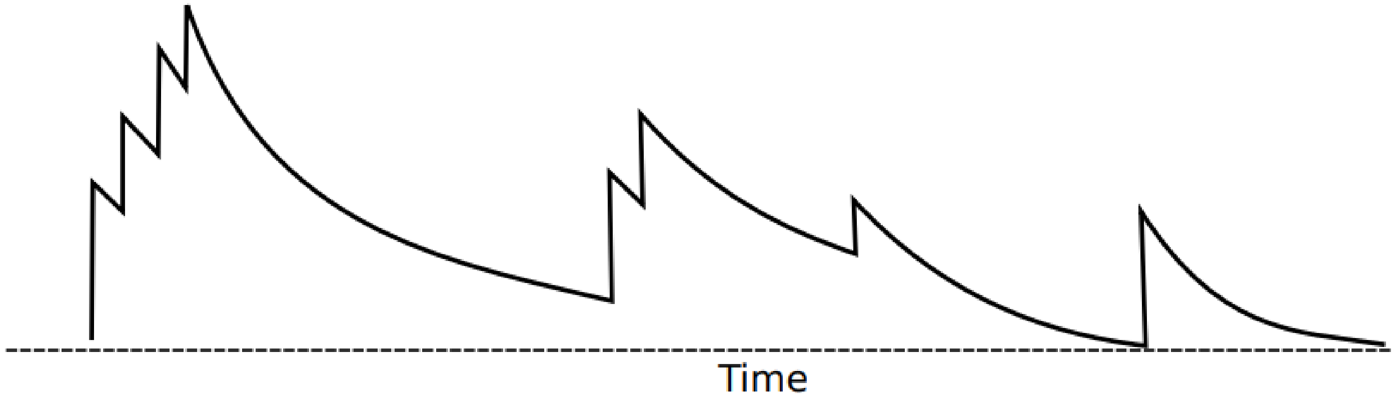


Figure 2.6: The eligibility trace for a state over time, as it is repeatedly visited.

We can incorporate the eligibility trace into the Sarsa algorithm (referred to as Sarsa(λ)), if we define $e(s,a)$ as the eligibility trace for the state s and action a then the Q-function update becomes:

$$Q(s,a) \leftarrow Q(s,a) + \alpha e(s,a)[r + \gamma Q(s',a') - Q(s,a)] \tag{2.6}$$

Reinforcement Learning

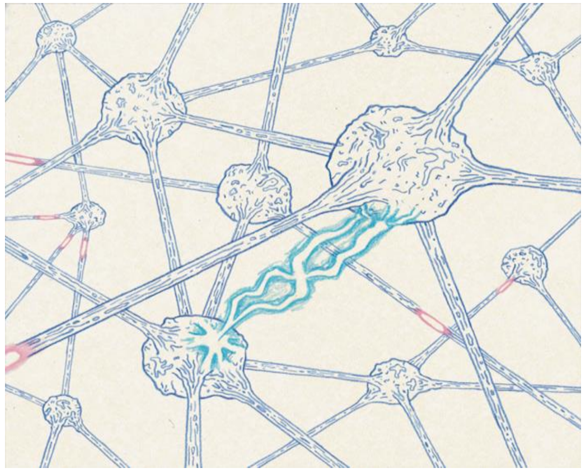
- Eligibility Traces

Algorithm 3 Sarsa(λ)

Initialize $Q(s, a)$ arbitrarily and $e(s, a) = 0$, for all s,a
Repeat (for each episode)
  Initialize s, a
  Repeat (for each step of episode):
    Take action a, observe r, s'
    Choose a' from s' using policy derived from Q
    $\delta \leftarrow r + \gamma Q(s', a') - Q(s, a)$
    $e(s, a) \leftarrow e(s, a) + 1$
    For all s,a:
      $Q(s, a) \leftarrow Q(s, a) + \alpha\delta e(s, a)$
      $e(s, a) \leftarrow \gamma\lambda e(s, a)$
    $s \leftarrow s'; a \leftarrow a'$
  until s is terminal

RL in the Brain

- BCM Theory

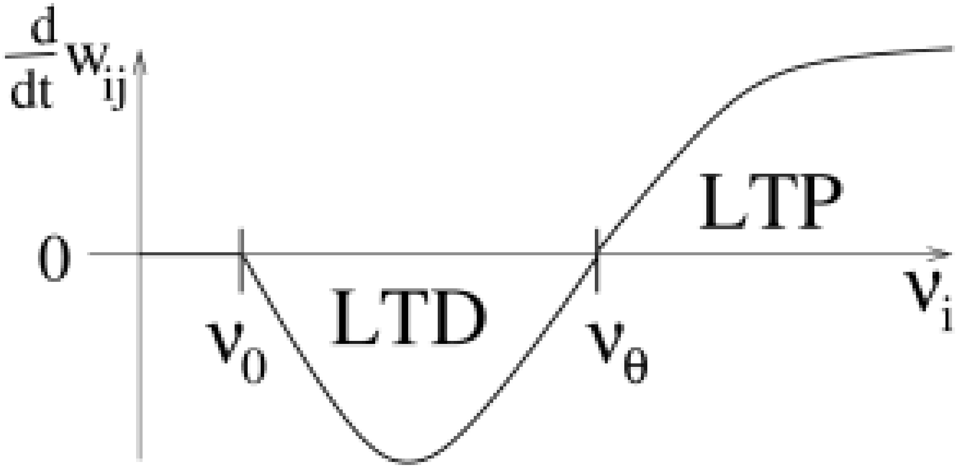


Hebb : “Fire together, Wire together”

→ LTP (long term potentiation)

BCM (Bienenstock, <u>Cooper</u>, Munro)

→ LTD (long term depression)



RL in the Brain

- STDP (Spike-timing Dependent Plasticity)

$$\Delta w = \begin{cases} A^{+}e^{-\Delta t/\tau^{+}} & \text{if } \Delta t \geq 0 \\ -A^{-}e^{\Delta t/\tau^{-}} & \text{if } \Delta t < 0 \end{cases}$$

Δw is the weight update

$\Delta t = t_{post} - t_{pre}$

A^{+},A^{-},τ^{+} and τ^{-} are constants that define how STDP is applied over time.

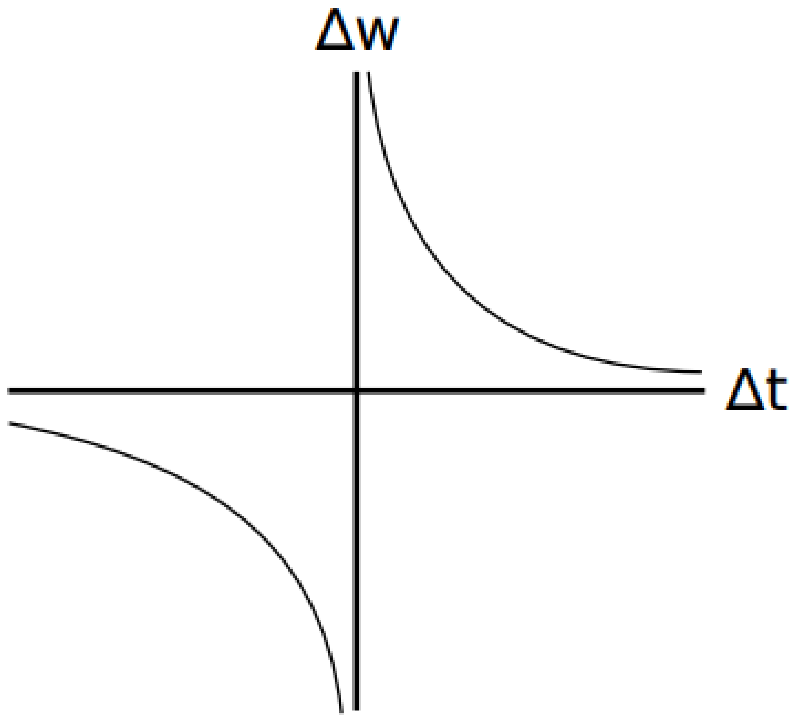
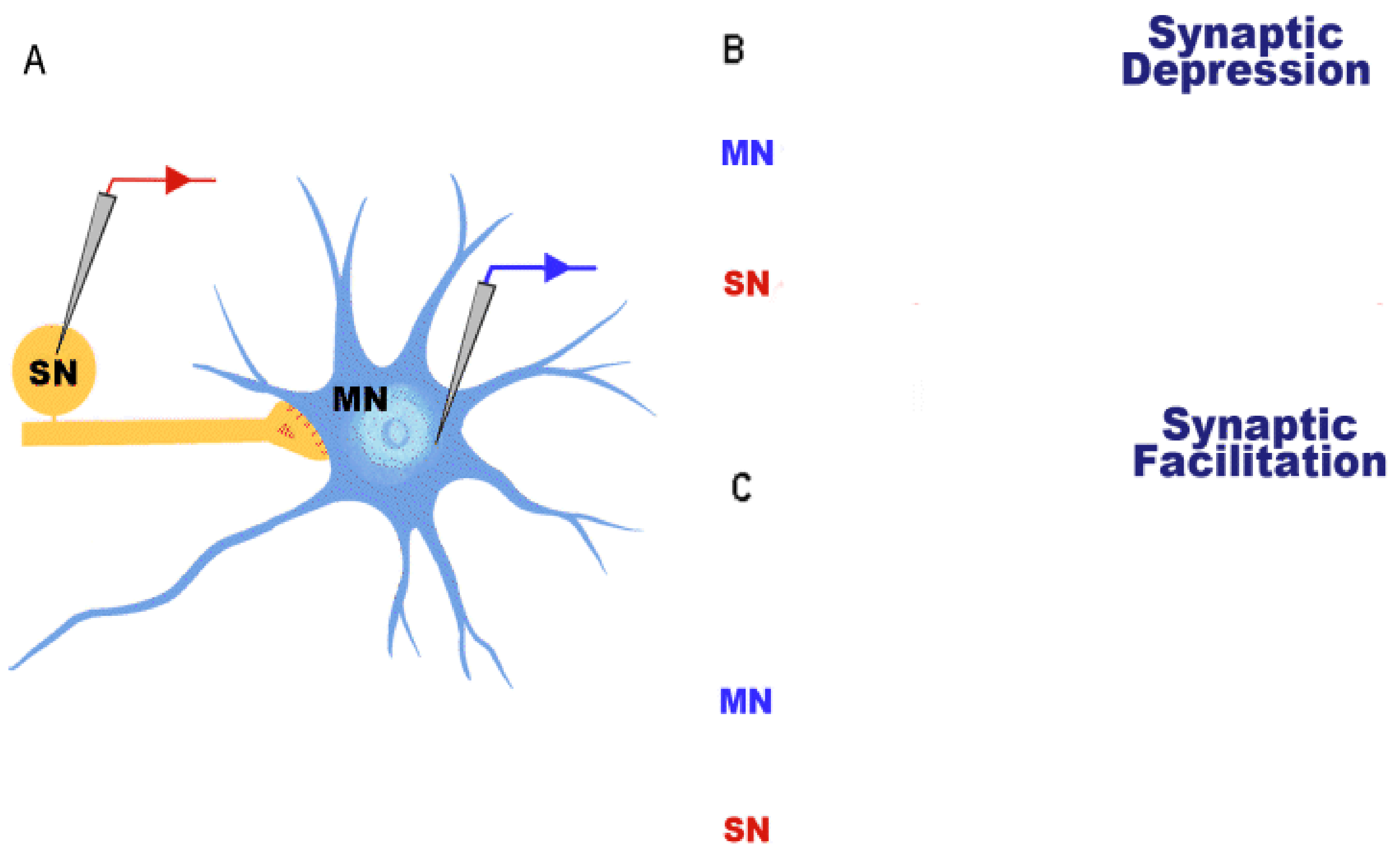


Figure 2.7: Graph showing how the weight update Δw relates to the $\Delta t = t_{post} - t_{pre}$ parameter.

RL in the Brain

- STDP

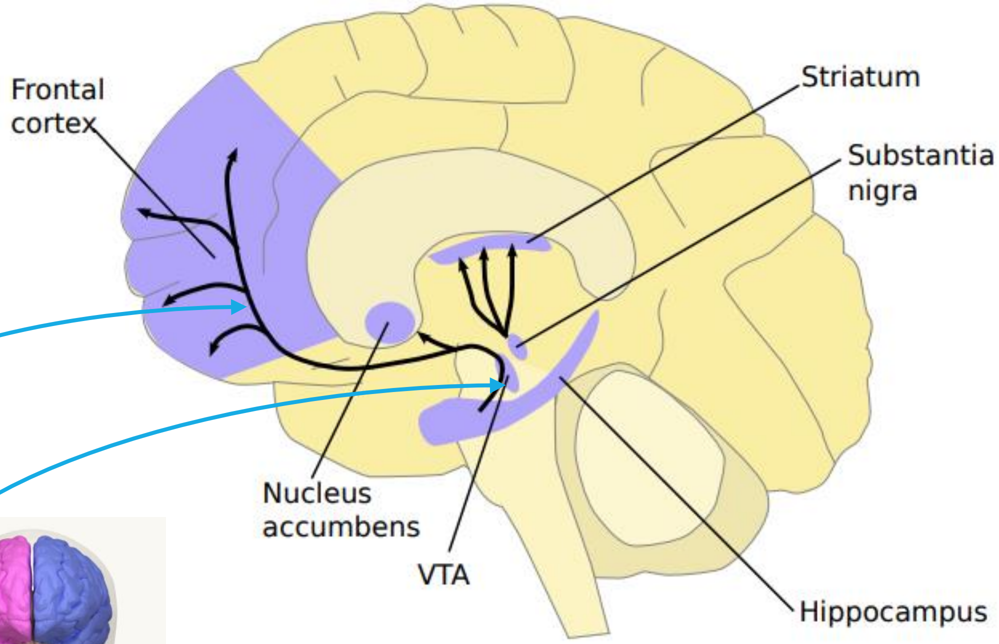


RL in the Brain

- Dopamine Modulated STDP

This is the pathway of the dopamine!

Where the most of dopaminergic neurons are! (VTA)



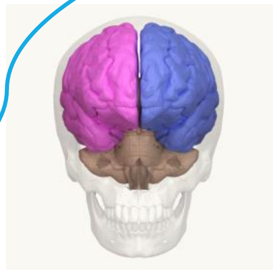


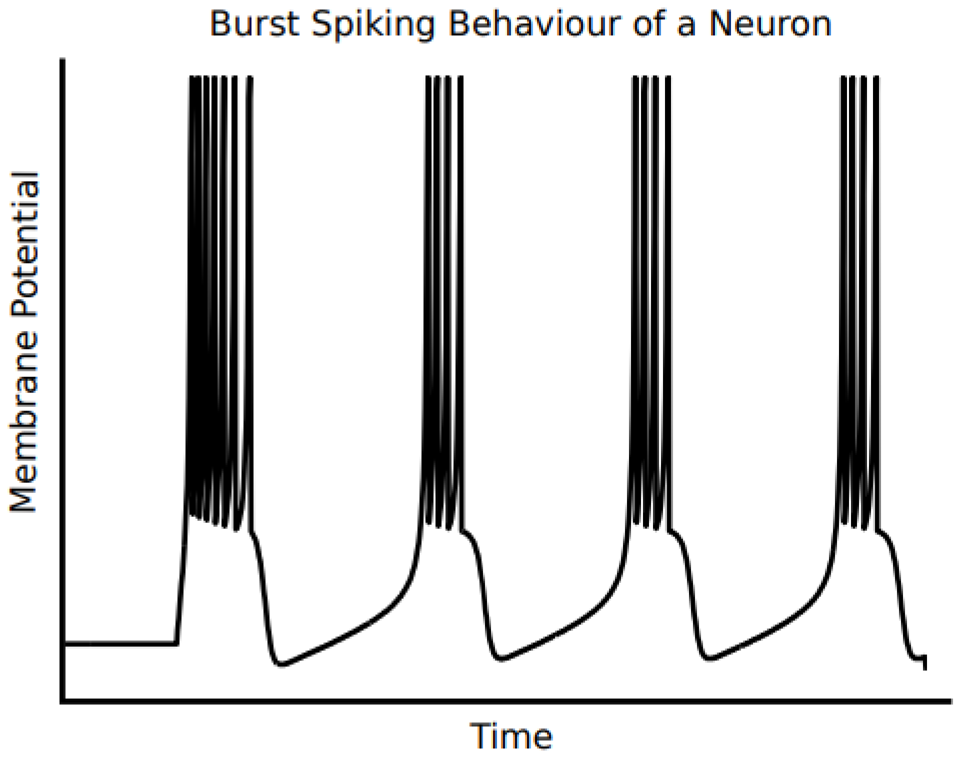
Figure 2.8: The main dopamine pathways in the human brain.

RL in the Brain

- Dopamine Modulated STDP

Two different firing pattern(dopaminergic neurons)

1. Background firing (stimulus X)
2. Burst firing (stimulus O)



RL in the Brain

$$\dot{c} = -c/\tau_c + STDP(\tau)\delta(t - t_{pre/post})$$

- Synaptic tag
  → For distal reward problem

decay rate of
the eligibility trace

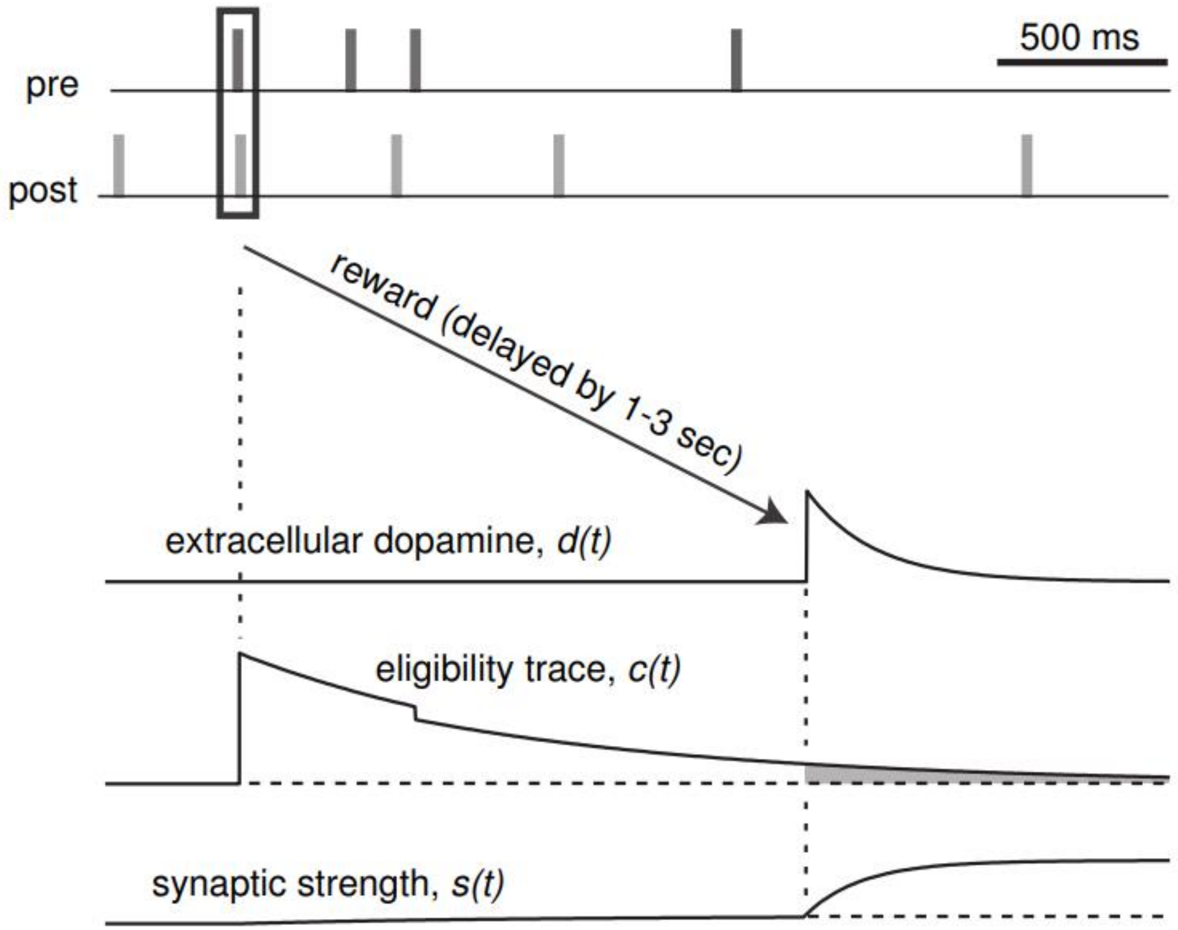
Dirac delta function

RL in the Brain

- Synaptic tag
  - → For distal reward problem

$$\dot{s} = cd$$

Where s is the synapse strength and d is the current level of dopamine.



RL in the Brain

- Dual-path model

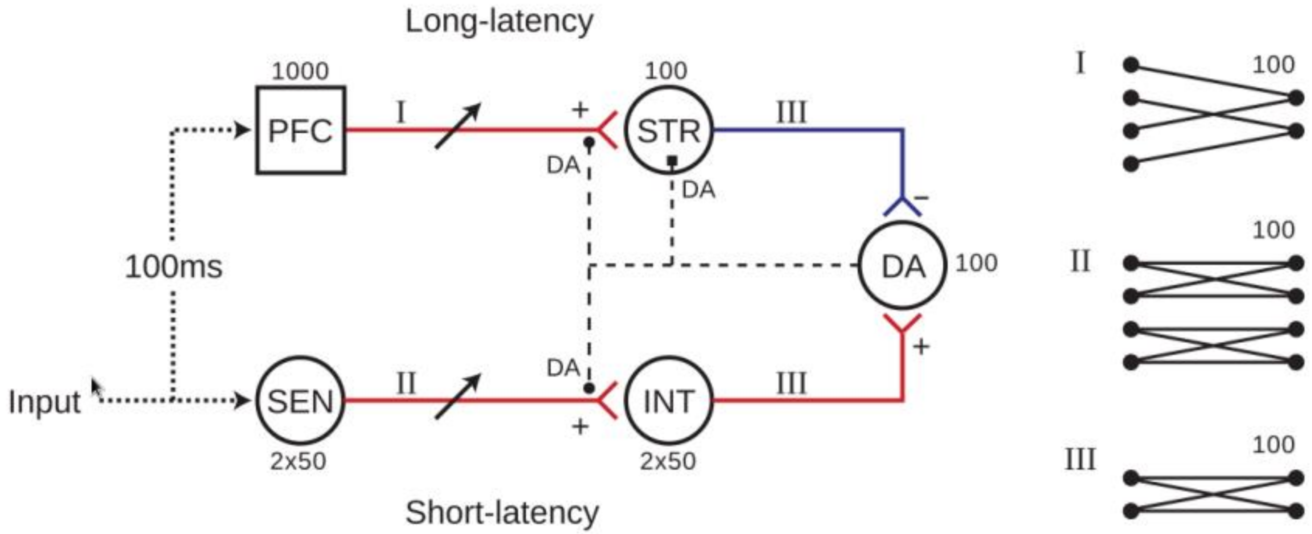


Figure 2.12: The network architecture used by Chorley & Seth [31], red lines represent excitatory connections and blue represent inhibitory connections. When neurons in the DA module fire then dopamine is released which causes STDP of the PFC→STR and SEN→INT pathways. The mean firing rate of the STR module is also modulated by the amount of dopamine.

RL in the Brain

- Dual-path model

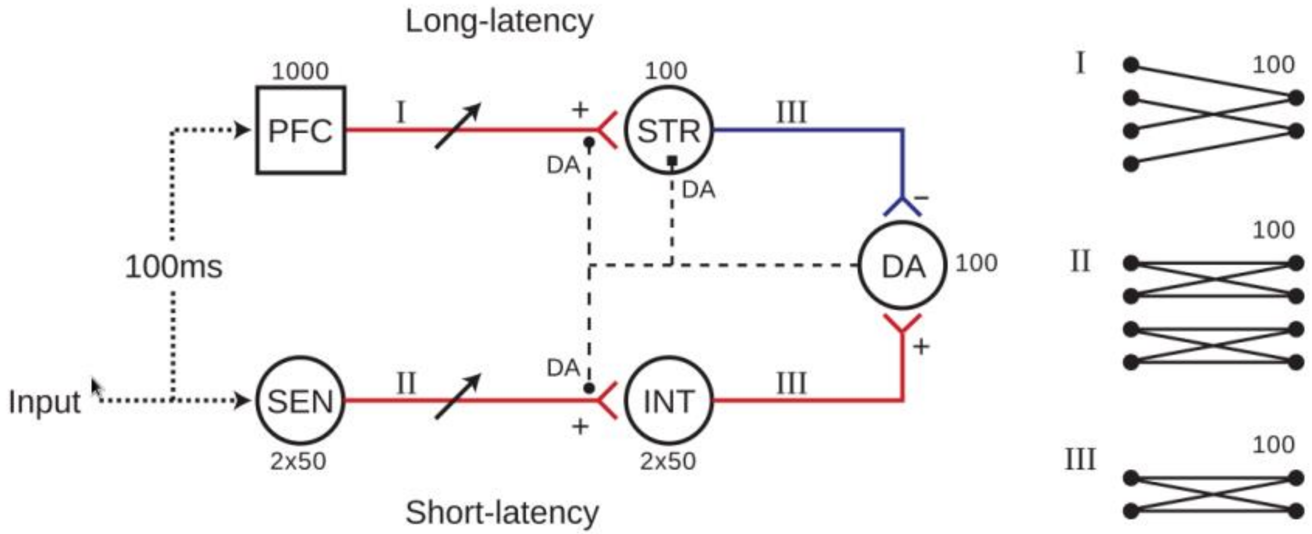


Figure 2.12: The network architecture used by Chorley & Seth [31], red lines represent excitatory connections and blue represent inhibitory connections. When neurons in the DA module fire then dopamine is released which causes STDP of the PFC→STR and SEN→INT pathways. The mean firing rate of the STR module is also modulated by the amount of dopamine.

Neural encoding

- Rate coding
- Population coding
- Temporal coding

Discussion

- Out of date(2015)
- bridge between RL and SNN