Metalearning and Neuromodulation

Kenji Doya

Introduction

- A computational theory on the role of neuromodulators.
- Relates the metaparameters of reinforcement learning with neuromodulators.
- The proposed roles of neuromodulators are:

Dopamine (DA)	Signals the error in reward prediction.
Serotonin (5-HT)	Controls the time-scale of reward prediction.
Noradrenaline (NA)	Controls the randomness in action selection.
Acetylcholine (ACh)	Controls the speed of memory update.

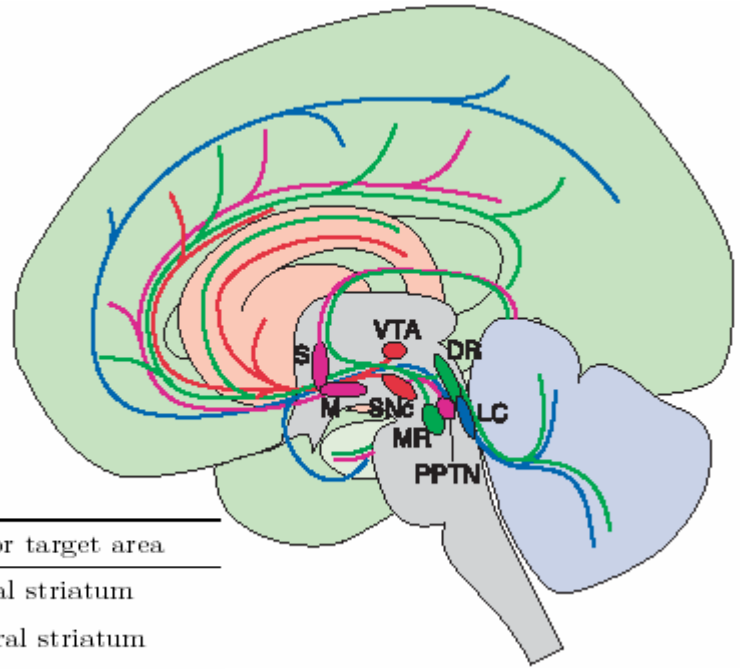
Division of Talk

- Background
- Hypothetical roles of neuromodulators.
- Dynamic interactions of neuromodulators.
- Conclusions
- References

Part I - Background

Origin and Targets of Neuromodulators

DA	Red
5-HT	Green
NA	Blue
ACh	Magenta



neuromodulator	origin of projection	major target area
dopamine (DA)	substantia nigra, pars compacta (SNc)	dorsal striatum
	ventral tegmental area (VTA)	ventral striatum
		frontal cortex
serotonin (5-HT)	dorsal raphe nucleus (DR)	cortex, striatum
		cerebellum
	median raphe nucleus (MR)	hippocampus
noradrenaline (NA)	locus coeruleus (LC)	cortex, hippocampus
(norepinephrine, NE)		cerebellum
acetylcholine (ACh)	Meynert nucleus (M)	cortex, amygdala
	medial septum (S)	hippocampus
	pedunculopontine tegmental	SNc, thalamus
	nucleus (PPTN)	superior colliculus

What is a Metaparameter?

• Parameters which globally effect the way in which system parameters change during learning.

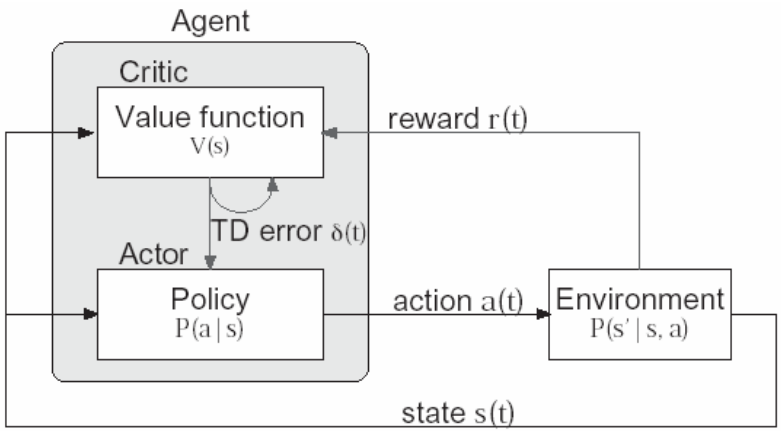
Examples are:

• Speed of learning.

• The size of noise for exploration.

• The time-scale for prediction of future rewards.

Reinforcement Learning Algorithm (RLA) – (1)



Goal: To find an *optimal policy* that maximizes the expected sum of future rewards.

$s(t) \in \{s_1, s_2, \ldots, s_n\}$;

$a(t) \in \{a_1, a_2, \ldots, a_n\}$;

	Deterministic	Stochastic
Policy	G(s)	P(a|s)
State	S(t+1) = F(s(t), a(t))	P(s(t+1)|s(t), a(t))
Reward	R(t+1) = R(s(t), a(t))	P(r(t+1)|s(t), a(t))

RLA – (2) State Value Function and TD Error

The *value function* is defined as:

$$V(s(t)) = E[r(t+1) + \gamma r(t+2) + \gamma^2 r(t+3) + ...].$$

Consistency Condition:

$$V(s(t-1)) = E[r(t) + \gamma r(t+1) + \gamma^2 r(t+2) + ...] = E[r(t) + \gamma V(s(t))].$$

Deviation from consistency condition is the *temporal difference* (TD) error.

$$\delta(t) = r(t) + \gamma V(s(t)) - V(s(t-1))$$

Correction to *value function* is proportional to δ.

$$\Delta V(s(t-1)) \propto \delta(t).$$

RLA – (3) Action Value Function and Policy

The *action value function* Q(s,a) is defined as:

$$Q(s(t), a) = E[r(t+1) + \gamma V(s(t+1))|a(t) = a]$$

Policy is defined via the *action value function* Q:

$$P(a_i|s) = \frac{\exp[\beta Q(s, a_i)]}{\sum_{j=1}^{m} \exp[\beta Q(s, a_j)]}.$$

β- Inverse Temperature

β = 0: Action Selection is random

In the limit of $\beta \rightarrow \infty$, action selection is deterministic:

$$a(t) = \arg\max_a Q(s(t), a).$$

The *action value function* is updated by:

$$\Delta Q(s(t-1), a(t-1)) \propto \delta(t)$$

RLA – (4) Global Learning Signal & Metaparameters

V(s) and Q(s,a) are often represented as weighted sum of basis functions:

$$V(s) = \sum_j v_j b_j(s) \qquad Q(s,a) = \sum_k w_k c_k(s,a)$$

The weight parameters are updated as:

$$\Delta v_j = \alpha \delta(t) b_j(s(t-1)) \qquad \Delta w_k = \alpha \delta(t) c_k(s(t-1), a(t-1))$$

• Change in parameters like v_j and w_k, is dependent on global parameters like learning rate α, inverse temperature β, discount factor γ, and TD error δ. These global parameters are called as *metaparameters*.

Part II – Hypothetical Role of Neuromodulators

Key Hypotheses

- Dopamine (DA) signals the TD error δ.
- Serotonin (5-HT) controls the discount factor γ.
- Noradrenaline (NA) controls the inverse temprature β.
- Acetylcholine (ACh) controls the learning rate α.

Dopamine and Reward Prediction

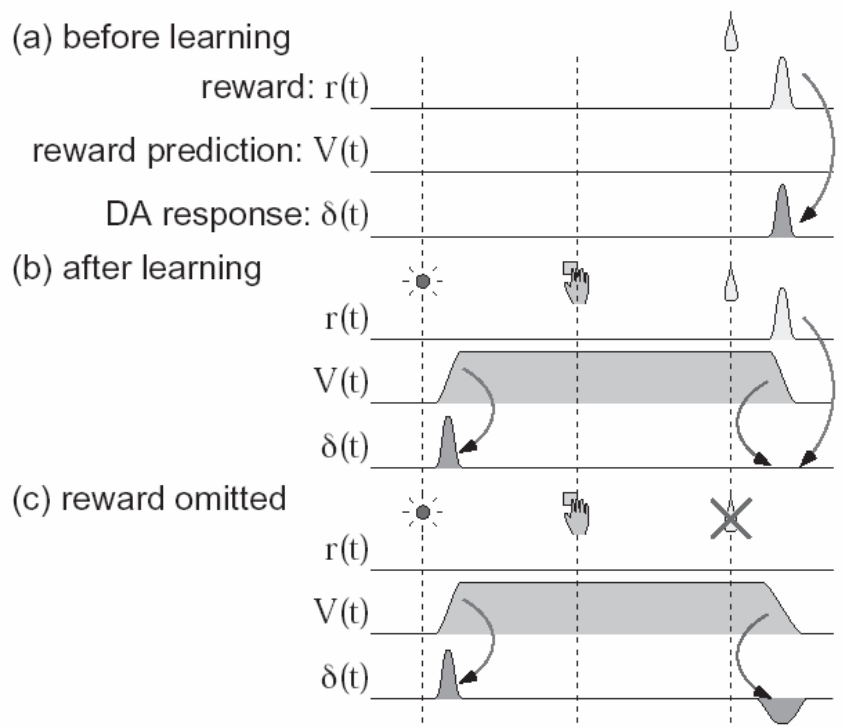
$$\delta(t) = r(t) + \gamma V(s(t)) - V(s(t-1))$$

Before Learning: V(t) = 0;

→ δ(t) = r(t).

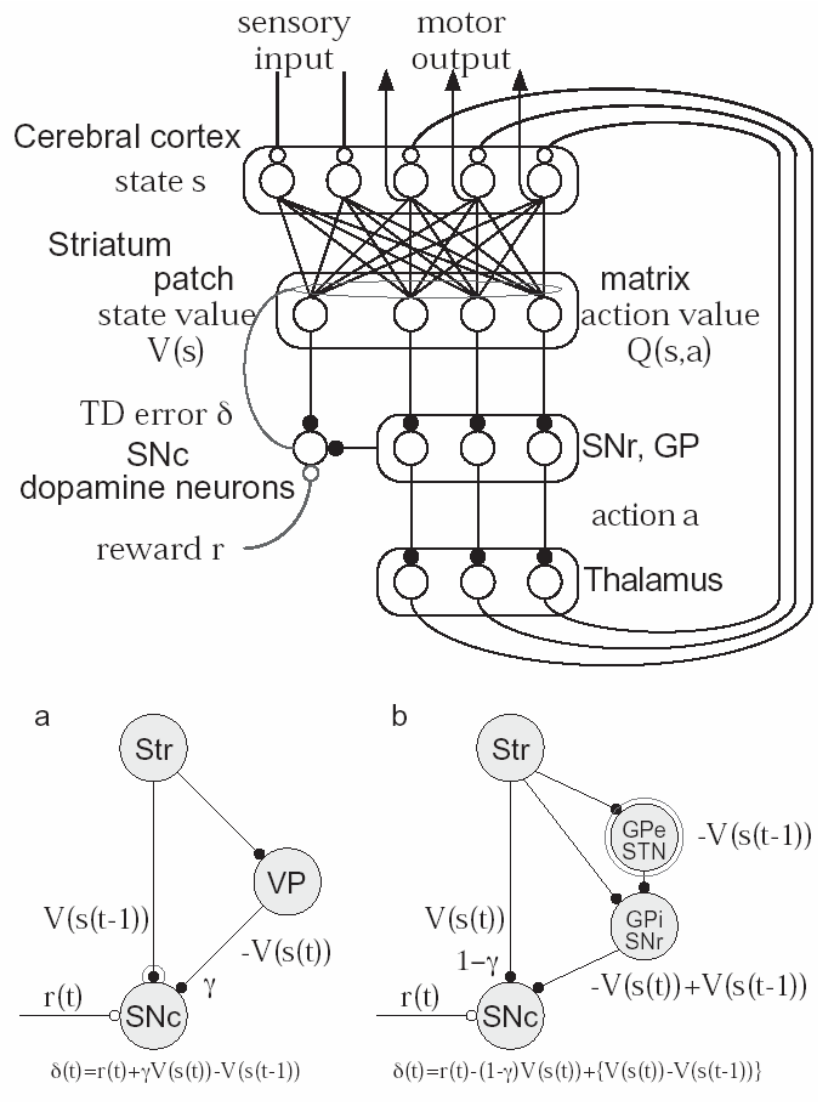
• After Learning: V(t) builds up at cue; this causes a pulse in δ; at reward delivery, r(t) is cancelled by drop in V(t).

• Reward Omitted: Drop in δ, due to drop in predicted reward V(t).

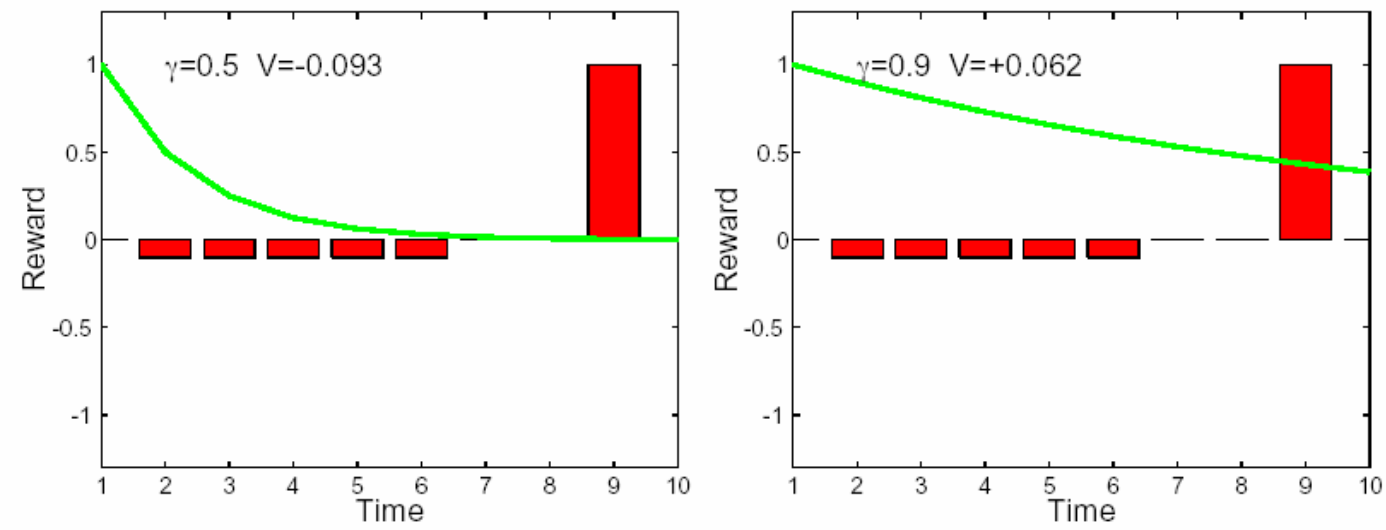


Reinforcement Learning Model of Basal Ganglia

Site	Role
Cortex	State s
Striatum	IP site of Basal Ganglia
Patch	V(s)
Striosome	Q(s,a)
SNc	δ
SNr, GP	Action Selection



Serotonin and Time Scale of Reward Prediction

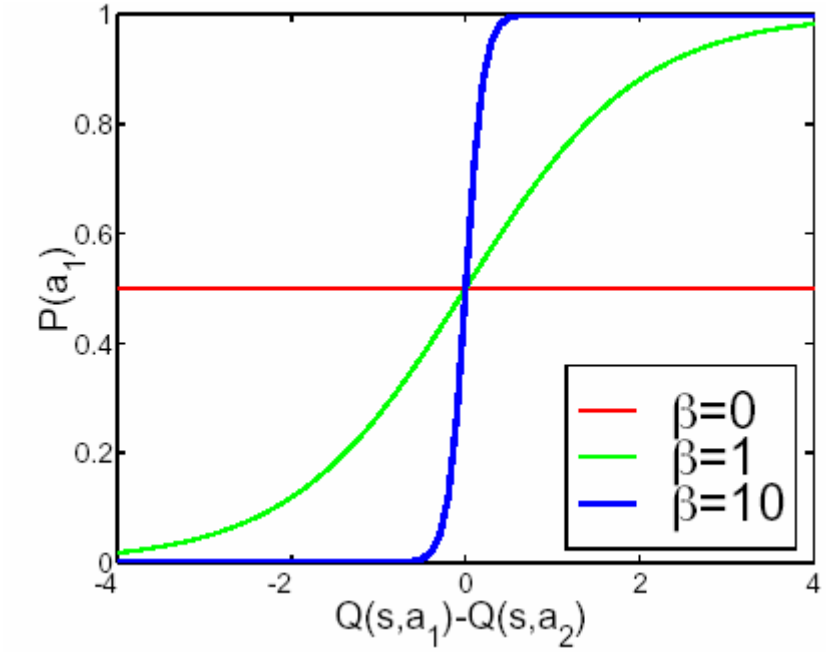


Real-life Strategies:

- Pay immediate cost (negative reward) to achieve future rewards.
- Avoid positive immediate reward if linked with large future negative reward.
- Rewards have to be acquired not too far in the future.

Noradrenaline and Randomness of Action Selection

- Lower value of β results is random action selection.
- Higher values make action selection more deterministic.
- In the attention task of monkeys, it is shown that the LC (Locus Coeruleus) neuron activity is closely correlated with the accuracy of action selection.
- Higher level of NA means higher β.



Acetylcholine and Memory Update

- ACh controls the learning rate α.

- Small α makes the learning slow, larger α can result in fast overwrite of learned information, too large α can make the learning unstable.

- ACh controls balance of storage and update of memory (Hasselmo et. al.).

- ACh modulates the synaptic plasticity in hippocampus, cortex and striatum.

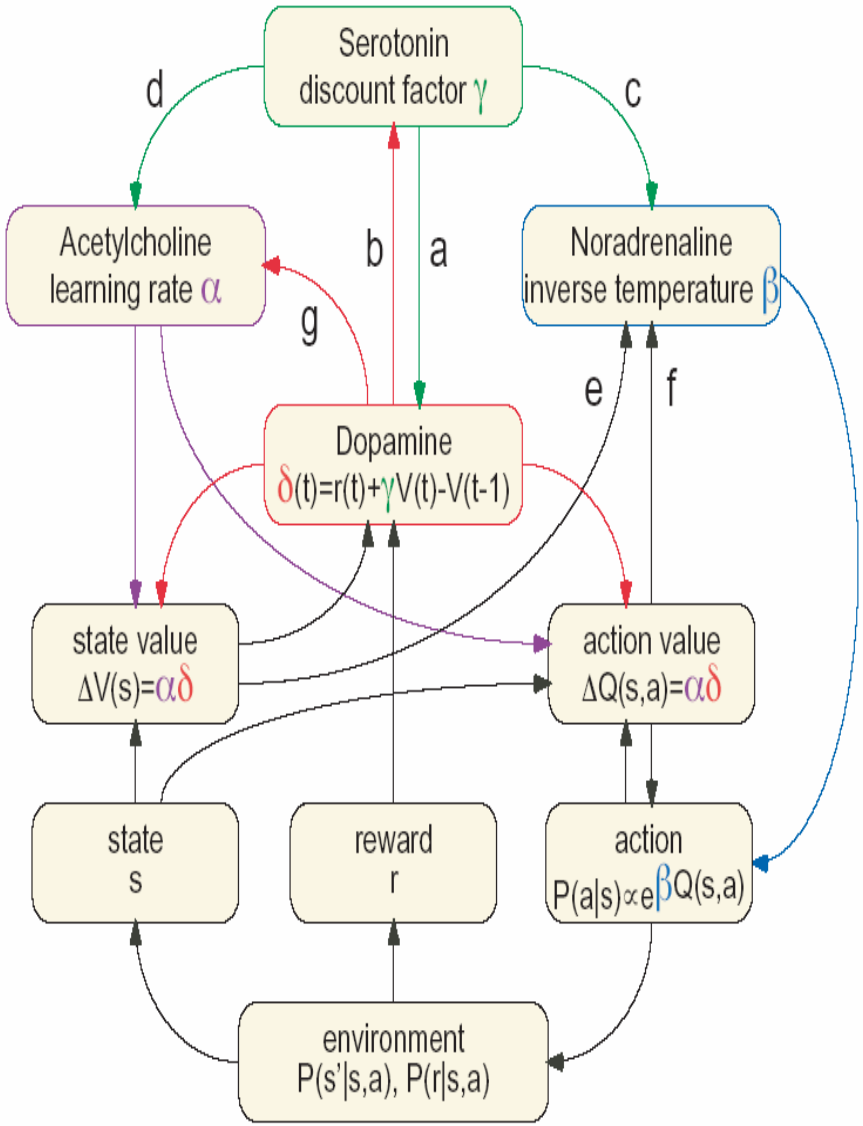
- Loss of cholinergic neurons is related to Alzheimer's disease.

- Memory storage for high Ach, retrieval for low Ach.

Part III – Dynamic Interactions of Neuromodulators

Interdependencies of Neuromodulators

- If V(t) is *positive*, 5-HT has *facilitatory* effect on DA and vice-versa (a).
- High variability of DA activity should have an inhibitory effect on 5-HT (b).
- Inhibitory effects of 5-HT on NA and ACh (c & d).
- β is reduced when the action value function has a high variance (f)
- Frequent change in the TD error (δ) will have an inhibitory effect on learning rate (α). (g)



Conclusions

- A unified theory to relate neuromodulators with metaparameters.
- A reinforcement learning model of the Basal Ganglia is presented.
- Interdependencies of neuromodulators are described.

References

• K. Doya, Metalearning and Neuromodulation, *Neural Networks*, 15, 495-506.