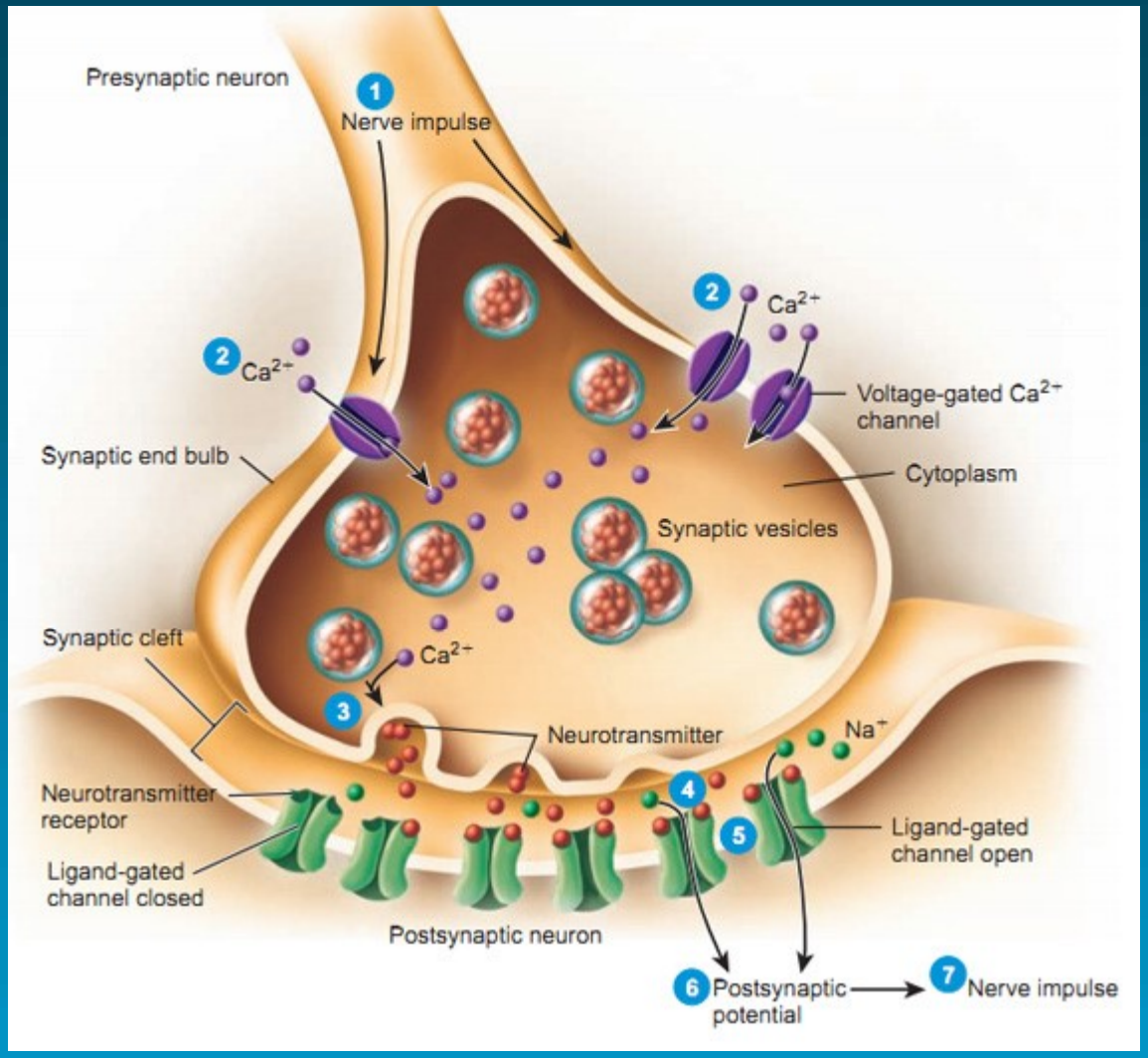
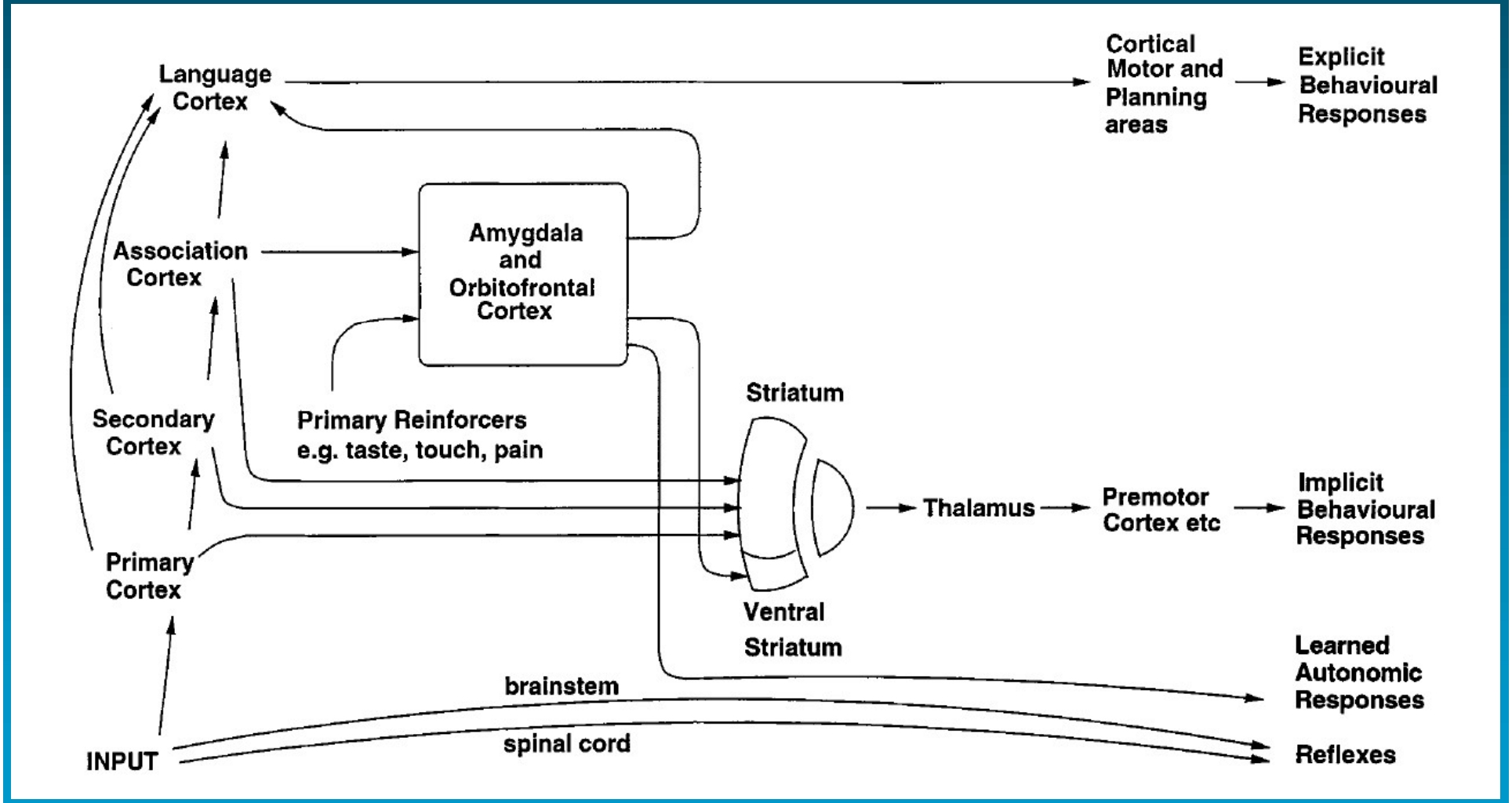
wor orld

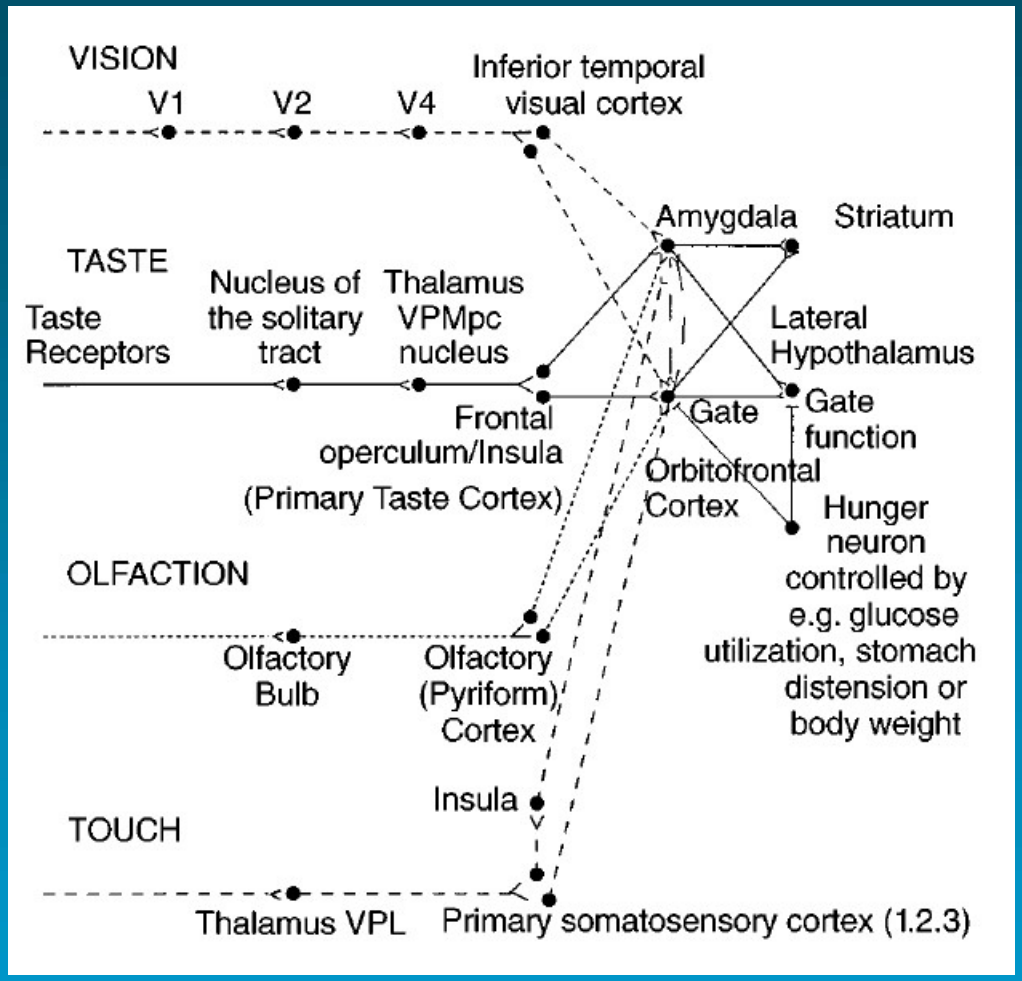


dimensional reduction

The Brain and Emotion, précis in Behavioral and Brain Sciences 23:177-234 (2000)

Emotive valence would be assigned in the amygdala and OFC




VISION
V1
V2
V4
Inferior temporal visual cortex
Amygdala
Striatum
TASTE
Taste Receptors
Nucleus of the solitary tract
Thalamus VPMpc nucleus
Lateral Hypothalamus
Gate
Gate function
Frontal operculum/Insula (Primary Taste Cortex)
Orbitofrontal Cortex
Hunger neuron controlled by e.g. glucose utilization, stomach distension or body weight
OLFACTION
Olfactory Bulb
Olfactory (Pyriform) Cortex
Insula
TOUCH
Thalamus VPL
Primary somatosensory cortex (1.2.3)

Stimulation of the amygdala produces a very high-dimensional constellation of behavioural responses…

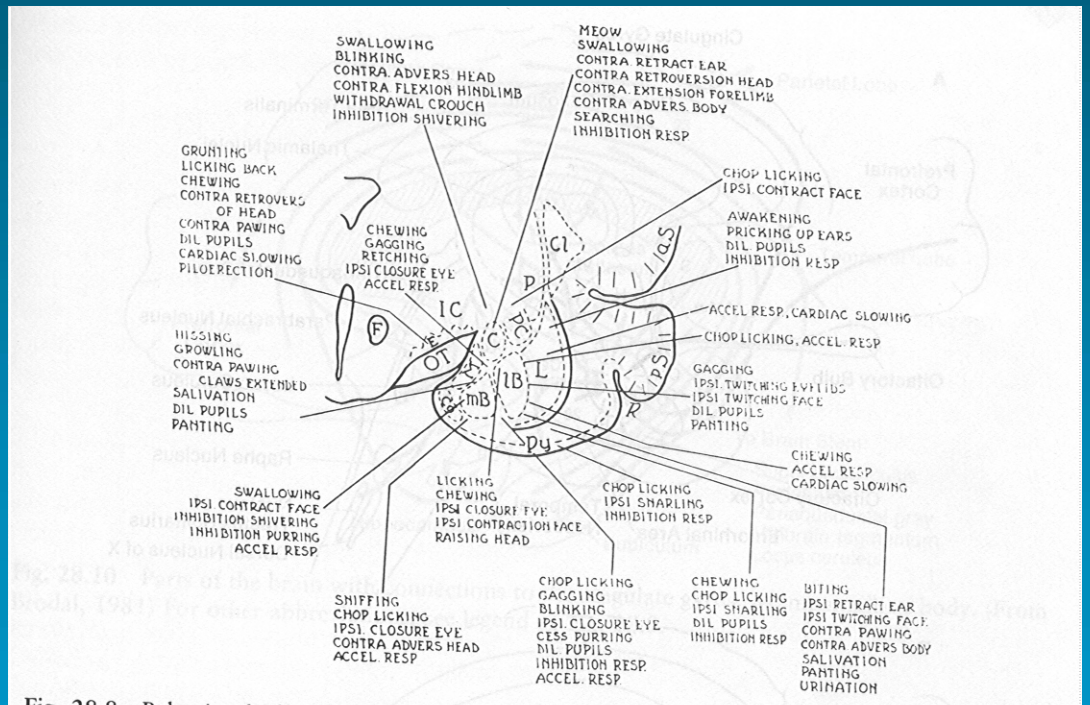
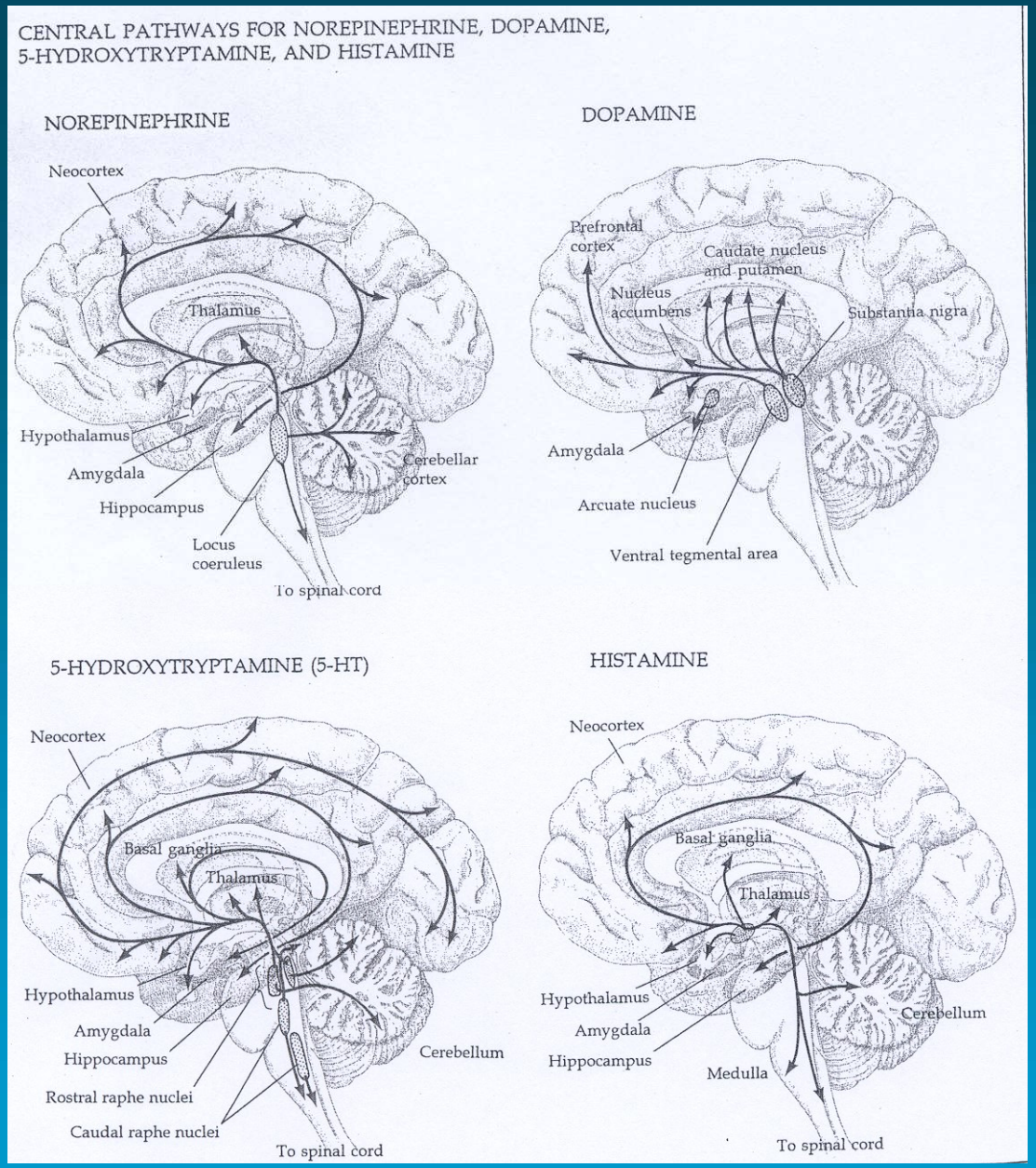


Fig. 28.8 Behavioral effects produced in the awake monkey by focal electrical stimulation at the sites shown. Abbreviations for nuclei of the amygdalar complex: C, central; PC, paracentral; M. medial; co, cortical; mB, mediobasal; lB, lateral basal; L, lateral. Other regions: IC, internal capsule; OT, optic tract; P, putamen; Cl, claustrum; aS, anterior sylvian; pS, posterior sylvian; R, rhinal sulcus; Py, pyriform cortex; F, fornix. (From MacLean and Delgado, 1953)

Chemical


CENTRAL PATHWAYS FOR NOREPINEPHRINE, DOPAMINE, 5-HYDROXYTRYPTAMINE, AND HISTAMINE
NOREPINEPHRINE
Neocortex
Thalamus
Hypothalamus
Amygdala
Hippocampus
Locus coeruleus
Cerebellar cortex
To spinal cord
DOPAMINE
Prefrontal cortex
Caudate nucleus and putamen
Nucleus accumbens
Substantia nigra
Amygdala
Arcuate nucleus
Ventral tegmental area
5-HYDROXYTRYPTAMINE (5-HT)
Neocortex
Basal ganglia
Thalamus
Hypothalamus
Amygdala
Hippocampus
Rostral raphe nuclei
Caudal raphe nuclei
Cerebellum
To spinal cord
HISTAMINE
Neocortex
Basal ganglia
Thalamus
Hypothalamus
Amygdala
Hippocampus
Medulla
Cerebellum
To spinal cord

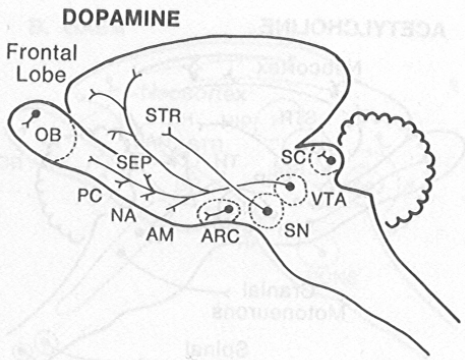


Fig. 24.10 Distribution of dopamine-containing neurons in the rat brain. For abbreviations, see legend to Fig. 24.9.

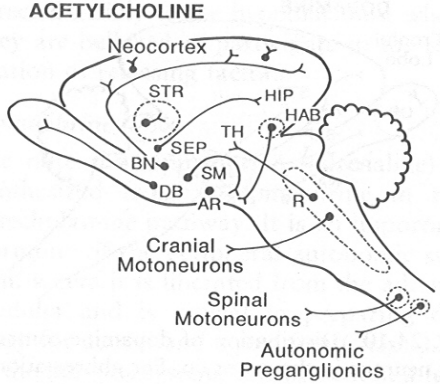
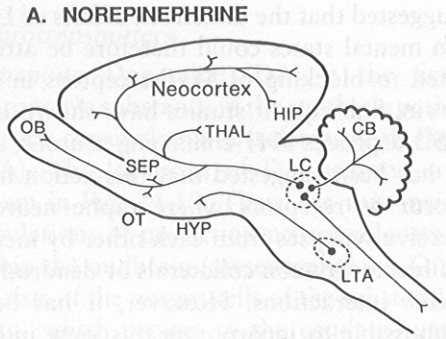


Fig. 24.11 Distribution of cholinergic cell groups and their projections in the rat brain. For abbreviations, see legend to Fig. 24.9.



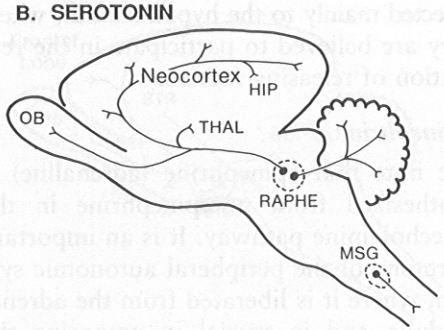


Fig. 24.9 Maps of the distribution of cell groups containing different neurotransmitters in the mammalian brain. A sagittal view of the rat brain is shown for this and succeeding figures. **A.** Distribution of norepinephrine-containing neurons and their axonal projections. **B.** Distribution of serotonin-containing neurons and their projections. These are discussed in the text under the category of central state circuits. Abbreviations for this and the following maps: AM, amygdala; AR, arcuate nucleus; ARC, arcuate nucleus; BN, basal nucleus; DB, diagonal band; DCN, deep cerebellar nuclei; DH, dorsal horn, DRG, dorsal root ganglion; EPN, endopeduncular nucleus; GP, globus pallidus; HAB, habenula; HIP, hippocampus; HYP, hypothalamus; LC, locus ceruleus; LTA, lateral tegmental area; MED, medulla; MSG, medullary serotonin group; NA, nucleus accumbens; OB, olfactory bulb; OT, olfactory tubercle; PC, pyriform cortex; PERI-V., periventricular gray; R, reticular nucleus; SC, superior colliculus; SEP, septum; SM, stria medullaris; SN, substantia nigra; STR, striatum; TH or THAL, thalamus; VTA, ventral tegmental area.

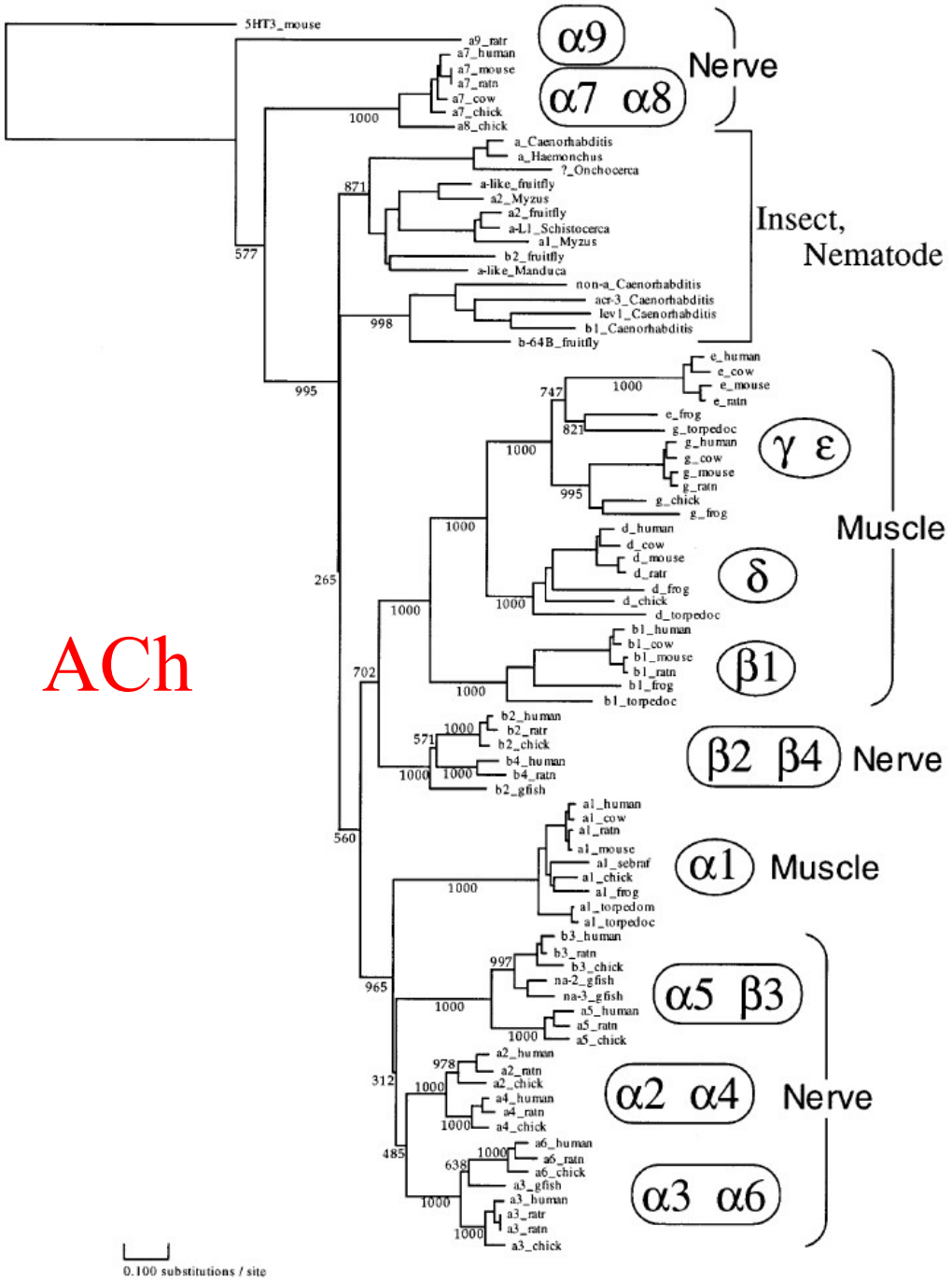


FIG. 1.—The phylogenetic tree for the nAChR subunits. The bootstrap values are indicated at the corresponding nodes.

ACh

Very old stuff

Bull Acad Natl Med. 1998;182(7):1505-14; discussion 1515-6.

[Evolution of monoamine receptors and the origin of motivational and emotional systems in vertebrates]

[Article in French]

Vincent JD, Cardinaud B, Vernier P.

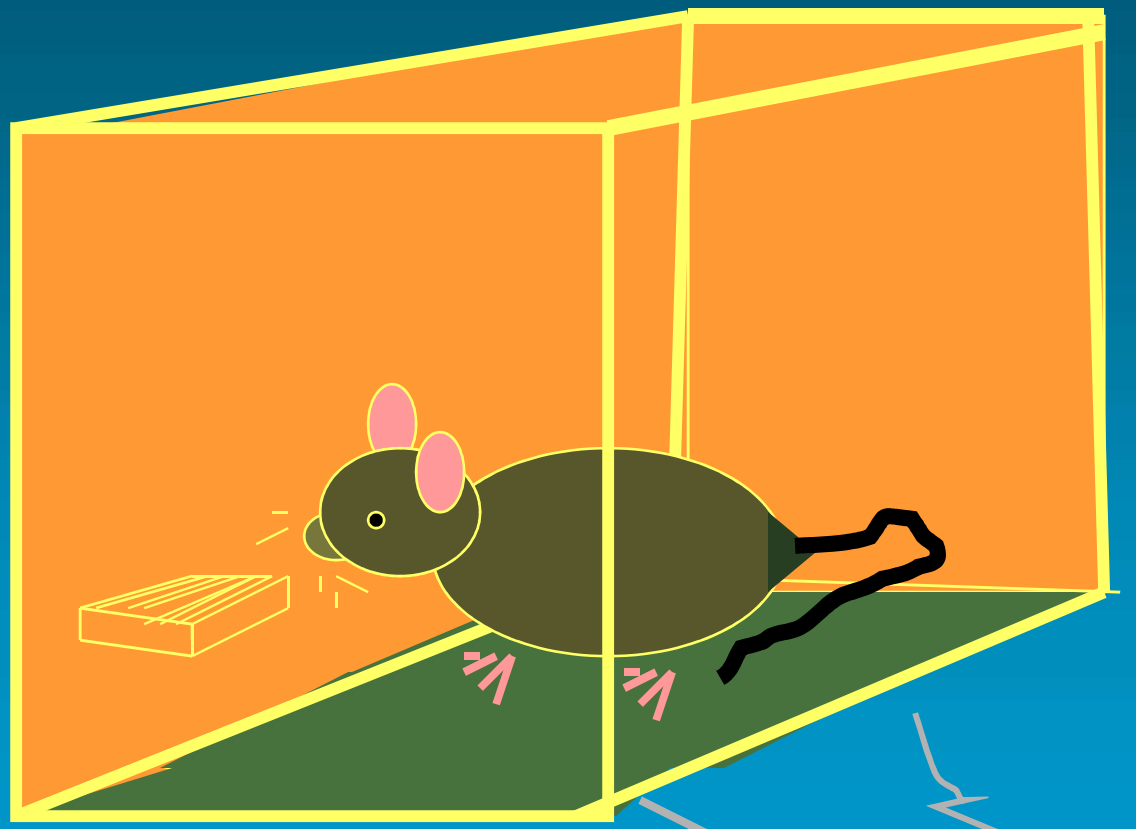
IMPC, CNRS, Valbonne.

MA → DA + NA

The evolving vertebrate nervous system was accompanied by major gene duplication events generating novel organs and a sympathetic system. Vertebrate neural pathways synthesizing catecholamine neurotransmitters (dopamine and noradrenaline), were subsequently recruited to process increased information demands by mediating psychomotor functions such as selective attention/predictive reward and emotional drive via the activation of multiple G-protein linked catecholamine receptor subtypes. Here we show that the evolution of these receptor-mediated events were similarly driven by forces of gene duplication, at the cephalochordate/vertebrate transition. In the cephalochordate Amphioxus, a sister group to vertebrates, a single catecholamine receptor gene was found, which based on molecular phylogeny and functional analysis formed a monophyletic group with both vertebrate dopamine D1 and beta adrenergic receptor classes. In addition, the presence of dopamine but not of noradrenaline was assayed in Amphioxus. In contrast, two distinct genes homologous to jawed vertebrate dopamine D1 and beta adrenergic receptor genes were extant in representatives of the earliest craniates, lamprey and hagfish, paralleling high dopamine and noradrenaline content throughout the brain. These data suggest that a D1/beta receptor gene duplication was required to elaborate novel catecholamine psychomotor adaptive responses and that a noradrenergic system specifically emerged at the origin of vertebrate evolution.

Reinforcement Learning (RL): Psychology and animal behavior literature

- B. F. Skinner, 1938, The Behavior of Organisms, New York: D. Appleton-Century Publishers.
- Reward strengthens likelihood of animal response.
- Rats are better at learning, e.g., mazes, when they receive a reward.



Skinner box

Reinforcement Learning (RL) *in Artificial Intelligence*

- The ***reinforcement learning (RL) problem*** is the problem faced by an agent that learns behavior through trial-and-error interactions with its environment. It consists of an agent that exists in an environment described by a set S of possible states, a set A of possible actions, and a ***reward*** (or punishment) r_t that the agent receives each time t after it takes an action in a state. (Alternatively, the reward might not occur until after a ***sequence*** of actions have been taken.)
  - It is typically assumed that the environment is non-deterministic.
  - Agent evaluation (in terms of rewards) may be interleaved with learning.

- **The objective of an RL agent is to maximize its cumulative reward received over its lifetime.**

A Few Definitions

- ***(time) step*** – the agent is in a state, s_t, takes action a, and that moves the agent to a next state, s_{t+1}. After getting to s_{t+1} , the agent receives a reward, r_t.
- ***trial*** – this is the RL term used for an ***episode***. A trial consists of a sequence of steps that terminates when either:
  - **the agent enters a terminal/goal state, or**
  - **a predetermined time limit (number of steps) has been reached.**
- ***terminal (or absorbing) state*** – a state from which the agent does not leave, and which includes a final reward or punishment. A ***goal state*** is an example of a terminal state.

Markov Decision Processes (MDPs)

- **Unless stated otherwise, we will assume this is a *Markov decision process* (MDP). There are two functions, the transition function, $\delta(s_t, a_t) = s_{t+1}$ which defines the next state given the current state and action, and the reward function, $r(s_t, a_t)$ which provides a reward for taking this action in this state (or, alternatively, $r(s_t)$). *In an MDP, δ and r depend only on the current state and action, not on earlier states or actions.* If the environment is non-deterministic, then you also want to know $p(s_t, a_t, s_{t+1})$, i.e., the probability of going from s_t to s_{t+1} by taking action a_t. All of this information combined is called a *model of the environment*.**
  - **The model may or may not be known to the agent.**
  - **The model may or may not be learned by the agent. The latter case is called *model-free* reinforcement learning. This is the type of RL that we will study.**

Discount Factor

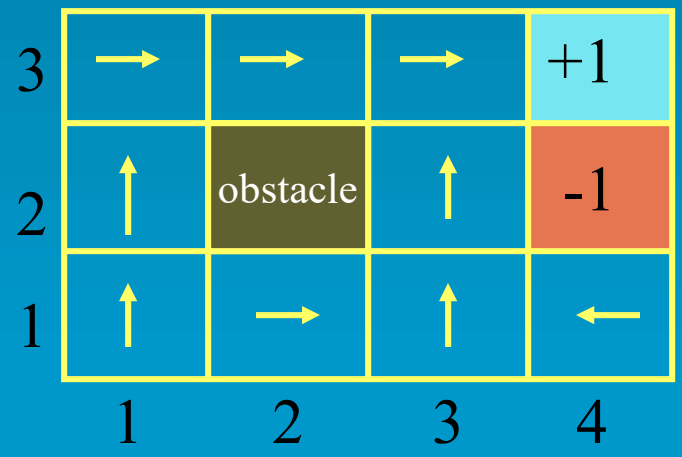
For agents with a very long (modeled as infinite) lifetime, a discount factor is useful. Future rewards are discounted.

1. A discount factor γ makes future rewards less valuable than current rewards.
2. It ensures that the total reward will converge to a finite, reasonable amount.

A "Policy"

A *policy* is a *complete* mapping from *every* state to the action to be taken in that state.

In a gridworld, we can consider a square to be a *state*.



Objective of RL

- **The objective of reinforcement learning (RL) is to try to find an *optimal policy*. A *policy*, Π: S→A , is a *complete* mapping from every state to the action to be taken in that state.**
  - **For simple problems, a policy (also called a *control strategy*) may be implemented as a lookup table.**
  - ***An optimal policy is one that leads to optimal behavior for solving the problem, i.e., it is the policy that results in the highest cumulative reward over time.* In other words, define the *discounted cumulative reward* achieved by policy Π from initial state s_t as:**

Value of a policy

$$V^{\Pi}(s_t) \equiv r_t + \gamma r_{t+1} + \gamma^2 r_{t+2} + \cdots = \sum_{i=0}^{\infty} \gamma^i r_{t+i}$$

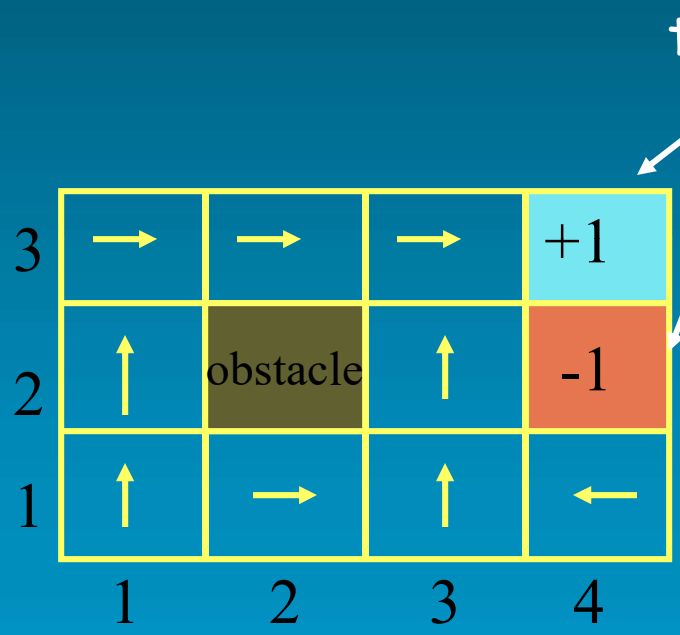
Policy Π is followed always; $0 <= \gamma < 1$ is a discount factor

- **Then an optimal policy is one that maximizes the discounted cumulative reward, and is defined as:**

$$\Pi^* \equiv \arg\max_{\Pi} V^{\Pi}(s), \forall s$$

V*(s) is the *maximum discounted cumulative reward*, which is obtained by starting in state s and following Π*.

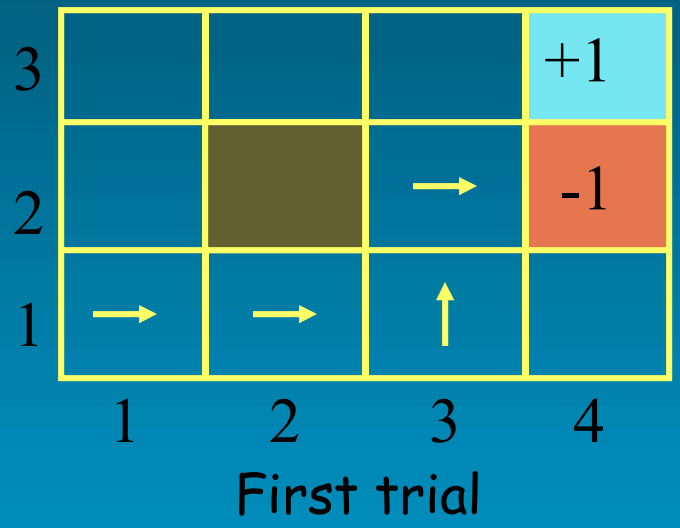
An Example of an Optimal Policy



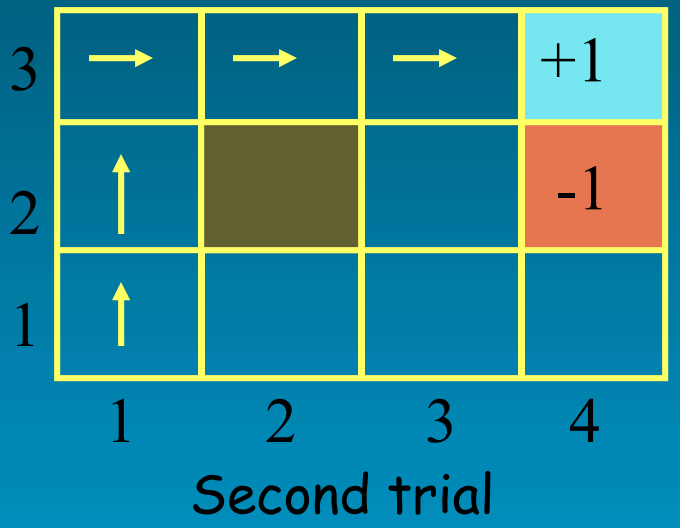
Assumes reward is –0.04 in all non-terminal states.
Rewards for terminal states (4,3) and (4,2) are shown.
Assumes no discounting.

Note: ***There may be more than one optimal policy.***
Can you think of another optimal policy here?

An Example of Trials While Learning an Optimal Policy



First trial



Second trial

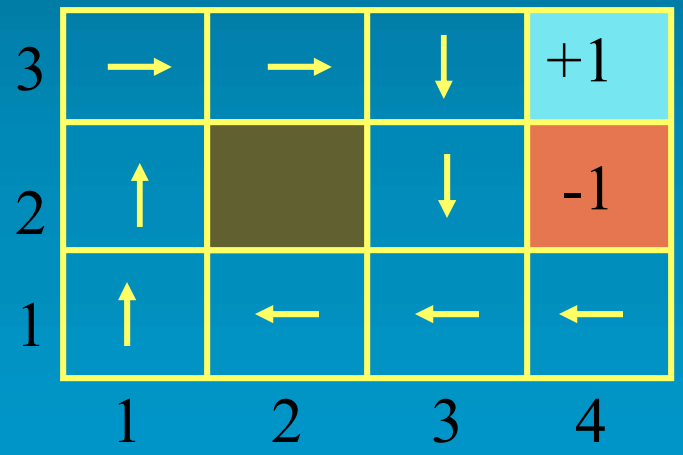
Example trials on the way to learning an optimal policy:

$(1,1)_{-0.04} \rightarrow (2,1)_{-0.04} \rightarrow (3,1)_{-0.04} \rightarrow (3,2)_{-0.04} \rightarrow (4,2)_{-1}$ First trial

$(1,1)_{-0.04} \rightarrow (1,2)_{-0.04} \rightarrow (1,3)_{-0.04} \rightarrow (2,3)_{-004} \rightarrow (3,3)_{-0.04} \rightarrow (4,3)_{+1}$ Second trial

Maximum Trial Length

Typically one sets a maximum number of steps per trial. The following policy gives an example why:



Comment on Setting the Rewards

- The choice of rewards you give the agent can determine how quickly it will learn. For example,
  - **If you give a reward of 0.99 for every state that leads directly to the goal, and a reward of 0 for every other state, then you are giving a great deal of prior knowledge to your agent, and it can learn very fast because little learning is required. In essence, you are *teaching* the agent how to get to the goal by carefully selecting your rewards.**
  - **If you give relatively equal rewards (e.g., close to 0) from all states other than the terminal states, it will take the agent a long time to learn. The previous two slides give an example of this.**
- For your projects, you probably want to do something in the middle of these two extremes.

Two Popular Reinforcement Learning Algorithms

Temporal Difference (TD) Learning
(Sutton, 1984)

- The objective is to ***learn*** an estimate of the ***utility*** of all states. The utility is the expected sum of rewards from this state on, i.e., it is a measure (really, an estimate) of V*(s).
- Once the agent has learned an estimated utility for each state, it can use this utility for deciding which action to take next – ***it will choose the action that leads to the next state with the highest utility.***

Temporal Difference Learning

- The objective is to learn an estimate of the ***utility*** of all states. The utility is the expected sum of rewards from this state on.
- **Key idea**: Use insight from ***dynamic programming*** to adjust the utility of a state based on the immediate reward and the utility of the next state.
  - $U(s) \leftarrow U(s) + \alpha(r(s) + \gamma U(s') - U(s))$

the observed successor state

learning rate

reward obtained in state i

Essence: (1- α)*(old) + α*(new)

U(s) is an estimate of V(s), which is the maximum discounted cumulative reward starting in state s.*

A Simple TD Learning Algorithm

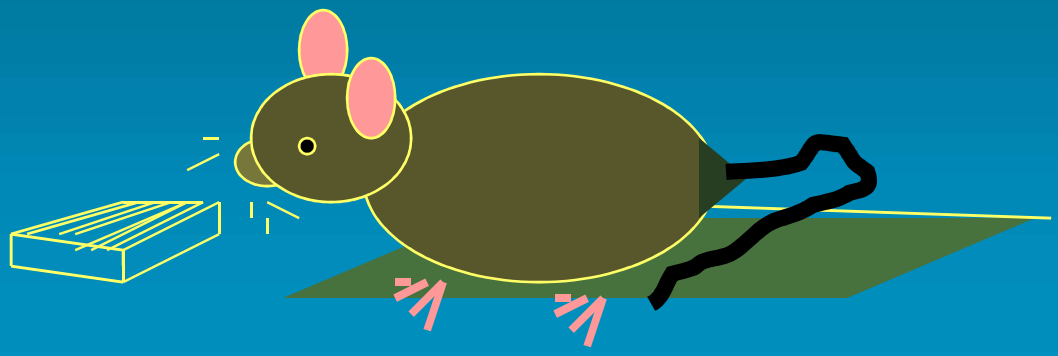
- Initialize $U(s) = 0$ for all non-terminal states s. For terminal states, $U(s) = r(s)$. Start in a designated initial state s_0. (We assume all other states are reachable from s_0.)
- For each transition $\delta(s, a) = s'$ and reward $r(s)$ for going from state s to state s', do:
  - $U(s) \leftarrow U(s) + \alpha(r(s) + U(s') - U(s))$
- Repeat above step until the difference in successive values (before/after update) of U is less than or equal to some small desired ε (called ***convergence***).

Active learning in an unknown environment

- The TD learner just described is a ***passive learner***, i.e., a learner that observes the state and reward sequences and estimates the expected sum of rewards in all non-terminal states that it visits. After learning the utilities, actions can be chosen based on those utilities.
- An ***active learner*** must consider what actions to take, what their outcomes may be, and how they affect the rewards achieved. An active learner takes actions while it learns. Only an active learner can handle a dynamic environment.

Active learning

s_1, a_1, r_1, learn, s_2, a_2, r_2, learn, ...



Here, the reward is a function of the state ***and*** action, i.e., $r_i (s_i, a_i)$.

Interaction between world and an active learning robot

- World: you are in state 34. You have three possible actions from this state.
- Robot: I'll take action 2.
- World: you are in state 77. Your immediate reward is –7. You have two possible actions from this state.
- Robot: I'll take action 1.
- World: you are in state 34. Your immediate reward is 3. You have three possible actions from this state.
- ……

An Active Learning Algorithm: Q-Learning (Watkins, 1992)

Objective of Q-Learning

Let **Q*(s, a)** be the ***maximum, discounted, cumulative reward for taking action a in state s, and then continuing to choose actions optimally*** **(**according to Π*). This is analogous to V*(s), the maximum discounted cumulative reward**,** which is obtained by starting in state s and following Π*(s).

Note that:

$$V^*(s) = \max_{\hat{a}} Q^*(s, \hat{a})$$

Note that â is a variable here

Assume δ(s, a) = s'. Then Q*(s, a) can be defined recursively as:

$$Q^*(s,a) = r(s,a) + \gamma(V^*(\delta(s,a))) = r(s,a) + \gamma(V^*(s')) = r(s,a) + \gamma(Q^*(s',a'))$$

where $a' = \arg\max_{\hat{a}} Q^*(s', \hat{a})$

<u>Objective:</u> **Learn Q(s, a), which estimates Q*(s, a).**

Q-Learning Update Formula

- Learn an action value function Q mapping state-action pairs to the expected utility of the sequence starting with that state/action pair. There is no need to learn the functions δ(s, a) or r(s, a), or p(s, a, s ’), i.e., the model of the environment.
- The procedure **UPDATE-Q-VALUE(s, a)** is:

$$Q(s,a) \leftarrow (1-\alpha)Q(s,a) + \alpha(r(s,a) + \gamma \max_{\hat{a}} Q(s',\hat{a}))$$

Recall from previous slide: $Q^*(s,a) = r(s,a) + \gamma(Q^*(s',a'))$ $\quad a' = \arg\max_{\hat{a}} Q^*(s',\hat{a})$

- The Q-value is related to the utility value U by:

$$U(s) = \max_{\hat{a}} Q(s,\hat{a})$$

As for the Neural Implementation Peter Dayan notes that…

RL needs signal that

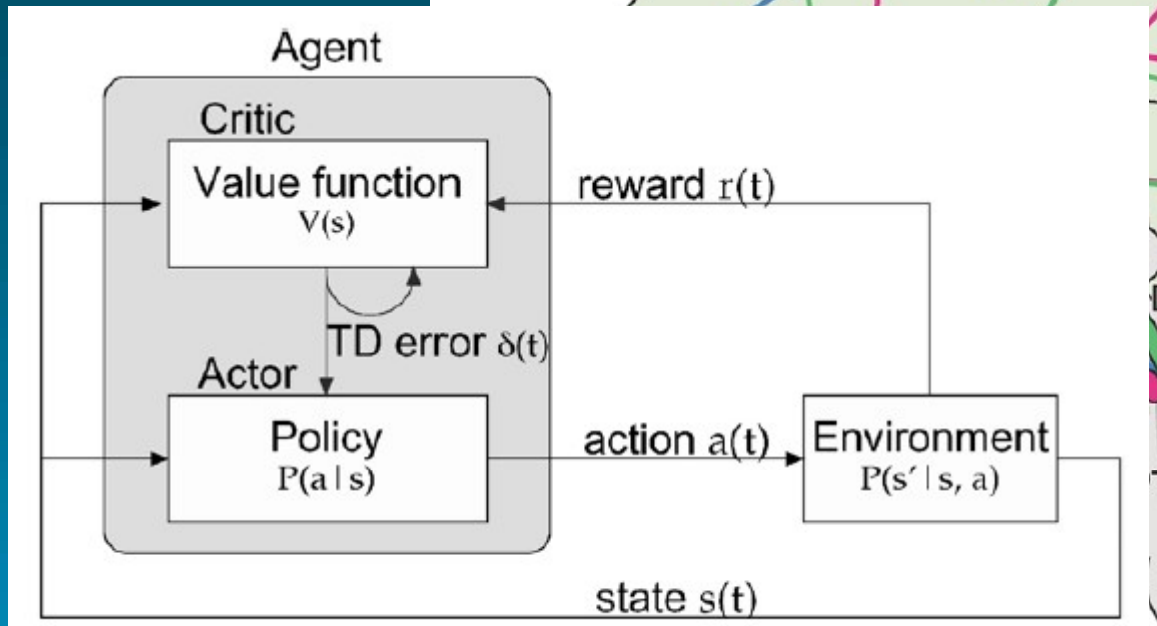
- Respond to affective contingencies
- Affect learning of predictions and actions
- Are essentially scalar
- Broadcast their information multimodally

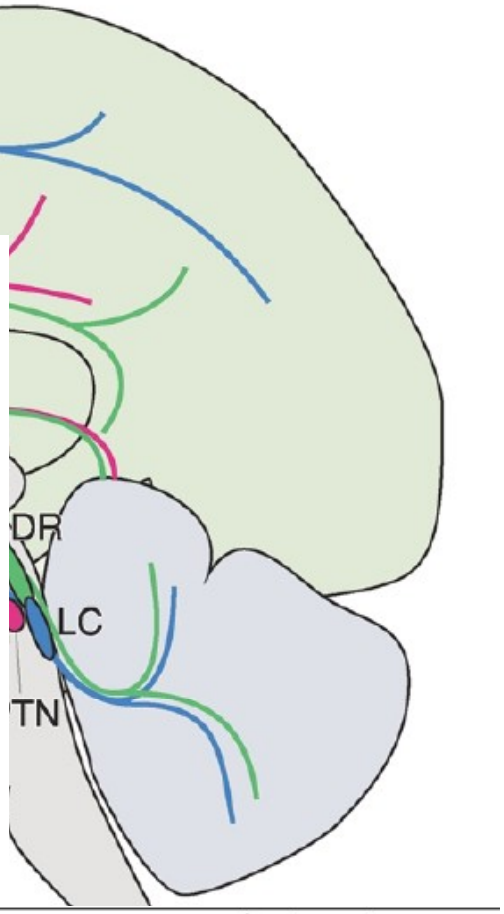
Neuromodulators in fact

- Respond to reinforcers and surprise
- Are known to affect synaptic plasticity
- Come from small mid-brain nuclei
- Have extensive arborization throughout the brain

Kenji Doya

Neural Networks 15 (2002) 495–506





neuromodulator	origin of projection	major target area
	medial septum (3)	hippocampus
	pedunculopontine tegmental nucleus (PPTN)	SNc, thalamus
		superior colliculus

1. Dopamine signals the TD error δ.
2. Serotonin controls the discount factor γ.
3. Noradrenaline controls the inverse temperature β.
4. Acetylcholine controls the learning rate α.

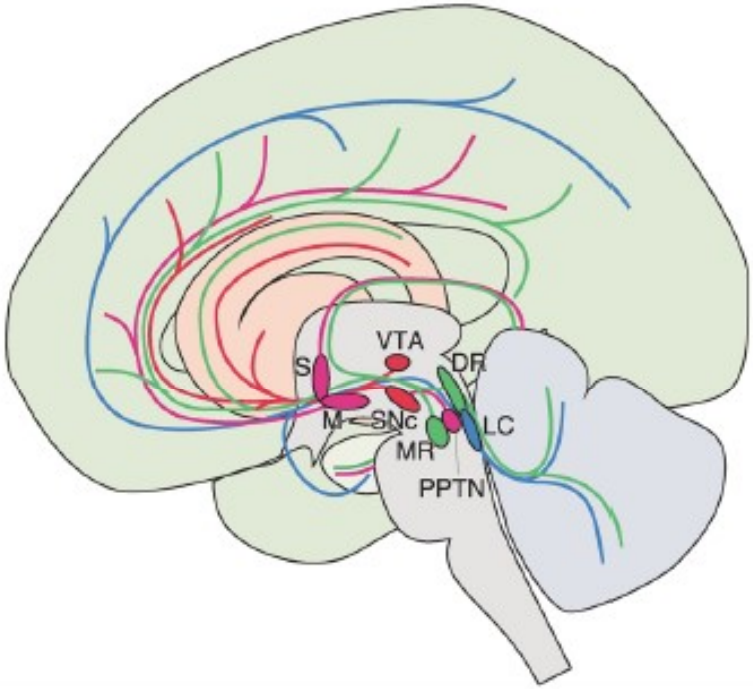
Neural Networks 15 (2002) 495–506

2002 Special issue

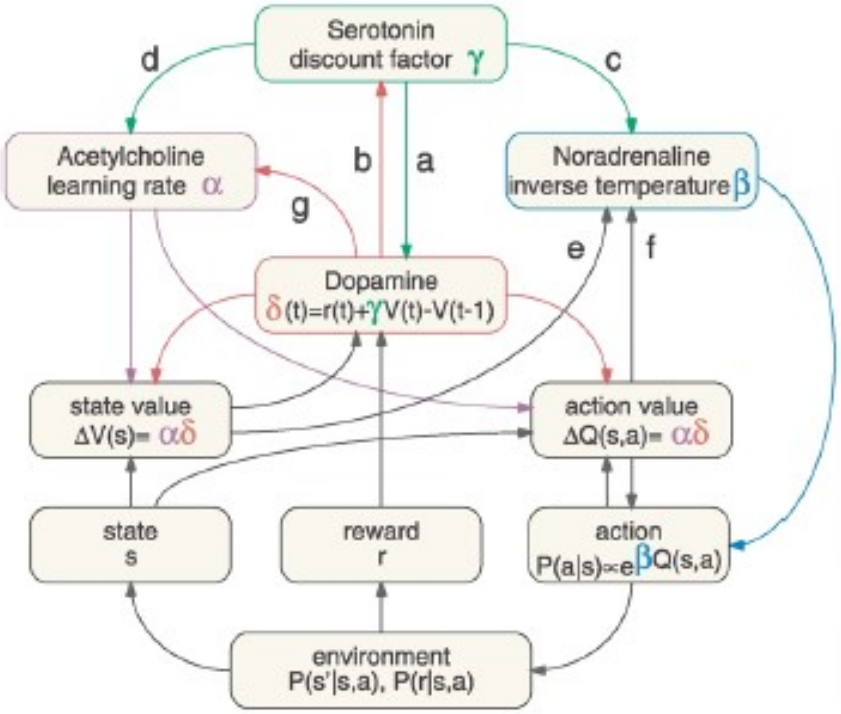
Metalearning and neuromodulation

Kenji Doya*

1. Dopamine represents the global learning signal for prediction of rewards and reinforcement of actions.
2. Serotonin controls the balance between short-term and long-term prediction of reward.
3. Noradrenaline controls the balance between wide exploration and focused execution.
4. Acetylcholine controls the balance between memory storage and renewal.



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
		cerebellum
acetylchol. ACh	Meynert nucleus (M)	cortex, amygdala
	medial septum (S)	hippocampus
	pedunculopontine tegmental nucleus (PPTN)	SNc, thalamus
		superior colliculus



1. Dopamine signals the TD error δ.
2. Serotonin controls the discount factor γ.
3. Noradrenaline controls the inverse temperature β.
4. Acetylcholine controls the learning rate α.

Temporal difference models describe higher-order learning in humans

Ben Seymour[1], John P. O'Doherty[1], Peter Dayan[2], Martin Koltzenburg[3], Anthony K. Jones[4], Raymond J. Dolan[1], Karl J. Friston[1] & Richard S. Frackowiak[1,5]

NATURE | VOL 429 | 10 JUNE 2004 | www.nature.com/nature

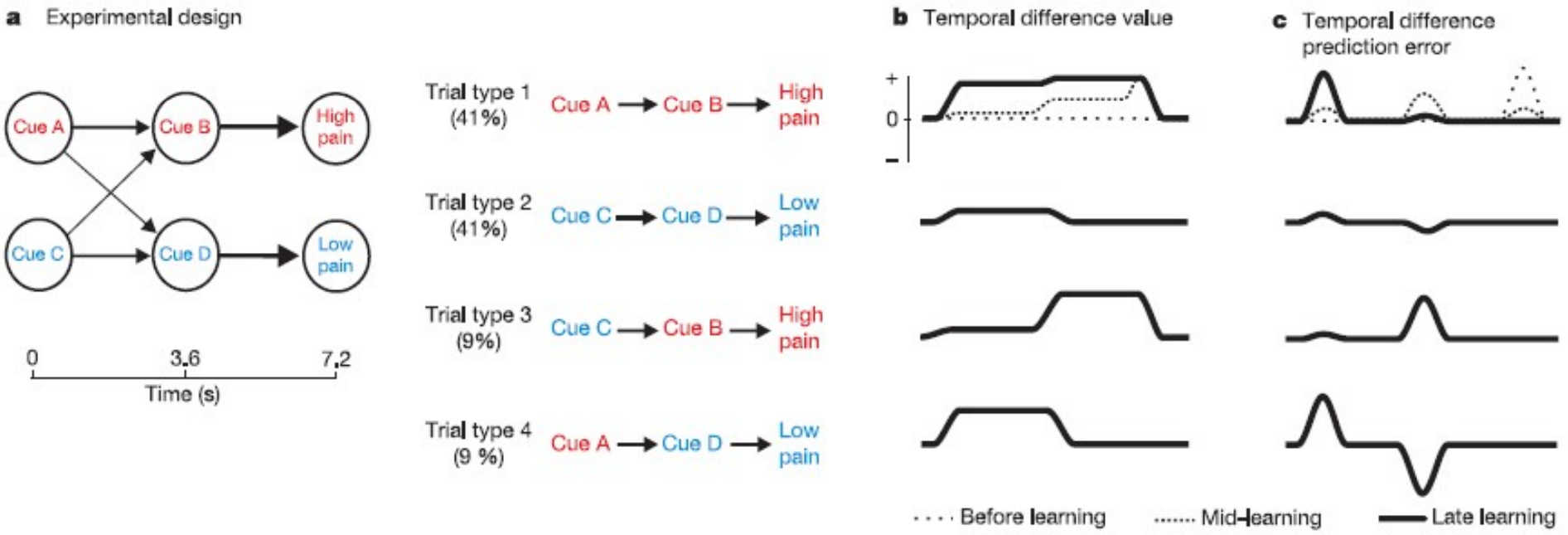


Figure 1 Experimental design and temporal difference model. **a**, The experimental design expressed as a Markov chain, giving four separate trial types. **b**, Temporal difference value. As learning proceeds, earlier cues learn to make accurate value predictions (that is, weighted averages of the final expected pain). **c**, Temporal difference prediction error; during learning the prediction error is transferred to earlier cues as they acquire the ability to make predictions. In trial types 3 and 4, the substantial change in prediction elicits a large positive or negative prediction error. (For clarity, before and mid-learning are shown only for trial type 1.)

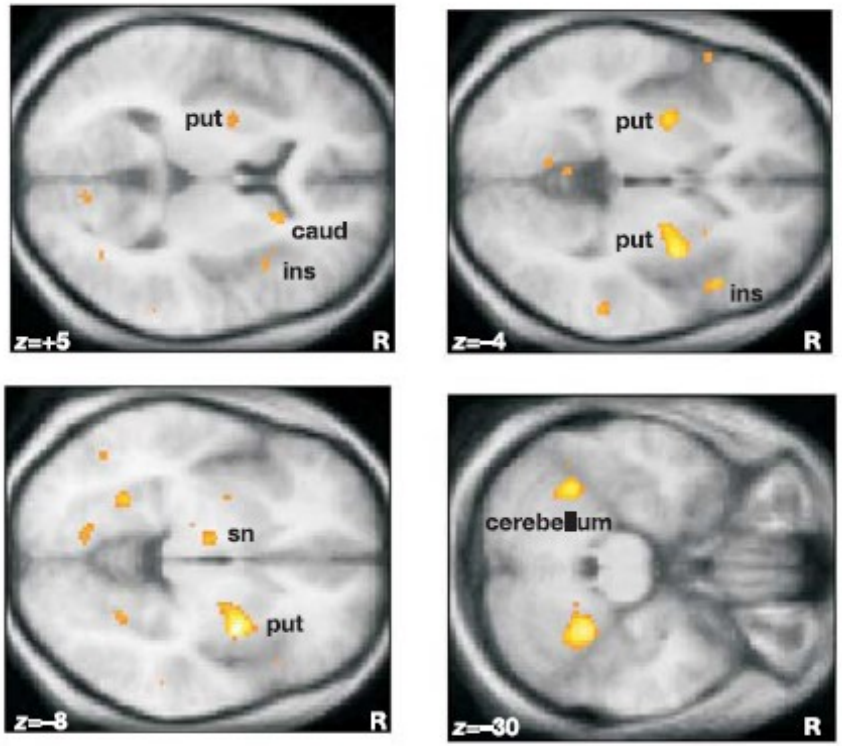


Figure 2 Temporal difference prediction error (statistical parametric maps). Areas coloured yellow/orange show significant correlation with the temporal difference prediction error. Yellow represents the greatest correlation. Peak activations (MNI coordinates and statistical *z* scores) are: right ventral putamen (put; (32, 0, −8), $z = 5.38$); left ventral putamen (put; (−30, −2, −4), $z = 3.93$); right head of caudate (caud; (18, 20, 6), $z = 3.75$); left substantia nigra (sn; (−10, −10, −8), $z = 3.52$); right anterior insula (ins; (46, 22, −4), $z = 3.71$); right cerebellum ((28, −46, −30), $z = 4.91$); and left cerebellum ((−34, −52, −28), $z = 4.42$). R indicates the right side.

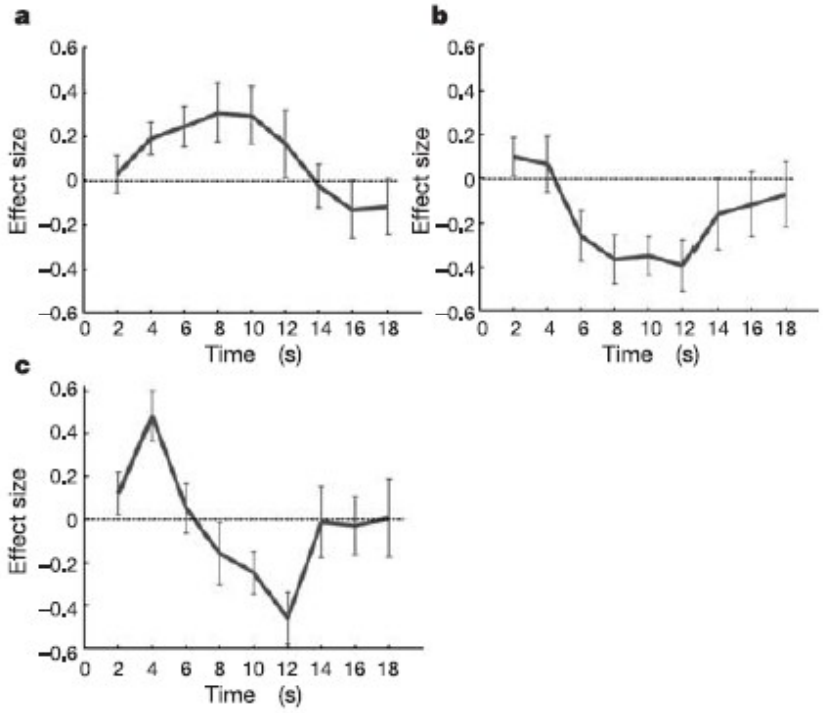


Figure 3 Temporal difference prediction error (impulse responses). Time course of the impulse response (±s.e.m.) to higher-order prediction error in the right ventral putamen. **a**, Positive prediction error (contrast of trial types 3 and 2). **b**, Negative prediction error (contrast of trial types 4 and 1). **c**, Biphasic prediction error; positive at the first cue, becoming negative at the second (contrast of trial types 4 and 2).

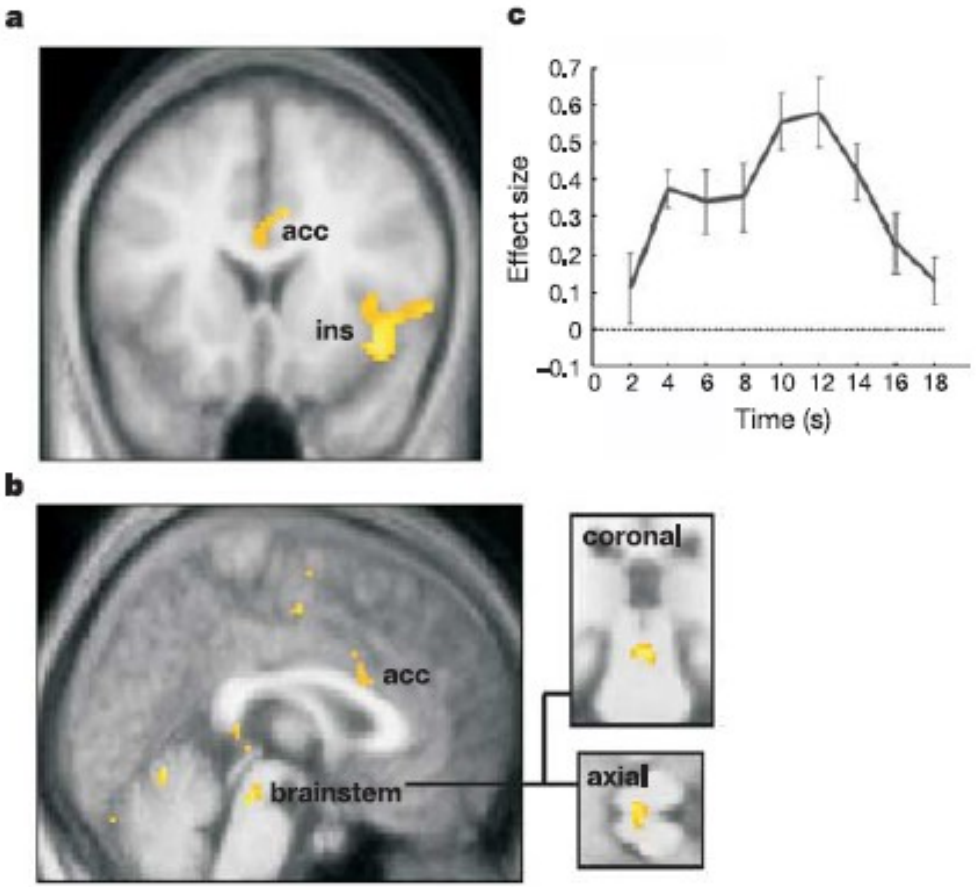


Figure 4 Temporal difference value (statistical parametric maps and impulse response in the right anterior insula). **a**, **b**, Areas showing significant correlation with the temporal difference value. Peak activations (MNI coordinates and statistical *z* scores) are: right anterior insula (ins; (42, 16, −14), $z = 4.16$); brainstem ((0, −28, −18), $z = 3.89$); and anterior cingulate cortex (acc; (8, 12, 32), $z = 3.82$). Coronal and axial slices of brainstem activation are shown, highlighting localization to dorsal raphe nucleus. **c**, Time course of impulse response (±s.e.m.) in right anterior insula cortex, from contrast of trial types 1 and 2.