Plasticity

- Metaplasticity
  - Stability vs. change
  - Cascade model of synaptic plasticity
- Reinforcement learning

Metaplasticity

- The rules by which the plastic processes change can themselves change in time
- This change is not necessarily expressed as a modification in the efficacy of normal synaptic transmission
- It manifests as a change in the ability to induce subsequent synaptic plasticity such as long-term potentiation or depression

  (Abraham & Bear, 1996)

Metaplasticity

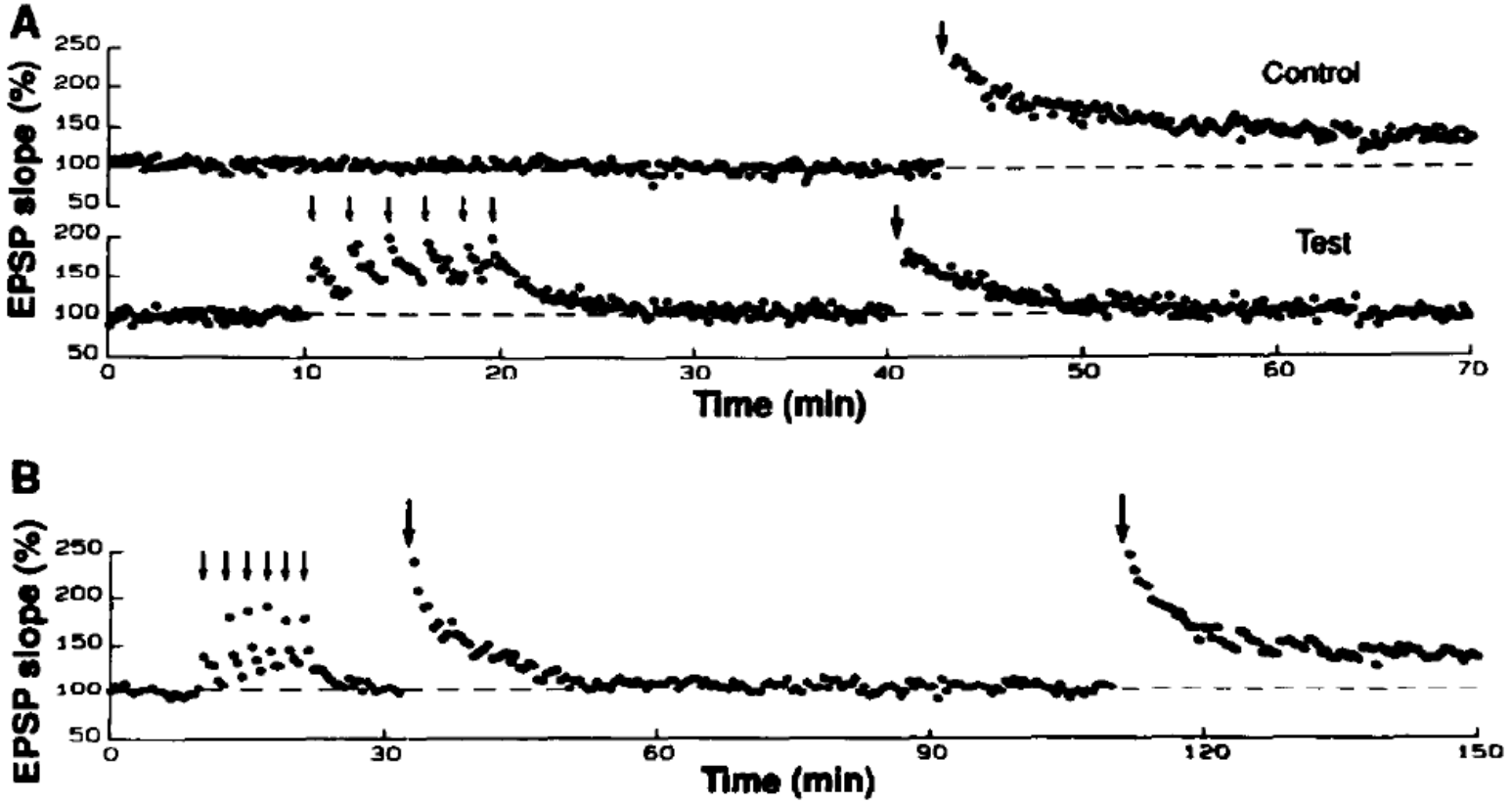


Fig. 1. *Effect of prior stimulation on long-term potentiation (LTP) in area CA1 of the hippocampus.* (A) *The population excitatory postsynaptic potential (EPSP) was recorded and two separate pathways (control and test) were stimulated alternately. At the times indicated by the small downward arrows, weak tetani (30 Hz, 0.15 s) were delivered to the test pathway. Although this stimulation did not produce a lasting change in synaptic effectiveness, it did inhibit induction of LTP by a strong tetanus delivered 20 minutes later (indicated by the large downward arrow). LTP on the control path was unaffected.* (B) *The inhibition of LTP caused by prior stimulation was transient, lasting no more than about an hour.* Figure adapted, with permission, from Ref. 2.

Metaplasticity

- Long term plasticity has been blocked by the weak tetanic pulses
- The thresholds for plasticity are dynamic
- Similar phenomenon has been observed for LTD (Cristie & Abraham, 1992)
- Biochemical factors: neuromodulators (cathecolamins) and hormones affect synaptic plasticity

Metaplasticity

- These factors could regulate the way plasticity changes in different stages
  - Wakefulness → synaptic potentiation
  - Slow-wave activity → synaptic depression
  - Sleep → synaptic downscaling (?)
- But there must be intrinsic synaptic processes

Metaplasticity

- One way of thinking about this process is that synapses are not in a continuum of states but that they are discrete (Montgomery & Madison, 1992)

- They showed for instance that in CA3 pyramidal cells, cell pairs connected by all silent synaptic connections cannot be depressed immediately following their unsilencing

- There are “hidden” variables that define the state of the synapse

Metaplasticity

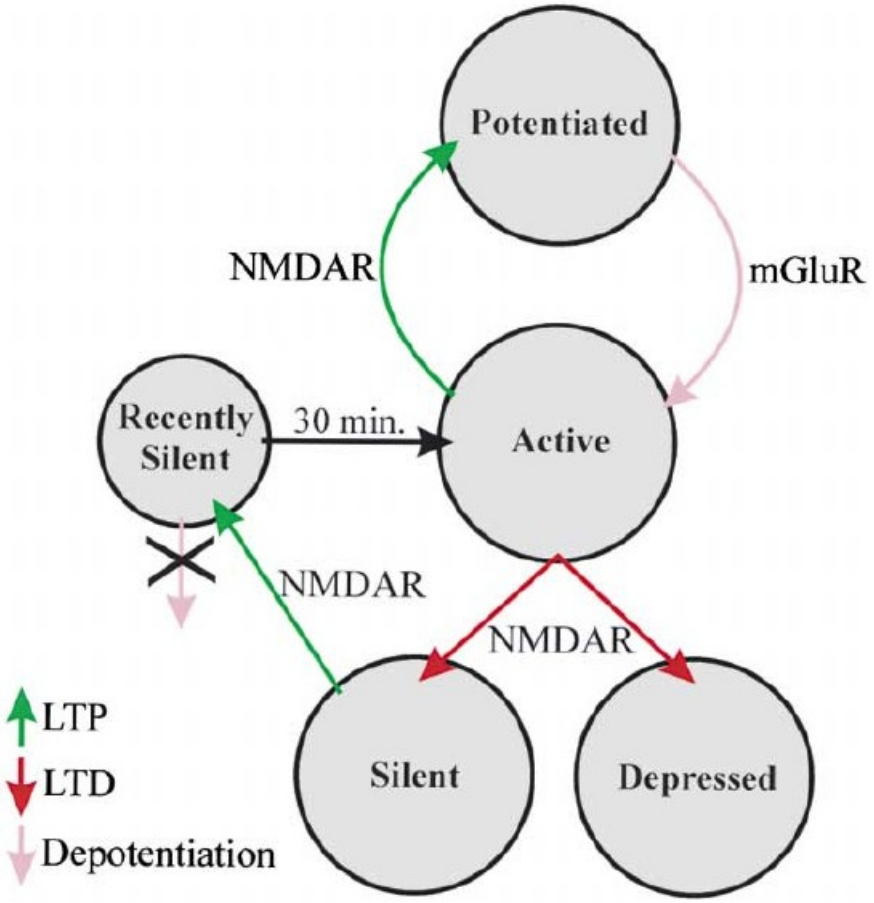


Figure 8. A Model of State-Dependent Plastic Potential of Synaptic Transmission

Metaplasticity

- More recent models are even more complicated
- What are the behavioral manifestations of metaplasticity?
  - Memory can be stored fast fast but it is robust to perturbations and long-lived
  - More specifically: forgetfulness behaves with a power law (Wixted & Ebessen 1991, 1997)

Metaplasticity

- Wixted & Ebessen, 1997

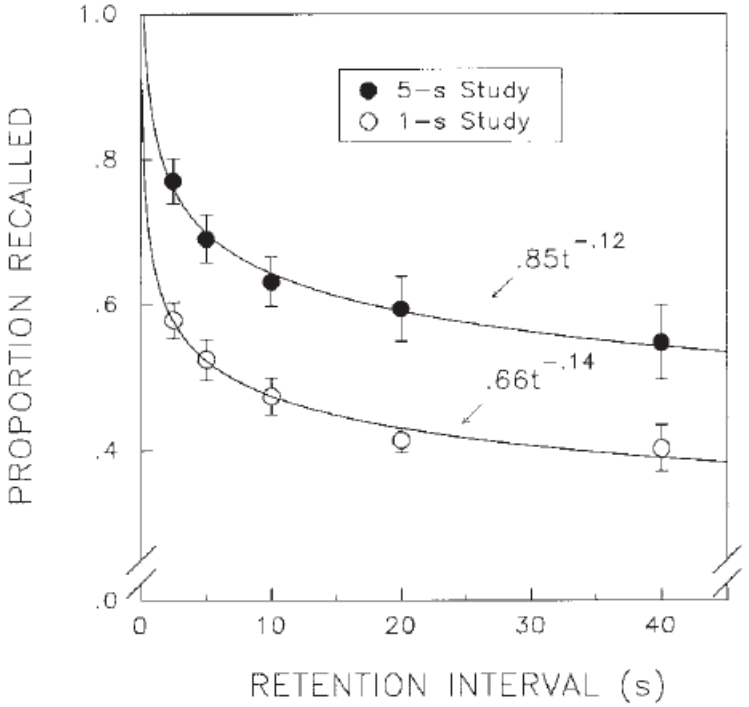


Figure 4. Mean proportion correct recall for the fast and slow conditions of the free recall experiment. The solid curves represent the best fitting power function.

Metaplasticity

- The memory trace depends on a signal-to -noise ratio (SNR)
- Let us consider a system with N_{syn} synapses
- Each time there is an event, a fraction q goes from weak to strong or viceversa
- Event rate is r
- After a time t the probability that a synapse has NOT been modified is $exp(-qrt)$
- Noise is proportional to $N_{syn}^{1/2}$

Metaplasticity

- Signal-to-noise ratio is

$$S/N = qN_{syn}\exp(-qrt)/\sqrt{N_{syn}} = q\sqrt{N_{syn}}\exp(-qrt)$$

(exponential forgetting)

- Time for forget (SNR=1)

$$t_{\max} \approx \ln(q\sqrt{N_{syn}})/(qr)$$

- This number goes one in a time
- This grows very slowly with the number of synapses

Metaplasticity

- How can we have tails that are so long?
- Power laws can be approximated by several exponential functions with well separated time scales
- Cascade models of synaptic plasticity (Fusi et al., 2005)

Metaplasticity

- Cascade models of synaptic plasticity (Fusi et al., 2005)

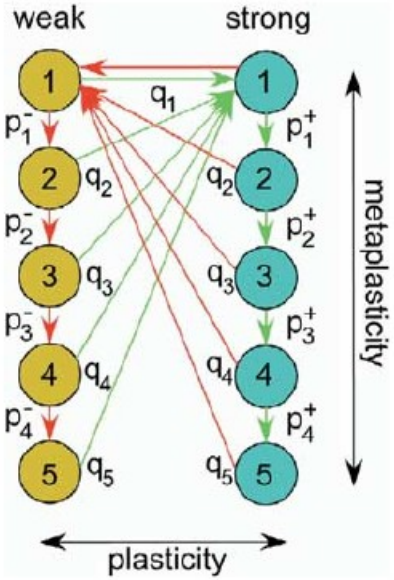


Figure 2. Schematic of a Cascade Model of Synaptic Plasticity

$$q_l = x^{l-1}$$

$$p_l^{+/-} = \frac{x^l}{1 - x}$$

There are two levels of synaptic strength, weak (brown) and strong (turquoise), denoted by + and –. Associated with each of these strengths is a cascade of n states (n = 5 in this example). Transitions between state i of the ± cascade and state 1 of the opposite cascade take place with probability q_i (arrows pointing up and to the left or right), corresponding to conventional synaptic plasticity. Transitions with probabilities $p_i^{\pm}$ link the states within the ± cascades (downward arrows), corresponding to metaplasticity.

- x is a constant <= 1/2 (usually x=1/2)

Metaplasticity

- *Plasticity rules*
- When the conditions for synaptic strengthening are met, a synapse in state i of the weak cascade makes a transition to state 1 of the strong cascade with probability q_i
- When the conditions for synaptic weakening are met, a synapse in state i of the strong cascade makes a transition to state 1 of the weak cascade with probability q_i

Metaplasticity

- *Metaplasticity rules*
- When the conditions for synaptic strengthening are met, a synapse in state i of the strong cascade makes a transition to state i+1 of the strong cascade with probability p_i^+
- When the conditions for synaptic weakening are met, a synapse in state i of the weak cascade makes a transition to state i+1 of the weak cascade with probability p_i^-

Metaplasticity

- With these rules, if strengthening and weakening events occur with the same rate, then all the states are occupied with the same probability
- We choose x=1/2 because it is the largest value compatible with $p_1^{+/-} < 1$
- How long the memories will last?

Metaplasticity

- Numerical simulation:
  - 10,000 synapses
  - Initial states from uniform distribution
  - Memory: half of the synapses have a strengthening event and the other half a weakening event
  - After that synapses have random strengthening and weakening events with rate *f*
  - The memory trace is proportional to the difference between the number of synapses in each condition (weak or strong) with respect to the state after memory storage

Metaplasticity

- Numerical simulation:

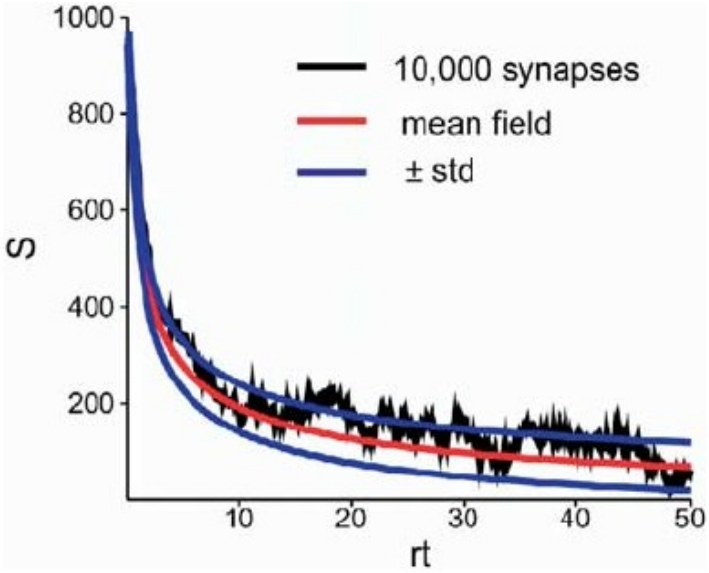


Figure 3. Memory Signal as a Function of Time following Storage of a Memory Trace

The black curve shows the memory signal obtained from simulating 10,000 synapses described by the cascade model of Figure 2, except with ten states per cascade. The red curve is the value obtained from a mean-field calculation, and the blue lines indicate one standard deviation away from this curve.

Metaplasticity

- How does the SNR depends on the number of steps in the cascade?

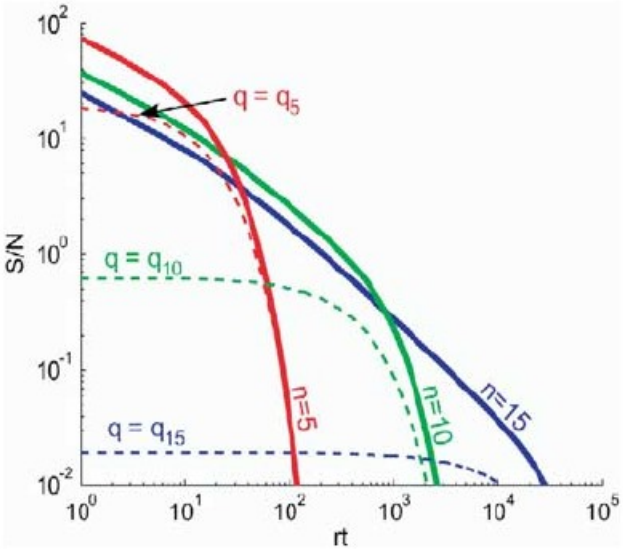
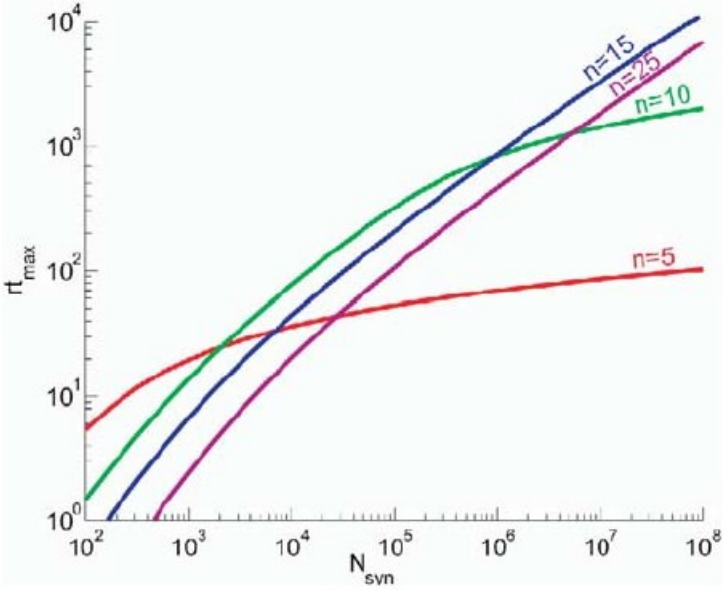


Figure 4. Signal-to-Noise Ratio as a Function of Time

Decay of the signal-to-noise ratios of memory traces stored by cascade models of different sizes (solid curves) and binary models with different transition probabilities (dashed curves). The solid curves for the cascade models initially decay as a power-law, but this changes to an exponential decay at a time determined by the smallest transition probability q_n in the model. Increasing n, and hence decreasing $q_n = 2^{-n+1}$, expands the range over which the power-law applies. The binary models shown have transition probabilities set to the minimum transition probability in the n = 5, 10, and 15 cascade models (red, green, and blue curves, respectively). All these curves correspond to memory storage with 10^5 synapses.

Metaplasticity

- How does memory retention depends on the number of synapses?

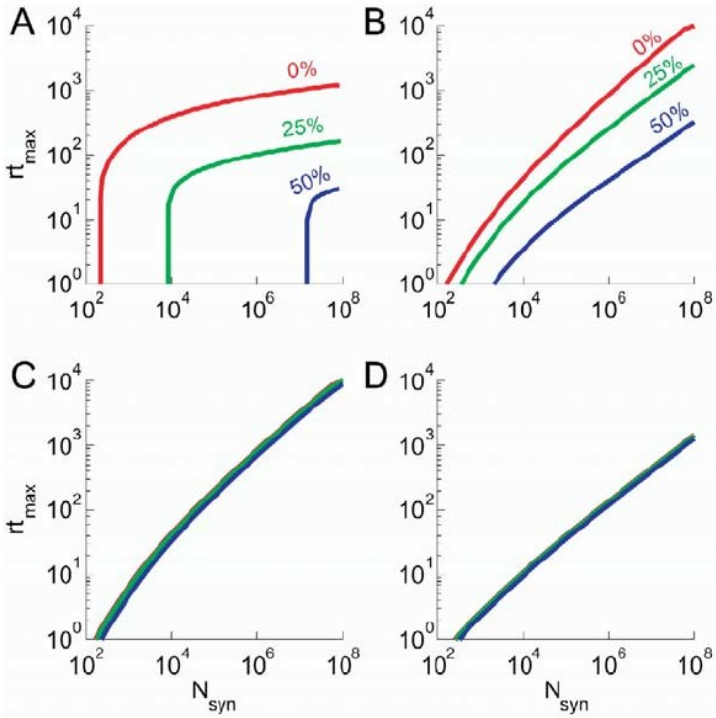


Memory lifetime (in units of 1/*r*) for different size cascade models versus the number of synapses used in storage. The optimal number of cascade states depends on the number of synapses being used for memory storage.

Metaplasticity

- What happens if the balance between potentiation and depression is broken?

Default parameters Optimized parameters



Non-cascade

Cascade

Reinforcement learning

- Up to now the environment is a source of stimuli (or the object of motor actions)
- But learning must involve some feedback about whether the behavior generated by the learning process is giving “right” or “wrong” results
- Reinforcement learning (RL) is the problem faced by an agent that learns behavior through trial-and-error interactions with its environment

Reinforcement learning

- 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

Reinforcement learning

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

- **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.

Reinforcement learning

- The system is defined by two functions
  - $\delta(s_t,a_t)=s_{t+1}$: transition function
  - $r(s_t,a_t)$: reward function. Sometimes written as $r(s_t,s_{t+1})$
  - If the environment is stochastic we could have $p(s_t,a_{t,}s_{t+1})$: probability of going to state s_{t+1}

  - The model is not necessarily known to the agent
  - For agents with long lifetime, a discount factor is useful
  - Future rewards are less valuable

Reinforcement learning

- The policy: is a *complete* mapping from *every* state to the action to be taken in that state
- The objective of reinforcement learning (RL) is to try to find an *optimal policy*
- The optimal policy is the one that maximizes the *cumulative discounted reward*

Reinforcement learning

- 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.

Reinforcement learning

- Temporal difference algorithm (Sutton, 1984)
- The agent first estimates the *utility* of a state. This is the sum of the rewards beginning the path from this state
- The from each initial condition it moves to the neighboring state with the maximum utility

How to estimate the utility of a state?

Reinforcement learning

- Adjust the utility of a state based on the immediate reward and the utility of the next state

- $U(s) + \alpha(r(s,s') + \gamma U(s') - U(s)) \rightarrow U(s)$
  - $r(s,s')$: reward of $s \rightarrow s'$
  - $U(s')$ utility of successor state
  - α: learning rate
  - γ: discount factor

Reinforcement learning

- 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 δ(s, a) = s′ and reward r(s,s′) for going from state s to state s′, do:

$$U(s) + \alpha(r(s,s') + \gamma U(s') - U(s)) \rightarrow U(s)$$

- 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***).

Reinforcement learning

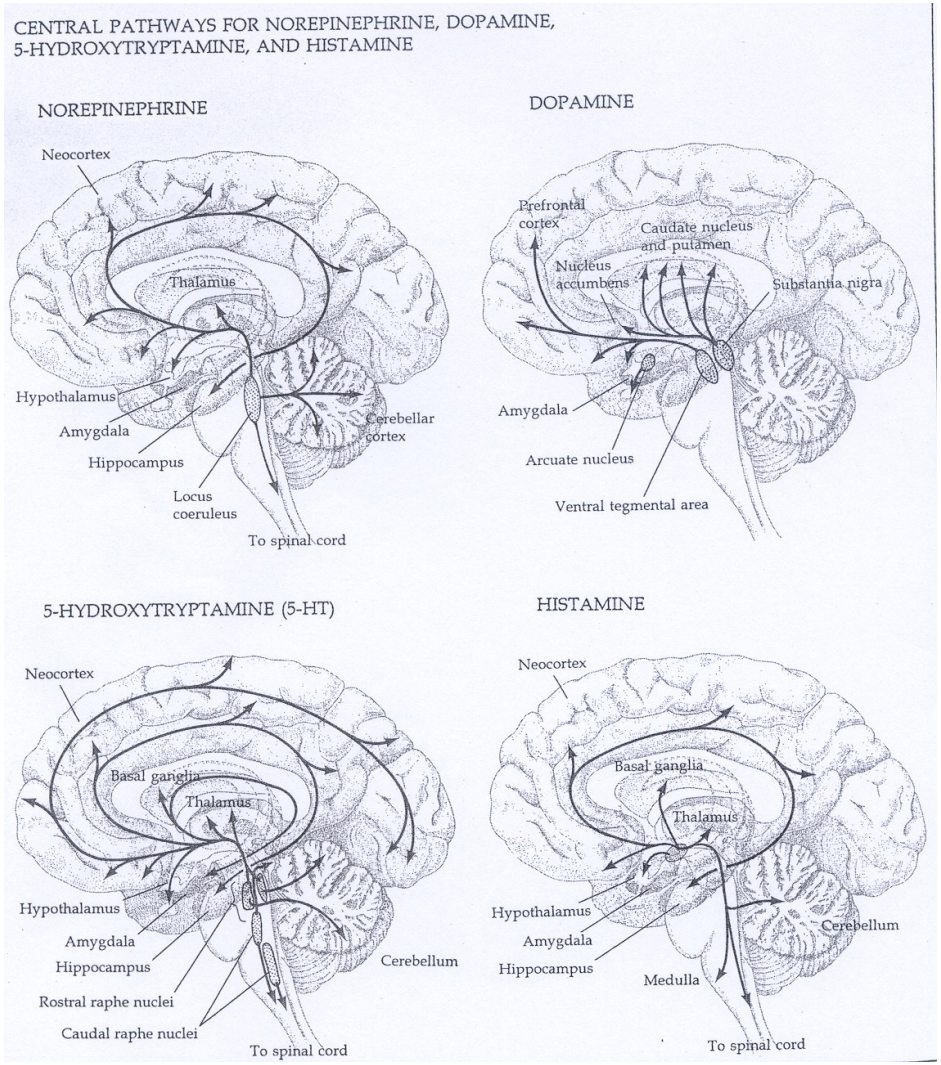
- This is a *passive* algorithm
- An ***active learner*** must consider what actions to take, what their outcomes may be, and how they affect the rewards achieved. An active learner takes actions while it learns. Only an active learner can handle a dynamic environment
  (Watkins, 1992)

Reinforcement learning

- RL needs a signal the satisfies the following criteria:
  - Responds to affective contingencies
  - Affects learning of predictions and actions
  - Is essentially scalar
  - Broadcasts their information multimodally

- Neuromodulators verify most of these points:
  - Respond to reinforcers and surprise
  - Are known to affect synaptic plasticity
  - Come from small mid-brain nuclei
  - Have extensive arborization throughout the brain

Reinforcement learning



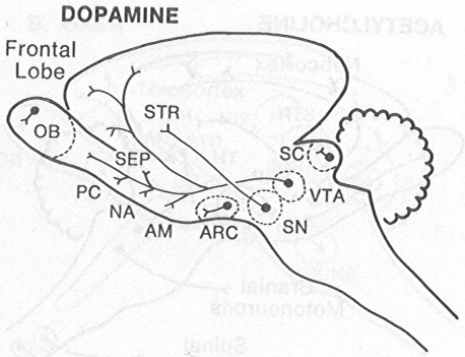


Fig. 24.10 Distribution of dopamine-containing neurons in the rat brain. For abbreviations, see legend to Fig. 24.9.

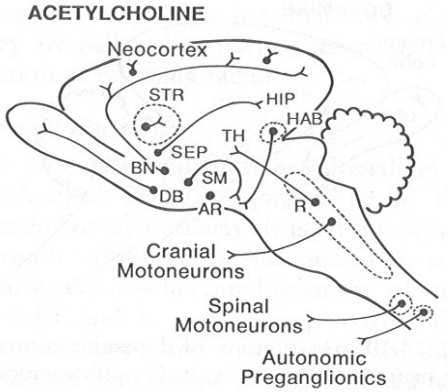
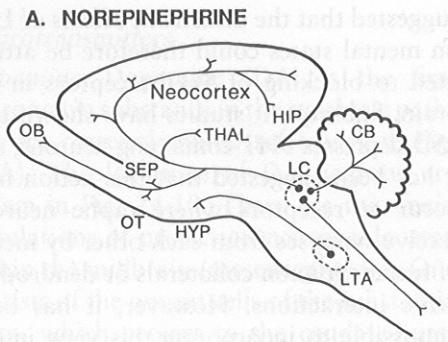


Fig. 24.11 Distribution of cholinergic cell groups and their projections in the rat brain. For abbreviations, see legend to Fig. 24.9.



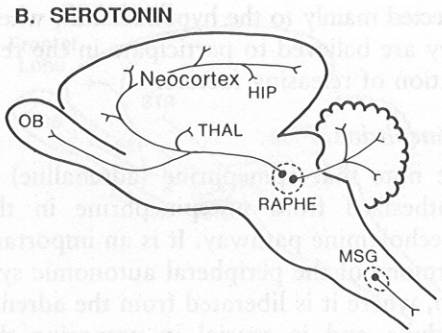
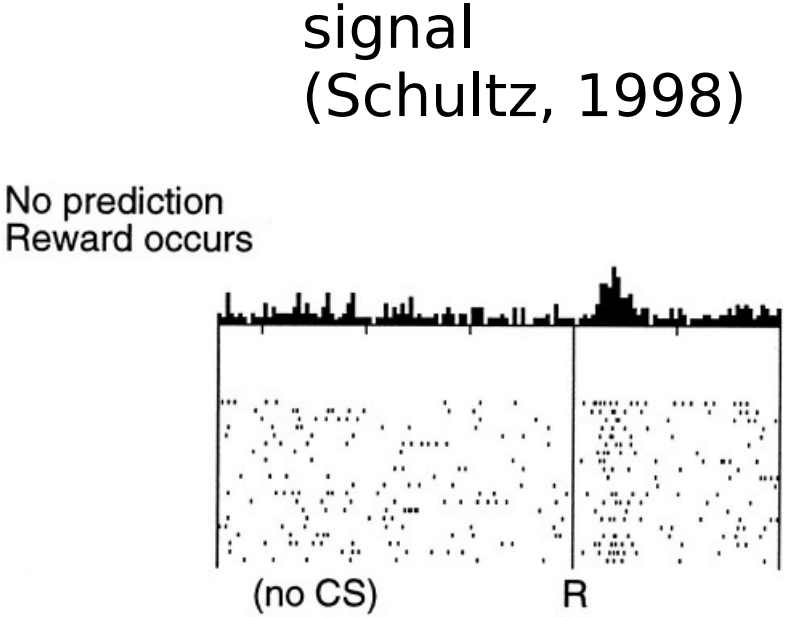
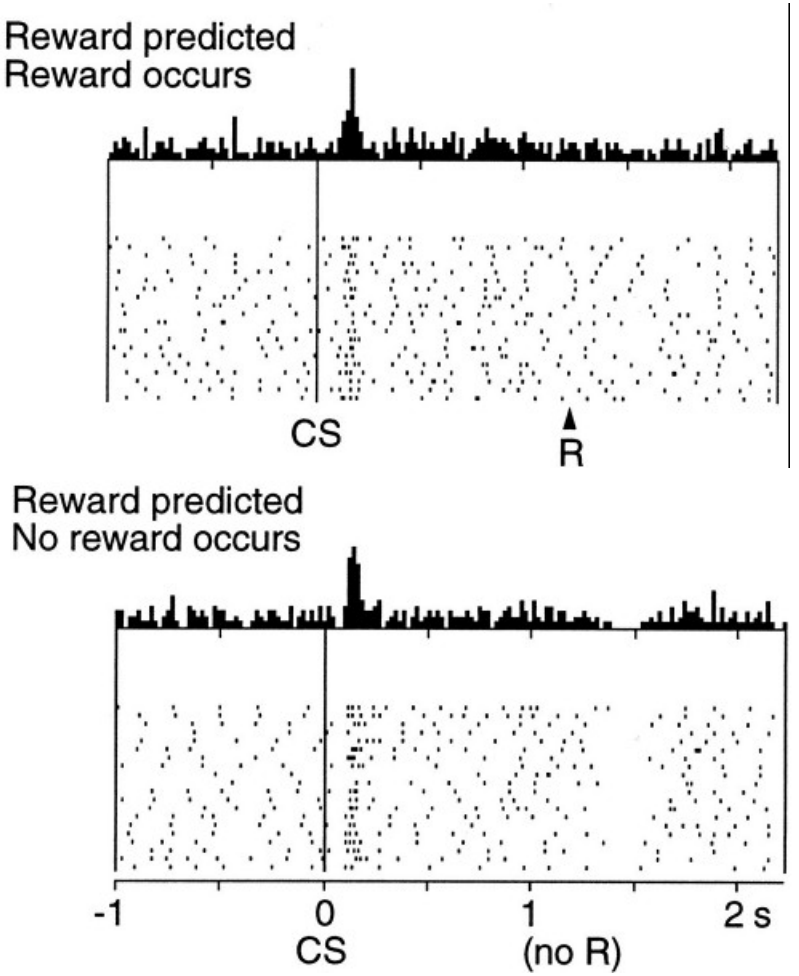


Fig. 24.9 Maps of the distribution of cell groups containing different neurotransmitters in the mammalian brain. A sagittal view of the rat brain is shown for this and succeeding figures. **A.** Distribution of norepinephrine-containing neurons and their axonal projections. **B.** Distribution of serotonin-containing 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.

Reinforcement learning

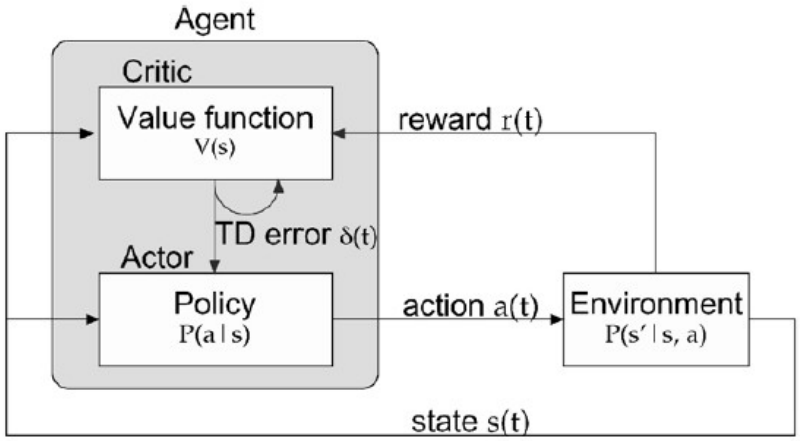
- Dopamine generates a *predictive* reward signal
(Schultz, 1998)





Reinforcement learning

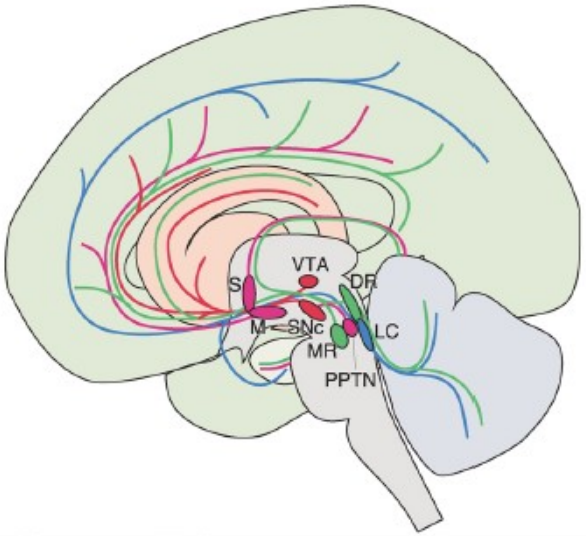
- And what are the other neuromodulators good for? (Doja, *Metalearning and neuromodulation,* 2002)



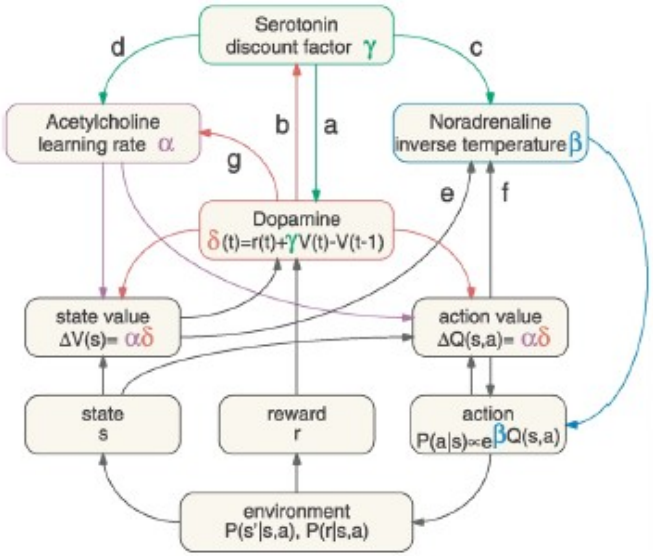
1. Dopamine signals the TD error δ.
2. Serotonin controls the discount factor γ.
3. Noradrenaline controls the inverse temperature β.
4. Acetylcholine controls the learning rate α.

Reinforcement learning

- And what are the other neuromodulators good for? (Doja, *Metalearning and neuromodulation*, 2002)



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



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.

Reinforcement learning

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

Prediction error: $\delta = r + V(s_{t+1}) - V(s_t)$

$V(s_{t+1}) = V(s_t) + \alpha\delta$

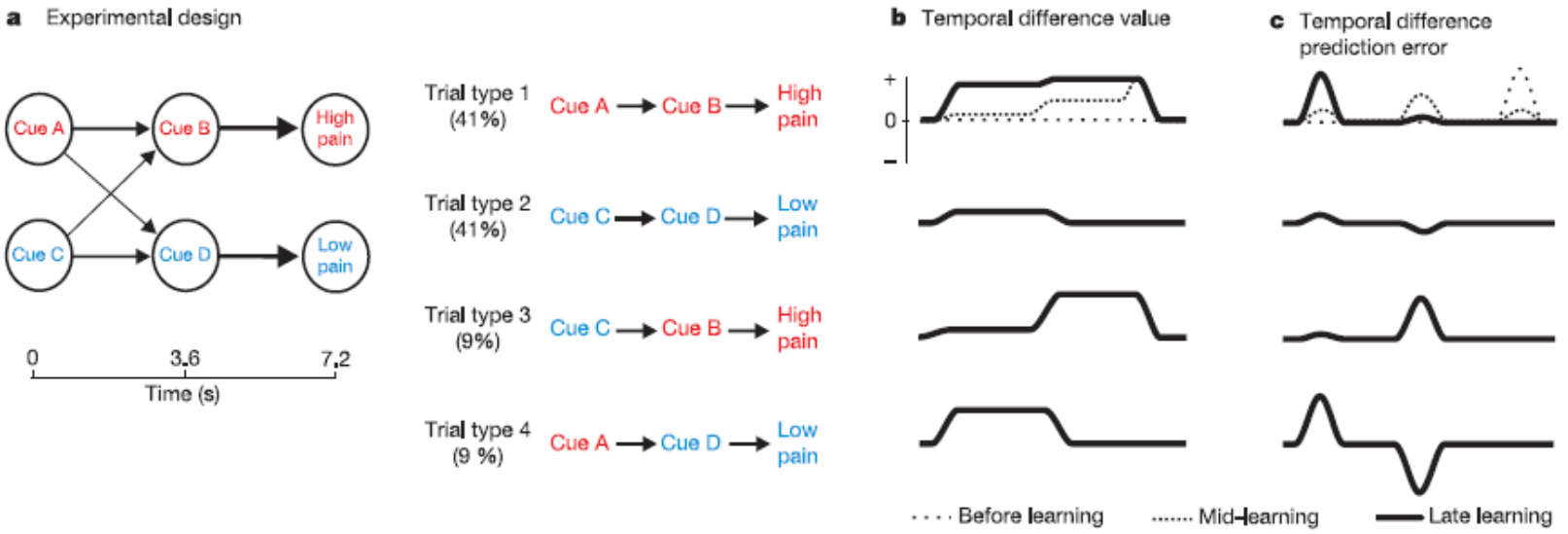


Figure 1 Experimental design and temporal difference model. **a**, The experimental design expressed as a Markov chain, giving four separate trial types. **b**, Temporal difference value. As learning proceeds, earlier cues learn to make accurate value predictions (that is, weighted averages of the final 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.)

Reinforcement learning

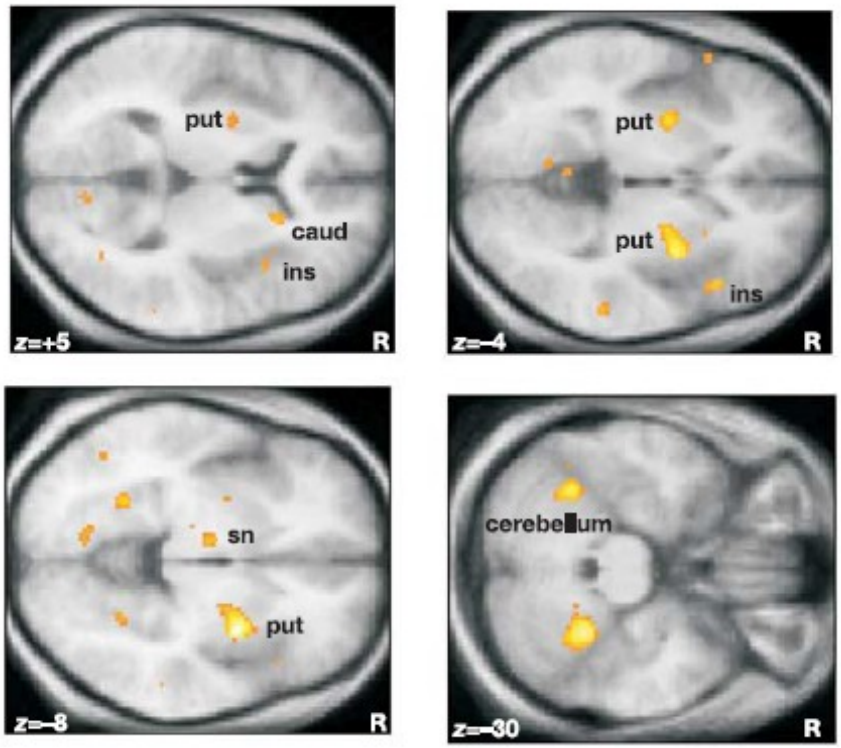


Figure 2 Temporal difference prediction error (statistical parametric maps). Areas coloured yellow/orange show significant correlation with the temporal difference prediction error. Yellow represents the greatest correlation. Peak activations (MNI coordinates and statistical *z* scores) are: right ventral 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.

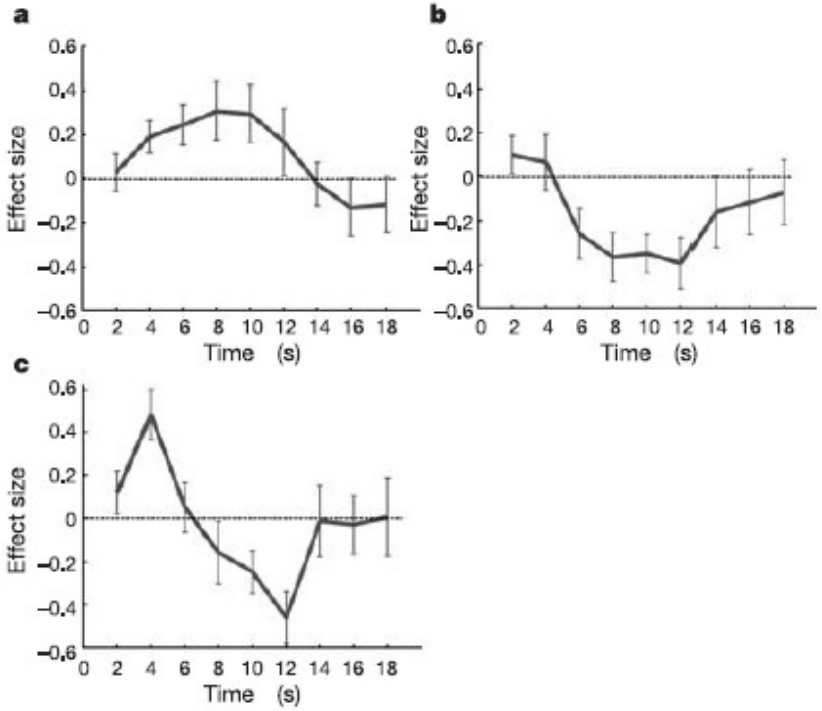


Figure 3 Temporal difference prediction error (impulse responses). Time course of the impulse response (±s.e.m.) to higher-order prediction error in the right ventral putamen. **a**, Positive prediction error (contrast of trial types 3 and 2). **b**, Negative prediction error (contrast of trial types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