10703 Deep Reinforcement Learning

Reinforcement Learning in Humans and Animals

Tom Mitchell

October 29, 2018

Reading: Barto & Sutton Chapter 15

ML
MACHINE LEARNING
DEPARTMENT

Outline

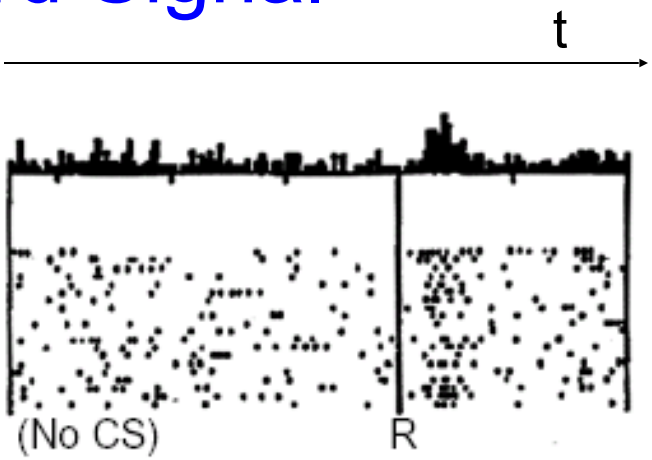
- RL in primates
- RL in humans
- Error signals and predictive coding

Reward based learning in primates

Dopamine As Reward Signal

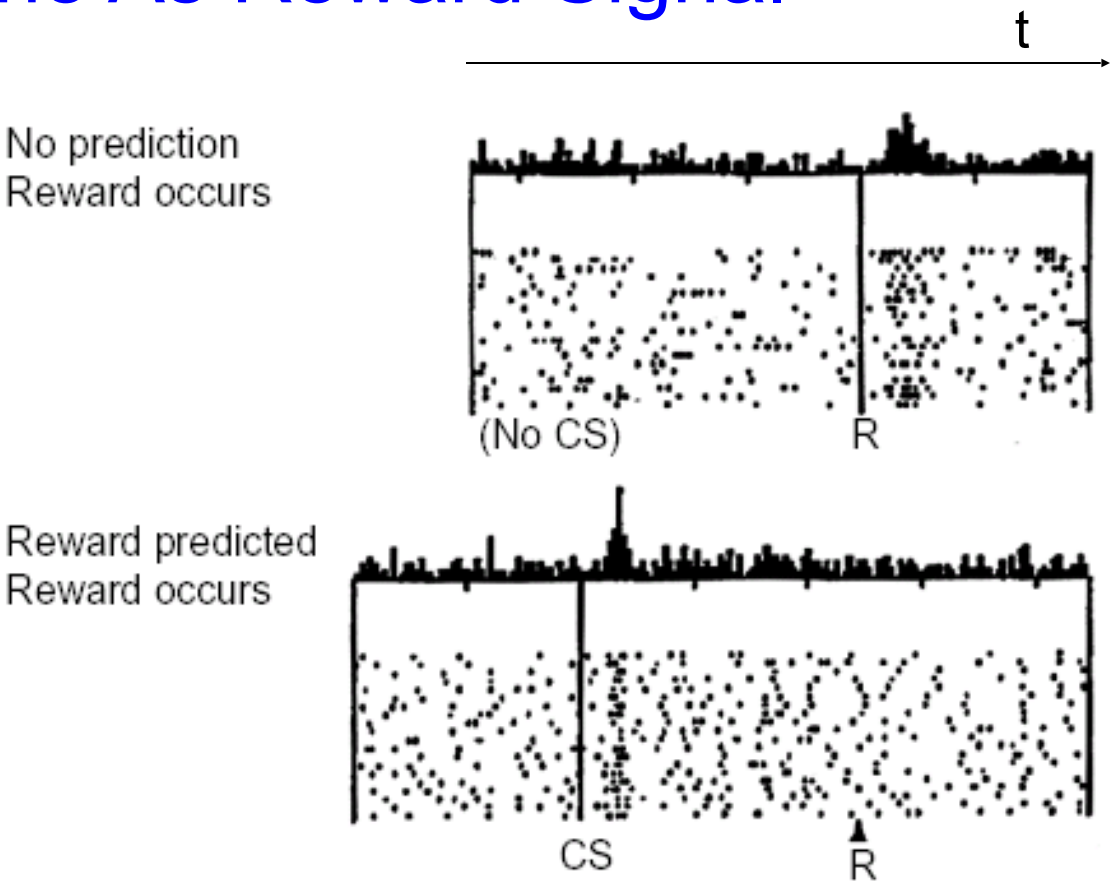
[Schultz et al., *Science*, 1997]

No prediction
Reward occurs



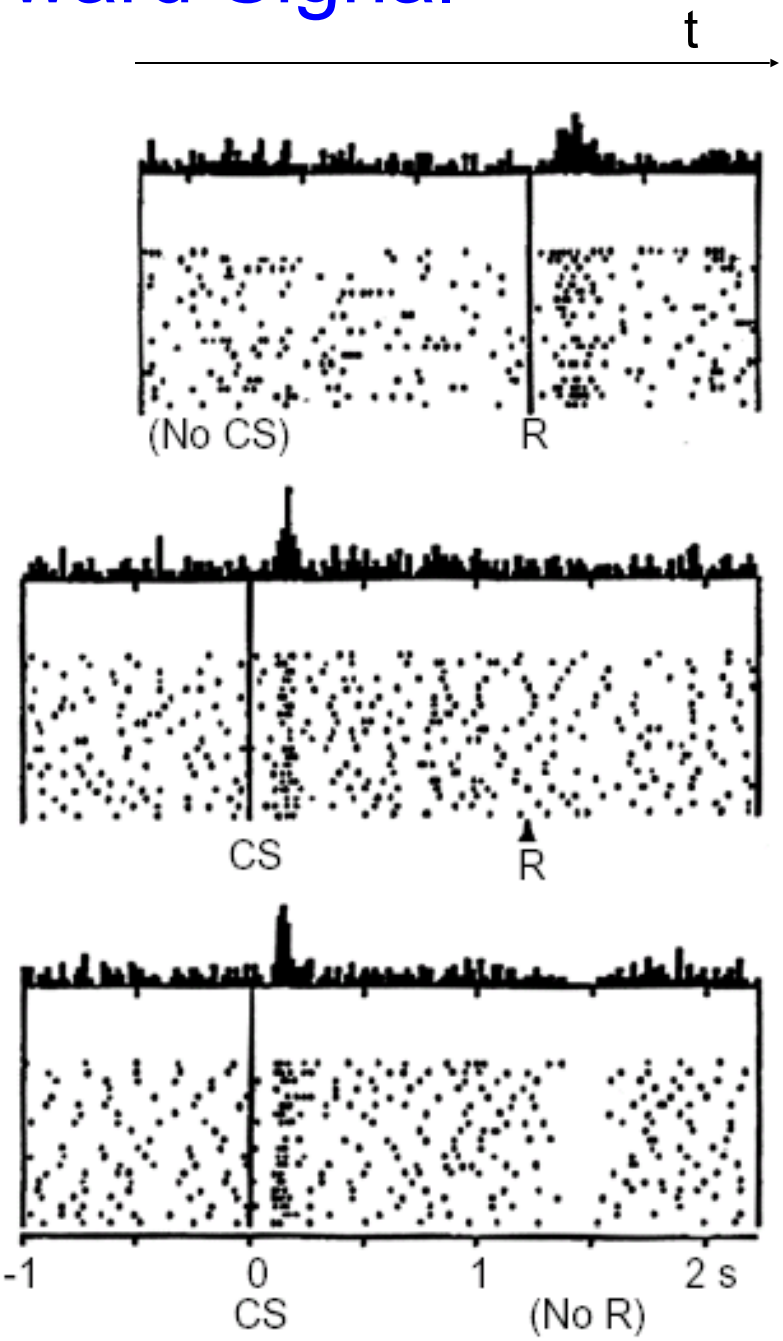
Dopamine As Reward Signal

[Schultz et al., *Science*, 1997]



Dopamine As Reward Signal

[Schultz et al., *Science*, 1997]



$$\text{error} = r_t + \gamma V(s_{t+1}) - V(s_t)$$

Reward based learning in humans

RL Models for Human Learning

[Seymore et al., Nature 2004]

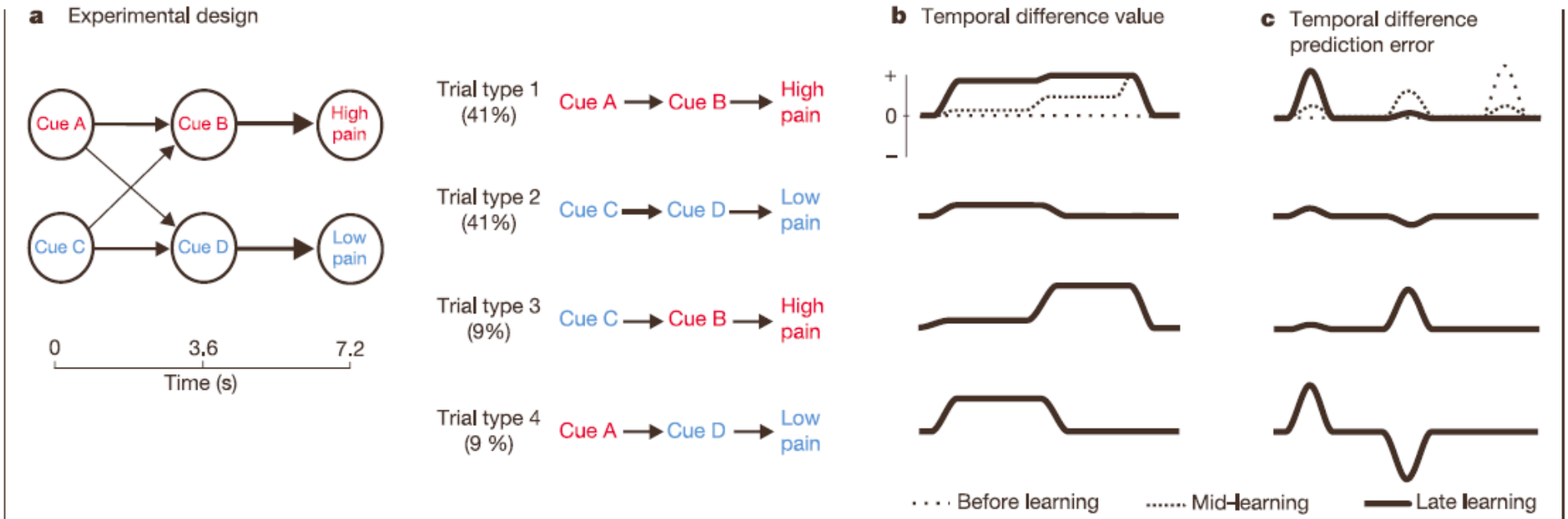
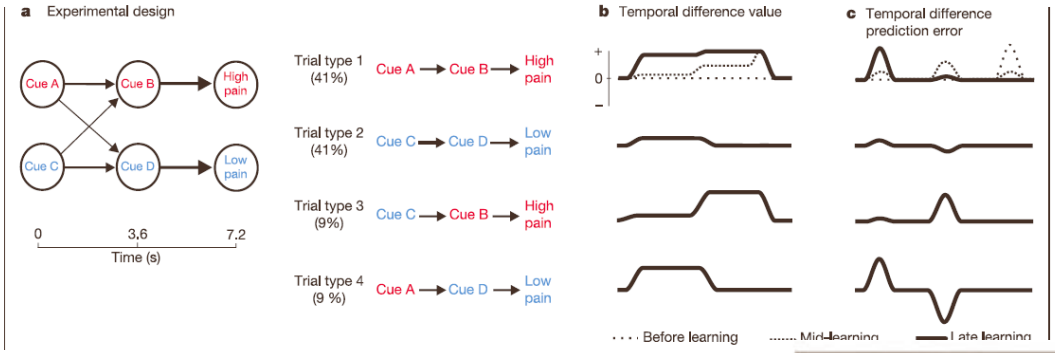


Figure 1 Experimental design and temporal difference model. **a**, The experimental design expressed as a Markov chain, giving four separate trial types. **b**, Temporal difference value. As learning proceeds, earlier cues learn to make accurate value predictions (that is, weighted averages of the final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



[Seymore et al., Nature 2004]

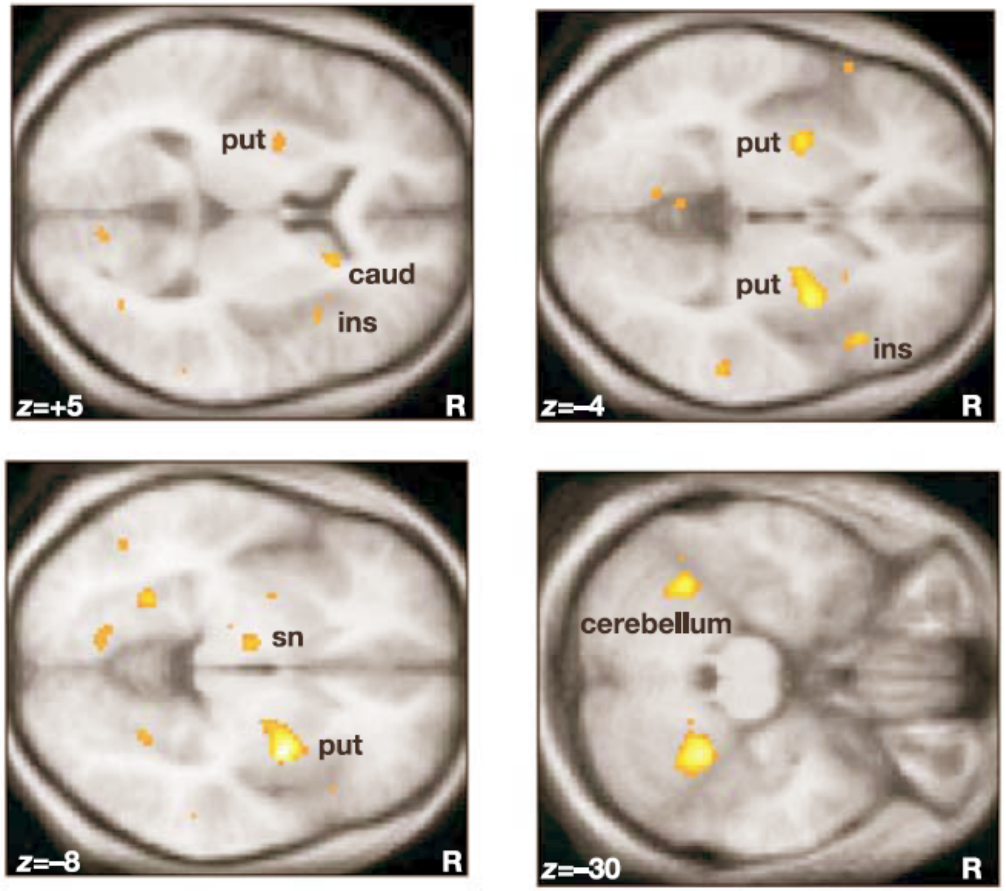


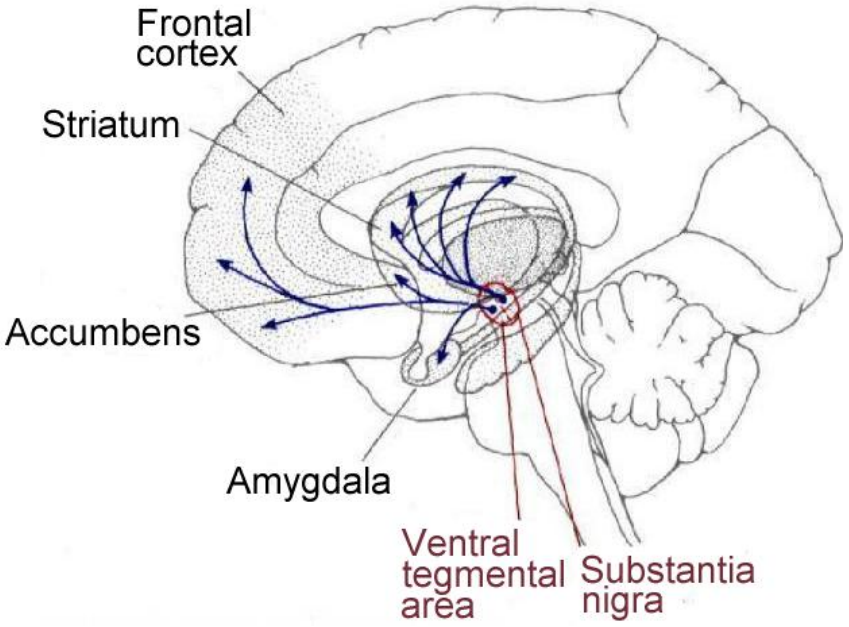
Figure 2 Temporal difference prediction error (statistical parametric maps). Areas coloured yellow/orange show significant correlation with the temporal difference prediction error. Yellow represents the greatest correlation. Peak activations (MNI

One Theory of RL in the Brain

from [Nieuwenhuis et al.]

- Basal ganglia monitor events, predict future rewards
- When prediction revised upward (downward), causes increase (decrease) in activity of midbrain dopaminergic neurons, influencing ACC

- This dopamine-based activation somehow results in revising the reward prediction function. Possibly through direct influence on Basal ganglia, and via prefrontal cortex



Axonal arbor of a single neuron producing dopamine as a neurotransmitter. These axons make synaptic contacts with a huge number of dendrites of neurons in targeted brain areas.

Adapted from *The Journal of Neuroscience*, Matsuda, Furuta, Nakamura, Hioki, Fujiyama,

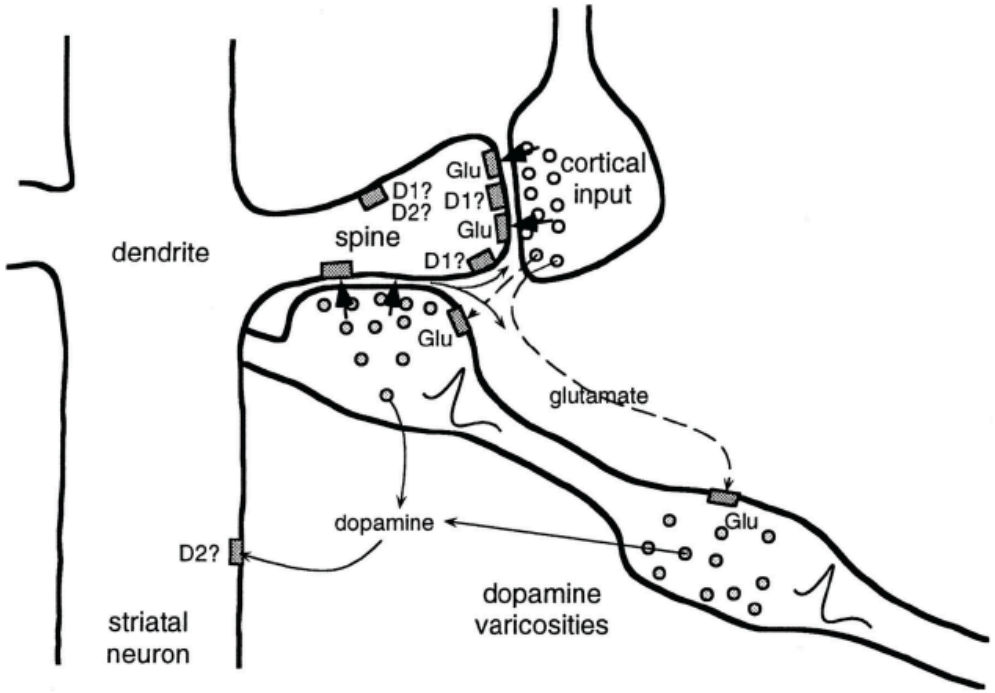
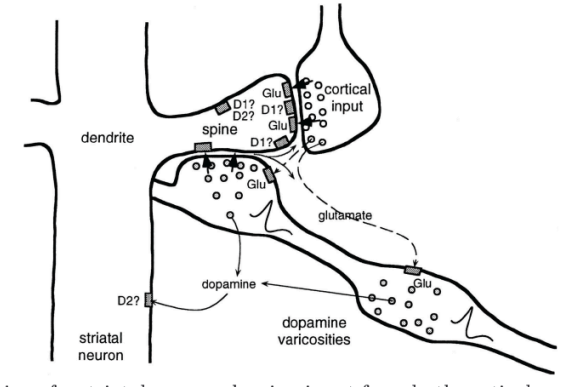


Figure 15.1: Spine of a striatal neuron showing input from both cortical and dopamine neurons. Axons of cortical neurons influence striatal neurons via corticostriatal synapses releasing the neurotransmitter glutamate at the tips of spines covering the dendrites of striatal neurons. An axon of a VTA or SNpc dopamine neuron is shown passing by the spine (from the lower right). "Dopamine varicosities" on this axon release dopamine at or near the spine stem, in an arrangement that brings together presynaptic input from cortex, postsynaptic activity of the striatal neuron, and dopamine, making it possible that several types of learning rules govern the

Neuron Level Learning Mechanisms

- Hebbian learning
  - fire together → wire together
- Spike Timing Dependent Plasticity (STDP)
  - if incoming neuron fires *before* outgoing then *strengthen* connection
  - if incoming neuron fires *after* outgoing then *weaken* connection



- Reward modulated STDP
  - less understood
  - in some neurons, appears STDP occurs only if neuromodulator (e.g., dopamine) activity follows firing within time up to 10 sec

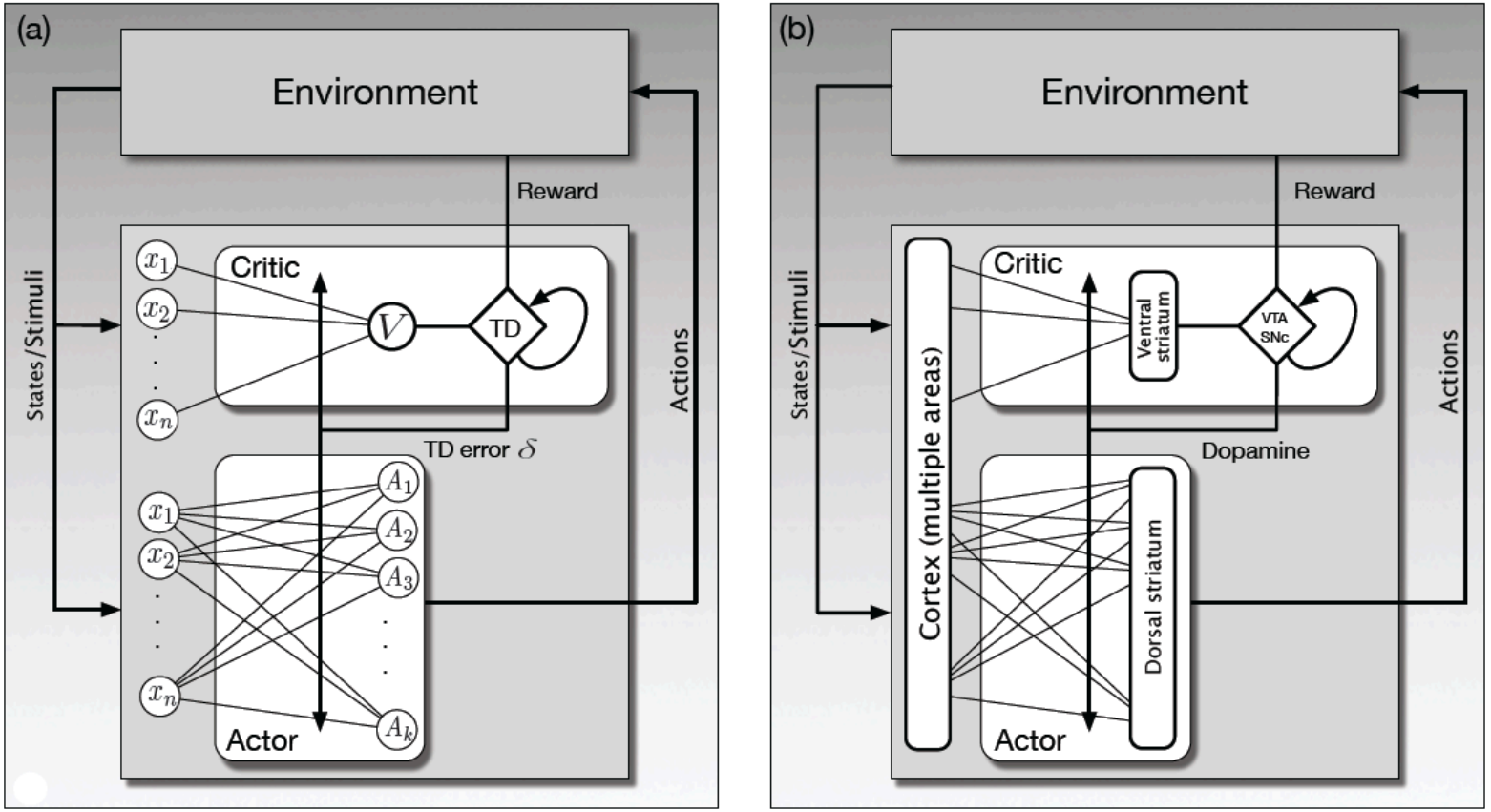


Figure 15.5: Actor–critic ANN and a hypothetical neural implementation. a) Actor–critic algorithm as an ANN. The actor adjusts a policy based on the TD error δ it receives from the critic; the critic adjusts state-value parameters using the same δ. The critic produces a TD error from the reward signal, R, and the current change in its estimate of state values. The actor does not have direct access to the reward signal, and the critic does not have direct access to the action. b) Hypothetical neural implementation of an actor–critic algorithm. The actor and the value-learning part of the critic are respectively placed in the dorsal and ventral subdivisions of the striatum. The TD error is transmitted by dopamine neurons located in the VTA and SNpc to modulate changes in synaptic efficacies of input from cortical areas to the ventral and dorsal striatum. Adapted from Frontiers in Neuroscience, vol. 2(1), 2008, Y. Takahashi, G. Schoenbaum, and Y. Niv, Silencing the critics: Understanding the effects of cocaine sensitization on dorsolateral and ventral striatum in the context of an Actor/Critic model.

Summary: Temporal Difference ML Model Predicts Dopaminergic Neuron Acitivity during Learning

- Evidence now of neural reward signals from
  - Direct neural recordings in monkeys
  - fMRI in humans (1 mm spatial resolution)
  - EEG in humans (1-10 msec temporal resolution)
- Dopaminergic responses encode Temporal Difference error
- Some differences, and efforts to refine the model
  - How/where is the value function encoded in the brain?
  - Study timing (e.g., basal ganglia learns faster than PFC ?)
  - Role of prior knowledge, rehearsal of experience, multi-task learning?

Predictive Coding

Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive-field effects

Rajesh P. N. Rao[1] and Dana H. Ballard[2]

[1] *The Salk Institute, Sloan Center for Theoretical Neurobiology and Computational Neurobiology Laboratory, 10010 N. Torrey Pines Road, La Jolla, California 92037, USA*

[2] *Department of Computer Science, University of Rochester, Rochester, New York 14627-0226, USA*

Correspondence should be addressed to R.P.N.R. (rao@salk.edu)

We describe a model of visual processing in which feedback connections from a higher- to a lower-order visual cortical area carry predictions of lower-level neural activities, whereas the feedforward connections carry the residual errors between the predictions and the actual lower-level activities. When exposed to natural images, a hierarchical network of model neurons implementing such a model developed simple-cell-like receptive fields. A subset of neurons responsible for carrying the residual errors showed endstopping and other extra-classical receptive-field effects. These results suggest that rather than being exclusively feedforward phenomena, nonclassical surround effects in the visual cortex may also result from cortico-cortical feedback as a consequence of the visual system using an efficient hierarchical strategy for encoding natural images.

ML
MACHINE LEARNING
DEPARTMENT

[Rao & Ballard, *Nature,* 1999]

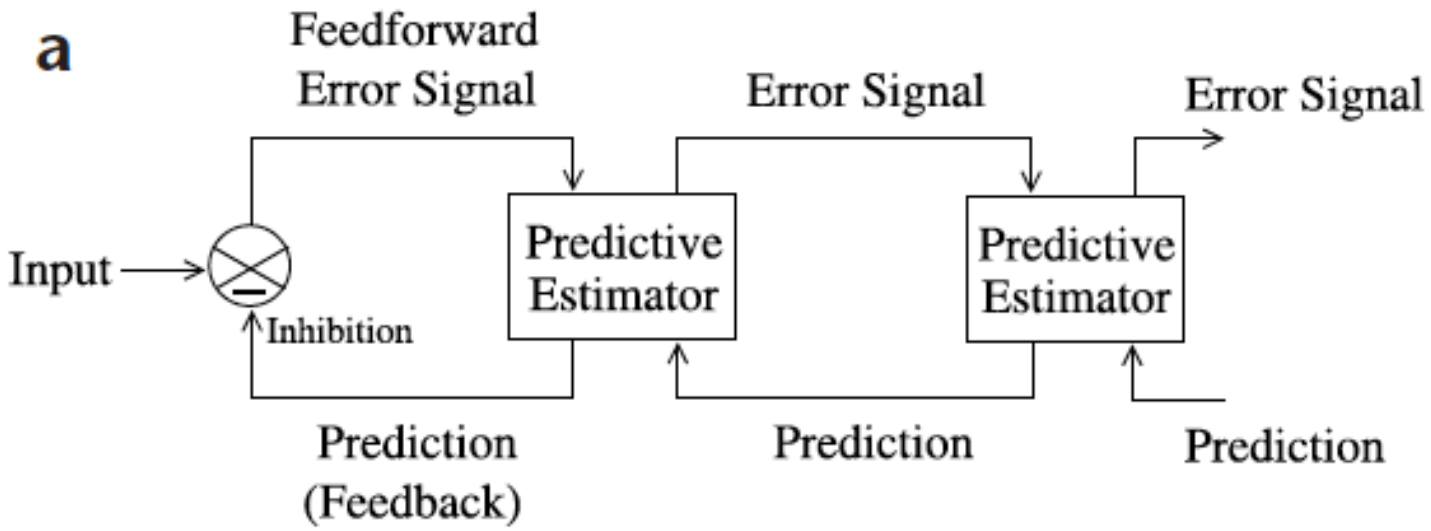
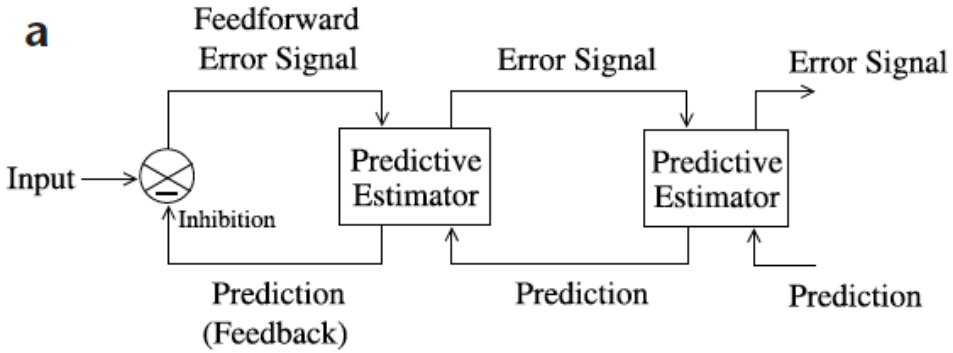
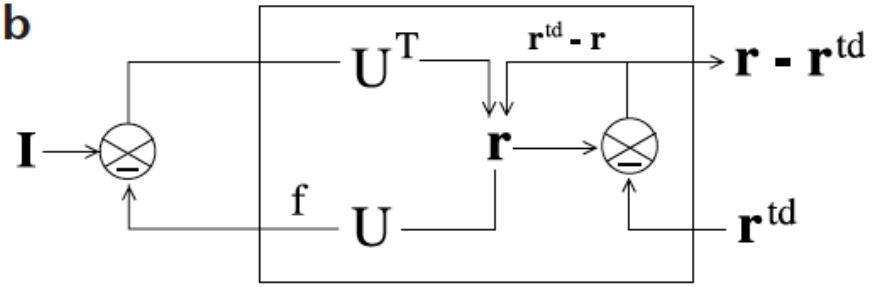


Fig. 1. Hierarchical network for predictive coding. **(a)** General architecture of the hierarchical predictive coding model. At each hierarchical level, feedback pathways carry predictions of neural activity at the lower level, whereas feedforward pathways carry residual errors between the predictions and actual neural activity. These errors are used by the predictive estimator (PE) at each level to correct its current estimate of the input signal and generate the next prediction. **(b)** Components of a PE module, composed of

[Rao & Ballard, 1999]





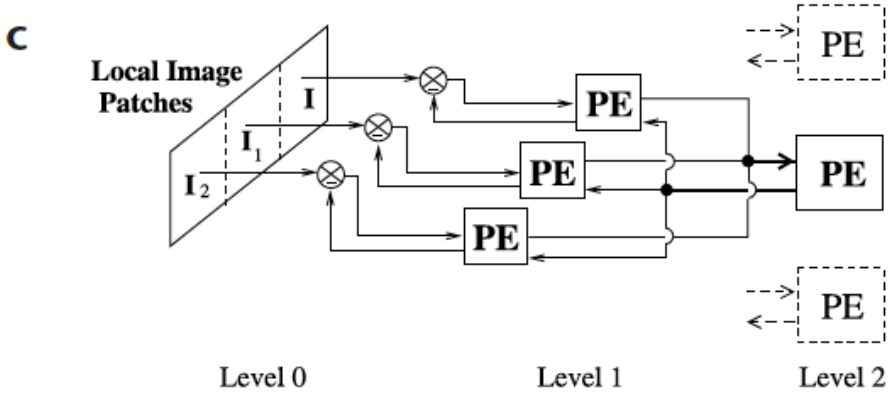
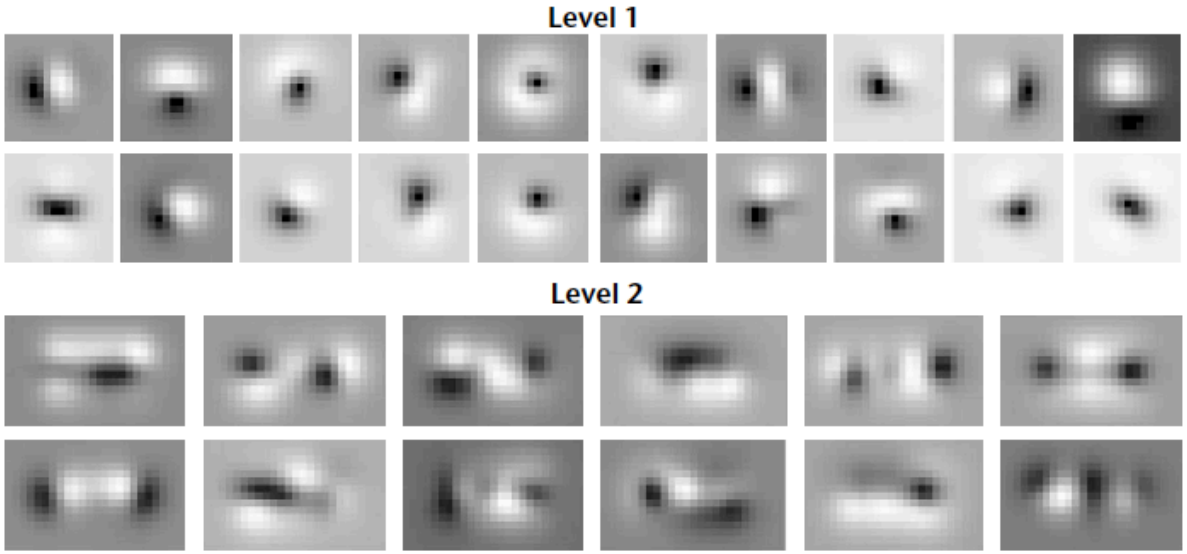


Fig. 1. Hierarchical network for predictive coding. **(a)** General architecture of the hierarchical predictive coding model. At each hierarchical level, feedback pathways carry predictions of neural activity at the lower level, whereas feedforward pathways carry residual errors between the predictions and actual neural activity. These errors are used by the predictive estimator (PE) at each level to correct its current estimate of the input signal and generate the next prediction. **(b)** Components of a PE module, composed of feedforward neurons encoding the synaptic weights U^T, neurons whose responses $\mathbf{r}$ maintain the current estimate of the input signal, feedback neurons encoding U and conveying the prediction $f(U\mathbf{r})$ to the lower level, and error-detecting neurons computing the difference $(\mathbf{r} - \mathbf{r}^{td})$ between the current estimate $\mathbf{r}$ and its top-down prediction $\mathbf{r}^{td}$ from a higher level. **(c)** A three-level hierarchical network used in the simulations. An input image was analyzed by three level-1 PE modules, each predicting its own local image patch. The responses $\mathbf{r}$ of all three level-1 modules were input to the level-2 module. This convergence of lower-level inputs to a higher-level module increases receptive-field size of neurons as one ascends the hierarchy, with the receptive field at the highest level spanning the entire input image.



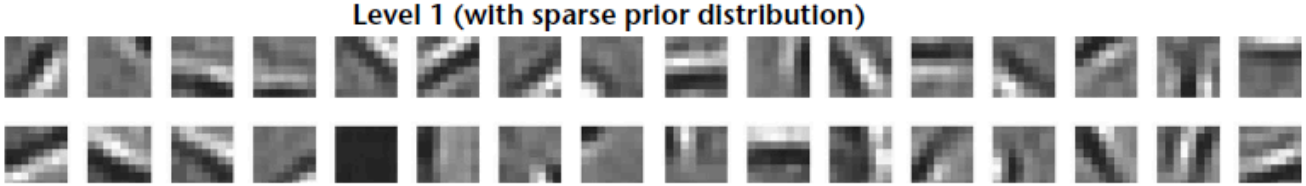


Fig. 2. Receptive fields of feedforward model neurons after training on natural images. **(a)** Five natural images used for training the three-level hierarchical network of Fig. 1c (Methods). The two upper boxes in the bottom right corner show relative sizes (16 x 16 and 16 x 26 pixels) of level-1 and level-2 receptive fields respectively. **(b)** Learned synaptic weights (RF weighting profiles) of 20 of the 32 feedforward model neurons in the level-1 module analyzing the central image region. Flanking image regions were analyzed by two other level-1 modules (**Fig. 1c**), each with 32 feedforward model neurons (Methods). Values for these synapses, which form rows of the matrix U^T, can be positive (excitatory, bright regions) or negative (inhibitory, dark regions). These RF profiles resemble classical oriented-edge/bar detectors characteristic of simple cells[2]. **(c)** RF profiles of 12 of the 128 level-2 feedforward model neurons. **(d)** Localized RF profiles resembling Gabor wavelets obtained by using a sigmoidal nonlinearity in the generative model, along with a sparse kurtotic prior distribution for the network activities (Methods). All 32 level-1 feedforward model neurons are shown; Gaussian windowing of inputs (as in b) was not necessary in this case.